Geometry-Aware Statistical Learning and Causal Inference for Complex Observational Data

Yikun Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Yen-Chi Chen, Chair
Thomas S. Richardson
Zaid Harchaoui

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Geometry-Aware Statistical Learning and Causal Inference for Complex Observational Data

Yikun Zhang

Chair of the Supervisory Committee:
Yen-Chi Chen
Department of Statistics

Motivated by cosmic web detection and related problems in astronomy, this dissertation develops statistically principled and computationally efficient methods for analyzing complex observational data. Three major challenges arise from such data: (i) nonlinear geometric structure in the observational domain, (ii) high dimensionality and missingness in measured features, and (iii) confounding and other causal assumption violations when interpreting scientific relationships from observational studies.

The first part of the dissertation addresses the geometric challenge in (i). In Chapter 2, we introduce directional density ridges as statistical surrogates for representing low-dimensional high-density structures on the sphere. These surrogates provide geometry-aware models for cosmic nodes and filaments while respecting the spherical geometry of sky observations. We establish stability and statistical consistency guarantees for directional density ridge estimation based on kernel smoothing. In Chapter 3, we derive practical algorithms for recovering directional density ridges from data and study their linear convergence properties. In Chapter 4, we apply the statistical and computational framework of directional density ridges to galaxy and quasar observations from the Sloan Digital Sky Survey IV, producing a tomographic cosmic web catalog.

The second part of the dissertation develops statistical inference methods for the high-

dimensional, incomplete, and confounded observational settings described in (ii) and (iii). In Chapter 5, we study high-dimensional linear regression with missing outcomes and propose an efficient debiased inference procedure for low-dimensional functionals of the regression function. When the missingness probabilities are consistently estimated at the observed data points, our proposed estimator is asymptotically normal and semiparametrically efficient among all asymptotically linear estimators. We apply this method to study associations between galactic stellar mass and nearby cosmic web structures. In Chapter 6, we turn from association analysis to causal inference for continuous treatments in observational studies, focusing on settings where the usual positivity condition may fail. We propose novel identification and estimation strategies for the dose-response curve and its derivative under an additive structural model for the counterfactual outcome. Using tools from nonparametric set estimation, we construct inverse probability weighted and doubly robust estimators that remain unbiased under certain positivity violations. We further introduce an extended estimand of the classical dose-response curve based on the nearest feasible treatment policy regime, which adapts to the geometry of the conditional treatment support and accounts for treatment-level heterogeneity. Across these causal settings, we establish asymptotic normality at standard nonparametric rates of convergence, enabling valid statistical inference. We illustrate the proposed causal framework with an application to the IllustrisTNG simulation data, studying the effect of local environment on galactic star formation rate and providing a new causal perspective on the "Nature versus Nurture" debate in galaxy evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

When this eight-year journey at the University of Washington (UW) finally comes to an end, I am heartily grateful to all the people who have guided me, supported me in my studies and life in the U.S. and Shanghai during my two gap years, and accompanied me through this long, memorable, and rewarding adventure.

First and foremost, I extend my deepest gratitude to my advisor, Prof. Yen-Chi Chen, for his guidance, support, and mentorship throughout my time at UW. He has always been someone who offers me entirely new perspectives on research, teaching, and life. Beginning doctoral-level research was not easy for me, but with his patience, intelligence, and optimism, he helped shape me into an independent researcher and, more importantly, a better human. To me, Yen-Chi is not only a research mentor but also a life advisor.

I also wish to express my warm appreciation to the other members of my doctoral supervisory committee, each of whom has played a complementary and impactful role in my doctoral life. I am deeply appreciative of Prof. Thomas Richardson for mentoring me through my preliminary exam, supporting my academic job applications, opening the door to causal inference, and demonstrating to me what it truly means to be a scholar. I am sincerely thankful to Prof. Zaid Harchaoui for his genuine advice and support, and for teaching me numerous skills and ideas that have enriched my training in ways beyond what my advisor could provide. I am much obliged to Prof. Alex Luedtke, whose Advanced Statistical Inference sequence helped me overcome my fear of hard-core statistical theory, and whose introduction to the TGIF group at UW gave me the invaluable opportunity to present and discuss my causal inference research. I am grateful to Prof. Armeen Taeb for introducing me to his connections at conferences and for bringing me so much joy through our conversations

and tennis matches. I also thank Prof. Matthew McQuinn for serving as the GSR on my committee and as an external endorser for my astronomy-related applications.

I have also been fortunate to learn from several wonderful collaborators during my doctoral studies. I am extremely grateful to Prof. Andrea Rotnitzky for her guidance and for teaching me how to transform my rudimentary idea into a high-quality research manuscript. My gratitude also goes to Prof. Rafael S. de Souza, Caren Marzban, and Alexander Giessing for our thoughtful discussions in astronomy, causal inference, and high-dimensional statistics. Moreover, my appreciation goes to Prof. Xinwei Shen for introducing me to the new and fascinating world of distributional learning and for our friendship. In addition, I would like to thank Prof. Tim Althoff and Ken Gu for inviting me to their data science research project. Special thanks are also due to Wesley Lee, Steven Wilkins-Reeves, and Aude Hofleitner, who hosted and mentored me for more than half a year within the Central Applied Science team at Meta. Because of you, as well as the many amazing colleagues and interns with whom I was fortunate to interact, my internship at Meta became the best industry experience that I had during the past eight years.

Beyond my doctoral committee and collaborators, I have benefited tremendously from my interactions with many faculty and staff members at UW. In chronological order, I would like to thank Prof. Ema Perkovic for supporting my PhD application twice; Prof. Michael Perlman and June Morita for opening the door to classical statistics and giving me my very first jobs at UW; Prof. Marina Meila for hosting me in her geometric data analysis group; Prof. Jon Wellner for answering my many spontaneous questions in empirical processes; Prof. Abel Rodriguez for mentoring me in my role as a pre-doctoral instructor and supporting my academic job search; Prof. Tyler McCormick for being a firm supporter of my survival analysis project; Prof. Elena Erosheva for providing feedback on my spatial confounding research; Prof. Daniela Witten for the physically and intellectually rewarding runs; Prof. Adrian Dobra for answering my drop-in questions; Prof. Carlos Cinelli for

discussing sensitivity analysis with me; Prof. Fang Han for his amazing and kind students; and Prof. Marco Carone for his insightful comments on my causal inference work. I am also grateful to the staff of the Department of Statistics at UW: Ellen Reynolds, Kristine Chan, Tracy Pham, Veronica Bae, Asa Sourdiffe, and Vickie Graybeal. You are the people who worked behind the scenes to make our department, and my doctoral studies, run as smoothly as possible. Lastly, I would like to thank the Department of Mathematics at UW for offering me a PhD position during the COVID-19 pandemic, allowing me to defer it twice, and eventually accepting my decision to decline the offer. A special thank you goes to Prof. Krzysztof Burdzy for teaching me probability theory and supporting my PhD application twice.

Although my acknowledgment may already be long, I would not forgive myself if I failed to recognize the wonderful friends that I have met along this journey. I begin with my thanks to Yen-Chi's group: Gang Cheng, Zhen Miao, Daniel Suen, Steven Wilkins-Reeve, Kunhui Zhang, Yanjiao Yang, Haoyang Wu, Jerry Wei, and Jen-Chieh Teng. In particular, I thank Gang, Zhen, and Kunhui for their helpful life advice and genuine encouragement; Daniel for his willingness to listen to and comment on my premature ideas; and Steve for his referral and our collaboration at Meta. I also thank the students in Fang's group: Zhexiao Lin, Yandi Shen, Ziming Lin, Leon Tran, and Keunwoo Lim, especially Zhexiao, for being a reliable conference buddy and an endless fountain of resources. To my other friends and peers, thank you for setting examples that I have long admired and for sharing so many joyful moments with me. Your dedication and accomplishments reminded me of the reputation of UW Statistics and inspired me to work hard throughout my doctoral studies, hoping that my own work might live up to the efforts of those who came before me:

• Those more senior than me: Richard Guo, Lang Liu, Wenyu Chen, Hanyu Zhang, Sam Koelle, Kenny Zhang, Zhenman Yuan, Apara Venkat, Jess Kunke, James Buenfil, Alex Kokot, Vydhourie R T Thiyageswaran, Lars van der Laan, Paweł Morzywołek, Ronak Mehta,

Jillian Fisher, Shuntao Chen, Zhaoqi Li, and Tianran Li;

• My PhD and master's cohort: Kayla Irish, Dasha Petrov, Saksham Jain, Alex Bank, Ronan Perry, Ethan Ancell, Juejue Wang, Shirley Mathur, Fengjie Chen, Qilong Chen, Dehai Liu, Xinyu Gao, Mars Gao, and Peter Liu;

• Those more junior than me: Wenhao Pan, Qijia He, Simon Nguyen, Dwright Xu, Joshua Yang, Zihang Yu, Jaxon Ren, Peng Zhang, Lexi Liu, Zhaoxing Wu, Jingfan Xu, Mateus Piovezan Otto, Anqi Zhao, Yuhan Qian, Yuhong Li, Austin Sibu, Danielle Tsao, Olivia Lucile Mcgough, and Grant Hopkins;

• Those whom I met in industry: Tao Li (at Trip.com), Zeyuan Cai (at Trip.com), Jiayue Xue (at Trip.com), Bo Lin (at Uber), Xu Chen (at Meta), Jikai Jin (at Meta), Yao Song (at Meta), Bowen Zhang (at Meta), Zibin Chen (at Meta), Xuelin Yang (at Meta), Zoey Zhu (at Meta), Till Saenger (at Meta), and Weihang Xu (at Meta).

I know that I cannot list everyone who has offered me kind encouragement, genuine support, and timely help. If your deserved credit is not explicitly given here, please know that this was not intentional, and that you hold a special place in my heart.

Lastly, and most importantly, I am eternally thankful to my family for their unconditional love, unwavering support, invaluable sacrifices, and careful nurturing. I am deeply indebted to my grandparents for opening the door to the realm of mathematics for me and for nurturing my passion and determination for science. I feel beholden to my aunt for always pointing out my shortcomings and offering helpful suggestions whenever I faced difficult decisions. Finally, I owe an immeasurable debt of gratitude to my parents, who have consistently advised and supported me in all my pursuits. They created a liberal environment in which I could explore the world, and they encouraged me whenever I felt low. Without their sacrifices and steady support, I would never have achieved what I have in life. I would like to take this opportunity to say that I love you all.

Data Acknowledgments

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics — Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

DEDICATION

To my family, especially in loving memory of my grandfather (1937–2017), who laid its foundation.

Chapter 1

INTRODUCTION

1.1 Motivation

Modern astronomical sky surveys provide an increasingly detailed view of the Universe (York et al., 2000; Colless et al., 2003; Le Fèvre et al., 2005; Scoville et al., 2007; Jones et al., 2009; Driver et al., 2011; Guzzo et al., 2014; Bryant et al., 2015; Ivezić et al., 2019). By mapping galaxies, quasars, and other celestial objects across vast regions of the sky, these surveys create unprecedented opportunities to study the origin, structure, and evolution of the Universe. At the same time, the observational nature of these data creates statistical challenges. First, the astronomical objects are not naturally observed in a flat Euclidean space. The sky locations of astronomical objects lie on the celestial sphere, and redshift surveys are more faithfully viewed through spherical or light-cone geometry. Furthermore, many scientific signals of interest, such as galaxy clusters and cosmic filaments discussed below, appear as low-dimensional geometric structures embedded in noisy observations. Second, modern sky surveys record a large number of measured and derived features for each observed object, yielding data that are high-dimensional, noisy, and potentially incomplete. Third, as in many scientific domains, astronomical observations are not generated by randomized experiments. When scientific questions move beyond association toward causal interpretation, confounding and violations of standard identification assumptions become central obstacles. These challenges motivate the central question of this dissertation: how can we draw statistically valid conclusions from astronomical data while accounting for geometry, high dimensionality, missingness, and causal assumption violations?

To address this question, the dissertation develops statistical and computational methods for complex observational data, with astronomy serving as the main motivating domain.

The key astronomical concept throughout this dissertation is the cosmic web, a large-scale physical pattern through which matter in the Universe is not uniformly distributed but rather forms an interconnected network structure (Bond et al., 1996). This structure arises from gravitational instability and anisotropic collapse in the early Universe, and is commonly described in terms of four components: dense galaxy clusters (also known as cosmic nodes), elongated cosmic filaments, vast and flat cosmic sheets or walls, and the large regions of near-emptiness known as cosmic voids (Zel'Dovich, 1970). Among these components, we focus primarily on cosmic filaments and detect them using galaxies and quasars observed by the Sloan Digital Sky Survey (SDSS-IV; Ahumada et al. 2020). Filaments are scientifically important because they trace one-dimensional high-density structures that connect clusters and delineate the boundaries of cosmic voids. Accurate filament detection can thus support studies of cosmic voids, which provide rich resources to infer the cosmology and the nature of dark matter halos (Aragón-Calvo et al., 2007; Aragon-Calvo and Yang, 2014; Kreisch et al., 2022). Filaments are also relevant for inferring the stellar properties of nearby galaxies (Tempel et al., 2014; Chen et al., 2017; Malavasi et al., 2022) and detecting weak signals such as cosmic microwave background lensing (He et al., 2018). Despite the undoubted importance of cosmic filaments, identifying them from observed or simulated galaxy data is difficult because of their intricate geometric patterns (Cautun et al., 2014; Libeskind et al., 2018).

Motivated by these scientific questions and statistical challenges, this dissertation develops along three themes. The first is geometry-aware structure estimation. We propose statistical surrogates for cosmic nodes and filaments and develop algorithms for detecting cosmic web structures while respecting the spherical geometry of sky observations. The second is valid inference with high-dimensional and incomplete measurements. After reconstructing a cosmic web catalog, we develop efficient statistical methods for studying associations between galaxy properties and nearby cosmic web structures while accounting for missing outcomes and regularization bias. The third is causal inference for continuous treatments under positivity violations. We develop bias-corrected estimators and introduce new coun-

terfactual treatment regimes, with an application to the problem of separating "nature" and "nurture" effects in galaxy evolution using IllustrisTNG simulation data.

1.2 Contributions and Overview

The remainder of the dissertation is organized as follows. Chapter 2–Chapter 4 develop the statistical model, algorithmic theory, and astronomical catalog data for geometry-aware cosmic web detection. Chapter 5 and Chapter 6 then move from geometric reconstruction to statistical and causal inference with complex observational data.

Modeling the Cosmic Web Through Directional Density Ridges: In Chapter 2, we introduce a novel statistical surrogate for cosmic filaments called "directional density ridges". This surrogate respects the spherical geometry of sky observations while effectively modeling high-density features of the underlying galaxy distribution. We establish stability and statistical consistency guarantees for directional density ridge estimation via kernel smoothing.

Cosmic Web Detection With Geometry-Aware Algorithms: In Chapter 3, we derive practical algorithms, directional mean shift (DMS) and directional subspace constrained mean shift (DirSCMS), for recovering directional local modes and density ridges from data. We study their convergence properties and establish linear convergence guarantees under suitable regularity conditions.

These two chapters are based on my published work and preprints with Yen-Chi Chen (Zhang and Chen, 2021b,a, 2023, 2026).

SDSS-IV Cosmic Web Catalog: In Chapter 4, we apply the proposed directional density ridge framework and associated algorithms to galaxy and quasar observations from Data Release 16 of SDSS-IV. The resulting catalog provides a tomographic reconstruction of cosmic nodes and filaments over a wide redshift range. By turning the geometry-aware statistical framework into a data product that can be used in subsequent scientific analyses, this chapter serves as a bridge between the methodological and inferential parts of the dissertation. This chapter is based on my published work with Rafael S. de Souza and Yen-Chi

Chen (Zhang et al., 2022), together with the associated catalog data publication (Zhang, 2022).

High-Dimensional Statistical Inference With the Cosmic Web: In Chapter 5, we develop efficient inference methods for high-dimensional linear regression with missing outcomes, motivated by the problem of studying associations between galactic stellar mass and nearby cosmic web structures. The proposed method uses a convex debiasing program to balance bias and variance optimally when estimating low-dimensional functionals of a high-dimensional regression function. Provided that the missingness probabilities (or propensity scores) are consistently estimated at in-sample data points, the resulting debiased estimator is asymptotically normal and semiparametrically efficient among all asymptotically linear estimators. This chapter is based on my joint work with Alexander Giessing and Yen-Chi Chen (Zhang et al., 2023b).

Causal Inference For Galaxy Evolution Under Positivity Violations: In Chapter 6, we turn from association to causal inference for continuous treatments in observational studies, where the usual positivity condition may fail. We develop novel identification and estimation strategies for the dose-response curve and its derivative under an additive structural model for the counterfactual outcome. Using tools from nonparametric set estimation, we construct inverse probability weighted and doubly robust estimators that remain unbiased under certain positivity violations. We also introduce a new counterfactual regime defined by the nearest feasible treatment policy (NFTP), which extends the classical dose-response curve by adapting to the geometry of the conditional treatment support. The empirical application uses IllustrisTNG simulation data to study the effect of local environment on galactic star formation rate, providing a causal perspective on the "Nature versus Nurture" debate in galaxy evolution. This chapter is based on my joint work with Andrea Rotnitzky and Yen-Chi Chen (Zhang et al., 2024; Zhang and Chen, 2025).

Broader Impacts: Although the motivating examples come primarily from astronomy, the methodological ideas in this dissertation have broader impacts in statistics and beyond. Directional density ridges are relevant for data on spheres and product manifolds, including

directional, spatial, and spatio-temporal data. The debiasing method for high-dimensional regression with missing outcomes applies to studies in which the scientific target is a low-dimensional functional of a high-dimensional model. The causal inference framework for continuous treatments under positivity violations is applicable to observational studies in epidemiology, economics, environmental science, and other domains where some treatment levels are infeasible for some covariate strata.

Other Explorations: This dissertation does not include several additional projects: (1) my collaborative work with Tim Althoff and Ken Gu on building a benchmark for evaluating language model agents for statistical analysis (Gu et al., 2024); (2) two collaborative works with Caren Marzban, Bond Nicholas, and Michael Richman on practical introductions to causal inference for meteorologists (Marzban et al., 2025a,b); (3) my work with Steven Wilkins-Reeves, Wesley Lee, and Aude Hofleitner on transfer regression learning (Zhang et al., 2026); and (4) my work with Yanjiao Yang, Xinwei Shen, and Yen-Chi Chen on imputation (Yang et al., 2026). I served as a secondary contributor to the first two lines of work, as the primary author of the third project, which was conducted during my internship at Meta, and as a main contributor to the fourth.

1.3 Notation

Although each chapter introduces problem-specific notation as needed, we summarize here the conventions that recur throughout the dissertation.

Bold lower-case letters, such as $\boldsymbol{x}$ and $\boldsymbol{\mu}$, denote vectors, while bold upper-case letters, such as $\boldsymbol{X}$, denote random vectors. Non-bold capital letters typically denote scalar random variables. For a vector $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ and $q \geq 1$, we write $||\boldsymbol{u}||_q = \left(\sum_{j=1}^d |u_j|^q\right)^{1/q}$; in particular, $||\boldsymbol{u}||_\infty = \max_{1\leq j\leq d} |u_j|$ and $||\boldsymbol{u}||_0 = \sum_{j=1}^d \mathbb{1}_{\{u_j \neq 0\}}$ denotes the number of nonzero entries. For an index set S, the subvector of $\boldsymbol{u}$ restricted to S is denoted by $\boldsymbol{u}_S$, and $\boldsymbol{u}^T$ denotes its transpose. For a random variable Z and $q \geq 1$, we write $||Z||_q = (\mathbb{E}|Z|^q)^{1/q}$.

More generally, for $\alpha \in (0,2]$, the ψ_α-Orlicz norm is

$$||Z||_{\psi_\alpha} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{|Z|^\alpha}{t^\alpha}\right)\right] \leq 2\right\},$$

with $\alpha = 2$ and $\alpha = 1$ corresponding to the usual sub-Gaussian and sub-exponential tail scales. We write $X \perp\!\!\!\perp Y$ when random elements X and Y are independent. The set of real numbers is denoted by $\mathbb{R}$, and the q-dimensional unit sphere embedded in $\mathbb{R}^{q+1}$ is written as

$$\Omega_q := \left\{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1\right\},$$

and ω_q denotes the corresponding surface measure. For a matrix $A = (A_{jk}) \in \mathbb{R}^{m\times n}$, we write $||A||_{\max} = \max_{j,k}|A_{jk}|$ for its entrywise maximum norm, $||A||_F = \left(\sum_{j,k} A_{jk}^2\right)^{1/2}$ for its Frobenius norm, and $||A||_2 = \sup_{||\boldsymbol{u}||_2=1} ||A\boldsymbol{u}||_2$ for its operator norm. We use $||\cdot||_{\max}$ analogously for vectors or derivative arrays by taking the largest absolute entry. More generally, $\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \mathcal{X}, \mathcal{T}$, and $\mathcal{Y}$ denote generic sample spaces or supports. For a smooth function f, we use ∇f and $\nabla^2 f$ for its gradient and Hessian, with partial gradients such as $\nabla_{\boldsymbol{x}} f$ used when the arguments are partitioned.

We use P and $\mathbb{E}$ for probability and expectation under the relevant data-generating distribution, adding subscripts when the underlying law needs to be emphasized. Conditional laws and densities are written with subscripts, such as $\mathrm{P}_{T|\boldsymbol{X}}$ and $p_{T|\boldsymbol{X}}$, and lower-case letters denote densities whenever they exist. For an independent and identically distributed (i.i.d.) sample $U_1, \ldots, U_n$, the empirical measure is written as $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(U_i)$ and the empirical process as $\mathbb{G}_n f = \sqrt{n}\,(\mathbb{P}_n - \mathrm{P})\,f$. We use $\mathbb{1}_{\{\cdot\}}$ for indicator functions.

For deterministic sequences, $a_n = O(b_n)$ and $a_n = o(b_n)$ have their usual asymptotic meanings. For random sequences, $O_P(\cdot)$ and $o_P(\cdot)$ denote stochastic order in probability. We also use $a_n \lesssim b_n$ or equivalently $b_n \gtrsim a_n$ when a_n is bounded above by a constant multiple of b_n, and $a_n \asymp b_n$ when both bounds hold.

For the causal inference discussion, $Y^{(t)}$ denotes the potential outcome under treatment level t, the dose-response curve is written as $m(t) = \mathbb{E}[Y^{(t)}]$, and its derivative effect curve as $\theta(t) = m'(t)$ when it is well-defined. When the support of the treatment depends on

covariates, $\mathcal{T}(\boldsymbol{x})$ denotes the support of $T|\boldsymbol{X} = \boldsymbol{x}$ and $\mathcal{X}(t)$ denotes the support of $\boldsymbol{X}|T = t$. Additional notation specific to individual chapters is introduced locally when first needed.

Chapter 2

STATISTICAL THEORY FOR DIRECTIONAL MODES AND DENSITY RIDGES

This chapter develops the statistical foundation for modeling cosmic web structures (specifically, nodes and filaments) in data supported on the unit sphere Ω_q. We define local modes and density ridges of a directional density function f supported on Ω_q. In the astronomical applications motivating this dissertation, these objectives provide geometry-aware statistical surrogates for cosmic nodes and filaments observed on the celestial sphere. We construct kernel-based estimators for these geometric features and establish consistency guarantees for the estimated directional mode and ridge sets. The algorithmic problem of computing these estimators in practice is deferred to Chapter 3, where we study mean shift and subspace constrained mean shift algorithms adapted to flat, spherical, and their product-space geometries.

2.1 Introduction

Modern scientific datasets increasingly exhibit complex geometric structures, often arising from observations collected on nonlinear domains or from mixtures of variables with different geometric constraints. A central statistical task is to extract low-dimensional summaries that are both scientifically interpretable and geometrically faithful (Izenman, 2012; Chazal and Michel, 2021). While many methods exist for characterizing low-dimensional structures (Fefferman et al., 2016; Wasserman, 2018), we focus on local modes and density ridges (also known as principal curves or surfaces in Ozertem and Erdogmus 2011), which provide density-based summaries of high-probability regions and filamentary structures in point cloud data. These objects are attractive for both statistical and computational reasons. Statistically, they

can be consistently estimated by kernel density estimators (KDEs; Romano 1988; Mokkadem and Pelletier 2003; Genovese et al. 2014) under regularity conditions. Practically, the mean shift algorithm (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002) and its subspace constrained variant (Ozertem and Erdogmus, 2011; Ghassabeh, 2013) are well-suited for identifying these features in Euclidean spaces.

However, many datasets in practice do not naturally reside in a flat Euclidean space. Directional observations, for example, are supported on a nonlinear q-dimensional sphere Ω_q, or more generally on a topological space that is homeomorphic to a sphere. A simple example is provided by periodic dihedral angles in protein structures, which are naturally represented on a unit circle ($q = 1$) (Zimmermann and Hansmann, 2006). More complex settings involve variables defined on spaces with different geometries. In astronomy, for instance, each observed object is commonly recorded by a three-element tuple: (right ascension, declination, redshift) (Dawson et al., 2016; Brown et al., 2018). The first two coordinates specify a position on the celestial sphere ($q = 2$), whereas redshift reflects radial distance from Earth. More broadly, such observations fall within the class of spatio-temporal data, in which directional, Euclidean, and periodic components may appear jointly (Hall et al., 2006). These geometric constraints make standard Euclidean KDEs and mean shift algorithms inadequate without modification.

To address this issue, we formulate the sets of local modes and density ridges intrinsically for densities supported on spheres and related product spaces. Let f be a density on Ω_q satisfying $\int_{\Omega_q} f(\boldsymbol{x})\,\omega_q(d\boldsymbol{x}) = 1$, where ω_q denotes the volume measure on Ω_q. We define local modes and directional density ridges using the Riemannian gradient and Hessian of f, so that the resulting population objects are compatible with the spherical geometry. We further extend these definitions to densities on product spaces $\mathcal{S}_1 \times \mathcal{S}_2$, where each factor $\mathcal{S}_j$ may be either a Euclidean space $\mathbb{R}^{D_j}$ or a sphere Ω_{D_j}. For a density f on such a product space, normalized as $\int_{\mathcal{S}_1 \times \mathcal{S}_2} f(\boldsymbol{x}, \boldsymbol{y})\,\omega(d\boldsymbol{x}, d\boldsymbol{y}) = 1$, the measure ω is induced by the product Riemannian structure (Pennec, 2006).

Although local modes and density ridges have been studied extensively in $\mathbb{R}^D$ (Parzen,

Figure 2.1: Simulated dataset $\{(\theta_i, \phi_i)\}_{i=1}^{1000}$ on Ω_2 and $\Omega_1 \times \Omega_1$. Although the two coordinates (θ_i, ϕ_i) of the data in both panels are periodic, the underlying supports have different topologies. Consequently, mode-seeking and ridge-finding procedures must be adapted to the geometry of the sample space.

1962; Chernoff, 1964; Hall et al., 1992; Vieu, 1996; Hall et al., 2001; Genovese et al., 2014; Chen et al., 2014; Chen et al., 2015, 2016; Qiao and Polonik, 2016; Sasaki et al., 2018; Chacón, 2020; Qiao, 2021; Qiao and Polonik, 2021; Qiao, 2025), their generalizations to Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$ introduce several new challenges. The first challenge is geometric, where Ω_q has nonlinear curvature. Furthermore, the topology of $\mathcal{S}_1 \times \mathcal{S}_2$ may differ fundamentally from that of $\mathbb{R}^D$ or Ω_q when one or both factors $\mathcal{S}_j, j = 1, 2$ are directional. Consider a random sample $\{(\theta_i, \phi_i)\}_{i=1}^{1000}$ with $\theta_i \in [0, 2\pi)$ and $\phi_i = 0$. Depending on the problem context, this dataset may be viewed as either points from a circle on the sphere Ω_2, with (θ_i, ϕ_i) as longitude and latitude, or on the torus $\Omega_1 \times \Omega_1$. The periodicity constraints of Ω_2 and $\Omega_1 \times \Omega_1$ are different, and these two spaces are not homeomorphic. Consequently, mode and ridge estimation must respect the intrinsic geometry and topology of the support. The second challenge is inferential, where a coordinatewise extension of Euclidean mode-seeking or ridge-finding methods can fail to identify joint geometric features on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$. For example, consider two densities f_1, f_2 on $\Omega_1 \times \Omega_1$ with the sets of local modes given by $\mathcal{M} = \{(0,0), (0, \frac{3\pi}{4}), (\frac{\pi}{2}, 0), (\frac{\pi}{2}, \frac{3\pi}{4})\}$ and $\mathcal{M}' = \{(0, \frac{3\pi}{4}), (\frac{\pi}{2}, 0)\}$, respectively. Although $\mathcal{M}, \mathcal{M}'$ differ, their marginal local modes, which are $0, \frac{\pi}{2}$ on the first coordinate and $0, \frac{3\pi}{4}$ on the second coordinate, are identical. Any procedure that searches for modes separately

along each component would therefore be unable to distinguish the two joint structures. This unidentifiability motivates methods that define and estimate modes and ridges jointly on Ω_q and $\mathcal{S}_1 \times \mathcal{S}_2$, rather than reducing the problem to separate marginal analyses.

The contributions of this chapter are as follows. We first review the kernel density estimators and differential-geometric tools needed to work on Ω_q and $\mathcal{S}_1 \times \mathcal{S}_2$. We then define directional local modes and density ridges as Riemannian analogues of their Euclidean counterparts. Finally, we study the stability and statistical consistency of these geometric sets under kernel smoothing. These results provide the population and large-sample justification for the geometry-aware algorithms developed in Chapter 3 and, ultimately, for the cosmic web reconstruction pipeline studied in Chapter 4.

2.2 Preliminaries

This section reviews the KDEs for directional and directional-linear data as well as some differential geometry concepts on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$.

2.2.1 Kernel Density Estimation With Euclidean Data

Let $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ be a random sample from a distribution P with density p supported on the Euclidean space $\mathbb{R}^D$. The Euclidean KDE at point $\boldsymbol{x} \in \mathbb{R}^D$ with a kernel function K and bandwidth parameter $h \equiv h(n)$ is written as (Wand and Jones, 1994; Wasserman, 2006; Scott, 2015; Chen, 2017):

$$\widehat{p}_n(\boldsymbol{x}) = \frac{1}{nh^D} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{2.1}$$

The kernel $K : \mathbb{R}^D \to \mathbb{R}$ is typically chosen as a unimodal function satisfying the following standard properties:

- (K1) $\int_{\mathbb{R}^D} K(\boldsymbol{x}) d\boldsymbol{x} = 1$.
- (K2) $K(\boldsymbol{x})$ is (radially) symmetric, *i.e.*, $\int_{\mathbb{R}^D} \boldsymbol{x}\, K(\boldsymbol{x})\, d\boldsymbol{x} = 0$.

- (K3) $\int_{\mathbb{R}^D} ||\boldsymbol{x}||_2^2 K(\boldsymbol{x}) d\boldsymbol{x} < \infty$, where $||\cdot||_2$ is the usual L_2 norm in $\mathbb{R}^D$.

One common way to construct a multivariate kernel $K(\boldsymbol{x})$ with the above properties is to derive it from a kernel profile as follows:

$$K(\boldsymbol{x}) = c_{k,D} \cdot k\left(||\boldsymbol{x}||_2^2\right), \tag{2.2}$$

where $c_{k,D}$ is the normalizing constant such that K satisfies (K1) and the function $k : [0, \infty) \to [0, \infty)$ is called the *profile* of the kernel. This kernel form is often used in deriving (subspace constrained) mean shift algorithms; see Section 3.2. An important example of the profile function is $k_N(x) = \exp\left(-\frac{x}{2}\right)$ for $x \geq 0$, leading to the multivariate Gaussian kernel $K_N(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{||\boldsymbol{x}||_2^2}{2}\right)$.

Another approach to designing a multivariate kernel function is to leverage the product kernel technique as $K(\boldsymbol{x}) = K_1(x_1) \cdots K_D(x_D)$, where $K_1, ..., K_D$ are kernel functions defined on $\mathbb{R}$ satisfying the properties (K1-3). This choice leads to a multivariate KDE as:

$$\widehat{p}_n(\boldsymbol{x}) = \frac{1}{nh^D} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{i,1}}{h}\right) \cdots K_D\left(\frac{x_D - X_{i,D}}{h}\right). \tag{2.3}$$

In fact, the multivariate Gaussian kernel K_N can be obtained by defining its kernel profile as $k_N(x) = \exp\left(-\frac{x}{2}\right)$ for $x \geq 0$ or taking $K_1(x) = \cdots = K_D(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. In practice, the multivariate KDE (2.1) with Gaussian kernel is the most popular nonparametric density estimator with Euclidean data.

A crucial practical step in applying KDE is selecting the bandwidth parameter h. Common approaches aim to minimize the mean integrated squared error (MISE):

$$\text{MISE}(h) = \mathbb{E} \int_{\mathbb{R}^D} \left[\widehat{p}_h(\boldsymbol{x}) - p(\boldsymbol{x})\right]^2 d\boldsymbol{x}$$

or its leading asymptotic approximation, use a rule of thumb (Silverman, 1986), apply cross-validation (Rudemo, 1982; Bowman, 1984; Hall, 1983; Stone, 1984), or employ plug-in methods (Sheather and Jones, 1991). We refer the interested reader to Jones et al. (1996); Sheather (2004) and Chapter 6.5 of Scott (2015) for comprehensive reviews.

2.2.2 Kernel Density Estimation With Directional Data

Euclidean KDEs (2.1) are not directly appropriate for directional data because they ignore the geometry and periodicity of the support; see Section A.1 for a detailed discussion. Fortunately, the theory of kernel density estimation with directional data has been well-studied in the literature (Beran, 1979; Hall et al., 1987; Bai et al., 1988; Zhao and Wu, 2001; García-Portugués, 2013; Pewsey and García-Portugués, 2021). Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q \subset \mathbb{R}^{q+1}$ now be a random sample generated from an underlying directional density function f on Ω_q with $\int_{\Omega_q} f(\boldsymbol{x})\, \omega_q(d\boldsymbol{x}) = 1$, where ω_q is the Lebesgue measure on Ω_q. The directional KDE is given by:

$$\widehat{f}_h(\boldsymbol{x}) = \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) = \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right), \tag{2.4}$$

where L is a directional kernel (*i.e.*, a rapidly decaying nonnegative function defined on $(-\delta_L, \infty) \subset \mathbb{R}$ for some constant $\delta_L > 0$), $h > 0$ is the bandwidth parameter, and $c_{L,q}(h)$ is a normalizing constant satisfying

$$c_{L,q}(h)^{-1} = \int_{\Omega_q} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{y}}{h^2}\right) \omega_q(d\boldsymbol{y}) = O(h^q). \tag{2.5}$$

Remark 2.1. *The distance metric used by the directional KDE (2.4) on Ω_q is identical to the standard Euclidean metric in the ambient space $\mathbb{R}^{q+1}$. This is because the standard Euclidean metric $||\cdot||_2$ of $\mathbb{R}^{q+1}$ is topologically equivalent (but not strongly equivalent) to the geodesic distance $d_g(\cdot, \cdot)$ on Ω_q due to the following equality:*

$$||\boldsymbol{x} - \boldsymbol{y}||_2 = 2 \sin\left(\frac{d_g(\boldsymbol{x}, \boldsymbol{y})}{2}\right). \tag{2.6}$$

See Section C.1.5 in Ok (2007) for the definition of equivalence of metrics. Hence, the distance metric in (2.4) is indeed intrinsic on Ω_q and adapted to its geometry.

As in Euclidean KDEs, bandwidth selection plays a central role in the performance of directional KDEs (Hall et al., 1987; Bai et al., 1988; Taylor, 2008; Marzio et al., 2011; Oliveira et al., 2012; García-Portugués, 2013; Saavedra-Nieves and María Crujeiras, 2020). By contrast, the specific choice of the kernel is typically less important; see, *e.g.*, page 72 of

Wasserman (2006) and Section 6.3.2 in Scott (2015) for an explanation. A popular candidate is the so-called von Mises kernel $L(r) = e^{-r}$, which is the directional analogue of the Gaussian kernel. Its name originates from the famous q-von Mises-Fisher distribution on Ω_q, which is denoted by $\text{vMF}(\boldsymbol{\mu}, \nu)$ and has density

$$f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \nu) = C_q(\nu) \cdot \exp(\nu \boldsymbol{\mu}^T \boldsymbol{x}) \quad \text{with} \quad C_q(\nu) = \frac{\nu^{\frac{q-1}{2}}}{(2\pi)^{\frac{q+1}{2}} \mathcal{I}_{\frac{q-1}{2}}(\nu)}, \tag{2.7}$$

where $\boldsymbol{\mu} \in \Omega_q$ is the directional mean, $\nu \geq 0$ is the concentration parameter, and $\mathcal{I}_\alpha(\nu)$ is the modified Bessel function of the first kind of order ν. For further details on the statistical properties of the von Mises-Fisher distribution and directional KDE, we refer the interested reader to Mardia and Jupp (2000); Banerjee et al. (2005); García-Portugués (2013).

2.2.3 Kernel Density Estimation on $\mathcal{S}_1 \times \mathcal{S}_2$

Given the observed data $\{\boldsymbol{Z}_i\}_{i=1}^n = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^n$ on the product space $\mathcal{S}_1 \times \mathcal{S}_2$, we estimate the underlying density f using a product-kernel KDE as (Hall et al., 2006; García-Portugués et al., 2013, 2015):

$$\widehat{f}_{\boldsymbol{h}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} K_1\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_1}\right) K_2\left(\frac{\boldsymbol{y} - \boldsymbol{Y}_i}{h_2}\right), \tag{2.8}$$

where $K_j : \mathcal{S}_j \to \mathbb{R}$ is the kernel function for $j = 1, 2$, and each element of $\boldsymbol{h} = (h_1, h_2)$ is a bandwidth parameter. The form of the radially symmetric kernel K_j depends on the geometry of $\mathcal{S}_j$:

$$K_j(\boldsymbol{u}) = C_{k_j, D_j}(h_j) \cdot k_j\left(||\boldsymbol{u}||_2^2\right) = \begin{cases} \frac{C_{k,D_j}}{h_j^{D_j}} \cdot k\left(||\boldsymbol{u}||_2^2\right) & \text{if } \mathcal{S}_j = \mathbb{R}^{D_j}, \\ C_{L,D_j}(h_j) \cdot L\left(\frac{||\boldsymbol{u}||_2^2}{2}\right) & \text{if } \mathcal{S}_j = \Omega_{D_j}, \end{cases} \tag{2.9}$$

for $j = 1, 2$, where k and L are the profiles of the linear and directional kernels, respectively, while $C_{k,D_j}, C_{L,D_j}(h_j)$ are normalizing constants. A key distinction between linear and directional kernels in (2.9) is whether the normalizing constants are separable from the bandwidth parameters. In particular, if $\mathcal{S}_1$ is directional, then $K_1\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h_1}\right) = C_{L,D_1}(h_1) \cdot L\left(\frac{||\boldsymbol{x}-\boldsymbol{X}_i||_2^2}{2h_1^2}\right) =$

$C_{L,D_1}(h_1) \cdot L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h_1^2}\right)$, which aligns with the usual KDE formulation for directional data in (2.4). The estimator $\widehat{f}_{\boldsymbol{h}}$ is naturally defined not only on $\mathcal{S}_1 \times \mathcal{S}_2$ but also on its ambient space. Finally, if $\mathcal{S}_1, \mathcal{S}_2$ are both directional or Euclidean, $\widehat{f}_{\boldsymbol{h}}$ can also be defined via a spherically symmetric kernel (García-Portugués and Meilán-Vila, 2024).

As before, the commonly used linear and directional kernel profiles are $k(s) = e^{-s/2}$ and $L(r) = e^{-r}$, corresponding to the Gaussian kernel and von Mises kernel, respectively. Applying Gaussian and/or von Mises kernels to K_1 and K_2 in (2.8), the KDE simplifies to

$$\begin{aligned} \widehat{f}_{\boldsymbol{h}}(\boldsymbol{z}) &= \frac{C(\boldsymbol{H})}{n} \sum_{i=1}^{n} \exp\left[-\frac{(\boldsymbol{z}-\boldsymbol{Z}_i)^\top \boldsymbol{H}^{-1}(\boldsymbol{z}-\boldsymbol{Z}_i)}{2}\right] \\ \text{with} \quad \boldsymbol{H} &= \mathtt{Diag}\left(h_1^2 \boldsymbol{I}_{D_1+\mathbb{1}_{\{\mathcal{S}_1=\Omega_{D_1}\}}}, h_2^2 \boldsymbol{I}_{D_2+\mathbb{1}_{\{\mathcal{S}_2=\Omega_{D_2}\}}}\right), \end{aligned} \tag{2.10}$$

where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}_1 \times \mathcal{S}_2$, $\boldsymbol{I}_D$ is the identity matrix in $\mathbb{R}^{D\times D}$, and $C(\boldsymbol{H}) \equiv \prod_{j=1}^{2} C_{k_j,D_j}(h_j)$ is the normalizing constant from (2.8) and (2.9). For simplicity, we consider a (block) diagonal bandwidth matrix $\boldsymbol{H}$ (Wand and Jones, 1993), though our theory also applies to the general positive definite bandwidth matrix. Additionally, when $\mathcal{S}_j = \Omega_{D_j}$, it is more natural to apply a single bandwidth parameter h_j to each coordinate on Ω_{D_j} due to the isotropic geometry.

2.2.4 Riemannian Gradient, Hessian, and Exponential Map on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$

Because the unit hypersphere Ω_q is a nonlinear manifold, derivatives of a smooth function f on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$ should be interpreted on the corresponding tangent spaces. The Riemannian gradient and Hessian differ from the ordinary ambient gradient and Hessian, but they can be expressed through ambient derivatives followed by suitable projections.

• **Riemannian Gradient on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$:** We first discuss Ω_q and then extend the notation to the product space $\mathcal{S}_1 \times \mathcal{S}_2$. Let $T_{\boldsymbol{x}}$ be the tangent space of Ω_q at point $\boldsymbol{x} \in \Omega_q$, which consists of all the vectors tangent to Ω_q at $\boldsymbol{x}$. Given a smooth function $f : \Omega_q \to \mathbb{R}$, its *Riemannian gradient* $\mathtt{grad}\, f(\boldsymbol{x}) \in T_{\boldsymbol{x}}$ is defined as:

$$\langle \boldsymbol{v}, \mathtt{grad}\, f(\boldsymbol{x}) \rangle_{\boldsymbol{x}} = df_{\boldsymbol{x}}(\boldsymbol{v}) \tag{2.11}$$

for any (unit) vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$, where $\langle \cdot, \cdot \rangle_{\boldsymbol{x}}$ is the inner product (or Riemannian metric) in $T_{\boldsymbol{x}}$ and $df_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \mathbb{R}$ is the differential operator of f at $\boldsymbol{x} \in \Omega_q$; see, *e.g.*, Section 3.1 in Banyaga and Hurtubise (2004) for more details. Note that the Riemannian metric on Ω_q coincides with the standard inner product $\langle \cdot, \cdot \rangle$ in the ambient space $\mathbb{R}^{q+1}$; see Section 3.6.1 in Absil et al. (2008). If f is smooth in an open neighborhood containing Ω_q and we consider $\texttt{grad}\, f(\boldsymbol{x}), \boldsymbol{v} \in T_{\boldsymbol{x}}$ as vectors in $\mathbb{R}^{q+1}$, then the inner product in $T_{\boldsymbol{x}}$ reduces to the usual one in $\mathbb{R}^{q+1}$ and the Riemannian gradient $\texttt{grad}\, f(\boldsymbol{x})$ can be expressed in terms of the total gradient $\nabla f(\boldsymbol{x})$ as:

$$\texttt{grad}\, f(\boldsymbol{x}) = \left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}), \tag{2.12}$$

where $\boldsymbol{I}_{q+1} \in \mathbb{R}^{(q+1)\times(q+1)}$ is the identity matrix. This equation shows that the Riemannian gradient is the projection of the Euclidean/ambient gradient $\nabla f(\boldsymbol{x})$ onto the tangent space $T_{\boldsymbol{x}}$ at $\boldsymbol{x} \in \Omega_q$.

Similarly, the differential of f at $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}_1 \times \mathcal{S}_2$ is a linear map $df_{\boldsymbol{z}} : T_{\boldsymbol{z}} \equiv T_{\boldsymbol{x}}(\mathcal{S}_1) \times T_{\boldsymbol{y}}(\mathcal{S}_2) \to \mathbb{R}$ given by $df_{\boldsymbol{z}}(\boldsymbol{v}) = \frac{d}{dt} f\left(\gamma(t)\right)\big|_{t=0}$ for any smooth curve $\gamma : [0,1] \to \mathcal{S}_1 \times \mathcal{S}_2$ and tangent vector $\boldsymbol{v} = (\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) \in T_{\boldsymbol{z}} \equiv T_{\boldsymbol{x}}(\mathcal{S}_1) \times T_{\boldsymbol{y}}(\mathcal{S}_2)$. Then, the *Riemannian gradient* $\texttt{grad}\, f(\boldsymbol{z})$ is a vector field in the tangent space $T_{\boldsymbol{z}}$ defined as:

$$df_{\boldsymbol{z}}(\boldsymbol{v}) = \langle \texttt{grad}\, f(\boldsymbol{z}), (\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) \rangle = \texttt{grad}\, f(\boldsymbol{z})^\top \boldsymbol{v}, \tag{2.13}$$

where the inner product $\langle \cdot, \cdot \rangle$ is the usual one in the smallest ambient Euclidean space containing $\mathcal{S}_1 \times \mathcal{S}_2$. From (2.13), we obtain a vector form of $\texttt{grad}\, f(\boldsymbol{x}, \boldsymbol{y})$ as:

$$\texttt{grad}\, f(\boldsymbol{z}) = \mathcal{P}_{\boldsymbol{z}} \nabla f(\boldsymbol{z}) \;\text{ with }\; \mathcal{P}_{\boldsymbol{z}} = \begin{pmatrix} \mathcal{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \mathcal{P}_2 \end{pmatrix} \;\text{ and }\; \begin{cases} \mathcal{P}_1 = \boldsymbol{I}_{D_1 + \mathbb{1}_{\{\mathcal{S}_1 = \Omega_{D_1}\}}} - \boldsymbol{x}\boldsymbol{x}^\top \cdot \mathbb{1}_{\{\mathcal{S}_1 = \Omega_{D_1}\}}, \\ \mathcal{P}_2 = \boldsymbol{I}_{D_2 + \mathbb{1}_{\{\mathcal{S}_2 = \Omega_{D_2}\}}} - \boldsymbol{y}\boldsymbol{y}^\top \cdot \mathbb{1}_{\{\mathcal{S}_2 = \Omega_{D_2}\}}. \end{cases} \tag{2.14}$$

Here, $\mathcal{P}_{\boldsymbol{z}}$ is the projection matrix onto the tangent space $T_{\boldsymbol{z}}$ and $\boldsymbol{I}_D$ denotes an identity matrix in $\mathbb{R}^{D \times D}$. Thus, $\mathcal{P}_{\boldsymbol{z}}$ reduces to the identity matrix when $\mathcal{S}_1 \times \mathcal{S}_2$ is Euclidean, while it acts as a projection operator when either of $\mathcal{S}_j, j = 1, 2$ is directional.

• **Riemannian Hessian on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$:** The *Riemannian Hessian* $\mathcal{H}f(\boldsymbol{x})$ at point $\boldsymbol{x} \in \Omega_q$ is a symmetric bilinear map from the tangent space $T_{\boldsymbol{x}}$ into itself defined by

$$\mathcal{H}f(\boldsymbol{x})\,[\boldsymbol{v}] = \bar{\nabla}_{\boldsymbol{v}} \mathtt{grad}\, f(\boldsymbol{x}) \tag{2.15}$$

for any $\boldsymbol{v} \in T_{\boldsymbol{x}}$, where $\bar{\nabla}_{\boldsymbol{v}}$ is the Riemannian connection on Ω_q; see Section 5.5 in Absil et al. (2008). Similar to $\mathtt{grad}\, f(\boldsymbol{x})$, the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ has the following explicit formula when viewed in the ambient Euclidean space $\mathbb{R}^{q+1}$:

$$\mathcal{H}f(\boldsymbol{x}) = (\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T\boldsymbol{x} \cdot \boldsymbol{I}_{q+1}\right](\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T), \tag{2.16}$$

where $\nabla f(\boldsymbol{x})$ and $\nabla\nabla f(\boldsymbol{x})$ are the total gradient and Hessian of f in $\mathbb{R}^{q+1}$. This formula can be derived via the Riemannian connection and Weingarten map on Ω_q (Absil et al. 2013 and Proposition 5.3.2 in Absil et al. 2008) or geodesics on Ω_q (Zhang and Chen, 2021b).

Analogously, the Riemannian Hessian of $f : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}$ is also a symmetric linear map $\mathcal{H}f(\boldsymbol{z})$ of the tangent space $T_{\boldsymbol{z}} \equiv T_{\boldsymbol{x}}(\mathcal{S}_1) \times T_{\boldsymbol{y}}(\mathcal{S}_2)$ into itself defined as $\mathcal{H}f(\boldsymbol{z})[\boldsymbol{v}] = \bar{\boldsymbol{\nabla}}_{\boldsymbol{v}} \mathtt{grad}\, f(\boldsymbol{z})$ for any tangent vector $\boldsymbol{v} = (\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) \in T_{\boldsymbol{z}}$, where $\bar{\boldsymbol{\nabla}}_{\boldsymbol{v}}$ is the Riemannian connection on $\mathcal{S}_1 \times \mathcal{S}_2$. As $\mathcal{S}_1 \times \mathcal{S}_2$ can be viewed as a (nonlinear) submanifold in its ambient Euclidean space, the Riemannian Hessian $\mathcal{H}f(\boldsymbol{z})$ on $\mathcal{S}_1 \times \mathcal{S}_2$ can be written as:

$$\begin{aligned}
\mathcal{H}f(\boldsymbol{z})[\boldsymbol{v}] &= \mathcal{P}_{\boldsymbol{z}}\Big(\nabla_{\boldsymbol{x}} \mathtt{grad}\, f(\boldsymbol{z}),\, \nabla_{\boldsymbol{y}} \mathtt{grad}\, f(\boldsymbol{z})\Big)[\boldsymbol{v}] \\
&= \mathcal{P}_{\boldsymbol{z}}\Big[\nabla^2 f(\boldsymbol{z}) - \mathtt{Diag}\,(\mathcal{A}_1, \mathcal{A}_2)\Big][\boldsymbol{v}] \quad \text{with} \quad \begin{cases} \mathcal{A}_1 = \boldsymbol{x}^\top \nabla_{\boldsymbol{x}} f(\boldsymbol{z}) \boldsymbol{I}_{D_1+1} \cdot \mathbb{1}_{\{\mathcal{S}_1 = \Omega_{D_1}\}}, \\ \mathcal{A}_2 = \boldsymbol{y}^\top \nabla_{\boldsymbol{y}} f(\boldsymbol{z}) \boldsymbol{I}_{D_2+1} \cdot \mathbb{1}_{\{\mathcal{S}_2 = \Omega_{D_2}\}}, \end{cases}
\end{aligned} \tag{2.17}$$

where $\mathtt{Diag}\,(\mathcal{A}_1, \mathcal{A}_2)$ is a (block) diagonal matrix, and we plug in the formula (2.11) of $\mathtt{grad}\, f(\boldsymbol{z})$ in the second equality. $\mathcal{A}_j$ is the curvature correction for the jth component space. It is nonzero only when $\mathcal{S}_j$ is directional, and it vanishes when $\mathcal{S}_j$ is Euclidean. Using $\boldsymbol{v} \in T_{\boldsymbol{z}}$ and the symmetric property of $\mathcal{H}f(\boldsymbol{z})$, we express the Hessian operator $\mathcal{H}f(\boldsymbol{z})$ as a matrix $\mathcal{H}f(\boldsymbol{z}) = \mathcal{P}_{\boldsymbol{z}}\Big[\nabla^2 f(\boldsymbol{z}) - \mathtt{Diag}\,(\mathcal{A}_1, \mathcal{A}_2)\Big]\mathcal{P}_{\boldsymbol{z}}$. Finally, both $\mathtt{grad}\, f(\boldsymbol{z})$ and $\mathcal{H}f(\boldsymbol{z})$ reduce to the total gradient $\nabla f(\boldsymbol{z})$ and Hessian $\nabla^2 f(\boldsymbol{z})$ when $\mathcal{S}_1, \mathcal{S}_2$ are flat Euclidean spaces.

Throughout Chapters 2 and 3, for a smooth function g we write $||g||_{\infty}^{(j)}$ for the supremum of the absolute values of its order-j derivatives; for directional functions, these derivatives are interpreted in local coordinates on Ω_q. In particular, $||g||_{\infty}^{(2)}$ controls the largest absolute entry of the corresponding Euclidean or Riemannian Hessian, and we set $||g||_{\infty,4}^{*} = \max_{0\leq j\leq 4} ||g||_{\infty}^{(j)}$.

• **Exponential Map:** An *exponential map* $\texttt{Exp}_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \Omega_q$ at $\boldsymbol{x} \in \Omega_q$ is a mapping that takes a vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$ to a point $\boldsymbol{y} := \texttt{Exp}_{\boldsymbol{x}}(\boldsymbol{v}) \in \Omega_q$ along the curve φ with $\varphi(0) = \boldsymbol{x}, \varphi(1) = \boldsymbol{y}$ and $\varphi'(0) = \boldsymbol{v}$. Here, $\varphi : [0,1] \to \Omega_q$ is a curve of minimum length between $\boldsymbol{x}$ and $\boldsymbol{y}$ (*i.e.*, the so-called geodesic on Ω_q). Intuitively, evaluating the exponential map $\texttt{Exp}_{\boldsymbol{x}}$ at $\boldsymbol{v}$ on Ω_q means starting at $\boldsymbol{x}$ and moving along the geodesic, or great circle, in the direction of $\boldsymbol{v}$ until reaching a point $\boldsymbol{y}$ whose geodesic distance from $\boldsymbol{x}$ equals $||\boldsymbol{v}||_2$. As Ω_q is a compact Riemannian manifold, the exponential map $\texttt{Exp}_{\boldsymbol{x}}$ is a diffeomorphism (smooth bijection) from a neighborhood of $\mathbf{0} \in T_{\boldsymbol{x}}$ to its image on Ω_q; see Lemma 6.16 in Lee (2006). The inverse of an exponential map (or logarithmic map) is defined within a neighborhood $U \subset \Omega_q$ around $\boldsymbol{x}$ as a mapping $\texttt{Exp}_{\boldsymbol{x}}^{-1} : U \to T_{\boldsymbol{x}}$ such that $\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$ represents the vector in $T_{\boldsymbol{x}}$ starting at $\boldsymbol{x}$ and pointing toward $\boldsymbol{y}$, whose length equals the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. For $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}_1 \times \mathcal{S}_2$, the exponential map $\texttt{Exp}_{\boldsymbol{z}} : T_{\boldsymbol{z}} \to \mathcal{S}_1 \times \mathcal{S}_2$ is a concatenation of the exponential map in each component space as $\texttt{Exp}_{\boldsymbol{z}}(\boldsymbol{v}) = \big(\texttt{Exp}_{\boldsymbol{x}}(\boldsymbol{v}_{\boldsymbol{x}}), \texttt{Exp}_{\boldsymbol{y}}(\boldsymbol{v}_{\boldsymbol{y}})\big)$, where $\boldsymbol{v} = (\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) \in T_{\boldsymbol{z}}$ as well as $\texttt{Exp}_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \mathcal{S}_1$ and $\texttt{Exp}_{\boldsymbol{y}} : T_{\boldsymbol{y}} \to \mathcal{S}_2$.

2.3 Mode and Ridge Estimation on Ω_q or $\mathcal{S}_1 \times \mathcal{S}_2$

This section defines order-d density ridges on Ω_q and $\mathcal{S}_1 \times \mathcal{S}_2$, and then introduces the corresponding plug-in estimators based on the KDEs reviewed in Section 2.2.

2.3.1 Euclidean Density Ridge and its Stability

We first review the definition of an order-d density ridge for a smooth density p supported on the Euclidean space $\mathbb{R}^D$ and its stability result (Eberly, 1996; Genovese et al., 2014). Consider the spectral decomposition of the Hessian $\nabla\nabla p(\boldsymbol{x})$ as $\nabla\nabla p(\boldsymbol{x}) = V(\boldsymbol{x})\Lambda(\boldsymbol{x})V(\boldsymbol{x})^T$, where $V(\boldsymbol{x}) = [\boldsymbol{v}_1(\boldsymbol{x}), ..., \boldsymbol{v}_D(\boldsymbol{x})] \in \mathbb{R}^{D\times D}$ is a real orthogonal matrix with the eigenvectors

of $\nabla\nabla p(\boldsymbol{x})$ as its columns and $\Lambda(\boldsymbol{x}) = \texttt{Diag}\left[\lambda_1(\boldsymbol{x}), ..., \lambda_D(\boldsymbol{x})\right] \in \mathbb{R}^{D\times D}$ is a diagonal matrix with $\lambda_1(\boldsymbol{x}) \geq \cdots \geq \lambda_D(\boldsymbol{x})$. Given that $V_d(\boldsymbol{x}) = [\boldsymbol{v}_{d+1}(\boldsymbol{x}), ..., \boldsymbol{v}_D(\boldsymbol{x})] \in \mathbb{R}^{D\times(D-d)}$, we let $U_d(\boldsymbol{x}) \equiv V_d(\boldsymbol{x})V_d(\boldsymbol{x})^T$ be the projection matrix onto the column space of $V_d(\boldsymbol{x})$ and $U_d^{\perp}(\boldsymbol{x}) = \boldsymbol{I}_D - V_d(\boldsymbol{x})V_d(\boldsymbol{x})^T = V_\diamond(\boldsymbol{x})V_\diamond(\boldsymbol{x})^T$ be the projection matrix onto the complement space, where $V_\diamond(\boldsymbol{x}) = [\boldsymbol{v}_1(\boldsymbol{x}), ..., \boldsymbol{v}_d(\boldsymbol{x})] \in \mathbb{R}^{D\times d}$ and $\boldsymbol{I}_D$ is the identity matrix in $\mathbb{R}^{D\times D}$. Then, the order-d principal gradient $G_d(\boldsymbol{x})$ (or projected gradient in Genovese et al. 2014; Chen et al. 2015) is defined as:

$$G_d(\boldsymbol{x}) = U_d(\boldsymbol{x})\nabla p(\boldsymbol{x}) = V_d(\boldsymbol{x})V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x}), \tag{2.18}$$

and $U_d^{\perp}(\boldsymbol{x})\nabla p(\boldsymbol{x})$ will be called the residual gradient. The order-d density ridge is then defined as

$$\begin{aligned} R_d &= \left\{\boldsymbol{x} \in \mathbb{R}^D : G_d(\boldsymbol{x}) = \boldsymbol{0}, \lambda_{d+1}(\boldsymbol{x}) < 0\right\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^D : V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x}) = \boldsymbol{0}, \lambda_{d+1}(\boldsymbol{x}) < 0\right\}. \end{aligned} \tag{2.19}$$

In particular, the order-0 ridge R_0 is equivalent to the set of local modes of p as:

$$\mathcal{M} := R_0 = \left\{\boldsymbol{x} \in \mathbb{R}^D : \nabla p(\boldsymbol{x}) = \boldsymbol{0}, \lambda_1(\boldsymbol{x}) < 0\right\}. \tag{2.20}$$

For a point $\boldsymbol{x} \in \mathbb{R}^D$, we define its projection onto a ridge R_d by $\pi_{R_d}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}\in R_d} ||\boldsymbol{x} - \boldsymbol{y}||_2$ and its distance to R_d by $d_E(\boldsymbol{x}, R_d) = ||\boldsymbol{x} - \pi_{R_d}(\boldsymbol{x})||_2 = \min_{\boldsymbol{y}\in R_d} ||\boldsymbol{x} - \boldsymbol{y}||_2$. Note that the projection from point $\boldsymbol{x} \in \mathbb{R}^D$ to R_d may not be unique. A standard condition ensuring uniqueness of this projection is expressed through the *reach* (Federer, 1959; Cuevas, 2009) as:

$$\texttt{reach}(R_d) = \inf\left\{\delta > 0 : \forall \boldsymbol{x} \in R_d \oplus \delta, \boldsymbol{x} \text{ has a unique projection onto } R_d\right\}, \tag{2.21}$$

where $R_d \oplus \delta = \cup_{\boldsymbol{x}\in R_d}\text{Ball}_D(\boldsymbol{x}, \delta)$ and $\text{Ball}_D(\boldsymbol{x}, \delta) = \left\{\boldsymbol{z} \in \mathbb{R}^D : ||\boldsymbol{z} - \boldsymbol{x}||_2 \leq \delta\right\}$ is a D-dimensional ball of radius δ centered at $\boldsymbol{x}$. To ensure that R_d is well behaved, we impose the following assumptions on the underlying density p around a small neighborhood of R_d.

Assumption 2.1 (Differentiability)**.** *The density function p is bounded and at least four times differentiable with bounded partial derivatives up to the fourth order for every $\boldsymbol{x} \in \mathbb{R}^D$.*

Assumption 2.2 (Eigengap). *There exist constants $\rho > 0$ and $\beta_0 > 0$ such that $\lambda_{d+1}(\boldsymbol{y}) \leq -\beta_0$ and $\lambda_d(\boldsymbol{y}) - \lambda_{d+1}(\boldsymbol{y}) \geq \beta_0$ for any $\boldsymbol{y} \in R_d \oplus \rho$.*

Assumption 2.3 (Path smoothness). *Under the same $\rho, \beta_0 > 0$ in Assumption 2.2, there exists another constant $\beta_1 \in (0, \beta_0)$ such that*

$$D^{\frac{3}{2}} \left|\left|U_d^{\perp}(\boldsymbol{y}) \nabla p(\boldsymbol{y})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{y})\right|\right|_{\max} \leq \frac{\beta_0^2}{2},$$
$$d \cdot D^{\frac{3}{2}} \left|\left|\nabla p(\boldsymbol{x})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x})\right|\right|_{\max} \leq \beta_0(\beta_0 - \beta_1)$$

for all $\boldsymbol{y} \in R_d \oplus \rho$ and $\boldsymbol{x} \in R_d$.

Assumption 2.1 is a standard smoothness condition for ridge estimation. Assumption 2.2 is a curvature condition on the true density p, ensuring that p is "strongly concave" around R_d inside the $(D-d)$-dimensional linear space spanned by the columns of $V_d(\boldsymbol{y})$. We call this property "subspace constrained strong concavity". This condition is also important in establishing the linear convergence of the SCGA and SCMS algorithms; see Remark 3.3 for details. Assumption 2.3 prevents the gradient and third-order derivatives of p from becoming too steep around the ridge R_d. These assumptions are also imposed by Genovese et al. (2014) for characterizing a quadratic behavior of p around R_d and ensuring the stability of R_d, as well as by Chen et al. (2015) to avoid the degenerate normal spaces of R_d. Consequently, R_d is a d-dimensional manifold that contains neither intersections nor endpoints; see also Lemma A.1 in the Appendix. The inequalities in Assumption 2.3 depend on both the ambient dimension D and the intrinsic dimension d of the ridge R_d. As D or d increases, these inequalities become more restrictive. This behavior reflects the *curse of dimensionality* in nonparametric ridge estimation.

Given Assumptions 2.1–2.3, the ridge R_d is stable under small perturbations of the underlying density p and its derivatives, as summarized in the following lemma. The stability of R_d is measured by the Hausdorff distance defined as:

$$\texttt{Haus}(A, B) = \inf\left\{\epsilon > 0 : A \subset B \oplus \epsilon \text{ and } B \subset A \oplus \epsilon\right\}, \tag{2.22}$$

where A, B are two sets in $\mathbb{R}^D$.

Lemma 2.1 (Theorem 4 in Genovese et al. 2014). *Suppose that Assumptions 2.1, 2.2, and 2.3 hold for two densities p_1, p_2. When $||p_1 - p_2||^*_{\infty,3}$ is sufficiently small, we have*

$$\texttt{Haus}(R_{d,1}, R_{d,2}) = O\left(||p_1 - p_2||^*_{\infty,2}\right),$$

where $R_{d,1}$ and $R_{d,2}$ are the d-ridges of p_1 and p_2, respectively.

2.3.2 Directional Density Ridge and its Stability

We now extend the Euclidean definition (2.19) to the unit hypersphere Ω_q with directional density $f : \Omega_q \to \mathbb{R}$ using its Riemannian gradient $\texttt{grad}\, f(\boldsymbol{x})$ and Hessian $\mathcal{H}f(\boldsymbol{x})$ on Ω_q. Specifically, consider the spectral decomposition of $\mathcal{H}f(\boldsymbol{x})$ as $\mathcal{H}f(\boldsymbol{x}) = \underline{V}(\boldsymbol{x})\underline{\Lambda}(\boldsymbol{x})\underline{V}(\boldsymbol{x})^T$, where $\underline{V}(\boldsymbol{x}) = \left[\boldsymbol{x}, \underline{\boldsymbol{v}}_1(\boldsymbol{x}), ..., \underline{\boldsymbol{v}}_q(\boldsymbol{x})\right] \in \mathbb{R}^{(q+1)\times(q+1)}$ is a real orthogonal matrix whose columns $\underline{\boldsymbol{v}}_1(\boldsymbol{x}), ..., \underline{\boldsymbol{v}}_q(\boldsymbol{x})$ are eigenvectors of $\mathcal{H}f(\boldsymbol{x})$ associated with the eigenvalues $\underline{\lambda}_1(\boldsymbol{x}) \geq \cdots \geq \underline{\lambda}_q(\boldsymbol{x})$ and lie within the tangent space $T_{\boldsymbol{x}}$ at $\boldsymbol{x} \in \Omega_q$, and $\underline{\Lambda}(\boldsymbol{x}) = \texttt{Diag}\left[0, \underline{\lambda}_1(\boldsymbol{x}), ..., \underline{\lambda}_q(\boldsymbol{x})\right]$. Note that the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ has a unit eigenvector $\boldsymbol{x}$ that is orthogonal to $T_{\boldsymbol{x}}$ and corresponds to eigenvalue 0. Let $\underline{V}_d(\boldsymbol{x}) = \left[\underline{\boldsymbol{v}}_{d+1}(\boldsymbol{x}), ..., \underline{\boldsymbol{v}}_q(\boldsymbol{x})\right] \in \mathbb{R}^{(q+1)\times(q-d)}$ be the last $q-d$ columns of $\underline{V}(\boldsymbol{x})$, *i.e.*, the unit eigenvectors inside the tangent space $T_{\boldsymbol{x}}$ corresponding to the $q-d$ smallest eigenvalues of $\mathcal{H}f(\boldsymbol{x})$. Let $\underline{U}_d(\boldsymbol{x}) = \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T$ be the projection matrix onto the linear space spanned by the columns of $\underline{V}_d(\boldsymbol{x})$, and $\underline{U}_d^{\perp}(\boldsymbol{x}) = \boldsymbol{I}_{q+1} - \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T$. We define the order-$d$ principal Riemannian gradient $\underline{G}_d(\boldsymbol{x})$ by

$$\begin{aligned}\underline{G}_d(\boldsymbol{x}) = \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T\texttt{grad}\, f(\boldsymbol{x}) &= \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T\left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla f(\boldsymbol{x}) \\ &= \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T\nabla f(\boldsymbol{x}),\end{aligned} \tag{2.23}$$

where the last equality follows from the fact that the columns of $\underline{V}_d(\boldsymbol{x})$ are orthogonal to the unit vector $\boldsymbol{x}$. The order-d density ridge on Ω_q, or *directional density ridge*, is defined as

$$\begin{aligned}\underline{R}_d &= \left\{\boldsymbol{x} \in \Omega_q : \underline{G}_d(\boldsymbol{x}) = \boldsymbol{0}, \underline{\lambda}_{d+1}(\boldsymbol{x}) < 0\right\} \\ &= \left\{\boldsymbol{x} \in \Omega_q : \underline{V}_d(\boldsymbol{x})^T\texttt{grad}\, f(\boldsymbol{x}) = \boldsymbol{0}, \underline{\lambda}_{d+1}(\boldsymbol{x}) < 0\right\}.\end{aligned} \tag{2.24}$$

Our definition of density ridges on Ω_q can be generalized to any smooth function f supported on an arbitrary Riemannian manifold. It also follows that the order-0 directional

density ridge $\underline{R}_0$ is the set of local modes of f on Ω_q as:

$$\underline{\mathcal{M}} := \underline{R}_0 = \{\boldsymbol{x} \in \Omega_q : \texttt{grad}\, f(\boldsymbol{x}) = \boldsymbol{0}, \underline{\lambda}_1(\boldsymbol{x}) < 0\}\,. \tag{2.25}$$

To study the stability of the directional density ridge $\underline{R}_d$, we first extend f from its support Ω_q to $\mathbb{R}^{q+1} \setminus \{\boldsymbol{0}\}$ by defining

$$f(\boldsymbol{x}) \equiv f\left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}\right) \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^{q+1} \setminus \{\boldsymbol{0}\}. \tag{2.26}$$

This extension in (2.26) has also been used by Zhao and Wu (2001); García-Portugués et al. (2013); García-Portugués (2013). Then, we impose the following regularity conditions.

Assumption 2.1. *Under the extension (2.26) of the directional density f, the total gradient $\nabla f(\boldsymbol{x})$, total Hessian matrix $\nabla\nabla f(\boldsymbol{x})$, and third-order derivative tensor $\nabla^3 f(\boldsymbol{x})$ in $\mathbb{R}^{q+1}$ exist and are continuous on $\mathbb{R}^{q+1} \setminus \{\boldsymbol{0}\}$ and square integrable on Ω_q. Furthermore, f has bounded fourth-order derivatives on Ω_q.*

Assumption 2.2. *There exist constants $\underline{\rho} > 0$ and $\underline{\beta}_0 > 0$ such that $\underline{\lambda}_{d+1}(\boldsymbol{y}) \leq -\underline{\beta}_0$ and $\underline{\lambda}_d(\boldsymbol{y}) - \underline{\lambda}_{d+1}(\boldsymbol{y}) \geq \underline{\beta}_0$ for any $\boldsymbol{y} \in \left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$.*

Assumption 2.3. *Under the same $\underline{\rho}, \underline{\beta}_0 > 0$ in Assumption 2.2, there exists another constant $\underline{\beta}_1 \in \left(0, \underline{\beta}_0\right)$ such that*

$$\sqrt{2} \cdot q^{\frac{3}{2}} \left|\left|\underline{U}_d^{\perp}(\boldsymbol{y})\texttt{grad}\, f(\boldsymbol{y})\right|\right|_2 \left|\left|\nabla^3 f(\boldsymbol{y})\right|\right|_{\max} \leq \frac{\underline{\beta}_0^2}{2},$$
$$d \cdot q^{\frac{3}{2}} \left|\left|\nabla f(\boldsymbol{x})\right|\right|_2 \cdot \left|\left|\nabla^3 f(\boldsymbol{x})\right|\right|_{\max} \leq \underline{\beta}_0 \left(\underline{\beta}_0 - \underline{\beta}_1\right)$$

for all $\boldsymbol{y} \in \left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$ and $\boldsymbol{x} \in \underline{R}_d$.

Recall that $\underline{R}_d \oplus \underline{\rho} = \cup_{\boldsymbol{x} \in \underline{R}_d} \text{Ball}_{q+1}\left(\boldsymbol{x}, \underline{\rho}\right)$ is a $\underline{\rho}$-neighborhood of the directional ridge $\underline{R}_d$ in the ambient space $\mathbb{R}^{q+1}$. The discussion of Assumptions 2.1–2.3 in Section 2.3.1 also applies to their directional counterparts 2.1–2.3, except that the eigengap assumption 2.2 is imposed on eigenvalues $\underline{\lambda}_1(\boldsymbol{x}) \geq \cdots \geq \underline{\lambda}_q(\boldsymbol{x})$ within the tangent space $T_{\boldsymbol{x}}$ at $\boldsymbol{x} \in \Omega_q$. However, since the only eigenvalue of $\mathcal{H}f(\boldsymbol{x})$ associated with the eigenvector outside the tangent space $T_{\boldsymbol{x}}$

is 0, Assumption 2.2 also applies to the relevant spectrum of $\mathcal{H}f(\boldsymbol{x})$ in the ambient space $\mathbb{R}^{q+1}$.

Remark 2.2 (Connection to Solution Manifolds)**.** *Example 4 in Chen (2022) showed that any Euclidean density ridge R_d defined in (2.19) is a concrete example of a solution manifold $\mathcal{M}_S = \{\boldsymbol{x} \in \mathbb{R}^D : \Psi(\boldsymbol{x}) = 0\}$ with $\Psi : \mathbb{R}^D \to \mathbb{R}^{D-d}$ being a vector-valued function. The directional density ridge $\underline{R}_d$ in (2.24) also takes the form of the solution manifold $\mathcal{M}_S$, where we may rewrite $\underline{R}_d = \{\boldsymbol{x} \in \mathbb{R}^{q+1} : \Psi(\boldsymbol{x}) = 0\}$ with $\Psi : \mathbb{R}^{q+1} \to \mathbb{R}^{q+1-d}$ defined by:*

$$\Psi(\boldsymbol{x}) = \begin{bmatrix} \underline{\boldsymbol{v}}_{d+1}(\boldsymbol{x})^T \nabla f(\boldsymbol{x}) \\ \vdots \\ \underline{\boldsymbol{v}}_q(\boldsymbol{x})^T \nabla f(\boldsymbol{x}) \\ \boldsymbol{x}^T \boldsymbol{x} - 1 \end{bmatrix},$$

recalling that $\underline{\boldsymbol{v}}_{d+1}(\boldsymbol{x}), ..., \underline{\boldsymbol{v}}_q(\boldsymbol{x})$ are the last $q-d$ eigenvectors of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ of the directional density f. Moreover, the assumptions imposed here 2.1–2.3 in the Euclidean ridge case and 2.1–2.3 in the directional ridge case imply all the required assumptions in Chen (2022), i.e., *the differentiability of Ψ and non-degeneracy of the normal space of $\mathcal{M}_S$; see (d) of Lemmas A.1 and A.2 in the Appendix. Therefore, the statistical properties and normal gradient flows of a generic solution manifold $\mathcal{M}_S$ apply to the Euclidean and directional density ridges R_d or $\underline{R}_d$ here.*

Analogously to the Euclidean case in Lemma 2.1, we obtain the following stability theorem for directional density ridges. To measure the distance between two directional ridges $\underline{R}_d, \underline{\widetilde{R}}_d \subset \Omega_q$ defined by the directional densities f and $\widetilde{f}$, we adopt the definition (2.22) of Hausdorff distance between two sets in the ambient Euclidean space $\mathbb{R}^{q+1}$. Note that the Euclidean norm used in the definition (2.22) is upper bounded by the geodesic distance when the sets of interest lie on Ω_q; see also (2.6). This property is used in our proof of Theorem 2.2; see Section A.4 for details.

Theorem 2.2. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold for the directional density f and that Assumption 2.1 holds for $\widetilde{f}$. When $\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,3}$ is sufficiently small,*

(a) Assumptions 2.2 and 2.3 hold for $\widetilde{f}$.

(b) $\texttt{Haus}(\underline{R}_d, \widetilde{\underline{R}}_d) = O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}\right)$.

(c) $\texttt{reach}(\widetilde{\underline{R}}_d) \geq \min\left\{\underline{\rho}/2, \frac{\min\left\{\underline{\beta}_1, 1\right\}^2}{\underline{A}_2\left(||f||^{(3)}_\infty + ||f||^{(4)}_\infty\right)}\right\} + O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,3}\right)$ *for a constant* $\underline{A}_2 > 0$.

The stability result for the density ridge $\mathcal{R}_d$ on $\mathcal{S}_1 \times \mathcal{S}_2$ follows similarly from Lemma 2.1 and Theorem 2.2, so we omit the details.

2.3.3 Plug-in Estimator by KDE

Replacing the unknown density p that generates the Euclidean data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \mathbb{R}^D$ with the Euclidean KDE $\widehat{p}_n$ in the definition (2.19) of density ridges gives the natural plug-in estimator of the true ridge R_d as:

$$\widehat{R}_d = \left\{\boldsymbol{x} \in \mathbb{R}^D : \widehat{V}_d(\boldsymbol{x})^T \nabla \widehat{p}_n(\boldsymbol{x}) = \boldsymbol{0}, \widehat{\lambda}_{d+1}(\boldsymbol{x}) < 0\right\}.$$

To control the statistical behavior of the estimated ridge $\widehat{R}_d$, we make the following assumptions on the kernel of its form (2.2):

Assumption E1. *We assume that the kernel profile* $k : [0, \infty) \to [0, \infty)$ *is non-increasing and at least three times continuously differentiable with bounded fourth-order partial derivatives as well as*

$$\int_{\mathbb{R}^D} ||\boldsymbol{x}||_2^2 \cdot k\left(||\boldsymbol{x}||_2^2\right) d\boldsymbol{x}, \quad \int_{\mathbb{R}^d} -k'\left(||\boldsymbol{x}||_2^2\right) d\boldsymbol{x} < \infty, \quad \textit{and} \quad \int_{\mathbb{R}^d} \left|k^{(\alpha)}\left(||\boldsymbol{x}||_2^2\right)\right|^2 d\boldsymbol{x} < \infty$$

with $\alpha = 0, 1, 2, 3$.

Assumption E2. *Let*

$$\mathcal{K}_E = \left\{\boldsymbol{y} \mapsto K^{(\alpha)}\left(\frac{\boldsymbol{x} - \boldsymbol{y}}{h}\right) = D^{[\alpha]}\left[c_{k,D} \cdot k\left(\left|\left|\frac{\boldsymbol{x} - \boldsymbol{y}}{h}\right|\right|_2^2\right)\right] : \boldsymbol{x} \in \mathbb{R}^D, |[\alpha]| = 0, 1, 2, 3\right\}.$$

We assume that $\mathcal{K}_E$ is a bounded VC (subgraph) class of measurable functions on $\mathbb{R}^D$; that is, there exist constants $A, \upsilon > 0$ such that for any $0 < \epsilon < 1$,

$$\sup_Q N\left(\mathcal{K}_E, L_2(Q), \epsilon \left|\left|F\right|\right|_{L_2(Q)}\right) \leq \left(\frac{A}{\epsilon}\right)^{\upsilon},$$

where $N(\mathcal{K}_E, L_2(Q), \epsilon)$ is the ϵ-covering number of the normed space $\left(\mathcal{K}_E, \left|\left|\cdot\right|\right|_{L_2(Q)}\right)$, Q is any probability measure on $\mathbb{R}^D$, and F is an envelope function of $\mathcal{K}_E$. Here, the norm $\left|\left|F\right|\right|_{L_2(Q)}$ is defined as $\left[\int_{\mathbb{R}^D} |F(\boldsymbol{x})|^2 dQ(\boldsymbol{x})\right]^{\frac{1}{2}}$.

Remark 2.3. *Recall that the* ϵ-covering number *$N(\mathcal{F}, L_2(Q), \epsilon)$ is defined as the minimal number of $L_2(Q)$-balls $\left\{g : \left|\left|g - f\right|\right|_{L_2(Q)} < \epsilon\right\}$ of radius ϵ needed to cover the (function) class $\mathcal{F}$. One popular concept for controlling uniform covering number $\sup_Q N\left(\mathcal{F}, L_2(Q), \epsilon \left|\left|F\right|\right|_{L_2(Q)}\right)$ is the notion of Vapnik-Červonenkis (subgraph) classes, or simply VC classes. Starting from collections of sets, we say that a collection $\mathcal{C}$ of subsets of the sample space $\mathcal{X}$* picks out *a certain subset of the finite set $\{x_1, ..., x_n\} \subset \mathcal{X}$ if it can be written as $C \cap \{x_1, ..., x_n\}$ for some $C \in \mathcal{C}$. The collection is said to* shatter *$\{x_1, ..., x_n\}$ if $\mathcal{C}$ picks out each of its 2^n subsets. The* VC-index *$V(\mathcal{C})$ of $\mathcal{C}$ is the smallest n for which no set of size n is shattered by $\mathcal{C}$. A collection $\mathcal{C}$ of measurable sets is called a* VC class *if its index $V(\mathcal{C})$ is finite. To generalize this concept to a class $\mathcal{F}$ of real-valued and measurable functions defined on $\mathcal{X}$, we say that $\mathcal{F}$ is a* VC subgraph class *if the collection of all subgraphs of the functions in $\mathcal{F}$ forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$. An important property of VC (subgraph) classes is that their ϵ-covering numbers grow polynomially in $\frac{1}{\epsilon}$ as stated in Assumption E2; see Theorem 2.6.4 in van der Vaart and Wellner (1996). More in-depth discussion on VC classes can be found in Section 2.6 of the same book.*

Assumption E1 can be relaxed to require that the kernel profile k be three times continuously differentiable except at a finite number of points on $[0, \infty)$. Such a relaxation allows us to include the Epanechnikov and other compactly supported kernels. The integrability assumption on k in Assumption E1 is similar to conditions (K1) and (K3) in Section 2.2.1 for the purpose of bounding the expectations and variances of the KDE $\nabla\widehat{p}_n(\boldsymbol{x})$ and its (partial)

derivatives. Assumption E2 controls the complexity of the kernel and its (partial) derivatives, which is essential in establishing the uniform convergence of $\widehat{p}_n$ and its derivatives to the corresponding quantities of p as in (2.27) below.

Given Assumptions E1 and E2, the techniques in Giné and Guillou (2002); Einmahl and Mason (2005); Chacón et al. (2011) can be used to show the uniform consistency of the Euclidean KDE $\widehat{p}_n$ and its derivatives as:

$$||\widehat{p}_n - p||_\infty^{(k)} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+2k}}}\right) \quad \text{for } k = 0, ..., 3. \tag{2.27}$$

For the directional density ridge $\underline{R}_d$ in (2.24), we similarly obtain a plug-in estimator from the directional KDE $\widehat{f}_h$ as:

$$\widehat{\underline{R}}_d = \left\{\boldsymbol{x} \in \Omega_q : \widehat{\underline{V}}_d(\boldsymbol{x})^T \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) = \boldsymbol{0}, \widehat{\underline{\lambda}}_{d+1}(\boldsymbol{x}) < 0\right\}. \tag{2.28}$$

To regularize the statistical behavior of the estimated directional ridge $\widehat{\underline{R}}_d$, we consider the following assumptions that generalize Assumptions E1 and E2:

Assumption D1. *Assume that $L : (-\delta_L, \infty) \to [0, \infty)$ is a bounded and three times continuously differentiable function with a bounded fourth-order derivative on $(-\delta_L, \infty) \subset \mathbb{R}$ for some constant $\delta_L > 0$ such that*

$$0 < \int_0^\infty |L^{(\ell)}(r)|^k r^{\frac{q}{2}-1} dr < \infty \quad \textit{for all } q \geq 1, k = 1, 2, \textit{ and } \ell = 0, 1, 2, 3.$$

Assumption D2. *Let*

$$\mathcal{K}_D = \left\{\boldsymbol{u} \mapsto K\left(\frac{\boldsymbol{z}-\boldsymbol{u}}{h}\right) : \boldsymbol{u}, \boldsymbol{z} \in \Omega_q, h > 0, K(\boldsymbol{x}) = D^{[\tau]}L\left(\frac{||\boldsymbol{x}||_2^2}{2}\right), |[\tau]| = 0, 1, 2, 3\right\}.$$

We assume that $\mathcal{K}_D$ is a bounded VC (subgraph) class of measurable functions on Ω_q; that is, there exist constants $A, \upsilon > 0$ such that for any $0 < \epsilon < 1$,

$$\sup_Q N\left(\mathcal{K}_D, L_2(Q), \epsilon||F||_{L_2(Q)}\right) \leq \left(\frac{A}{\epsilon}\right)^\upsilon,$$

where $N(\mathcal{K}_D, L_2(Q), \epsilon)$ is the ϵ-covering number of the normed space $\left(\mathcal{K}_D, ||\cdot||_{L_2(Q)}\right)$, Q is any probability measure on Ω_q, and F is an envelope function of $\mathcal{K}_D$. Here, the norm $||F||_{L_2(Q)}$ is defined as $\left[\int_{\Omega_q} |F(\boldsymbol{x})|^2 dQ(\boldsymbol{x})\right]^{\frac{1}{2}}$.

The differentiability assumption in Assumption D1 can be relaxed to require that L be (three times) continuously differentiable except on a set of points with Lebesgue measure 0 on $[0, \infty)$. Assumptions D1 and 2.1 are standard conditions for establishing the pointwise convergence rates of the directional KDE and its derivatives (Hall et al., 1987; Klemelä, 2000; Zhao and Wu, 2001; García-Portugués et al., 2013; García-Portugués, 2013). Under these two assumptions, $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2$ appearing in the step sizes $\underline{\eta}_{n,h}^{(t)}$ or $\underline{\eta}_{n,h}^{(t)'}$ of the directional mean shift or SCMS algorithm can also be shown to diverge at the order $O(h^{-2}) + O_P\left(\frac{1}{nh^{q+2}}\right)$ as $nh^q \to \infty$ and $h \to 0$; see Section 3.4 and Section 3.6 for details. Assumption D2 regularizes the complexity of the kernel L and its derivatives as in Assumption E2 to obtain the uniform convergence rates of the directional KDE and its derivatives; see (2.29) below. One can justify via integration by parts that the von-Mises kernel $L(r) = e^{-r}$ and many compactly supported kernels satisfy Assumptions D1 and D2.

Given Assumptions D1 and D2, the techniques in Hall et al. (1987); Bai et al. (1988); Zhao and Wu (2001); García-Portugués et al. (2013); García-Portugués (2013); Zhang and Chen (2021b) can be used to show that

$$\left|\left|\widehat{f}_h - f\right|\right|_\infty^{(k)} = \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\bar{\nabla}^k \widehat{f}_h(\boldsymbol{x}) - \bar{\nabla}^k f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2k}}}\right) \tag{2.29}$$

for $k = 0, ..., 3$, where $\bar{\nabla} \equiv \bar{\nabla}_{\boldsymbol{v}}$ is the Riemannian connection on Ω_q with $\boldsymbol{v} \in T_{\boldsymbol{x}}$ so that $\bar{\nabla} f(\boldsymbol{x}) = \texttt{grad}\, f(\boldsymbol{x})$, $\bar{\nabla}^2 f(\boldsymbol{x}) = \mathcal{H} f(\boldsymbol{x})$, and $\bar{\nabla}^3 f(\boldsymbol{x}) = \bar{\nabla}\mathcal{H} f(\boldsymbol{x})$; see Section 5.3 in Absil et al. (2008) and Chapter 4 in Lee (2018). Specifically, the uniform convergence rates for the Riemannian gradient and Hessian of the directional KDE $\widehat{f}_h$ are derived in Theorem 4 of Zhang and Chen (2021b). Finally, Lemma B.2 in Zhang and Chen (2026) also outlines the uniform convergence rates for the Riemannian gradient and Hessian of the KDE $\widehat{f}_{\boldsymbol{h}}$ within the tangent space/bundle on $\mathcal{S}_1 \times \mathcal{S}_2$.

2.4 Statistical Consistency Results

2.4.1 Mode Consistency on Ω_q

Consistency of local mode estimation has been established for the usual Euclidean KDE by Chen et al. (2016), where the authors showed that with probability tending to 1, the number of estimated local modes is the same as the number of true local modes under appropriate assumptions. Moreover, the convergence rate of the Hausdorff distance (a common distance between two sets) between the collection of local modes and its estimator is derived. Here we state the consistency of estimating local modes of a directional density f supported on Ω_q by the local modes of the directional KDE $\widehat{f}_h$.

Let $||f||^*_{\infty,3}$ denote the uniform functional norm of the partial derivatives of the directional density f on the compact manifold Ω_q up to the third order. This quantity is finite under Assumption D1. Let $\widehat{\underline{\mathcal{M}}}_n = \{\widehat{\underline{\boldsymbol{m}}}_1, ..., \widehat{\underline{\boldsymbol{m}}}_{\widehat{K}_n}\}$ be the collection of local modes of $\widehat{f}_h$ and $\underline{\mathcal{M}} = \{\underline{\boldsymbol{m}}_1, ..., \underline{\boldsymbol{m}}_K\}$ be the collection of local modes of f. Here, $\widehat{K}_n$ is the number of estimated local modes and K is the number of true local modes. We consider the following assumptions.

Assumption M1. *There exists $\lambda_* > 0$ such that*

$$0 < \lambda_* \leq |\lambda_1(\underline{\boldsymbol{m}}_j)|, \quad \text{for all } j = 1, ..., K,$$

where $0 > \lambda_1(\boldsymbol{x}) \geq \cdots \geq \lambda_q(\boldsymbol{x})$ are the q most negative eigenvalues of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$.

Assumption M2. *There exist $\Theta_1, \rho_* > 0$ such that*

$$\left\{\boldsymbol{x} \in \Omega_q : ||\texttt{Tang}(\nabla f(\boldsymbol{x}))||_{\max} \leq \Theta_1, \lambda_1(\boldsymbol{x}) \leq -\frac{\lambda_*}{2} < 0\right\} \subset \underline{\mathcal{M}} \oplus \rho_*,$$

where λ_* is defined in Assumption M1 and $0 < \rho_* < \min\left\{\sqrt{2 - 2\cos\left(\frac{3\lambda_*}{2||f||^*_{\infty,3}}\right)}, 2\right\}$.

Assumption M1 is imposed so that every local mode of f is isolated from other critical points; see Lemma 3.2 in Banyaga and Hurtubise (2004). The assumption also guarantees

that the number of local modes of f supported on the compact manifold Ω_q is finite. As noted by Chen et al. (2016), Assumption M1 always holds when f is a Morse function on Ω_q. Assumption M2 controls the behavior of f so that points with near-zero (Riemannian) gradients and negative eigenvalues of $\mathcal{H}f(\boldsymbol{x})$ within the tangent space $T_{\boldsymbol{x}}$ must be close to local modes. See the paper by Chen et al. (2016) for a detailed discussion. The constant $\sqrt{2-2\cos\left(\frac{3\lambda_*}{2||f||^*_{\infty,3}}\right)}$ is selected so that the great-circle distance from $\underline{\boldsymbol{m}}_k$ to the boundary of $\underline{\boldsymbol{m}}_k \oplus \rho_*$, that is, $\arccos(\underline{\boldsymbol{m}}_k^T \boldsymbol{x})$ with $\boldsymbol{x} \in \partial S_k$, is less than $\frac{3\lambda_*}{2||f||^*_{\infty,3}}$ for any $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$, where $S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \underline{\boldsymbol{m}}_k||_2 \leq \rho_*\}$ and $\partial S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \underline{\boldsymbol{m}}_k||_2 = \rho_*\}$.

Theorem 2.3. *Assume that Assumptions D1, D2, M1, and M2 hold. For any $\delta \in (0,1)$, when h is sufficiently small and n is sufficiently large,*

(a) there must be at least one estimated local mode $\widehat{\underline{\boldsymbol{m}}}_k$ within $S_k = \underline{\boldsymbol{m}}_k \oplus \rho_*$ for every $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$, and

(b) the collection of estimated modes satisfies $\widehat{\underline{\mathcal{M}}}_n \subset \underline{\mathcal{M}} \oplus \rho_*$ and there is a unique estimated local mode $\widehat{\underline{\boldsymbol{m}}}_k$ within $S_k = \underline{\boldsymbol{m}}_k \oplus \rho_*$

with probability at least $1-\delta$. In total, when h is sufficiently small and n is sufficiently large, there exist constants $A_3, B_3 > 0$ such that

$$\mathbb{P}\left(\widehat{K}_n \neq K\right) \leq B_3 e^{-A_3 n h^{q+4}}.$$

(c) The Hausdorff distance between the collection of local modes and its estimator satisfies

$$\texttt{Haus}\left(\underline{\mathcal{M}}, \widehat{\underline{\mathcal{M}}}_n\right) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right),$$

as $h \to 0$ and $nh^{q+2} \to \infty$.

The proof of Theorem 2.3 appears in Appendix D.4 of Zhang and Chen (2021b). The theorem states that, asymptotically, the set of estimated local modes is close to the set of true local modes and there is a one-to-one correspondence between pairs of estimated and

true local modes. Thus, the local modes of the directional KDE consistently estimate the local modes of the population directional density. The mode consistency for the KDE $\widehat{f}_{\boldsymbol{h}}$ on $\mathcal{S}_1 \times \mathcal{S}_2$ is stated in Theorem B.3 in Zhang and Chen (2026).

2.4.2 Ridge Consistency on Ω_q

Given the plug-in estimator $\widehat{\underline{R}}_d$ in (2.28) of the ridge provided by the directional KDE $\widehat{f}_h$, we apply Theorem 2.2 to obtain consistency under the Hausdorff distance.

Theorem 2.4. *Suppose that Assumptions 2.1–2.3 and D1–D2 hold. Then, when* $\left|\left|\widehat{f}_h - f\right|\right|^*_{\infty,3}$ *is sufficiently small,*

$$\texttt{Haus}(\widehat{\underline{R}}_d, \underline{R}_d) = O\left(h^2\right) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right).$$

The proof follows directly from Theorem 2.2 and the uniform convergence rates of the directional KDE $\widehat{f}_h$ in (2.29), so we omit the details. Finally, Theorem B.4 in Zhang and Chen (2026) outlines the ridge consistency for $\widehat{f}_{\boldsymbol{h}}$ on $\mathcal{S}_1 \times \mathcal{S}_2$.

Chapter 3

ALGORITHMIC THEORY FOR SUBSPACE CONSTRAINED MEAN SHIFT ALGORITHMS

Building on the statistical theory developed in Chapter 2, this chapter studies the computational problem of recovering local modes and density ridges from data with Euclidean and spherical geometry. The central procedures are the directional mean shift (DMS) algorithm for estimating directional modes and the directional subspace constrained mean shift (DirSCMS) algorithm for estimating directional density ridges. In the cosmic web application, these procedures are not merely numerical devices, because the DMS algorithm identifies high-density cosmic nodes and the DirSCMS algorithm traces filamentary structures on the celestial sphere. Therefore, establishing their convergence properties provides algorithmic support for the stability and reproducibility of the cosmic web catalog constructed in Chapter 4.

3.1 Introduction

Chapter 2 introduced directional modes and density ridges as effective geometric features of the underlying density function and established the statistical consistency of their KDE-based estimators. The present chapter turns to the complementary computational question: how can these geometric features be recovered reliably from observed point clouds?

In Euclidean spaces, the mean shift algorithm and its subspace constrained version provide natural mode-seeking and ridge-finding procedures for the KDE-based estimators. The classical mean shift algorithm with Euclidean data was proposed by Fukunaga and Hostetler (1975) and revisited for image classification in Comaniciu and Meer (2002); see Carreira-Perpiñán (2015) for a comprehensive review. Its convergence properties have been well-

studied in Cheng (1995); Li et al. (2007); Ghassabeh (2013, 2015); Arias-Castro et al. (2016); Wang et al. (2016). For common kernels, including Gaussian and Epanechnikov kernels, the algorithm typically exhibits linear convergence except under extreme bandwidth choices (Carreira-Perpiñán, 2007; Huang et al., 2018). Several acceleration strategies have also been proposed, including dynamic data updates, blurring, random sampling, and stochastic optimization variants (Zhang et al., 2006; Carreira-Perpiñán, 2006, 2008; Yuan and Li, 2009; Hyrien and Baran, 2016). For directional data, the mean shift (or DMS) algorithms have been developed in several forms Oba et al. (2005); Kafai et al. (2010); Kobayashi and Otsu (2010); Shou-Jen Chang-Chien et al. (2010); Chang-Chien et al. (2012); Yang et al. (2014). More broadly, Tuzel et al. (2005); Subbarao and Meer (2006, 2009); Cetingul and Vidal (2009); Caseiro et al. (2012); Ashizawa et al. (2017) proposed mean shift algorithms on manifolds using logarithmic and exponential maps, heat kernels, or direct log-density estimation via least squares. Nevertheless, global convergence and linear convergence guarantees for the DMS algorithm remain much less developed than their Euclidean counterparts.

The algorithmic theory is even more limited for the subspace constrained mean shift (SCMS) algorithm, which is designed to estimate density ridges rather than local modes. Existing work has established monotonicity properties of the density estimate and justified certain stopping criteria (Ghassabeh, 2013; Ghassabeh and Rudzicz, 2021), but a general convergence theory for SCMS and its directional extension is still incomplete. There are two main challenges in establishing such a theory. First, each SCMS iteration involves a projection matrix determined by the Hessian of the estimated density. Consequently, although the update resembles a projected gradient step, the algorithm is not a standard first-order optimization method: the projection subspace itself changes with the current iterate and depends on second-order information. Second, ridge estimation is intrinsically nonconvex. The target set is defined by both first-order projected gradients and second-order curvature constraints, so the convergence analysis should account for both the dynamics of the iterates and the local geometry of the ridge.

The need for geometry-aware algorithms is illustrated in Figure 3.1. Consider a synthetic

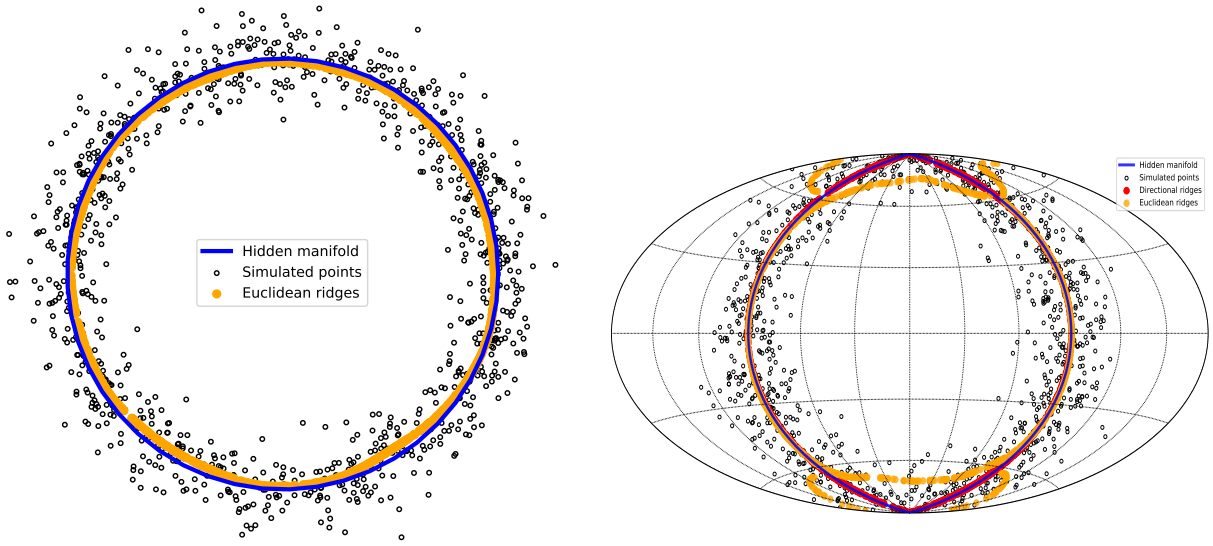


Figure 3.1: Density ridges estimated by Euclidean and directional SCMS algorithms on two synthetic datasets (drawn as black points) with hidden circular manifold structures (indicated by blue curves) on $\mathbb{R}^2$ and the unit sphere $\Omega_2 \subset \mathbb{R}^3$, respectively. **Left**: The orange points indicate the estimated ridge obtained by the Euclidean SCMS algorithm from the dataset on $\mathbb{R}^2$. **Right**: The red points represent the estimated directional ridge identified by our directional SCMS algorithm, while the orange points indicate the estimated ridge obtained by the Euclidean SCMS algorithm from the dataset on Ω_2. This panel is presented under the Hammer projection; see Appendix A.1 for more details.

dataset consisting of independent observations concentrated around a great circle on Ω_2 that connects the North and South Poles with additional noise. Applying the Euclidean SCMS algorithm directly to these directional data introduces substantial bias near high-latitude regions, where the Euclidean representation distorts the spherical geometry. In contrast, the proposed directional SCMS (DirSCMS) algorithm recovers a ridge that closely follows the underlying circular structure; see the right panel of Figure 3.1 for an illustration and Section A.1 for a more detailed discussion.

To address these challenges, this chapter provides an algorithmic foundation for the ridge-based statistical framework developed in Chapter 2. We first study the Euclidean SCMS

algorithm through a more general subspace constrained gradient ascent (SCGA) framework, which isolates the essential mechanism behind SCMS and yields local linear convergence guarantees under suitable regularity conditions. In particular, for a sequence $\{\widehat{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ generated by the Euclidean SCGA or SCMS algorithm, we establish bounds of the form

$$\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \widehat{\boldsymbol{x}}^{*}\right|\right|_2 \leq \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \widehat{\boldsymbol{x}}^{*}\right|\right|_2,$$

where $\widehat{\boldsymbol{x}}^{*}$ is the limit point of the sequence, and $0 < \Upsilon < 1$ is a constant. We then develop directional versions of SCGA and SCMS for data supported on Ω_q, where the update must respect the intrinsic spherical geometry of the sample space. For the resulting sequence $\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\}_{t=0}^{\infty}$, we prove analogous linear convergence guarantees with respect to the geodesic distance:

$$d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \widehat{\underline{\boldsymbol{x}}}^{*}) \leq \underline{\Upsilon}^t \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(0)}, \widehat{\underline{\boldsymbol{x}}}^{*}),$$

where $\widehat{\underline{\boldsymbol{x}}}^{*}$ is the convergence point, $0 < \underline{\Upsilon} < 1$ is a constant, and d_g is the geodesic distance on Ω_q. Similar arguments can be extended to directional-linear and directional-directional product spaces; since these extensions are not needed for the cosmic web reconstruction considered in this dissertation, we omit their detailed exposition and refer interested readers to Zhang and Chen (2026). The chapter concludes with several numerical experiments to consolidate our theory.

3.2 Directional Mean Shift Algorithm

We begin by reviewing the Euclidean mean shift algorithm. Given Assumption E1 and the Euclidean KDE $\widehat{p}_n(\boldsymbol{x}) = \frac{c_{k,D}}{nh^D} \sum_{i=1}^{n} k\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ with kernel (2.2), its gradient estimator takes the form as:

$$\begin{aligned}
\nabla \widehat{p}_n(\boldsymbol{x}) &= \frac{2c_{k,D}}{nh^{D+2}} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot k'\left(\left|\left|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right) \\
&= \frac{2c_{k,D}}{nh^{D+2}} \left[\sum_{i=1}^{n} -k'\left(\left|\left|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right] \left[\frac{\sum_{i=1}^{n} \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n} k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right],
\end{aligned} \tag{3.1}$$

where the first term is a variant of KDEs and the second term is the *mean shift* vector

$$\Xi_h(\boldsymbol{x}) = \frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}. \tag{3.2}$$

This factorization suggests that the mean shift vector aligns with the direction of maximum increase in $\widehat{p}_n$. Thus, moving a point along its mean shift vector successively yields an ascending path to a local mode. Let $\left\{\widehat{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ be the mean shift sequence with the Euclidean KDE $\widehat{p}_n$. Then, one iteration of the mean shift algorithm can be written as:

$$\begin{aligned}\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \widehat{\boldsymbol{x}}^{(t)} + \Xi_h\left(\widehat{\boldsymbol{x}}^{(t)}\right) &= \frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} \\ &= \widehat{\boldsymbol{x}}^{(t)} + \frac{1}{\frac{2c_{k,D}}{nh^{D+2}}\sum_{i=1}^n -k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} \cdot \nabla\widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)}),\end{aligned} \tag{3.3}$$

showing that the mean shift algorithm is a gradient ascent method with an adaptive step size

$$\eta_{n,h}^{(t)} = \frac{1}{\frac{2c_{k,D}}{nh^{D+2}}\sum_{i=1}^n -k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} = \frac{1}{\widehat{g}_n\left(\widehat{\boldsymbol{x}}^{(t)}\right)}. \tag{3.4}$$

Here, we denote by $\widehat{g}_n\left(\widehat{\boldsymbol{x}}^{(t)}\right) = \frac{2c_{k,D}}{nh^{D+2}}\sum_{i=1}^n -k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ the denominator of the adaptive step size $\eta_{n,h}^{(t)}$. Lemma 3.1 below shows that under Assumptions 2.1 and E1, $h^2\widehat{g}_n\left(\boldsymbol{x}\right)$ tends to a fixed constant with probability tending to 1 for any $\boldsymbol{x} \in \mathbb{R}^D$ as $nh^D \to \infty$ and $h \to 0$. Therefore, the step size $\eta_{n,h}^{(t)}$ has its asymptotic rate as $O(h^2)$ and tends to zero as $nh^D \to \infty$ and $h \to 0$ as well.

Lemma 3.1. *Assume that Assumptions 2.1 and E1 hold. For any fixed $\boldsymbol{x} \in \mathbb{R}^D$, the convergence rate of $\widehat{g}_n(\boldsymbol{x}) = \frac{2c_{k,D}}{nh^{D+2}}\sum_{i=1}^n -k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ is*

$$h^2\widehat{g}_n(\boldsymbol{x}) = -2c_{k,D} \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right) d\boldsymbol{u} + O\left(h^2\right) + O_P\left(\sqrt{\frac{1}{nh^D}}\right)$$

$$= O(1) + O\left(h^2\right) + O_P\left(\sqrt{\frac{1}{nh^D}}\right)$$

as $nh^D \to \infty$ *and* $h \to 0$.

To derive the DMS algorithm, we follow Yang et al. (2014) and assume that the directional kernel L is differentiable except at finitely many points on $[0, \infty)$. To locate the local modes of the directional KDE $\widehat{f}_h(\boldsymbol{x})$ in (2.4), we maximize $\widehat{f}_h(\boldsymbol{x})$ subject to the constraint $\boldsymbol{x}^T\boldsymbol{x} = 1$. By introducing a Lagrangian multiplier λ, we define the Lagrangian function as:

$$\widehat{\mathcal{L}}_h(\boldsymbol{x}, \lambda) = \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) + \lambda(1 - \boldsymbol{x}^T\boldsymbol{x}).$$

Taking the partial derivatives of $\widehat{\mathcal{L}}_h$ with respect to $\boldsymbol{x}$ and λ and setting them to zero yields

$$\frac{\partial \widehat{\mathcal{L}}_h}{\partial \boldsymbol{x}} = -\frac{c_{L,q}(h)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) - 2\lambda\boldsymbol{x} = 0 \quad \text{and} \quad 1 - \boldsymbol{x}^T\boldsymbol{x} = 0.$$

Solving for $\boldsymbol{x}$ leads to the fixed-point iteration $\boldsymbol{x}^{(t+1)} = F(\boldsymbol{x}^{(t)})$ for the DMS algorithm, where

$$F(\boldsymbol{x}) = -\frac{\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)\right|\right|_2}. \tag{3.5}$$

This fixed-point iteration can be expressed more compactly in terms of $\nabla\widehat{f}_h(\boldsymbol{x}) = -\frac{c_{L,q}(h)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i \cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ as:

$$\boldsymbol{x}^{(t+1)} = F(\boldsymbol{x}^{(t)}) = \frac{\nabla\widehat{f}_h(\boldsymbol{x}^{(t)})}{\left|\left|\nabla\widehat{f}_h(\boldsymbol{x}^{(t)})\right|\right|_2}, \tag{3.6}$$

where ∇ denotes the total gradient operator in the ambient Euclidean space $\mathbb{R}^{q+1}$.

3.3 The EM Perspective of Directional Mean Shift Algorithm

We now establish a connection between the DMS algorithm and the Expectation-Maximization (EM) framework (Dempster et al., 1977; Wu, 1983; McLachlan and Krishnan, 2008; Balakrishnan et al., 2017).

3.3.1 Directional Mean Shift Algorithm as a Generalized EM Algorithm

Given the observed directional data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$, we introduce the following mixture density on Ω_q as:

$$f_{L,q}(\boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - \boldsymbol{y}^T\boldsymbol{\nu}_i)\right], \tag{3.7}$$

where $\{\alpha_i, \boldsymbol{\nu}_i, \kappa_i\}_{i=1}^n$ are the parameters of this mixture model such that $\sum_{i=1}^{n} \alpha_i = 1$, $\alpha_i \geq 0$ for $i = 1, ..., n$, and $C_{q,L}(\kappa_i)$ is the normalizing constant such that

$$C_{q,L}(\kappa_i)^{-1} = \int_{\Omega_q} L\left[\kappa_i\left(1 - \boldsymbol{y}^T\boldsymbol{\nu}_i\right)\right] \omega_q(d\boldsymbol{y}) = w_{q-1}(\Omega_{q-1}) \int_{-1}^{1} L\left[\kappa_i(1-t)\right] \cdot (1-t^2)^{\frac{q}{2}-1} dt \tag{3.8}$$

with $\omega_q(\Omega_q)$ as the surface area of Ω_q for $q > 1$. Notably, the normalizing constant $C_{q,L}(\kappa_i)$ does not depend on the directional mean parameter $\boldsymbol{\nu}_i$. The parameters α_i and κ_i represent the weights (*i.e.*, mixing proportions) and concentration parameters around $\boldsymbol{\nu}_i$, respectively. Later, for the directional KDE, we set $\alpha_i = \frac{1}{n}$ and $\kappa_i = \frac{1}{h^2}$. Throughout this section, $\{\alpha_i, \kappa_i\}_{i=1}^n$ are treated as fixed.

Crucially, we define the directional mean parameter $\boldsymbol{\nu}_i := \boldsymbol{\nu}_i(\boldsymbol{\mu})$ through an unknown parameter $\boldsymbol{\mu} \in \Omega_q$ and a given observation $\boldsymbol{X}_i$ in the directional data as:

$$\boldsymbol{\nu}_i := \boldsymbol{\nu}_i(\boldsymbol{\mu}) = \left(\underbrace{\sqrt{\frac{1 - (\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{q}}, ..., \sqrt{\frac{1 - (\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{q}}}_{q \text{ times}}, \boldsymbol{\mu}^T\boldsymbol{X}_i \right)^T \in \Omega_q. \tag{3.9}$$

Given $\boldsymbol{\mu} \in \Omega_q$, the distribution in (3.7) is a mixture of n densities on Ω_q with directional means $\{\boldsymbol{\nu}_i(\boldsymbol{\mu})\}_{i=1}^n$, *fixed* mixing proportions $\{\alpha_i\}_{i=1}^n$ and *fixed* concentration parameters $\{\kappa_i\}_{i=1}^n$. Thus, varying the parameter $\boldsymbol{\mu}$ induces a spherical transformation of the mixture distribution on Ω_q; see Remark 3.1 for more details. Our objective is to obtain the maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ via an EM procedure. To establish the connection with the mean shift algorithm, we consider a hypothetical setting in which only a single observation $\boldsymbol{Y}_1$ from the mixture model (3.7) is observed. We now derive the corresponding EM updates for estimating $\boldsymbol{\mu}$ in this model and show that they coincide with the DMS iteration.

3.3.2 The EM Framework and its Connection to Directional Mean Shift

To derive the EM algorithm, we introduce a latent indicator variable $Z_1 \in \{1, ..., n\}$, where $Z_1 = i$ when $\boldsymbol{Y}_1$ is generated by the i-th component $C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - \boldsymbol{y}^T\boldsymbol{\nu}_i)\right]$ in (3.7).

- **E-Step**: The complete log-likelihood for $(\boldsymbol{Y}_1, Z_1)$ is given by

$$\log P(\boldsymbol{Y}_1, Z_1|\boldsymbol{\mu}) = \sum_{i=1}^{n} \mathbb{1}_{\{Z_1=i\}} \cdot \left\{\log \alpha_i + \log C_{q,L}(\kappa_i) + \log L\left[\kappa_i(1 - \boldsymbol{Y}_1^T\boldsymbol{\nu}_i)\right]\right\}. \tag{3.10}$$

Given $(\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)})$ at the t-th iteration, taking the expectation with respect to the posterior distribution of $Z_1|\left(\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}\right)$ yields the Q-function as:

$$\begin{aligned} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^{(t)}) &= \mathbb{E}_{Z_1|(\boldsymbol{Y}_1,\boldsymbol{\mu}^{(t)})}\left[\log P(\boldsymbol{Y}_1, Z_1|\boldsymbol{\mu})\right] \\ &= \sum_{i=1}^{n} \left[\log \alpha_i + \log C_{q,L}(\kappa_i)\right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) \\ &\quad + \sum_{i=1}^{n} \log L\left[\kappa_i(1 - \boldsymbol{Y}_1^T\boldsymbol{\nu}_i)\right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}). \end{aligned} \tag{3.11}$$

By Bayes' theorem, the posterior distribution $P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)})$ for $i = 1, ..., n$ is computed as:

$$\begin{aligned} P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) &= \frac{P(\boldsymbol{Y}_1|Z_1 = i, \boldsymbol{\mu}^{(t)}) \cdot P(Z_1 = i|\boldsymbol{\mu}^{(t)})}{P(\boldsymbol{Y}_1|\boldsymbol{\mu}^{(t)})} \\ &= \frac{\alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i\left(1 - \boldsymbol{Y}_1^T\boldsymbol{\nu}_i^{(t)}\right)\right]}{\sum_{\ell=1}^{n} \alpha_\ell \cdot C_{q,L}(\kappa_\ell) \cdot L\left[\kappa_\ell\left(1 - \boldsymbol{Y}_1^T\boldsymbol{\nu}_\ell^{(t)}\right)\right]}. \end{aligned} \tag{3.12}$$

- **M-Step**: To derive the M-step, we consider the specific case where $\boldsymbol{Y}_1 = \boldsymbol{e}_{q+1} = (0, ..., 0, 1)^T \in \Omega_q$. This choice ensures that $\boldsymbol{\nu}_i^T\boldsymbol{Y}_1 = \boldsymbol{\mu}^T\boldsymbol{X}_i$ for $i = 1, ..., n$, and the Q-function (3.11) reduces to

$$\begin{aligned} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^{(t)}) &= \sum_{i=1}^{n} \left[\log \alpha_i + \log C_{q,L}(\kappa_i)\right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) \\ &\quad + \sum_{i=1}^{n} \log L\left[\kappa_i(1 - \boldsymbol{\mu}^T\boldsymbol{X}_i)\right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}). \end{aligned} \tag{3.13}$$

Since $C_{q,L}(\kappa_i)$ does not depend on $\boldsymbol{\mu}$, maximizing (3.13) reduces to maximizing the second term in (3.13) subject to the constraint $\boldsymbol{\mu}^T\boldsymbol{\mu} = 1$. By introducing the Lagrangian function $\mathcal{L}(\boldsymbol{\mu}, \lambda) = \sum_{i=1}^n \log L\left[\kappa_i(1 - \boldsymbol{\mu}^T\boldsymbol{X}_i)\right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) + \lambda\left(1 - \boldsymbol{\mu}^T\boldsymbol{\mu}\right)$ and solving the first-order conditions, we obtain that

$$\begin{aligned}\widehat{\boldsymbol{\mu}} &= \frac{\sum_{i=1}^n P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) \cdot \frac{-\kappa_i \boldsymbol{X}_i \cdot L'\left[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)\right]}{L[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)]}}{\left|\left|\sum_{i=1}^n P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}^{(t)}) \cdot \frac{-\kappa_i \boldsymbol{X}_i \cdot L'[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)]}{L[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)]}\right|\right|_2} \\ &= \frac{-\sum_{i=1}^n \kappa_i\alpha_i \cdot C_{q,L}(\kappa_i) \cdot \boldsymbol{X}_i \cdot L'\left[\kappa_i\left(1 - \widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i\right)\right] \cdot \frac{L\left[\kappa_i\left(1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}\right)\right]}{L[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)]}}{\left|\left|\sum_{i=1}^n \kappa_i\alpha_i \cdot C_{q,L}(\kappa_i) \cdot \boldsymbol{X}_i \cdot L'\left[\kappa_i\left(1 - \widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i\right)\right] \cdot \frac{L\left[\kappa_i\left(1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}\right)\right]}{L[\kappa_i(1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i)]}\right|\right|_2},\end{aligned} \tag{3.14}$$

where we plug in the posterior distribution (3.12) in the second equality. Therefore, (3.14) defines a fixed-point iteration for obtaining $\boldsymbol{\mu}^{(t+1)}$ in the EM framework, and an exact M-step would iterate (3.14) until convergence.

Now, we take $\kappa_i = \frac{1}{h^2}$ and $\alpha_i = \frac{1}{n}$ for all $i = 1, ..., n$ in the fixed-point iteration equation (3.14). Then, the normalizing constant $C_{q,L}(\kappa_i)$ will no longer depend on the summation index i, and the fixed-point iteration becomes $\widehat{\boldsymbol{\mu}} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i}{h^2}\right) \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{L\left(\frac{1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i}{h^2}\right)}}{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i}{h^2}\right) \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{L\left(\frac{1-\widehat{\boldsymbol{\mu}}^T\boldsymbol{X}_i}{h^2}\right)}\right|\right|_2}$, or more specifically,

$$\widehat{\boldsymbol{\mu}}^{(k+1;t)} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{\mu}}^{(k;t)}}{h^2}\right) \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{\mu}}^{(k;t)}}{h^2}\right)}}{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{\mu}}^{(k;t)}}{h^2}\right) \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{\mu}}^{(k;t)}}{h^2}\right)}\right|\right|_2} \tag{3.15}$$

for $k = 0, 1, 2, ...$ Therefore, the complete EM algorithm consists of two nested loops: the outer one is the usual iteration over t, while the inner one iterates $\{\widehat{\boldsymbol{\mu}}^{(k;t)}\}_{k=0}^{\infty}$ under the fixed $\boldsymbol{\mu}^{(t)}$ for an exact M-step. If we choose the initial value for iterating (3.15) as $\widehat{\boldsymbol{\mu}}^{(0;t)} = \boldsymbol{\mu}^{(t)}$

and do a single iteration of (3.15), *i.e.*, $\boldsymbol{\mu}^{(t+1)} = \widehat{\boldsymbol{\mu}}^{(1;t)}$, we obtain that

$$\boldsymbol{\mu}^{(t+1)} = -\frac{\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{\left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)\right|\right|_2}. \tag{3.16}$$

This coincides exactly with the DMS updating formula (3.5). Thus, the DMS algorithm corresponds to an EM procedure in which the M-step is replaced by a single ascent step. In particular, it is a generalized EM algorithm. We will demonstrate that under mild regularity conditions on the directional kernel L, the Q-function $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^{(t)})$ is non-decreasing along this iteration, thereby validating the generalized EM interpretation.

For the von Mises kernel $L(r) = e^{-r}$, the derivative satisfies $L'(r) = -L(r)$. Consequently, the ratio $\frac{L'(r)}{L(r)}$ in (3.14) is constant, and the fixed-point equation resolves in a single step. With $\kappa_i = \frac{1}{h^2}$ and $\alpha_i = \frac{1}{n}$ for all $i = 1, ..., n$ as before, the resulting M-step of (3.14) is simplified as:

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{X}_i \exp\left(\frac{\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)}{\left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i \exp\left(\frac{\boldsymbol{X}_i^T\boldsymbol{\mu}^{(t)}}{h^2}\right)\right|\right|_2}. \tag{3.17}$$

This is precisely the DMS algorithm with the von Mises kernel. Therefore, the DMS algorithm with the von Mises kernel is an exact EM algorithm, rather than merely a generalized EM procedure.

Remark 3.1 (Group action interpretation)**.** *The above construction provides a concrete realization of the conjecture stated in Section 7.1 of Subbarao and Meer (2009), who suggested that nonlinear mean shift algorithms on Riemannian manifolds may admit an EM interpretation through a properly defined group action. Our derivation above shows that constructing such a group action is non-trivial even in the case of the sphere Ω_q. Specifically, we denote the original mixture density defined on the observed data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$ by $\bar{f}_{L,q}(\boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - \boldsymbol{y}^T\boldsymbol{X}_i)\right]$. The transformed mixture density $f_{L,q}$ in (3.7) can be expressed as a group action $R \in \mathrm{OG}(q+1)$ on $\bar{f}_{L,q}$ as:*

$$f_{L,q}(\boldsymbol{y}) = \bar{f}_{L,q}(R\boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - (R\boldsymbol{y})^T\boldsymbol{X}_i)\right] = \sum_{i=1}^{n} \alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - \boldsymbol{y}^T\boldsymbol{\nu}_i)\right],$$

where $\mathrm{OG}(q+1) = \left\{R \in \mathbb{R}^{(q+1)\times(q+1)} : RR^T = R^T R = I_{q+1}\right\}$ *is the orthogonal group of dimension* $(q+1)$*, where* $I_{q+1} \in \mathbb{R}^{(q+1)\times(q+1)}$ *is the identity matrix. Under the parameterization of* $\boldsymbol{\nu}_i$ *in* (3.9)*, the orthogonal transformation* R *is uniquely determined by the target parameter* $\boldsymbol{\mu}$ *as* $R = I_{q+1} - \frac{2(\boldsymbol{e}_{q+1}-\boldsymbol{\mu})(\boldsymbol{e}_{q+1}-\boldsymbol{\mu})^T}{||\boldsymbol{e}_{q+1}-\boldsymbol{\mu}||_2^2}$*, where* $\boldsymbol{e}_{q+1} = (0, ..., 0, 1)^T$ *is a standard basis vector in* $\mathbb{R}^{q+1}$*.*

3.3.3 Convergence of Directional Mean Shift as a Generalized EM Algorithm

In this subsection, we establish the convergence properties of the DMS algorithm by exploiting its interpretation as a generalized EM procedure. We first prove that the Q-function (3.13) is non-decreasing along the DMS trajectory (3.16) under mild regularity conditions on the kernel L. Since the directional KDE coincides with the observed likelihood under the EM formulation, it offers a new proof of the classical ascending property of density estimates along DMS sequences. Our argument differs fundamentally from existing approaches (*e.g.*, Kobayashi and Otsu 2010; Kafai et al. 2010), as it relies on the generalized EM structure rather than gradient-based arguments. Moreover, the EM perspective implies a global convergence property, where the DMS sequence converges to a stationary point of the directional KDE $\widehat{f}_h$ in (2.4) from almost any starting point on Ω_q. While the convergence to saddle points or local minima is theoretically possible, as is well-known for EM-type algorithms (see Section 3.6.1 and 3.6.2 of McClean et al. 2024), such points are unstable for maximization (Carreira-Perpiñán, 2007). In practice, the convergence is thus effectively to local modes of $\widehat{f}_h$.

The characterization of the DMS iteration as a generalized EM algorithm hinges on the fact that a single iteration of the inexact M-step (3.16) is sufficient to increase the Q-function (3.13).

Theorem 3.2. *Suppose that the directional kernel* $L : [0, \infty) \to [0, \infty)$ *is non-increasing, differentiable, and convex with* $0 < L(0) < \infty$*. Then, the Q-function* (3.13) *satisfies*

$$Q(\underline{\widehat{\boldsymbol{x}}}^{(t+1)} \mid \underline{\widehat{\boldsymbol{x}}}^{(t)}) \geq Q(\underline{\widehat{\boldsymbol{x}}}^{(t)} \mid \underline{\widehat{\boldsymbol{x}}}^{(t)}),$$

where

$$\widehat{\underline{\boldsymbol{x}}}^{(t+1)} = -\frac{\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T \widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T \widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right|\right|_2}$$

is from the iteration (3.16).

The proof of Theorem 3.2 relies on two elementary inequalities $\log y - \log x \geq \frac{y-x}{x}$ and $L(y) - L(x) \geq L'(x) \cdot (y - x)$, which follow respectively from the concavity of $\log(\cdot)$ as well as the convexity and differentiability of the kernel $L(\cdot)$; see Section B.2. The convexity and differentiability assumptions on kernel L are sufficient but not necessary. Alternative sufficient conditions for the iteration (3.16) to increase the Q-function can be formulated (*e.g.*, when $\log L(r)$ is convex).

Since the directional KDE coincides with the observed likelihood under the EM formulation (see Proposition B.1), the monotonicity of the Q-function (Theorem 1 in Dempster et al. 1977) implies the non-decreasing property of the density values along DMS sequences. The result is summarized as the theorem below, whose proof is in Section B.2.1.

Theorem 3.3. *Let* $\left\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\right\}_{t=0}^{\infty}$ *be the path of successive points defined by the DMS algorithm* (3.5). *Suppose that the directional kernel* $L : [0, \infty) \to [0, \infty)$ *is non-increasing, differentiable, and convex with* $0 < L(0) < \infty$. *Then, the sequence* $\left\{\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$ *is non-decreasing and thus converges.*

Beyond the convergence of density values, the EM framework enables us to establish the global convergence of the DMS sequence itself.

Theorem 3.4. *Let* $\left\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\right\}_{t=0}^{\infty}$ *be the path of successive points defined by the DMS algorithm* (3.5). *Suppose that the following conditions hold:*

(a) The directional KDE $\widehat{f}_h$ *has finitely many local modes on* Ω_q, *and these local modes are isolated from other critical points.*

(b) The directional kernel $L : [0, \infty) \to [0, \infty)$ *is non-increasing, differentiable, and convex with* $0 < L(0) < \infty$.

Then, the sequence $\left\{\underline{\widehat{\boldsymbol{x}}}^{(t)}\right\}_{t=0}^{\infty}$ *converges to a local mode of* $\widehat{f}_h$ *from almost all initializations* $\underline{\widehat{\boldsymbol{x}}}^{(0)} \in \Omega_q$.

The proof builds on the classical convergence theory for generalized EM algorithms (Wu, 1983). Convergence to a stationary point is guaranteed by compactness of Ω_q, differentiability of the observed log-likelihood, and discreteness of the set of local modes. We verify in Section B.3 that the two conditions in Theorem 3.4 imply these requirements. Notably, the set of unstable stationary points (*e.g.*, saddle points and local minima) of $\widehat{f}_h$ has zero Lebesgue measure on Ω_q. Thus, for nearly all starting points, the DMS algorithm identifies a local mode in practice. This global convergence result establishes a basin of attraction for each local mode.

3.4 Linear Convergence of Gradient Ascent Algorithms on Ω_q

We now discuss the linear convergence of gradient ascent algorithms on Ω_q, of which the DMS algorithm is a special case with an adaptive step size. Adopting the notation of Zhang and Sra (2016), a gradient ascent algorithm applied to an objective function f on Ω_q (a Riemannian manifold) is written as:

$$\boldsymbol{y}^{(t+1)} = \texttt{Exp}_{\boldsymbol{y}^{(t)}}\left(\eta \cdot \texttt{grad}\, f(\boldsymbol{y}^{(t)})\right). \tag{3.18}$$

Given a sequence $\left\{\boldsymbol{y}^{(t)}\right\}_{t=0}^{\infty}$ converging to $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$, the convergence is said to be linear if there exists a positive constant $\Upsilon < 1$ (rate of convergence) such that $\left|\left|\boldsymbol{y}^{(t+1)} - \underline{\boldsymbol{m}}_k\right|\right| \leq \Upsilon \left|\left|\boldsymbol{y}^{(t)} - \underline{\boldsymbol{m}}_k\right|\right|$ when t is sufficiently large (Boyd and Vandenberghe, 2004). In our context, the norm $||\cdot||$ refers to the geodesic (or great-circle) distance $d(\boldsymbol{x}, \boldsymbol{y}) = ||\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})||_2$ between two points $\boldsymbol{x}, \boldsymbol{y} \in \Omega_q$. An equivalent statement of linear convergence is that the algorithm takes $O(\log(1/\epsilon))$ iterations to converge to an ϵ-error of $\underline{\boldsymbol{m}}_k$.

Using the notation in Zhang and Sra (2016), we let $\zeta(1, c) \equiv \frac{c}{\tanh(c)}$. One can show by differentiating $\zeta(1, c)$ that $\zeta(1, c)$ is strictly increasing and $\zeta(1, c) > 1$ for any $c > 0$.

Theorem 3.5. *Assume that Assumptions D1 and M1 hold.*

(a) ***Linear convergence of gradient ascent with*** f*:* *Given a convergence radius* r_0 *with* $0 < r_0 \leq \sqrt{2 - 2\cos\left[\frac{3\lambda_*}{2(q+1)^{\frac{3}{2}}||f||^*_{\infty,3}}\right]}$*, the iterative sequence* $\{\boldsymbol{y}^{(t)}\}_{t=0}^{\infty}$ *defined by the population-level gradient ascent algorithm* (3.18) *satisfies*

$$d(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k) \leq \Upsilon^t \cdot d(\boldsymbol{y}^{(0)}, \underline{\boldsymbol{m}}_k) \quad \text{with} \quad \Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}},$$

whenever $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)||f||^*_{\infty,3}\zeta(1,r_0)}\right\}$ *and the initial point* $\boldsymbol{y}^{(0)} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ *for some* $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$*. We recall that* $||f||^*_{\infty,3}$ *is the uniform functional norm of the derivatives of the directional density* f *up to the third order,* $\lambda_* > 0$ *is defined in Assumption M1, and* $\underline{\mathcal{M}}$ *is the set of local modes of the directional density* f*.*

We further assume Assumption D2 in the sequel.

(b) ***Linear convergence of gradient ascent with*** $\widehat{f}_h$*:* *Let the sample-based gradient ascent update on* Ω_q *be* $\widehat{\boldsymbol{y}}^{(t+1)} = \texttt{Exp}_{\widehat{\boldsymbol{y}}^{(t)}}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}^{(t)})\right)$*. With the same choice of the convergence radius* $r_0 > 0$ *and* $\Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}}$ *as in (a), if* $h \to 0$ *and* $\frac{nh^{q+2}}{|\log h|} \to \infty$*, then for any* $\delta \in (0, 1)$*,*

$$d\left(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k\right) \leq \Upsilon^t \cdot d\left(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k\right) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$$

with probability at least $1 - \delta$*, whenever* $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)||f||^*_{\infty,3} \cdot \zeta(1,r_0)}\right\}$ *and the initial point* $\widehat{\boldsymbol{y}}^{(0)} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ *for some* $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$*.*

The proof of Theorem 3.5 is in Appendix B.3.1. As shown in part (a) of Theorem 3.5, the linear convergence radius of the gradient ascent algorithm (3.18) on Ω_q generally depends on the lower bound λ_* on absolute eigenvalues of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ within the tangent space $T_{\boldsymbol{x}}$, the uniform functional norm $||f||^*_{\infty,3}$ of the (partial) derivatives of f up to the third order, and the manifold dimension q.

In practice, we are more interested in the algorithmic convergence rate of sample-based gradient ascent algorithms with directional KDEs to the estimated local modes $\widehat{\underline{\mathcal{M}}}_n$. As

indicated by (2.29), the Hessian matrices of $\widehat{f}_h$ at its local modes have only negative eigenvalues within the corresponding tangent spaces given Assumption M1, sufficiently small h, and sufficiently large $\frac{nh^{q+4}}{|\log h|}$. In reality, unless the data configuration is highly abnormal, the local modes of directional KDEs are non-degenerate and $\widehat{f}_h$ is geodesically strongly concave (see Section B.3.1 for a precise definition) around small neighborhoods of the estimated local modes. Together with an application of smooth kernels, say the von Mises kernel, $\widehat{f}_h$ is β-smooth on Ω_q and, consequently, a sample-based gradient ascent algorithm with the directional KDE $\widehat{f}_h$ converges linearly to the estimated local modes around their small neighborhoods, given a proper step size.

Let $\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\}_{t=0}^{\infty}$ denote the sequence defined by the above DMS procedure. Later, by abuse of notation, we will use the same notation to denote the directional SCGA/SCMS sequence with $\widehat{f}_h$. As $\nabla\widehat{f}_h(\boldsymbol{x}) = -\frac{c_{L,q}(h)}{nh^2}\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$, we have already shown in Equation 3.6 that the DMS algorithm can be written as the following fixed-point iteration:

$$\widehat{\underline{\boldsymbol{x}}}^{(t+1)} = -\frac{\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\left|\left|\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right|\right|_2} \quad \text{or} \quad \widehat{\underline{\boldsymbol{x}}}^{(t+1)} = \frac{\nabla\widehat{f}_h\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)}{\left|\left|\nabla\widehat{f}_h\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)\right|\right|_2}.$$

We then write the DMS algorithm as a gradient ascent method on Ω_q with the iteration formula (Zhang and Sra, 2016):

$$\widehat{\underline{\boldsymbol{x}}}^{(t+1)} = \mathtt{Exp}_{\widehat{\underline{\boldsymbol{x}}}^{(t)}}\left(\underline{\eta}_{n,h}^{(t)}\cdot\mathtt{grad}\,\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right), \tag{3.19}$$

where the adaptive step size $\underline{\eta}_{n,h}^{(t)}$ is given by

$$\underline{\eta}_{n,h}^{(t)} = \arccos\left(\frac{\nabla\widehat{f}_h\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{\left|\left|\nabla\widehat{f}_h\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)\right|\right|_2}\right)\cdot\frac{1}{\left|\left|\mathtt{grad}\,\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2} = \frac{\theta_t}{\left|\left|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2\cdot\sin\theta_t}. \tag{3.20}$$

Here, we denote the angle between $\widehat{\underline{\boldsymbol{x}}}^{(t+1)}$ and $\widehat{\underline{\boldsymbol{x}}}^{(t)}$ (or equivalently, the angle between $\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})$ and $\widehat{\underline{\boldsymbol{x}}}^{(t)}$) by θ_t; see Section 5.2 in Zhang and Chen (2021b) for detailed derivations. Within some small neighborhoods around local modes of $\widehat{f}_h$, $\frac{\theta_t}{\sin\theta_t}\approx 1$ and the adaptive step size $\underline{\eta}_{n,h}^{(t)}$ is dominated by $\left|\left|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2$. The following lemma characterizes the asymptotic behavior of $\left|\left|\nabla\widehat{f}_h(\boldsymbol{x})\right|\right|_2$ on Ω_q and, consequently, of $\underline{\eta}_{n,h}^{(t)}$.

Lemma 3.6 (Lemma 10 in Zhang and Chen 2021b)**.** *Assume Assumptions D1 and 2.1. For any fixed* $\boldsymbol{x} \in \Omega_q$*, we have*

$$h^2 \cdot \left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 = f(\boldsymbol{x}) \cdot C_{L,q} + o(1) + O_P\left(\sqrt{\frac{1}{nh^q}}\right)$$

as $nh^q \to \infty$ *and* $h \to 0$*, where* $C_{L,q} = -\frac{\int_0^\infty L'(r) r^{\frac{q}{2}-1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} > 0$ *is a constant depending only on the kernel* L *and the dimension* q*.*

Lemma 3.6 indicates that $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 \to \infty$ with probability tending to 1 as $h \to 0$ and $nh^q \to \infty$ for any $\boldsymbol{x} \in \Omega_q$. The conclusion may seem counterintuitive at first glance, but one should be aware that the consistency of $\nabla \widehat{f}_h(\boldsymbol{x})$ holds only on its tangent component; see (2.29). The radial component of $\nabla \widehat{f}_h(\boldsymbol{x})$ that is perpendicular to Ω_q diverges, despite the fact that the true directional density f does not have any radial component. Using Lemma 3.6, one can argue that the adaptive step size $\underline{\eta}_{n,h}^{(t)}$ in (3.20) of the DMS algorithm as a gradient ascent method on Ω_q tends to zero at the rate $O(h^2)$ as $h \to 0$ and $nh^q \to \infty$. Hence, when n is sufficiently large and h is small but fixed, $\underline{\eta}_{n,h}^{(t)}$ satisfies the upper bound condition in Theorem 3.5 and therefore, the DMS algorithm converges linearly.

3.5 Linear Convergence of Subspace Constrained Mean Shift Algorithm

Building upon the mean shift algorithm in (3.3), the SCMS algorithm updates the sequence $\{\widehat{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ through the subspace constrained mean shift vector $\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \Xi_h(\widehat{\boldsymbol{x}}^{(t)})$ as:

$$\begin{aligned}\widehat{\boldsymbol{x}}^{(t+1)} &\leftarrow \widehat{\boldsymbol{x}}^{(t)} + \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \Xi_h(\widehat{\boldsymbol{x}}^{(t)}) \\ &= \widehat{\boldsymbol{x}}^{(t)} + \frac{1}{\frac{2c_{k,D}}{nh^{D+2}} \sum_{i=1}^{n} -k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i}{h}\right|\right|^2\right)} \cdot \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \nabla \widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)}). \end{aligned} \tag{3.21}$$

See Algorithm 1 in Section B.1 for the complete procedure. This also implies that the SCMS algorithm can be viewed as a sample-based SCGA method as:

$$\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \widehat{\boldsymbol{x}}^{(t)} + \eta_{n,h}^{(t)} \cdot \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \nabla \widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)}) \tag{3.22}$$

with the same adaptive step size $\eta_{n,h}^{(t)}$ as the Euclidean mean shift algorithm in (3.3). The formulation (3.22) sheds light on some (linear) convergence properties of the SCMS algorithm as we will demonstrate below.

To establish (linear) convergence results for the SCMS algorithm with Euclidean KDE $\widehat{p}_n$, it suffices to study the (linear) convergence of the sample-based SCGA algorithm with objective function $\widehat{p}_n$. To this end, we begin by studying the convergence of the population SCGA algorithm whose objective function is the underlying density p.

Let $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ be the sequence defined by the population SCGA algorithm and $\{\widehat{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ be the sequence defined by the sample-based SCGA algorithm. The population SCGA algorithm is defined by the iterative formula

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} + \eta \cdot V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)}), \tag{3.23}$$

where $\eta > 0$ is a (fixed) step size. The sample-based SCGA algorithm has the same iterative formula as (3.22), except that the standard sample-based SCGA algorithm normally uses a constant step size $\eta > 0$.

Remark 3.2. *In (3.21) and (3.22), we consider the SCMS algorithm as a sample-based SCGA iteration with an adaptive step size $\eta_{n,h}^{(t)}$. Our Lemma 3.1 suggests that $\eta_{n,h}^{(t)}$ tends to zero at a rate $O(h^2)$ as $nh^D \to \infty$ and $h \to 0$. However, once the sample size n is fixed and the bandwidth h is chosen, the step size $\eta_{n,h}^{(t)}$ is not only upper bounded but also uniformly lower bounded away from zero with respect to the iteration number t by Assumption E1 when the current iterative point $\widehat{\boldsymbol{x}}^{(t)}$ lies within the compact neighborhood $R_d \oplus \rho$. Note that $R_d \oplus \rho$ is compact because R_d is a finite union of connected and compact manifolds; see (d) of Lemma A.1. More importantly, these upper and lower bounds of $\eta_{n,h}^{(t)}$ when $\widehat{\boldsymbol{x}}^{(t)} \in R_d \oplus \rho$ are independent of the iteration number t. Therefore, conditioning on the case when the sample size n is sufficiently large, one can always select a small bandwidth h such that the adaptive step size $\eta_{n,h}^{(t)}$ of the SCMS algorithm is sufficiently small but not equal to zero.*

As revealed by the following proposition, Assumptions 2.1–2.3 ensure that as long as the step size $\eta > 0$ is small, the objective function p along any population SCGA sequence

$\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ is non-decreasing and the sequence itself converges to R_d when it is initialized within a small neighborhood of R_d.

Proposition 3.7 (Convergence of the SCGA Algorithm). *For any SCGA sequence* $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ *defined by* (3.23) *with* $0 < \eta < \frac{2}{D||p||_{\infty}^{(2)}}$, *the following properties hold.*

(a) Under Assumption 2.1, the objective function sequence $\{p(\boldsymbol{x}^{(t)})\}_{t=0}^{\infty}$ *is non-decreasing and converges.*

(b) Under Assumption 2.1, $\lim_{t\to\infty} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 = \lim_{t\to\infty} \left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right|\right|_2 = 0.$

(c) Under Assumptions 2.1–2.3, $\lim_{t\to\infty} d_E(\boldsymbol{x}^{(t)}, R_d) = 0$ *whenever* $\boldsymbol{x}^{(0)} \in R_d \oplus r_1$ *with the convergence radius* r_1 *satisfying*

$$0 < r_1 < \min\left\{\frac{\rho}{2},\ \frac{\beta_1^2}{A_2\left(||p||_{\infty}^{(3)} + ||p||_{\infty}^{(4)}\right)},\ \frac{\beta_1}{B_4(p)}\right\},$$

where $A_2 > 0$ *is a constant defined in (h) of Lemma A.1 while* $B_4(p) > 0$ *is a quantity depending on both the dimension* D *and functional norm* $||p||_{\infty,4}^{*}$ *up to the fourth-order (partial) derivatives of* p.

The proof of Proposition 3.7 can be found in Appendix B.4. We make two comments on the choice of the convergence radius r_1 in (c) of Proposition 3.7. The first two quantities in the upper bound of r_1 ensure that $r_1 \leq \texttt{reach}(R_d)$ and therefore, the projection of $\boldsymbol{x}^{(t)} \in R_d \oplus r_1$ onto R_d is well-defined. The last quantity in the upper bound of r_1 is critical to guarantee that the distances $\{d_E(\boldsymbol{x}^{(t)}, R_d)\}_{t=0}^{\infty}$ from the SCGA sequence $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ to the ridge R_d can be controlled by the norms $\left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2$ of order-d principal gradients for $t = 0, 1, ...$

Corollary 3.8 (Convergence of the SCMS Algorithm). *When the fixed sample size* n *is sufficiently large and the fixed bandwidth* h *is chosen to be sufficiently small, the following properties hold for the SCMS sequence* $\{\widehat{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ *with high probability under Assumptions 2.1–2.3 and Assumptions E1 and E2.*

(a) The Euclidean KDE sequence $\{\widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)})\}$ *is non-decreasing and thus converges.*

(b) $\lim_{t\to\infty} \left|\left|\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \nabla \widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)})\right|\right|_2 = \lim_{t\to\infty} \left|\left|\widehat{\boldsymbol{x}}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t)}\right|\right|_2 = 0.$

(c) $\lim_{t\to\infty} d_E(\widehat{\boldsymbol{x}}^{(t)}, \widehat{R}_d) = 0$ *whenever* $\widehat{\boldsymbol{x}}^{(0)} \in \widehat{R}_d \oplus r_1$ *with the convergence radius* $r_1 > 0$ *defined in (c) of Proposition 3.7.*

Corollary 3.8 is the sample-based version of Proposition 3.7. On the one hand, when $\frac{nh^{D+6}}{|\log h|}$ is sufficiently large and h is small enough, the estimated ridge $\widehat{R}_d$ also satisfies Assumptions 2.1–2.3 with high probability; see Lemma 2.1 and the uniform bounds (2.27) of $\widehat{p}_n$. On the other hand, the adaptive step size $\eta_{n,h}^{(t)}$ of the SCMS algorithm can always be smaller than the threshold $\frac{2}{D||p||_\infty^{(2)}}$ when the sample size n is sufficiently large and h is small; see Remark 3.2. Consequently, our arguments in Proposition 3.7 can be applied to establish the (local) convergence of the SCMS sequence here. In addition, we point out that Proposition 2 in Ghassabeh (2013) also proved results (a-b) of Corollary 3.8 under Assumption E1 and the convexity assumption on the kernel profile k. The difference is that our arguments hold when n is large and h is small while the extra convexity assumption in Ghassabeh (2013) enables the authors to prove results (a-b) universally for any choice of the bandwidth h.

By Proposition 3.7 and Corollary 3.8, it is now reasonable to denote the limiting points of the population and sample-based SCGA sequences $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ and $\{\widehat{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ by $\boldsymbol{x}^* \in R_d$ and $\widehat{\boldsymbol{x}}^* \in \widehat{R}_d$, respectively. Before stating our main linear convergence results, we introduce the concepts of Q-linear and R-linear convergence from the optimization literature; see, *e.g.*, Appendix A2 in Nocedal and Wright (2006).

Definition 3.9 (Linear Rate of Convergence)**.** *We say that the convergence of the sequence* $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ *to* $\boldsymbol{x}^*$ *is* Q-linear *if there exists a constant* $\Upsilon \in (0,1)$ *such that*

$$\frac{\left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right|\right|_2}{\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2} \leq \Upsilon \quad \textit{for all } t \textit{ sufficiently large.}$$

We say that the convergence is R-linear *if there is a sequence of nonnegative scalars* $\{\epsilon_t\}_{t=0}^{\infty}$

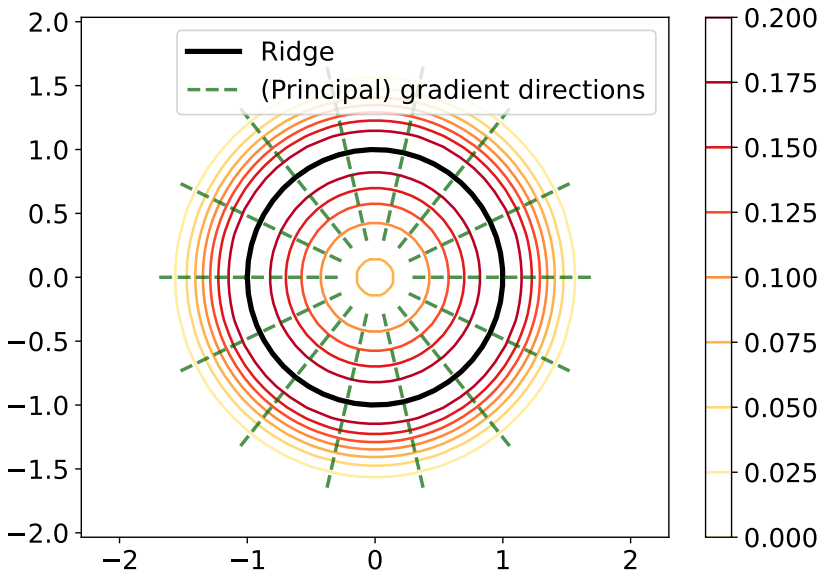


Figure 3.2: Contour lines of the density function (3.24) and its principal gradient flows.

such that

$$\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq \epsilon_t \text{ for all } t, \text{ and } \{\epsilon_t\}_{t=0}^{\infty} \text{ converges Q-linearly to zero.}$$

The linear convergence of the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$ will be established under the following local assumption.

Assumption A4 (Quadratic Behaviors of Residual Vectors)**.** *We assume that the SCGA sequence* $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$ *with step size* $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_{\infty}^{(2)}}\right\}$ *and* $\boldsymbol{x}^* \in R_d$ *as its limiting point satisfies*

$$\nabla p(\boldsymbol{x}^{(t)})^T U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \leq \frac{\beta_0}{4}\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2,$$
$$\left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})\right|\right|_2 \leq \beta_2\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2$$

for some constant $\beta_2 > 0$*, where* $\beta_0 > 0$ *is the constant defined in Assumption 2.2.*

Assumption A4 imposes a direct assumption on the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$, under which the residual vector $U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})$ and its inner product with the residual gradi-

ent $U^{\perp}(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})$ are upper bounded by a quadratic term $O\left(\left|\left|\boldsymbol{x}^{*}-\boldsymbol{x}^{(t)}\right|\right|_2^2\right)$. This assumption is imposed to guarantee that p is "subspace constrained strongly concave" around R_d; see also Remark 3.3. Our proof of Theorem 3.10 suggests that the residual vector $U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^{*}-\boldsymbol{x}^{(t)})$ is only required to be smaller than the first-order term $o\left(\left|\left|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{*}\right|\right|_2\right)$. For simplicity, we require it to be quadratic. When Assumption A4 fails to hold, the associated SCGA sequence can only converge sublinearly to R_d. Therefore, it is an essential element in the linear convergence of the SCGA algorithm, and we discuss some potentially weaker assumptions that imply Assumption A4 in Appendix B.5. Intuitively, the SCGA path converges to R_d following the direction of principal gradient $V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})$. To further gain more insight into the correctness of Assumption A4, we consider a special density function

$$p(u,v)=C_{p,2}\cdot\exp\left[-\left(u^2+v^2-1\right)^2\right] \quad \text{with} \quad C_{p,2}=\pi\left(\frac{\sqrt{\pi}}{2}+\int_0^1 e^{-t^2}dt\right) \tag{3.24}$$

on $\mathbb{R}^2$, whose one-dimensional ridge is $R_1=\{(u,v)\in\mathbb{R}^2: u^2+v^2=1\}$ by the definition (2.19). Some careful calculations suggest that its principal gradient $G_1(u,v)$ points towards the ridge R_1 in the direction (u,v) when $\frac{2-\sqrt{2}}{2}<u^2+v^2<1$ and in the direction $(-u,-v)$ when $1<u^2+v^2<\frac{2+\sqrt{2}}{2}$; see Figure 3.2 for a graphical illustration. Furthermore, the smallest eigenvalue of $\nabla\nabla p(u,v)$ is negative whenever $u^2+v^2>\frac{1}{2}$. Hence, the residual gradient $U_1^{\perp}(u,v)\nabla p(u,v)$ is perpendicular to the SCGA direction, and Assumption A4 naturally holds.

We now present our linear convergence results for the population and sample-based SCGA algorithms.

Theorem 3.10 (Linear Convergence of the SCGA Algorithm)**.** *Assume that Assumptions 2.1–2.3 and A4 hold throughout the theorem.*

(a) ***Q-Linear convergence of*** $\left|\left|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{*}\right|\right|_2$*: Consider a convergence radius* $r_2>0$

satisfying

$$0 < r_2 < \min\left\{\frac{\rho}{2}, \frac{\beta_1^2}{A_2\left(||p||_\infty^{(3)} + ||p||_\infty^{(4)}\right)}, \frac{\beta_1}{B_4(p)}, \frac{3\beta_0}{4D\left(6\,||p||_\infty^{(2)}\,\beta_2^2\rho + D^{\frac{1}{2}}\,||p||_\infty^{(3)}\right)}\right\},$$

where $A_2 > 0$ *is the constant defined in (h) of Lemma A.1 and* $B_4(p) > 0$ *is a quantity defined in (c) of Proposition 3.7 that depends on both the dimension* D *and the functional norm* $||p||_{\infty,4}^*$ *up to the fourth-order derivative of* p*. Whenever* $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ *and the initial point* $\boldsymbol{x}^{(0)} \in Ball_D(\boldsymbol{x}^*, r_2)$ *with* $\boldsymbol{x}^* \in R_d$*, we have that*

$$\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq \Upsilon^t \left|\left|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right|\right|_2 \quad \textit{with} \quad \Upsilon = \sqrt{1 - \frac{\beta_0\eta}{4}}.$$

(b) ***R-Linear convergence of*** $d_E(\boldsymbol{x}^{(t)}, R_d)$*: Under the same radius* $r_2 > 0$ *in (a), we have that whenever* $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ *and the initial point* $\boldsymbol{x}^{(0)} \in Ball_D(\boldsymbol{x}^*, r_2)$ *with* $\boldsymbol{x}^* \in R_d$*,*

$$d_E(\boldsymbol{x}^{(t)}, R_d) \leq \Upsilon^t \left|\left|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right|\right|_2 \quad \textit{with} \quad \Upsilon = \sqrt{1 - \frac{\beta_0\eta}{4}}.$$

We further assume Assumptions E1 and E2 in the remaining statements. If $h \to 0$ *and* $\frac{nh^{D+4}}{|\log h|} \to \infty$*,*

(c) ***Q-Linear convergence of*** $\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2$*: under the same radius* $r_2 > 0$ *and* $\Upsilon = \sqrt{1 - \frac{\beta_0\eta}{4}}$ *in (a), we have that*

$$\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+4}}}\right)$$

with probability tending to 1 whenever $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ *and the initial point* $\widehat{\boldsymbol{x}}^{(0)} \in Ball_D(\boldsymbol{x}^*, r_2)$ *with* $\boldsymbol{x}^* \in R_d$*.*

(d) ***R-Linear convergence of*** $d_E(\widehat{\boldsymbol{x}}^{(t)}, R_d)$*: under the same radius* $r_2 > 0$ *and* $\Upsilon = \sqrt{1 - \frac{\beta_0\eta}{4}}$ *in (a), we have that*

$$d_E(\widehat{\boldsymbol{x}}^{(t)}, R_d) \leq \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+4}}}\right)$$

with probability tending to 1 whenever $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ *and the initial point* $\widehat{\boldsymbol{x}}^{(0)} \in Ball_D(\boldsymbol{x}^*, r_2)$ *with* $\boldsymbol{x}^* \in R_d$.

The detailed proof of Theorem 3.10 can be found in Appendix B.4. Note that, as in (c) of Proposition 3.7, we elucidate a threshold value for the convergence radius $r_2 > 0$ in (a), under which the population SCGA algorithm converges linearly to R_d. The first three quantities in the threshold value are directly adopted from the upper bound of the convergence radius r_1 in (c) of Proposition 3.7, while the last term controls the "subspace constrained strong concavity" (3.26) of p within $R_d \oplus r_2$.

Remark 3.3. *Notice that the standard strong concavity assumption on the objective function (or density function)* p *is not sufficient to establish the linear convergence of the population SCGA algorithm* (3.23). *This is because, under the (quasi-)strong concavity assumption (Necoara et al., 2019), the objective function* p *would satisfy*

$$p(\boldsymbol{x}^*) - p(\boldsymbol{y}) \leq \nabla p(\boldsymbol{y})^T(\boldsymbol{x}^* - \boldsymbol{y}) - \frac{A_6}{2}||\boldsymbol{x}^* - \boldsymbol{y}||_2^2 \tag{3.25}$$

for some constant $A_6 > 0$, *and those standard proofs of the linear convergence of gradient ascent methods rely on this inequality; see Section 3.4 in Bubeck (2015). However, as indicated in our proof of Theorem 3.10, the linear convergence of the SCGA algorithm requires the following inequality instead:*

$$p(\boldsymbol{x}^*) - p(\boldsymbol{y}) \leq \nabla p(\boldsymbol{y})^T V_d(\boldsymbol{y}) V_d(\boldsymbol{y})^T(\boldsymbol{x}^* - \boldsymbol{y}) - \frac{A_7}{2}||\boldsymbol{x}^* - \boldsymbol{y}||_2^2 + o\left(||\boldsymbol{x}^* - \boldsymbol{y}||_2^2\right) \tag{3.26}$$

for some constant $A_7 > 0$, *where* $\boldsymbol{y}$ *is generally chosen to be* $\boldsymbol{x}^{(t)}$. *We call a function* p *satisfying* (3.26) *"subspace constrained strongly concave". Since*

$$\nabla p(\boldsymbol{y})^T(\boldsymbol{x}^* - \boldsymbol{y}) = \nabla p(\boldsymbol{y})^T V_d(\boldsymbol{y}) V_d(\boldsymbol{y})^T(\boldsymbol{x}^* - \boldsymbol{y}) + \nabla p(\boldsymbol{y})^T U_d^\perp(\boldsymbol{y})(\boldsymbol{x}^* - \boldsymbol{y}),$$

the strong concavity assumption (3.25) *will not imply the key inequality* (3.26) *for the linear convergence of the population SCGA algorithm unless the residual gradient term* $\nabla p(\boldsymbol{y})^T U_d^\perp(\boldsymbol{y})(\boldsymbol{x}^* -$

$\boldsymbol{y}$) *can be upper bounded by the second-order error term* $O\left(||\boldsymbol{x}^* - \boldsymbol{y}||_2^2\right)$. *The imposed eigen-gap Assumption 2.2 as well as Assumption A4 with its related discussion in Appendix B.5 fill in this gap, ensuring that such a quadratic upper bound holds on the residual gradients along the SCGA sequence.*

Corollary 3.11 (Linear Convergence of the SCMS Algorithm)**.** *Assume that Assumptions 2.1–2.3, A4, E1, and E2 hold. When the fixed sample size* n *is sufficiently large and the bandwidth* h *is chosen to be sufficiently small, there exists a convergence radius* $r_3 \in (0, r_2)$ *such that the SCMS sequence* $\left\{\widehat{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ *satisfies the following property with high probability:*

$$d_E(\widehat{\boldsymbol{x}}^{(t)}, \widehat{R}_d) \leq \left|\left|\widehat{\boldsymbol{x}}^{(t)} - \widehat{\boldsymbol{x}}^*\right|\right|_2 \leq \Upsilon_{n,h}^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \widehat{\boldsymbol{x}}^*\right|\right|_2$$
$$\text{with} \quad \Upsilon_{n,h} = \sqrt{1 - \frac{\beta_0 \widetilde{\eta}_{n,h}}{4}} \quad \text{and} \quad \widetilde{\eta}_{n,h} = \inf_t \eta_{n,h}^{(t)}$$

whenever $0 < \sup_t \eta_{n,h}^{(t)} \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ *and the initial point* $\widehat{\boldsymbol{x}}^{(0)} \in \widehat{R}_d \oplus r_3$.

Corollary 3.11 should also be regarded as the linear convergence of the sample-based SCGA algorithm to the estimated ridge $\widehat{R}_d$ defined by the Euclidean KDE $\widehat{p}_n$. Based on Assumptions E1 and E2 and the uniform bounds (2.27), $\widehat{p}_n$ together with its ridge $\widehat{R}_d$ and sample-based SCGA sequence $\left\{\widehat{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ satisfy Assumptions 2.1–2.3 and A4 with probability tending to 1 as $h \to 0$ and $\frac{nh^{D+6}}{|\log h|} \to \infty$. As a result, one can follow our argument in part (a) of Theorem 3.10 to establish the linear convergence of the sample-based SCGA algorithm with a fixed step size η satisfying $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$. Furthermore, when the fixed sample size n is sufficiently large and the bandwidth h is chosen to be small, the adaptive step size $\eta_{n,h}^{(t)}$ of the SCMS algorithm always falls below the threshold $\min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ for linear convergence but is also uniformly bounded away from zero with respect to the iteration number t; see Remark 3.2. By taking the infimum of the adaptive step size $\eta_{n,h}^{(t)}$ with respect to t, one can thus establish the linear convergence of the SCMS algorithm with rate of convergence $\Upsilon_{n,h} = \sqrt{1 - \frac{\beta_0 \widetilde{\eta}_{n,h}}{4}}$ and $\widetilde{\eta}_{n,h} = \inf_t \eta_{n,h}^{(t)}$.

3.6 *Directional Subspace Constrained Mean Shift Algorithm and its Linear Convergence*

Throughout this section, we denote by $\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\}_{t=0,1,\ldots} \subset \Omega_q$ the iterative sequence generated by our directional SCMS (DirSCMS) algorithm. The fixed-point iteration formula (3.6) of the DMS algorithm suggests the following formulation of the DirSCMS algorithm as:

$$\widehat{\underline{\boldsymbol{x}}}^{(t+1)} \leftarrow \widehat{\underline{\boldsymbol{x}}}^{(t)} + \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \cdot \frac{\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})}{\left|\left|\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2} \quad \text{and} \quad \widehat{\underline{\boldsymbol{x}}}^{(t+1)} \leftarrow \frac{\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{\left|\left|\widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right|\right|_2}. \tag{3.27}$$

The DirSCMS update can be similarly written as a fixed-point iteration

$$\widehat{\underline{\boldsymbol{x}}}^{(t+1)} = \frac{\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left|\left|\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)}}{\left|\left|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left|\left|\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)}\right|\right|_2}. \tag{3.28}$$

Algorithm 2 in Section B.1 provides the detailed steps of implementing the DirSCMS algorithm in practice.

Inspired by Proposition 2 in Ghassabeh (2013) for the Euclidean SCMS algorithm, we derive the ascending property of our DirSCMS algorithm (3.27) and two convergence results for stopping the algorithm in the following proposition. The proof is deferred to Appendix B.7, in which our argument is similar to but logically different from the proof of Proposition 2 in Ghassabeh (2013).

Proposition 3.12. *Assume that the directional kernel L is non-increasing, twice continuously differentiable, and convex with $L(0) < \infty$. Given the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{L,q}(h)}{n}\sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ and the DirSCMS sequence $\left\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\right\}_{t=0}^{\infty} \subset \Omega_q$ defined by (3.27) or (3.28), the following properties hold:*

(a) The estimated density sequence $\left\{\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$ is non-decreasing and thus converges.

(b) $\lim_{t\to\infty}\left|\left|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2 = 0.$

(c) If the kernel L is also strictly decreasing on $[0,\infty)$, then $\lim_{t\to\infty}\left|\left|\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right|\right|_2 = 0.$

Remark 3.4. *Our results (b) and (c) in Proposition 3.12 demonstrate that the stopping criterion of our DirSCMS algorithm can follow either the norm of the principal Riemannian gradient estimator* $\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2$ *or the (Euclidean) distance* $\left|\left|\underline{\widehat{\boldsymbol{x}}}^{(t+1)}-\underline{\widehat{\boldsymbol{x}}}^{(t)}\right|\right|_2$ *between two consecutive iterative points, where the latter requires a strictly decreasing kernel such as the von Mises kernel* $L(r)=e^{-r}$.

Motivated by the iterative formula (3.19) for the gradient ascent algorithm on Ω_q, we consider writing our DirSCMS algorithm as a variant of the SCGA algorithm on Ω_q with an iterative formula:

$$\underline{\widehat{\boldsymbol{x}}}^{(t+1)}=\mathtt{Exp}_{\underline{\widehat{\boldsymbol{x}}}^{(t)}}\left(\underline{\eta}_{n,h}^{(t)'}\cdot\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\mathtt{grad}\,\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right), \tag{3.29}$$

where $\mathtt{Exp}_{\boldsymbol{x}}$ is the exponential map at $\boldsymbol{x}\in\Omega_q$ and $\underline{\eta}_{n,h}^{(t)'}>0$ is the adaptive step size. As with the Euclidean SCMS algorithm and its SCGA representation (3.22), the formulation (3.29) will reveal the (linear) convergence properties of our DirSCMS algorithm. To derive an explicit formula for $\underline{\eta}_{n,h}^{(t)'}$, we recall the fixed-point equation (3.28) of our DirSCMS algorithm and compute the geodesic distance between $\underline{\widehat{\boldsymbol{x}}}^{(t+1)}$ and $\underline{\widehat{\boldsymbol{x}}}^{(t)}$ (one-step DirSCMS update) as:

$$\begin{aligned}\arccos\left(\left(\underline{\widehat{\boldsymbol{x}}}^{(t+1)}\right)^T\underline{\widehat{\boldsymbol{x}}}^{(t)}\right)&=\arccos\left(\frac{\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2}{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\cdot\underline{\widehat{\boldsymbol{x}}}^{(t)}\right|\right|_2}\right)\\&=\left|\left|\underline{\eta}_{n,h}^{(t)'}\cdot\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\mathtt{grad}\,\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2,\end{aligned}$$

where, in the second equality, we equate the geodesic distance between $\underline{\widehat{\boldsymbol{x}}}^{(t+1)}$ and $\underline{\widehat{\boldsymbol{x}}}^{(t)}$ to the norm of the tangent vector inside the exponential map in (3.29). This suggests that our DirSCMS algorithm is a (sample-based) SCGA algorithm on Ω_q with adaptive step size

$$\underline{\eta}_{n,h}^{(t)'}=\frac{\arccos\left(\frac{\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2}{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\cdot\underline{\widehat{\boldsymbol{x}}}^{(t)}\right|\right|_2}\right)}{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2}=\frac{\theta_t'}{\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\cdot\tan\theta_t'} \tag{3.30}$$

for $t=0,1,...$, where θ_t' denotes the angle between $\underline{\widehat{\boldsymbol{x}}}^{(t+1)}$ and $\underline{\widehat{\boldsymbol{x}}}^{(t)}$. Note that the above derivation is based on the orthogonality between $\underline{\widehat{\boldsymbol{x}}}^{(t)}$ and the order-d principal or projected

Riemannian gradient estimator

$$\widehat{\underline{G}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)}) = \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \mathtt{grad}\, \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) = \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}).$$

When our DirSCMS algorithm approaches the estimated ridge $\widehat{\underline{R}}_d$, θ_t' tends to 0 and $\frac{\theta_t'}{\tan\theta_t'}$ is approximately equal to 1. Thus, the step size $\underline{\eta}_{n,h}^{(t)'}$ is also controlled by $\left|\left|\nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2$ as in the DMS scenario; see Equation (3.20). Therefore, Lemma 3.6 is still useful for showing that the step size $\underline{\eta}_{n,h}^{(t)'}$ converges to 0 with probability tending to 1 when $h \to 0$ and $nh^q \to \infty$.

As shown in (3.29), our proposed DirSCMS algorithm is an example of the sample-based SCGA method with directional KDE $\widehat{f}_h$ on Ω_q and an adaptive step size $\eta_{n,h}^{(t)'}$. Our main focus in this subsection is to study the (linear) convergence of such an SCGA algorithm on Ω_q. We first consider the population SCGA algorithm on Ω_q, defined by the iterative formula

$$\underline{\boldsymbol{x}}^{(t+1)} = \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\underline{\eta} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right) \tag{3.31}$$

with a suitable choice of the step size $\underline{\eta} > 0$. The sample-based version substitutes the subspace constrained (or principal) Riemannian gradient $\underline{V}_d(\underline{\boldsymbol{x}})\underline{V}_d(\underline{\boldsymbol{x}})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}})$ with its estimator $\widehat{\underline{V}}_d(\underline{\boldsymbol{x}})\widehat{\underline{V}}_d(\underline{\boldsymbol{x}})^T \mathtt{grad}\, \widehat{f}_h(\underline{\boldsymbol{x}})$, which typically has a constant step size $\underline{\eta} > 0$; see (3.29). Throughout this section, we denote the sequence defined by the population SCGA algorithm with objective function f on Ω_q by $\{\underline{\boldsymbol{x}}^{(t)}\}_{t=0}^{\infty}$ and the sequence defined by the sample-based SCGA algorithm with objective function $\widehat{f}_h$ on Ω_q by $\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\}_{t=0}^{\infty}$.

Remark 3.5. *The definition (3.31) of the SCGA algorithm is applicable to any Riemannian manifold $\mathcal{M}$, rather than being restricted to the unit hypersphere Ω_q. The only requirement on $\mathcal{M}$ for (3.31) to be valid is that the exponential map $\mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}} : T_{\underline{\boldsymbol{x}}^{(t)}}(\mathcal{M}) \to \mathcal{M}$ is well-defined within a small neighborhood of $\mathbf{0}$ on the tangent space $T_{\underline{\boldsymbol{x}}^{(t)}}(\mathcal{M})$ for each $t \geq 0$. More importantly, Assumptions 2.1–2.3 and A4 can be generalized to any smooth function f supported on $\mathcal{M}$, and our (linear) convergence results are applicable to the SCGA algorithm (3.31) on $\mathcal{M}$ whose sectional curvature is lower bounded by a real number; see one of the key lemmas in our proofs (Lemma B.7).*

As with the SCGA algorithm in Euclidean space $\mathbb{R}^D$, the following proposition demonstrates that the SCGA algorithm (3.31) on Ω_q yields a non-decreasing sequence of the objective function f supported on Ω_q and an SCGA sequence that converges to the directional ridge $\underline{R}_d$, as long as the step size $\underline{\eta}$ is sufficiently small.

Proposition 3.13 (Convergence of the SCGA Algorithm on Ω_q)**.** *For any SCGA sequence* $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty} \subset \Omega_q$ *defined by* (3.31) *with* $0 < \underline{\eta} < \frac{2}{q||f||_{\infty}^{(2)}}$*, the following properties hold:*

(a) Under Assumption 2.1, the objective function sequence $\left\{f(\underline{\boldsymbol{x}}^{(t)})\right\}_{t=0}^{\infty}$ *is non-decreasing and thus converges.*

(b) Under Assumption 2.1, $\lim_{t\to\infty} \left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2 = \lim_{t\to\infty} d_g\left(\underline{\boldsymbol{x}}^{(t+1)}, \underline{\boldsymbol{x}}^{(t)}\right) = 0.$

(c) Under Assumptions 2.1–2.3, $\lim_{t\to\infty} d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{R}_d\right) = 0$ *whenever* $\underline{\boldsymbol{x}}^{(0)} \in \underline{R}_d \oplus \underline{r}_1$ *with the convergence radius* $\underline{r}_1$ *satisfying*

$$0 < \underline{r}_1 < \min\left\{\underline{\rho}/2, \frac{\min\left\{\underline{\beta}_1, 1\right\}^2}{\underline{A}_2\left(||f||_{\infty}^{(3)} + ||f||_{\infty}^{(4)}\right)}, 2\sin\left(\frac{\underline{\beta}_1}{2\underline{B}_4(f)}\right)\right\},$$

where $\underline{A}_2$ *is a constant defined in (h) of Lemma A.2 while* $\underline{B}_4(f) > 0$ *is a quantity depending on both the dimension* q *and the functional norm* $||f||_{\infty,4}^*$ *up to the fourth-order (partial) derivatives of* f*.*

The proof of Proposition 3.13 can be found in Appendix B.7. The upper bound for the convergence radius $\underline{r}_1 > 0$ has the same meaning as in Proposition 3.7 for the Euclidean SCGA algorithm, ensuring that $\underline{r}_1 \leq \mathtt{reach}(\underline{R}_d)$ and the distances from the SCGA sequence $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ on Ω_q to the directional ridge $\underline{R}_d$ can be upper bounded by the norms $\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2$ of order-d principal Riemannian gradients for all $t = 0, 1, \ldots$.

Corollary 3.14 (Convergence of the DirSCMS Algorithm)**.** *When the fixed sample size n is sufficiently large and the bandwidth h is chosen to be correspondingly small, the following properties hold for the DirSCMS sequence $\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\}_{t=0}^{\infty} \subset \Omega_q$ with high probability under Assumptions 2.1–2.3 and Assumptions D1 and D2:*

(a) The directional KDE sequence $\left\{\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$ is non-decreasing and thus converges.

(b) $\lim_{t\to\infty} \left|\left|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \mathtt{grad}\, \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right|\right|_2 = \lim_{t\to\infty} d_g\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right) = 0.$

(c) $\lim_{t\to\infty} d_g\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \widehat{\underline{R}}_d\right) = 0$ *whenever* $\widehat{\underline{\boldsymbol{x}}}^{(0)} \in \widehat{\underline{R}}_d \oplus \underline{r}_1$ *with the convergence radius* $\underline{r}_1 > 0$ *defined in (c) of Proposition 3.13.*

Corollary 3.14 should also be regarded as a convergence result for the sample-based SCGA algorithm on Ω_q. To justify Corollary 3.14, we know from Theorem 2.2 that Assumptions 2.1–2.3 also hold with high probability for the directional KDE $\widehat{f}_h$ and its estimated directional ridge $\widehat{\underline{R}}_d$ when $\frac{nh^{q+6}}{|\log h|}$ is sufficiently large and h is small enough. Further, by Lemma 3.6, the adaptive step size $\underline{\eta}_{n,h}^{(t)'}$ of our DirSCMS algorithm can be made smaller than the threshold value $\frac{2}{q||f||_{\infty}^{(2)}}$ in Proposition 3.13 while remaining uniformly bounded away from zero with respect to the iteration number t, given a sufficiently large but fixed sample n and a sufficiently small bandwidth h; recall Remark 3.2. As a result, Corollary 3.14 follows from Proposition 3.13. Notice that the statements in Proposition 3.12 are essentially the same as results (a-b) in Corollary 3.14 here. However, as in Proposition 2 of Ghassabeh (2013) for the Euclidean SCMS algorithm, Proposition 3.12 for the DirSCMS algorithm is established under the convexity assumption on the directional kernel L and holds for any sample size n and bandwidth h. By contrast, results (a-b) in Corollary 3.14 are asymptotic and probabilistic properties, in which we require $\frac{nh^{q+6}}{|\log h|} \to \infty$ and $h \to 0$.

According to Proposition 3.13 and Corollary 3.14, we can denote the limiting points of the population and sample-based SCGA algorithms on Ω_q by $\underline{\boldsymbol{x}}^* \in \underline{R}_d$ and $\widehat{\underline{\boldsymbol{x}}}^* \in \widehat{\underline{R}}_d$, respectively. The definition of the linear convergence of any converging sequence on Ω_q (or an arbitrary Riemannian manifold) is similar to the one in the flat Euclidean space $\mathbb{R}^D$ (see

Definition 3.9), except that the Euclidean distance is replaced with the geodesic distance on Ω_q in the definition; see Section 4.5 in Absil et al. (2008).

Using the notation in Zhang and Sra (2016), we let $\zeta(\kappa, c) \equiv \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)}$. Given that the sectional curvature is $\kappa = 1$ on Ω_q, we have $\zeta(1, c) = \frac{c}{\tanh(c)}$. One can show by differentiating $\zeta(1, c)$ that $\zeta(1, c)$ is strictly increasing with respect to c and $\zeta(1, c) > 1$ for any $c > 0$. Analogous to the Euclidean SCGA algorithms, we will establish the linear convergence of the SCGA sequence $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ on Ω_q (or any Riemannian manifold whose sectional curvature is lower bounded by a real number) as well as its sample-based version under the following local condition.

Assumption A4 (Quadratic Behaviors of Residual Vectors)**.** *We assume that the SCGA sequence* $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ *on* Ω_q *with step size* $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_{\infty}^{(2)} \cdot \zeta(1, \underline{\rho})}\right\}$ *and* $\underline{\boldsymbol{x}}^* \in \underline{R}_d$ *as its limiting point satisfies*

$$\left\langle \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}),\, \texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \right\rangle \leq \frac{\underline{\beta}_0}{4} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2,$$
$$\left|\left| \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \right|\right|_2 \leq \underline{\beta}_2 \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2$$

for some constant $\underline{\beta}_2 > 0$*, where* $\underline{\beta}_0 > 0$ *is the constant defined in Assumption 2.2 and* $\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \in T_{\underline{\boldsymbol{x}}^{(t)}}$ *is the logarithmic map.*

Assumption A4 serves as a generalization of its Euclidean counterpart Assumption A4 to Ω_q, which again requires a quadratic behavior of the residual vector $\underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)})\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*)$ within the tangent space $T_{\underline{\boldsymbol{x}}^{(t)}}$. Under this assumption, the objective (density) function f is "subspace constrained geodesically strongly concave" around the directional ridge $\underline{R}_d$; see also Remark 3.6. The discussion of potentially weaker assumptions that imply Assumption A4 in Appendix B.5 also applies in the manifold setting with some modifications; see Remark B.2. One intuitive example where Assumption A4 holds is presented in the second row of Figure 3.4, where the directional SCMS/SCGA iterative vector $\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*)$ is always orthogonal to the residual space $\underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)})$ for all $t \geq 0$ around the (estimated) ridge on Ω_q.

Theorem 3.15 (Linear Convergence of the SCGA Algorithm on Ω_q)**.** *Assume that Assumptions 2.1–2.3 and A4 hold throughout the theorem.*

(a) ***Q-Linear convergence of*** $d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*)$*: Consider a convergence radius* $\underline{r}_2 > 0$ *satisfying*

$$0 < \underline{r}_2 \leq \min\left\{\underline{\rho}/2, \frac{\underline{\beta}_1^2}{\underline{A}_2\left(||f||_\infty^{(3)} + ||f||_\infty^{(4)}\right)}, \frac{\underline{\beta}_1}{\underline{B}_4(f)}, \right.$$
$$\left. 2\sin\left[\frac{3\underline{\beta}_0}{8q\left(12\,||f||_\infty^{(2)}\,\underline{\beta}_2^2\arcsin\left(\underline{\rho}/2\right) + \sqrt{q}\,||f||_\infty^{(3)}\right)}\right]\right\},$$

where $\underline{A}_2 > 0$ *is the constant defined in (h) of Lemma A.2 and* $\underline{B}_4(f) > 0$ *is a quantity defined in (c) of Proposition 3.13 that depends on both the dimension* q *and the functional norm* $||f||_{\infty,4}^*$ *up to the fourth-order (partial) derivatives of* f*. Whenever* $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)}\cdot\zeta(1,\underline{\rho})}\right\}$ *and* $\underline{\boldsymbol{x}}^* \in \underline{R}_d$ *is the limiting point specified in Assumption A4 with* $\underline{\boldsymbol{x}}^{(0)} \in Ball_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$*, we have that*

$$d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*) \leq \underline{\Upsilon}^t \cdot d_g(\underline{\boldsymbol{x}}^{(0)}, \underline{\boldsymbol{x}}^*) \quad \text{with} \quad \underline{\Upsilon} = \sqrt{1 - \frac{\underline{\beta}_0\underline{\eta}}{4}}.$$

(b) ***R-Linear convergence of*** $d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{R}_d)$*: Under the same radius* $\underline{r}_2 > 0$ *in (a), we have that whenever* $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)}\cdot\zeta(1,\underline{\rho})}\right\}$ *and the initial point* $\underline{\boldsymbol{x}}^{(0)} \in Ball_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$ *with* $\underline{\boldsymbol{x}}^* \in \underline{R}_d$*,*

$$d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{R}_d) \leq \underline{\Upsilon}^t \cdot d_g(\underline{\boldsymbol{x}}^{(0)}, \underline{\boldsymbol{x}}^*) \quad \text{with} \quad \underline{\Upsilon} = \sqrt{1 - \frac{\underline{\beta}_0\underline{\eta}}{4}}.$$

We further assume Assumptions D1 and D2 in the remaining statements. Suppose that $h \to 0$ *and* $\frac{nh^{q+4}}{|\log h|} \to \infty$*.*

(c) ***Q-Linear convergence of*** $d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*)$*: Under the same radius* $\underline{r}_2 > 0$ *and* $\underline{\Upsilon} = \sqrt{1 - \frac{\underline{\beta}_0\underline{\eta}}{4}}$ *in (a), we have that*

$$d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*) \leq \underline{\Upsilon}^t \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right)$$

with probability tending to 1 whenever $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta(1,\underline{\rho})}\right\}$ *and the initial point* $\widehat{\underline{\boldsymbol{x}}}^{(0)} \in Ball_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$ *with* $\underline{\boldsymbol{x}}^* \in \underline{R}_d$.

(d) ***R-Linear convergence of*** $d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{R}_d)$*: Under the same radius* $\underline{r}_2 > 0$ *and* $\underline{\Upsilon} = \sqrt{1 - \frac{\underline{\beta}_0 \underline{\eta}}{4}}$ *in (a), we have that*

$$d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{R}_d) \leq \underline{\Upsilon}^t \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right)$$

with probability tending to 1 whenever $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta(1,\underline{\rho})}\right\}$ *and the initial point* $\widehat{\underline{\boldsymbol{x}}}^{(0)} \in Ball_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$ *with* $\underline{\boldsymbol{x}}^* \in \underline{R}_d$.

The detailed proof of Theorem 3.15 is in Appendix B.7. The theorem illuminates both the step size requirement and the convergence radius $\underline{r}_2 > 0$ for the linear convergence of SCGA algorithms on Ω_q. As with the Euclidean SCGA algorithms in Theorem 3.10, the upper bound of the convergence radius $\underline{r}_2$ consists of the three quantities adopted from Proposition 3.13 and a quantity controlling the "subspace constrained geodesically strong concavity" around the directional ridge $\underline{R}_d$.

Remark 3.6. *As with Euclidean SCGA algorithms, the geodesically strong concavity assumption (Zhang and Sra, 2016) on the objective function* f *is not sufficient to prove the linear convergence of the SCGA algorithm (3.31) on* Ω_q*. We instead establish the following "subspace constrained geodesically strong concavity" under Assumptions 2.1–2.3 and A4:*

$$f(\underline{\boldsymbol{x}}^*) - f(\boldsymbol{y}) \leq \left\langle \underline{V}_d(\boldsymbol{y})\underline{V}_d(\boldsymbol{y})^T \mathtt{grad}\, f(\boldsymbol{y}), \mathtt{Exp}_{\boldsymbol{y}}^{-1}(\underline{\boldsymbol{x}}^*) \right\rangle - A_8 \cdot d_g(\underline{\boldsymbol{x}}^*, \boldsymbol{y})^2 + o\left(d_g(\underline{\boldsymbol{x}}^*, \boldsymbol{y})^2\right) \quad (3.32)$$

for some constant $A_8 > 0$*, where* $\boldsymbol{y}$ *is generally chosen to be* $\underline{\boldsymbol{x}}^{(t)}$*. In fact, the most critical factors for establishing this property are the eigengap Assumption 2.2 and the quadratic behavior of residual vectors stated in Assumption A4.*

Corollary 3.16 (Linear Convergence of the DirSCMS Algorithm)**.** *Assume that Assumptions 2.1–2.3, A4, D1, and D2 hold. When the fixed sample size* n *is sufficiently large and the*

fixed bandwidth is chosen to be sufficiently small, there exists a convergence radius $\underline{r}_3 \in (0, \underline{r}_2)$ *such that the DirSCMS sequence* $\left\{\widehat{\underline{\boldsymbol{x}}}^{(t)}\right\}_{t=0}^{\infty}$ *satisfies*

$$d_g\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \widehat{\underline{R}}_d\right) \leq d_g\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \widehat{\underline{\boldsymbol{x}}}^{*}\right) \leq \underline{\Upsilon}_{n,h}^{t} \cdot d_g\left(\widehat{\underline{\boldsymbol{x}}}^{(0)}, \widehat{\underline{\boldsymbol{x}}}^{*}\right)$$

$$\textit{with} \quad \underline{\Upsilon}_{n,h} = \sqrt{1 - \frac{\underline{\beta}_0 \widetilde{\underline{\eta}}_{n,h}}{4}} \quad \textit{and} \quad \widetilde{\underline{\eta}}_{n,h} = \inf_t \underline{\eta}_{n,h}^{(t)'}$$

with high probability whenever $0 < \sup_t \underline{\eta}_{n,h}^{(t)'} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta(1,\underline{\rho})}\right\}$ *and the initial point* $\widehat{\underline{\boldsymbol{x}}}^{(0)} \in \left(\widehat{\underline{R}}_d \oplus \underline{r}_3\right) \cap \Omega_q$.

We also identify Corollary 3.16 as the linear convergence of the sample-based SCGA algorithm on Ω_q to the estimated directional ridge $\widehat{\underline{R}}_d$ defined by the directional KDE $\widehat{f}_h$. The corollary can be justified by noticing that, under Assumptions D1 and D2 and the uniform bounds (2.29), $\widehat{f}_h$ satisfies Assumptions 2.1–2.3 with probability tending to 1 as $h \to 0$ and $\frac{nh^{q+6}}{|\log h|} \to \infty$; see Theorem 2.2. With this fact, one can leverage our argument in part (a) of Theorem 3.15 to prove the linear convergence of the sample-based SCGA algorithm on Ω_q with a fixed step size $\underline{\eta}$ satisfying $0 < \underline{\eta} \leq \min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta(1,\underline{\rho})}\right\}$. Additionally, when the fixed sample size n is sufficiently large and the bandwidth is chosen to be sufficiently small, the adaptive step size $\underline{\eta}_{n,h}^{(t)'}$ of our DirSCMS algorithm in (3.30) always falls below the threshold value $\min\left\{\frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta(1,\underline{\rho})}\right\}$ for linear convergence by Lemma 3.6 but is also bounded away from zero; recall Remark 3.2. Taking the infimum of $\underline{\eta}_{n,h}^{(t)'}$ with respect to the iteration number t under a fixed n and h yields our results in Corollary 3.16.

3.7 Numerical Experiments

We conclude this chapter with numerical experiments that connect the preceding convergence theory to practical computation. First, we use simulated datasets to examine the empirical linear convergence behavior of the Euclidean and directional SCMS algorithms. Second, motivated by the cosmic web application developed further in Chapter 4, we apply the DMS algorithm to SDSS-IV galaxy observations and use the resulting trajectories to

detect basins of attraction associated with cosmic nodes. Detailed implementation choices, including parameter settings and numerical procedures, can be found in Section 5 of Zhang and Chen (2023, 2021a).

3.7.1 Simulation Studies for Euclidean SCMS Algorithm

To evaluate the algorithmic rate of convergence of the Euclidean SCMS algorithm (Algorithm 1), we generate the first simulated dataset by randomly drawing 1000 data points from a Gaussian mixture model with density $0.4 \cdot N(\boldsymbol{\mu}_1, \Sigma_1) + 0.6 \cdot N(\boldsymbol{\mu}_2, \Sigma_2)$, where $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (1,1)^T$, $\Sigma_1 = \texttt{Diag}\left(\frac{1}{4}, \frac{1}{4}\right)$, and $\Sigma_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}$. Another simulated dataset consists of 1000 data points randomly generated from an upper semicircle with radius 2 and i.i.d. Gaussian noise $N(0, 0.3^2)$. When applying Algorithm 1 with the estimated log-density to each of these two simulated datasets, we set the initial mesh equal to the simulated dataset and remove points whose density values are below 25% of the maximum density to obtain a cleaner ridge structure.

Figure 3.3 presents the Euclidean KDE plots, estimated density ridges from the Euclidean SCMS algorithm, and their (linear) convergence plots on the two simulated datasets. The linear trends of those plots in the second and third columns of Figure 3.3 empirically support the conclusions of Theorem 3.10 and Corollary 3.11 regarding the linear convergence of the Euclidean SCMS algorithm.

3.7.2 Simulation Studies for Directional SCMS Algorithm

As in our simulation study of the linear convergence of the Euclidean SCMS algorithm, we verify the linear convergence of our DirSCMS algorithm (Algorithm 2) on two different simulated datasets. One of them comprises 1000 data points randomly generated from a vMF mixture model $0.4 \cdot \text{vMF}\left(\underline{\boldsymbol{\mu}}_1, \nu_1\right) + 0.6 \cdot \text{vMF}\left(\underline{\boldsymbol{\mu}}_2, \nu_2\right)$ with $\underline{\boldsymbol{\mu}}_1 = (0,0,1)^T \in \Omega_2 \subset \mathbb{R}^3$, $\underline{\boldsymbol{\mu}}_2 = (1,0,0)^T \in \Omega_2 \subset \mathbb{R}^3$, and $\nu_1 = \nu_2 = 10$. The other simulated dataset is identical to the example in the right panel of Figure 3.1 and the underlying dataset in Figure A.2;

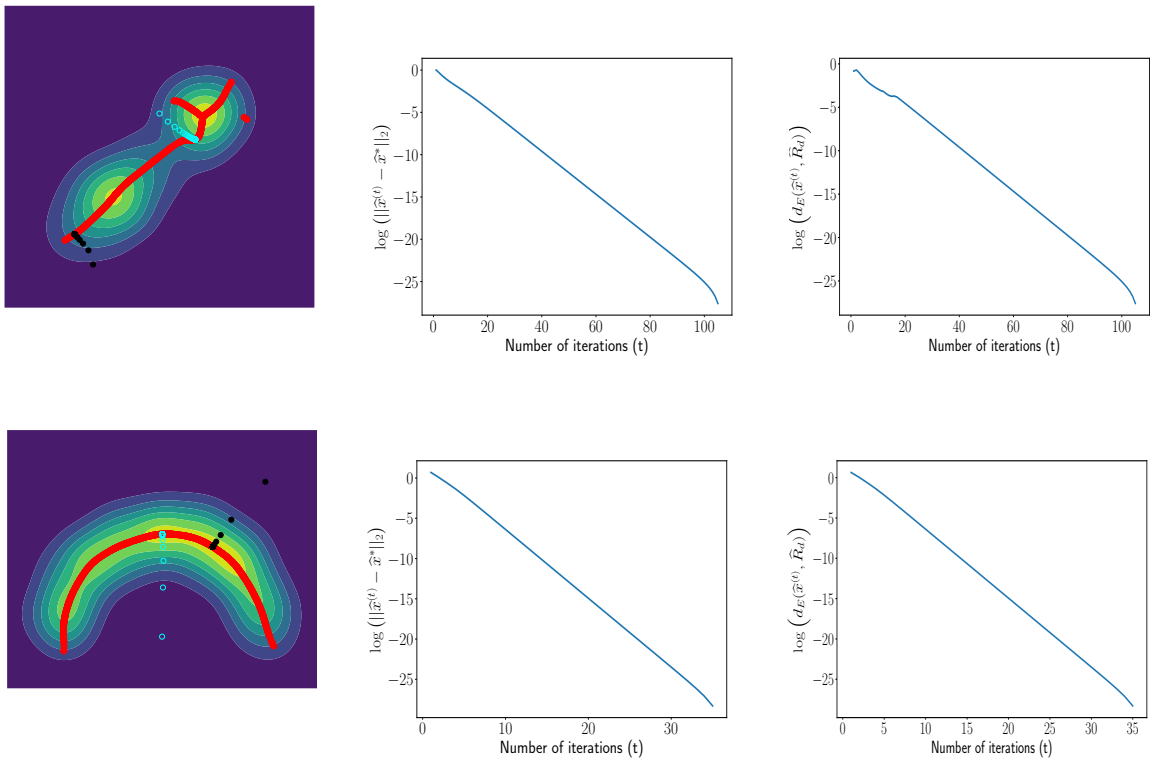


Figure 3.3: Density ridges estimated by the Euclidean SCMS algorithm on the two simulated datasets and their linear convergence plots. Horizontally, the first row displays the results of the simulated Gaussian mixture dataset, while the second row presents the results of the half circle simulated dataset. Vertically, the first column includes plots with Euclidean KDE, estimated ridges, and trajectories of SCMS sequences from two (randomly) chosen initial points. The second and third columns present the (linear) convergence plots for the log-distances of points in the highlighted sequences (indicated by hollow cyan points) to their limiting points or the estimated ridges.

it consists of 1000 randomly sampled points from a circle connecting the two poles on Ω_2, with i.i.d. additive Gaussian noise $N(0, 0.2^2)$ added to their Cartesian coordinates, followed by L_2 normalization onto Ω_2. In our implementation of Algorithm 2 with the directional log-density on the two simulated datasets, we set the initial mesh equal to the dataset and remove points whose density values are below 10% of the maximum density value from each mesh.

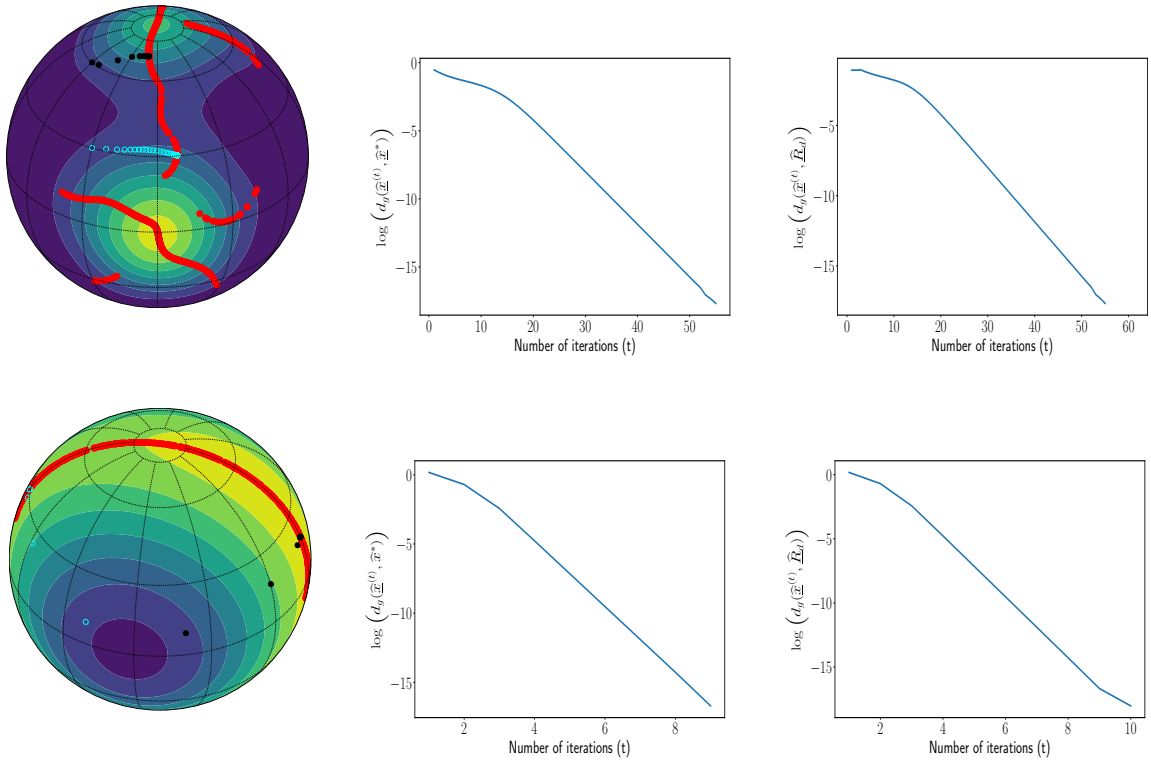


Figure 3.4: Density ridges estimated by the DirSCMS algorithm applied to the two simulated datasets and their linear convergence plots. Horizontally, the first row displays the results on the simulated vMF mixture dataset, while the second row presents the results on the circular simulated dataset on Ω_2. Vertically, the first column includes plots with directional KDE, estimated ridges, and trajectories of DirSCMS sequences from two (randomly) chosen initial points on Ω_2. The second and third columns present the convergence plots for the log-distances of points in the highlighted sequences (indicated by hollow cyan points) to their limiting points or the estimated ridges on Ω_2.

Figure 3.4 shows the directional KDE plots, estimated density ridges on Ω_2 from the DirSCMS algorithm, and their (linear) convergence plots on the aforementioned simulated datasets. Those linearly decreasing trends in the convergence plots, possibly after several pilot iterations, illustrate the local linear convergence of the DirSCMS algorithm that we proved in Theorem 3.15 and Corollary 3.16. Note that the minor perturbations at the tails of some linear convergence plots in Figure 3.4 are due to numerical precision errors.

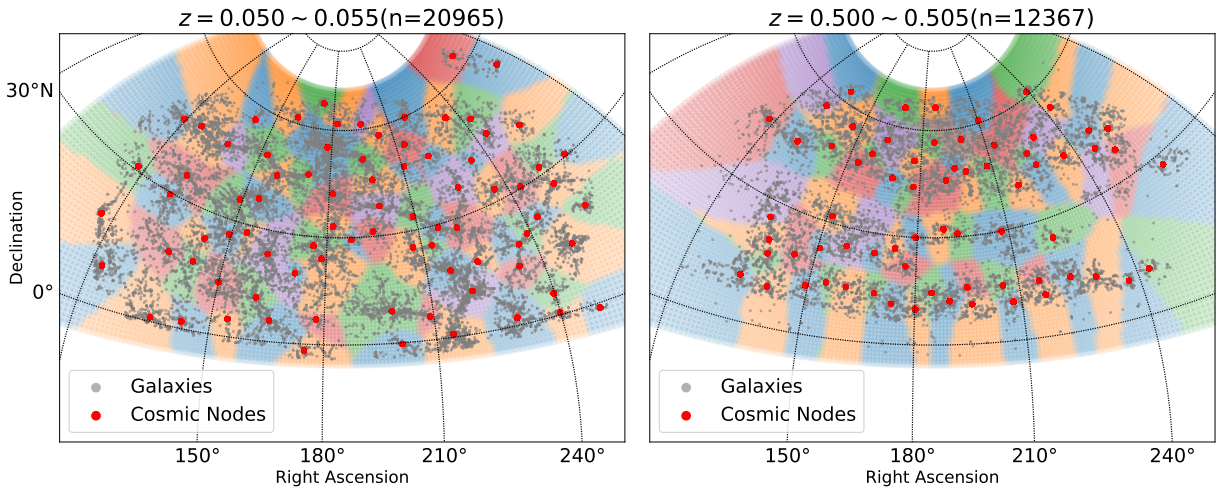


Figure 3.5: Basins of attraction of cosmic nodes detected by the DMS algorithm across two redshift slices from the Sloan Digital Sky Survey (SDSS). The sample size n denotes the number of observed galaxies in each slice.

3.7.3 Basins of Attraction For Cosmic Nodes

We apply the DMS algorithm to detect the basins of attraction of the cosmic nodes, which represent the gravitational catchment regions of the underlying density field. Unlike the Hamiltonian Monte Carlo approach of Valade et al. (2024), our DMS-based framework provides a non-parametric, computationally efficient alternative.

We utilize the galaxy observations from Data Release 16 (DR16) of the Sloan Digital Sky Survey (SDSS-IV; Ahumada et al. 2020). The galaxy samples and their associated stellar masses are obtained from the FIREFLY (Fitting IteRativEly For Likelihood analYsis) value-added catalog (Comparat et al., 2017; Wilkinson et al., 2017). We retain only the galaxies with reliable positive redshift measurements and non-missing stellar mass estimates. The stellar mass of each galaxy is inferred from SDSS spectra using the FIREFLY stellar population model, which estimates the stellar properties based on the Chabrier stellar initial mass function (Chabrier, 2003) and the MILES stellar library (Falcón-Barroso et al., 2011).

The location of each galaxy is recorded as (ξ_i, ϕ_i, z_i), where ξ_i is its right ascension (`RAsc`), ϕ_i is its declination (`DEC`), and z_i is its redshift value (`Z_NOQSO`). To mitigate redshift distortions along the line-of-sight directions (Sargent and Turner, 1977; Kaiser, 1987), we restrict our analysis to two thin redshift slices $0.05 \leq z < 0.055$ and $0.5 \leq z < 0.505$ within the North Galactic Cap ($100 <$ `RAsc` < 270 and $-5 <$ `DEC` < 70). These slices contain 20965 and 12367 galaxies, respectively. Within each redshift slice, the galaxy locations are naturally viewed as directional data on Ω_2. When applying the DMS algorithm, we employ the modified rule-of-thumb bandwidth selector in Eq. (14) of Zhang et al. (2022), multiplied by a factor of $1/3$ to induce mild undersmoothing and enhance sensitivity to local modes (cosmic nodes). This yields bandwidths $h \approx 0.077$ and $h \approx 0.404$ for the low and high redshift slices, respectively. The directional kernel is fixed to the von Mises kernel. Additionally, we remove the lowest 10% of galaxies in terms of estimated directional densities by (2.4) to reduce the influence of sparse regions before applying the DMS algorithm.

The resulting basins of attraction detected by the DMS algorithm are shown in Figure 3.5, where we obtain 83 basins of attraction for the low redshift slice $0.05 \leq z < 0.055$ and 69 basins of attraction for the high redshift slice $0.5 \leq z < 0.505$. In particular, the global convergence guarantee established in Section 3.3 ensures that almost every initialization on the North Galactic Cap converges to a local mode, providing theoretical support for the robustness of the detected cosmic nodes. Notably, local modes (cosmic nodes) do not necessarily coincide with geometric centers of their corresponding basins. Instead, both observed galaxies and grid points are attracted to these modes along gradient ascent trajectories of the directional KDE. Consequently, the identified basins of attraction can be interpreted as regions of dynamical influence in the large-scale galaxy density field.

To further explore astrophysical implications, we examine the relationship between the average stellar mass of galaxies and the mean density estimate within each detected basin of attraction. For each basin, the mean density estimate is computed as the average directional KDE value of galaxies within the basin, normalized by the overall mean directional KDE value of all galaxies in the corresponding redshift slice. Figure 3.6 presents scatter plots of

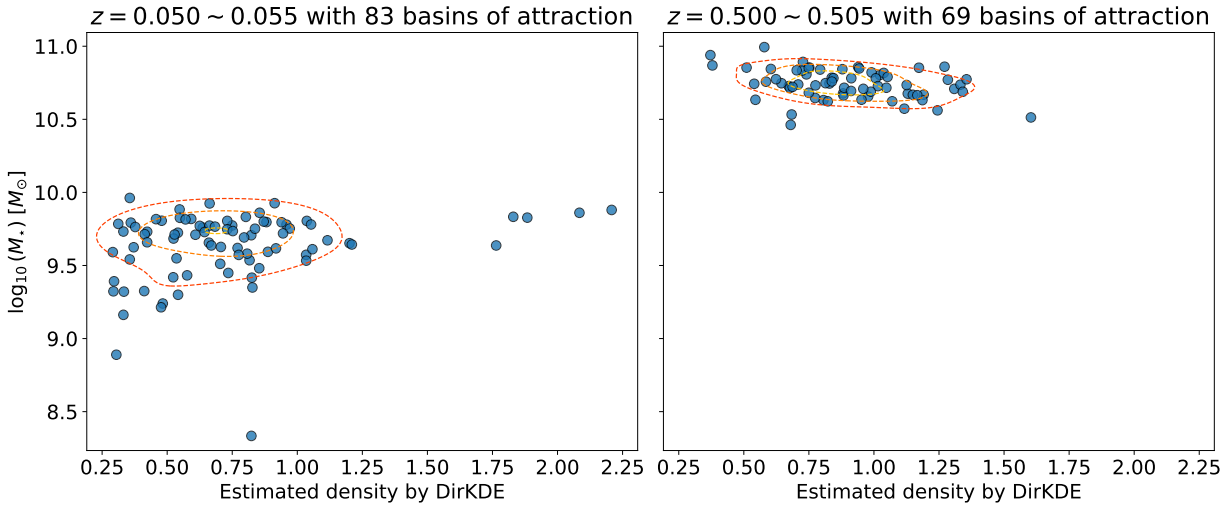


Figure 3.6: Scatter plot of the log-scale average stellar masses of galaxies versus the (normalized) mean estimated density obtained from the directional KDE (DirKDE; 2.4) across all identified basins of attraction for the two redshift slices. In each panel, the contour lines indicate the direct density estimate constructed from the displayed scatter points.

the log-scale average stellar mass against the (normalized) mean density for both redshift slices. Despite having the larger total number of galaxies, the basins of attraction in the low-redshift slice exhibit lower mean densities and lower average stellar masses than those in the high-redshift slice. Interestingly, within each slice, the average stellar mass remains relatively consistent across different basins, even when the mean estimated densities vary substantially across basins of attraction. This suggests that while the large-scale environment characterized by the basins of attraction dictates the spatial distribution of galaxies or matter, the intrinsic stellar properties may be governed by more localized physical processes.

Overall, this empirical study illustrates the practical relevance of the DMS algorithm for analyzing large-scale directional data and detecting structurally meaningful regions in complex scientific datasets.

Chapter 4

SDSS-IV COSMIC WEB CATALOG RECONSTRUCTION

This chapter applies the directional mode and ridge models developed in Chapter 2 and the algorithms studied in Chapter 3 to construct a public cosmic web catalog from the Data Release 16 (DR16) of the Sloan Digital Sky Survey IV (SDSS-IV; Ahumada et al. 2020)[1]. The catalog covers the redshift range $0 \leq z < 3$ and is generated by `SCONCE` (**S**pherical and **CON**ic **C**osmic w**E**b; Zhang et al. 2022), a geometry-aware cosmic web finder based on directional and conic extensions of the mean shift and SCMS algorithms.

4.1 Introduction

The preceding two chapters developed the statistical and algorithmic foundations for estimating modes and density ridges on spherical domains. This chapter turns those ideas into an astronomical data product. Using the notation and methodology established in Chapter 2 and Chapter 3, we describe the application-specific choices required to construct a cosmic web catalog from SDSS-IV DR16. These choices include assembling galaxy and quasar samples, partitioning the data into tomographic spherical slices, estimating directional ridges as cosmic filaments within each slice, masking the irregular survey footprint, and documenting the released catalog.

This chapter also serves as a bridge between the methodological and inferential parts of the dissertation. Chapter 2 and Chapter 3 justify the use of directional ridges as geometry-aware statistical surrogates for cosmic filaments. The present chapter applies those tools to produce a catalog of filamentary and nodal structures. Later, Chapter 5 uses such cosmic web information as scientific input for association analysis, and Chapter 6 then turns to

[1] See `https://www.sdss4.org/dr16`.

causal analyses of galaxy properties.

The chapter is organized as follows. Section 4.2 describes how the galaxy and quasar (QSO) samples are extracted and combined. Section 4.3 explains the tomographic slicing strategy, and Section 4.4 connects each slice to the directional ridge framework on Ω_2. We then detail the full catalog-construction pipeline in Section 4.5 and summarize the released data products in Section 4.6. The code for constructing the catalog is available at `https://github.com/zhangyk8/SDSS-Cosmic-Web-Catalog/`.

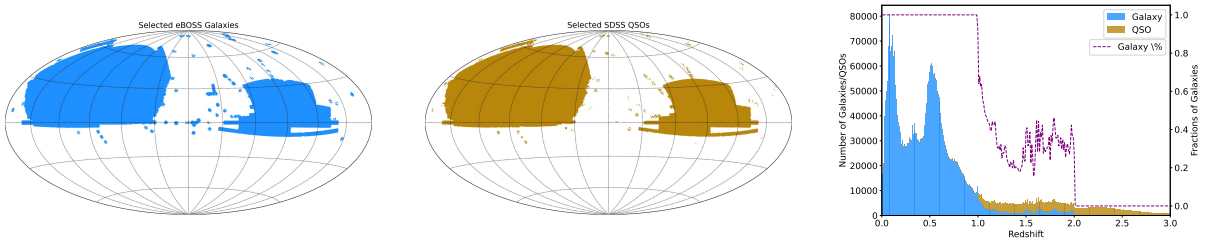
4.2 SDSS-IV Galaxy and QSO Samples

Compared with several existing cosmic web catalogs in the literature (Sousbie, 2011; Tempel et al., 2014; Einasto et al., 2014; Chen et al., 2016; Carrón Duque et al., 2022), our updated catalog uses galaxy and quasar (QSO) observations from Data Release 16 (DR16; Dawson et al. 2016; Ahumada et al. 2020)[2] of SDSS-IV (Ahumada et al., 2020). The goal of incorporating QSO observations is to exploit the depth and angular coverage of the SDSS-IV spectroscopic samples in order to recover filamentary structures over a wide redshift range. In this section, we describe in detail how these galaxy and QSO observations are selected and explain how they are combined into the data input for the DMS and DirSCMS algorithms in `SCONCE`.

4.2.1 Galaxy Observations

SDSS-IV contains three main survey programs: the Extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016), Mapping Nearby Galaxies at APO (MaNGA; Bundy et al. 2015), and the APO Galactic Evolution Experiment 2 (APOGEE-2; Majewski et al. 2017). Among these programs, eBOSS provides the galaxy sample most directly suited for reconstructing large-scale cosmic web structures. In DR16, the raw eBOSS spectra were processed with version `v5_13_0` of the redshift pipeline `idlspec2d` (Bolton et al., 2012; Daw-

[2]The final SDSS-IV release, Data Release 17 (Accetta et al., 2022), does not introduce additional observations that affect the construction considered here.



(a) Sky footprint of selected eBOSS Galaxies.

(b) Sky footprint of selected SDSS QSOs.

(c) Histograms of galaxies and QSOs across the redshift range.

Figure 4.1: Sky footprints of the selected eBOSS galaxies and SDSS QSOs, together with their distributions across redshift. The footprint plots are centered at RAsc 90° so that the left and right portions correspond to the North and South Galactic Caps, respectively.

son et al., 2013), which includes improvements to background fitting during extraction and updated stellar templates for flux calibration.

We select candidate galaxies from the eBOSS FIREFLY (Fitting IteRativEly For Likelihood analYsis) value-added catalog (Wilkinson et al., 2017; Comparat et al., 2017).[3] To avoid introducing additional uncertainty through ambiguous classifications or unreliable redshifts, we retain objects classified as galaxies with definite positive redshift estimates under the redshift pipeline. Specifically, we use the criteria `CLASS_NOQSO='galaxy'`, `RUN2D='26' or 'v5_13_0'`, and `Z_NOQSO > Z_ERR_NOQSO > 0`. The resulting sample contains 3,606,004 galaxies and covers approximately the redshift range $0 < z \leq 2$. Among these galaxies, 67.7% lie in the designated North Galactic Cap region, defined by $100° < \text{RAsc} < 270°$ and $-5° < \text{DEC} < 70°$. The sky footprint and redshift distribution of the selected galaxies are shown in Figure 4.1. Throughout this chapter, right ascension (RAsc) and declination (DEC) are recorded in J2000 degrees.

[3]The data file used here is https://data.sdss.org/sas/dr16/eboss/spectro/firefly/v1_1_1/sdss_eboss_firefly-dr16.fits.

4.2.2 Quasar (QSO) Observations

DR16 also released a large catalog of spectroscopically confirmed QSOs (Lyke et al., 2020), including 225,082 objects that were new relative to previous data releases. We use the quasar-only catalog, whose objects have high-confidence classifications based on visual inspection and deep neural network classification (Busca and Balland, 2018). We retain objects classified as QSOs with positive redshift values, using the criteria `IS_QSO_FINAL=1` and `Z>0`.[4] We further restrict the QSO sample to $1 \leq z < 3$, since the number of QSOs above $z = 3$ is too small to support reliable filament reconstruction in our tomographic pipeline. The resulting QSO sample contains 749,742 objects, of which 70.0% lie in the designated North Galactic Cap region; see Figure 4.1 for the sky footprint and redshift distribution.

4.2.3 Final Dataset

Detecting cosmic web structures with high fidelity requires a sufficiently dense set of observations within each region of interest. Since our catalog is constructed through a tomographic analysis with filaments and nodes estimated separately within redshift slices (see Section 4.3), each slice should contain enough observations to support local density and ridge estimation. Thus, we combine the selected galaxy and QSO samples without additional trimming, which creates an overlapping redshift range $1 \leq z < 2$ between the two samples. As shown in panel (c) of Figure 4.1, the direct combination does not create an abrupt increase in sample size near the boundaries $z = 1$ and $z = 2$, reducing the risk of artificial structures induced by the data merging step. Moreover, because `SCONCE` includes an explicit denoising step before filament estimation (see Section 4.5), retaining a large number of potentially noisy observations is typically preferable.

---

[4]The QSO data file used here is `https://data.sdss.org/sas/dr16/eboss/qso/DR16Q/DR16Q_v4.fits`.

4.3 Tomographic Spherical Slicing

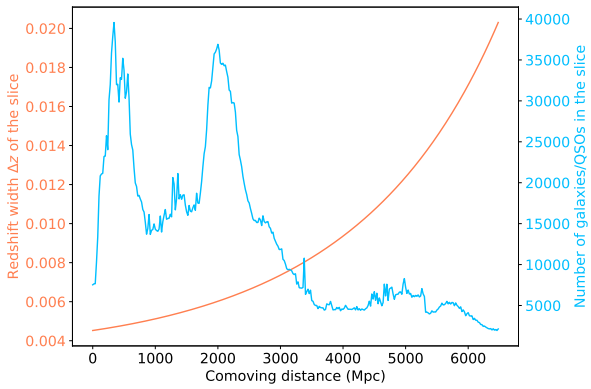
Each observation in our combined dataset in Section 4.2.3 has a unique spatial coordinate (RAsc,DEC,z) in our Universe. A common strategy in previous studies of cosmic web detection is to transform these coordinates into corresponding three-dimensional Cartesian coordinates in comoving-distance space, during which the redshift value is converted to a radial comoving distance based on a prespecified cosmological model (Tempel et al., 2014; Pfeifer et al., 2022). These full three-dimensional reconstructions, however, introduce several practical complications. The reconstruction is computationally demanding, and the observed galaxy density decreases with comoving distance. More importantly, recovering cosmic web structures in real space from redshift-space observations requires careful treatment of redshift distortions induced by peculiar velocities of the observed galaxies. These distortions are known as the finger-of-god (Sargent and Turner, 1977) and Kaiser (Kaiser, 1987) effects. In general, calibrating the redshift distortions is complicated and may not lead to a unique recovery of the real-space structures (Tempel et al., 2012). Recent studies on galaxy cluster outskirts further demonstrate that, because of the complex infall motions of galaxies toward filaments, simple statistical compression of finger-of-god elongations in virialized regions may not reliably recover filamentary structures (Kuchner et al., 2020, 2021).

Because of the aforementioned limitations of three-dimensional cosmic web reconstructions, we adopt a tomographic strategy similar to those used in Chen et al. (2016) and Carrón Duque et al. (2022). We partition the redshift range $0 < z < 3$ into non-overlapping spherical slices with equal comoving distances. Within each thin shell, the redshift values of galaxy/QSO observations are treated identically so that they can be regarded as a random directional sample on a fixed celestial sphere. We estimate the cosmic web structures, specifically nodes and filaments, within each spherical slice using the directional density ridge model implemented in `SCONCE`, and the slice-wise estimates are then combined into the final catalog.

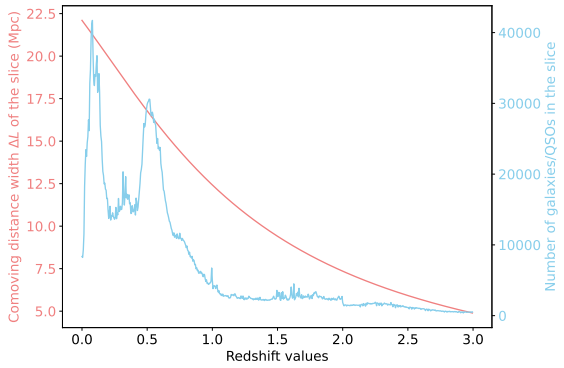
This tomographic construction has both advantages and limitations. On the one hand,

since different spherical slices are analyzed largely independently, the effect of local noise and survey inhomogeneity is largely confined to the corresponding redshift range. The approach also avoids imposing a potentially fragile correction for redshift distortions before filament estimation. On the other hand, slicing the data inevitably misses filamentary structures that are primarily aligned with the line of sight (Carrón Duque et al., 2022; Zhang et al., 2022). Furthermore, it cannot identify two-dimensional cosmic walls or sheets, because each slice is modeled as a two-dimensional nonlinear manifold (Lee, 2018). Nevertheless, filaments and nodes contain a substantial fraction of the cosmic-web mass budget over $0 < z < 3$ (Aragón-Calvo et al., 2010; Cautun et al., 2014; Zhu and Feng, 2017), and reliable recovery of these structures is the primary objective of the present catalog.

A remaining question is how to define the spherical slices. For low-redshift analyses, it is reasonable to directly divide the data into thin slices of equal redshift width, *e.g.*, $\Delta z = 0.005$. However, this choice becomes less appropriate when the target redshift range extends to $z = 3$, because equal redshift intervals correspond to unequal comoving-distance intervals in a cosmology of the expanding universe. As illustrated in panel (b) of Figure 4.2, slicing the data with equal width $\Delta z = 0.005$ produces much thinner comoving-distance shells at high redshift than at low redshift, leaving too few observations in many high-redshift slices for stable ridge estimation. As a result, we consider partitioning the sample into equal comoving-distance intervals of width $\Delta L = 20\,\mathrm{Mpc}$ as in Carrón Duque et al. (2022). The conversion from redshift to comoving distance uses the Planck15 ΛCDM cosmology with $H_0 = 67.7 km/(\,\mathrm{Mpc} \cdot s)$, $\Omega_m = 0.307$, and $\Omega_\Lambda = 0.693$ (Ade et al., 2016). Under this convention, the range $0 \leq z < 3$ corresponds approximately to $0 \leq L \lesssim 6500\,\mathrm{Mpc}$, yielding 325 slices of width 20 Mpc. As shown in panel (a) of Figure 4.2, the redshift width of a slice increases with redshift itself, ensuring enough observations in high-redshift shells for ridge (or filament) estimation. Notably, different reasonable cosmological parameter choices may shift the slice boundaries slightly, but these shifts are minor relative to the scale at which the catalog is constructed. In other words, our resulting catalog is independent of the cosmological model applied to the slicing process.



(a) Partition by comoving-distance width $\Delta L = 20\,\text{Mpc}$.



(b) Partition by redshift width $\Delta z = 0.005$.

Figure 4.2: Comparison between partitioning the combined sample into slices with equal comoving-distance width $\Delta L = 20\,\text{Mpc}$ and equal redshift width $\Delta z = 0.005$.

4.4 Filament Model Within Each Slice

After partitioning the combined sample into spherical slices, we analyze each slice using the directional density ridge framework developed in Chapter 2 and Chapter 3. For a fixed slice, let (α_i, δ_i) denote the RAsc and DEC of the i-th observation. We convert these angular coordinates into their associated Cartesian coordinates on Ω_2 by

$$\boldsymbol{X}_i = (\cos\delta_i \cos\alpha_i,\ \cos\delta_i \sin\alpha_i,\ \sin\delta_i) \in \Omega_2, \quad i = 1, ..., n. \tag{4.1}$$

Thus, the observations in each slice are treated as a directional sample on Ω_2.

We estimate the density generating the observations on Ω_2 using the directional KDE $\widehat{f}_h$ in (2.4), where the slice-specific bandwidth h is selected by the rule described in Section C.1. Then, the order-1 directional density ridges of $\widehat{f}_h$ are estimated by the DirSCMS algorithm from Section 3.6. In the resulting cosmic web catalog, the filamentary structures are represented by dense collections of converged DirSCMS points, which serve as discrete approximations to the estimated one-dimensional ridges. Furthermore, we identify candidate cosmic nodes within each slice as the union of local modes and filament intersections on Ω_2.

Specifically, the local modes of $\widehat{f}_h$ are estimated by the DMS procedure from Section 3.2, whereas filament intersections or knots are recovered using the metric graph reconstruction algorithm (Aanjaneya et al., 2011; Lecci et al., 2014); see Section 2.4 and Appendix B.4 in Zhang et al. (2022) for details. Finally, for each filamentary point, we compute an uncertainty measure given by its root mean square angular distance to the corresponding nonparametric bootstrap filament estimates on the same slice (Efron, 1981; Chen et al., 2015); see Appendix C in Zhang et al. (2022) for details.

4.5 Cosmic Web Detection Pipeline

Given the combined SDSS-IV galaxy and QSO sample, we construct the catalog separately on the North and South Galactic Caps. The pipeline consists of the following steps:

1. Convert redshift to comoving distance under the Planck15 cosmology and partition the combined sample over $0 \leq z < 3$ into 325 spherical slices of width $\Delta L = 20\,\text{Mpc}$; see Section 4.3.

2. Within each slice, restrict the observations to one of the two disconnected SDSS footprints:

   - **North Galactic Cap:** $\{(\alpha, \delta) : 100^\circ < \alpha < 270^\circ, -5^\circ \leq \delta < 70^\circ\}$.
   - **South Galactic Cap:**

   $$\{(\alpha, \delta) : 300^\circ < \alpha < 360^\circ, -15^\circ \leq \delta < 70^\circ\} \cup \{(\alpha, \delta) : 0^\circ \leq \alpha < 100^\circ, -15^\circ \leq \delta < 70^\circ\}.$$

   Here, (α, δ) denotes the (RAsc,DEC) coordinate system in J2000 degrees.

3. Select the bandwidth using (C.1) in Section C.1 with $A_0 = 0.25$, evaluate the directional KDE $\widehat{f}_h$ in (2.4) on all observations within the slice, and remove noisy observations according to the threshold rule (C.2).

4. Lay down a uniform angular grid over the North Galactic Cap with 170×75 initial mesh points and over the South Galactic Cap with 160×85 initial mesh points. The South Galactic Cap grid is understood on the wrapped RAsc interval $[300°, 360°) \cup [0°, 100°)$. To speed up convergence, evaluate $\widehat{f}_h$ at the mesh points after converting them to Cartesian coordinates on Ω_2 and discard the lowest-density 15% of mesh points on each cap.

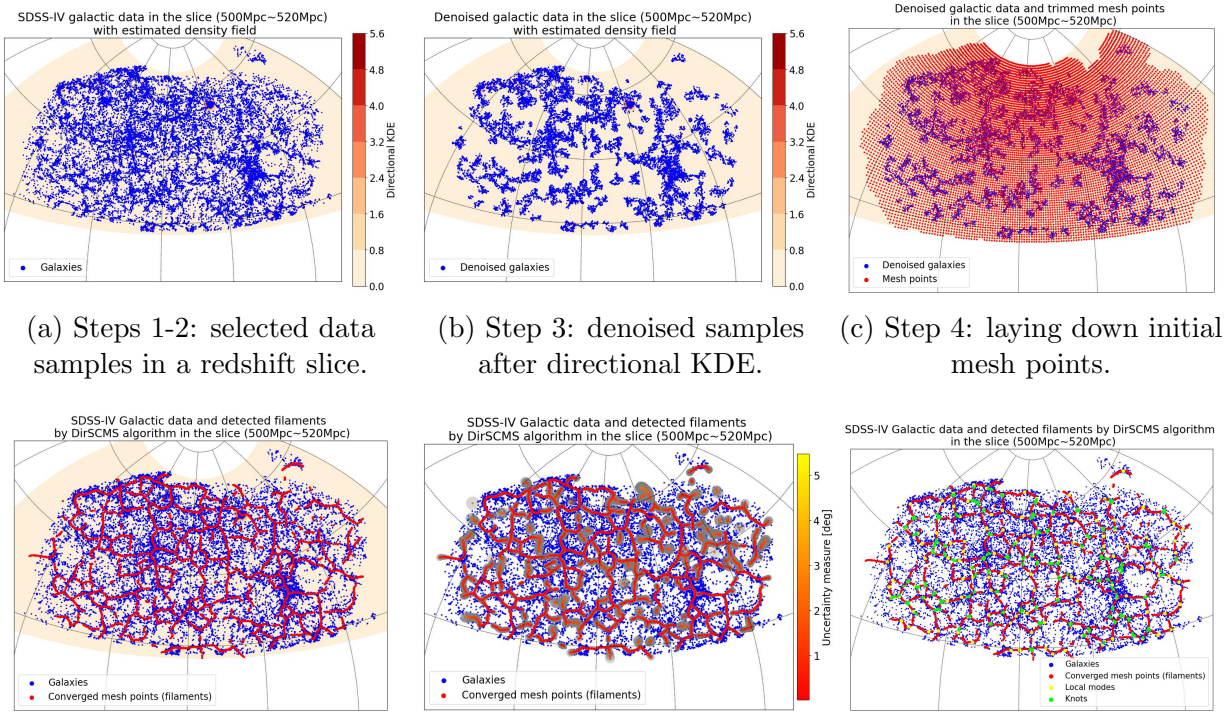
5. Apply the DirSCMS algorithm from Section 3.6 to estimate the filamentary ridges on each slice.

6. Quantify the uncertainty of each detected filamentary point using the nonparametric bootstrap.

7. Identify the local modes and knots from the detected filamentary points as cosmic nodes using the DMS and metric graph reconstruction algorithms.

8. Construct a mask from the sky coverage of the combined data within each slice and remove filamentary and nodal points that fall outside the observed region.

Here, the mask within each spherical slice is constructed using the HEALPix (Hierarchical Equal Area iso-Latitude Pixelization; Górski et al. 2005) scheme. In particular, we discretize the celestial sphere into 12×32^2 pixels, corresponding to $N_{\text{side}} = 32$, and count the number of observations in each pixel. For slices with comoving distance below 3000 Mpc, we retain only pixels containing more than two observations. For slices with comoving distance above 3000 Mpc, we do not remove pixels by this count threshold, which stabilizes the fraction of removed filamentary points at approximately 10% across the full range $0 \leq z < 3$. We then add the adjacent pixels of each retained pixel to the final mask, effectively expanding the retained region by approximately 1°. Finally, filamentary and nodal points outside the resulting mask are excluded from our released cosmic web catalog. Figure 4.3 illustrates the pipeline on the North Galactic Cap in the comoving-distance slice 500 Mpc–520 Mpc.



(a) Steps 1-2: selected data samples in a redshift slice.

(b) Step 3: denoised samples after directional KDE.

(c) Step 4: laying down initial mesh points.

(d) Step 5: filaments detected by DirSCMS.

(e) Step 6: filament uncertainty measures.

(f) Step 7: candidate cosmic nodes on detected filaments.

Figure 4.3: Detailed procedures of our cosmic web detection pipeline on the North Galactic Cap in the comoving-distance slice 500 Mpc–520 Mpc. The plots are centered at $\alpha_0 = 185°$ and $\delta_0 = 30°$ under the stereographic projection.

4.6 Cosmic Web Catalog Data

The final cosmic web catalog is publicly available on Zenodo at https://zenodo.org/records/6244866 (Zhang, 2022). The release follows the tomographic construction described above and contains detected filamentary structures and cosmic node candidates across 325 spherical slices spanning the redshift range $0 \leq z < 3$. Each slice corresponds to a comoving-distance shell of width 20 Mpc under the Planck15 cosmology.

The catalog is distributed in both CSV and FITS formats. It contains two primary data products: `Cosmic_filaments_2D_DirSCMS_new1`, which stores a discrete point repre-

sentation of the estimated filamentary ridges, and `Cosmic_local_modes_2D_DirMS_unique1`, which stores the estimated local modes of the smoothed galaxy/QSO density field. In our framework, these local modes form one class of candidate cosmic nodes. A second class is given by filament intersections, which are recorded through the `knot_label` column in the filament file.

Each row of the filament file corresponds to one estimated filamentary point on a particular spherical slice. The columns are organized as follows:

- `RAsc` and `DEC` give the sky position of the filamentary point in J2000 degrees.
- `z_low` and `z_high` specify the redshift interval of the slice, while `comov_dist_low` and `comov_dist_high` specify the corresponding comoving-distance shell.
- `bw` records the bandwidth used for directional KDE on that slice.
- `density` stores the proportional estimated density value at the filamentary point.
- `unc_meas` stores the bootstrap-based uncertainty measure.
- `grad_Dir1`, `grad_Dir2`, and `grad_Dir3` record the three Cartesian components of the Riemannian gradient of $\widehat{f}_h$ on Ω_2.
- `knot_label` indicates whether the point is classified as a knot, namely an intersection of multiple filamentary branches.

The data file for local modes uses analogous slice identifiers, together with the variables needed to describe the estimated modes.

This organization makes the catalog suitable for both visualization and downstream scientific analysis. Since the catalog is tomographic, users can subset the estimated structures by redshift interval or comoving-distance shell before cross-matching them with external galaxy samples. The pointwise density, gradient, and uncertainty fields further allow users

to distinguish prominent filamentary features from weakly supported ones, which is useful when the catalog is used as an input to subsequent astrophysical studies.

Chapter 5

HIGH-DIMENSIONAL LINEAR MODEL INFERENCE WITH MISSING OUTCOMES

After constructing the cosmic web catalog in Chapter 4, this chapter turns from geometric learning to statistical inference. The motivating scientific question is whether the stellar mass of a galaxy is significantly associated with its proximity to nearby cosmic filaments. Answering this question requires inference with high-dimensional features for observed galaxies and their incomplete stellar mass measurements. We develop a debiasing method for inference on the regression function of a sparse high-dimensional linear model when the outcome variable may be missing.

5.1 Introduction

As discussed in previous chapters, large-scale sky surveys provide rich spectroscopic and photometric features for galaxies and other astronomical objects, often together with value-added catalogs that report derived stellar properties of the observed galaxies, such as stellar mass (Blanton et al., 2005; Conroy et al., 2009; Chang et al., 2015). These associated features for each galaxy are typically high-dimensional and noisy, while the derived stellar mass may also be missing in the latest value-added catalog due to data contamination, target-selection limitations, or misclassification of galaxies as stars (Comparat et al., 2017). In the cosmic web application, this leads to a high-dimensional inference problem on a regression function (or conditional mean) with potentially missing outcomes: we wish to quantify the correlation between a galaxy's stellar mass and its distance to the nearby cosmic filaments while adjusting for many observed galaxy features. More generally, let $Y \in \mathbb{R}$ be an outcome variable and $X \in \mathbb{R}^d$ be a high-dimensional covariate vector. The object of interest is the regression

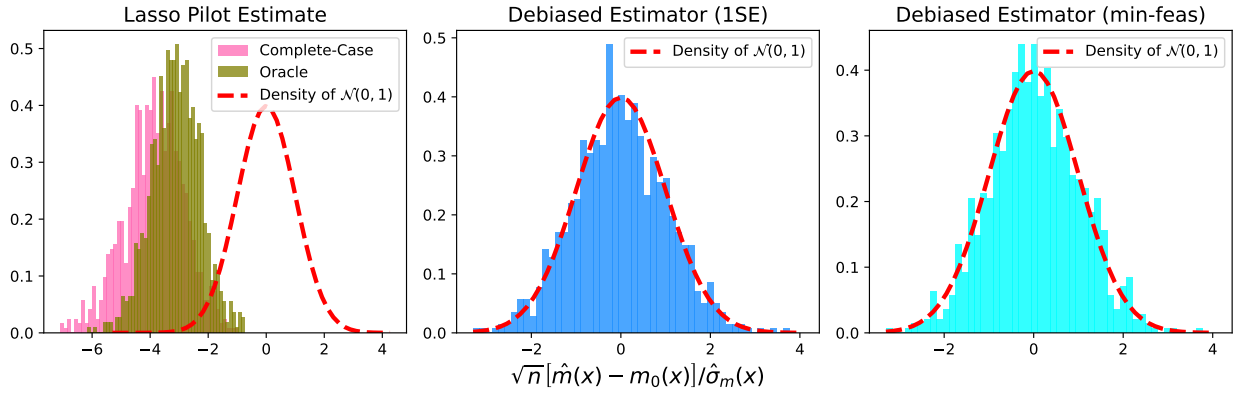


Figure 5.1: Comparison of our debiased estimators under two different choices of the tuning parameters ("1SE" and "min-feas") with the conventional Lasso estimates based on complete-case or oracle data. Here, we adopt the sparse β_0^{sp} and dense $x^{(4)}$ with $d = 1000, n = 900$, $X \sim \mathcal{N}_d(0, \Sigma^{\text{cs}})$, and $\epsilon \sim \mathcal{N}(0,1)$ under the MAR setting (5.23); see Section 5.4.1 for details.

function $m_0(x) = \mathbb{E}(Y|X = x)$.

When the outcome variables are fully observed in a data sample (which is known as the oracle data setting), statistical estimation of $m_0(x)$ is tractable with regularization and sparsity constraints under high-dimensional settings (Wainwright, 2019). One of the most well-studied approaches is Lasso (Tibshirani, 1996), which assumes the linear model and imposes L_1-regularization on the regression coefficients. When it comes to statistical inference, however, the Lasso solution leads to a biased estimate of $m_0(x)$ even when the linear model assumption is correct (van de Geer et al., 2014; Zhang and Zhang, 2014). Missing outcomes further exacerbate the bias of the Lasso solution; see the first panel of Figure 5.1 for an illustration. While this bias could be partially mitigated via sample-splitting or re-fitting, doing so would reduce the sample size and increase the computational cost, leading to less efficient estimates.

To tackle the challenges of high-dimensionality and missing outcomes, we impose the following structural assumptions throughout this chapter.

Assumption 5.1 (High-dimensional sparse linear regression model with MAR outcomes)**.**

(a) The data sample $\{(Y_i, R_i, X_i)\}_{i=1}^n$ *consists of independent and identically distributed (i.i.d.) observations drawn from* $(Y, R, X) \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ *generated by the linear model*

$$Y = X^T\beta_0 + \epsilon, \qquad \mathbb{E}\left(\epsilon \middle| X\right) = 0, \qquad \mathbb{E}\left(\epsilon^2 \middle| X\right) = \sigma_\epsilon^2, \tag{5.1}$$

where $||\beta_0||_0 = \sum_{k=1}^d \mathbb{1}_{\{\beta_{0k} \neq 0\}} = s_\beta \ll d$.

(b) The missingness indicator R *is conditionally independent of* Y *given* X.

Assumption 5.1(a) is standard in high-dimensional statistics. On the one hand, the sparse linear model can be extended to sparse additive models (Ravikumar et al., 2009) and partially linear models (Müller and van de Geer, 2015; Belloni et al., 2019). On the other hand, our proposed debiasing method can be generalized to handle heteroscedastic errors; see Section 5.2.1. However, developing theory for these generalizations is beyond the scope of this chapter. Assumption 5.1(b) is common in the missing data literature (Tsiatis, 2007; Little and Rubin, 2019) and is related to the ignorability or unconfoundedness condition in causal inference (Imbens, 2004). Under the MAR assumption, the propensity score $\pi(Y, X) := \mathrm{P}(R = 1|Y, X)$ depends only on the covariate vector X so that $\pi(Y, X) \equiv \pi(X) := \mathrm{P}(R = 1|X)$, which can thus be estimated from the fully observed data $\{(X_i, R_i)\}_{i=1}^n$ (Rosenbaum and Rubin, 1983).

Under Assumption 5.1, our proposed debiasing method directly infers the regression function $m_0(x) = x^T\beta_0$ by combining a Lasso pilot estimate of $\beta_0 \in \mathbb{R}^d$ based on the complete-case data with a bias correction term based on the weighted residuals of the Lasso regression. The weights depend on estimates of the propensity scores $\pi(X_i), i = 1, ..., n$ and are obtained as the solution to a convex debiasing program which trades off bias and variance in a mean-squared-error-optimal way; see Figure 5.1(b,c) for a preview.

A crucial difference between the existing work on debiasing the Lasso estimate in the literature (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014a,b) and our proposal is that we focus our inference on the scalar regression function $m_0(x) = x^T\beta_0$ rather than the high-dimensional regression (coefficient) vector $\beta_0 \in \mathbb{R}^d$. By

considering $m_0(x)$, we are able to conduct inference on any individual regression coefficient by setting x equal to a standard basis vector in $\mathbb{R}^d$ as well as on joint effects of several regression coefficients. Additionally, inferring $m_0(x)$ itself is more computationally efficient than inferring the regression vector β_0.

5.1.1 Contributions and Outline of the Chapter

1. Methodology: We describe the detailed procedures for our proposed debiasing method and reveal its computational feasibility through the dual formulation. We also provide interpretations of our debiasing method from the perspective of a bias-variance trade-off as well as Neyman near-orthogonality; see Section 5.2 for details.

2. Asymptotic Theory: We prove that our proposed debiased estimator is asymptotically unbiased and normally distributed as long as the propensity scores are pointwise consistently estimable (*i.e.*, sample-splitting or cross-fitting are not needed); see Section 5.3.4. Moreover, the asymptotic variance of our debiased estimator attains the semiparametric efficiency bound among all asymptotically linear estimators for any given dimension d. To establish the asymptotic normality of our debiased estimator, we derive the consistency of the Lasso pilot estimate with complete-case data and the dual solution of our debiasing program as building blocks; see Section 5.3.3.

3. Simulations and Real-World Applications: To demonstrate the finite-sample performance of our debiasing method, we compare it with the existing debiasing methods in the literature through extensive Monte Carlo experiments; see Section 5.4.1 and Section 5.4.2. We also apply the proposed debiasing method to the problem of inferring the stellar masses of galaxies in the Sloan Digital Sky Survey; see Section 5.5.

4. Algorithmic Implementation: We describe the implementation details of our debiasing method in Section D.1 and implement them in both the Python package "`Debias-Infer`" and the R package "`DebiasInfer`" (Zhang et al., 2023a). All the code for our experiments is available at `https://github.com/zhangyk8/Debias-Infer/tree/main/Paper_Code`.

The proofs of the results in this chapter are given in the corresponding paper (Zhang et al.,

2023b). To keep the dissertation focused on the main methodology, theory, and applications, we omit the proof details here.

5.2 Methodology

In this section, we outline our debiasing inference method and discuss a feasible solution to the key debiasing program through its dual form. Furthermore, we motivate our debiasing method from the perspectives of a bias-variance trade-off as well as Neyman near-orthogonality.

5.2.1 Debiasing Inference Procedure

Recall that we aim to conduct statistical inference on the regression function $m_0(x) = \mathbb{E}(Y|X = x) = x^T\beta_0$ given a random sample $\{(Y_i, R_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$, where Y_i is the outcome variable with missingness indicator R_i and X_i is the fully observed high-dimensional covariate vector for $i = 1, ..., n$. To this end, we propose the following debiasing procedure.

- **Step 1:** Compute the Lasso pilot estimate $\widehat{\beta} \in \mathbb{R}^d$ with complete-case data

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} R_i (Y_i - X_i^T \beta)^2 + \lambda \, ||\beta||_1 \right], \tag{5.2}$$

  where $\lambda > 0$ is a regularization parameter.

- **Step 2:** Obtain consistent estimates $\widehat{\pi}_i, i = 1, ..., n$ of the propensity scores $\pi_i = \pi(X_i) = \mathrm{P}(R_i = 1|X_i), i = 1, ..., n$ by any machine learning method (not necessarily a parametric model) on the data $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0,1\}$.

- **Step 3:** Solve for the debiasing weight vector $\widehat{\boldsymbol{w}} \equiv \widehat{\boldsymbol{w}}(x) \in \mathbb{R}^n$ through the following debiasing program with a tuning parameter $\gamma > 0$ as:

$$\min_{\boldsymbol{w} \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \widehat{\pi}_i w_i^2 : \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \cdot \widehat{\pi}_i \cdot X_i \right\|_\infty \leq \frac{\gamma}{n} \right\}. \tag{5.3}$$

- **Step 4:** Define the debiased estimator for $m_0(x)$ as:

$$\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) = x^T\widehat{\beta} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{w}_i(x)R_i\left(Y_i - X_i^T\widehat{\beta}\right). \tag{5.4}$$

- **Step 5:** Construct the asymptotic $(1-\tau)$-level confidence interval for $m_0(x)$ as:

$$\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) \pm \Phi^{-1}\left(1-\frac{\tau}{2}\right)\cdot\sigma_\epsilon\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{\pi}_i\widehat{w}_i(x)^2}\right], \tag{5.5}$$

  where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of $\mathcal{N}(0,1)$. If the noise level σ_ϵ^2 is unknown, then replace it by any consistent estimator $\widehat{\sigma}_\epsilon^2$.

The debiasing program in **Step 3** above is motivated by a conditional bias-variance trade-off; see Section 5.2.3. In Section 5.2.2, we derive the dual formulation of our debiasing program (5.3) in **Step 3**. The dual is an unconstrained convex program that facilitates both the theoretical analysis of the statistical properties and the practical implementation of our debiasing procedure. Finally, we prove in Section 5.3.4 that the debiased estimator $\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}})$ is asymptotically normal and that its asymptotic variance can be estimated consistently by $\widehat{\sigma}_\epsilon^2\sum_{i=1}^n\widehat{\pi}_i\widehat{w}_i^2(x)$. This guarantees the asymptotic validity of the confidence interval in **Step 5**.

Remark 5.1 (Debiasing program with heteroscedastic errors)**.** *Under heteroscedastic errors satisfying* $\mathbb{E}\left(\epsilon_i^2|X_i\right) = \sigma_{\epsilon,i}^2$*, our debiasing program* (5.3) *needs to be modified as follows:*

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}\widehat{\pi}_i\widehat{\sigma}_{\epsilon,i}^2 w_i^2 : \left\|x - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\cdot\widehat{\pi}_i\cdot X_i\right\|_\infty \le \frac{\gamma}{n}\right\}, \tag{5.6}$$

where $\widehat{\sigma}_{\epsilon,i}^2$ *is a proxy for* $\sigma_{\epsilon,i}^2$ *for* $i = 1,...,n$*. For the asymptotic analysis, it suffices to construct* $\widehat{\sigma}_{\epsilon,i}^2, i = 1,...,n$ *so that*

$$\sum_{i=1}^{n}\widehat{\pi}_i\left(\widehat{\sigma}_{\epsilon,i}^2 - \sigma_{\epsilon,i}^2\right)\widehat{w}_i^2 = o_P(1) \quad \text{or} \quad \frac{\sum_{i=1}^n\widehat{\pi}_i\left(\widehat{\sigma}_{\epsilon,i}^2 - \sigma_{\epsilon,i}^2\right)\widehat{w}_i^2}{\sum_{i=1}^n\widehat{\pi}_i\sigma_{\epsilon,i}^2\widehat{w}_i^2} = o_P(1).$$

One simple choice is the residual-based proxy $\widehat{\sigma}_{\epsilon,i}^2 = \left(Y_i - X_i^T\widehat{\beta}\right)^2$*, where* $\widehat{\beta}$ *is any consistent estimate of* β_0*, such as the Lasso pilot estimate from* **Step 1***, the refitted Lasso (Belloni*

and Chernozhukov, 2013), or the weighted adaptive Lasso (Wagener and Dette, 2013). Although this residual-based proxy may not be a consistent estimate of $\sigma^2_{\epsilon,i}$ for each $i = 1, ..., n$, it provides a practical plug-in approximation for the unknown error scale. More in-depth theoretical derivations are left for future work.

5.2.2 Dual Formulation of the Debiasing Program (5.3)

By standard results in convex optimization (Boyd and Vandenberghe, 2004), strong duality refers to the situation in which the primal program and its dual attain the same optimal objective value. Under this notion, we derive the following result.

Proposition 5.1. *The dual form of the debiasing program (5.3) is given by*

$$\min_{\ell \in \mathbb{R}^d} \left\{ \frac{1}{4n} \sum_{i=1}^{n} \widehat{\pi}_i \left[X_i^T \ell \right]^2 + x^T \ell + \frac{\gamma}{n} ||\ell||_1 \right\}. \tag{5.7}$$

If strong duality holds, then the solution $\widehat{\boldsymbol{w}}(x) \in \mathbb{R}^n$ to the primal debiasing program (5.3) and the solution $\widehat{\ell}(x) \in \mathbb{R}^d$ to the dual debiasing program (5.7) satisfy the following identity:

$$\widehat{w}_i(x) = -\frac{1}{2\sqrt{n}} \cdot X_i^T \widehat{\ell}(x), \quad i = 1, ..., n. \tag{5.8}$$

This result has two important implications. First, we can select the tuning parameter $\gamma > 0$ via cross-validation on the loss function of the dual program. Indeed, since the primal debiasing program (5.3) is a constrained quadratic program, it is difficult to select $\gamma > 0$ based on the primal formulation alone: if $\gamma > 0$ is too small, the program will be infeasible; if $\gamma > 0$ is too large, the estimate will have a non-negligible bias (see Section 5.2.3). Second, if strong duality holds, the debiased estimator (5.4) can be expressed via (5.8) in terms of the dual solution $\widehat{\ell}(x)$ as:

$$\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}) = x^T \widehat{\beta} - \frac{1}{2n} \sum_{i=1}^{n} R_i X_i^T \widehat{\ell}(x) \left(Y_i - X_i^T \widehat{\beta} \right), \tag{5.9}$$

where the second bias-correction term reduces to a scaled sample average of n random variables. This representation is the key to deriving the asymptotic normality of our debiased estimator; see Section 5.3.1 and Section 5.3.4 for details. Notice that strong duality holds whenever $\gamma > 0$ is sufficiently large; see Lemma 5.6 in Section 5.3.3 for more details.

5.2.3 Bias-Variance Trade-off Perspective

The design of our debiasing program (5.3) is motivated by controlling the conditional mean squared error $\mathbb{E}\left[\left(\sqrt{n}\cdot\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}})-\sqrt{n}\cdot m_0(x)\right)^2\Big|\boldsymbol{X}\right]$ of our debiased estimator (5.4) and can thus be interpreted as solving a (conditional) bias-variance trade-off. To explicate this motivation, we consider the generic debiased estimator $m^{\text{debias}}(x;\boldsymbol{w})$ from (5.4) with an arbitrary weight vector $\boldsymbol{w}=(w_1,...,w_n)^T\in\mathbb{R}^n$ as:

$$m^{\text{debias}}(x;\boldsymbol{w})=x^T\beta+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_iR_i\left(Y_i-X_i^T\beta\right), \tag{5.10}$$

where one would plug in the Lasso pilot estimate $\widehat{\beta}$ from (5.2) and the solution $\widehat{\boldsymbol{w}}\equiv\widehat{\boldsymbol{w}}(x)\in\mathbb{R}^n$ from our debiasing program (5.3) to obtain a concrete debiased estimator (5.4) in practice. We decompose the conditional mean squared error of (5.10) into three explicit terms as follows.

Proposition 5.2. *Under Assumption 5.1, the conditional mean squared error of* $\sqrt{n}\,m^{\text{debias}}(x;\boldsymbol{w})$ *given* $\boldsymbol{X}=(X_1,...,X_n)^T$ *has the following decomposition as:*

$$\begin{aligned}&\mathbb{E}\left[\left(\sqrt{n}\,m^{\text{debias}}(x;\boldsymbol{w})-\sqrt{n}\,m_0(x)\right)^2\Big|\boldsymbol{X}\right]\\&=\underbrace{\sigma_\epsilon^2\sum_{i=1}^{n}w_i^2\pi(X_i)}_{\textit{Conditional variance I}}+\underbrace{(\beta_0-\beta)^T\left[\sum_{i=1}^{n}w_i^2\pi(X_i)\left(1-\pi(X_i)\right)X_iX_i^T\right](\beta_0-\beta)}_{\textit{Conditional variance II}}\\&\quad+\underbrace{\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\pi(X_i)X_i-x\right)^T\sqrt{n}\,(\beta_0-\beta)\right]^2}_{\textit{Conditional bias}}.\end{aligned} \tag{5.11}$$

By Hölder's inequality, we can upper bound the "Conditional bias" term as follows:

$$\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\pi(X_i)X_i-x\right)^T\sqrt{n}\,(\beta_0-\beta)\right]^2\le\left[\left|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\pi(X_i)X_i-x\right|\right|_\infty\sqrt{n}\,||\beta_0-\beta||_1\right]^2.$$

The debiasing program (5.3) is hence designed to optimize the weights $w_i, i=1,...,n$ so that the estimated "Conditional variance I" term $\sum_{i=1}^{n}\widehat{\pi}_iw_i^2$ is minimized as the objective function

while an upper bound on the estimated "Conditional bias" term $\left|\left|x - \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \widehat{\pi}_i X_i\right|\right|_\infty$ is well-controlled by the constraint. We ignore the "Conditional variance II" term because, by Lemma 5.7, it is asymptotically negligible and dominated by the "Conditional variance I" term if $\beta = \widehat{\beta}$ (the solution to the Lasso program (5.2)) and $\boldsymbol{w} = \widehat{\boldsymbol{w}}(x)$ (the solution to the debiasing program (5.3)). Similar bias-variance trade-offs also appear in Cai et al. (2023) for generalized linear models and Giessing and Wang (2023) for quantile regression models. However, the motivation for the procedures in these two papers is ad hoc and not grounded in a rigorous decomposition of the conditional mean squared error. Upon minimizing the (conditional) variance, we expect our debiased estimator to be asymptotically more efficient than other estimators; see our discussion in Section 5.3.4.

5.2.4 Neyman Near-Orthogonalization Viewpoint

Our debiasing method in Section 5.2.1 can also be motivated from the perspective of Neyman near-orthogonalization (Neyman, 1959, 1979; Chernozhukov et al., 2018). To this end, we regard the regression function $m \equiv m(x) \in \mathbb{R}$ as the parameter of interest and the regression vector $\beta \in \mathbb{R}^d$ as a high-dimensional nuisance parameter. Under Assumption 5.1, the true values of these two parameters are $m_0 = m_0(x) = x^T \beta_0$ and β_0, respectively. We define the generic score function by

$$\Xi_x(Y, R, X; m, \beta) = m - x^T \beta - \sqrt{n} \cdot w \cdot R(Y - X^T \beta),$$

where $w \in \mathbb{R}$ is a symbolic weight. Given the observed data $\{(Y_i, R_i, X_i)\}_{i=1}^n$, arbitrary $\boldsymbol{w} \in \mathbb{R}^n$, and arbitrary $\beta \in \mathbb{R}^d$, the generic debiased estimator $m^{\text{debias}}(x; \boldsymbol{w})$ solves the sample-based estimating equation

$$\frac{1}{n} \sum_{i=1}^n \Xi_x(Y_i, R_i, X_i; m^{\text{debias}}, \beta) = m^{\text{debias}}(x; \boldsymbol{w}) - x^T \beta - \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \cdot R_i \left(Y_i - X_i^T \beta\right) = 0. \quad (5.12)$$

Provided that Assumption 5.1 holds and $\boldsymbol{w} \in \mathbb{R}^n$ satisfies the box constraint in (5.3), the Neyman near-orthogonalization condition given $\boldsymbol{X} = (X_1, ..., X_n)^T \in \mathbb{R}^{n \times d}$ at (m_0, β_0) re-

quires that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Xi_x(Y_i,R_i,X_i;m_0,\beta_0)\Big|\boldsymbol{X}\right]=0,\\
&\sup_{\beta\in\mathcal{T}_n}\left|\left\{\frac{\partial}{\partial\beta}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Xi_x(Y_i,R_i,X_i;m,\beta)\Big|\boldsymbol{X}\right]\Big|_{(m_0,\beta_0)}\right\}^T(\beta-\beta_0)\right|\le\frac{\delta_n}{\sqrt{n}},
\end{aligned}\tag{5.13}$$

where $\mathcal{T}_n$ is a properly shrinking neighborhood of β_0 and $\delta_n = o(1)$; see Definition 2.2 in Chernozhukov et al. (2018). Indeed, the first condition obviously holds true, while the second condition is also satisfied because for any $\beta\in\mathcal{T}_n$,

$$\begin{aligned}
&\left|\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\mathbb{E}\left[\Xi_x(Y_i,R_i,X_i;m,\beta)|\boldsymbol{X}\right]\big|_{(m_0,\beta_0)}\right\}^T(\beta-\beta_0)\right|\\
&=\left|\left[x-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\cdot\pi(X_i)X_i\right]^T(\beta_0-\beta)\right|\\
&\overset{\text{(a)}}{\text{“}\le\text{”}}\left|\left|x-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i\cdot\widehat{\pi}_i\cdot X_i\right|\right|_\infty||\beta-\beta_0||_1\\
&\overset{\text{(b)}}{\le}\frac{\gamma}{n}||\beta-\beta_0||_1\le\frac{\delta_n}{\sqrt{n}},
\end{aligned}$$

where the approximate inequality (a) follows from Hölder's inequality and substituting the estimated propensity scores $\widehat{\pi}_i, i = 1,...,n$ while (b) holds since $\boldsymbol{w}\in\mathbb{R}^n$ satisfies the box constraint in (5.3). Therefore, the nuisance realization set $\mathcal{T}_n$ can be chosen as $\left\{\beta\in\mathcal{B}\subset\mathbb{R}^d : ||\beta-\beta_0||_1\le\frac{\sqrt{n}\delta_n}{\gamma}\right\}$ for some convex set $\mathcal{B}$ containing β_0. We show in Theorem 5.4 that the Lasso pilot estimate $\widehat{\beta}$ satisfies $\left|\left|\widehat{\beta}-\beta_0\right|\right|_1 = O_P\left(s_\beta\sqrt{\frac{\log d}{n}}\right)$ and thus, our final debiased estimator (5.4) satisfies the Neyman near-orthogonality condition (5.13) with a fine-tuned parameter $\gamma>0$.

By Theorem 3.1 in Chernozhukov et al. (2018), the (asymptotic) variance of $m^{\text{debias}}(x;\boldsymbol{w})$ is $\sigma_\epsilon^2\,\mathbb{E}\left[\sum_{i=1}^n w_i^2R_i^2|\boldsymbol{X}\right] = \sigma_\epsilon^2\sum_{i=1}^n w_i^2\pi(X_i)$. Therefore, our debiasing program (5.3) is designed to yield the debiased estimator with the smallest (estimated) variance among all the estimators that (approximately) satisfy the constraints in (5.13). Consequently, we can expect that our debiased estimator is more efficient than most other estimators even without

explicitly constructing an efficient score function satisfying the exact Neyman orthogonality; see Section 2.2.2 in Chernozhukov et al. (2018) with references therein. Finally, in light of (5.13), our debiasing method de-correlates the Lasso pilot regression from the propensity score estimation and the optimization of weights for the debiased estimator so that we are able to establish the asymptotic normality of our debiased estimator without resorting to sample-splitting.

5.3 Asymptotic Theory

In this section, we study the consistency and asymptotic normality of our debiasing inference method in Section 5.2.1. The high-level motivation and an overview of the technical results are given in Section 5.3.1, with more detailed analyses in the subsequent subsections.

5.3.1 Heuristics and Overview

In order to derive the asymptotic normality of our debiased estimator (5.4), we need to analyze the asymptotic behavior of $\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right]$. Under Assumption 5.1 and the definition (5.4) of $\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}})$, we deduce that

$$\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] = \sum_{i=1}^{n} \widehat{w}_i(x) R_i \epsilon_i + \left[x - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widehat{w}_i(x) R_i X_i\right]^T \sqrt{n}\left(\widehat{\beta} - \beta_0\right),$$

where $\epsilon_i, i = 1, ..., n$ are i.i.d. noise variables under model (5.1). From the above expression, it is unclear how one can apply the central limit theorem (CLT) or related asymptotic theories to derive the asymptotic distribution of $\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right]$, because the limiting distribution of the weight vector $\widehat{\boldsymbol{w}} \equiv \widehat{\boldsymbol{w}}(x) = (\widehat{w}_1(x), ..., \widehat{w}_n(x))^T \in \mathbb{R}^n$ is intractable and may not even exist. However, the dual formulation in Proposition 5.1 provides a promising approach to this problem. By the dual representation (5.9) of the debiased estimator, we have that

$$\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] \tag{5.14}$$

$$= -\frac{1}{2\sqrt{n}} \sum_{i=1}^{n} R_i \epsilon_i X_i^T \widehat{\ell}(x) + \left[x + \frac{1}{2n} \sum_{i=1}^{n} R_i X_i^T \widehat{\ell}(x) X_i \right]^T \sqrt{n} \left(\beta_0 - \widehat{\beta} \right)$$

$$= -\frac{1}{2\sqrt{n}} \sum_{i=1}^{n} R_i \epsilon_i X_i^T \ell_0(x) + \text{“Bias terms”}, \tag{5.15}$$

where $\ell_0(x) \in \mathbb{R}^d$ is the deterministic solution to the population dual program for which the dual solution $\widehat{\ell}(x) \in \mathbb{R}^d$ is consistent; see (5.18) in Section 5.3.2 for its definition. The leading term in (5.14) is a sample average of n i.i.d. scalar random variables, which can be handled via the standard CLT, while the “Bias terms” in (5.14) are of order $o_P(1)$ under mild regularity conditions; see Theorem 5.8.

The rest of this section corroborates the above heuristic arguments. After introducing the notation and necessary assumptions, we derive the following main results:

- **Consistency of the Lasso pilot estimate** (5.2)**:** Under a suitable choice of $\lambda > 0$, it holds that $\left\| \widehat{\beta} - \beta_0 \right\|_2 = O_P \left(\frac{1}{\kappa_R^2} \sqrt{\frac{s_\beta \log d}{n}} \right)$ when $\log d = o(n)$, where $\kappa_R^2 > 0$ is the minimum eigenvalue of the complete-case Gram matrix $\mathbb{E}\left(RXX^T\right)$; see Theorem 5.4 for details.

- **Consistency of the solution to the dual program** (5.7)**:** Under a suitable choice of $\gamma > 0$, we deduce that $\left\| \widehat{\ell}(x) - \ell_0(x) \right\|_2 = O_P \left(\frac{1}{\kappa_R^3} \sqrt{\frac{s_\ell(x) \log d}{n}} + \left(\frac{1+\kappa_R^2}{\kappa_R^4} \right) r_\ell + \frac{r_\pi \sqrt{s_\ell(x)}}{\kappa_R^4} \right)$, where $\ell_0(x)$ is the solution to the population dual program, $r_\ell > 0$ is a parameter for the sparse approximation to $\ell_0(x)$, and $r_\pi > 0$ is the rate of convergence of $\max_{1 \le i \le n} |\widehat{\pi}_i - \pi(X_i)|$; see Theorem 5.5 and its related discussions.

- **Asymptotic normality of the debiased estimator** (5.4)**:** Combining the above consistency results with other mild regularity conditions, we prove that

$$\sqrt{n} \left[\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}) - m_0(x) \right] \xrightarrow{d} \mathcal{N}\left(0,\, \sigma_m^2(x)\right)$$

  and discuss the semiparametric efficiency of $\sigma_m^2(x) = \lim_{n \to \infty} \sigma_\epsilon^2 \cdot x^T \left[\mathbb{E}\left(RXX^T\right) \right]^{-1} x$; see Theorem 5.8 for details.

5.3.2 Notations and Assumptions

We enumerate the regularity conditions under the high-dimensional sparse linear model in Assumption 5.1 for our subsequent theoretical analysis.

Assumption 5.2 (Sub-Gaussian covariate)**.** *The covariate vector $X \in \mathbb{R}^d$ is sub-Gaussian,* i.e., $\left|\left|u^T X\right|\right|_{\psi_2} \lesssim \left[\mathbb{E}\left(u^T X X^T u\right)\right]^{1/2}$ *for any* $u \in \mathbb{R}^d$.

We introduce the sub-Gaussian condition as Assumption 5.2 on the covariate vector X in order to control the exponential tail behavior of X by its second moment and facilitate our proofs. Notice that the covariate vector $X \in \mathbb{R}^d$ is not necessarily centered in our definition of a sub-Gaussian covariate vector. We denote the s-sparse maximum eigenvalue of the population Gram matrix $\mathbb{E}\left(XX^T\right) \in \mathbb{R}^{d\times d}$ by

$$\phi_s := \sup_{u\in\Omega_{d-1}, ||u||_0\leq s} \mathbb{E}\left[(X^T u)^2\right]. \tag{5.16}$$

In particular, $\mathbb{E}(X_{ik}^2) \leq \max_{1\leq i\leq n}\max_{1\leq k\leq d}\mathbb{E}(X_{ik}^2) \leq \phi_1$ for each covariate vector $X_i = (X_{i1}, ..., X_{id})^T \in \mathbb{R}^d$. It is noteworthy that under the high-dimensional setting with $s \ll d$, the s-sparse maximum eigenvalue ϕ_s is substantially smaller than the unrestricted maximum eigenvalue ϕ_d of $\mathbb{E}\left(XX^T\right)$ and can even be independent of the dimension d.

Assumption 5.3 (Sub-Gaussian noise)**.** *The noise variable $\epsilon \in \mathbb{R}$ is sub-Gaussian,* i.e., $||\epsilon||_{\psi_2} \lesssim [\mathbb{E}(\epsilon^2)]^{1/2} = \sigma_\epsilon$ *with $\sigma_\epsilon^2 > 0$ being the noise level.*

We introduce Assumption 5.3 to simplify the theoretical analysis. In principle, this assumption can be relaxed to a lower-order moment condition; see Belloni and Chernozhukov (2013). We conduct numerical experiments with our debiasing method in which the sub-Gaussian noise assumption is violated in Section 5.4.3 and demonstrate that our proposed method is robust to various heavy-tailed noise distributions. To establish the consistency of the Lasso pilot estimate (5.2) and the dual solution $\widehat{\ell}(x) \in \mathbb{R}^d$ to (5.7), we impose the following eigenvalue lower bound on the complete-case population Gram matrix $\mathbb{E}\left(RXX^T\right) \in \mathbb{R}^{d\times d}$.

Assumption 5.4 (Lower eigenvalue bound on $\mathbb{E}\left(RXX^T\right)$)**.** *The minimum eigenvalue of the complete-case Gram matrix* $\mathbb{E}\left(RXX^T\right) \in \mathbb{R}^{d\times d}$ *is bounded away from zero,* i.e.*, there exists a constant* $\kappa_R > 0$ *such that*

$$\inf_{v\in\Omega_{d-1}} \mathbb{E}\left[R(X^Tv)^2\right] \geq \kappa_R^2.$$

Assumption 5.4 ensures that the regression vector $\beta_0 \in \mathbb{R}^d$ is identifiable. We explicitly introduce the "constant" $\kappa_R > 0$ to allow for situations in which the minimum eigenvalue of $\mathbb{E}\left(RXX^T\right)$ may decrease (slowly) as $d, n \to \infty$. In particular, κ_R serves as a lower bound tracking the smallest eigenvalue of $\mathbb{E}\left(RXX^T\right)$ and will appear explicitly in the rates derived in the subsequent analysis.

When proving the consistency of the Lasso pilot estimate (5.2), however, we allow the common relaxation of Assumption 5.4 that the minimum eigenvalue of $\mathbb{E}\left(RXX^T\right)$ is only bounded away from zero over the cone $\mathcal{C}_1(S_\beta, 3)$:

$$\inf_{v\in\mathcal{C}_1(S_\beta,3)\cap\Omega_{d-1}} \mathbb{E}\left[R(X^Tv)^2\right] \geq \kappa_R^2, \tag{5.17}$$

where $S_\beta \subset \{1, ..., d\}$ is an index set for nonzero entries of $\beta_0 \in \mathbb{R}^d$ while $\mathcal{C}_q(S, \vartheta) = \left\{v \in \mathbb{R}^d : ||v_{S^c}||_q \leq \vartheta\, ||v_S||_q\right\}$ is a cone of dominant coordinates (Section 7.2 in Koltchinskii 2011) for some index set $S \subset \{1, ..., d\}$ and the parameter $\vartheta > 0$ with respect to the norm $||\cdot||_q$ for $q \geq 1$. Here, $v_S \in \mathbb{R}^{|S|}$ denotes the subvector with elements of v indexed by S and v_{S^c} is defined in an analogous manner. This relaxed assumption (5.17) is known as the restricted eigenvalue condition in the literature; see Section 7.3.1 in Wainwright (2019) and also Bickel et al. (2009); van de Geer and Bühlmann (2009). We can show that under Assumption 5.2, the sample-based complete-case Gram matrix $\frac{1}{n}\sum_{i=1}^n R_iX_iX_i^T$ is close to the population one with high probability, so (5.17) holds interchangeably for these two Gram matrices in the setting of this chapter.

As for propensity score estimation, we introduce a mild assumption on the consistency of the estimated propensity scores $\widehat{\pi}_i$ with the true propensity scores $\pi_i = \pi(X_i)$ for $i = 1, ..., n$.

Assumption 5.5 (Stochastic control of the propensity score estimation)**.** *Given any* $n \geq 1$ *and* $\delta \in (0, 1)$*, there exists* $r_\pi \equiv r_\pi(n, \delta) > 0$ *such that* $\mathrm{P}\left(\max_{1\leq i\leq n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta$.

Assumption 5.5 is a non-asymptotic formulation of the stochastic boundedness of the maximum estimation errors of the propensity scores evaluated at the in-sample covariate vectors $X_i, i = 1, ..., n$. In particular, it follows from Assumption 3.2 in Chakrabortty et al. (2019) by a union bound under mild rate conditions. It can also be replaced by the in-sample ℓ_1-prediction consistency condition $\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}|\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta$ without affecting the subsequent results. In applications, the exact rate of consistency r_π depends on how the propensity scores are estimated, whether by parametric, nonparametric, or semiparametric methods. We derive an explicit rate formula for r_π under the Lasso-type generalized linear regression estimator (D.3) in Proposition D.2 of Section D.3 and discuss the corresponding rates for other semiparametric or nonparametric methods in Section D.3.1. We demonstrate in Theorem 5.5 and Theorem 5.8 below that as long as $r_\pi(n, \delta) \to 0$ for some $\delta \to 0$ as $n \to \infty$, the solution to the dual program (5.7) is consistent and consequently, the debiased estimator (5.4) is asymptotically normal; see Section 5.3.3 and Section 5.3.4 for details. The specific rate at which $r_\pi \to 0$ is essentially irrelevant, as shown in (5.21). These results unveil the possibility of combining our debiased estimator with consistent estimates of the propensity scores obtained from any machine learning approach; see Section 5.4.4 for numerical experiments.

Remark 5.2. *In contrast to the existing literature on regression models with missing outcomes, incomplete covariates, or potential outcomes (Rosenbaum and Rubin, 1983; Robins et al., 1994; Chakrabortty et al., 2019), we do not require the positivity condition* $\pi(x) = \mathrm{P}(R = 1|X = x) > \pi_{\min} > 0$ *for all* $x \in \mathrm{support}(X) \subset \mathbb{R}^d$. *As noted by D'Amour et al. (2021), the positivity condition is too stringent in the context of high-dimensional covariates. Instead, we merely assume that the complete-case population Gram matrix* $\mathbb{E}\left(RXX^T\right)$ *is positive definite as in Assumption 5.4, which is valid even when the positivity condition is violated. For example,* $\mathbb{E}\left(RXX^T\right)$ *will be positive definite if* $\mathrm{support}(X)$ *has a nonzero Lebesgue measure in* $\mathbb{R}^d$ *and* $\pi(x)$ *is positive within a subset of* $\mathrm{support}(X)$ *with a nonzero Lebesgue measure.*

The population dual program for any $x \in \mathbb{R}^d$ is defined by taking $n \to \infty$ in (5.7) under the true propensity scores as:

$$\min_{\ell \in \mathbb{R}^d} \left\{ \frac{1}{4} \mathbb{E}\left[R \left(X^T \ell \right)^2 \right] + x^T \ell \right\}. \tag{5.18}$$

Under Assumption 5.4, the unique solution to the above population dual program (5.18) is given by

$$\ell_0(x) := -2 \left[\mathbb{E}\left(R X X^T \right) \right]^{-1} x \in \mathbb{R}^d. \tag{5.19}$$

The population dual solution $\ell_0(x)$ already appears as the "orthogonalization parameter" in the debiased machine learning literature (e.g. Chernozhukov et al., 2018, Example 2.1) and (up to a scaling factor) as the "least favorable sub-model" in Bellec and Zhang (2022). However, in both cases, $\ell_0(x)$ does not arise as the solution to a dual program whose corresponding primal program solves a bias-variance trade-off; see Section 5.2.3. We assume that $\ell_0(x)$ is approximately sparse in the following sense; see also Assumption 5.6.

Definition 5.3 (r_ℓ-approximation to $\ell_0(x)$)**.** *For any $x \in \mathbb{R}^d$, we define an r_ℓ-approximation $\widetilde{\ell}(x)$ to the population dual solution $\ell_0(x)$ as:*

$$\widetilde{\ell}(x) \equiv \widetilde{\ell}(x; r_\ell) := \underset{\ell(x) \in \mathbb{R}^d}{\arg\min} \left\{ ||\ell(x)||_0 : ||\ell(x) - \ell_0(x)||_2 \leq r_\ell \, ||\ell_0(x)||_2 \right\}, \tag{5.20}$$

where $r_\ell \in \left[0, \frac{1}{2}\right]$ controls the accuracy of approximation.

Remark 5.3. *Under the stronger Assumption 5.4, $\mathbb{E}(RXX^T)$ also satisfies the $\mathcal{C}_1(S_\ell(x), 3)$-restricted eigenvalue condition with some parameter $\kappa_R^2 > 0$ defined in (5.17), where $S_\ell(x) :=$* support $\left(\widetilde{\ell}(x) \right) = \left\{ 1 \leq k \leq d : \widetilde{\ell}_k(x) \neq 0 \right\}$.

A larger value of r_ℓ gives rise to a sparser approximation $\widetilde{\ell}(x)$ to the vector $\ell_0(x)$ of interest. Furthermore, the consistency of $\widehat{\ell}(x)$ requires the approximation parameter r_ℓ for $\widetilde{\ell}(x)$ to converge to 0 at a certain rate as $n \to \infty$; see Assumption 5.6 below and the discussion after Theorem 5.5 in Section 5.3.3. In other words, as the sample size n increases, the unique population dual solution $\ell_0(x)$ must tend to a sparse vector in order for the consistency result to hold.

Assumption 5.6 (Sparsity of $\widetilde{\ell}(x)$)**.** *The r_ℓ-approximation $\widetilde{\ell}(x) \in \mathbb{R}^d$ to the population dual solution $\ell_0(x) \in \mathbb{R}^d$ exists and satisfies $s_\ell(x) := \left|\left|\widetilde{\ell}(x)\right|\right|_0 \ll \min\{n, d\}$.*

The existence of $\widetilde{\ell}(x)$ and the sparsity quantity $s_\ell(x)$ in Assumption 5.6 reflect the effective sparsity of the population dual solution $\ell_0(x) = -2\left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x$. Intuitively, $s_\ell(x)$ is small when the query point $x \in \mathbb{R}^d$ interacts effectively with a limited subset of the covariate vector $X \in \mathbb{R}^d$ after accounting for both the missingness pattern in $Y \in \mathbb{R}$ and the dependence structure among covariates in X. Such a situation arises, for example, when the inverse complete-case Gram matrix $\left[\mathbb{E}\left(RXX^T\right)\right]^{-1}$ is (approximately) sparse and the query point x itself is concentrated on only a few coordinates. These sparsity structures on $\left[\mathbb{E}\left(RXX^T\right)\right]^{-1}$ occur naturally in many high-dimensional statistical models, including graphical Lasso models (Friedman et al., 2008), autoregressive or short-range dependent processes, and settings where the propensity score $\mathrm{P}(R = 1|X)$ depends only on a small subset of covariates in X. Section D.2 furnishes several sufficient conditions and motivating examples under which Assumption 5.6 holds.

5.3.3 Consistency Results

In this section, we present the consistency results for our Lasso pilot estimate (5.2) and the solution to the dual program (5.7). Additionally, we discuss the strong duality theory of our debiasing inference method and validate the asymptotic negligibility of the "Conditional variance II" term in Proposition 5.2 of Section 5.2.3.

Theorem 5.4 (Consistency of the Lasso pilot estimate with complete-case data)**.** *Let $\delta \in (0, 1)$ and $A_1, A_2 > 0$ be some absolute constants. Suppose that Assumptions 5.1, 5.2, and 5.4 (more precisely, Eq.(5.17)) hold and n is sufficiently large so that*

$$\sqrt{\frac{s_\beta \log(ed/s_\beta)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} < \min\left\{\frac{\kappa_R^2}{A_1 \phi_{s_\beta}}, 1\right\}.$$

(a) If $\lambda > \frac{4}{n}\left|\left|\sum_{i=1}^n R_i X_i \epsilon_i\right|\right|_\infty > 0$ and $d \geq s_\beta + 2$, then with probability at least $1 - \delta$,

$$\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 \lesssim \frac{\sqrt{s_\beta}\lambda}{\kappa_R^2}.$$

(b) If, in addition, Assumption 5.3 holds and $\lambda = A_2\sigma_\epsilon\sqrt{\frac{\phi_1 \log(d/\delta)}{n}}$*, then with probability at least* $1-\delta$*,*

$$\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 \lesssim \frac{\sigma_\epsilon}{\kappa_R^2}\sqrt{\frac{\phi_1 s_\beta \log(d/\delta)}{n}}.$$

Remark 5.4. *If* $\log d = o(n)$ *with* $d > n$ *and* $\sigma_\epsilon, \phi_1 > 0$ *are independent of* n *and* d*, then we obtain the asymptotic consistency result* $\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^2}\sqrt{\frac{s_\beta \log d}{n}}\right)$ *by setting* $\delta = \frac{1}{d}$*.*

Theorem 5.4 implies that the Lasso pilot estimate $\widehat{\beta}$ in (5.2) based on the complete-case data is consistent. This property is a consequence of our MAR assumption and the lower eigenvalue bound condition on $\mathbb{E}(RXX^T)$ (Assumption 5.4), because the conditional distribution of Y given X remains unchanged on the complete-case data ($R = 1$) and the loss function in (5.2) is "strictly convex" on the cone in which $\widehat{\beta}$ lies.

Next, we present the consistency result for the dual solution $\widehat{\ell}(x)$ to the sparse r_ℓ-approximation $\widetilde{\ell}(x)$. Among other things, the rate of consistency for $\widehat{\ell}(x)$ depends on the rate of consistency $r_\pi = r_\pi(n, \delta)$ for estimating propensity scores.

Theorem 5.5 (Consistency of the dual solution $\widehat{\ell}(x)$)**.** *Let* $\delta \in (0,1)$*,* $x \in \mathbb{R}^d$ *be a fixed covariate vector, and* $A_1, A_2, A_3 > 0$ *be some large absolute constants. Suppose that Assumptions 5.1, 5.2, 5.4, 5.5, and 5.6 hold as well as*

$$\sqrt{\frac{s_\ell(x)\log(ed/s_\ell(x))}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + r_\pi < \min\left\{\frac{\kappa_R^2}{A_1\phi_{s_\ell(x)}}, 1\right\} \quad \text{and} \quad d \geq s_\ell(x) + 2.$$

We denote $r_\gamma \equiv r_\gamma(n,\delta) = \frac{\sqrt{\phi_1}}{\kappa_R}||x||_2\sqrt{\frac{\log(d/\delta)}{n}} + \frac{\sqrt{\phi_1\phi_{s_\ell(x)}}||x||_2}{\kappa_R^2}\cdot r_\pi$*. If* $\frac{\gamma}{n} = A_2 \cdot r_\gamma$ *and*

$$r_\ell \leq \min\left\{\frac{1}{2}, \left[\left(\frac{A_2^2\kappa_R^6\phi_1}{A_3^2||x||_2\phi_{s_\ell(x)}}\right)^{\frac{1}{4}}\left(\frac{\log(d/\delta)}{n}\right)^{\frac{1}{4}} + \left(\frac{A_2^2\phi_1\kappa_R^4}{A_3^2\phi_{s_\ell(x)}||x||_2^2}\right)^{\frac{1}{4}}\sqrt{r_\pi}\right]\right\},$$

then with probability at least $1-\delta$*,*

$$\left|\left|\widehat{\ell}(x) - \widetilde{\ell}(x)\right|\right|_2 \lesssim \frac{||x||_2}{\kappa_R^3}\left[\sqrt{\frac{\phi_1 s_\ell(x)\log(d/\delta)}{n}} + \frac{\phi_{s_\ell(x)}}{\kappa_R}\cdot r_\ell + \frac{\sqrt{\phi_1\phi_{s_\ell(x)}s_\ell(x)}}{\kappa_R}\cdot r_\pi\right].$$

Remark 5.5. *If* $\log d = o(n)$, $||x||_2 = O(1)$, *and* $\phi_1, \phi_{s_\ell(x)}$ *are independent of* n *and* d, *then we obtain the asymptotic consistency result*

$$\left|\left|\widehat{\ell}(x) - \widetilde{\ell}(x)\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} + \frac{r_\ell}{\kappa_R^4} + \frac{r_\pi\sqrt{s_\ell(x)}}{\kappa_R^4}\right)$$

by setting $\delta = \frac{1}{d}$. *This asymptotic rate of convergence comprises two parts. The first part is of the order* $O_P\left(\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} + \frac{r_\ell}{\kappa_R^4}\right)$, *which is the oracle rate of convergence for* $\left|\left|\widehat{\ell}(x) - \widetilde{\ell}(x)\right|\right|_2$ *when the true propensity scores are known. The second part is of the order* $O_P\left(\frac{r_\pi\sqrt{s_\ell(x)}}{\kappa_R^4}\right)$, *which is related to the consistency of the propensity score estimation.*

By the triangle inequality, this result implies that

$$\begin{aligned}\left|\left|\widehat{\ell}(x) - \ell_0(x)\right|\right|_2 &\leq \left|\left|\widehat{\ell}(x) - \widetilde{\ell}(x)\right|\right|_2 + \left|\left|\widetilde{\ell}(x) - \ell_0(x)\right|\right|_2 \\ &= O_P\left(\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} + \left(\frac{1+\kappa_R^2}{\kappa_R^4}\right) r_\ell + \frac{r_\pi\sqrt{s_\ell(x)}}{\kappa_R^4}\right).\end{aligned}$$

Thus, the dual solution $\widehat{\ell}(x)$ is consistent for the population dual solution $\ell_0(x)$ under the setup in Remark 5.5 if $\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} \to 0$, $\left(\frac{1+\kappa_R^2}{\kappa_R^4}\right) r_\ell \to 0$, and $\frac{r_\pi\sqrt{s_\ell(x)}}{\kappa_R^4} \to 0$ as $n \to \infty$. In particular, we only need to consistently estimate the propensity scores $\pi_i, i = 1, ..., n$ because the rate of consistency r_π depends on the in-sample prediction; see Remark 5.8 below.

The assumptions in Theorem 5.5 are also sufficient to establish strong duality between the primal debiasing program (5.3) and its dual (5.7) as follows.

Lemma 5.6 (Sufficient conditions for strong duality)**.** *Let* $\delta \in (0, 1)$. *Under the assumptions of Theorem 5.5, there exists* $\gamma > 0$ *such that* $\frac{\gamma}{n} \asymp r_\gamma \equiv r_\gamma(n, \delta)$ *and strong duality as well as the relation* (5.8) *in Proposition 5.1 hold with probability at least* $1 - \delta$.

The consistency results for the Lasso pilot estimate $\widehat{\beta}$ and the dual solution $\widehat{\ell}(x)$ also imply that under mild regularity conditions, the "Conditional variance II" term in Proposition 5.2 is asymptotically negligible; see Lemma 5.7. This result completes the motivation of the debiasing program (5.3) from the perspective of a (conditional) bias-variance trade-off in Section 5.2.3.

Lemma 5.7 (Negligible conditional variance term)**.** *Suppose that Assumptions 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6 hold as well as*

$$\left[s_\beta \log(d/s_\beta) + s_\ell(x) \log(d/s_\ell(x))\right]^2 = O\left(\sqrt{n}\right), \ \frac{||x||_2}{\kappa_R^2}\sqrt{\phi_{2s_\ell(x)}\phi_{2s_\beta}} = O(1), \ \text{and} \ r_\ell = O(1).$$

If $\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 = o_P(1)$ *and* $\left|\left|\widehat{\ell}(x) - \widetilde{\ell}(x)\right|\right|_2 = o_P(1)$*, then the "Conditional variance II" term in Proposition 5.2 is asymptotically negligible in the sense that*

$$\left(\beta_0 - \widehat{\beta}\right)^T \left[\sum_{i=1}^n \widehat{w}_i^2(x)\widehat{\pi}_i\left(1-\widehat{\pi}_i\right) X_i X_i^T\right]\left(\beta_0 - \widehat{\beta}\right) = o_P(1).$$

5.3.4 Asymptotic Normality

In this section, we leverage the consistency results from Section 5.3.3 to establish the asymptotic normality of the debiased estimator $\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}})$ in (5.4). As mentioned in Section 5.3.1, the key observation for applying the CLT is the linear expansion (5.14) based on the dual solution.

Theorem 5.8 (Asymptotic normality of the debiased estimator)**.** *Let $x \in \mathbb{R}^d$ be a fixed query point with $x \neq 0$ and $s_{\max} = \max\{s_\beta, s_\ell(x)\}$. Suppose that Assumptions 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6 hold and $\lambda, \gamma, r_\ell > 0$ are specified as in Theorem 5.4(b) and Theorem 5.5, respectively. Furthermore, we assume that $\sigma_\epsilon \phi_1 \phi_{s_{\max}}^{3/2} ||x||_2 = O(1)$,*

$$\frac{(1+\kappa_R^2)s_{\max}\log(nd)}{\kappa_R^4} = o\left(\sqrt{n}\right), \quad \text{and} \quad \frac{(1+\kappa_R^4)\sqrt{\log(nd)}}{\kappa_R^6}\left(\sqrt{s_{\max}}\cdot r_\ell + s_{\max} r_\pi\right) = o(1). \tag{5.21}$$

Then, when $\sigma_m^2(x) = \lim_{n\to\infty} \sigma_\epsilon^2 \cdot x^T \left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x$ exists, we have that

$$\sqrt{n}\left[\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}) - m_0(x)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_m^2(x)\right).$$

We make four remarks about this paramount result.

Remark 5.6 (Semiparametric efficiency)**.** *Given any fixed dimension $d > 0$, the asymptotic variance*

$$\sigma_{m,d}^2(x) = \sigma_\epsilon^2 \cdot x^T \left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x = \sigma_\epsilon^2 \cdot x^T \left\{\mathbb{E}\left[\pi(X)XX^T\right]\right\}^{-1} x$$

achieves the semiparametric efficiency bound for our debiased estimator among all asymptotically linear estimators with the MAR outcome (Müller and Keilegom, 2012). It also attains the efficiency bound for de-sparsified Lasso in Theorem 3 of Jankova and van de Geer (2018) under Gaussian covariates and noise. This key property demonstrates the superiority of our debiasing inference method in terms of statistical efficiency.

Remark 5.7 (Requirement on the query point $x \in \mathbb{R}^d$)**.** *We require* $||x||_2 = O(1)$ *to control higher-order remainder terms and establish asymptotic normality of* $\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}})$*. This requirement may seem contradictory to the standard result* $||x||_2 = O_P\left(\sqrt{d}\right)$ *in high-dimensional statistics when* x *is drawn from a sub-Gaussian distribution; see Theorem 3.1.1 in Vershynin (2018). However, for those scenarios where* $||x||_2$ *grows with dimension* d*, we can always apply our debiasing method to the normalized query point* $\frac{x}{||x||_2}$*. Since the solution* $\widehat{\boldsymbol{w}} \equiv \widehat{\boldsymbol{w}}\left(\frac{x}{||x||_2}\right)$ *to our debiasing program (5.3) is adaptive to the normalization, the resulting debiased estimator* $\widehat{m}^{\text{debias}}\left(\frac{x}{||x||_2}; \widehat{\boldsymbol{w}}\right)$ *and associated asymptotic confidence intervals can be rescaled with* $||x||_2$ *in order to obtain asymptotically valid confidence intervals for* $m_0(x) = x^T\beta_0$*.*

Remark 5.8 (In-sample consistency of the propensity score estimation)**.** *Our debiased estimator is asymptotically normal as long as the propensity scores are pointwise consistently estimated at in-sample data points, because the rate requirement (5.21) on* r_π *relies only on the in-sample prediction. Furthermore, unlike double/debiased machine learning (Chernozhukov et al., 2018), there is no need to leverage sample-splitting or cross-fitting when estimating the nuisance propensity score* $\pi(X) = \mathrm{P}(R = 1|X)$ *in order to guarantee the asymptotic normality. While we derive* $r_\pi = O_P\left(\frac{\log d}{\sqrt{n}}\right)$ *under the Lasso-type generalized linear regression estimator (D.3) in Proposition D.2 of Section D.3, this in-sample prediction rate of consistency is, in general, slower than those rates from more complicated machine learning methods such as deep neural networks (Farrell et al., 2021), random forests (Gao et al., 2022), and support vector machines with universal kernels (such as the Gaussian radial basis function) (Steinwart, 2001). The higher learnability of these propensity score estimation methods leads*

to better performance of our debiasing method; see Section 5.4.4 for numerical experiments.

Remark 5.9 (Necessary sparsity condition)**.** *Compared with the consistency results in Theorem 5.5, we impose a stricter growth condition* $s_{\max} = \max\{s_\beta, s_\ell(x)\} = o\left(\frac{\sqrt{n}}{\log d}\right)$ *on the sparsity level in order to establish the asymptotic normality, provided that other model parameters* $\sigma_\epsilon, \phi_{s_{\max}}, \kappa_R$ *are independent of* n *and* d*. The requirement on the sparsity level* s_β *is a standard and essentially necessary condition in the high-dimensional debiased inference literature in pursuit of asymptotic normality (van de Geer et al., 2014; Javanmard and Montanari, 2014a, 2018; Cai et al., 2023); see also the discussion in Section 8.6 of Jankova and van de Geer (2018). The condition on the sparsity level* $s_\ell(x)$ *of the population dual solution* $\ell_0(x) = -2\left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x$ *to (5.18) is specific to the present scalar functional formulation, because we establish the asymptotic normality of our debiased estimator for* $m_0(x) = x^T\beta_0$ *rather than for a debiased estimator of the full regression vector* $\beta_0 \in \mathbb{R}^d$*. Notably, Bellec and Zhang (2022) imposed a slightly weaker sparsity condition on essentially the same object as* $\ell_0(x)$ *and only required* $s_\ell(x) = o\left(\frac{n}{\log d}\right)$*. We plan to investigate in future research whether we can further improve our proof technique and weaken our sparsity condition to match theirs.*

To utilize the asymptotic normality of Theorem 5.8 and conduct inference on $m_0(x)$ in practice, we need to estimate the asymptotic variance $\sigma_m^2(x)$ or $\sigma_{m,d}^2(x)$ under the current dimension d. Proposition 5.9 below verifies that, when the noise level σ_ϵ^2 is known, $\sigma_\epsilon^2 \sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2(x)$ is a consistent estimator of $\sigma_{m,d}^2(x)$.

Proposition 5.9 (Consistent estimate of the asymptotic variance)**.** *Let* $x \in \mathbb{R}^d$ *be a fixed query point. Suppose that Assumptions 5.1, 5.2, 5.4, 5.5, and 5.6 hold and* $\gamma, r_\ell > 0$ *are specified as in Theorem 5.5. Furthermore, we assume that* $||x||_2^2 \phi_{s_\ell(x)}^2 = O(1)$*,* $\frac{(1+\kappa_R^3)}{\kappa_R^5}\sqrt{\frac{s_\ell(x)\log(nd)}{n}} = o(1)$*, and* $\frac{(1+\kappa_R^4)}{\kappa_R^6}\left[r_\ell + r_\pi\sqrt{s_\ell(x)}\right] = o(1)$*. Then,*

$$\left|\sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2(x) - x^T\left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x\right| = o_P(1).$$

This result again motivates the design of our debiasing program as in Section 5.2.3 and supports the formulation of our asymptotically valid confidence interval (5.5). When the noise level σ_ϵ^2 is unknown in practice, one can always replace it with a consistent estimator $\widehat{\sigma}_\epsilon^2$ without losing any asymptotic efficiency and obtain the final estimator $\widehat{\sigma}_\epsilon^2 \sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2(x)$ for the asymptotic variance. There are various approaches available in the literature to construct such an estimator for σ_ϵ^2 (Zhang, 2010; Fan et al., 2012; Dicker, 2012; Bayati et al., 2013; Belloni and Chernozhukov, 2013; Reid et al., 2016). In our simulation studies and real-world application, we use the scaled Lasso (Sun and Zhang, 2012) with automatically selected regularization parameter $\lambda > 0$ to obtain $\widehat{\sigma}_\epsilon^2$ (and the Lasso pilot estimate $\widehat{\beta}$ simultaneously); see Section D.1 for details.

5.4 Experiments

In this section, we evaluate the empirical performance of our proposed debiasing inference method in Section 5.2.1 and compare it with several existing high-dimensional inference methods in the literature through comprehensive simulation studies.

5.4.1 Basic Simulation Design and Methods to Be Compared

We generate the i.i.d. data $\{(Y_i, R_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ from the following linear model

$$Y_i = X_i^T \beta_0 + \epsilon_i \quad \text{with} \quad X_i \perp\!\!\!\perp \epsilon_i, \quad Y_i \perp\!\!\!\perp R_i | X_i, \quad \text{and} \quad X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma), \tag{5.22}$$

where $d = 1000$ and $n = 900$ unless stated otherwise. In order to investigate the effect of different covariance structures $\Sigma \in \mathbb{R}^{d \times d}$, sparsity patterns of the true regression vector $\beta_0 \in \mathbb{R}^d$, and designs of the query point $x \in \mathbb{R}^d$ on the finite-sample performance of the candidate high-dimensional inference methods, we consider all possible combinations of the following simulation designs.

Choices of the covariance matrix: We consider two standard designs of $\Sigma = (\Sigma_{jk})_{j,k=1}^d$ from the literature:

(i) The circulant symmetric matrix Σ^{cs} in Javanmard and Montanari (2014a) defined by $\Sigma_{jj} = 1$, $\Sigma_{jk} = 0.1$ when $j+1 \le k \le j+5$ or $j+d-5 \le k \le j+d-1$ with $\Sigma_{jk} = 0$ elsewhere for $j \le k$, and $\Sigma_{jk} = \Sigma_{kj}$;

(ii) The Toeplitz (or autoregressive) matrix Σ^{ar} in van de Geer et al. (2014) defined by $\Sigma_{jk} = 0.9^{|j-k|}$.

Designs of the regression vector: For the true value of $\beta_0 \in \mathbb{R}^d$, we consider three different scenarios as:

(i) $\beta_0^{sp} = \left(\underbrace{\sqrt{5}, ..., \sqrt{5}}_{5}, 0, ..., 0\right)^T \in \mathbb{R}^d$ is sparse;

(ii) $\beta_0^{de} \propto \left(1, \frac{1}{\sqrt{2}}, ..., \frac{1}{\sqrt{d}}\right)^T \in \mathbb{R}^d$ is dense with $\left|\left|\beta_0^{de}\right|\right|_2 = 5$;

(iii) $\beta_0^{pd} \propto \left(1, \frac{1}{2}, ..., \frac{1}{d}\right)^T \in \mathbb{R}^d$ is pseudo-dense with $\left|\left|\beta_0^{pd}\right|\right|_2 = 5$.

Designs of the query point: We conduct experiments with the following four different choices of the query point $x \in \mathbb{R}^d$:

(i) $x^{(1)} = (1, 0, ..., 0)^T \in \mathbb{R}^d$ to infer the individual effect of the first (significantly nonzero) component of β_0;

(ii) $x^{(2)} = \left(1, \frac{1}{2}, \frac{1}{4}, 0, 0, 0, \frac{1}{2}, \frac{1}{8}, 0, ..., 0\right)^T \in \mathbb{R}^d$ to infer the joint effects of a few components of β_0;

(iii) $x^{(3)} = \left(0, ..., 0, \underbrace{1}_{100^{th}}, 0, ..., 0\right)^T \in \mathbb{R}^d$ to infer an inactive (*i.e.*, truly zero or close to zero) component of β_0;

(iv) $x^{(4)} = \left(1, \frac{1}{2^2}, ..., \frac{1}{d^2}\right)^T \in \mathbb{R}^d$ for the circulant symmetric covariance cases, and $x^{(4)} = \left(1, \frac{1}{2}, ..., \frac{1}{d}\right)^T \in \mathbb{R}^d$ for the Toeplitz covariance cases. The purpose of choosing a relatively dense query point $x^{(4)}$ is to study the joint effect of all the components of β_0;

(v) $x^{(5)} = \left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, ..., \frac{1}{\sqrt{d}}\right)^T \in \mathbb{R}^d$, which is used to examine how a fully dense query point and the corresponding population dual solution $\ell_0(x)$ affect the inference results.

MAR mechanism: After $Y_i, i = 1, ..., n$ are sampled according to (5.22), we define the missingness indicators $R_i, i = 1, ..., n$ for $Y_i, i = 1, ..., n$ through the MAR mechanism by

$$\mathrm{P}(R_i = 1|X_i) = \frac{1}{1 + \exp\left(-1 + X_{i7} - X_{i8}\right)} \quad \text{for} \quad i = 1, ..., n. \tag{5.23}$$

In general, the above MAR mechanism yields around 28% of missingness for the outcome variables $Y_i, i = 1, ..., n$, making the complete-case data even more high-dimensional. Additional simulation results under a simpler MCAR setting are deferred to Section D.4.2.

Noise distribution: We first generate the noise variables $\epsilon_i, i = 1, ..., n$ independently from $\mathcal{N}(0, 1)$. Other noise distributions that violate the sub-Gaussian noise condition (Assumption 5.3) are considered in Section 5.4.3.

Methods to be compared: To demonstrate the statistical efficiency of our proposed debiasing method (**"Debias"**), we compare it with several existing inference methods in the literature. The detailed implementation of our proposed method can be found in Section D.1, where we also explain the three different criteria "min-CV", "1SE", and "min-feas" for selecting the tuning parameter $\gamma > 0$ via cross-validation. For a fair comparison, we implement the first group of comparative methods below on (i) the complete-case (CC) data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ with $R_i = 1$ for $i = 1, ..., n$, (ii) the inverse probability weighted (IPW) data $\left(\frac{X_i}{\sqrt{\widehat{\pi}_i}}, \frac{Y_i}{\sqrt{\widehat{\pi}_i}}\right) \in \mathbb{R}^d \times \mathbb{R}$ with $R_i = 1$ for $i = 1, ..., n$, and (iii) the oracle fully observed data $(X_i, Y_i), i = 1, ..., n$. Here, $\widehat{\pi}_i, i = 1, ..., n$ are the propensity scores estimated by the default Lasso-type logistic regression as described in Section D.3.

- **"DL-Jav"** is the debiased Lasso proposed by Javanmard and Montanari (2014a) that solves d convex programs for an unbiased estimator for β_0 and a consistent estimate for the asymptotic covariance matrix. The method is implemented by the source code `sslasso` provided in their paper. We select the tuning parameters as in their Section 5.1.

- **"DL-vdG"** is the debiased Lasso proposed by van de Geer et al. (2014) that solves d nodewise regressions for the debiased estimator of β_0 and the estimate of its asymptotic covariance matrix. This method is implemented via the function `lasso.proj` in the R package `hdi` (Dezeure et al., 2015).

- **"R-Proj"** is the ridge projection method proposed by Bühlmann (2013) that produces a bias-corrected estimator of β_0 via a ridge regression estimator. This method is implemented by the function `ridge.proj` in the R package `hdi`. We initialize the estimation of β_0 with an estimate obtained from the scaled Lasso. Since it is difficult to obtain a consistent estimate of the asymptotic covariance matrix from the function `ridge.proj`, we only run "R-Proj" when the query point is $x^{(1)}$ or $x^{(3)}$. This means that we only conduct inference on single coordinates of β_0 under these scenarios.

- **"Refit"** first runs the scaled Lasso to obtain a pilot estimate $\widehat{\beta}$ and then computes the final least-squares estimate based on covariates in the support set of $\widehat{\beta}$ (Belloni and Chernozhukov, 2013).

Additionally, we implement several competing high-dimensional inference methods that handle the missing-outcome scenarios:

- **"DDR"** is an ℓ_1-regularized debiased and doubly robust estimator of β_0 based on a high-dimensional adaptation of the traditional doubly robust construction when the outcome Y is possibly MAR (Chakrabortty et al., 2019).

- **"SAS"** is a surrogate-assisted imputation-based inference approach for β_0 with a high-dimensional predictor X in the semi-supervised learning setting (*e.g.*, Y is MCAR) (Hou et al., 2023). For a fair comparison, we apply their method to our simulation setting with no surrogate outcomes.

- **"SSL-AIPW"** is an AIPW method for estimating β_0 in the semi-supervised learning (SSL) setting with MAR outcomes (Tian et al., 2024).

Confidence intervals for "DL-Jav", "DL-vdG", "R-Proj", "DDR", "SAS", and "SSL-AIPW" are constructed according to their asymptotic normality results, while the confidence intervals from "Refit" are based on the assumption that the selected model by the Lasso pilot estimate $\widehat{\beta}$ contains the true support set.

5.4.2 Simulation Results Under the Gaussian Noise Distribution

We evaluate the performance of our proposed debiasing method and the existing inference methods stated above via the average absolute bias $\left|\widehat{m}^{\text{debias}}(x) - m_0(x)\right|$, the average coverage probability of 95% confidence intervals, and the average length of confidence intervals over 1000 Monte Carlo experiments for each simulation scenario stated in Section 5.4.1.

We provide comparative plots in terms of the three evaluation metrics and normality assessments for a selected simulation setting in Figure 5.2, while deferring the full simulation results for both the circulant symmetric and Toeplitz (or autoregressive) covariance designs to Figure D.1 and Table D.1-D.2 in Section D.4.1, respectively. In summary, our proposed debiasing method achieves bias reduction comparable to that of the debiased Lasso "DL-Jav" and the Lasso refitting "Refit", while yielding relatively shorter confidence intervals and more accurate coverage probabilities. Although the ridge projection method "R-Proj" sometimes produces better coverage probabilities than other methods, it consistently produces larger biases and more conservative confidence intervals. Among the competing methods for missing outcomes, "DDR" and "SSL-AIPW" fail to maintain asymptotic normality under our simulation settings, with "SSL-AIPW" exhibiting particularly large absolute bias. Our debiasing method performs comparably to "SAS", but provides more accurate coverage for the resulting confidence intervals. Finally, under the fully dense query point $x^{(5)}$, only "DL-Jav", "DL-vdG", "DDR", and our proposed debiasing method achieve approximately the desired coverage probabilities, suggesting that our method remains robust even when the sparsity assumption on the population dual solution $\ell_0(x)$ is violated.

The superiority of our debiasing method is more evident when the query point $x \in \mathbb{R}^d$ has many nonzero entries. The debiased estimator from our method is asymptotically

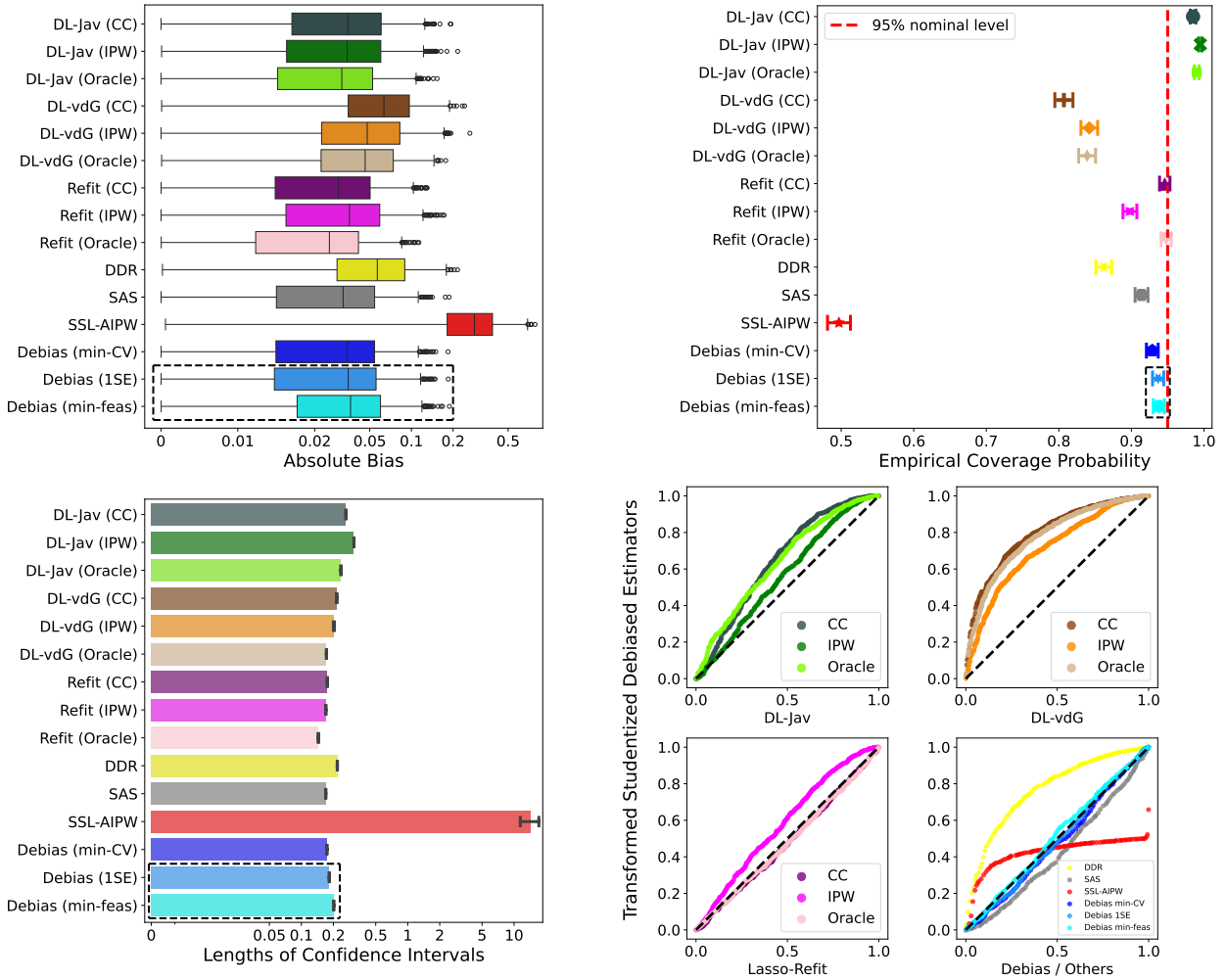


Figure 5.2: Simulation results with sparse β_0^{sp} and sparse $x^{(2)}$ for the circulant symmetric covariance matrix Σ^{cs} under the Gaussian noise $\mathcal{N}(0,1)$ and MAR setting (5.23). **Top Left:** Boxplots of absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Bottom Left:** Average lengths of confidence intervals with standard error bars. **Bottom Right:** Four comparative "QQ-plots" of the transformed studentized debiased estimators $\Phi\left(\sqrt{n}\cdot\widehat{\sigma}_n^{-1}(x)\left[\widehat{m}(x)-m_0(x)\right]\right)$ obtained from different debiasing methods, where $\widehat{\sigma}_n(x)^2$ is the estimated (asymptotic) variance of $\widehat{m}(x)$ by each method and $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$. We also highlight the recommended rules "1SE" and "min-feas" for our proposed debiasing method via dashed rectangles in the first three panels.

semiparametrically efficient for arbitrary x in Theorem 5.8, and its asymptotic variance can be naturally estimated as in Proposition 5.9. Hence, the asymptotic normality of our

debiased estimator can be better approximated in finite samples than that of the competing methods. All these simulation results corroborate the theoretical properties of our debiasing method derived in Section 5.3 and demonstrate its effectiveness in conducting valid statistical inference with missing outcomes.

5.4.3 Simulation Results Under Heavy-Tailed Noise Distributions

The consistency and asymptotic normality results for our debiasing inference method are derived under the sub-Gaussian noise condition (Assumption 5.3), and our previous simulation results in Section 5.4.2 are accordingly based on normally distributed noise. We demonstrate through additional simulations that the asymptotic normality of our debiased estimator (5.4) remains valid even when the distribution of the noise ϵ in model (5.1) is not sub-Gaussian. To this end, we follow the simulation designs in Section 5.4.1 but explore two heavy-tailed noise distributions:

(i) ϵ follows a Laplace distribution with mean 0 and scale parameter $\frac{1}{\sqrt{2}}$. Note that ϵ becomes sub-exponentially distributed and has variance $\mathrm{Var}(\epsilon) = 1$.

(ii) ϵ follows a mean-zero t-distribution with 2 degrees of freedom. In this case, ϵ is not even sub-exponential and has infinite variance.

We present selected plots and asymptotic normality assessments in Figure 5.3 and Figure 5.4, while again deferring the full comparative simulation results for both the circulant symmetric and Toeplitz covariance designs to Table D.3-D.4 and Table D.5-D.6 in Section D.4.1, respectively. In short, the debiased estimator appears to be approximately normally distributed in finite samples even when the noise distribution is heavy-tailed, while other competing methods except for "SAS" fail to maintain their asymptotic normality. Other conclusions from these simulation results are similar to those described in Section 5.4.2.

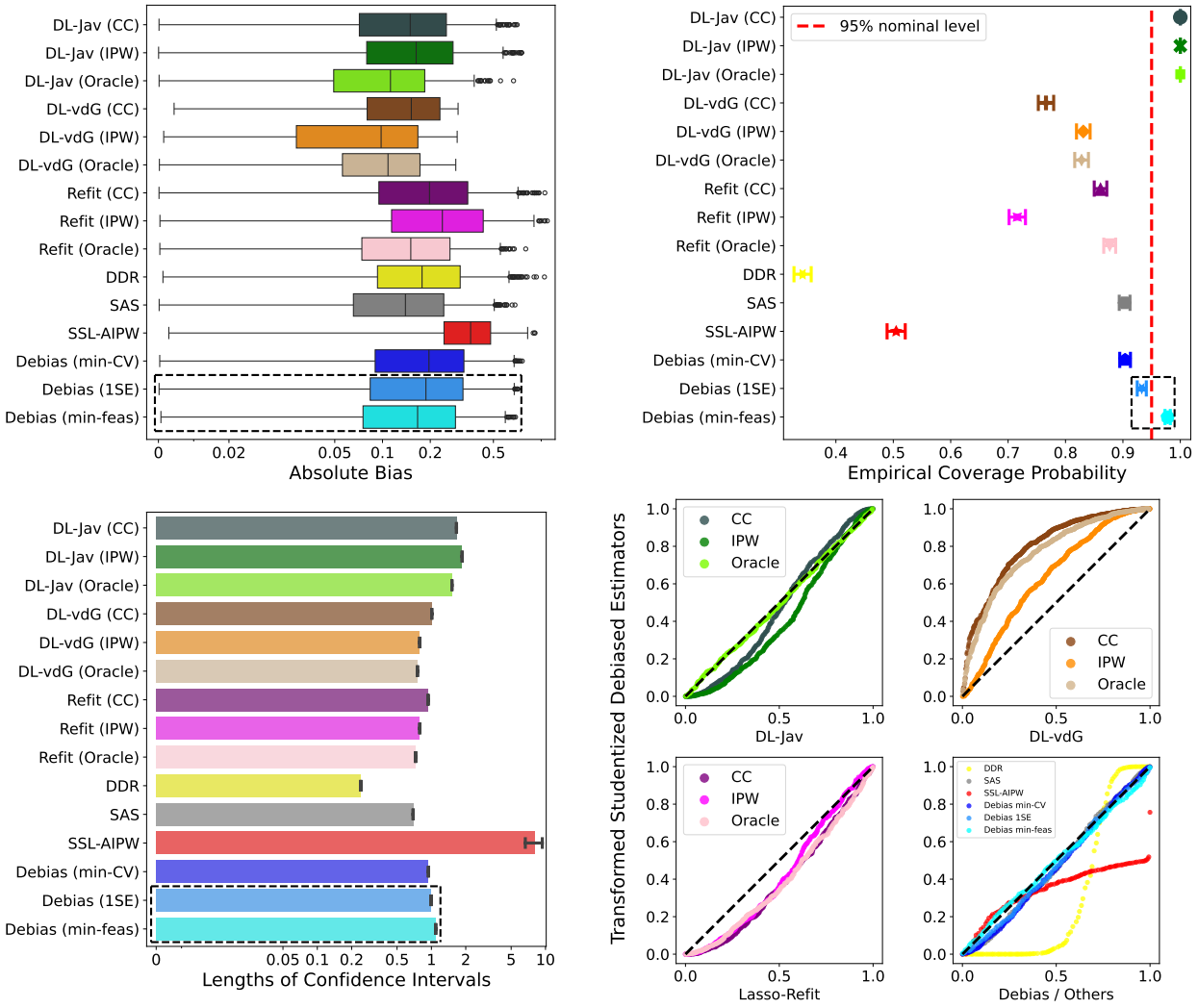


Figure 5.3: Simulation results with dense β_0^{de} and sparse $x^{(2)}$ for the circulant symmetric covariance matrix Σ^{cs} under Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$-distributed noise and the MAR setting (5.23). **Top Left:** Boxplots of the absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Bottom Left:** Average lengths of confidence intervals with standard error bars. **Bottom Right:** Four comparative "QQ-plots" of the transformed studentized debiased estimators $\Phi\left(\sqrt{n} \cdot \widehat{\sigma}_n^{-1}(x)\left[\widehat{m}(x) - m_0(x)\right]\right)$ obtained from different debiasing methods, where $\widehat{\sigma}_n(x)^2$ is the estimated (asymptotic) variance of $\widehat{m}(x)$ by each method and $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$.

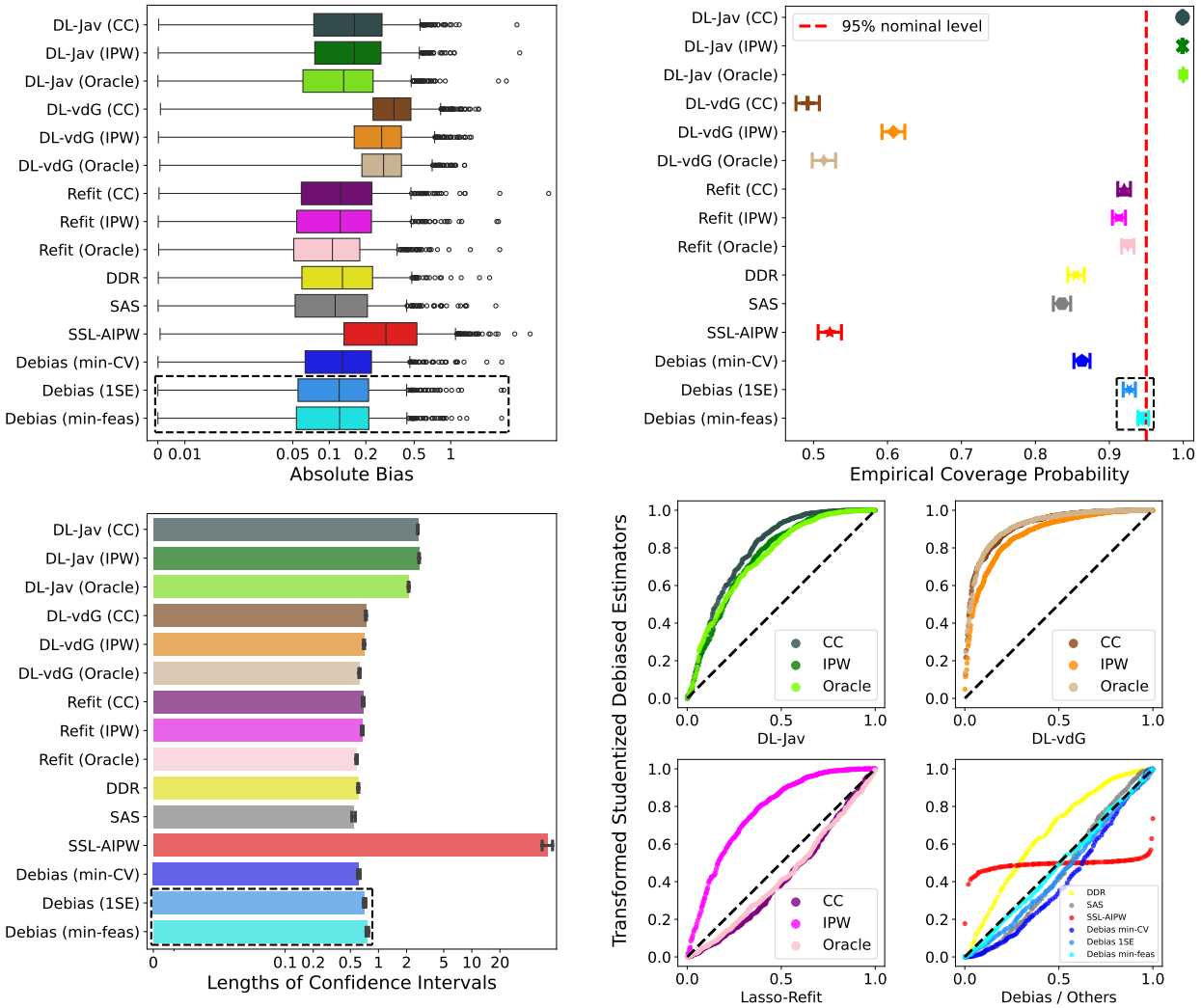


Figure 5.4: Simulation results with pseudo-dense β_0^{pd} and dense $x^{(4)}$ for the Toeplitz (or autoregressive) covariance matrix Σ^{ar} under t_2-distributed noise and the MAR setting (5.23). **Top Left:** Boxplots of the absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Bottom Left:** Average lengths of confidence intervals with standard error bars. **Bottom Right:** Four comparative "QQ-plots" of the transformed studentized debiased estimators $\Phi\left(\sqrt{n}\cdot\widehat{\sigma}_n^{-1}(x)\left[\widehat{m}(x)-m_0(x)\right]\right)$ obtained from different debiasing methods, where $\widehat{\sigma}_n(x)^2$ is the estimated (asymptotic) variance of $\widehat{m}(x)$ by each method and $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$.

5.4.4 Simulation Results for the Proposed Debiasing Method With Nonparametrically Estimated Propensity Scores

As shown in Theorem 5.5 and Theorem 5.8, the consistency and asymptotic normality of the proposed debiasing method do not require any parametric assumption on the propensity

score $\mathrm{P}(R = 1|X)$ under the MAR condition. Therefore, we now conduct additional simulation studies for our debiasing method in which the propensity scores $\pi(X_i), i = 1, ..., n$ are estimated by the following nonlinear/nonparametric machine learning methods based on the observed data $\{(X_i, R_i)\}_{i=1}^n$. All of these methods are implemented in the `scikit-learn` package (Pedregosa et al., 2011) in Python.

Naive Bayes ("NB"): We adapt the Gaussian naive Bayes method (Zhang, 2004) to model the propensity scores through $\pi(X_i) = \mathrm{P}(R_i = 1|X_{i1}, ..., X_{id}) = \frac{\mathrm{P}(R_i=1)\cdot\prod_{k=1}^d \mathrm{P}(X_{ik}|R_i=1)}{\mathrm{P}(X_{i1},...,X_{id})}$ for any $i = 1, ..., n$, where each class-conditional density $\mathrm{P}(X_{ik}|R_i = r)$ with $r \in \{0, 1\}$ is modeled by a Gaussian density whose parameters are estimated by maximum likelihood.

Random Forest ("RF"): We implement the random forest method (Breiman, 2001) with 100 trees, bootstrap samples, and the Gini impurity to measure the quality of a split.

Support Vector Machine ("SVM"): We fit a support vector machine (Chen et al., 2005) with the Gaussian radial basis function to classify the missingness indicator R_i based on the observed covariate vector $X_i \in \mathbb{R}^d$ for $i = 1, ..., n$. The outputs of the trained SVM are regarded as preliminary estimates of the propensity scores.

Neural Network ("NN"): We train a neural network model (Hinton, 1990) with two hidden layers of size 80×50 and use the rectified linear unit function $h(x) = \max\{x, 0\}$ as the activation function. The learning rate η_{NN} is initially set to 0.001 as long as the training loss keeps decreasing and is adaptively changed to $\eta_{\mathrm{NN}}/5$ when two consecutive epochs fail to decrease the training loss.

For the above methods, we also consider calibrated versions obtained through Platt scaling (Platt, 1999), which fits an extra logistic regression on the estimated propensity scores $\widehat{\pi}_i, i = 1, ..., n$. However, since the errors in the estimated propensity scores and the final performance of the proposed debiased estimator worsen after this calibration, we choose not to report them here.

The covariate vectors $X_i \in \mathbb{R}^d, i = 1, ..., n$ are again sampled independently from $\mathcal{N}_d(\mathbf{0}, \Sigma^{\mathrm{cs}})$ with $d = 1000$ and $n = 900$, and the noise variables $\epsilon_i, i = 1, ..., n$ are generated independently from $\mathcal{N}(0, 1)$ as in Section 5.4.1. To increase the complexity of estimating the propen-

sity scores, we generate the missingness indicators $R_i, i = 1, ..., n$ for the outcome variables $Y_i, i = 1, ..., n$ through a different MAR mechanism from the one in Section 5.4.1 as:

$$\mathrm{P}(R_i = 1|X_i) = \Phi\left(-4 + \sum_{k=1}^{K} Z_{ik}\right), \tag{5.24}$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0, 1)$ and the vector $(Z_{i1}, ..., Z_{iK})$ contains all polynomial combinations of the first eight components $X_{i1}, ..., X_{i8}$ of the covariate vector X_i with degrees less than or equal to two (*i.e.*, including the linear, quadratic, and one-way interaction terms).

The simulation results for our debiasing method under the oracle propensity scores ("Oracle"), the Lasso-type logistic regression (D.4) ("LR"), and the aforementioned nonparametric methods for propensity score estimation are shown in Figure 5.5. We provide more comprehensive results with the evaluation metrics from Section 5.4.2 and an additional metric, the average mean absolute error ("Avg-MAE"), for the estimated propensity scores in Figure D.3, Table D.7, as well as additional "QQ-plots" in Figure D.2 of Section D.4.1.

The main conclusion from this simulation study is that the performance of our debiasing method can be further improved when the propensity scores are better estimated. Under the MAR mechanism (5.24) and Gaussian design for covariate vectors $X_i, i = 1, ..., n$, neural network and Gaussian naive Bayes models outperform other machine learning methods in estimating the propensity scores and generally lead to subsequent debiased estimators with smaller biases, better coverage probabilities, and shorter confidence intervals. Conversely, the Lasso-type logistic regression is misspecified and consequently degrades the performance of the resulting debiased estimator. Another interesting phenomenon from these simulation results is that our debiasing method sometimes performs better when using the estimated propensity scores $\widehat{\pi}_i, i = 1, ..., n$ than when using the oracle propensity scores $\pi_i = \pi(X_i), i = 1, ..., n$. Such a paradox has been analyzed in propensity score matching (Rosenbaum, 1987), in IPW estimation for causal inference (Robins et al., 1992; Su et al., 2023), and in other general parameter estimation problems in the presence of a nuisance parameter (Henmi and Eguchi, 2004; Hitomi et al., 2008; Lok, 2021). A rigorous study of this paradox for our debiasing method is beyond the scope of this chapter.

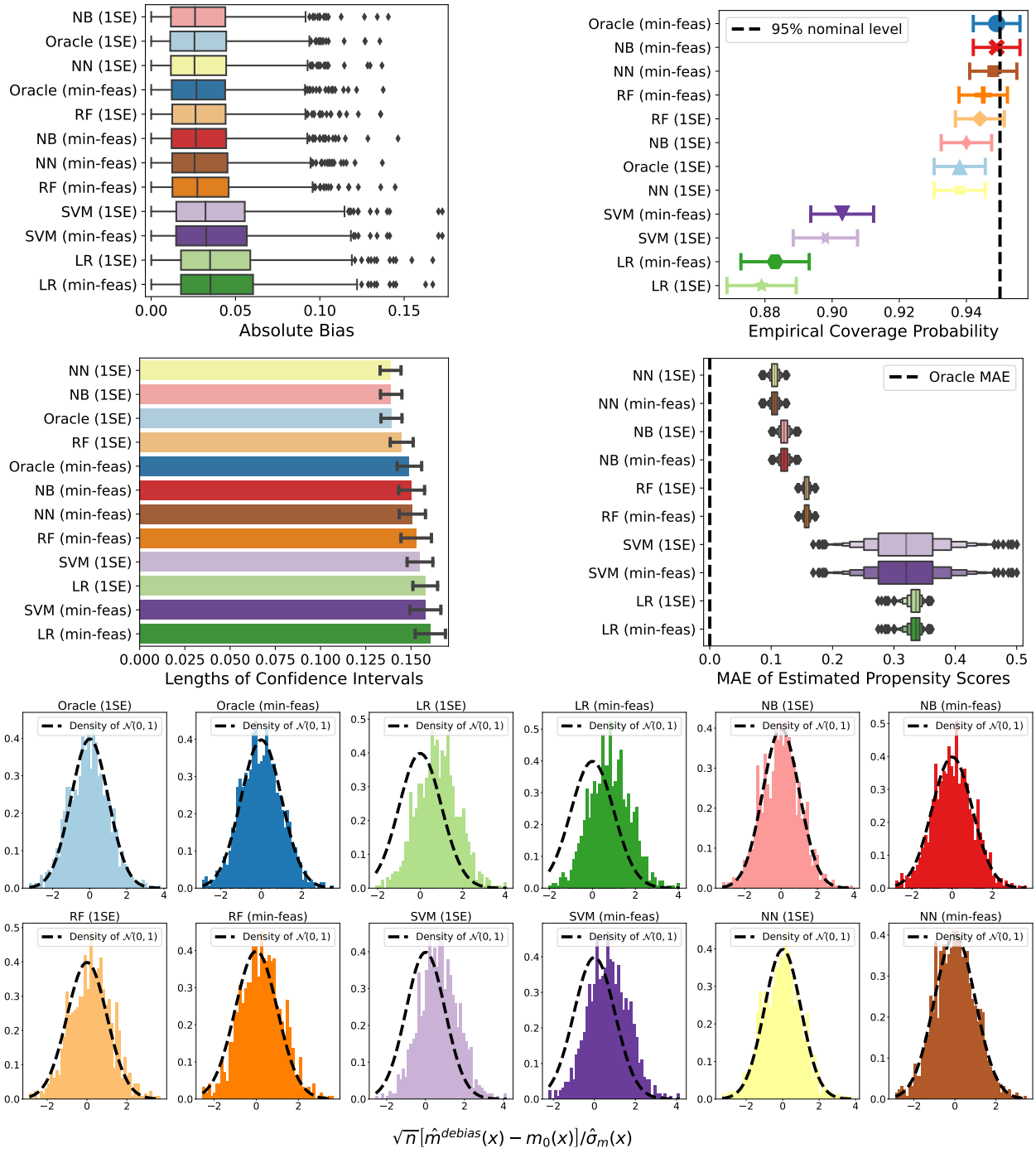


Figure 5.5: Simulation results for our debiasing method with sparse β_0^{sp} and (weakly) dense $x^{(4)}$ when the propensity scores are estimated by various machine learning methods under the new MAR setting (5.24). **Top Left:** Boxplots of the absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Middle Left:** Average lengths of confidence intervals with standard deviation bars. **Middle Right:** Letter-valued plots of the MAE of estimated propensity scores. **Bottom Two Rows:** Histograms of the studentized debiased estimators under different methods. We also order the performance metrics according to their means in the first four panels.

5.5 *Stellar Mass Inference Problem*

We showcase a real-world application of our proposed debiasing method to stellar mass inference for selected galaxies in the Sloan Digital Sky Survey, Fourth Phase and Data Release 16 (SDSS-IV DR16; Ahumada et al. 2020).

5.5.1 Background and Study Design

As discussed in Chapter 4, SDSS-IV DR16 contains three main survey programs, among which the eBOSS survey and its predecessor BOSS cover a broad range of cosmological scales and include observations of a vast number of galaxies. While a couple of value-added catalogs for the estimated stellar masses of observed galaxies based on the BOSS spectra are available (Blanton et al., 2005; Conroy et al., 2009; Chang et al., 2015), we focus on the eBOSS Firefly value-added catalog (Comparat et al., 2017)[1] for our stellar mass inference study given its timeliness and completeness. In more detail, this catalog fits combinations of single-burst stellar population models to spectroscopic data for objects that have reliable positive redshift measurements and are classified as galaxies by SDSS-IV DR16, following an iterative best-fitting process controlled by the Bayesian Information Criterion to produce stellar masses and other stellar properties of the galaxies (Wilkinson et al., 2017). A fraction of the stellar mass estimates in the Firefly catalog are missing because of the limited use of the observational run in SDSS-IV DR16 dedicated to galaxy targets, data contamination, and the misclassification of galaxies as stars (Comparat et al., 2017). Given that we incorporate various spectroscopic and photometric properties of the observed galaxies into the covariate vector for our stellar mass inference task, it is reasonable to assume that the estimated stellar masses are missing at random.

Since nearly all existing catalogs yield only point estimates of stellar mass, we intend to leverage our debiasing inference method to answer two scientific questions:

- *Question 1:* How can we produce uncertainty measures (or confidence intervals) for the

[1]See https://www.sdss4.org/dr17/spectro/galaxy_firefly/.

estimated stellar mass of a newly observed galaxy based on its observed spectroscopic and photometric properties?

- *Question 2:* Is it statistically significant that the stellar mass is negatively correlated with its distance to the nearby cosmic filament structures?

To this end, we obtain the spectroscopic and photometric properties of $n = 1185$ observed galaxies within a thin redshift slice $0.4 \sim 0.4005$ from the SDSS-IV DR16 database as basic covariates. Such a thin redshift slice helps control redshift distortions caused by the Kaiser effect (*i.e.*, galaxies falling toward the galaxy cluster; Kaiser 1987) and the "Finger-of-God" effect (*i.e.*, the elongation of galaxy distributions along the line-of-sight direction; Jackson 1972). Among the galaxies in the selected redshift slice, 30.2% of their stellar masses are missing in the Firefly catalog under the "Chabrier" initial stellar mass function (Chabrier, 2003) and the "MILES" stellar population model (Maraston and Strömbäck, 2011). In addition, the covariate vector for each galaxy consists of right ascension, declination, ugriz bands and their measurement errors, model and cmodel magnitudes for ugriz bandpasses, galactic extinction corrections for ugriz bandpasses, median signal-to-noise per pixel within each ugriz bandpass, original and best-fit template spectra projected onto ugriz filters as well as their inverse variances, sky flux in each of the ugriz imaging filters, and signal-to-noise squared for spectrographs #1 and #2 at $g = 20.20$, $r = 20.20$, and $i = 20.20$ for SDSS spectrograph spectra. To capture nonlinear patterns in stellar mass estimation and handle extreme values, we generate additional covariates by applying the logarithmic transformations $x \mapsto \text{sign}(x) \cdot \log(1 + |x|)$ to ugriz bands, galactic extinction corrections, and flux features. We also remove covariates whose absolute correlation coefficients exceed 0.95 before generating univariate B-spline basis covariates of polynomial order 3 with 40 knots. We adopt B-spline basis covariates instead of the usual polynomial combinations because covariates generated by univariate B-spline bases tend to be less linearly correlated and can facilitate the subsequent inference task. To address *Question 2* above, we compute the angular diameter distances from the galaxies to the two-dimensional spherical cosmic

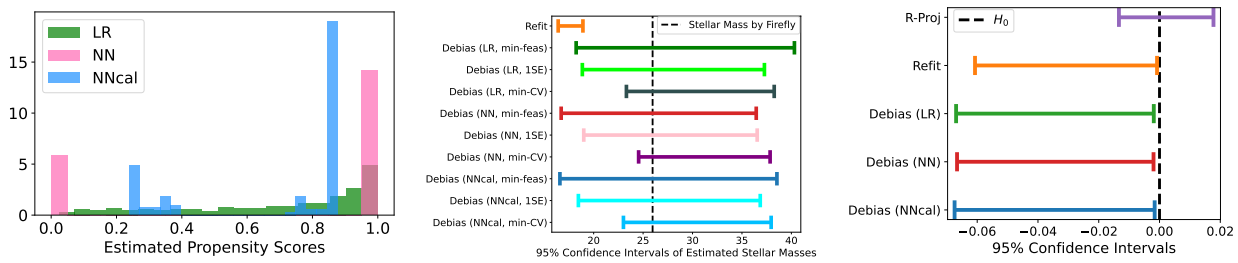


Figure 5.6: Stellar mass inference with our debiasing and related methods. **Left Panel:** Estimated propensity scores by "LR", "NN", and "NNcal". **Middle Panel:** 95% confidence intervals by different methods for the estimated stellar mass of the new galaxy, addressing *Question 1*. **Right Panel:** 95% confidence intervals by different methods for the estimated regression coefficient associated with the distance to cosmic filaments, addressing *Question 2*.

filaments constructed from SDSS-IV DR16 data by the directional subspace constrained mean shift algorithm in Chapter 4. We also control for the confounding effects of nearby galaxy clusters by including the angular diameter distances from galaxies to the estimated local modes and intersections of filaments as extra covariates. The final dimension of the covariate vector for each galaxy in the selected redshift slice is $d = 1409$, and we take the natural logarithm of the Firefly stellar mass of each galaxy as the outcome variable.

5.5.2 Results

To address *Question 1*, we randomly select a galaxy in the nearby redshift slice $0.4005 \sim 0.401$ as a new observation and construct its associated covariate vector as the query point $x^{(Q1)}$ following the procedures in Section 5.5.1. As for *Question 2*, we set the query point as $x^{(Q2)} = (0, 0, 1, 0, ..., 0)^T \in \mathbb{R}^d$, where the only nonzero entry corresponds to the covariate of the angular diameter distance to cosmic filaments. We estimate the propensity scores (*i.e.*, the probabilities of having a non-missing stellar mass) for the galaxies in the redshift slice by the Lasso-type logistic regression ("LR") in Section D.3 and the neural network ("NN") model described in Section 5.4.4. Nevertheless, as the left panel of Figure 5.6 shows, the estimated propensity scores from the neural network model are too extreme, *i.e.*, close to 0

or 1, so we also consider calibrating them through Platt scaling (Platt, 1999), denoted by "NNcal".

Figure 5.6 displays the inference results from our proposed debiasing method and the classical approaches in Section 5.4.1 applied to the complete-case data. Due to the sparse structure of the design matrix created by B-spline bases, "DL-Jav" and "DL-vdG" fail to produce reliable inference results. For *Question 1*, the 95% confidence intervals obtained from our debiasing method cover the stellar mass of the new galaxy in the Firefly catalog regardless of how we estimate the propensity scores or select the tuning parameter, while the Lasso refitting method underestimates the stellar mass by a statistically significant amount. For *Question 2*, the 95% confidence intervals produced by our debiasing and Lasso refitting methods for the regression coefficient of the distance to cosmic filaments are both below 0, aligning with the existing findings that galaxies are less massive when they are farther away from cosmic filaments (Alpaslan et al., 2016; Kraljic et al., 2018; Malavasi et al., 2022). By contrast, the 95% confidence interval obtained from the ridge projection method contains zero and cannot support a statistically significant conclusion about the relationship between the stellar masses of galaxies and their distances to cosmic filaments. These results demonstrate the usefulness of our proposed debiasing method in practice.

Chapter 6

CAUSAL INFERENCE FOR CONTINUOUS TREATMENTS UNDER POSITIVITY VIOLATIONS

The preceding chapters develop geometry-aware tools for reconstructing cosmic web structures and using the resulting catalog in statistical analyses of galaxy properties. Chapter 5, in particular, studies associations or correlations between the proximity of a galaxy to the cosmic filaments and its stellar mass. However, it is well-known in statistics that association or correlation does not imply causation (Holland, 1986; Pearl, 2009). Therefore, this chapter turns to causal inference and develops rigorous methods for continuous treatments when the usual positivity condition may fail. The motivating scientific question is whether a galaxy's local environment causally affects its star formation rate (SFR), a central component of the long-standing "Nature versus Nurture" debate in galaxy evolution.

Because obtaining reliable SFR estimates and other required covariates of the observed galaxies from SDSS alone is difficult, the empirical illustration in this chapter uses the IllustrisTNG cosmological simulation (Pillepich et al., 2018a,b; Springel et al., 2018) rather than the SDSS-IV cosmic web catalog constructed in Chapter 4. This choice allows us to evaluate the proposed causal estimands and their estimators in a setting where the relevant galaxy properties and environmental variables are available in a controlled physical simulation.

6.1 Introduction

In observational studies, causal effects do not always result from standard binary interventions but rather arise from continuous treatments or exposures. The causal effect of a continuous treatment variable $T \in \mathcal{T} \subset \mathbb{R}$ on an outcome $Y \in \mathcal{Y} \subset \mathbb{R}$ is typically summarized by the dose-response curve $t \mapsto m(t) = \mathbb{E}\left[Y^{(t)}\right]$, which characterizes the average

outcome if all individuals in the population were assigned to a treatment level $T = t$ (Robins et al., 2000; Neugebauer and van der Laan, 2007; Díaz and van der Laan, 2013; Kennedy et al., 2017). Here, $Y^{(t)}$ is the potential outcome (Rubin, 1974) that would have been observed under treatment level $T = t$. Additionally, a set of covariates $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ may be observed, capturing potential confounding variables that influence both the treatment T and the outcome Y. Under standard identification conditions (Assumptions 6.1 and 6.2), the dose-response curve coincides with the covariate-adjusted regression function $t \mapsto \mathbb{E}\left[\mu(t, \boldsymbol{X})\right]$, where $\mu(t, \boldsymbol{x}) = \mathbb{E}\left(Y | T = t, \boldsymbol{X} = \boldsymbol{x}\right)$, making it identifiable from the observable data for any $t \in \mathcal{T}$.

However, a central difficulty in observational studies is that the positivity condition may fail. For continuous treatments, the standard positivity condition requires that each treatment level of interest be feasible, with non-negligible conditional density, for all covariate values in the target population.

Assumption 6.1 (Positivity)**.** *The conditional density $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ is bounded above and away from zero almost surely for all $(t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}$.*

The positivity condition (6.1) may be violated in observational studies for two main reasons: (i) theoretically, it is impossible for some individuals with specific covariate values to receive certain treatment levels, and (ii) practically, a finite data sample may fail to collect individuals exposed to certain treatment levels; see Cole and Hernán (2008); Westreich and Cole (2010); Petersen et al. (2012) for related discussions. These issues are particularly pervasive in the context of continuous treatments. For instance, air pollution levels may vary significantly across larger regions while remaining relatively consistent within smaller areas, resulting in individuals in the same location being exposed to only one level of pollution. This discrepancy, where spatial covariates change at a finer scale than the exposure measurement, leads to violations of the positivity condition (Paciorek, 2010; Schnell and Papadogeorgou, 2020; Keller and Szpiro, 2020). When Assumption 6.1 is violated, the counterfactual mean $\mathbb{E}\left[Y^{(t)}\right]$ is not identifiable from the observed data distribution. As an illustrative example, we

consider a binary covariate $X \in \{0,1\}$ with $\mathbb{P}(X=1)=\frac{1}{2}$ and the conditional distribution of $T|X=x$ as $(1-x)\cdot\text{Uniform}[-1,1]+x\cdot\text{Uniform}[-1,0]$. Under the two potential outcome models

$$\textbf{Model 1: } Y^{(t)} = X\cdot t+\epsilon, \quad \textbf{Model 2: } Y^{(t)} = (X\cdot t+\epsilon)\cdot \mathbb{1}_{\{t\le 0\}} + \epsilon\cdot \mathbb{1}_{\{t>0\}},$$

where ϵ is mean-zero noise independent of (T,X), the observed distribution of (Y,T,X) is identical under standard identification conditions (Assumption 6.2), but the positivity condition is violated for $X=1$ when $T=t>0$. More importantly, the dose-response curves $t\mapsto m(t)=\mathbb{E}\left[Y^{(t)}\right]$ under these two models differ:

$$\textbf{Model 1: } \mathbb{E}\left[Y^{(t)}\right]=\frac{t}{2} \text{ for all } t\in[-1,1], \quad \textbf{Model 2: } \mathbb{E}\left[Y^{(t)}\right]=\frac{t}{2}\cdot\mathbb{1}_{\{-1\le t\le 0\}}.$$

This chapter develops several remedies for causal inference with continuous treatments under positivity violations. First, in Section 6.2, we introduce a novel identification strategy for the counterfactual dose-response curve $m(t)$ and its derivative $\theta(t)=\frac{d}{dt}\mathbb{E}\left[Y^{(t)}\right]$ based on conditional expectations and an additive structural model assumption on $Y^{(t)}$. Leveraging this strategy, we propose an integral estimator of $m(t)$ and a localized estimator of $\theta(t)$, both of which remain consistent even when the positivity condition (6.1) fails in some regions of $\mathcal{T}\times\mathcal{X}$.

Second, in Section 6.3 and Section 6.4, we show that conventional inverse probability weighting (IPW) and doubly robust (DR) estimators remain biased under positivity violations, even when $m(t)$ and $\theta(t)$ are identifiable under an additive structural model. The bias arises from a discrepancy between the conditional covariate support $\mathcal{X}(t)$ at treatment level t and the marginal covariate support $\mathcal{X}$ of the target population. We use nonparametric set estimation to construct bias-corrected IPW and DR estimators for $\theta(t)$ and, consequently, bias-corrected estimators for $m(t)$ with the integral estimator strategy.

Finally, we argue in Section 6.5–Section 6.7 that, when positivity (Assumption 6.1) fails, the classical dose-response curve may no longer be the most scientifically meaningful target estimand. We thus consider a new counterfactual regime defined by the *nearest feasible*

treatment policy (NFTP). Given the target level $t_0 \in \mathcal{T}$, the NFTP intervention assigns each unit with covariates $\boldsymbol{x} \in \mathcal{X}$ the treatment level in its conditional feasible set $\mathcal{T}(\boldsymbol{x})$ that is closest to t_0, where $\mathcal{T}(\boldsymbol{x})$ denotes the support of the conditional treatment distribution of T given $\boldsymbol{X} = \boldsymbol{x}$. The induced mean potential outcome curve coincides with the classical dose-response curve when positivity holds, because then t_0 is feasible for every covariate value. When positivity fails, the policy modifies only unsupported treatment assignments while preserving the original target population. This distinguishes the NFTP estimand from propensity score trimming approaches, which address positivity violations by changing the population under study (van der Laan and Petersen, 2007; Branson et al., 2023). The NFTP is also distinct from modified treatment policies (Muñoz and van der Laan, 2012; Haneuse and Rotnitzky, 2013), which intervene by modifying each unit's natural treatment value rather than by projecting an externally specified target level onto the unit-specific feasible support.

To conclude this chapter, we apply the proposed framework to the IllustrisTNG cosmological simulation data to study the causal effect of local environment on galactic SFR. Although the preceding chapters use SDSS observations to reconstruct and analyze cosmic web structures, conducting a reliable causal analysis of SFR that includes the other required covariates is substantially more difficult with data from such a sky survey. The IllustrisTNG simulation provides a suitable complementary test bed. It records galaxy properties, halo information, and related environmental measurements across cosmic time (or redshift), allowing us to examine positivity violations and compare the proposed NFTP-based estimators with existing dose-response analyses.

6.2 Basic Setup, Identification, and Integral Estimation Strategy

Following the potential outcome framework (Rubin, 1974), we let $Y^{(t)}$ denote the potential outcome that would have been observed under treatment level $T = t$. The mean counterfactual outcome for the continuous treatment variable can be written as $t \mapsto m(t) = \mathbb{E}\left[Y^{(t)}\right]$, and its derivative effect curve is defined as $t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y^{(t)}\right]$. We first state some basic

identification assumptions.

Assumption 6.2 (Basic causal conditions)**.** *For any $t \in \mathcal{T}$, the following conditions hold:*

(a) (Consistency) $T = t$ implies that $Y^{(t)} = Y$.

(b) (Ignorability or unconfoundedness) $Y^{(t)}$ is conditionally independent of T given $\boldsymbol{X}$.

Assumption 6.3 (Treatment variation)**.** *The conditional variance of T given any $\boldsymbol{X} = \boldsymbol{x} \in \mathcal{X}$ is strictly positive,* i.e.*, $\text{Var}(T|\boldsymbol{X} = \boldsymbol{x}) > 0$ for any $\boldsymbol{x} \in \mathcal{X}$.*

Assumption 6.2(a) is the usual consistency condition that requires the observed outcome to agree with the potential outcome under the treatment actually received (Kennedy et al., 2017). Implicit in this formulation is that the intervention $T = t$ is well-defined, ruling out interference and relevant hidden versions of treatment as in the stable unit treatment value assumption (SUTVA; Page 19 of Cox 1958 and Rubin 1980). For continuously distributed treatments, where regular conditional distributions of (potential) outcomes given $T = t$ are generally defined only for almost every $t \in \mathcal{T}$, we would formulate Assumption 6.2(a) as $Y^{(T)} = Y$ almost surely. The ignorability condition in Assumption 6.2(b) was first generalized to continuous treatments by Hirano and Imbens (2004), implying that the potential outcome variable is independent of the treatment within each specific stratum of covariates. In spatial confounding, this condition holds when spatial locations are adjusted for (Gilbert et al., 2023). It further ensures that the mean potential outcome under $T = t$ is invariant across treatment levels, conditional on $\boldsymbol{X} = \boldsymbol{x}$. Finally, Assumption 6.3 is crucial for identifying the conditional mean outcome (or regression) function $\mu(t, \boldsymbol{x}) = \mathbb{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$ over a non-degenerate region of $\mathcal{T} \times \mathcal{X}$. Notice that when T is binary and model (6.4) below is assumed, it can be relaxed to the "strict overlap with positive probability" condition in Section 4.2 of D'Amour et al. (2021). In the context of a continuous treatment T, we demonstrate how $\text{Var}(T|\boldsymbol{X} = \boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \mathcal{X}$ can lead to ambiguous definitions of the associated dose-response curves in Example E.1 of Section E.1. In cases where Assumption 6.3 fails,

we also derive nonparametric bounds on $m(t)$ and its derivative $\theta(t)$ in Section E.1 under an additive confounding model (6.4) stated in Section 6.2.1.

Unfortunately, Assumptions 6.2 and 6.3 alone are insufficient to identify the dose-response curve $t \mapsto m(t) = \mathbb{E}\left[Y^{(t)}\right]$ with observable data in the standard manner, such as through the covariate-adjusted regression function (or g-formula; Robins 1986) $\mathbb{E}\left[\mu(t, \boldsymbol{X})\right] = \int_{\mathcal{X}} \mu(t, \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})$. The key reason is that without the positivity condition in Assumption 6.1, the conditional mean outcome function $\mu(t, \boldsymbol{x}) = \mathbb{E}\left(Y | T = t, \boldsymbol{X} = \boldsymbol{x}\right)$ is ill-defined in those regions of $\mathcal{T} \times \mathcal{X}$ outside the support $\mathcal{E}$ of the joint density $p_{T,\boldsymbol{X}}(t, \boldsymbol{x})$. The same argument applies to the identification of the derivative effect curve $t \mapsto \theta(t) = \frac{d}{dt}\mathbb{E}\left[Y^{(t)}\right]$. To resolve this identification issue, we introduce an additional assumption to identify the derivative curve $\theta(t)$, subsequently linking the identification of $m(t)$ to $\theta(t)$ via an integral formula.

Assumption 6.4 (Extrapolation condition)**.** *The function* $\mathbb{E}\left[Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}\right]$ *(or one of its versions) is continuously differentiable with respect to* t *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \in \mathcal{X}$*. Moreover, for* P_T*-almost every* $t \in \mathcal{T}$*, the following equalities hold true:*

$$\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)} | \boldsymbol{X}\right]\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)} | \boldsymbol{X}\right] \middle| T = t\right]. \tag{6.1}$$

Additionally, it holds true that $\mathbb{E}(Y) = \mathbb{E}\left[m(T)\right]$.

The first equality of (6.1) essentially reflects the interchangeability of expectation and partial derivative operations, which holds under mild conditions. In particular, this equality is valid if $\left|\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)} | \boldsymbol{X}\right]\right|$ is bounded by some integrable function $\bar{\mu}(\boldsymbol{X})$ with respect to the distribution of $\boldsymbol{X}$; see Theorem 1.1 and Example 1.8 in Shao (2003). The key identification condition in Assumption 6.4 is captured by the second equality in (6.1), which resembles the (strong) parallel-trends condition imposed by the difference-in-differences design (Abadie, 2005; Callaway et al., 2024). As demonstrated by Proposition 6.1 below, it connects the hypothetical potential outcome regime and the observable data distribution. More importantly, this assumption allows identification of $\theta(t)$ by extrapolating from the support $\mathcal{X}(t)$ of $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$, where positivity holds, to the entire marginal support $\mathcal{X}$ of $p_{\boldsymbol{X}}(\boldsymbol{x})$. We will

also verify in Section 6.2.1 that an additive confounding model (6.4) and its variant satisfy the second equality of (6.1).

Remark 6.1. *As an alternative to Assumption 6.4, we may also assume that the potential outcome $Y^{(t)}$ is continuously differentiable with respect to t almost surely and that, for every $t_0 \in \mathcal{T}$, there exists a neighborhood $\mathcal{N}(t_0)$ and an integrable random variable G_{t_0} such that $\sup_{t\in\mathcal{N}(t_0)} \left|\frac{\partial}{\partial t}Y^{(t)}\right| \leq G_{t_0}$ almost surely. Then, $\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}Y^{(t)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial}{\partial t}Y^{(t)}\middle|\boldsymbol{X}\right]\right]$. If, in addition,*

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\partial}{\partial t}Y^{(t)}\middle|\boldsymbol{X}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial}{\partial t}Y^{(t)}\middle|\boldsymbol{X}\right]\middle|T=t\right],$$

then this provides an alternative formulation of the extrapolation condition in Assumption 6.4. Smoothness assumptions imposed directly on $Y^{(t)}$ have also been considered, for example, in Assumption 3 of Rothenhäusler and Yu (2019) for identifying incremental effects.

Proposition 6.1 (Identifications for $\theta(t)$ and $m(t)$)**.** *Under Assumptions 6.2, 6.3, and 6.4, we have that for P_T-almost every $t \in \mathcal{T}$,*

$$\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)}|\boldsymbol{X}\right]\middle|T=t\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{X})\middle|T=t\right] := \theta_C(t), \tag{6.2}$$

where $\mu(t,\boldsymbol{x}) = \mathbb{E}\left(Y|T=t,\boldsymbol{X}=\boldsymbol{x}\right)$. Furthermore, if the marginal support $\mathcal{T}$ is connected, then

$$m(t) = \mathbb{E}\left[m(T) + \int_{\widetilde{t}=T}^{\widetilde{t}=t}\theta(\widetilde{t})\,d\widetilde{t}\right] = \mathbb{E}(Y) + \mathbb{E}\left\{\int_{\widetilde{t}=T}^{\widetilde{t}=t}\mathbb{E}\left[\frac{\partial}{\partial t}\mu(\widetilde{t},\boldsymbol{X})\middle|T=\widetilde{t}\right]\,d\widetilde{t}\right\}. \tag{6.3}$$

The proof of Proposition 6.1 is in Section E.2.1. These identification formulas suggest that the derivative curve $\theta(t)$ is identifiable through $\theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{X})\middle|T=t\right]$, and we only require $\frac{\partial}{\partial t}\mu(t,\boldsymbol{x}) = \frac{\partial}{\partial t}\mathbb{E}\left(Y|T=t,\boldsymbol{X}=\boldsymbol{x}\right)$ to be well-defined whenever $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) > 0$ for any $t \in \mathcal{T}$. This serves as our key technique to bypass the restrictive positivity condition in Assumption 6.1 in the context of continuous treatments. Furthermore, once $\theta(t)$ is identifiable from observable data, we can relate the identification and estimation of the dose-response curve $m(t)$ back to $\theta(t)$ by the fundamental theorem of calculus. All the

quantities on the right-hand sides of (6.2) and (6.3) are estimable from the observed data $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ even without the positivity condition in Assumption 6.1; see Section 6.2.2 for more details.

6.2.1 Motivating Example: Additive Confounding Model

The key running example in this chapter is the additive confounding model defined as:

$$Y^{(t)} = \bar{m}(t) + \eta(\boldsymbol{X}) + \epsilon, \tag{6.4}$$

where $\bar{m}(t)$ is the primary treatment effect of interest, $\eta(\boldsymbol{X})$ is the random effect, and $\epsilon \in \mathbb{R}$ is an independent noise variable with $\mathbb{E}(\epsilon) = 0$.

Under model (6.4) and Assumption 6.2(a), the outcome model becomes $Y = \bar{m}(T) + \eta(\boldsymbol{X}) + \epsilon$. Such an additive form is a common working model in spatial confounding problems (Paciorek, 2010; Schnell and Papadogeorgou, 2020) and is also known as the geoadditive structural equation model (Kammann and Wand, 2003; Thaden and Kneib, 2018; Wiecha and Reich, 2024), where $\boldsymbol{X} \in \mathbb{R}^d$ are the spatial locations (usually with $d = 2$) or other spatially correlated covariates that affect both the treatment T and the outcome Y. We summarize some key properties of the additive confounding model (6.4) in the following proposition, whose proof is in Section E.2.2. In particular, Proposition 6.2(c,d) verifies Assumption 6.4 under model (6.4).

Proposition 6.2 (Properties of the additive confounding model)**.** *Let* $\mu(t, \boldsymbol{x}) = \mathbb{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x}) = \bar{m}(t) + \eta(\boldsymbol{x})$ *under the additive confounding model* (6.4)*. Assume that* $\bar{m}(t)$ *is differentiable for* P_T*-almost every* $t \in \mathcal{T}$*. The following results hold for any* $t \in \mathcal{T}$ *under Assumptions 6.2 and 6.3:*

(a) $\mathbb{E}[\mu(t, \boldsymbol{X})] = \bar{m}(t)$ *when* $\mathbb{E}[\eta(\boldsymbol{X})] = 0$.

(b) $\theta(t) \neq \frac{d}{dt}\mathbb{E}[\mu(t, \boldsymbol{X})|T = t]$ *when* $\frac{d}{dt}\mathbb{E}[\eta(\boldsymbol{X})|T = t] \neq 0$.

(c) $\theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{X})\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{X})\middle|T = t\right] = \theta_C(t)$.

(d) $\mathbb{E}(Y) = \mathbb{E}\left[\mu(T, \boldsymbol{X})\right] = \mathbb{E}\left[m(T)\right]$, *where* $m(t) = \mathbb{E}\left[Y^{(t)}\right]$.

Remark 6.2 (Identification of $m(t)$ under model (6.4))**.** *When the additive confounding model* (6.4) *holds, the dose-response curve* $m(t) = \mathbb{E}[Y^{(t)}]$ *is indeed identifiable under Assumption 6.2 via the standard G-computation formula* $\mathbb{E}\left[\mu(t, \boldsymbol{X})\right] = \bar{m}(t) + \mathbb{E}\left[\eta(\boldsymbol{X})\right]$ *(Robins, 1986; Gill and Robins, 2001), leading to the regression adjustment (RA) estimator of* $m(t)$ *as:*

$$\widehat{m}_{RA}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(t, \boldsymbol{X}_i), \tag{6.5}$$

where $\widehat{\mu}(t, \boldsymbol{x})$ *is any consistent (additive) regression estimator (Stone, 1985) of the conditional mean outcome function* $\mu(t, \boldsymbol{x}) = \mathbb{E}\left(Y | T = t, \boldsymbol{X} = \boldsymbol{x}\right)$. *However, if the positivity condition in Assumption 6.1 fails at certain points* $(t, \boldsymbol{X}_i) \in \mathcal{T} \times \mathcal{X}$, $\widehat{m}_{RA}(t)$ *may become unstable or even inconsistent. Our identification formula* (6.3) *thus offers an alternative approach to identifying* $m(t)$ *under model* (6.4). *The associated integral estimator* (6.8) *in Section 6.2.2 aggregates derivative effects over* $\mathcal{T}$*, producing a more stable and consistent estimate.*

In addition to the additive confounding model (6.4), Assumption 6.4 holds under the following more general model:

$$Y^{(t)} = \bar{m}(t) + \eta(\boldsymbol{X}_1, \boldsymbol{X}_2) + q(\boldsymbol{X}_1, t) \cdot \Phi(\boldsymbol{X}_2) + \epsilon \quad \text{with} \quad \mathbb{E}\left[\Phi(\boldsymbol{X}_2) \middle| \boldsymbol{X}_1, T\right] = 0 \quad \text{almost surely}, \tag{6.6}$$

where the covariate vector is $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^T \in \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^d$, and $\eta : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}, q : \mathcal{X}_1 \times \mathcal{T} \to \mathbb{R}, \Phi : \mathcal{X}_2 \to \mathbb{R}$ are deterministic functions. We further assume that $t \mapsto \bar{m}(t)$ and $t \mapsto q(\boldsymbol{x}_1, t)$ are continuously differentiable with respect to t for $\mathrm{P}_{\boldsymbol{X}_1}$-almost every $\boldsymbol{x}_1$, and their derivatives satisfy the integrability conditions needed to interchange differentiation and expectation.

6.2.2 Proposed Integral Estimator of $m(t)$

Given the observed data $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$, our estimation strategy for addressing violations of the positivity condition in Assumption 6.1 is motivated by three critical insights.

- **Insight 1: Local estimation on the observed support.** Rather than estimating $\mu(t, \boldsymbol{x})$ and its partial derivative $\frac{\partial}{\partial t}\mu(t, \boldsymbol{x})$ over the entire product space $\mathcal{T} \times \mathcal{X}$, it suffices to estimate these functions over the joint support of $(T, \boldsymbol{X})$, where the observed data provide sufficient local information. Under appropriate smoothness, local treatment variation, and other regularity conditions, $\frac{\partial}{\partial t}\mu(t, \boldsymbol{x})$ can be consistently estimated without extrapolating the outcome regression to unsupported treatment-covariate pairs.

- **Insight 2: Estimation of $\theta(t)$ via $\theta_C(t)$.** The localized form $\theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \boldsymbol{X})\middle| T = t\right]$ averages $\frac{\partial}{\partial t}\mu(t, \boldsymbol{x})$ with respect to the conditional distribution $\mathrm{P}_{\boldsymbol{X}|T=t}$. Consequently, estimation of $\theta_C(t)$ requires sufficient accuracy of the derivative estimator under this conditional distribution, rather than uniformly over the entire marginal covariate support $\mathcal{X}$.

- **Insight 3: Integral relation between $m(t)$ and $\theta(t)$.** The integral relation (6.3) suggests that, if $\theta(\widetilde{t})$ can be estimated accurately over the treatment values $\widetilde{t}$ connecting the observed treatment levels T_i to the target level t, then $m(t)$ can be estimated without directly evaluating $\mu(t, \boldsymbol{X}_i)$ at treatment-covariate pairs outside the joint support, even when $p_{T,\boldsymbol{X}}(t, \boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \mathcal{X}$.

Based on these insights, the RA estimator of $\theta(t)$ under model (6.4) without assuming the positivity condition is given by

$$\widehat{\theta}_{\mathrm{C,RA}}(t) = \int \widehat{\beta}(t, \boldsymbol{x})\, d\widehat{F}_{\boldsymbol{X}|T}(\boldsymbol{x}|t), \tag{6.7}$$

where $\widehat{\beta}(t, \boldsymbol{x})$ and $\widehat{F}_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ are consistent estimators of $\beta(t, \boldsymbol{x}) = \frac{\partial}{\partial t}\mu(t, \boldsymbol{x})$ and the conditional cumulative distribution function (CDF) $\mathrm{P}_{\boldsymbol{X}|T}(\boldsymbol{x}|t) := F_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$, respectively. By (6.3), the integral RA estimator of $m(t)$ under model (6.4) can be written as:

$$\widehat{m}_{\mathrm{C,RA}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i + \int_{\widetilde{t}=T_i}^{\widetilde{t}=t} \widehat{\theta}_{\mathrm{C,RA}}(\widetilde{t})\, d\widetilde{t}\right]. \tag{6.8}$$

Both estimators (6.7) and (6.8) are consistent even when the positivity condition (Assumption 6.1) is violated (Zhang et al., 2024).

6.3 *Estimation Issues of IPW Estimators Under the Additive Confounding Model* (6.4)

Although the causal quantities $m(t)$ and $\theta(t)$ are identifiable under the additive confounding model (6.4), the IPW formulae

$$\lim_{h\to 0}\mathbb{E}\left[\frac{Y\cdot K\left(\frac{T-t}{h}\right)}{h\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right]=\mathbb{E}\left[\mu(t,\boldsymbol{X})\right]\quad\text{and}\quad\lim_{h\to 0}\mathbb{E}\left[\frac{Y\left(\frac{T-t}{h}\right)K\left(\frac{T-t}{h}\right)}{\kappa_2 h^2\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right]=\mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{X})\right],\tag{6.9}$$

are typically biased without positivity due to the support discrepancy between $\mathcal{X}(t)$ from the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ for $t\in\mathcal{T}$ and $\mathcal{X}$ from the marginal density $p_{\boldsymbol{X}}(\boldsymbol{x})$.

6.3.1 Assumptions

We introduce some regularity conditions for our theoretical analysis. Let $\mathcal{J}\subset\mathcal{T}\times\mathcal{X}$ be the support of the joint density $p_{T,\boldsymbol{X}}(t,\boldsymbol{x})$, $\mathcal{J}^\circ$ be the interior of $\mathcal{J}$, and $\partial\mathcal{J}$ be the boundary of $\mathcal{J}$.

Assumption 6.5 (Differentiability of the conditional mean outcome function)**.** *For any $(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}$ so that $\mu(t,\boldsymbol{x})=\mathbb{E}(Y|T=t,\boldsymbol{X}=\boldsymbol{x})$ is well-defined, it holds that*

(a) $\mu(t,\boldsymbol{x})$ is at least three times continuously differentiable with respect to t.

(b) $\mu(t,\boldsymbol{x})$ and all of its partial derivatives are uniformly bounded on $\mathcal{T}\times\mathcal{X}$.

(c) There exist constants $\sigma,c_1>0$ such that $\mathrm{Var}(Y|T=t,\boldsymbol{X}=\boldsymbol{x})>\sigma^2$ and $\mathbb{E}|Y|^{2+c_1}<\infty$.

Assumption 6.6 (Differentiability of the density functions)**.** *For any $(t,\boldsymbol{x})\in\mathcal{J}$, it holds that*

(a) The joint density $p_{T,\boldsymbol{X}}(t,\boldsymbol{x})$ and the conditional density $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ are at least three times continuously differentiable with respect to t.

(b) $p_{T,\boldsymbol{X}}(t,\boldsymbol{x})$, $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$, $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$, as well as all of the partial derivatives of $p_{T,\boldsymbol{X}}(t,\boldsymbol{x})$ and $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ are bounded and continuous up to the boundary $\partial\mathcal{J}$.

(c) *The support* $\mathcal{T}$ *of the marginal density* $p_T(t)$ *is compact and* $p_T(t)$ *is uniformly bounded away from 0 within* $\mathcal{T}$.

Assumption 6.7 (Regular kernel conditions)**.** *A kernel function* $K : \mathbb{R} \to [0, \infty)$ *is bounded and compactly supported on* $[-1, 1]$ *with* $\int_{\mathbb{R}} K(t)\, dt = 1$ *and* $K(t) = K(-t)$*. In addition, it holds that*

(a) $\kappa_j := \int_{\mathbb{R}} u^j K(u)\, du < \infty$ *and* $\nu_j := \int_{\mathbb{R}} u^j K^2(u)\, du < \infty$ *for all* $j = 1, 2, ...$

(b) K *is a second-order kernel,* i.e., $\kappa_1 = 0$ *and* $\kappa_2 > 0$.

(c) $\mathcal{K} = \left\{ t' \mapsto \left(\frac{t'-t}{h}\right)^{k_1} K\left(\frac{t'-t}{h}\right) : t \in \mathcal{T}, h > 0, k_1 = 0, 1 \right\}$ *is a bounded VC-type class of measurable functions on* $\mathbb{R}$.

Assumptions 6.5 and 6.6 are common smoothness conditions for derivative estimation with kernel smoothing methods (Gasser and Müller, 1984; Mack and Müller, 1989; Wand and Jones, 1994; Wasserman, 2006). These assumptions can be relaxed by the Hölder continuity condition. The uniform lower bound on $p_T(t)$ within its support $\mathcal{T}$ in Assumption 6.6(c) is only needed when we establish the uniform consistency of our proposed estimators and identify the derivative effect curve $\theta(t)$ when the positivity condition is violated. Assumption 6.7(a,b) are properties of commonly used kernel functions rather than regularity conditions, such as the triangular kernel $K(u) = (1 - |u|)\, \mathbb{1}_{\{|u| \le 1\}}$ and Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)\, \mathbb{1}_{\{|u| \le 1\}}$. Finally, the VC-type condition in Assumption 6.7(c) is only required when we are interested in the uniform consistency of our proposed estimators over $\mathcal{T}$.

6.3.2 Main Inconsistency Results

To examine the biases in (6.9), we can equivalently analyze the following oracle IPW estimators of $m(t)$ and $\theta(t)$ defined as:

$$\widetilde{m}_{\mathrm{IPW}}(t) = \frac{1}{nh} \sum_{i=1}^{n} \frac{Y_i \cdot K\left(\frac{T_i - t}{h}\right)}{p_{T|\boldsymbol{X}}(T_i|\boldsymbol{X}_i)} \quad \text{and} \quad \widetilde{\theta}_{\mathrm{IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{Y_i \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{\kappa_2 \cdot p_{T|\boldsymbol{X}}(T_i|\boldsymbol{X}_i)}, \tag{6.10}$$

where the estimated conditional density $\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ is replaced by the true one $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$.

Proposition 6.3 (Inconsistency of IPW estimators)**.** *Suppose that Assumptions 6.2, 6.3, 6.5, 6.6(c), and 6.7(a-b) hold under the additive confounding model (6.4). Assume also that when the bandwidth h is small, the Lebesgue measure of the symmetric difference set satisfies $|\mathcal{X}(t+uh)\triangle\mathcal{X}(t)| = |[\mathcal{X}(t+uh)\setminus\mathcal{X}(t)]\cup[\mathcal{X}(t)\setminus\mathcal{X}(t+uh)]| = o(1)$ for any $t\in\mathcal{T}$ and $u\in\mathbb{R}$. Then, when h is small, the expectation of $\widetilde{m}_{\mathrm{IPW}}(t)$ in (6.10) is given by*

$$\mathbb{E}\left[\widetilde{m}_{\mathrm{IPW}}(t)\right] = \bar{m}(t)\cdot\rho(t) + \omega(t) + o(1),$$

where $\rho(t) = \mathrm{P}\left(\boldsymbol{X}\in\mathcal{X}(t)\right)$ and $\omega(t) = \mathbb{E}\left[\eta(\boldsymbol{X})\mathbb{1}_{\{\boldsymbol{X}\in\mathcal{X}(t)\}}\right]$. If, in addition, there exists a constant $A_h > 0$ depending on h such that

$$\int_{\mathbb{R}} \mathbb{E}\left\{\left[\bar{m}(t)+\eta(\boldsymbol{X})\right]\left[\mathbb{1}_{\{\boldsymbol{X}\in\mathcal{X}(t+uh)\setminus\mathcal{X}(t)\}} - \mathbb{1}_{\{\boldsymbol{X}\in\mathcal{X}(t)\setminus\mathcal{X}(t+uh)\}}\right]\right\} u\cdot K(u)\,du = O(A_h) \quad (6.11)$$

for any $t\in\mathcal{T}$ and $u\in\mathbb{R}$ when h is small, then the expectation of $\widetilde{\theta}_{\mathrm{IPW}}(t)$ in (6.9) is given by

$$\mathbb{E}\left[\widetilde{\theta}_{\mathrm{IPW}}(t)\right] = \bar{m}'(t)\cdot\rho(t) + O\left(\frac{A_h}{h}\right).$$

We emphasize that the IPW estimators in (6.10) have two layers of bias. First, if $\frac{A_h}{h}\to 0$ as $h\to 0$ (see also Remark 6.3 below), then the results in Proposition 6.3 imply that

$$\lim_{h\to 0}\mathbb{E}\left[\widetilde{m}_{\mathrm{IPW}}(t)\right] = \lim_{h\to 0}\mathbb{E}\left[\frac{Y\cdot K\left(\frac{T-t}{h}\right)}{h\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right] = \bar{m}(t)\cdot\rho(t)+\omega(t)\neq m(t),$$
$$\lim_{h\to 0}\mathbb{E}\left[\widetilde{\theta}_{\mathrm{IPW}}(t)\right] = \lim_{h\to 0}\mathbb{E}\left[\frac{Y\left(\frac{T-t}{h}\right)K\left(\frac{T-t}{h}\right)}{h^2\cdot\kappa_2\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right] = \bar{m}'(t)\cdot\rho(t)\neq\theta(t),$$

where we recall that $m(t) = \bar{m}(t)+\mathbb{E}\left[\eta(\boldsymbol{X})\right]$ and $\theta(t) = \bar{m}'(t)$ from (6.4). Second, if $\frac{A_h}{h}$ does not converge to 0, then the bias of $\widetilde{\theta}_{\mathrm{IPW}}(t)$ will be larger and can even diverge to infinity as $h\to 0$. In reality, the estimation biases or inconsistencies of IPW estimators in (6.10) are due to the discrepancy between the conditional support $\mathcal{X}(t)$ of $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ and the marginal support $\mathcal{X}$ of $p_{\boldsymbol{X}}(\boldsymbol{x})$. To correct for the bias of IPW estimators, it is necessary to address the geometric discrepancy, a solution to which will be elaborated upon in Section 6.4.

Finally, since both RA and IPW estimators cannot be used to identify and estimate $m(t)$ and $\theta(t)$ due to identification and inconsistency issues, the previously studied DR estimators for $m(t)$ and $\theta(t)$ would not provide meaningful results as well when positivity is violated.

Remark 6.3. *The condition (6.11) is best viewed as a regularity property rather than a substantive modeling assumption. This is because as $h \to 0$, the differences between two sets $\mathcal{X}(t+uh) \setminus \mathcal{X}(t)$ and $\mathcal{X}(t) \setminus \mathcal{X}(t+uh)$ shrink to 0 for any $t \in \mathcal{T}$ and $u \in \mathbb{R}$. Additionally, when the expectation in (6.11) is independent of u, one can deduce by the second-order kernel property of K that the left-hand side of (6.11) is 0. Hence, as $h \to 0$, the left-hand side of (6.11) should converge to 0 at a certain rate depending on h.*

6.4 Nonparametric Inference on $\theta(t)$ Under Positivity Violations

In this section, we present our solution for addressing the estimation biases of IPW estimators for the dose-response curve $m(t)$ and its derivative $\theta(t)$, as described in Section 6.3, when the positivity condition in Assumption 6.1 is violated. Specifically, our proposed IPW and DR estimators for $\theta(t)$ under the additive confounding model (6.4) rely on a consistent estimation of the interior region of the support of the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$. Our approach establishes a connection between the classical support estimation problem and a contemporary causal inference challenge, namely the dose-response curve estimation problem.

6.4.1 Bias-Corrected IPW and DR Estimators of $\theta(t)$

Recall from (6.10) and Proposition 6.3 that the oracle IPW estimator of $\theta(t)$ is the sample average of the IPW quantity $\Xi_t(Y,T,\boldsymbol{X}) = \frac{Y\left(\frac{T-t}{h}\right)K\left(\frac{T-t}{h}\right)}{h^2\cdot\kappa_2\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}$, and it is biased for estimating the quantity of interest $\theta(t) = \bar{m}'(t)$ even under model (6.4). In particular, $\mathbb{E}\left[\Xi_t(Y,T,\boldsymbol{X})\right]$ converges to $\bar{m}'(t)\cdot\rho(t)$ as $h \to 0$ under some mild regularity conditions, where $\rho(t) = \mathrm{P}\left(\boldsymbol{X} \in \mathcal{X}(t)\right)$ for any $t \in \mathcal{T}$. The first step toward removing the bias of $\mathbb{E}\left[\Xi_t(Y,T,\boldsymbol{X})\right]$ is to decouple the quantity of interest $\theta(t) = \bar{m}'(t)$ from the nuisance function $\rho(t)$. To this end,

we consider a modified IPW quantity defined as:

$$\widetilde{\Xi}_t(Y,T,\boldsymbol{X}) = \Xi_t(Y,T,\boldsymbol{X}) \cdot \frac{p_{\boldsymbol{X}|T}(\boldsymbol{X}|t)}{p_{\boldsymbol{X}}(\boldsymbol{X})} = \frac{Y\left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right) p_{\boldsymbol{X}|T}(\boldsymbol{X}|t)}{h^2 \cdot \kappa_2 \cdot p_{T,\boldsymbol{X}}(T,\boldsymbol{X})}, \tag{6.12}$$

in which we multiply the original IPW quantity $\Xi_t(Y,T,\boldsymbol{X})$ by a density ratio $\frac{p_{\boldsymbol{X}|T}(\boldsymbol{X}|t)}{p_{\boldsymbol{X}}(\boldsymbol{X})}$. The following proposition demonstrates that the remaining bias in $\mathbb{E}\left[\widetilde{\Xi}_t(Y,T,\boldsymbol{X})\right]$ can be disentangled from the quantity of interest $\theta(t) = \bar{m}'(t)$ in an additive form.

Proposition 6.4. *Suppose that Assumptions 6.2, 6.3, 6.5, 6.6(c), and 6.7(a-b) hold under the additive confounding model (6.4). Then, when the bandwidth h is small, the expectation of the modified IPW quantity (6.12) is given by*

$$\begin{aligned}\mathbb{E}\left[\widetilde{\Xi}_t(Y,T,\boldsymbol{X})\right] &= \bar{m}'(t) + O(h^2) \\ &\quad + \int_{\mathbb{R}} \mathbb{E}\left\{[\bar{m}(t+uh)+\eta(\boldsymbol{X})]\left[\mathbb{1}_{\{\boldsymbol{X}\in\mathcal{X}(t+uh)\setminus\mathcal{X}(t)\}} - \mathbb{1}_{\{\boldsymbol{X}\in\mathcal{X}(t)\setminus\mathcal{X}(t+uh)\}}\right] \middle| T=t\right\} u\cdot K(u)\,du.\end{aligned}$$

Proposition 6.4 reveals that the estimation bias of the modified IPW quantity (6.12) results from the support discrepancy between $\mathcal{X}(t)$ and the integration range $\mathcal{X}(t+uh)$ for a given integration variable $u \in \mathbb{R}$; see Figure 6.1 for an illustration. As shown in Proposition 6.3, this additive bias may not always shrink at the rate $O(h^2)$ as $h \to 0$. To further reduce the bias of the modified IPW quantity (6.12) to $O(h^2)$ without assuming positivity, we address the support discrepancy of (6.12) by restricting the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ to its interior region, defining it as $p_\zeta(\boldsymbol{x}|t)$, and refining (6.12) as:

$$\widetilde{\Xi}_{t,\zeta}(Y,T,\boldsymbol{X}) = \frac{Y\left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right) p_\zeta(\boldsymbol{X}|t)}{h^2 \cdot \kappa_2 \cdot p_{T,\boldsymbol{X}}(T,\boldsymbol{X})}. \tag{6.13}$$

Essentially, the only requirement for defining the ζ-interior conditional density $p_\zeta(\boldsymbol{x}|t)$ is that its support satisfies the following condition:

$$\{\boldsymbol{x}\in\mathcal{X}(t) : p_\zeta(\boldsymbol{x}|t) > 0\} \subset \mathcal{X}(t+\delta) \quad \text{for any} \quad \delta\in[-h,h]. \tag{6.14}$$

We propose two approaches for defining $p_\zeta(\boldsymbol{x}|t)$ and leave other options to interested readers.

1. Support Shrinking Approach: Let $\mathcal{X}(t) \ominus \zeta = \left\{\boldsymbol{x} \in \mathcal{X}(t) : \inf_{\boldsymbol{x}_1 \in \partial\mathcal{X}(t)} ||\boldsymbol{x} - \boldsymbol{x}_1||_2 \geq \zeta\right\}$ denote the set of interior points of $\mathcal{X}(t)$ that are at least a distance ζ away from the boundary of $\mathcal{X}(t)$. Then, for a tuning parameter $\zeta > 0$, we define the ζ-interior conditional density by

$$p_\zeta(\boldsymbol{x}|t) = \frac{p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \cdot \mathbb{1}_{\{\boldsymbol{x} \in \mathcal{X}(t) \ominus \zeta\}}}{\int_{\mathcal{X}(t)\ominus\zeta} p_{\boldsymbol{X}|T}(\boldsymbol{x}_1|t)\, d\boldsymbol{x}_1} \propto p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \cdot \mathbb{1}_{\{\boldsymbol{x} \in \mathcal{X}(t)\ominus\zeta\}}. \tag{6.15}$$

This interior density is indeed the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ restricted to the interior of its support $\mathcal{X}(t)$. Its estimator $\widehat{p}_\zeta(\boldsymbol{x}|t)$ can be constructed using a support estimator $\widehat{\mathcal{X}}(t)$ and constraining the conditional density estimator $\widehat{p}_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ within the region $\widehat{\mathcal{X}}(t) \ominus \zeta$.

2. Level Set Approach: Let $\mathcal{L}_\zeta(t) = \left\{\boldsymbol{x} \in \mathcal{X}(t) : p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \geq \zeta\right\}$ be the ζ-upper level set of the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$. Then, we define the ζ-interior conditional density as:

$$p_\zeta(\boldsymbol{x}|t) = \frac{p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \cdot \mathbb{1}_{\left\{\boldsymbol{x} \in \mathcal{L}_\zeta(t)\right\}}}{\int_{\mathcal{L}_\zeta(t)} p_{\boldsymbol{X}|T}(\boldsymbol{x}_1|t)\, d\boldsymbol{x}_1} \propto p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \cdot \mathbb{1}_{\left\{\boldsymbol{x} \in \mathcal{L}_\zeta(t)\right\}}. \tag{6.16}$$

The level set approach restricts the conditional density $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ to the high-density region, which is generally located away from the support boundary. We may construct the estimator $\widehat{p}_\zeta(\boldsymbol{x}|t)$ using a level set estimator $\widehat{\mathcal{L}}_\zeta(t) = \left\{\boldsymbol{x} \in \mathcal{X}(t) : \widehat{p}_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \geq \zeta\right\}$ and constraining $\widehat{p}_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ to $\widehat{\mathcal{L}}_\zeta(t)$.

We further specialize condition (6.14) for the above two approaches by introducing the following smoothness condition on the conditional support $\mathcal{X}(t)$.

Assumption 6.8 (Smoothness condition on $\mathcal{X}(t)$)**.** *For any $\delta \in \mathbb{R}$ and $t \in \mathcal{T}$, there exists an absolute constant $A_0 > 0$ such that either (i) $\mathcal{X}(t) \ominus (A_0|\delta|) \subset \mathcal{X}(t+\delta)$ for the support shrinking approach or (ii) $\mathcal{L}_{A_0|\delta|}(t) \subset \mathcal{X}(t+\delta)$ for the level set approach.*

To some extent, Assumption 6.8 can be viewed as a Lipschitz condition of the conditional support $\mathcal{X}(t)$. It can be satisfied when the Euclidean norm of the gradient $\left|\left|\nabla_{\boldsymbol{x}} p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)\right|\right|_2$ is bounded away from 0 at the boundary of $\mathcal{X}(t)$ (Cadre, 2006) or under the standardness condition in Cuevas (1990); Cuevas and Fraiman (1997). This assumption allows us to ignore the boundary discrepancy as long as we do not evaluate our IPW quantity (6.13) near the boundary; see Figure 6.1 for a graphical illustration.

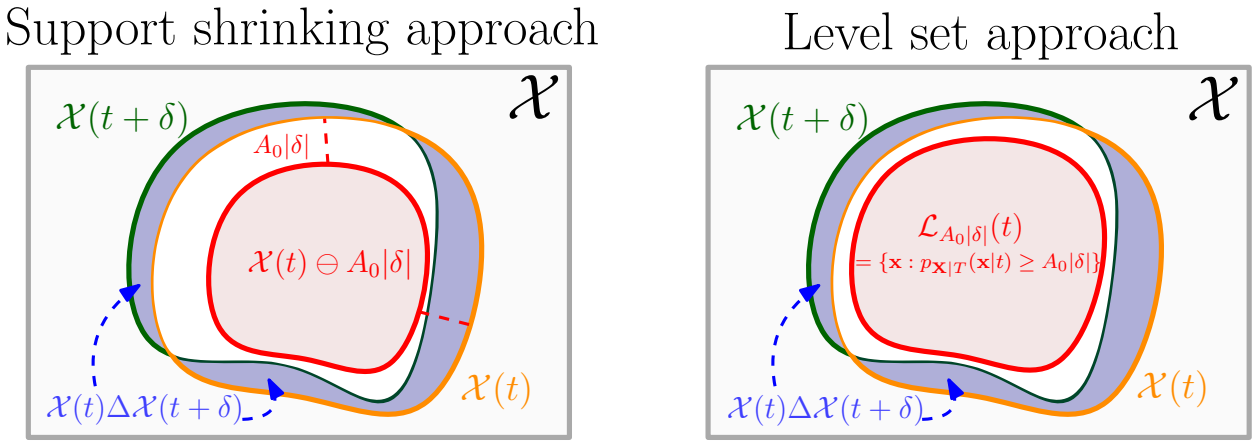


Figure 6.1: Graphical illustrations of the support discrepancy between $\mathcal{X}(t)$ and $\mathcal{X}(t+\delta)$ for $t \in \mathcal{T}$ as well as Assumption 6.8, where δ can equal $uh \in \mathbb{R}$.

Proposition 6.5. *Suppose that Assumptions 6.2, 6.3, 6.5, 6.6(c), 6.7(a-b), and 6.8 hold under the additive confounding model (6.4). Then, when* $h \in \left(0, \frac{\zeta}{A_0}\right)$ *is small with* $A_0 > 0$ *defined in Assumption 6.8, the expectation of the modified IPW quantity (6.13) is given by* $\mathbb{E}\left[\widetilde{\Xi}_{t,\zeta}(Y, T, \boldsymbol{X})\right] = \bar{m}'(t) + \frac{h^2\kappa_4}{6\kappa_2} \cdot \bar{m}^{(3)}(t) + O\left(h^3\right)$.

This result demonstrates that the expectation of our newly modified IPW quantity $\widetilde{\Xi}_{t,\zeta}(Y, T, \boldsymbol{X})$ in (6.13) converges to the quantity of interest $\theta(t) = \bar{m}'(t)$ at the standard rate $O(h^2)$ as $h \to 0$ under the additive confounding model (6.4). Notice that the tuning parameter $\zeta = \zeta_n > 0$ in (6.15) is allowed to converge to 0 as $n \to \infty$, provided that the condition $|\delta| \leq h = h_n < \frac{\zeta_n}{A_0}$ holds under Assumption 6.8. In this regime, the population on which the causal parameter $\theta(t)$ is defined remains asymptotically unchanged.

Given this newly modified IPW quantity (6.13), we propose the bias-corrected IPW estimator of $\theta(t)$ without the positivity condition as:

$$\widehat{\theta}_{\mathrm{C,IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{Y_i \left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right) \widehat{p}_\zeta(\boldsymbol{X}_i|t)}{\kappa_2 \cdot \widehat{p}(T_i, \boldsymbol{X}_i)}, \tag{6.17}$$

where $\widehat{p}(t, \boldsymbol{x})$ is a consistent estimator of the joint density $p_{T,\boldsymbol{X}}(t, \boldsymbol{x})$ and $\widehat{p}_\zeta(\boldsymbol{x}|t)$ is an estimated ζ-interior conditional density.

Finally, we combine the modified RA estimator (6.7) with our bias-corrected IPW estimator (6.17) to propose our bias-corrected DR estimator of $\theta(t)$ as:

$$\begin{aligned}\widehat{\theta}_{\mathrm{C,DR}}(t) = \frac{1}{nh^2}\sum_{i=1}^{n}\frac{\left(\frac{T_i-t}{h}\right)K\left(\frac{T_i-t}{h}\right)\widehat{p}_{\zeta}(\boldsymbol{X}_i|t)}{\kappa_2\cdot\widehat{p}(T_i,\boldsymbol{X}_i)}\left[Y_i-\widehat{\mu}(t,\boldsymbol{X}_i)-(T_i-t)\cdot\widehat{\beta}(t,\boldsymbol{X}_i)\right] \\ +\int\widehat{\beta}(t,\boldsymbol{x})\cdot\widehat{p}_{\zeta}(\boldsymbol{x}|t)\,d\boldsymbol{x}.\end{aligned} \tag{6.18}$$

For the RA component of $\widehat{\theta}_{\mathrm{C,DR}}(t)$, we replace the original conditional CDF estimator $\widehat{F}_{\boldsymbol{X}|T}$ in (6.7) with the estimated ζ-interior conditional density $\widehat{p}_{\zeta}$. This modification is necessary because the IPW component of $\widehat{\theta}_{\mathrm{C,DR}}(t)$ is defined through $\widehat{p}_{\zeta}$. Both the RA and IPW components need to match in the definition of $\widehat{\theta}_{\mathrm{C,DR}}(t)$ for consistency.

As shown by Theorem 6.6 in Section 6.4.2, $\widehat{\theta}_{\mathrm{C,DR}}(t)$ enjoys a doubly robust inference property when positivity fails under the additive confounding model (6.4). Furthermore, both $\widehat{\theta}_{\mathrm{C,DR}}(t)$ and $\widehat{\theta}_{\mathrm{C,IPW}}(t)$ remain consistent for $\theta(t)$ when positivity holds, because the (population) factor $\frac{K\left(\frac{T-t}{h}\right)\cdot p_{\zeta}(\boldsymbol{X}|t)}{p_{T,\boldsymbol{X}}(T,\boldsymbol{X})}$ in (6.17) reduces to $\frac{K\left(\frac{T-t}{h}\right)}{p_{T|\boldsymbol{X}}(t|\boldsymbol{X})}$ asymptotically as $\zeta\to 0$ under positivity. Therefore, $\widehat{\theta}_{\mathrm{C,DR}}(t)$ with nonparametric nuisance function estimation enjoys two layers of robustness: double robustness against model misspecification under positivity, and robustness to violations of positivity.

Remark 6.4. *While both the support shrinking and level set approaches are valid, we recommend the level set approach in practice, as support estimation is notoriously difficult in nonparametric statistics (Devroye and Wise, 1980). Additionally, selecting an appropriate ζ for the support shrinking method is nontrivial. In contrast, level set estimation has been studied for decades (Cuevas and Fraiman, 1997; Cadre, 2006), and the threshold can be chosen,* e.g., *as $\zeta = 0.5\cdot\max\left\{\widehat{p}_{\boldsymbol{X}|T}(\boldsymbol{X}_i|t) : i = 1,...,n\right\}$. In practice, $\widehat{\theta}_{\mathrm{C,DR}}(t)$ is fairly insensitive to a broad range of ζ values. Users can adjust the multiplier 0.5 in this rule, where smaller values increase the effective sample size but raise the risk of violating condition (6.14). Determining an optimal ζ that trades off finite-sample precision and violation of (6.14) is beyond the scope of this chapter.*

6.4.2 Asymptotic Theory

The following theorem summarizes the asymptotic properties of our DR estimator (6.18) of $\theta(t)$ under the additive confounding model (6.4) without assuming the positivity condition.

Theorem 6.6 (Asymptotic properties of $\widehat{\theta}_{\mathrm{C,DR}}(t)$ without positivity)**.** *Suppose that Assumptions 6.2, 6.3, 6.5, 6.6, 6.7, and 6.8 hold under the additive confounding model (6.4), and the support $\mathcal{X} \subset \mathbb{R}^d$ of the marginal density $p_{\boldsymbol{X}}$ is compact. In addition, $\widehat{\mu}, \widehat{\beta}, \widehat{p}_\zeta, \widehat{p}$ are constructed on a data sample independent of $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$. For any fixed $t \in \mathcal{T}$, we let $\bar{\mu}(t, \boldsymbol{x})$, $\bar{\beta}(t, \boldsymbol{x})$, $\bar{p}_\zeta(\boldsymbol{x}|t)$, and $\bar{p}(t, \boldsymbol{x})$ be fixed bounded functions to which $\widehat{\mu}(t, \boldsymbol{x})$, $\widehat{\beta}(t, \boldsymbol{x})$, $\widehat{p}_\zeta(\boldsymbol{x}|t)$, and $\widehat{p}(t, \boldsymbol{x})$ converge. If, in addition, we assume that*

(a) $\bar{p}, \bar{p}_\zeta$ satisfy Assumptions 6.6 and 6.8 as well as $\sqrt{nh^3}\,||\widehat{p}_\zeta(\boldsymbol{X}|t) - \bar{p}_\zeta(\boldsymbol{X}|t)||_{L_2} = o(1)$;

(b) either (i) $\bar{\mu} = \mu$ and $\bar{\beta} = \beta$ or (ii) $\bar{p} = p$;

(c) $\sqrt{nh}\left[||\widehat{p}_\zeta(\boldsymbol{X}|t) - \bar{p}_\zeta(\boldsymbol{X}|t)||_{L_2} + \sup_{|u-t|\le h} ||\widehat{p}(u, \boldsymbol{X}) - p(u, \boldsymbol{X})||_{L_2}\right]\left[||\widehat{\mu}(t, \boldsymbol{X}) - \mu(t, \boldsymbol{X})||_{L_2} + h\left|\left|\widehat{\beta}(t, \boldsymbol{X}) - \beta(t, \boldsymbol{X})\right|\right|_{L_2}\right] = o_P(1)$,

then

$$\sqrt{nh^3}\left[\widehat{\theta}_{\mathrm{C,DR}}(t) - \theta(t)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\phi_{C,h,t}\left(Y_i, T_i, \boldsymbol{X}_i; \bar{\mu}, \bar{\beta}, \bar{p}, \bar{p}_\zeta\right) + \sqrt{h^3}\left[\int \bar{\beta}(t, \boldsymbol{x}) \cdot \bar{p}_\zeta(\boldsymbol{x}|t)\, d\boldsymbol{x} - \theta(t)\right]\right\} + o_P(1)$$

when $nh^7 \to c_3$ for some finite number $c_3 \ge 0$, where

$$\phi_{C,h,t}\left(Y, T, \boldsymbol{X}; \bar{\mu}, \bar{\beta}, \bar{p}, \bar{p}_\zeta\right) = \frac{\left(\frac{T-t}{h}\right) K\left(\frac{T-t}{h}\right) \cdot \bar{p}_\zeta(\boldsymbol{X}|t)}{\sqrt{h} \cdot \kappa_2 \cdot \bar{p}(T, \boldsymbol{X})} \cdot \left[Y - \bar{\mu}(t, \boldsymbol{X}) - (T-t) \cdot \bar{\beta}(t, \boldsymbol{X})\right].$$

Furthermore,

$$\sqrt{nh^3}\left[\widehat{\theta}_{\mathrm{C,DR}}(t) - \theta(t) - h^2 B_{C,\theta}(t)\right] \xrightarrow{d} \mathcal{N}\left(0, V_{C,\theta}(t)\right)$$

with $V_{C,\theta}(t) = \lim_{h\to 0} \mathbb{E}\left[\phi^2_{C,h,t}\left(Y, T, \boldsymbol{X}; \bar{\mu}, \bar{\beta}, \bar{p}, \bar{p}_\zeta\right)\right]$ and

$$B_{C,\theta}(t) = \begin{cases} \frac{\kappa_4}{6\kappa_2} \int \left\{ \frac{3\frac{\partial}{\partial t} p(t,\boldsymbol{x}) \cdot \bar{m}''(t) + p(t,\boldsymbol{x})\left[\bar{m}^{(3)}(t) - 3\frac{\partial}{\partial t}\log \bar{p}(t,\boldsymbol{x}) \cdot \bar{m}''(t)\right]}{\bar{p}(t,\boldsymbol{x})} \right\} \bar{p}_\zeta(\boldsymbol{x}|t)\, d\boldsymbol{x} & \text{when } \bar{\mu} = \mu \text{ and } \bar{\beta} = \beta, \\ \frac{\kappa_4}{6\kappa_2} \cdot \bar{m}^{(3)}(t) & \text{when } \bar{p} = p. \end{cases}$$

We emphasize that the limiting quantity $\bar{p}_\zeta$ of $\widehat{p}_\zeta$ in Theorem 6.6 need not be the true interior conditional density p_ζ, but may be a sufficiently smooth surrogate. A fast-rate estimator of $\bar{p}_\zeta$ can be obtained, for example, by (i) using a fixed-bandwidth kernel density estimator (KDE), in which case $\bar{p}_\zeta$ corresponds to the quantity implied by the smoothed density $\mathbb{E}[\widehat{p}_w]$ with $\widehat{p}_w$ being the KDE with bandwidth w, or (ii) adopting a parametric model for $\widehat{p}_\zeta(\boldsymbol{x}|t)$, where $\bar{p}_\zeta$ is defined as the Kullback-Leibler projection of the true conditional density $p(\boldsymbol{x}|t)$ onto the parametric family. Both approaches achieve a $\sqrt{n}$ rate of convergence to their respective targets. Thus, although the typical rate of convergence for the conditional interior density function estimator $\widehat{p}_\zeta$ to the true interior conditional density p_ζ is of order $O\left(n^{-\frac{2}{5+d}}\right)$ (Cuevas and Fraiman, 1997; Tsybakov, 1997), which appears slower than the requirement $\sqrt{nh}\,||\widehat{p}_\zeta(\boldsymbol{X}|t) - \bar{p}_\zeta(\boldsymbol{X}|t)||_{L_2} = o_P(1)$, Theorem 6.6 remains applicable because we only need the convergence toward $\bar{p}_\zeta$, not to the true p_ζ.

Note that $V_{C,\theta}(t)$ in Theorem 6.6 can be estimated by

$$\widehat{V}_{C,\theta}(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\phi_{C,h,t}\left(Y_i, T_i, \boldsymbol{X}_i; \widehat{\mu}, \widehat{\beta}, \widehat{p}, \widehat{p}_\zeta\right) + \sqrt{h^3}\left[\int \widehat{\beta}(t,\boldsymbol{x})\cdot\widehat{p}_\zeta(\boldsymbol{x}|t)\,d\boldsymbol{x} - \widehat{\theta}_{\mathrm{C,DR}}(t)\right]\right\}^2$$

and choosing the bandwidth h to be of order $O\left(n^{-\frac{1}{5}}\right)$ ensures valid inference. As a corollary, we can plug either IPW (6.17) or DR (6.18) estimators into our integral formula (6.8) to obtain the integral IPW or DR estimators of the dose-response curve $m(t)$ under model (6.4).

6.5 Nearest Feasible Treatment Policy

In this section, we formally define the NFTP target estimand and show that it is identified by the g-formula for the NFTP regime, under consistency and no-unmeasured confounding assumptions but without requiring the positivity condition.

6.5.1 Causal Estimand Induced by NFTP

The central idea of how NFTP resolves the non-identifiability of the static intervention $T = t_0$ in the dose-response curve is to always restrict the treatment assignment within the feasible

set $\mathcal{T}(\boldsymbol{x})$ for each unit with covariates $\boldsymbol{x} \in \mathcal{X}$. At a high level, the NFTP intervention targeted at t_0 is a possibly dynamic treatment regime that assigns to each unit with $\boldsymbol{X} = \boldsymbol{x}$ the closest value $t_0^\star(\boldsymbol{x})$ to t_0 in the support $\mathcal{T}(\boldsymbol{x})$ of the conditional distribution of T given $\boldsymbol{X} = \boldsymbol{x}$. Since T and $\boldsymbol{X}$ take values in $\mathcal{T} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^d$ respectively, both equipped with their Borel σ-fields $\mathcal{B}(\mathbb{R})$ and $\mathcal{B}(\mathbb{R}^d)$, a regular conditional distribution of T given $\boldsymbol{X}$ exists. In the sequel, we fix one such probability kernel

$$\Psi(\boldsymbol{x}, A) = \mathrm{P}(T \in A | \boldsymbol{X} = \boldsymbol{x}), \qquad \text{for any } A \in \mathcal{B}(\mathbb{R}).$$

For every $\boldsymbol{x} \in \mathcal{X}$, we define the conditional treatment support as:

$$\mathcal{T}(\boldsymbol{x}) := \operatorname{supp}(\Psi(\boldsymbol{x}, \cdot)) = \{t \in \mathbb{R} : \Psi(\boldsymbol{x}, (t-\epsilon, t+\epsilon)) > 0 \text{ for every } \epsilon > 0\}. \tag{6.19}$$

While it is defined through a chosen probability kernel for every $\boldsymbol{x}$, $\mathcal{T}(\boldsymbol{x})$ and the regular conditional distribution of T given $\boldsymbol{X}$ are determined by the joint law of $(T, \boldsymbol{X})$ only $\mathrm{P}_{\boldsymbol{X}}$-almost everywhere. Furthermore, $\mathcal{T}(\boldsymbol{x})$ is a nonempty closed subset of $\mathbb{R}$. We use $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$ interchangeably with the fixed kernel $\Psi(\boldsymbol{x}, \cdot)$. When $\Psi(\boldsymbol{x}, \cdot)$ is absolutely continuous with respect to the Lebesgue measure, we write $p_{T|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ for a specified jointly measurable density. In particular, by (6.19), $t \in \mathcal{T}(\boldsymbol{x})$ if and only if $\int_{t-\epsilon}^{t+\epsilon} p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\, du > 0$ for every $\epsilon > 0$. Any modification of $p_{T|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ on a Lebesgue measure-zero set would not change the definition of $\mathcal{T}(\boldsymbol{x})$. We also impose the following uniqueness and measurability assumption to disambiguate the definition of $t_0^\star(\boldsymbol{x})$.

Assumption 6.9 (Unique measurable projection)**.** *For $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{X}$ and a fixed $t_0 \in \mathbb{R}$, the projection set*

$$\Pi(t_0, \boldsymbol{x}) := \operatorname*{arg\,min}_{t \in \mathcal{T}(\boldsymbol{x})} |t - t_0|$$

is a singleton denoted $t_0^\star(\boldsymbol{x})$. Furthermore, defining $t_0^\star(\boldsymbol{x})$ arbitrarily on the exceptional $\mathrm{P}_{\boldsymbol{X}}$-null set, the map $\boldsymbol{x} \mapsto t_0^\star(\boldsymbol{x})$ is measurable.

Given a fixed t_0, under Assumption 6.9, we set $t_0^\star(\boldsymbol{x})$ to any arbitrary but fixed value in $\mathbb{R}$ on the exceptional $\mathrm{P}_{\boldsymbol{X}}$-null set on which the projection is not uniquely defined. This arbitrary

treatment assignment rule on the $\mathrm{P}_{\boldsymbol{X}}$-null set does not affect any population quantity defined below.

Let $Y^{(t_0^\star)}$ denote the potential outcome under the measurable dynamic regime $\boldsymbol{x} \mapsto t_0^\star(\boldsymbol{x})$. Our target estimand, throughout referred to as the NFTP estimand at t_0 (or simply the NFTP estimand when the targeted point is clear), is

$$\bar{m}_0(t_0) := \mathbb{E}\left[Y^{(t_0^\star)}\right]. \tag{6.20}$$

We refer to the map $t_0 \mapsto \bar{m}_0(t_0)$ as the NFTP curve. Note that (6.20) reduces to the classical dose-response curve $t_0 \mapsto m(t_0)$ when positivity holds for all t_0, in the sense that $t_0^\star(\boldsymbol{X}) = t_0$ almost surely. In such a case, the NFTP intervention at each t_0 becomes the static intervention that assigns treatment level t_0 to the entire population.

6.5.2 Positivity-Lean Identification

Throughout this section, we continue to work under Assumption 6.2. Assumption 6.2(a) connects counterfactual and factual outcomes (Cox, 1958; Rubin, 1980), while Assumption 6.2(b) essentially says that within covariate levels, treatment T is as good as randomized. To account for the feasible support, however, we slightly refine Assumption 6.2(b) by requiring that $Y^{(t)} \perp\!\!\!\perp T \mid \boldsymbol{X} = \boldsymbol{x}$ for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$ and every $t \in \mathcal{T}(\boldsymbol{x})$.

For a continuous treatment, the conditional mean $\mathbb{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$ and related functionals are determined only for $\mathrm{P}_{T,\boldsymbol{X}}$-almost every $(t, \boldsymbol{x})$. Thus, we need to impose additional regularity conditions below to disambiguate their definitions and evaluations at any specific, possibly boundary, point $t_0^\star(\boldsymbol{x})$.

Assumption 6.10 (Support-continuous mean identifications)**.** *There exist jointly Borel measurable functions $\mu, \mu^Y : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ satisfying the following conditions.*

(a) $\mu(T, \boldsymbol{X})$ is a version of $\mathbb{E}(Y|T, \boldsymbol{X})$, and $\mu^Y(T, \boldsymbol{X})$ is a version of $\mathbb{E}\left(Y^{(T)}|\boldsymbol{X}\right)$. Moreover, for the measurable dynamic rule $\boldsymbol{x} \mapsto t_0^\star(\boldsymbol{x})$,

$$\mathbb{E}\left[Y^{(t_0^\star)}\middle|\boldsymbol{X}\right] = \mu^Y(t_0^\star(\boldsymbol{X}), \boldsymbol{X}) \quad \textit{almost surely.}$$

(b) *For* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x}$*,* $t \mapsto \mu(t, \boldsymbol{x})$ *and* $t \mapsto \mu^Y(t, \boldsymbol{x})$ *are continuous at* $t_0^\star(\boldsymbol{x})$ *relative to* $\mathcal{T}(\boldsymbol{x})$*. That is, for every sequence* $t_n \in \mathcal{T}(\boldsymbol{x})$ *with* $t_n \to t_0^\star(\boldsymbol{x})$*, we have that* $\mu^Y(t_n, \boldsymbol{x}) \to \mu^Y(t_0^\star(\boldsymbol{x}), \boldsymbol{x})$ *and* $\mu(t_n, \boldsymbol{x}) \to \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})$*.*

Assumption 6.10 selects versions of the conditional means needed for evaluation at the possibly $\mathrm{P}_{T,\boldsymbol{X}}$-null point $t_0^\star(\boldsymbol{X})$. Since $T \in \mathcal{T}(\boldsymbol{X})$ almost surely, Assumption 6.2, together with Assumption 6.10(a), implies $\mu^Y(T, \boldsymbol{X}) = \mu(T, \boldsymbol{X})$ almost surely. Hence, for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$, $\mu^Y(t, \boldsymbol{x}) = \mu(t, \boldsymbol{x})$ for $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$-almost every t. Because $t_0^\star(\boldsymbol{x}) \in \mathcal{T}(\boldsymbol{x})$ and both functions are continuous at $t_0^\star(\boldsymbol{x})$ relative to $\mathcal{T}(\boldsymbol{x})$ by Assumption 6.10(b), it follows that $\mu^Y(t_0^\star(\boldsymbol{x}), \boldsymbol{x}) = \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})$ for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$. In the sequel, μ and μ^Y denote the versions specified in Assumption 6.10. Their values outside the feasible support are immaterial.

Assumptions 6.2 and 6.10, together with the projection condition in Assumption 6.9, yield the following identification of the NFTP estimand. Throughout, we define the feasibility covariate set from the conditional support $\mathcal{T}(\boldsymbol{x})$ as:

$$\mathcal{X}_0(t_0) := \{\boldsymbol{x} \in \mathcal{X} : t_0 \in \mathcal{T}(\boldsymbol{x})\} = \bigcap_{m \geq 1} \left\{ \boldsymbol{x} \in \mathcal{X} : \mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}} \left(T \in \left(t_0 - \frac{1}{m}, t_0 + \frac{1}{m} \right) \right) > 0 \right\}.$$

Proposition 6.7 (Identification of the NFTP estimand)**.** *Suppose Assumptions 6.9, 6.2, and 6.10 hold and* $\mathbb{E}|Y^{(t_0^\star)}| < \infty$*. Then,*

$$\begin{aligned} \bar{m}_0(t_0) := \mathbb{E}\left[Y^{(t_0^\star)}\right] &= \int_{\mathcal{X}} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x}) \, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\ &= \int_{\mathcal{X}_0(t_0)} \mu(t_0, \boldsymbol{x}) \, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) + \int_{\mathcal{X} \setminus \mathcal{X}_0(t_0)} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x}) \, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}). \end{aligned} \tag{6.21}$$

The first line of (6.21) follows directly from the tower property under Assumptions 6.2 and 6.10. This formula is known as the g-computation formula (or Robins' g-formula; Robins 1986, 1987; Gill and Robins 2001) applied to the NFTP intervention at t_0. The second line of (6.21) is derived from the uniqueness in Assumption 6.9 so that $t_0^\star(\boldsymbol{x}) = t_0$ when $\boldsymbol{x} \in \mathcal{X}_0(t_0)$. In particular, we divide the population of interest into two subpopulations, where the first comprises units for whom the target level t_0 is feasible and the second consists of those for whom t_0 is infeasible. This decomposition demonstrates that the identifying g-formula is

a weighted average of the standard g-formula for identifying the outcome mean under the static regime $T = t_0$ in the first subpopulation and the g-formula for identifying the outcome mean under the dynamic regime $\boldsymbol{x} \mapsto t_0^\star(\boldsymbol{x})$ in the second subpopulation.

6.6 Stochastic Tilted Intervention and the NFTP Limit

In this section, we introduce a class of stochastic interventions constructed by tilting the observed conditional treatment law given the covariates towards a given treatment level t_0, and indexed by a tilting parameter $\hbar > 0$. We demonstrate that, under regularity conditions, the tilted interventions in our class converge to the NFTP intervention at t_0 as the tilting parameter tends to 0.

6.6.1 PLPT-Tilted Intervention and Associated Causal Estimand

For each $t_0 \in \mathbb{R}$ and $\hbar \in (0, \infty)$, let $t \mapsto \eta(t; t_0, \hbar)$ be a map satisfying the following condition.

Condition 6.1 (Positivity-lean-point-targeted tilting function)**.** *For each fixed $t_0 \in \mathbb{R}$ and $\hbar > 0$:*

(a) $t \mapsto \eta(t; t_0, \hbar)$ is Borel measurable, finite, and strictly positive on $\mathbb{R}$,

(b) $0 < \int_{\mathcal{T}} \eta(u; t_0, \hbar)\, \mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(du) < \infty$ for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$,

(c) for every $0 \le \rho_1 < \rho_2$, $\inf_{t \in \mathcal{T}: |t - t_0| \le \rho_1} \eta(t; t_0, \hbar) > 0$ for all sufficiently small $\hbar$ and

$$\frac{\sup_{t \in \mathcal{T}: |t - t_0| \ge \rho_2} \eta(t; t_0, \hbar)}{\inf_{t \in \mathcal{T}: |t - t_0| \le \rho_1} \eta(t; t_0, \hbar)} \to 0 \quad \text{as } \hbar \to 0.$$

For any measurable set $A \in \mathcal{B}(\mathbb{R})$, we define a new conditional law $Q_{\hbar, t_0}(\cdot | \boldsymbol{X} = \boldsymbol{x})$ for treatment given $\boldsymbol{X} = \boldsymbol{x}$ as:

$$Q_{\hbar, t_0}(A | \boldsymbol{X} = \boldsymbol{x}) := \frac{\int_A \eta(t; t_0, \hbar)\, \mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(dt)}{\int_{\mathcal{T}} \eta(u; t_0, \hbar)\, \mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(du)}. \tag{6.22}$$

On the fixed exceptional null set with $\int_{\mathcal{T}} \eta(u; t_0, \hbar)\, \mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(du) = 0$, we define $Q_{\hbar, t_0}$ to equal $\Psi = \mathrm{P}_{T|\boldsymbol{X}}$. By definition, $Q_{\hbar, t_0}(\cdot | \boldsymbol{X} = \boldsymbol{x})$ is a tilted transformation of the conditional

distribution $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$ of the observed treatment given the covariates $\boldsymbol{x}$. Condition 6.1(c) also implies that for each fixed t_0, the tilting function $t \mapsto \eta(t; t_0, \hbar)$ increasingly favors treatment values close to the target level t_0 as $\hbar \to 0$. We thus refer to stochastic interventions governed by laws $Q_{\hbar,t_0}(\cdot|\boldsymbol{X} = \boldsymbol{x})$ constructed using tilts η satisfying Condition 6.1 as *positivity-lean-point-targeted* (PLPT) tilted interventions.

A canonical class satisfying this condition is $\eta(t; t_0, \hbar) \propto \exp\left[-\frac{\ell(|t-t_0|)}{a_\hbar}\right]$ with $a_\hbar > 0$ and $a_\hbar \to 0$ as $\hbar \to 0$, where $\ell : [0, \infty) \to [0, \infty)$ is continuous and strictly increasing. For such functions, it holds that for $0 \leq \rho_1 < \rho_2$,

$$\frac{\sup_{t\in\mathcal{T}:|t-t_0|\geq\rho_2} \eta(t; t_0, \hbar)}{\inf_{t\in\mathcal{T}:|t-t_0|\leq\rho_1} \eta(t; t_0, \hbar)} \lesssim \exp\left[-\frac{\ell(\rho_2) - \ell(\rho_1)}{a_\hbar}\right] \to 0.$$

For instance, $u \mapsto \ell(u) = u$ with $a_\hbar = \hbar$ gives a Laplace-type PLPT-tilting, while $u \mapsto \ell(u) = u^2$ with $a_\hbar = 2\hbar^2$ leads to a Gaussian-type PLPT-tilting. Additionally, the reflected exponential tilting function recently proposed by Schindl et al. (2024); Schindl and Wasserman (2025) also leads to a PLPT-tilted intervention. In contrast, the standard exponential tilt and related incremental interventions (Díaz and Hejazi, 2020; Schindl et al., 2024; Rakshit et al., 2024) are not PLPT-tilted interventions, because they correspond to tilting functions that do not satisfy Condition 6.1.

Since $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$ admits Lebesgue density $t \mapsto p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$, the PLPT-tilted intervention law $Q_{\hbar,t_0}(\cdot|\boldsymbol{X} = \boldsymbol{x})$ has density $t \mapsto q_{\hbar,t_0}(t|\boldsymbol{x})$ given by

$$q_{\hbar,t_0}(t|\boldsymbol{x}) = \frac{\eta(t; t_0, \hbar) \cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} \eta(u; t_0, \hbar) \cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\, du}. \tag{6.23}$$

Let $Y^{(q_{\hbar,t_0})}$ denote the potential outcome under treatment drawn from the general tilted distribution (6.23) and let $\bar{m}_\hbar(t_0) := \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$. We refer to the map $t_0 \mapsto \bar{m}_\hbar(t_0)$ as the tilted intervention curve.

Because $Q_{\hbar,t_0}(\cdot|\boldsymbol{X} = \boldsymbol{x})$ is, by construction, absolutely continuous with respect to $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$ for every $\boldsymbol{x} \in \mathcal{X}$, the counterfactual mean $\bar{m}_\hbar(t_0)$ for each fixed $\hbar$ is identifiable under Assumption 6.2. Specifically,

$$\bar{m}_\hbar(t_0) = \int_{\mathcal{X}} \int_{\mathcal{T}} \mu(t, \boldsymbol{x}) \cdot Q_{\hbar,t_0}(dt|\boldsymbol{X} = \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}); \tag{6.24}$$

see Muñoz and van der Laan (2012) and Theorem 1 in Kennedy (2019); Díaz and Hejazi (2020).

6.6.2 Convergence to NFTP as $\hbar \to 0$

We next prove that the NFTP intervention coincides with the limit of the PLPT-tilted intervention law as the tilting parameter converges to 0. Specifically, we establish the weak convergence of $Q_{\hbar,t_0}(\cdot|\boldsymbol{X}=\boldsymbol{x})$ to a point mass $\delta_{t_0^\star(\boldsymbol{x})}(\cdot)$ as $\hbar \to 0$. We then use this fact to demonstrate that $\lim_{\hbar\to 0} \bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(t_0^\star)}\right]$.

Theorem 6.8. *Suppose that Assumptions 6.9 and Condition 6.1 hold. Then, for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$,*

$$Q_{\hbar,t_0}(\cdot|\boldsymbol{X}=\boldsymbol{x}) \xrightarrow{d} \delta_{t_0^\star(\boldsymbol{x})}(\cdot) \quad \text{as} \quad \hbar \to 0,$$

where "$\xrightarrow{d}$" stands for the convergence in law.

The proof of Theorem 6.8 is in Section E.5. The theorem shows that, under the mild regularity conditions, the PLPT-tilted intervention converges to the NFTP intervention as the tilting parameter goes to 0. In Remark E.1 of Section E.5, we also show that if Condition 6.1 fails because $\Pi(t_0,\boldsymbol{x}) = \arg\min_{t\in\mathcal{T}(\boldsymbol{x})} |t-t_0|$ is not a singleton, then the thesis of Theorem 6.8 is weakened to $Q_{\hbar,t_0}\left(\{t\in\mathcal{T} : d(t,\Pi(t_0,\boldsymbol{x})) > \epsilon\}\,\middle|\,\boldsymbol{X}=\boldsymbol{x}\right) \to 0$ as $\hbar\to 0$ for every $\epsilon > 0$, where $d(t,\Pi(t_0,\boldsymbol{x})) = \inf_{t'\in\Pi(t_0,\boldsymbol{x})} |t-t'|$.

However, this weak convergence result alone does not imply the convergence of the tilted intervention mean to the NFTP estimand and their associated identification formulas. We need the following uniform integrability condition on (counterfactual) conditional means under the tilted intervention law (6.22).

Assumption 6.11 (Tilted uniform integrability)**.** *There exists a constant $\hbar_0 > 0$ such that, for each $f \in \{\mu, \mu^Y\}$ and for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$,*

$$\lim_{M\to\infty} \sup_{0<\hbar\le\hbar_0} \int_{\mathcal{T}} |f(t,\boldsymbol{x})| \cdot \mathbb{1}\{|f(t,\boldsymbol{x})| > M\}\, Q_{\hbar,t_0}(dt|\boldsymbol{x}) = 0.$$

Moreover, the family $\left\{\boldsymbol{x} \mapsto \int_{\mathcal{T}} f(t, \boldsymbol{x})\, Q_{\hbar,t_0}(dt|\boldsymbol{x}) : 0 < \hbar \leq \hbar_0\right\}$ *is uniformly integrable with respect to* $\mathrm{P}_{\boldsymbol{X}}$.

One sufficient condition for Assumption 6.11 is that $|\mu(t, \boldsymbol{x})|$ and $\left|\mu^Y(t, \boldsymbol{x})\right|$ are upper bounded by some $\mathrm{P}_{\boldsymbol{X}}$-integrable envelope function $\boldsymbol{x} \mapsto M(x)$ over all $t \in \mathcal{T}(\boldsymbol{x})$, but the current form of Assumption 6.11 also accommodates unbounded (counterfactual) conditional means under light-tailed tilted laws.

Proposition 6.9. *Suppose that Condition 6.1 as well as Assumptions 6.9, 6.10, and 6.11 hold. Then,*

(a) $\lim_{\hbar \to 0} \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right] = \mathbb{E}\left[Y^{(t_0^\star)}\right]$.

(b) $\lim_{\hbar \to 0} \int_{\mathcal{X}} \int_{\mathcal{T}} \mu(t, \boldsymbol{x}) \cdot Q_{\hbar,t_0}(dt|\boldsymbol{X} = \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathcal{X}} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})$.

The proof of Proposition 6.9 is in Section E.5. This proposition is particularly important for various reasons. First, it establishes that the NFTP counterfactual mean at t_0 can be approximated without requiring knowledge of the projection of t_0 into the support of the conditional distribution of T given X. In this sense, it provides an automatic approximation to the counterfactual mean under the NFTP regime. Second, exact identification via the g-formula depends on $t_0^\star(\boldsymbol{x})$, and any attempt to estimate directly the g-formula would necessitate recovering this quantity, an inherently difficult problem, as it involves estimating the boundary of the support of T given X. In contrast, the identifying formula for the fixed-$\hbar$ PLPT-tilted intervention targeted at t_0 is a pathwise differentiable parameter that provides a smooth approximation of the g-formula identifying the NFTP estimand at t_0, a non-regular parameter. This approximation is the basis for the influence-function-based estimation and inference procedures for the NFTP mean developed in Section 6.8.

Example 6.1 (Illustrative Gaussian example)**.** *Consider the following structural equation*

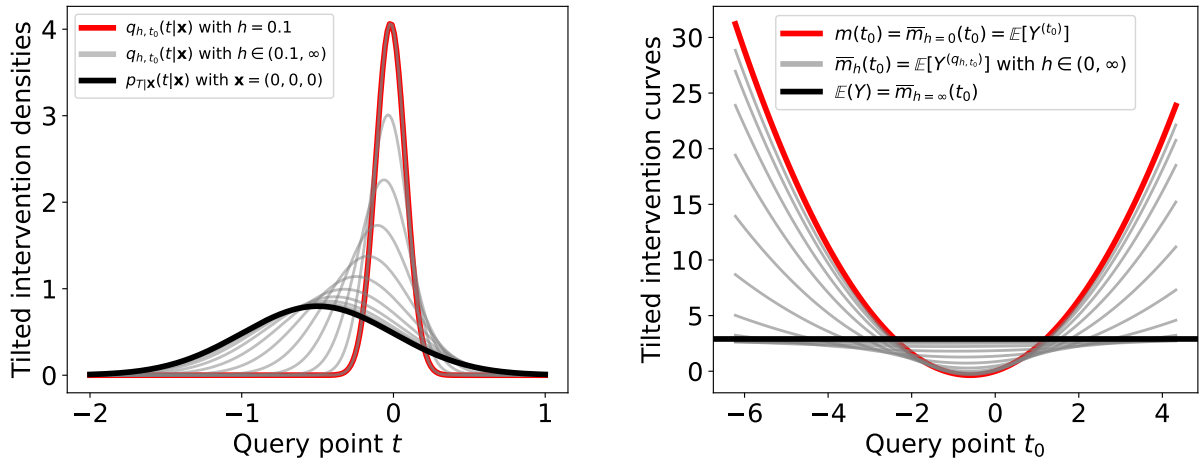


Figure 6.2: Illustrations of a Gaussian-type tilted intervention distribution (6.23) and its associated outcome mean curve for varying $\hbar$; see Example 6.1 for details.

model

$$\begin{aligned} \boldsymbol{X} &= (X_1, X_2, X_3)^T \sim \mathcal{N}_3\left(\mathbf{0}, \mathbf{I}_3\right), \\ T &= X_1 + X_2 - 0.5 + 0.5\epsilon_T, \\ Y &= 1.2\,T + T^2 + TX_1 + 1.2\,\boldsymbol{\xi}^T\boldsymbol{X} + \epsilon_Y, \end{aligned} \tag{6.25}$$

with $\boldsymbol{\xi} = \left(1, \frac{1}{4}, \frac{1}{9}\right)^T$, *where* ϵ_T *and* ϵ_Y *are standard normal random variables that are mutually independent and independent of all other variables. It can be easily checked that under this model,* $\mathbb{E}\left[Y^{(t_0)}\right] = 1.2t_0 + t_0^2$ *and* $\mathbb{E}(Y) = 2.9$.

Consider the Gaussian-type PLPT-tilted intervention with tilting function $\eta(t; t_0, \hbar) = \frac{1}{\hbar}K\left(\frac{t-t_0}{\hbar}\right)$ *where* $K(u) := \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$. *As indicated before, this tilting function satisfies Condition 6.1. Furthermore, the (counterfactual) conditional mean* $\mu^Y(t, \boldsymbol{x}) = 1.2t + t^2 + tx_1 + 1.2\boldsymbol{\xi}^T\boldsymbol{x}$ *satisfies Assumption 6.10 by the Gaussian design in (6.25). Since the conditional distribution* $T|\boldsymbol{X} = \boldsymbol{x}$ *is normally distributed as* $\mathcal{N}(x_1 + x_2 - 0.5, 0.25)$, *this Gaussian-type PLPT-tilted intervention is also Gaussian; specifically,* $Q_{\hbar,t_0}(\cdot|\boldsymbol{x})$ *is equal to* $\mathcal{N}\left(\frac{t_0+(4x_1+4x_2-2)\hbar^2}{1+\hbar^2}, \frac{\hbar^2}{1+\hbar^2}\right)$. *The outcome mean under the stochastic intervention* $Q_{\hbar,t_0}(\cdot|\boldsymbol{x})$

has a closed-form expression as:

$$\bar{m}_\hbar(t_0) = \frac{2.6\hbar^2 + 1.2t_0^2}{4\hbar^2 + 1} + \frac{t_0^2 - 4\hbar^2 t_0 + 36\hbar^4}{(4\hbar^2 + 1)^2}.$$

Notice that, as $\hbar \to 0$, the right hand side of the last equation converges to $1.2t_0 + t_0^2$, which is precisely equal to $\mathbb{E}\left[Y^{(t_0)}\right]$. On the other hand, as $\hbar \to \infty$, the expression converges to $\mathbb{E}(Y) = 2.9$. Figure 6.2 displays this Gaussian-type PLPT-tilted intervention densities and the associated mean outcome curve $t_0 \mapsto \bar{m}_\hbar(t_0)$ for varying $\hbar \in [0, \infty)$.

6.7 Kernel-Smoothed Tilting and Approximation Bias

In this section, we study a concrete construction of the tilted intervention distribution through kernel smoothing in Section 6.7.1 and quantify the associated approximation bias between the tilted intervention curve and its NFTP limit with respect to $\hbar$ in Section 6.7.2.

6.7.1 Kernel-Smoothed Tilting

Condition 6.2 (Second-order exponential-type kernel)**.** *The kernel function $K : \mathbb{R} \to (0, \infty)$ is continuous and bounded, satisfying the following conditions:*

(a) $\int_{\mathbb{R}} K(u)\,du = 1$, $\int_{\mathbb{R}} u\,K(u)\,du = 0$, *and* $\kappa_2 := \int_{\mathbb{R}} u^2 K(u)\,du < \infty$.

(b) there exist constants $\gamma > 0$, $0 < c_K \le C_K < \infty$, and $0 < R_K < \infty$ such that $c_K \exp(-|u|^\gamma) \le K(u) \le C_K \exp(-|u|^\gamma)$ when $|u| \ge R_K$.

As indicated before, Condition 6.2 implies Condition 6.1 with $\eta(t; t_0, \hbar) = \frac{1}{\hbar} K\left(\frac{t - t_0}{\hbar}\right)$. Whenever the conditional law has density $p_{T|\boldsymbol{X}}(\cdot|\boldsymbol{x})$, the resulting tilted intervention law $Q_{\hbar,t_0}(\cdot|\boldsymbol{x})$ also has a density

$$q_{\hbar,t_0}(t|\boldsymbol{x}) = \frac{K\left(\frac{t-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}. \tag{6.26}$$

Kernel smoothing has been widely used in causal inference for continuous treatments. For instance, Hirano and Imbens (2004); Kallus and Zhou (2018) applied kernel-based IPW

form $\mathbb{E}\left[\frac{Y \cdot K\left(\frac{T-t_0}{h}\right)}{h \cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right]$ as in (6.9), while Example 3 in Bibaut and van der Laan (2017) and Branson et al. (2023) proposed a regression-adjustment form

$$\psi_h(t_0) := \int_{\mathcal{X}} \int_{\mathcal{T}} \mu(t, \boldsymbol{x}) \cdot \frac{1}{h} K\left(\frac{t - t_0}{h}\right) dt \, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \tag{6.27}$$

to address the non-pathwise differentiability of $\mathbb{E}\left[Y^{(t_0)}\right]$. The kernel-smoothed tilted intervention (6.26) differs conceptually from these approaches. Rather than smoothing the outcome regression, it smooths the conditional treatment distribution $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}$. The resulting estimand $\mathbb{E}\left[Y^{(q_\hbar, t_0)}\right]$ thus remains identifiable even when the standard positivity condition fails; recall Section 6.6.1. Another distinction is visible in Example 6.1. Under the structural equation model (6.25), the regression-based approximation (6.27) yields $\psi_h(t_0) = 1.2t_0 + t_0^2 + h^2$, which converges to $m(t_0)$ as $h \to 0$ but diverges as $h \to \infty$. Conversely, the kernel-smoothed tilted intervention mean curve remains well-defined for all $\hbar \in [0, \infty)$ and converges to the observational mean as $\hbar \to \infty$. This gives the tilted framework a principled interpolation between the static-intervention regime and the observational regime.

6.7.2 Rate of Convergence for Approximating the NFTP Limit

We now characterize the rate of convergence for the kernel-smoothed tilted intervention mean curve to the NFTP mean curve as $\hbar \to 0$ under positivity violations. The support-continuous mean condition in Assumption 6.10 and tilted uniform integrability condition in Assumption 6.11 are not sufficient for explicating the rate, and we need the following weak positivity and support geometry condition around the closest point $t_0^\star(\boldsymbol{x})$ to the target level t_0.

Assumption 6.12 (Weak positivity and support geometry)**.** *For $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{X}$, the conditional probability kernel $\Psi(\boldsymbol{x}, dt)$ is absolutely continuous with a density representative $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$. For any fixed $t_0 \in \mathbb{R}$, there exist absolute constants $\epsilon_0 > 0$, $0 < p_{\min} \leq p_{\max} < \infty$, and $C_g < \infty$ such that $p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \leq p_{\max}$ for $\mathrm{P}_{T,\boldsymbol{X}}$-almost every $(t, \boldsymbol{x})$. Furthermore:*

(a) *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \in \mathcal{X}_0(t_0)$, $p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \geq p_{\min}$ *almost everywhere for* $t \in (t_0 - \epsilon_0, t_0 + \epsilon_0) \subset \mathcal{T}(\boldsymbol{x})$;

(b) *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$, *let*

$$I(\boldsymbol{x}) = \begin{cases} [t_0^\star(\boldsymbol{x}) - \epsilon_0, t_0^\star(\boldsymbol{x})], & \text{if } t_0 > t_0^\star(\boldsymbol{x}), \\ [t_0^\star(\boldsymbol{x}), t_0^\star(\boldsymbol{x}) + \epsilon_0], & \text{if } t_0 < t_0^\star(\boldsymbol{x}) \end{cases}.$$

Then, $I(\boldsymbol{x}) \subset \mathcal{T}(\boldsymbol{x})$, $p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \geq p_{\min}$ *almost everywhere for* $t \in I(\boldsymbol{x})$.

(c) *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$ *and every* $t \in \mathcal{T}(\boldsymbol{x})$, $|t - t_0^\star(\boldsymbol{x})| \leq C_g \left(|t - t_0| - |t_0 - t_0^\star(\boldsymbol{x})|\right)$.

Notice that Assumption 6.12 is much weaker than the standard positivity condition, since the lower bound on $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ is only imposed around $t_0^\star(\boldsymbol{x})$ for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{X}$. It allows "holes" in $\mathcal{T}(\boldsymbol{x})$ and its additional support components away from the interval in which $t_0^\star(\boldsymbol{x})$ lies (Schindl et al., 2024). At the same time, it also prevents these components from being arbitrarily close and creating near ties with the selected projection. One simple sufficient condition for Assumption 6.12(c) is to require $\mathcal{T}(\boldsymbol{x})$ to be a closed interval for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$. Then, $t_0^\star(\boldsymbol{x})$ is the endpoint closest to t_0, and all feasible treatment values lie on the same side of t_0 with $C_g = 1$.

Assumption 6.13 (Smoothness of μ and $p_{T|\boldsymbol{X}}$)**.** *Under the notation of Assumption 6.12 and for a fixed* $t_0 \in \mathbb{R}$*:*

(a) *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \in \mathcal{X}_0(t_0)$, *the selected representatives* $t \mapsto \mu(t, \boldsymbol{x})$ *and* $t \mapsto p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ *are twice continuously differentiable with respect to* $t \in (t_0 - \epsilon_0, t_0 + \epsilon_0) \subset \mathcal{T}(\boldsymbol{x})$. *Moreover, there exists a constant* $C_s < \infty$ *such that, for* $j = 0, 1, 2$,

$$\sup_{\substack{\boldsymbol{x} \in \mathcal{X}_0(t_0) \\ |t - t_0| < \epsilon_0}} \left\{ \left| \frac{\partial^j}{\partial t^j} \mu(t, \boldsymbol{x}) \right| + \left| \frac{\partial^j}{\partial t^j} p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \right| \right\} \leq C_s,$$

where the supremum is taken over the $\mathrm{P}_{\boldsymbol{X}}$*-probability-one set on which the stated smoothness holds;*

(b) *there exists an* $\mathrm{P}_{\boldsymbol{X}}$*-integrable function* $\boldsymbol{x} \mapsto M_\mu(\boldsymbol{x})$ *such that* $|\mu(t, \boldsymbol{x})| \leq M_\mu(\boldsymbol{x})$ *for all* $t \in \mathcal{T}(\boldsymbol{x})$ *and* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x}$*;*

(c) *there is a constant* $L_\mu \in (0, \infty)$ *such that for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$ *and every* $t \in \mathcal{T}(\boldsymbol{x})$*,*

$$|\mu(t, \boldsymbol{x}) - \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})| \leq L_\mu \, |t - t_0^\star(\boldsymbol{x})| \, .$$

Assumption 6.13 provides the regularity needed to control the approximation bias. It ensures valid second-order expansions of μ and $p_{T|\boldsymbol{X}}$ around t_0 on $\mathcal{X}_0(t_0)$, while the integrable envelope and Lipschitz conditions control μ near the projected treatment $t_0^\star(\boldsymbol{x})$ outside $\mathcal{X}_0(t_0)$. These conditions, together with Condition 6.2 and Assumption 6.14 below, yield the corresponding kernel approximation rates.

Assumption 6.14 (Margin condition)**.** *Fix* $t_0 \in \mathbb{R}$*. Let* $\zeta(\boldsymbol{x}) := \inf_{t \in \mathcal{T}(\boldsymbol{x})} |t - t_0| < \infty$ *for* $\mathrm{P}_{\boldsymbol{X}}$*-almost every* $\boldsymbol{x}$*. There exist constants* $C_\zeta, \omega, r_0 > 0$ *such that for all* $0 < r \leq r_0$*,*

$$\mathrm{P}_{\boldsymbol{X}} \left(\{ \boldsymbol{X} \notin \mathcal{X}_0(t_0) : 0 < \zeta(\boldsymbol{X}) \leq r \} \right) \leq C_\zeta \cdot r^\omega .$$

If the tail exponent γ *for the kernel in Condition 6.2 satisfies* $0 < \gamma < 1$*, assume additionally that* $\mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{ \zeta(\boldsymbol{X}) > r_0 \right\} \right] < \infty$*.*

Assumption 6.14 is known as the Tsybakov-type margin condition (Mammen and Tsybakov, 1999; Tsybakov, 2004; Audibert and Tsybakov, 2007). It controls the amount of probability mass carried by covariates $\boldsymbol{x}$ for which the target level t_0 is infeasible but lies close to the conditional treatment support $\mathcal{T}(\boldsymbol{x})$. This condition is standard in classification problems and closely related to the ω-regular condition used in density level-set estimation (Tsybakov, 1997). In our context, it is essential for controlling how rapidly the tilted intervention law approaches the nearest feasible treatment value $t_0^\star(\boldsymbol{x})$ as $\hbar \to 0$. The exponent ω characterizes the local amount of covariate mass near the feasibility boundary $\partial\mathcal{X}_0(t_0)$. Most commonly, when $\partial\mathcal{X}_0(t_0)$ is locally a smooth $(d-1)$-dimensional manifold and $\boldsymbol{X}$ has a positive, bounded density $p_{\boldsymbol{X}}(\boldsymbol{x})$ around a neighborhood of this boundary, then Assumption 6.14 holds with $\omega = 1$.

Theorem 6.10. *Suppose that Assumptions 6.9, 6.2, 6.10, 6.12, 6.13, and 6.14 hold. Assume further that the kernel function $K : \mathbb{R} \to [0, \infty)$ satisfies Condition 6.2. Then, as $\hbar \to 0$, the kernel-smoothed tilted intervention curve satisfies that*

$$\bar{m}_\hbar(t_0) = \bar{m}_0(t_0) + \Delta_{\text{int}}(\hbar) + \Delta_{\text{proj}}(\hbar),$$

where the interior approximation bias $\Delta_{\text{int}}(\hbar)$ is

$$\Delta_{\text{int}}(\hbar) = \frac{\kappa_2 \hbar^2}{2} \int_{\mathcal{X}_0(t_0)} \left[\frac{\partial^2}{\partial t^2} \mu(t_0, \boldsymbol{x}) + 2 \frac{\partial}{\partial t} \mu(t_0, \boldsymbol{x}) \cdot \frac{\partial}{\partial t} \log p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) \right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) + o(\hbar^2)$$

and the projected approximation bias $\Delta_{\text{proj}}(\hbar)$ satisfies

$$|\Delta_{\text{proj}}(\hbar)| = \begin{cases} O\left(\hbar^\gamma\right) & \text{when } 0 < \gamma \leq 1 \text{ or } \gamma > 1,\ \omega > \gamma - 1, \\ O(\hbar^\gamma |\log \hbar|) & \text{when } \gamma > 1,\ \omega = \gamma - 1, \\ O(\hbar^{\omega+1}) & \text{when } \gamma > 1,\ 0 < \omega < \gamma - 1 \end{cases}$$

where κ_2, γ, ω are specified in Condition 6.2 and Assumption 6.14, respectively.

The proof of Theorem 6.10 is in Section E.6, where all pointwise density values and derivatives refer to the twice continuously differentiable representatives fixed in Assumption 6.13(a). The approximation rate derived in Theorem 6.10 is particularly useful for the inference procedure for the NFTP estimand $\bar{m}_0(t_0) = \mathbb{E}\left[Y^{(t_0^\star)}\right]$, where the tilting parameter $\hbar$ is selected via a data-driven selection rule in Section 6.8.3. Theorem 6.10 shows that the bias has two components: $\Delta_{\text{int}}(\hbar)$, a regular kernel smoothing bias in the interior region of the support and $\Delta_{\text{proj}}(\hbar)$, a new bias caused by the violation of the positivity condition.

In the regular case where $\omega = 1$ in Assumption 6.14, using the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$, for which $\gamma = 2$, yields an approximation bias rate $O(\hbar^2 |\log \hbar|)$. This aligns with the standard second-order bias rate for kernel-smoothed estimators up to a logarithmic factor. Finally, the proof of Theorem 6.10 also explains why kernels with unbounded support are essential for approximating the full-population NFTP estimand. If $K : \mathbb{R} \to [0, \infty)$ has a bounded support, then for covariates $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$, the support of the localized kernel

$t \mapsto K\left(\frac{t-t_0}{\hbar}\right)$ will eventually have no overlap with the feasible treatment support $\mathcal{T}(\boldsymbol{x})$ as $\hbar \to 0$. The normalizing denominator of $q_{\hbar,t_0}(t|\boldsymbol{x})$ in (6.26) thus becomes zero, making the tilted intervention ill-defined for such units. In contrast, an unbounded-support kernel K assigns positive, though possibly exponentially small, weight to every feasible treatment value. This allows the tilted intervention to remain well-defined for the full target population and to converge to the NFTP limit without changing the population of interest.

6.8 Inference for Tilted Intervention Curve and NFTP Limit

This section develops pointwise estimation and inference for the tilted intervention mean $\bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ for fixed $\hbar > 0$ and, along a deterministic sequence $\hbar = \hbar_n \to 0$, for its NFTP limit $\bar{m}_0(t_0) = \lim_{\hbar \to 0} \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$. Throughout this section, the conditional probability kernel $\Psi(\boldsymbol{x}, dt)$ is assumed to admit a jointly measurable Lebesgue density representative $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$. The theory is stated for the kernel-smoothed tilted intervention (6.26) for simplicity.

6.8.1 Influence-Function-Based Estimation

For any fixed $\hbar > 0$ and $t_0 \in \mathcal{T}$, the PLPT-tilted intervention estimand $\bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ with identification (6.24) is pathwise differentiable (Bickel et al., 1998). We begin by deriving its efficient influence function (EIF) and constructing the corresponding one-step estimator. In addition, we discuss several practical strategies for approximating the conditional expectations that arise in the estimator. For measurable functions $\mu, p_{T|\boldsymbol{X}}$ and (6.26), we define the non-centered score as:

$$\begin{aligned}\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}}) := \frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y - \int_{\mathcal{T}} \mu(t_1,\boldsymbol{X}) \cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1\right] \\ + \int_{\mathcal{T}} \mu(t_1,\boldsymbol{X}) \cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1,\end{aligned} \tag{6.28}$$

and $\mathbb{E}\left[\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}})\right] = \bar{m}_\hbar(t_0)$ by (6.24).

Proposition 6.11. *Fix $\hbar > 0$ and $t_0 \in \mathcal{T}$. Suppose that the identifying functional in (6.24) is finite and*

$$V_{\text{eff},\bar{m}_\hbar}(t_0) := \text{Var}\left[\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}})\right] < \infty.$$

Then, $\bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ is pathwise differentiable under a nonparametric observed data model, with efficient influence function

$$(Y,T,\boldsymbol{X}) \mapsto \Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}}) - \bar{m}_\hbar(t_0).$$

The corresponding nonparametric efficiency bound can be written as:

$$V_{\text{eff},\bar{m}_\hbar}(t_0) = \mathbb{E}\left\{\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left(\text{Var}\left(Y|T,\boldsymbol{X}\right) + \left[\mu(T,\boldsymbol{X}) - \mathbb{E}_{Q_{\hbar,t_0}}\left(\mu(T,\boldsymbol{X})|\boldsymbol{X}\right)\right]^2\right)\right\}$$
$$+ \text{Var}\left\{\mathbb{E}_{Q_{\hbar,t_0}}\left[\mu(T,\boldsymbol{X})|\boldsymbol{X}\right]\right\},$$

where $\mathbb{E}_{Q_{\hbar,t_0}}\left[\mu(T,\boldsymbol{X})|\boldsymbol{X}\right] = \int_{\mathcal{T}} \mu(t_1,\boldsymbol{X}) \cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1$.

The proof of Proposition 6.11 follows standard arguments used for exponential tilting in Díaz and Hejazi (2020); Kennedy (2024); Schindl et al. (2024) and is therefore omitted.

By Proposition 6.11, we propose the following *one-step* estimator of $\bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ as:

$$\widehat{m}_\hbar(t_0) = \frac{1}{n}\sum_{i=1}^{n} \Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{q}_{\hbar,t_0}(T_i|\boldsymbol{X}_i)}{\widehat{p}_{T|\boldsymbol{X}}(T_i|\boldsymbol{X}_i)}\left[Y_i - \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X}_i)\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\, dt_1\right] + \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X}_i)\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\, dt_1\right\}, \tag{6.29}$$

where $\widehat{\mu}$ and $\widehat{p}_{T|\boldsymbol{X}}$ estimate the conditional mean outcome function $\mu(t,\boldsymbol{x}) = \mathbb{E}(Y|T = t,\boldsymbol{X} = \boldsymbol{x})$ and the conditional treatment density $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$, respectively. The estimated tilted density

$$\widehat{q}_{\hbar,t_0}(t|\boldsymbol{x}) = \frac{K\left(\frac{t-t_0}{\hbar}\right)\cdot\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\cdot\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{x})\, du} \tag{6.30}$$

is completely determined by the estimated conditional treatment density $\widehat{p}_{T|\boldsymbol{X}}$. In practice, $\widehat{p}_{T|\boldsymbol{X}}, \widehat{\mu}$ may need to be trained on an independent dataset using sample splitting or cross-fitting (Schick, 1986; Newey and Robins, 2018; Chernozhukov et al., 2018); see Theorem 6.13

and Section E.3.1 for details. To approximate the NFTP estimand $\bar{m}_0(t_0) = \mathbb{E}\left[Y^{(t_0^\star)}\right]$ through the estimator of $\mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ in its limit, one naturally seeks small values of $\hbar$ for (6.29). However, as we establish in Proposition 6.12 below, the variance of $\widehat{m}_\hbar(t_0)$ increases as $\hbar$ decreases. Thus, $\hbar$ should be chosen to balance decreasing approximation bias against increasing variance. This trade-off motivates the data-driven selection rule developed in Section 6.8.3.

For comparison, the Regression-Adjustment (RA) estimator is written as:

$$\widehat{m}_{\hbar,\mathrm{RA}}(t_0) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}} \widehat{\mu}(t_1, \boldsymbol{X}_i)\cdot \widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\, dt_1 \tag{6.31}$$

with the same nuisance function estimation as in (6.29). Although (6.31) may achieve smaller root mean squared error than (6.29) when both $\widehat{\mu}$ and $\widehat{p}_{T|\boldsymbol{X}}$ are consistent, it typically exhibits non-negligible first-order bias at the $\sqrt{n}$ or $\sqrt{n\hbar}$ scale. In contrast, the one-step estimator (6.29) is more robust to the model misspecification of $\mu(t, \boldsymbol{x})$ and admits straightforward asymptotic normal inference; see Theorem 6.13 and related discussion.

• **Integral Approximation:** Both the RA and one-step estimators require evaluation of the integral $\int_{\mathcal{T}} \widehat{\mu}(t_1, \boldsymbol{X}_i)\cdot \widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\, dt_1$, which may be analytically intractable. Following Schindl et al. (2024) for exponential tilting, we approximate this integral by a Riemann sum as:

$$\begin{aligned}\int_{\mathcal{T}} \widehat{\mu}(t_1, \boldsymbol{X})\cdot \widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1 &= \frac{\int_{\mathcal{T}} \widehat{\mu}(t_1, \boldsymbol{X})\cdot \widehat{p}_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\cdot \eta(t_1; t_0, \hbar)\, dt_1}{\int_{\mathcal{T}} \widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\cdot \eta(u; t_0, \hbar)\, du} \\ &\approx \frac{\sum_{j=1}^{N} \widehat{\mu}(\widetilde{t}_j, \boldsymbol{X})\cdot \widehat{p}_{T|\boldsymbol{X}}(\widetilde{t}_j|\boldsymbol{X})\cdot \eta(\widetilde{t}_j; t_0, \hbar)}{\sum_{j=1}^{N} \widehat{p}_{T|\boldsymbol{X}}(\widetilde{t}_j|\boldsymbol{X})\cdot \eta(\widetilde{t}_j; t_0, \hbar)},\end{aligned} \tag{6.32}$$

where $a_0 = \min\{T_1, ..., T_n\}$, $a_N = \max\{T_1, ..., T_n\}$, $a_j = a_0 + \frac{j\cdot(a_N - a_0)}{N}$ for $j = 1, ..., N-1$, and $\widetilde{t}_j = \frac{a_{j-1}+a_j}{2}$ or any value in $[a_{j-1}, a_j]$ for $j = 1, ..., N$. For tilting functions with exponential tails, such as $t \mapsto \eta(t; t_0, \hbar) \propto \exp\left(-\left|\frac{t-t_0}{\hbar}\right|^\gamma\right)$, we employ a log-sum-exp trick to maintain numerical stability when $\hbar > 0$ is small; see Section E.3.3 for details. This approximation step is essential for the estimators (6.29) and (6.31) to approach $t_0 \mapsto \bar{m}_0(t_0)$ through the g-formula (6.21).

6.8.2 *Minimax and Asymptotic Inference Theory*

To clarify how the variance of (6.29) depends on the tilting parameter $\hbar > 0$, we first derive a minimax lower bound for the mean squared error of estimating $\bar{m}_\hbar(t_0)$ at a fixed $t_0 \in \mathcal{T}$ under the kernel-smoothed tilted distribution (6.26). We restrict the model class through the following standard regularity condition on the conditional treatment density $p_{T|\boldsymbol{X}}$ and the outcome Y.

Assumption 6.15 (Non-degenerate local mass and outcome bounds)**.** *Fix $t_0 \in \mathcal{T}$. There exist a measurable set $\mathcal{A}_0(t_0)$ with $\mathrm{P}_{\boldsymbol{X}}(\mathcal{A}_0(t_0)) > 0$ and constants $\epsilon_1 > 0$, $0 < p_{\min} \leq p_{\max} < \infty$, $B_{\max} < \infty$, and $\sigma_{\min} > 0$ such that:*

(a) for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$, $(t_0 - \epsilon_1, t_0 + \epsilon_1) \subset \mathcal{T}(\boldsymbol{x})$, and $p_{\min} \leq p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \leq p_{\max}$ for Lebesgue-almost every $t \in (t_0 - \epsilon_1, t_0 + \epsilon_1)$.

(b) $|Y| \leq B_{\max}$ almost surely, and there exists a jointly measurable version $v(T, \boldsymbol{X})$ of $\mathrm{Var}(Y|T, \boldsymbol{X})$ satisfying $v(t, \boldsymbol{x}) \geq \sigma^2_{\min}$ for Lebesgue-almost every $t \in (t_0 - \epsilon_1, t_0 + \epsilon_1)$ and $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$.

Assumption 6.15(a) requires that a positive fraction of the covariate distribution $\mathrm{P}_{\boldsymbol{X}}$ has a uniformly non-degenerate neighborhood around the target treatment level t_0. For such covariate values $\boldsymbol{x}$, $t_0^\star(\boldsymbol{x}) = t_0$ and Assumption 6.15(a) can be viewed as a strengthening of the "weak positivity" condition in Assumption 6.12. Assumption 6.15(b) regularizes the moment and variation of Y on the same local region for the minimax lower bound below.

Proposition 6.12. *Let $\mathcal{P}$ denote the class of models satisfying Assumption 6.15 and containing at least one distribution P_0 for which a version of its conditional variance satisfies*

$$v_0(T, \boldsymbol{X}) = \mathrm{Var}_{\mathrm{P}_0}(Y|T, \boldsymbol{X}) \geq \sigma^2_{\min} + \epsilon_2$$

for some constant $\epsilon_2 > 0$, Lebesgue-almost every $t \in (t_0 - \epsilon_1, t_0 + \epsilon_1)$, and $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$. Suppose further that the kernel function satisfies Condition 6.2. For any

deterministic sequence $\hbar = \hbar_n \to 0$ *with* $n\hbar_n \to \infty$,

$$\inf_{\widehat{m}_\hbar(t_0)} \sup_{\mathrm{P}\in\mathcal{P}} \mathbb{E}_{\mathrm{P}} \left[\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0;\mathrm{P})\right]^2 \gtrsim \frac{1}{n\hbar},$$

where $\bar{m}_\hbar(t_0;\mathrm{P})$ *denotes the identifying functional in* (6.26).

The proof of Proposition 6.12 is in Section E.7. When $\mathcal{T} \subset \mathbb{R}$ is compact and Assumption 6.15 holds uniformly over $\mathcal{T}$, the same argument yields the lower bound

$$\inf_{\widehat{m}_\hbar} \sup_{\mathrm{P}\in\mathcal{P}} \mathbb{E}_{\mathrm{P}} \left\{ \int_{\mathcal{T}} \left[\widehat{m}_\hbar(t) - \bar{m}_\hbar(t)\right]^2 dt \right\} \gtrsim \frac{1}{n\hbar}$$

for the mean integrated squared error. This result confirms that the effective sample size is of order $n\hbar$ for estimating $\bar{m}_\hbar(t_0)$. Thus, as the tilted intervention becomes sharper with $\hbar \to 0$, the variance of any estimator must necessarily increase. This variance inflation is the fundamental obstacle to direct inference on the NFTP estimand $\bar{m}_0(t_0)$, and it is precisely what motivates the bias-variance tradeoff studied below. We now establish the asymptotic theory of the proposed one-step estimator $\widehat{m}_\hbar(t_0)$ in (6.29).

Theorem 6.13. *Let* $t_0 \in \mathcal{T}$ *be fixed and* $\hbar = \hbar_n$ *be deterministic. Suppose that Assumptions 6.2, 6.10, and Condition 6.2 hold. Assume further that* $\widehat{\mu}, \widehat{p}_{T|\boldsymbol{X}}$ *are constructed on a data sample independent of* $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$, *and*

(a) $\sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2} = o_P(1)$ *and* $\sup_{t\in\mathcal{T}} ||\widehat{\mu}(t,\boldsymbol{X}) - \widetilde{\mu}(t,\boldsymbol{X})||_{L_2} = o_P(1)$;

(b) The limiting function $\widetilde{\mu}(t,\boldsymbol{x})$ *in condition (b) is not necessarily equal to the true quantity* $\mu(t,\boldsymbol{x})$ *but* $\widehat{\mu}, \mu, \widetilde{\mu}$ *are all upper bounded for any* $(t,\boldsymbol{x}) \in \mathcal{J}$;

(c) $\sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2} \sup_{t\in\mathcal{T}} ||\widehat{\mu}(t,\boldsymbol{X}) - \mu(t,\boldsymbol{X})||_{L_2} + \sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2}^2 = o_P\left(\sqrt{\frac{V_{\mathrm{eff},\bar{m}_\hbar}(t_0)}{n}}\right)$, *where* $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)$ *is the nonparametric efficiency bound in Proposition 6.11 with* μ *replaced by* $\widetilde{\mu}$;

(d) $V_{\mathrm{eff},\bar{m}_\hbar}(t_0) > 0$ *and* $V_{\mathrm{eff},\bar{m}_\hbar}(t_0) = o(n)$.

Then, as $n \to \infty$ *and* $\hbar = \hbar_n \to 0$*, the estimator* (6.29) *of* $\bar{m}_\hbar(t_0)$ *under the kernel-smoothed tilted intervention* (6.26) *satisfies*

$$\sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\,[\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0)] \xrightarrow{d} \mathcal{N}(0,1).$$

The proof of Theorem 6.13 can be found in Section E.8. Two critical implications from the asymptotic results in Theorem 6.13 are noteworthy:

1. Although the one-step estimator (6.29) is derived from the EIF, it requires only second-order nuisance rates (as condition (d) in Theorem 6.13) but does not possess classical model double robustness (Smucler et al., 2019). In particular, the consistency of $\widehat{m}_\hbar(t_0)$ requires $\widehat{p}_{T|\boldsymbol{X}}$ to converge to the true conditional density $p_{T|\boldsymbol{X}}$, while $\widehat{\mu}$ may be misspecified.

2. As $\hbar \to 0$ with $n \to \infty$, the product rate requirement in condition (d) of Theorem 6.13 is weaker than the usual $o_P\left(n^{-\frac{1}{2}}\right)$ requirement for regular $\sqrt{n}$-consistent estimators. Thus, sufficiently fast convergence of $\widehat{p}_{T|\boldsymbol{X}}$ allows slow or even inconsistent estimation of μ; see also Corollary 6.14. For fixed $\hbar > 0$, however, both nuisance estimators $\widehat{\mu}, \widehat{p}_{T|\boldsymbol{X}}$ need to be consistent, and the product rate requirement in condition (d) reduces to $o_P\left(n^{-\frac{1}{2}}\right)$.

• **Statistical Inference for the fixed-$\hbar$ tilted intervention curve:** For fixed $\hbar > 0$, the asymptotic variance $V_{\text{eff},\bar{m}_\hbar}(t_0)$ of $\widehat{m}_\hbar(t_0)$ can be estimated by the empirical EIF variance as:

$$\begin{aligned}\widehat{V}_{\text{eff},\bar{m}_\hbar}(t_0) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ &\frac{\widehat{q}^2_{\hbar,t_0}(T_i|\boldsymbol{X}_i)}{\widehat{p}^2_{T|\boldsymbol{X}}(T_i|\boldsymbol{X}_i)}\left[Y_i - \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X}_i)\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\,dt_1\right]^2 \\ &+ \left[\int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X}_i)\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\,dt_1 - \widehat{m}_\hbar(t_0)\right]^2\Bigg\}.\end{aligned} \tag{6.33}$$

Accordingly, an asymptotically valid $(1-\tau)$-level confidence interval for $\bar{m}_\hbar(t_0)$ is

$$\left[\widehat{m}_\hbar(t_0) - z_{1-\frac{\tau}{2}}\cdot\sqrt{\frac{\widehat{V}_{\text{eff},\bar{m}_\hbar}(t_0)}{n}},\ \widehat{m}_\hbar(t_0) + z_{1-\frac{\tau}{2}}\cdot\sqrt{\frac{\widehat{V}_{\text{eff},\bar{m}_\hbar}(t_0)}{n}}\right],$$

where $z_{1-\frac{\tau}{2}}$ is the $\left(1-\frac{\tau}{2}\right)$ quantile of $\mathcal{N}(0,1)$. If the kernel function $K : \mathbb{R} \to [0,\infty)$ is additionally Lipschitz and the rate requirements of Theorem 6.13 hold uniformly over $t_0 \in \mathcal{T}$, then our proof of Theorem 6.13 in Section E.8 suggests that

$$\sup_{t_0\in\mathcal{T}}\left|\sqrt{\frac{n}{\widehat{V}_{\text{eff},\bar{m}_\hbar}(t_0)}}\left[\widehat{m}_\hbar(t_0)-\bar{m}_\hbar(t_0)\right]-\sqrt{n}\left(\mathbb{P}_n-\mathrm{P}\right)\frac{\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})}{\sqrt{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\right|=o_P(1),$$

where $\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}}) := \frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y-\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]+\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1$. Consequently, simultaneous confidence bands for $t_0 \mapsto \bar{m}_\hbar(t_0)$ over $t_0 \in \mathcal{T}$ can be constructed using a multiplier bootstrap method as in Belloni et al. (2015); Kennedy (2019); Mauro et al. (2020).

Combining Theorem 6.13 with our bias expansion in Theorem 6.10 yields the following corollary for valid inference on the NFTP estimand $\bar{m}_0(t_0) = \lim_{\hbar\to 0}\bar{m}_\hbar(t_0)$, whose proof is in Section E.8. This result is crucial because it justifies the validity of (6.29) for inference on the static intervention effect even when the standard positivity assumption fails.

Corollary 6.14. *Suppose that the assumptions in Theorem 6.10 and Theorem 6.13 hold. Then, we have that* $\sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\left[\widehat{m}_\hbar(t_0)-\bar{m}_0(t_0)\right] \xrightarrow{d} \mathcal{N}(0,1)$ *when* $n\to\infty$, $\frac{n\hbar^4}{V_{\text{eff},\bar{m}_\hbar}(t_0)}\to 0$, *and*

$$\frac{1}{V_{\text{eff},\bar{m}_\hbar}(t_0)}\cdot\begin{cases} n\hbar^{2\gamma}\to 0 & \text{when } 0<\gamma\le 1 \text{ or } \gamma>1,\ \omega>\gamma-1,\\ n\hbar^{2\gamma}(\log\hbar)^2\to 0 & \text{when } \gamma>1,\ \omega=\gamma-1,\\ n\hbar^{2\omega+2}\to 0 & \text{when } \gamma>1,\ 0<\omega<\gamma-1.\end{cases}$$

This corollary makes explicit the bias-variance trade-off for inference on the NFTP estimand. The tilting parameter $\hbar$ needs to be small enough so that the approximation bias $\bar{m}_\hbar(t_0)-\bar{m}_0(t_0)$ is asymptotically negligible after normalization by the standard error, but not so small that the effective sample size $\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}$ along the tilted path degenerates. This observation directly motivates the data-driven selection rule in Section 6.8.3 below. If Assumption 6.15(a) holds for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$, then the asymptotic variance satisfies $V_{\text{eff},\bar{m}_\hbar}(t_0) \asymp \frac{1}{\hbar}$. Together with the regular boundary condition with $\omega = 1$ in Assumption 6.14, Corollary 6.14 reduces, under the Gaussian kernel with $\gamma = 2$, to the classical

undersmoothing rate condition $n\hbar^5|\log\hbar|^2 \to 0$, along with $n\hbar \to \infty$ for variance stability. This matches the standard bandwidth condition for pointwise inference under a second-order kernel up to a logarithmic factor; see Section 5.7 in Wasserman (2006).

6.8.3 Data-Driven Selection of the Tilting Parameter

In practice, choosing $\hbar > 0$ too small may lead to unstable variance estimates $\widehat{V}_{\text{eff},\bar{m}_\hbar}(t_0)$ and poor finite-sample behavior. To address this issue, we propose a data-driven selection rule for $\hbar$ inspired by a Lepski-type bias-variance trade-off (Lepskii, 1991; Goldenshluger and Lepski, 2011). The guiding principle is to select the largest value of $\hbar$ for which the residual bias remains negligible relative to the corresponding standard deviation. This yields a stable estimator while preserving valid inference for the NFTP estimand $\bar{m}_0(t_0)$.

1. Initialize the tilting parameter $\hbar_0$ at a small value, such as $\hbar_0 = 10^{-2}\,\widehat{\sigma}_T\,n^{-\frac{1}{4}}$, and iteratively double its value $\hbar_{k+1} \leftarrow 2\hbar_k$ until either $\hbar_k > \hbar_{\max}$ or the average estimated variance of $\widehat{m}_\hbar(t_0)$ over the grid $\mathcal{D}_T \subset \mathcal{T}$ stops decreasing, *i.e.*, $\sum_{t_0\in\mathcal{D}_T} \widehat{V}_{\text{eff},\bar{m}_{\hbar_k}}(t_0) > \sum_{t_0\in\mathcal{D}_T} \widehat{V}_{\text{eff},\bar{m}_{\hbar_{k+1}}}(t_0)$. We set the upper bound $\hbar_{\max} = 10\,\widehat{\sigma}_T\,n^{-\frac{1}{5}}$ according to a scaled version of Silverman's rule of thumb, where $\widehat{\sigma}_T$ is the sample standard deviation of $\{T_1, ..., T_n\}$.

2. For each candidate $\hbar_k$, compute

$$\widehat{\text{Bias}}(\hbar_k) = \sqrt{\frac{1}{|\mathcal{D}_T|}\sum_{t_0\in\mathcal{D}_T}\left[\widehat{m}_{\hbar_k}(t_0) - \widehat{m}_{\hbar_{k+1}}(t_0)\right]^2} \quad \text{and} \quad \widehat{\text{SD}}(\hbar_k) = \sqrt{\frac{1}{n|\mathcal{D}_T|}\sum_{t_0\in\mathcal{D}_T}\widehat{V}_{\text{eff},\bar{m}_{\hbar_k}}(t_0)}.$$

3. Select the largest $\hbar_k$ such that $\frac{\widehat{\text{Bias}}(\hbar_k)}{\widehat{\text{SD}}(\hbar_k)} \le \epsilon_T$ with a default choice $\epsilon_T = 0.1$.

Step 3 is the central component of the procedure. By selecting the largest possible $\hbar$ satisfying the ratio criterion, we ensure that the estimated bias remains negligible relative to the estimated asymptotic standard deviation, aligning with the conditions of Corollary 6.14. The specific bias and standard deviation estimators in Step 2 are inspired by Section 4.4 in Powell

and Stoker (1996). Cross-validation-based bias estimates (van der Laan and Dudoit, 2003; Díaz and van der Laan, 2013) could also be used, but are typically more computationally expensive.

6.9 "Nature versus Nurture" in Galaxy Evolution

We apply the proposed RA and one-step estimators for the NFTP intervention curve to study the causal effect of local environment on galactic star formation rate (SFR) using data from the IllustrisTNG simulation (Pillepich et al., 2018a,b; Springel et al., 2018). This analysis provides a new causal perspective on the long-standing "Nature versus Nurture" question in galaxy evolution, *i.e.*, whether galaxy formation and evolution are driven primarily by internal processes (nature) or by external environmental factors (nurture).

6.9.1 Data Description

We follow the data setup of Mucesh et al. (2024) to conduct our causal analysis. For some background, IllustrisTNG consists of hydrodynamical simulations of galaxy formation and evolution in large cosmological volumes, implemented with the moving-mesh code `AREPO` (Springel, 2010). Our analysis uses the TNG100 simulation across cosmic time from redshift $z \approx 3$ to $z \approx 0$. The simulation volume is a periodic box with side length $75\,h_{\text{hubble}}^{-1}$Mpc under a flat ΛCDM cosmology (Ade et al., 2016), where the dimensionless Hubble constant is $h_{\text{hubble}} = 0.6774$. Earlier Illustris simulations, which preceded IllustrisTNG, have been shown to reproduce several observed galaxy properties at low and intermediate redshifts (Vogelsberger et al., 2014; Sijacki et al., 2015), supporting their use for methodological investigations.

We obtain a sample of 18,629 galaxies with stellar mass $M_\star > 10^9 M_\odot$ by matching friends-of-friend halos with `SUBFIND` subhalos (Springel et al., 2005; Dolag et al., 2009). Subhalos with no detectable dark matter or with a non-cosmological origin are excluded. Here, $M_\odot \approx 1.9884 \times 10^{30}$kg denotes one solar mass. For the treatment variable T, we adopt the local environment proxy from Section 5.1 in Mucesh et al. (2024) as the 10-th

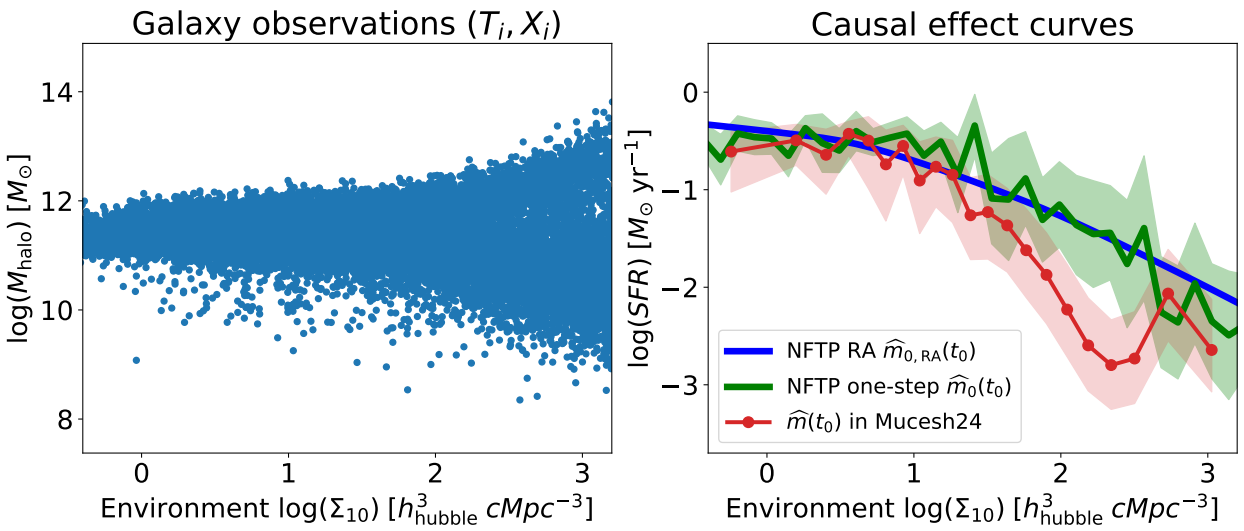


Figure 6.3: Results on the IllustrisTNG galaxy data at redshift $z = 0$. **Left:** Scatter plot of the confounder (halo mass) against the treatment variable (environment density). **Right:** Estimated tilted intervention curves by our proposed RA and one-step estimators under data-driven tilting parameters, together with the dose-response curve estimated by Mucesh et al. (2024).

nearest neighbor density $\Sigma_{10} = \frac{10}{(4\pi/3)r_{10}^3}$, where r_{10} is the three-dimensional distance from the galaxy to its 10-th nearest neighbor prior to applying the stellar mass threshold. The outcome variable Y is the instantaneous SFR recorded by the simulation for each galaxy. To adjust for the internal processes (nature), we use the halo mass as our confounder X. A more comprehensive astronomical assessment of this causal graph, including additional confounders and mediating mechanisms, is left for future work.

6.9.2 Results

We primarily examine the causal effect of local environment on galactic SFR at redshift $z = 0$ based on the selected IllustrisTNG galaxy data. The left panel of Figure 6.3 presents a log-scale scatter plot of the treatment (environmental density) against the confounder (halo mass). This plot reveals clear evidence of potential positivity violations: for a given halo mass (*e.g.*, $x = \log(M_{\text{halo}}) = 10M_\odot$), the conditional treatment density $p_{T|X}(t|x)$ is

effectively zero in some regions where the marginal density $p_T(t)$ remains strictly positive (*e.g.*, $t = \log(\Sigma_{10}) < 1\, h^3_{\text{hubble}}\text{cMpc}^{-3}$). Hence, traditional dose-response curve estimators are likely biased or entirely infeasible for this dataset. Since our proposed RA and one-step estimators for the NFTP intervention curve remain valid even with positivity violations, we apply them with data-driven tilting parameters to the galaxy data and compare the results against the causal dose-response curve estimated by Mucesh et al. (2024). As displayed in the right panel of Figure 6.3, both our proposed estimators and the one by Mucesh et al. (2024) identify a decreasing trend in the galactic SFR as the local environment density increases. The results for our estimators are robust regardless of whether RKSO or RKDE methods are used to estimate $\widehat{p}_{T|X}$. However, the tilted intervention curves produced by our estimators decrease more gradually than the dose-response curve estimated by Mucesh et al. (2024). In the presence of positivity violations, standard dose-response estimators rely on the validity of extrapolating the outcome regression model $\mu(t, x) = \mathbb{E}(Y|T = t, X = x)$ beyond the empirical joint support of (T, X), which often incurs severe bias. In contrast, our estimators target the NFTP curve, which assigns each galaxy the closest feasible environment level given its halo mass, yielding an estimand that is directly identifiable and more stable to estimate. Additionally, the shaded region reported by Mucesh et al. (2024) represents the Monte Carlo standard error, rather than a statistically valid confidence band. By contrast, the shaded regions for our one-step estimator $\widehat{m}_{\hbar}(t_0)$ represent theoretically valid pointwise 95% confidence intervals. In Section E.4, we provide extensive comparisons between our estimators and the estimated curves by Mucesh et al. (2024) across various redshifts z and tilting parameters $\hbar$. Overall, our estimators reveal significant downward trends at low redshifts, indicating a negative causal effect of local environment on galactic SFR. The magnitudes of these negative effects attenuate as redshift increases beyond $z > 0.5$. Unlike the results in Mucesh et al. (2024), our estimated curves do not yield a strong positive causal trend between environment and SFR at high redshifts, which is consistent with prior research (Scoville et al., 2013; Ziparo et al., 2014; Darvish et al., 2016). Nevertheless, more in-depth analyses are required in the future to fully disentangle the relative contributions of nature

and nurture to galaxy evolution, particularly to validate the chosen environmental proxy and to assess whether conditioning on halo mass sufficiently controls for internal processes.

Chapter 7

CONCLUDING REMARKS

This dissertation develops statistical methods for extracting low-dimensional structures, conducting high-dimensional inference, and drawing causal conclusions from complex observational data, with astronomy as the primary motivating domain. Together, the contributions address challenges that range from geometric structure estimation to high-dimensional inference and causal analysis, and they open up several directions for future work.

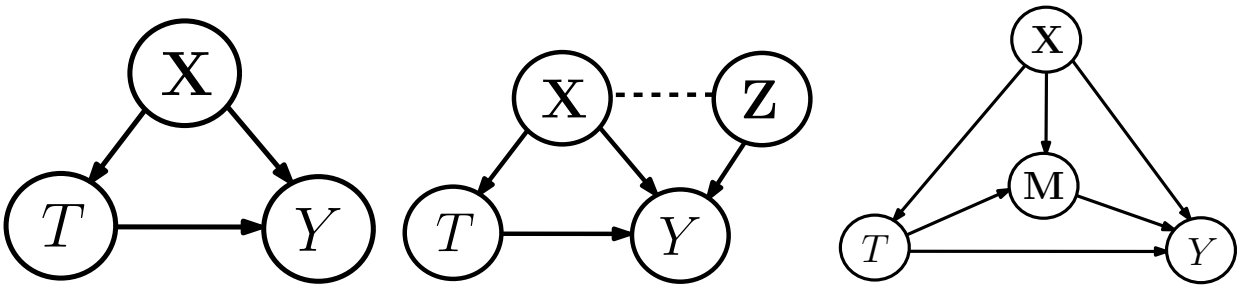
1. Cosmic Void Detection via Delaunay Triangulation: Chapter 2–Chapter 4 focus primarily on reconstructing cosmic filaments from SDSS-IV data. A natural next step is to extend the catalog to incorporate cosmic voids, which are large regions of near-emptiness in the Universe. Since cosmic voids are less affected by nonlinear gravitational effects than high-density structures, they provide valuable probes for constraining cosmological parameters and studying the nature of dark energy (Betancort-Rijo et al., 2009; Sutter et al., 2014; Lee and Park, 2009; Cai et al., 2015). Existing void-finding algorithms typically identify cosmic voids either by growing spheres centered at low-density locations (Sánchez et al., 2016) or by applying Delaunay triangulation directly to raw observed galaxies or simulated halos (Zhao et al., 2016). The former approach can be restrictive because cosmic voids need not be spherical, whereas the latter may be sensitive to noise and sampling variation in the raw point cloud data.

The filament catalog developed in this dissertation offers an alternative strategy. Instead of applying Delaunay triangulation directly to galaxies, we could apply it to the denoised filamentary points in the cosmic web catalog from Chapter 4. This approach is able to identify cosmic voids with adaptive shapes whose boundaries are supported by estimated filamentary structures. Future work could validate these candidate cosmic voids by assessing

the statistical significance of gravitational effects around them (Sánchez et al., 2016), enrich the current cosmic web catalog with this cosmic void component, and use the identified cosmic voids for downstream inference on cosmological parameters, potentially with modern machine learning tools (Kreisch et al., 2022).

2. High-dimensional Inference With Missing Covariates: Chapter 5 studies high-dimensional inference when outcomes are missing. In many astronomical applications, however, covariates may also be incomplete. For example, when fitting the mass profile of the Milky Way, some velocity and position features for globular clusters and nearby dwarf galaxies may be unavailable (Eadie et al., 2015). Extending the proposed debiasing framework to settings with both missing outcomes and covariates would therefore broaden its applicability. One promising direction is to combine the bias-corrected inference procedure in Chapter 5 with a covariate imputation strategy (Carpenter et al., 2023). Such an extension would need to account for the interactions between imputation error, missing-outcome adjustment, and high-dimensional regularization bias. Related ideas have been developed for high-dimensional linear regression with blockwise missing covariates (Xue et al., 2021; Song et al., 2024), suggesting that a unified framework for simultaneous missing outcomes and covariates may be feasible.

3. Causal Inference for Continuous Treatments Under More Complex Causal Graphs: Chapter 6 considers observational studies in which the relevant confounding variables $\boldsymbol{X}$ are observed and controlled, as illustrated by the causal graph in Figure 7.1(a). In practice, more complicated causal relationships may exist among the collected variables. Additional variables $\boldsymbol{Z}$ may directly affect the outcome while being correlated with the confounders, as in Figure 7.1(b). Alternatively, intermediate variables $\boldsymbol{M}$ may lie on the causal pathway from the treatment T to the outcome Y, as in the mediation problems illustrated in Figure 7.1(c) (Huber et al., 2020). In astronomy, such settings arise naturally when the effect of a galaxy's local environment on its star formation rate is mediated by its proximity to, or interaction with, cosmic web structures. To enable valid causal inference under general causal graphs, it would be of research interest to generalize our proposed identification and



(a) Basic case with only confounding variables.

(b) Extension 1 with independent effect variables.

(c) Extension 2 with causal mediators.

Figure 7.1: Causal directed acyclic graphs in different scenarios. Here, T is the treatment variable, Y is the outcome of interest, $\boldsymbol{X}$ consists of the confounding variables, and $\boldsymbol{M}$ is a mediator.

estimation strategies, as well as the new NFTP regime, to these richer causal settings under positivity violations.

4. Causal Discovery in Astronomy: Chapter 6 and the preceding future direction assume that the relevant causal graph is known or scientifically specified in advance. In many astronomical applications, however, the causal graph itself may be uncertain. This raises the potential of combining causal inference with causal discovery or structure learning methods. Although causal discovery has been extensively studied in statistics, computer science, and related fields (Richardson, 1996; Spirtes and Zhang, 2016; Zanga et al., 2022), its use in astronomy remains comparatively limited (Pasquato et al., 2023).

Future work could investigate how causal discovery methods can be adapted to astronomical data, where observational constraints, selection effects, measurement errors, and complex physical mechanisms all complicate graph learning processes. A particularly promising direction is to combine domain knowledge from astrophysics with statistically principled structure learning procedures, using the learned or partially learned graph as a basis for downstream causal estimation. Such a framework would move beyond estimating causal effects under a fixed graph and toward a broader pipeline for causal scientific discovery from astronomical surveys.

BIBLIOGRAPHY

M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. J. Guibas, and D. Morozov. Metric graph reconstruction from noisy data. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, pages 37–46, 2011.

A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

P. A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the riemannian hessian. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, pages 361–368. Springer Berlin Heidelberg, 2013.

K. Accetta, C. Aerts, V. Silva Aguirre, R. Ahumada, N. Ajgaonkar, N. Filiz Ak, S. Alam, C. Allende Prieto, A. Almeida, F. Anders, et al. The seventeenth data release of the sloan digital sky surveys: Complete release of manga, mastar, and apogee-2 data. *The Astrophysical Journal Supplement Series*, 259(2):35, 2022.

P. A. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. Banday, R. Barreiro, J. Bartlett, N. Bartolo, et al. Planck 2015 results-xiii. cosmological parameters. *Astronomy & Astrophysics*, 594:A13, 2016.

A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, R. Arcodia, E. Armengaud, M. Aubert, et al. The 16th data release of the sloan

digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020.

M. Alpaslan, M. Grootes, P. M. Marcum, C. Popescu, R. Tuffs, J. Bland-Hawthorn, S. Brough, M. J. I. Brown, L. J. M. Davies, S. P. Driver, B. W. Holwerda, L. S. Kelvin, M. A. Lara-López, Á. R. López-Sánchez, J. Loveday, A. Moffett, E. N. Taylor, M. Owers, and A. S. G. Robotham. Galaxy And Mass Assembly (GAMA): stellar mass growth of spiral galaxies in the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 457 (3):2287–2300, apr 2016.

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

M. A. Aragon-Calvo and L. F. Yang. The hierarchical nature of the spin alignment of dark matter haloes in filaments. *Monthly Notices of the Royal Astronomical Society*, 440: L46–L50, may 2014.

M. A. Aragón-Calvo, R. van de Weygaert, B. J. T. Jones, and J. M. van der Hulst. Spin Alignment of Dark Matter Halos in Filaments and Walls. *Astrophysical Journal Letters*, 655(1):L5–L8, jan 2007.

M. A. Aragón-Calvo, R. van de Weygaert, and B. J. T. Jones. Multiscale phenomenology of the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 408(4):2163–2187, nov 2010.

E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.

M. Ashizawa, H. Sasaki, T. Sakai, and M. Sugiyama. Least-squares log-density gradient clustering for riemannian manifolds. In *Proceedings of the Artificial Intelligence and Statistics*, pages 537–546. PMLR, 2017.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007.

Z. Bai, C. Rao, and L. Zhao. Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27(1):24–39, 1988.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 02 2017.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46):1345–1382, 2005.

A. Banyaga and D. Hurtubise. *Lectures on Morse homology*, volume 29. Springer Science & Business Media, 2004.

M. Bayati, M. A. Erdogdu, and A. Montanari. Estimating lasso risk and noise level. *Advances in Neural Information Processing Systems*, 26:944–952, 2013.

P. C. Bellec and C.-H. Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

A. Belloni, V. Chernozhukov, and K. Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758, 2019.

R. Beran. Exponential models for directional data. *Annals of Statistics*, 7(6):1162–1178, 11 1979.

J. Betancort-Rijo, S. G. Patiri, F. Prada, and A. E. Romano. The statistics of voids as a tool to constrain cosmological parameters: σ_8 and Γ. *Monthly Notices of the Royal Astronomical Society*, 400(4):1835–1849, 2009.

A. F. Bibaut and M. J. van der Laan. Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. *arXiv preprint arXiv:1706.07408*, 2017.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer New York, 1998.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

M. R. Blanton, D. J. Schlegel, M. A. Strauss, J. Brinkmann, D. Finkbeiner, M. Fukugita, J. E. Gunn, D. W. Hogg, Ž. Ivezić, G. Knapp, et al. New york university value-added galaxy catalog: a galaxy catalog based on new public surveys. *The Astronomical Journal*, 129(6):2562, 2005.

A. S. Bolton, D. J. Schlegel, É. Aubourg, S. Bailey, V. Bhardwaj, J. R. Brownstein, S. Burles, Y.-M. Chen, K. Dawson, D. J. Eisenstein, J. E. Gunn, G. R. Knapp, C. P. Loomis, R. H. Lupton, C. Maraston, D. Muna, A. D. Myers, M. D. Olmstead, N. Padmanabhan, I. Pâris, W. J. Percival, P. Petitjean, C. M. Rockosi, N. P. Ross, D. P. Schneider, Y. Shu, M. A. Strauss, D. Thomas, C. A. Tremonti, D. A. Wake, B. A. Weaver, and W. M. Wood-Vasey. Spectral Classification and Redshift Measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey. *Astronomical Journal*, 144(5):144, nov 2012.

J. R. Bond, L. Kofman, and D. Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380(6575):603–606, apr 1996.

S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

N. Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 3, 2020.

A. W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

Z. Branson, E. H. Kennedy, S. Balakrishnan, and L. Wasserman. Causal effect estimation after propensity score trimming with continuous treatments. *arXiv preprint arXiv:2309.00706*, 2023.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

A. Brown, A. Vallenari, T. Prusti, J. De Bruijne, C. Babusiaux, C. Bailer-Jones, M. Biermann, D. W. Evans, L. Eyer, F. Jansen, et al. Gaia data release 2-summary of the contents and survey properties. *Astronomy & Astrophysics*, 616:A1, 2018.

J. J. Bryant, M. S. Owers, A. S. G. Robotham, S. M. Croom, S. P. Driver, M. J. Drinkwater, N. P. F. Lorente, L. Cortese, N. Scott, M. Colless, A. Schaefer, E. N. Taylor, I. S. Konstantopoulos, J. T. Allen, I. Baldry, L. Barnes, A. E. Bauer, J. Bland-Hawthorn, J. V. Bloom, A. M. Brooks, S. Brough, G. Cecil, W. Couch, D. Croton, R. Davies, S. Ellis, L. M. R. Fogarty, C. Foster, K. Glazebrook, M. Goodwin, A. Green, M. L. Gunawardhana, E. Hampton, I. T. Ho, A. M. Hopkins, L. Kewley, J. S. Lawrence, S. G. Leon-Saval, S. Leslie, R. McElroy, G. Lewis, J. Liske, Á. R. López-Sánchez, S. Mahajan, A. M. Medling, N. Metcalfe, M. Meyer, J. Mould, D. Obreschkow, S. O'Toole, M. Pracy, S. N. Richards, T. Shanks, R. Sharp, S. M. Sweet, A. D. Thomas, C. Tonini, and C. J. Walcher. The

SAMI Galaxy Survey: instrument specification and target selection. *Monthly Notices of the Royal Astronomical Society*, 447(3):2857–2879, mar 2015.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(4):231–357, January 2015.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212 – 1242, 2013.

P. Bühlmann and S. van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

K. Bundy, M. A. Bershady, D. R. Law, R. Yan, N. Drory, N. MacDonald, D. A. Wake, B. Cherinka, J. R. Sánchez-Gallego, A.-M. Weijmans, et al. Overview of the sdss-iv manga survey: mapping nearby galaxies at apache point observatory. *The Astrophysical Journal*, 798(1):7, 2015.

Y. Burago, M. Gromov, and G. Perel'man. A.d. alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys*, 47(2):1–58, apr 1992.

N. Busca and C. Balland. Quasarnet: Human-level spectral classification and redshifting with deep neural networks. *arXiv preprint arXiv:1808.09955*, 2018.

B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4): 999–1023, 2006.

T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

T. T. Cai, Z. Guo, and R. Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, 118(542): 1319–1332, 2023.

Y.-C. Cai, N. Padilla, and B. Li. Testing gravity using cosmic voids. *Monthly Notices of the Royal Astronomical Society*, 451(1):1036–1055, 2015.

B. Callaway, A. Goodman-Bacon, and P. H. Sant'Anna. Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research, 2024.

J. R. Carpenter, J. Bartlett, T. Morris, A. Wood, M. Quartagno, and M. G. Kenward. *Multiple imputation and its application*. John Wiley & Sons, 2023.

M. Á. Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 153–160, 2006.

M. Á. Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, may 2007.

M. Á. Carreira-Perpiñán. Generalised blurring mean-shift algorithms for nonparametric clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

M. Á. Carreira-Perpiñán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

J. Carrón Duque, M. Migliaccio, D. Marinucci, and N. Vittorio. A novel cosmic filament catalogue from SDSS data. *Astronomy & Astrophysics*, 659:A166, mar 2022.

R. Caseiro, J. F. Henriques, P. Martins, and J. Batista. Semi-intrinsic mean shift on riemannian manifolds. In *Computer Vision – ECCV 2012*, pages 342–355, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

M. Cautun, R. van de Weygaert, B. J. T. Jones, and C. S. Frenk. Evolution of the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 441(4):2923–2973, jul 2014.

H. E. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1896–1902. IEEE, 2009.

G. Chabrier. Galactic stellar and substellar initial mass function. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795, jul 2003.

J. E. Chacón. The modal age of statistics. *International Statistical Review*, 88(1):122–141, 2020.

J. E. Chacón, T. Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807–840, 2011.

A. Chakrabortty, J. Lu, T. T. Cai, and H. Li. High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*, 2019.

Y.-Y. Chang, A. van der Wel, E. da Cunha, and H.-W. Rix. Stellar masses and star formation rates for 1 m galaxies from sdss+ wise. *The Astrophysical Journal Supplement Series*, 219 (1):8, 2015.

S.-J. Chang-Chien, W.-L. Hung, and M.-S. Yang. On mean shift-based clustering for circular data. *Soft Computing*, 16(6):1043–1060, jun 2012.

Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021.

P.-H. Chen, C.-J. Lin, and B. Schölkopf. A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry*, 21(2):111–136, 2005.

Y. Chen and Y. Yang. The one standard error rule for model selection: Does it work? *Stats*, 4(4):868–892, 2021.

Y.-C. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

Y.-C. Chen. Solution manifold and its statistical applications. *Electronic Journal of Statistics*, 16(1):408–450, 2022.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Generalized Mode and Ridge Estimation. *arXiv e-prints*, art. arXiv:1406.1803, jun 2014.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *Annals of Statistics*, 43(5):1896–1928, 10 2015.

Y.-C. Chen, S. Ho, P. E. Freeman, C. R. Genovese, and L. Wasserman. Cosmic web reconstruction through density ridges: method and algorithm. *Monthly Notices of the Royal Astronomical Society*, 454(1):1140–1156, nov 2015.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.

Y.-C. Chen, S. Ho, J. Brinkmann, P. E. Freeman, C. R. Genovese, D. P. Schneider, and L. Wasserman. Cosmic web reconstruction through density ridges: catalogue. *Monthly Notices of the Royal Astronomical Society*, 461(4):3896–3909, oct 2016.

Y.-C. Chen, S. Ho, R. Mandelbaum, N. A. Bahcall, J. R. Brownstein, P. E. Freeman, C. R. Genovese, D. P. Schneider, and L. Wasserman. Detecting effects of filaments on galaxy properties in the Sloan Digital Sky Survey III. *Monthly Notices of the Royal Astronomical Society*, 466(2):1880–1893, apr 2017.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

H. Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16 (1):31–41, 1964.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

N. R. Chrisman. Calculating on a round planet. *International Journal of Geographical Information Science*, 31(4):637–657, 2017.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Business & Economic Statistics*, pages 1–26, 2025.

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.

M. Colless, B. A. Peterson, C. Jackson, J. A. Peacock, S. Cole, P. Norberg, I. K. Baldry, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, C. Collins, W. Couch, N. Cross, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, W. Sutherland, and K. Taylor. The 2dF Galaxy Redshift Survey: Final Data Release. *arXiv e-prints*, art. astro-ph/0306581, jun 2003.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

J. Comparat, C. Maraston, D. Goddard, V. Gonzalez-Perez, J. Lian, S. Meneses-Goytia, D. Thomas, J. R. Brownstein, R. Tojeiro, A. Finoguenov, A. Merloni, F. Prada, M. Salvato, G. B. Zhu, H. Zou, and J. Brinkmann. Stellar population properties for 2 million galaxies from SDSS DR14 and DEEP2 DR4 from full spectral fitting. *arXiv e-prints*, art. arXiv:1711.06575, nov 2017.

C. Conroy, J. E. Gunn, and M. White. The propagation of uncertainties in stellar population synthesis modeling. i. the relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *The Astrophysical Journal*, 699(1):486, 2009.

D. R. Cox. *Planning of Experiments*. Wiley, 1958.

A. Cuevas. On pattern analysis in the non-convex case. *Kybernetes*, 19(6):26–33, 1990.

A. Cuevas. Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85, 2009.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, pages 2300–2312, 1997.

B. Darvish, B. Mobasher, D. Sobral, A. Rettura, N. Scoville, A. Faisst, and P. Capak. The effects of the local environment and stellar mass on galaxy quenching to z 3. *The Astrophysical Journal*, 825(2):113, 2016.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

K. S. Dawson, D. J. Schlegel, C. P. Ahn, S. F. Anderson, É. Aubourg, S. Bailey, R. H. Barkhouser, J. E. Bautista, A. Beifiori, A. A. Berlind, et al. The baryon oscillation spectroscopic survey of sdss-iii. *The Astronomical Journal*, 145(1):10, 2013.

K. S. Dawson, J.-P. Kneib, W. J. Percival, S. Alam, F. D. Albareti, S. F. Anderson, E. Armengaud, É. Aubourg, S. Bailey, J. E. Bautista, et al. The sdss-iv extended baryon oscillation spectroscopic survey: overview and early data. *The Astronomical Journal*, 151 (2):44, 2016.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data

via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

L. Devroye and G. L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

I. Díaz and N. S. Hejazi. Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):661–683, 2020.

I. Díaz and M. J. van der Laan. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, 1(2):171–192, 2013.

L. H. Dicker. Residual variance and the signal-to-noise ratio in high-dimensional linear models. *arXiv preprint arXiv:1209.0012*, 2012.

M. do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Dover Books on Mathematics. Dover Publications, 2016.

K. Dolag, S. Borgani, G. Murante, and V. Springel. Substructures in hydrodynamical cluster simulations. *Monthly Notices of the Royal Astronomical Society*, 399(2):497–514, 2009.

S. P. Driver, D. T. Hill, L. S. Kelvin, A. S. G. Robotham, J. Liske, P. Norberg, I. K. Baldry, S. P. Bamford, A. M. Hopkins, J. Loveday, J. A. Peacock, E. Andrae, J. Bland-Hawthorn, S. Brough, M. J. I. Brown, E. Cameron, J. H. Y. Ching, M. Colless, C. J. Conselice, S. M. Croom, N. J. G. Cross, R. de Propris, S. Dye, M. J. Drinkwater, S. Ellis, A. W. Graham, M. W. Grootes, M. Gunawardhana, D. H. Jones, E. van Kampen, C. Maraston, R. C. Nichol, H. R. Parkinson, S. Phillipps, K. Pimbblet, C. C. Popescu, M. Prescott,

I. G. Roseboom, E. M. Sadler, A. E. Sansom, R. G. Sharp, D. J. B. Smith, E. Taylor, D. Thomas, R. J. Tuffs, D. Wijesinghe, L. Dunne, C. S. Frenk, M. J. Jarvis, B. F. Madore, M. J. Meyer, M. Seibert, L. Staveley-Smith, W. J. Sutherland, and S. J. Warren. Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release. *Monthly Notices of the Royal Astronomical Society*, 413(2):971–995, may 2011.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

G. M. Eadie, W. E. Harris, and L. M. Widrow. Estimating the galactic mass profile in the presence of incomplete data. *The Astrophysical Journal*, 806(1):54, 2015.

D. Eberly. *Ridges in Image and Data Analysis*. Computational Imaging and Vision. Springer Netherlands, 1996.

B. Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158, 1981.

M. Einasto, E. Tago, H. Lietzen, C. Park, P. Heinämäki, E. Saar, H. Song, L. J. Liivamägi, and J. Einasto. Tracing a high redshift cosmic web with quasar systems. *Astronomy & Astrophysics*, 568:A46, 2014.

U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33(3):1380–1403, 06 2005.

J. Falcón-Barroso, P. Sánchez-Blázquez, A. Vazdekis, E. Ricciardelli, N. Cardiel, A. J. Cenarro, J. Gorgas, and R. F. Peletier. An updated MILES stellar library and stellar population models. *Astronomy & Astrophysics*, 532:A95, aug 2011.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer Birkhäuser Basel, 2013.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1): 32–40, 1975.

W. Gao, F. Xu, and Z.-H. Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022.

E. García-Portugués. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7:1655–1685, 2013.

E. García-Portugués and A. Meilán-Vila. Kernel density estimation with polyspherical data and its applications. *arXiv preprint arXiv:2411.04166*, 2024.

E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga. Kernel density estimation for directional-linear data. *Journal of Multivariate Analysis*, 121:152 – 175, 2013.

E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga. Central limit theorems for directional and linear random variables with applications. *Statistica Sinica*, pages 1207–1229, 2015.

T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, pages 171–185, 1984.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *Annals of Statistics*, 42(4):1511–1545, 08 2014.

Y. A. Ghassabeh. On the convergence of the mean shift algorithm in the one-dimensional space. *Pattern Recognition Letters*, 34(12):1423–1427, 2013.

Y. A. Ghassabeh. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis*, 135:1–10, 2015.

Y. A. Ghassabeh and F. Rudzicz. Modified subspace constrained mean shift algorithm. *Journal of Classification*, 38(1):27–43, 2021.

A. Giessing and J. Wang. Debiased inference on heterogeneous quantile treatment effects with regression rank scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1561–1588, 08 2023.

B. Gilbert, A. Datta, J. A. Casey, and E. L. Ogburn. A causal inference framework for spatial confounding. *arXiv preprint arXiv:2112.14946*, 2023.

R. D. Gill and J. M. Robins. Causal inference for complex longitudinal data: the continuous case. *The Annals of Statistics*, 29(6):1785–1811, 2001.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6): 907–921, 2002.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608 – 1632, 2011.

K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophysical Journal*, 622(2):759–771, apr 2005.

K. Gu, R. Shang, R. Jiang, K. Kuang, R.-J. Lin, D. Lyu, Y. Mao, Y. Pan, T. Wu, J. Yu, Y. Zhang, T. M. Zhang, L. Zhu, M. A. Merrill, J. Heer, and T. Althoff. Blade: Benchmarking language model agents for data-driven science. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971, 2024.

C. Gupta, S. Balakrishnan, and A. Ramdas. Path length bounds for gradient descent and flow. *Journal of Machine Learning Research*, 22(68):1–63, 2021.

L. Guzzo, M. Scodeggio, B. Garilli, B. R. Granett, A. Fritz, U. Abbas, C. Adami, S. Arnouts, J. Bel, M. Bolzonella, D. Bottini, E. Branchini, A. Cappi, J. Coupon, O. Cucciati, I. Davidzon, G. De Lucia, S. de la Torre, P. Franzetti, M. Fumana, P. Hudelot, O. Ilbert, A. Iovino, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Małek, F. Marulli, H. J. McCracken, L. Paioro, J. A. Peacock, M. Polletta, A. Pollo, H. Schlagenhaufer, L. A. M. Tasca, R. Tojeiro, D. Vergani, G. Zamorani, A. Zanichelli, A. Burden, C. Di Porto, A. Marchetti,

C. Marinoni, Y. Mellier, L. Moscardini, R. C. Nichol, W. J. Percival, S. Phleps, and M. Wolk. The VIMOS Public Extragalactic Redshift Survey (VIPERS). An unprecedented view of galaxies and large-scale structure at $0.5 < z < 1.2$. *Astronomy & Astrophysics*, 566:A108, jun 2014.

P. Hall. Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, pages 1156–1174, 1983.

P. Hall, G. S. Watson, and J. Cabrara. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 12 1987.

P. Hall, W. Qian, and D. M. Titterington. Ridge finding from noisy data. *Journal of Computational and Graphical Statistics*, 1(3):197–211, 1992.

P. Hall, L. Peng, and C. Rau. Local likelihood tracking of fault lines and boundaries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):569–582, 2001.

P. Hall, H.-G. Müller, and P.-S. Wu. Real-time density and mode estimation with application to time-dynamic mode tracking. *Journal of Computational and Graphical Statistics*, 15 (1):82–100, 2006.

S. Haneuse and A. Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine*, 32(30):5260–5277, 2013.

A. Haris, N. Simon, and A. Shojaie. Generalized sparse additive models. *Journal of Machine Learning Research*, 23(70):1–56, 2022.

S. He, S. Alam, S. Ferraro, Y.-C. Chen, and S. Ho. The detection of the imprint of filaments on cosmic microwave background lensing. *Nature Astronomy*, 2(5):401–406, 2018.

M. Henmi and S. Eguchi. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91(4):929–941, 2004.

G. E. Hinton. Connectionist learning procedures. In *Machine learning*, pages 555–610. Elsevier, 1990.

K. Hirano and G. W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004.

K. Hitomi, Y. Nishiyama, and R. Okui. A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory*, 24(6):1717–1728, 2008.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge Univ. Press, 1991.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, 2 edition, 2012.

J. Hou, Z. Guo, and T. Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction. *Journal of Machine Learning Research*, 24(265):1–58, 2023.

J. Huang and C.-H. Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *The Journal of Machine Learning Research*, 13(1):1839–1864, 2012.

K. Huang, X. Fu, and N. Sidiropoulos. On convergence of epanechnikov mean shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3191–3198, 2018.

M. Huber, Y.-C. Hsu, Y.-Y. Lee, and L. Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840, 2020.

O. Hyrien and A. Baran. Fast nonparametric density-based clustering of large data sets using a stochastic approximation mean-shift algorithm. *Journal of Computational and Graphical Statistics*, 25(3):899–916, 2016.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

M. C. Irwin. *Smooth dynamical systems*, volume 17. World Scientific, 2001.

Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.

A. J. Izenman. Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):439–446, 2012.

J. C. Jackson. A critique of Rees's theory of primordial gravitational radiation. *Monthly Notices of the Royal Astronomical Society*, 156:1P, jan 1972.

J. Jankova and S. van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014a.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

D. H. Jones, M. A. Read, W. Saunders, M. Colless, T. Jarrett, Q. A. Parker, A. P. Fairall, T. Mauch, E. M. Sadler, F. G. Watson, D. Burton, L. A. Campbell, P. Cass, S. M. Croom, J. Dawe, K. Fiegert, L. Frankcombe, M. Hartley, J. Huchra, D. James, E. Kirby, O. Lahav, J. Lucey, G. A. Mamon, L. Moore, B. A. Peterson, S. Prior, D. Proust, K. Russell,

V. Safouris, K.-I. Wakamatsu, E. Westra, and M. Williams. The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures. *Monthly Notices of the Royal Astronomical Society*, 399(2):683–698, oct 2009.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

M. Kafai, Y. Miao, and K. Okada. Directional mean shift and its application for topology classification of local 3d structures. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 170–177. IEEE, 2010.

N. Kaiser. Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society*, 227(1):1–21, 1987.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2018.

E. Kammann and M. P. Wand. Geoadditive models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 52(1):1–18, 2003.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

J. P. Keller and A. A. Szpiro. Selecting a scale for spatial confounding adjustment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3):1121–1143, 2020.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

J. Klemelä. Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, 73(1):18–40, 2000.

T. Kobayashi and N. Otsu. Von mises-fisher mean shift for clustering on a hypersphere. In *20th International Conference on Pattern Recognition*, pages 2130–2133. IEEE, 2010.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

K. Kraljic, S. Arnouts, C. Pichon, C. Laigle, S. de la Torre, D. Vibert, C. Cadiou, Y. Dubois, M. Treyer, C. Schimd, et al. Galaxy evolution in the metric of the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 474(1):547–571, 2018.

C. D. Kreisch, A. Pisani, F. Villaescusa-Navarro, D. N. Spergel, B. D. Wandelt, N. Hamaus, and A. E. Bayer. The GIGANTES dataset: precision cosmology from voids in the machine learning era. *The Astrophysical Journal*, 935(2):100, 2022.

U. Kuchner, A. Aragón-Salamanca, F. R. Pearce, M. E. Gray, A. Rost, C. Mu, C. Welker, W. Cui, R. Haggar, C. Laigle, A. Knebe, K. Kraljic, F. Sarron, and G. Yepes. Mapping and characterization of cosmic filaments in galaxy cluster outskirts: strategies and forecasts for observations from simulations. *Monthly Notices of the Royal Astronomical Society*, 494 (4):5473–5491, jun 2020.

U. Kuchner, A. Aragón-Salamanca, A. Rost, F. R. Pearce, M. E. Gray, W. Cui, A. Knebe, E. Rasia, and G. Yepes. Cosmic filaments in galaxy cluster outskirts: quantifying finding

filaments in redshift space. *Monthly Notices of the Royal Astronomical Society*, 503(2): 2065–2076, may 2021.

O. Le Fèvre, G. Vettolani, B. Garilli, L. Tresse, D. Bottini, V. Le Brun, D. Maccagni, J. P. Picat, R. Scaramella, M. Scodeggio, A. Zanichelli, C. Adami, M. Arnaboldi, S. Arnouts, S. Bardelli, M. Bolzonella, A. Cappi, S. Charlot, P. Ciliegi, T. Contini, S. Foucaud, P. Franzetti, I. Gavignaud, L. Guzzo, O. Ilbert, A. Iovino, H. J. McCracken, B. Marano, C. Marinoni, G. Mathez, A. Mazure, B. Meneux, R. Merighi, S. Paltani, R. Pellò, A. Pollo, L. Pozzetti, M. Radovich, G. Zamorani, E. Zucca, M. Bondi, A. Bongiorno, G. Busarello, F. Lamareille, Y. Mellier, P. Merluzzi, V. Ripepi, and D. Rizzo. The VIMOS VLT deep survey. First epoch VVDS-deep survey: 11 564 spectra with $17.5 \leq \mathrm{IAB} \leq 24$, and the redshift distribution over $0 \leq \mathrm{z} \leq 5$. *Astronomy & Astrophysics*, 439(3):845–862, sep 2005.

F. Lecci, A. Rinaldo, and L. Wasserman. Statistical analysis of metric graph reconstruction. *Journal of Machine Learning Research*, 15(1):3425–3446, 2014.

J. Lee and D. Park. Constraining the dark energy equation of state with cosmic voids. *The Astrophysical Journal*, 696(1):L10, 2009.

J. M. Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

J. M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, second edition, 2012.

J. M. Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

O. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

X. Li, Z. Hu, and F. Wu. A note on the convergence of the mean shift. *Pattern Recognition*, 40(6):1756–1762, 2007.

N. I. Libeskind, R. van de Weygaert, M. Cautun, B. Falck, E. Tempel, T. Abel, M. Alpaslan, M. A. Aragón-Calvo, J. E. Forero-Romero, R. Gonzalez, S. Gottlöber, O. Hahn, W. A. Hellwing, Y. Hoffman, B. J. T. Jones, F. Kitaura, A. Knebe, S. Manti, M. Neyrinck, S. E. Nuza, N. Padilla, E. Platen, N. Ramachandra, A. Robotham, E. Saar, S. Shandarin, M. Steinmetz, R. S. Stoica, T. Sousbie, and G. Yepes. Tracing the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 473(1):1195–1217, jan 2018.

J. K. Lindsey. *Statistical analysis of stochastic processes in time*, volume 14. Cambridge University Press, 2004.

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 3rd edition, 2019.

S. Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

J. J. Lok. How estimating nuisance parameters can reduce the variance (with consistent variance estimation). *arXiv preprint arXiv:2109.02690*, 2021.

B. W. Lyke, A. N. Higley, J. McLane, D. P. Schurhammer, A. D. Myers, A. J. Ross, K. Dawson, S. Chabanier, P. Martini, N. G. Busca, et al. The sloan digital sky survey quasar catalog: Sixteenth data release. *The Astrophysical Journal Supplement Series*, 250(1):8, 2020.

Y. Mack and H.-G. Müller. Derivative estimation in nonparametric regression with random predictor variable. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 59–72, 1989.

S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy, C. A. Prieto, R. Barkhouser, D. Bizyaev, B. Blank, S. Brunner, A. Burton, R. Carrera, et al. The apache point observatory galactic evolution experiment (apogee). *The Astronomical Journal*, 154(3):94, 2017.

N. Malavasi, M. Langer, N. Aghanim, D. Galárraga-Espinosa, and C. Gouin. Relative effect of nodes and filaments of the cosmic web on the quenching of galaxies and the orientation of their spin. *Astronomy & Astrophysics*, 658:A113, feb 2022.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

C. Maraston and G. Strömbäck. Stellar population models at high spectral resolution. *Monthly Notices of the Royal Astronomical Society*, 418(4):2785–2811, 2011.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, UK, 2000.

C. Marzban, Y. Zhang, N. Bond, and M. Richman. A practical introduction to regression-based causal inference in meteorology (i): All confounders measured. *arXiv preprint arXiv:2506.18808*, 2025a.

C. Marzban, Y. Zhang, N. Bond, and M. Richman. A practical introduction to regression-based causal inference in meteorology (ii): Unmeasured confounders. *arXiv preprint arXiv:2506.18652*, 2025b.

M. D. Marzio, A. Panzera, and C. C. Taylor. Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, 141(6):2156–2173, 2011.

J. A. Mauro, E. H. Kennedy, and D. Nagin. Instrumental variable methods using dynamic interventions. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183 (4):1523–1551, 2020.

A. McClean, Z. Branson, and E. H. Kennedy. Nonparametric estimation of conditional incremental effects. *Journal of Causal Inference*, 12(1):20230024, 2024.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2 edition, 2008.

A. Mokkadem and M. Pelletier. The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM: Probability and Statistics*, 7:1–21, 2003.

S. Mucesh, W. G. Hartley, C. M. Gilligan-Lee, and O. Lahav. Nature versus nurture in galaxy formation: the effect of environment on star formation with causal machine learning. *arXiv preprint arXiv:2412.02439*, 2024.

P. Müller and S. van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42(2):580–608, 2015.

U. U. Müller and I. V. Keilegom. Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics*, 6(none):1200 – 1219, 2012.

I. D. Muñoz and M. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

R. Neugebauer and M. van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

W. K. Newey and J. R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

J. Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and Statsitics*, pages 213–234, 1959.

J. Neyman. C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 41(1/2):1–21, 1979.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition, 2006.

S. Oba, K. Kato, and S. Ishii. Multi-scale clustering for gene expression profiling data. In *Proceedings of Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*, pages 210–217. IEEE, 2005.

E. A. Ok. *Real Analysis with Economic Applications*, volume 10. Princeton University Press, 2007.

M. Oliveira, R. M. Crujeiras, and A. Rodríguez-Casal. A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis*, 56(12):3898–3908, 2012.

U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12(34):1249–1286, 2011.

C. J. Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107–125, 2010.

E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

M. Pasquato, Z. Jin, P. Lemos, B. L. Davis, and A. V. Macciò. Causa prima: cosmology meets causal discovery for the first time. *arXiv preprint arXiv:2311.15160*, 2023.

J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

A. Pewsey and E. García-Portugués. Recent advances in directional statistics. *Test*, 30(1): 1–58, 2021.

S. Pfeifer, N. I. Libeskind, Y. Hoffman, W. A. Hellwing, M. Bilicki, and K. Naidoo. COWS: A filament finder for Hessian cosmic web identifiers. *arXiv e-prints*, art. arXiv:2201.04624, jan 2022.

A. Pillepich, D. Nelson, L. Hernquist, V. Springel, R. Pakmor, P. Torrey, R. Weinberger, S. Genel, J. P. Naiman, F. Marinacci, et al. First results from the illustristng simulations: the stellar mass content of groups and clusters of galaxies. *Monthly Notices of the Royal Astronomical Society*, 475(1):648–675, 2018a.

A. Pillepich, V. Springel, D. Nelson, S. Genel, J. Naiman, R. Pakmor, L. Hernquist, P. Torrey, M. Vogelsberger, R. Weinberger, et al. Simulating galaxy formation with the illustristng model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, 2018b.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

B. Polyak. Gradient methods for the minimisation of functionals. *Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

J. L. Powell and T. M. Stoker. Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, 75(2):291–316, 1996.

W. Qiao. Asymptotic confidence regions for density ridges. *Bernoulli*, 27(2):946–975, 2021.

W. Qiao. Confidence regions for filamentary structures. *Information and Inference: A Journal of the IMA*, 14(4):iaaf030, 2025.

W. Qiao and W. Polonik. Theoretical analysis of nonparametric filament estimation. *Annals of Statistics*, 44(3):1269–1297, 06 2016.

W. Qiao and W. Polonik. Algorithms for ridge estimation with convergence guarantees. *arXiv preprint arXiv:2104.12314*, 2021.

P. Rakshit, A. Levis, and L. Keele. Local effects of continuous instruments without positivity. *arXiv preprint arXiv:2409.07350*, 2024.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

J. Renegar. *A mathematical view of interior-point methods in convex optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2001.

T. S. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, page 454–461, San Francisco, CA, USA, 1996.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

J. M. Robins. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987.

J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

J. P. Romano. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, 16(2):629–647, 1988.

P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

M. Rosenblatt. Short range and long range dependence. *Asymptotic Laws and Methods in Stochastics: A Volume in Honour of Miklós Csörgő*, pages 283–294, 2015.

D. Rothenhäusler and B. Yu. Incremental causal effects. *arXiv preprint arXiv:1907.13258*, 2019.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill New York, third edition, 1976.

P. Saavedra-Nieves and R. María Crujeiras. Nonparametric estimation of directional highest density regions. *arXiv e-prints*, art. arXiv:2009.08915, sep 2020.

C. Sánchez, J. Clampitt, A. Kovacs, B. Jain, J. García-Bellido, S. Nadathur, D. Gruen, N. Hamaus, D. Huterer, P. Vielzeuf, et al. Cosmic voids and void lensing in the dark energy survey science verification data. *Monthly Notices of the Royal Astronomical Society*, page stw2745, 2016.

W. L. Sargent and E. L. Turner. A statistical method for determining the cosmological density parameter from the redshifts of a complete sample of galaxies. *Astrophysical Journal, Part 2-Letters to the Editor, vol. 212, Feb. 15, 1977, p. L3-L7.*, 212:L3–L7, 1977.

H. Sasaki, T. Kanamori, A. Hyvärinen, G. Niu, and M. Sugiyama. Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios. *Journal of Machine Learning Research*, 18(180):1–47, 2018.

A. Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139 – 1151, 1986.

K. Schindl and L. Wasserman. Causal geodesy: Counterfactual estimation along the path between correlation and causation. *arXiv preprint arXiv:2508.08499*, 2025.

K. Schindl, S. Shen, and E. H. Kennedy. Incremental effects for continuous exposures. *arXiv preprint arXiv:2409.11967*, 2024.

P. Schnell and G. Papadogeorgou. Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths. *Annals of Applied Statistics*, 14:2069–2095, 12 2020.

D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2015.

N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J.-P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, and N. D. Tyson. The cosmic evolution survey (COSMOS): Overview. *The Astrophysical Journal Supplement Series*, 172(1):1–8, sep 2007.

N. Scoville, S. Arnouts, H. Aussel, A. Benson, A. Bongiorno, K. Bundy, M. Calvo, P. Capak, M. Carollo, F. Civano, et al. Evolution of galaxies and their environments at z= 0.1–3 in cosmos. *The Astrophysical Journal Supplement Series*, 206(1):3, 2013.

J. Shao. *Mathematical Statistics*. Springer Science & Business Media, 2003.

S. J. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 11 2004.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):683–690, 1991.

Shou-Jen Chang-Chien, M. Yang, and Wen-Liang Hung. Mean shift-based clustering for directional data. In *Proceedings of the Third International Workshop on Advanced Computational Intelligence*, pages 367–372, Aug 2010.

D. Sijacki, M. Vogelsberger, S. Genel, V. Springel, P. Torrey, G. F. Snyder, D. Nelson, and L. Hernquist. The Illustris simulation: the evolving population of black holes across cosmic time. *Monthly Notices of the Royal Astronomical Society*, 452(1):575–596, sep 2015.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

E. Smucler, A. Rotnitzky, and J. M. Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

J. Snyder, P. Voxland, and G. S. (U.S.). *An Album of Map Projections*. Number 1453 in An Album of Map Projections. U.S. Government Printing Office, 1989.

S. Song, Y. Lin, and Y. Zhou. Semi-supervised inference for block-wise missing data without imputation. *Journal of Machine Learning Research*, 25(99):1–36, 2024.

T. Sousbie. The persistent cosmic web and its filamentary structure - I. Theory and implementation. *Monthly Notices of the Royal Astronomical Society*, 414(1):350–383, jun 2011.

P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied informatics*, 3(3), 2016.

V. Springel. E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. *Monthly Notices of the Royal Astronomical Society*, 401(2):791–851, 2010.

V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435(7042):629–636, jun 2005.

V. Springel, R. Pakmor, A. Pillepich, R. Weinberger, D. Nelson, L. Hernquist, M. Vogelsberger, S. Genel, P. Torrey, F. Marinacci, et al. First results from the illustristng simulations: matter and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 475(1):676–698, 2018.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, pages 1285–1297, 1984.

C. J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

F. Su, W. Mou, P. Ding, and M. Wainwright. When is the estimated propensity score better? high-dimensional analysis and bias correction. *arXiv preprint arXiv:2303.17102*, 2023.

R. Subbarao and P. Meer. Nonlinear mean shift for clustering over analytic manifolds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 1168–1175. IEEE, 2006.

R. Subbarao and P. Meer. Nonlinear mean shift over riemannian manifolds. *International journal of computer vision*, 84(1):1–20, 2009.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

P. Sutter, A. Pisani, B. D. Wandelt, and D. H. Weinberg. A measurement of the Alcock–Paczyński effect using cosmic voids in the SDSS. *Monthly Notices of the Royal Astronomical Society*, 443(4):2983–2990, 2014.

Z. Tan and C.-H. Zhang. Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics*, 47(5):2567 – 2600, 2019.

C. C. Taylor. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52(7):3493–3500, 2008.

E. Tempel, E. Tago, and L. Liivamägi. Groups and clusters of galaxies in the sdss dr8-value-added catalogues. *Astronomy & Astrophysics*, 540:A106, 2012.

E. Tempel, R. S. Stoica, V. J. Martínez, L. J. Liivamägi, G. Castellan, and E. Saar. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS. *Monthly Notices of the Royal Astronomical Society*, 438(4):3465–3482, mar 2014.

H. Thaden and T. Kneib. Structural equation models for dealing with spatial confounding. *The American Statistician*, 72(3):239–252, 2018.

Y. Tian, P. Wu, and Z. Tan. Semi-supervised regression analysis with model misspecification and high-dimensional data. *arXiv preprint arXiv:2406.13906*, 2024.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007.

A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 1, pages 18–25. IEEE, 2005.

A. Valade, N. I. Libeskind, D. Pomarede, R. B. Tully, Y. Hoffman, S. Pfeifer, and E. Kourkchi. Identification of basins of attraction in the local universe. *Nature Astronomy*, 8(12):1610–1616, 2024.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, 2014.

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, Berkeley, CA, 2003.

M. J. van der Laan and M. L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1): Article3, 2007.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

R. Vershynin. *High-dimensional Probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

P. Vieu. A note on density mode estimation. *Statistics & Probability Letters*, 26(4):297–307, 1996.

M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, S. Bird, D. Nelson, and L. Hernquist. Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509(7499):177–182, may 2014.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, dec 2007.

J. Wagener and H. Dette. The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, 22:137–154, 2013.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

M. Wand and M. Jones. *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

M. P. Wand and M. C. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528, 1993.

X. Wang, W. Qiu, and J. Wu. Convergence and stability analysis of mean-shift algorithm on large data sets. *Statistics and Its Interface*, 9(2):159–170, 2016.

L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag, Berlin, Heidelberg, 2006.

L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.

D. Westreich and S. R. Cole. Invited commentary: positivity in practice. *American Journal of Epidemiology*, 171(6):674–677, 2010.

N. Wiecha and B. J. Reich. Two-stage spatial regression models for spatial confounding. *arXiv preprint arXiv:2404.09358*, 2024.

D. M. Wilkinson, C. Maraston, D. Goddard, D. Thomas, and T. Parikh. FIREFLY (Fitting IteRativEly For Likelihood analYsis): a full spectral fitting code. *Monthly Notices of the Royal Astronomical Society*, 472(4):4297–4326, dec 2017.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

F. Xue, R. Ma, and H. Li. Semi-supervised statistical inference for high-dimensional linear regression with blockwise missing data. *arXiv preprint arXiv:2106.03344*, 2021.

M.-S. Yang, S.-J. Chang-Chien, and H.-C. Kuo. On mean shift clustering for directional data on a hypersphere. In *Proceedings of the Artificial Intelligence and Soft Computing*, pages 809–818, Cham, 2014. Springer International Publishing.

Y. Yang, Y. Zhang, X. Shen, and Y.-C. Chen. Emputation: Identification-guided neural imputation framework. *arXiv preprint arXiv:2607.05279*, 2026.

D. G. York et al. The Sloan Digital Sky Survey: Technical summary. *Astronomical Journal*, 120(3):1579–1587, 2000.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102:315–323, 2014.

X.-T. Yuan and S. Z. Li. Stochastic gradient kernel density mode-seeking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1926–1931, 2009.

A. Zanga, E. Ozkirimli, and F. Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Y. B. Zel'Dovich. Gravitational instability: An approximate theory for large density perturbations. *Astronomy and Astrophysics*, 5:84–89, 1970.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

H. Zhang. The optimality of naive bayes. In V. Barr and Z. Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press, 2004.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

K. Zhang, J. T. Kwok, and M. Tang. Accelerated convergence using dynamic mean shift. In *Proceedings of the European Conference on Computer Vision*, pages 257–268. Springer Berlin Heidelberg, 2006.

Y. Zhang. SDSS-IV cosmic web catalog, June 2022. URL https://doi.org/10.5281/zenodo.6244866.

Y. Zhang and Y.-C. Chen. The em perspective of directional mean shift algorithm. *arXiv preprint arXiv:2101.10058*, 2021a.

Y. Zhang and Y.-C. Chen. Kernel smoothing, mean shift, and their learning theory with directional data. *Journal of Machine Learning Research*, 22(154):1–92, 2021b.

Y. Zhang and Y.-C. Chen. Linear convergence of the subspace constrained mean shift algorithm: from euclidean to directional data. *Information and Inference: A Journal of the IMA*, 12(1):210–311, 2023.

Y. Zhang and Y.-C. Chen. Doubly robust inference on causal derivative effects for continuous treatments. *arXiv preprint arXiv:2501.06969*, 2025.

Y. Zhang and Y.-C. Chen. Mode and ridge estimation in euclidean and directional product spaces: A mean shift approach. *Journal of Computational and Graphical Statistics*, 35(1): 101–110, 2026.

Y. Zhang, R. S. de Souza, and Y.-C. Chen. Sconce: a cosmic web finder for spherical and conic geometries. *Monthly Notices of the Royal Astronomical Society*, 517(1):1197–1217, 2022.

Y. Zhang, A. Giessing, and Y.-C. Chen. *DebiasInfer: Efficient Inference on High-Dimensional Linear Model with Missing Outcomes*, 2023a. URL https://CRAN.R-project.org/package=DebiasInfer. R package version 0.2.1.

Y. Zhang, A. Giessing, and Y.-C. Chen. Efficient inference on high-dimensional linear models with missing outcomes. *arXiv preprint arXiv:2309.06429*, 2023b.

Y. Zhang, Y.-C. Chen, and A. Giessing. Nonparametric inference on dose-response curves without the positivity condition. *arXiv preprint arXiv:2405.09003*, 2024.

Y. Zhang, S. Wilkins-Reeves, W. Lee, and A. Hofleitner. Transfer learning through conditional quantile matching. *arXiv preprint arXiv:2602.02358*, 2026.

C. Zhao, C. Tao, Y. Liang, F.-S. Kitaura, and C.-H. Chuang. Dive in the cosmic web: voids with delaunay triangulation from discrete matter tracer distributions. *Monthly Notices of the Royal Astronomical Society*, 459(3):2670–2680, 2016.

L. Zhao and C. Wu. Central limit theorem for integrated square error of kernel estimators of spherical density. *Science in China A: Mathematics*, 44(4):474–483, apr 2001.

W. Zhu and L.-L. Feng. Evolution of mass and velocity field in the cosmic web: Comparison between baryonic and dark matter. *The Astrophysical Journal*, 838(1):21, 2017.

O. Zimmermann and U. H. Hansmann. Support vector machines for prediction of dihedral angle regions. *Bioinformatics*, 22(24):3009–3015, 2006.

F. Ziparo, P. Popesso, A. Finoguenov, A. Biviano, S. Wuyts, D. Wilman, M. Salvato, M. Tanaka, K. Nandra, D. Lutz, et al. Reversal or no reversal: the evolution of the star formation rate–density relation up to z 1.6. *Monthly Notices of the Royal Astronomical Society*, 437(1):458–474, 2014.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Limitations of Euclidean KDE in Handling Directional Data

In this section, we demonstrate with examples and simulation studies that it is inadequate to analyze angular or directional data with Euclidean KDE (2.1) and SCMS algorithm (Algorithm 1). Consider a directional data sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_2$ generated from a directional density f on Ω_2. In real-world applications, the random observations $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ on Ω_2 are commonly represented by their angular coordinates $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ with $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}) \in [-180°, 180°) \times [-90°, 90°]$ or equivalently, $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}) \in [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ for $i = 1, ..., n$, where $\{Y_{i,1}\}_{i=1}^n$ are longitudes and $\{Y_{i,2}\}_{i=1}^n$ are latitudes.

A.1.1 Case I: Density Estimation

As the angular coordinates $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n\}$ of the directional dataset $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_2$ take values in a subset $[-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ of the flat Euclidean space $\mathbb{R}^2$, it is tempting to apply the Euclidean KDE on $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n\}$ to construct a density estimator as:

$$\widehat{p}_n(\boldsymbol{y}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{y} - \boldsymbol{Y}_i}{h}\right) = \begin{cases} \frac{c_{k,2}}{nh^2} \sum_{i=1}^{n} k\left(\left|\left|\frac{\boldsymbol{y}-\boldsymbol{Y}_i}{h}\right|\right|_2^2\right) := \widehat{p}_n^{(1)}(\boldsymbol{y}), \\ \frac{1}{nh^2} \sum_{i=1}^{n} K_1\left(\frac{y_1 - Y_{i,1}}{h}\right) \cdot K_2\left(\frac{y_2 - Y_{i,2}}{h}\right) := \widehat{p}_n^{(2)}(\boldsymbol{y}), \end{cases} \tag{A.1}$$

where $\widehat{p}_n^{(1)}$ uses a radial symmetric kernel with profile k, and $\widehat{p}_n^{(2)}$ leverages a product kernel. However, the Euclidean KDEs in (A.1) (both $\widehat{p}_n^{(1)}$ and $\widehat{p}_n^{(2)}$) exhibit two potential drawbacks when applied to directional data.

• First, $\widehat{p}_n(\boldsymbol{y})$ in (A.1) is an estimator of the directional density f under its angular representation $p_f : [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to \mathbb{R}$. Here, f is 2π-periodic in its first coordinate and

π-periodic in its second coordinate. Then, the bias of $\widehat{p}_n(\boldsymbol{y})$ in estimating $p_f(\boldsymbol{y})$ is

$$\mathbb{E}\left[\widehat{p}_n(\boldsymbol{y})\right] - p_f(\boldsymbol{y}) = \frac{h^2}{2} C_K^2 \Delta p_f(\boldsymbol{y}) + o(h^2),$$

where $C_K^2 = \int_{\mathbb{R}^2} ||\boldsymbol{y}||_2^2 \, K(\boldsymbol{y}) d\boldsymbol{y}$ and $\Delta p_f = \frac{\partial^2 p_f}{\partial y_1^2} + \frac{\partial^2 p_f}{\partial y_2^2}$ is the Laplacian of p_f; see Chen (2017) for details. However, the second-order partial derivative $\frac{\partial^2 p_f}{\partial y_1^2}$ along the lines of constant latitude (or parallels) would tend to infinity as we approach the north and south poles, given that the first-order partial derivative $\frac{\partial p_f}{\partial y_1}$ is bounded. One method to justify this claim is that the curvatures of these parallels, which are equivalent to the reciprocals of their radii, tend to infinity as these radii shrink. In addition, one should recall that the curvature of a function $y = g(x)$ is defined as $\varkappa := \frac{|g''|}{(1+g'^2)^{\frac{3}{2}}}$. Therefore, applying (A.1) to estimate the angular representation p_f of the directional density f will produce large bias as the estimator $\widehat{p}_n$ approaches the high-latitude regions (around the north and south poles); see also Panel (c) of Figure A.2.

• Second, the Euclidean KDE $\widehat{p}_n$ leverages the Euclidean distances between any query point $\boldsymbol{y} \in [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and observations $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n\} \subset [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ under their angular coordinates to construct the density estimates, instead of using the (intrinsic) geodesic distances. Note that the Euclidean distance in the angular coordinate system is not equivalent to the Euclidean distance in the ambient Euclidean space $\mathbb{R}^3$ containing the directional data on Ω_2. As a result, some observations that have dramatically different geodesic distances to density query points can have the same density contributions in $\widehat{p}_n$, as illustrated in Example A.1.

Example A.1. *Suppose that we want to estimate the density values at* $\boldsymbol{y}_1 = (0, 0)$ *and* $\boldsymbol{y}_2 = (0, \frac{\pi}{2} - \epsilon)$*, where* $\epsilon > 0$ *is small. Consider a random sample consisting of only two observations* $\boldsymbol{Y}_1 = (\frac{\pi}{2}, 0)$ *and* $\boldsymbol{Y}_2 = (\frac{\pi}{2}, \frac{\pi}{2} - \epsilon)$*. If we use the Euclidean distance, the distance between* $(\boldsymbol{y}_1, \boldsymbol{Y}_1)$ *and the distance between* $(\boldsymbol{y}_2, \boldsymbol{Y}_2)$ *are the same. Therefore, when we use the Euclidean KDE* $\widehat{p}_n$ *to estimate the underlying density, the contribution of* $\boldsymbol{Y}_1$ *to* $\boldsymbol{y}_1$ *will be the same as the contribution of* $\boldsymbol{Y}_2$ *to* $\boldsymbol{y}_2$*. Nevertheless, their geodesic distances are very different, because* $d_g(\boldsymbol{y}_1, \boldsymbol{Y}_1) = \frac{\pi}{2}$ *while* $d_g(\boldsymbol{y}_2, \boldsymbol{Y}_2) = \arccos\left[\sin^2\left(\frac{\pi}{2} - \epsilon\right)\right] = \arccos\left[\frac{1+\cos(2\epsilon)}{2}\right]$

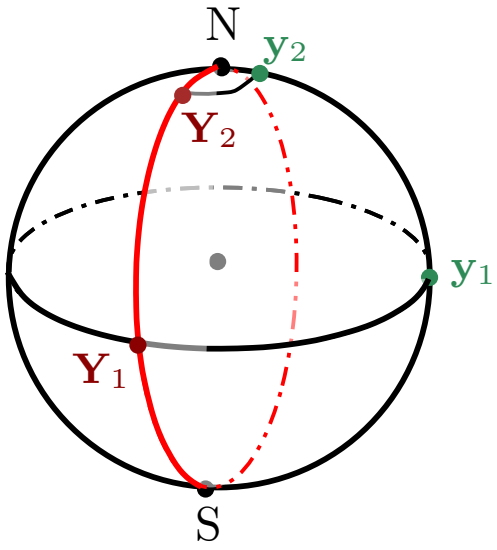


Figure A.1: Graphical illustration of geodesic distances between $\boldsymbol{y}_1$ and $\boldsymbol{Y}_1$ as well as $\boldsymbol{y}_2$ and $\boldsymbol{Y}_2$.

is a quantity close to zero; see Figure A.1 for a graphical illustration. It explains, from a different angle, why the Euclidean KDE $\widehat{p}_n$ will have a large bias in estimating the underlying density when the query point $\boldsymbol{y}$ is within the high-latitude region.

A.1.2 Case II: Ridge-Finding Problem

Consider the following simulated example of identifying a density ridge via the Euclidean SCMS algorithm (Algorithm 1) and our proposed directional SCMS algorithm (Algorithm 2). We generate 1000 data points $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{1000}\} \subset \Omega_2$ uniformly from a great circle connecting the North and South Poles of Ω_2, add i.i.d. Gaussian noise $N(0, 0.2^2)$ to their Cartesian coordinates, and normalize the resulting points back onto Ω_2 using the L_2 norm. The angular coordinates of these simulated points are denoted by $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{1000}\} \subset [-180°, 180°) \times [-90°, 90°]$. Figure A.2 presents the result of applying both the Euclidean SCMS algorithm (with the Gaussian kernel) to angular coordinates and the directional SCMS algorithm (with the von Mises kernel) to Cartesian coordinates of our simulated dataset. As shown in panel (b) of Figure A.2, the Euclidean SCMS algorithm exhibits high bias in es-

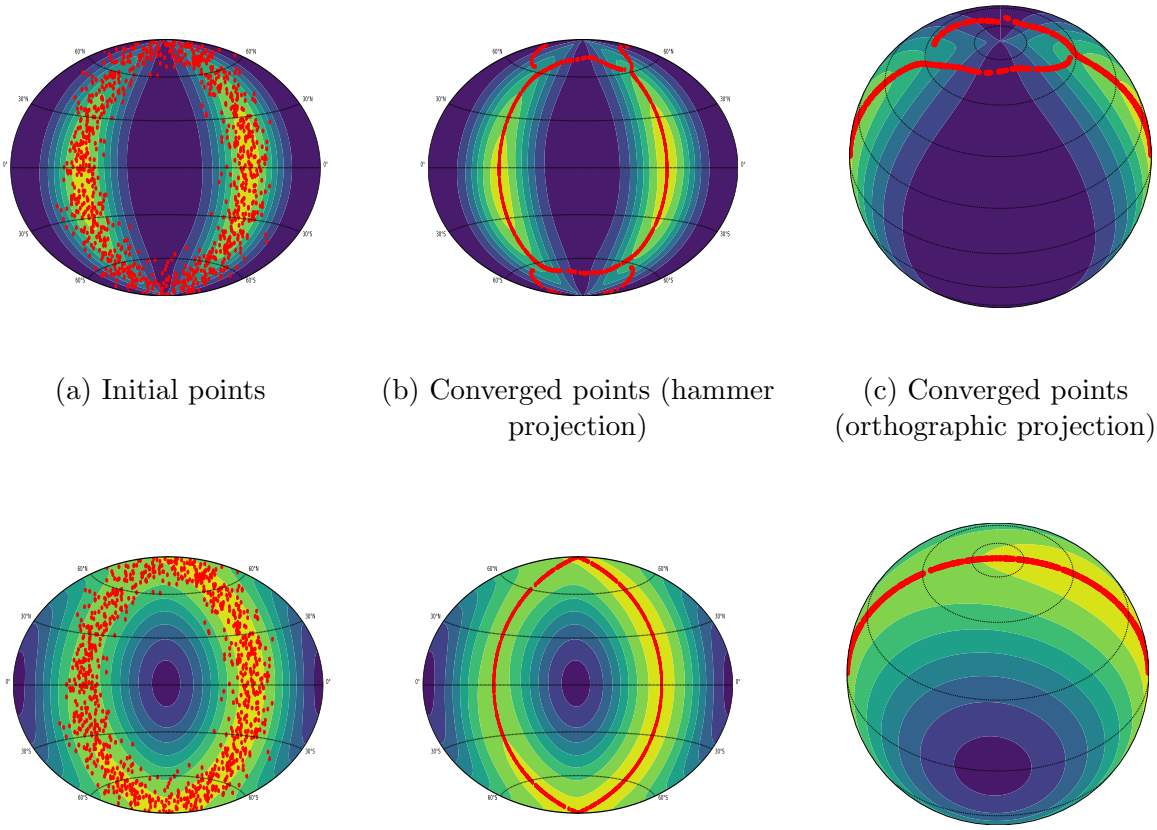
timating the true circular structure near the two poles of Ω_2, while our directional SCMS algorithm is able to recover the true circular structure with negligible error. The density plot in panel (c) of Figure A.2 exhibits two nonsmooth peaks at the North Pole because the Hessian of the underlying density is unbounded in angular coordinates; recall our discussion in Section A.1.1. This also explains the chaotic behavior of the Euclidean KDE in high-latitude regions.

At this point, some readers may have a natural concern: why we do not directly apply the Euclidean SCMS algorithm to the Cartesian coordinates $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$ of the available data points? We discuss the potential downsides of this approach from two different aspects.

1. The Euclidean SCMS algorithm is not intrinsically designed for handling the directional data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$. Directly applying the algorithm to these Cartesian coordinates leads to an estimated ridge not lying on Ω_q. While the L_2 normalization is able to standardize the ridge points back to Ω_q, this standardization process will inevitably introduce extra bias.

2. When estimating the underlying density of $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$, we know from the KDE literature (Chacón et al., 2011; Scott, 2015) that the (uniform) rates of convergence of the Euclidean KDE and its derivatives depend on the dimension $(q+1)$ of the ambient space instead of the intrinsic dimension q of directional data. This dimensionality effect also appears in the (linear) convergence of the downstream SCMS algorithm, which, for instance, shrinks the upper bounds of the (linear) convergence radius and step size threshold in Theorem 3.10. Thus, analyzing directional data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$ with the Euclidean KDE and SCMS algorithm will slow down the statistical and algorithmic rates of convergence of the density estimators as well as reduce the accuracy of the resulting ridge in recovering the underlying structure inside the dataset.

To support the preceding explanations, we extend our simulation study in Figure A.2 as follows. We vary the maximum latitude attained by the underlying (intrinsic) circular structure on Ω_2 from $45°$ to $90°$ while keeping the circle parallel to the original great



(a) Initial points

(b) Converged points (hammer projection)

(c) Converged points (orthographic projection)

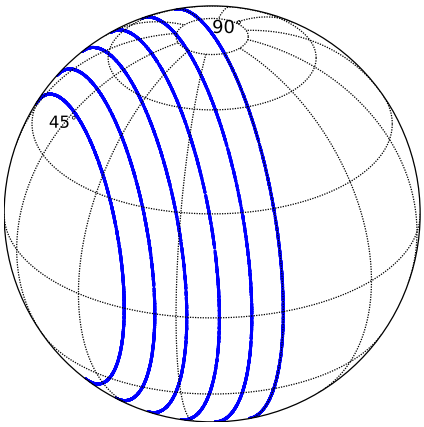
(d) Initial points

(e) Converged points (hammer projection)

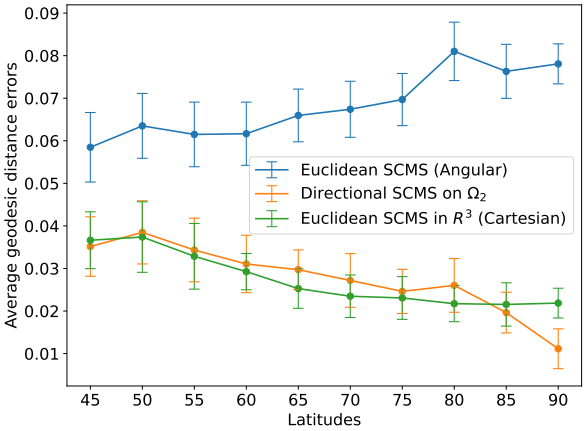
(f) Converged points (orthographic projection)

Figure A.2: Euclidean and directional SCMS algorithms performed on the simulated dataset. **Panels (a)-(c):** Outcomes of the Euclidean SCMS algorithm with the contour plot for the Euclidean KDE. **Panels (d)-(f):** Outcomes of our directional SCMS algorithm with the contour plot for the directional KDE. **Panels (a)-(b) and (d)-(e)** are displayed using Hammer projections (page 160 in Snyder et al. 1989), while **Panels (c) and (f)** are presented under the orthographic projections.

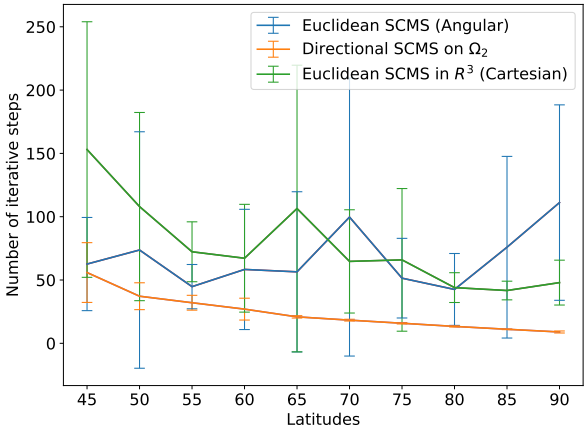
circle connecting the North and South Poles of Ω_2; see panel (a) in Figure A.3 for an illustration. For each of these underlying circles, we follow the same sampling scheme as in Figure A.2 by sampling 1000 points uniformly on the circle, adding i.i.d. Gaussian noise $N(0, 0.2^2)$ to their Cartesian coordinates, and normalizing them back onto Ω_2 using the



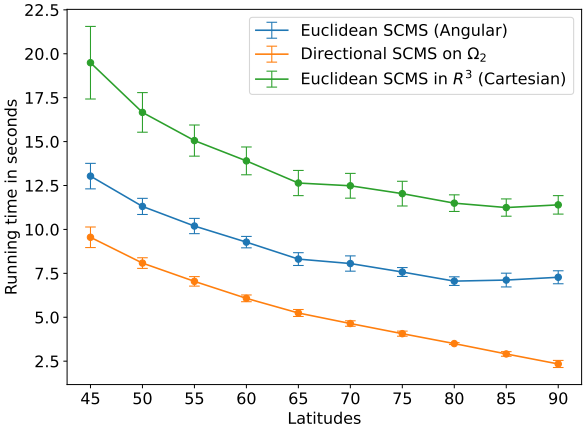
(a) Depiction of underlying true circular structures.



(b) Average geodesic distance errors.



(c) Number of iteration steps.



(d) Running time.

Figure A.3: Euclidean and directional SCMS algorithms applied to the simulated datasets whose true structures are circles on Ω_2 attaining their maximum latitudes from 45° to 90°, respectively. The dots on each line plot in the panels (b-d) are the means of the associated statistics for the repeated experiments, while the error bars indicate their corresponding standard deviations.

L_2 norm. The Cartesian coordinates of the simulated points from each circular structure are denoted by $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{1000}\} \subset \Omega_2$ while their angular coordinates are represented by $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{1000}\} \subset [-180°, 180°) \times [-90°, 90°]$. Then, we apply our directional SCMS algorithm to $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{1000}\}$ from each of these simulated datasets. Moreover, the Euclidean SCMS algorithm is applied to both the angular coordinates $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{1000}\}$ and Cartesian coordinates $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{1000}\}$ from each of these simulated datasets, where we consider $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_{1000}\}$ as a dataset in the ambient space $\mathbb{R}^3$ in the latter case. Here, the sets of initial points for the Euclidean and directional SCMS algorithms are the simulated datasets themselves. Finally, we compute the average geodesic distance errors on Ω_2 from the resulting ridges to the corresponding true circular structures. To reduce the randomness of our simulation studies, we also repeat the above sampling and experimental procedures 20 times for each true circular structure.

We present our comparisons of the Euclidean and directional SCMS algorithms based on three metrics in Figure A.3: (i) average geodesic distance errors between the estimated ridges and the true circular structures, (ii) the number of iteration steps, and (iii) the running time. Notice that, as the latitudes of the underlying circular structures increase, the distance errors of (Euclidean) ridges based on the Euclidean SCMS algorithm applied to the angular coordinates $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_{1000}\}$ increase. Conversely, the distance errors of directional ridges and the ridges based on the Euclidean SCMS algorithm in $\mathbb{R}^3$ decrease when the true circular structures lie at higher latitudes on Ω_2; see panel (b) of Figure A.3. While the performance of our directional SCMS algorithm and that of the Euclidean SCMS algorithm in $\mathbb{R}^3$ are almost indistinguishable in terms of the average geodesic distance errors, our directional SCMS algorithm significantly outperforms the Euclidean SCMS algorithm with regard to computational efficiency; see panels (c)–(d) of Figure A.3. Note that the Euclidean SCMS algorithm exhibits high variance in the number of iteration steps under the repeated experiments because each simulated dataset may contain outliers that are far from the true circular structure on Ω_2, and the Euclidean SCMS algorithm requires exceptionally many iterations to converge when initialized from these outliers. Our directional SCMS algorithm,

however, is more stable in its iteration count because it is adapted to the geometry of Ω_2.

Other potential issues of analyzing directional data with Euclidean methods and ignoring the curvature of Ω_2 can be found in Chrisman (2017). In summary, it is highly inadequate and inefficient to handle directional data with the Euclidean KDE and SCMS algorithm, which calls for the need to introduce the directional KDE (2.4) and propose our well-designed SCMS algorithm for analyzing directional data (Algorithm 2).

A.2 Normal Space of the Euclidean Density Ridge

We first introduce some notation for the rest of this section. Given a matrix-valued function $B : \mathbb{R}^D \to \mathbb{R}^{m\times n}$, its gradient $\nabla B(\boldsymbol{x})$ will be an $m \times n \times D$ array defined as $[B(\boldsymbol{x})]_{ijk} = \frac{\partial}{\partial x_k} B(\boldsymbol{x})_{ij}$. The derivative of B in the direction of a vector $\boldsymbol{z} \in \mathbb{R}^D$ is defined as:

$$B'(\boldsymbol{x};\boldsymbol{z}) \equiv \lim_{\epsilon\to 0} \frac{B(\boldsymbol{x}+\epsilon\boldsymbol{z}) - B(\boldsymbol{x})}{\epsilon} = \nabla B(\boldsymbol{x})\boldsymbol{z}.$$

When the matrix $A(\boldsymbol{x}) = \nabla\nabla f(\boldsymbol{x}) \in \mathbb{R}^{D\times D}$, we will use the notation $\nabla\nabla f'(\boldsymbol{x};\boldsymbol{z}) = \nabla^3 f(\boldsymbol{x})\boldsymbol{z} \equiv \boldsymbol{z}^T\nabla^3 f(\boldsymbol{x})$ interchangeably to denote its directional derivative along $\boldsymbol{z}$.

Recall that an order-d ridge of the density p in $\mathbb{R}^D$ is the collection of points defined as:

$$R_d = \left\{\boldsymbol{x} \in \mathbb{R}^D : G_d(\boldsymbol{x}) = \boldsymbol{0}, \lambda_{d+1}(\boldsymbol{x}) < 0\right\} = \left\{\boldsymbol{x} \in \mathbb{R}^D : V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x}) = \boldsymbol{0}, \lambda_{d+1}(\boldsymbol{x}) < 0\right\}.$$

Lemma A.1 below shows that under Assumptions 2.1–2.3, the Jacobian matrix $\nabla\left[V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x})\right]$ has rank $D-d$ at every point of R_d, and R_d is a d-dimensional manifold by the implicit function theorem (Rudin, 1976). Consequently, the row space of $\nabla\left[V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x})\right] \in \mathbb{R}^{(D-d)\times D}$ spans the normal space to R_d.

Define $M(\boldsymbol{x}) = \nabla\left[V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x})\right]^T = [\boldsymbol{m}_{d+1}(\boldsymbol{x}), ..., \boldsymbol{m}_D(\boldsymbol{x})] \in \mathbb{R}^{D\times(D-d)}$, the derivation in pages 60-63 of Eberly (1996) shows that

$$\boldsymbol{m}_k(\boldsymbol{x}) = \left[\lambda_k(\boldsymbol{x})\boldsymbol{I}_D + \sum_{i=1}^{d} \frac{\boldsymbol{v}_i(\boldsymbol{x})^T\nabla p(\boldsymbol{x})}{\lambda_k(\boldsymbol{x}) - \lambda_i(\boldsymbol{x})} \cdot \boldsymbol{v}_i(\boldsymbol{x})^T\nabla^3 p(\boldsymbol{x})\right]\boldsymbol{v}_k(\boldsymbol{x}) \tag{A.2}$$

for $k = d+1, ..., D$, and the column space of $M(\boldsymbol{x})$ spans the normal space to R_d. Let

$$\Lambda_0(\boldsymbol{x}) = \mathtt{Diag}\left[\lambda_{d+1}(\boldsymbol{x}), ..., \lambda_D(\boldsymbol{x})\right],$$

$$\Lambda_i(\boldsymbol{x}) = \texttt{Diag}\left[\frac{1}{\lambda_{d+1}(\boldsymbol{x}) - \lambda_i(\boldsymbol{x})}, ..., \frac{1}{\lambda_D(\boldsymbol{x}) - \lambda_i(\boldsymbol{x})}\right],$$
$$T_i(\boldsymbol{x}) = \left[\boldsymbol{v}_i(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right] \cdot \boldsymbol{v}_i(\boldsymbol{x})^T \nabla^3 p(\boldsymbol{x})$$

for $i = 1, ..., d$. Then,

$$M(\boldsymbol{x}) = V_d(\boldsymbol{x})\Lambda_0(\boldsymbol{x}) + \sum_{i=1}^{d} T_i(\boldsymbol{x}) V_d(\boldsymbol{x})\Lambda_i(\boldsymbol{x}). \tag{A.3}$$

The columns of $M(\boldsymbol{x})$ are not orthonormal. We therefore use the orthonormalization in Chen et al. (2015) to construct $N(\boldsymbol{x})$ whose columns are orthonormal and span the same column space as $M(\boldsymbol{x})$ in the following steps. Under the condition that $M(\boldsymbol{x}) = \nabla \left[V_d(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right]^T$ has full rank $D-d$ at every point $\boldsymbol{x} \in R_d$ (see Lemma A.1), $M(\boldsymbol{x})^T M(\boldsymbol{x})$ is positive definite, and we perform the Cholesky decomposition on it, *i.e.*,

$$M(\boldsymbol{x})^T M(\boldsymbol{x}) = J(\boldsymbol{x}) J(\boldsymbol{x})^T, \tag{A.4}$$

where $J(\boldsymbol{x}) \in \mathbb{R}^{(D-d)\times(D-d)}$ is a lower-triangular matrix whose diagonal elements are positive. We then define

$$N(\boldsymbol{x}) = M(\boldsymbol{x}) \left[J(\boldsymbol{x})^T\right]^{-1}. \tag{A.5}$$

Notice that $M(\boldsymbol{x}), N(\boldsymbol{x}), J(\boldsymbol{x})$ intrinsically depend on the dimension d of the ridge R_d, but we do not explicate these dependencies in their notations. As discussed in Chen et al. (2015), $M(\boldsymbol{x})$ might not be unique because the eigenvalues of $\nabla\nabla p(\boldsymbol{x})$ can have multiplicity greater than 1. Any collection of linearly independent unit eigenvectors of $\nabla\nabla p(\boldsymbol{x})$ fits into the above construction for $M(\boldsymbol{x})$. However, as will be shown later, this nonuniqueness of $M(\boldsymbol{x})$ does not affect the results below, as we only require the smoothness of $M(\boldsymbol{x})^T M(\boldsymbol{x})$ to develop a lower bound of $\texttt{reach}(R_d)$.

Lemma A.1. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. Given that $M(\boldsymbol{x})$ and $N(\boldsymbol{x})$ are defined in (A.3) and (A.5), we have the following properties:*

(a) $N(\boldsymbol{x})$ and $M(\boldsymbol{x})$ have the same column space. In addition,

$$N(\boldsymbol{x})N(\boldsymbol{x})^T = M(\boldsymbol{x}) \left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1} M(\boldsymbol{x})^T.$$

That is, $N(\boldsymbol{x})N(\boldsymbol{x})^T$ is the projection matrix onto the columns of $M(\boldsymbol{x})$.

(b) The columns of $N(\boldsymbol{x})$ are orthonormal to each other.

(c) For $\boldsymbol{x} \in R_d$, the column space of $N(\boldsymbol{x})$ is normal to the (tangent) direction of R_d at $\boldsymbol{x}$.

(d) For all $\boldsymbol{x} \in R_d$, $\texttt{rank}(N(\boldsymbol{x})) = \texttt{rank}(M(\boldsymbol{x})) = D - d$. Moreover, R_d is a d-dimensional manifold that contains neither intersections nor endpoints. Namely, R_d is a finite union of connected and compact manifolds.

(e) For $\boldsymbol{x} \in R_d$, all the $(D-d)$ nonzero singular values of $M(\boldsymbol{x})$ are greater than $\beta_1 > 0$ and therefore,

$$\left|\left|\left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1}\right|\right|_2 \leq \frac{1}{\beta_1^2} \quad \text{and} \quad \left|\left|\left[J(\boldsymbol{x})^T\right]^{-1}\right|\right|_2 \leq \frac{1}{\beta_1}.$$

(f) When $||\boldsymbol{x} - \boldsymbol{y}||_2$ is sufficiently small and $\boldsymbol{x}, \boldsymbol{y} \in R_d \oplus \rho$,

$$\left|\left|N(\boldsymbol{x})N(\boldsymbol{x})^T - N(\boldsymbol{y})N(\boldsymbol{y})^T\right|\right|_{\max} \leq A_0 \left(||p||_\infty^{(3)} + ||p||_\infty^{(4)}\right)^2 ||\boldsymbol{x} - \boldsymbol{y}||_2$$

for some constant $A_0 > 0$.

(g) Assume that another density function q also satisfies Assumptions 2.1–2.3 and $||p - q||_{\infty,3}^*$ is sufficiently small. Then

$$\left|\left|N_p(\boldsymbol{x})N_p(\boldsymbol{x})^T - N_q(\boldsymbol{x})N_q(\boldsymbol{x})^T\right|\right|_{\max} \leq A_1 \cdot ||p - q||_{\infty,3}^*$$

for some constant $A_1 > 0$ and any $\boldsymbol{x} \in R_d$, where $N_p(\boldsymbol{x})$ is the matrix defined in (A.5) with the underlying density p.

(h) The reach of R_d satisfies

$$\texttt{reach}(R_d) \geq \min\left\{\frac{\rho}{2}, \frac{\beta_1^2}{A_2\left(||p||_\infty^{(3)} + ||p||_\infty^{(4)}\right)}\right\}$$

for some constant $A_2 > 0$.

Lemma A.1 is extended from Lemma 2 in Chen et al. (2015) to handle the density ridge R_d with $1 \leq d < D$. Since Assumptions 2.1–2.3 imply the imposed conditions of Lemma 2 in Chen et al. (2015), our proof of Lemma A.1 essentially follows from their arguments with some minor modifications.

Proof of Lemma A.1. We adopt and generalize parts of the proof of Lemma 2 in Chen et al. (2015).

(a) This property is a natural corollary of the Cholesky decomposition as:

$$N(\boldsymbol{x})N(\boldsymbol{x})^T = M(\boldsymbol{x})\left[J(\boldsymbol{x})^T\right]^{-1} J(\boldsymbol{x})^{-1}M(\boldsymbol{x})^T = M(\boldsymbol{x})\left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1} M(\boldsymbol{x})^T.$$

(b) Some direct calculations show that

$$\begin{aligned} N(\boldsymbol{x})^T N(\boldsymbol{x}) &= J(\boldsymbol{x})^{-1}M(\boldsymbol{x})^T M(\boldsymbol{x})\left[J(\boldsymbol{x})^T\right]^{-1} \\ &= J(\boldsymbol{x})^{-1}M(\boldsymbol{x})^T M(\boldsymbol{x})\left[J(\boldsymbol{x})^T\right]^{-1} J(\boldsymbol{x})^{-1}J(\boldsymbol{x}) \\ &= J(\boldsymbol{x})^{-1}M(\boldsymbol{x})^T M(\boldsymbol{x})\left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1} J(\boldsymbol{x}) \\ &= \boldsymbol{I}_{D-d}. \end{aligned}$$

(c) It can be proved by the argument of Lemma 1 in Chen et al. (2015). Or, we define an arbitrary parametrized curve $\gamma : (-\epsilon, \epsilon) \to \mathbb{R}^D$ lying within R_d for some $\epsilon > 0$. Then $\gamma'(t)$ aligns with the tangent direction at $\gamma(t) \in R_d$. Since $V_d(\gamma(t))^T \nabla p(\gamma(t)) = \mathbf{0}$, taking the derivative with respect to t gives us that

$$0 = \frac{d}{dt}\left[V_d(\gamma(t))^T \nabla p(\gamma(t))\right] = \nabla\left[V_d(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right] \cdot \gamma'(t)$$

with $\boldsymbol{x} = \gamma(t)$. Hence, by the arbitrariness of $\gamma(t)$, the column of $M(\boldsymbol{x}) = \nabla\left[V_d(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right]^T$ is normal to the tangent direction of R_d at $\boldsymbol{x}$.

(d) We prove that the $(D-d)$ nonzero singular values of $M(\boldsymbol{x}) \in \mathbb{R}^{D\times(D-d)}$ are bounded away from 0. Recall that

$$M(\boldsymbol{x}) = V_d(\boldsymbol{x})\Lambda_0(\boldsymbol{x}) + \sum_{i=1}^{d} T_i(\boldsymbol{x})V_d(\boldsymbol{x})\Lambda_i(\boldsymbol{x})$$

with

$$\begin{aligned}
\Lambda_0(\boldsymbol{x}) &= \texttt{Diag}\left[\lambda_{d+1}(\boldsymbol{x}), ..., \lambda_D(\boldsymbol{x})\right], \\
\Lambda_i(\boldsymbol{x}) &= \texttt{Diag}\left[\frac{1}{\lambda_{d+1}(\boldsymbol{x}) - \lambda_i(\boldsymbol{x})}, ..., \frac{1}{\lambda_D(\boldsymbol{x}) - \lambda_i(\boldsymbol{x})}\right], \\
T_i(\boldsymbol{x}) &= \left[\boldsymbol{v}_i(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right] \cdot \boldsymbol{v}_i(\boldsymbol{x})^T \nabla^3 p(\boldsymbol{x})
\end{aligned}$$

for $i = 1, ..., d$. Under Assumptions 2.2 and 2.3,

$$\begin{aligned}
&\left|\left|\sum_{i=1}^{d} T_i(\boldsymbol{x}) V_d(\boldsymbol{x}) \Lambda_i(\boldsymbol{x})\right|\right|_2 \\
&\leq \sum_{i=1}^{d} ||T_i(\boldsymbol{x})||_2 \cdot ||V_d(\boldsymbol{x})||_2 \cdot \frac{1}{\beta_0} \quad \text{by Assumption 2.2} \\
&\leq \sum_{i=1}^{d} \left|\left|\left[\boldsymbol{v}_i(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right] \cdot \boldsymbol{v}_i(\boldsymbol{x})^T \nabla^3 p(\boldsymbol{x})\right|\right|_2 \cdot \frac{1}{\beta_0} \quad \text{since } ||V_d(\boldsymbol{x})||_2 = 1 \\
&\leq \frac{d\,||\nabla p(\boldsymbol{x})||_2 \cdot D^{\frac{3}{2}}\,||\nabla^3 p(\boldsymbol{x})||_{\max}}{\beta_0} \\
&\leq \beta_0 - \beta_1 \quad \text{by Assumption 2.3.}
\end{aligned}$$

It shows that all the singular values of $\sum_{i=1}^{d} T_i(\boldsymbol{x}) V_d(\boldsymbol{x}) \Lambda_i(\boldsymbol{x})$ are less than $\beta_0 - \beta_1$. Moreover, under Assumption 2.2 again, all the $(D-d)$ nonzero singular values of $V_d(\boldsymbol{x})\Lambda_0(\boldsymbol{x})$ are greater than β_0. By Theorem 3.3.16 in Horn and Johnson (1991), we know that all the $(D-d)$ nonzero singular values of $M(\boldsymbol{x})$ are greater than $\beta_0 - (\beta_0 - \beta_1) = \beta_1 > 0$. Therefore, $\texttt{rank}(M(\boldsymbol{x})) = \texttt{rank}(N(\boldsymbol{x})) = D - d$. The rest of the proof follows directly from Claim 4 in Chen et al. (2015).

(e) By the proof of (d), we already know that all the $(D-d)$ nonzero singular values of $M(\boldsymbol{x})$ are greater than $\beta_1 > 0$. Thus, $\left|\left|\left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1}\right|\right|_2 \leq \frac{1}{\beta_1^2}$, and

$$\begin{aligned}
\left|\left|\left[J(\boldsymbol{x})^T\right]^{-1}\right|\right|_2 = \sqrt{\frac{1}{\sigma_{\min}\left(J(\boldsymbol{x})J(\boldsymbol{x})^T\right)}} &= \sqrt{\frac{1}{\sigma_{\min}\left(M(\boldsymbol{x})^T M(\boldsymbol{x})\right)}} \\
&= \sqrt{\left|\left|\left[M(\boldsymbol{x})^T M(\boldsymbol{x})\right]^{-1}\right|\right|_2} \leq \frac{1}{\beta_1},
\end{aligned}$$

where $\sigma_{\min}(A)$ is the smallest singular value of matrix A.

Finally, the proofs of properties (f), (g), and (h) are essentially the same as the corresponding claims in Chen et al. (2015). We thus omit them. □

As we have discussed in Remark 2.2, property (d) of Lemma A.1 demonstrates that Assumptions 2.1–2.3 for the ridge R_d are sufficient to imply the critical full-rank condition of its normal space in Chen (2022) in order for R_d to be a well-defined solution manifold.

A.3 Normal Space of Directional Density Ridge

Recall that an order-d density ridge of a directional density f on $\Omega_q = \{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1\}$ is the set of points defined as:

$$\underline{R}_d = \left\{\boldsymbol{x} \in \Omega_q : \underline{G}_d(\boldsymbol{x}) = \boldsymbol{0}, \underline{\lambda}_{d+1}(\boldsymbol{x}) < 0\right\} = \left\{\boldsymbol{x} \in \Omega_q : \underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x}) = \boldsymbol{0}, \underline{\lambda}_{d+1}(\boldsymbol{x}) < 0\right\}. \tag{A.6}$$

Lemma A.2 below shows that under Assumptions 2.1–2.3, the Jacobian matrices $\nabla\left[\underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x})\right] \in \mathbb{R}^{(q-d)\times(q+1)}$ and $\left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla\left[\underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x})\right]^T \in \mathbb{R}^{(q+1)\times(q-d)}$ (*i.e.*, projecting the columns of $\nabla\left[\underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x})\right]^T$ onto the tangent space $T_{\boldsymbol{x}}$) both have rank $q-d$ at every point on $\underline{R}_d$, and $\underline{R}_d$ will be a d-dimensional submanifold on Ω_q by the implicit function theorem (Rudin, 1976; Lee, 2012). Analogously to the discussion of the normal space of a Euclidean density ridge in Appendix A.2, we define

$$\begin{aligned}\underline{M}(\boldsymbol{x}) = \nabla\left[\underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x})\right]^T &= \nabla\left[\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})\right]^T \\ &= \left(\underline{\boldsymbol{m}}_{d+1}(\boldsymbol{x}), ..., \underline{\boldsymbol{m}}_q(\boldsymbol{x})\right) \in \mathbb{R}^{(q+1)\times(q-d)}.\end{aligned}$$

Different from the Euclidean density ridge case, it is the column space of

$$\underline{M}_E(\boldsymbol{x}) = \left(\nabla\left[\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})\right]^T, \boldsymbol{x}\right) = \left(\underline{\boldsymbol{m}}_{d+1}(\boldsymbol{x}), ..., \underline{\boldsymbol{m}}_q(\boldsymbol{x}), \boldsymbol{x}\right) \in \mathbb{R}^{(q+1)\times(q+1-d)} \tag{A.7}$$

that spans the normal space of $\underline{R}_d$ within the ambient space $\mathbb{R}^{q+1}$. It can be seen from our

Remark 2.2 that the rows of

$$\nabla\Psi(\boldsymbol{x}) = \begin{pmatrix} \nabla\left[\underline{\boldsymbol{v}}_{d+1}(\boldsymbol{x})^T\nabla f(\boldsymbol{x})\right] \\ \vdots \\ \nabla\left[\underline{\boldsymbol{v}}_{q}(\boldsymbol{x})^T\nabla f(\boldsymbol{x})\right] \\ 2\boldsymbol{x} \end{pmatrix}$$

spans the normal space of the solution manifold $\underline{R}_d$; see also Lemma 1 in Chen (2022). Consequently, the column space of $\left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\underline{M}(\boldsymbol{x}) = \left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla\left[\underline{V}_d(\boldsymbol{x})^T\texttt{grad}\, f(\boldsymbol{x})\right]^T$ spans the normal space of $\underline{R}_d$ within the tangent space $T_{\boldsymbol{x}}$ at each $\boldsymbol{x} \in \underline{R}_d \subset \Omega_q$. The technique in pages 60-63 of Eberly (1996) is still valid to argue that

$$\begin{aligned} &\underline{\boldsymbol{m}}_k(\boldsymbol{x}) \\ &= \left[\underline{\lambda}_k(\boldsymbol{x})\boldsymbol{I}_{q+1} + \sum_{i=1}^{d}\frac{\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla f(\boldsymbol{x})}{\underline{\lambda}_k(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}\cdot\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla^3 f(\boldsymbol{x}) + \frac{\boldsymbol{x}^T\nabla f(\boldsymbol{x})}{\underline{\lambda}_k(\boldsymbol{x})}\cdot\boldsymbol{x}^T\nabla^3 f(\boldsymbol{x})\right]\underline{\boldsymbol{v}}_k(\boldsymbol{x}) \\ &= \left[\underline{\lambda}_k(\boldsymbol{x})\boldsymbol{I}_{q+1} + \sum_{i=1}^{d}\frac{\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla f(\boldsymbol{x})}{\underline{\lambda}_k(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}\cdot\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla^3 f(\boldsymbol{x})\right]\underline{\boldsymbol{v}}_k(\boldsymbol{x}) \end{aligned} \tag{A.8}$$

for $k = d+1, ..., q$, where we use the fact that $\boldsymbol{x}^T\nabla f(\boldsymbol{x}) = \boldsymbol{x}^T\texttt{grad}\, f(\boldsymbol{x}) = 0$ on Ω_q under the extension of f as in (A1). Let

$$\begin{aligned} \underline{\Lambda}_0(\boldsymbol{x}) &= \texttt{Diag}\left[\underline{\lambda}_{d+1}(\boldsymbol{x}), ..., \underline{\lambda}_q(\boldsymbol{x})\right], \\ \underline{\Lambda}_i(\boldsymbol{x}) &= \texttt{Diag}\left[\frac{1}{\underline{\lambda}_{d+1}(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}, ..., \frac{1}{\underline{\lambda}_q(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}\right], \\ \underline{T}_i(\boldsymbol{x}) &= \left[\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla f(\boldsymbol{x})\right]\cdot\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla^3 f(\boldsymbol{x}) \end{aligned}$$

for $i = 1, ..., d$. Then,

$$\underline{M}(\boldsymbol{x}) = \underline{V}_d(\boldsymbol{x})\underline{\Lambda}_0(\boldsymbol{x}) + \sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x}). \tag{A.9}$$

As in the Euclidean data case, the columns of $\underline{M}_E(\boldsymbol{x}) = [\underline{M}(\boldsymbol{x}), \boldsymbol{x}] \in \mathbb{R}^{(q+1)\times(q+1-d)}$ are not orthonormal, and we again leverage the orthonormalization technique in Chen et al.

(2015) to construct $\underline{N}(\boldsymbol{x})$ that shares the same column space with $\underline{M}_E(\boldsymbol{x})$ but has orthonormal columns. That is, under the condition that $\underline{M}_E(\boldsymbol{x})$ has full column rank $q+1-d$ at every point $\boldsymbol{x} \in \underline{R}_d$ (see Lemma A.2),

$$\underline{N}(\boldsymbol{x}) = \underline{M}_E(\boldsymbol{x}) \left[\underline{J}(\boldsymbol{x})^T\right]^{-1} \tag{A.10}$$

with the Cholesky decomposition $\underline{M}_E(\boldsymbol{x})^T \underline{M}_E(\boldsymbol{x}) = \underline{J}(\boldsymbol{x})\underline{J}(\boldsymbol{x})^T$, where $\underline{J}(\boldsymbol{x}) \in \mathbb{R}^{(q+1-d)\times(q+1-d)}$ is a lower-triangular matrix whose diagonal elements are positive. Finally, the non-uniqueness of $\underline{M}_E(\boldsymbol{x})$ will not affect our subsequent discussion of the properties of directional density ridges.

Lemma A.2. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. Given that $\underline{M}(\boldsymbol{x})$, $\underline{M}_E(\boldsymbol{x}) = [\underline{M}(\boldsymbol{x}), \boldsymbol{x}]$, and $\underline{N}(\boldsymbol{x})$ are defined in (A.9) and (A.10), we have the following properties:*

(a) $\underline{M}_E(\boldsymbol{x})$ and $\underline{N}(\boldsymbol{x})$ have the same column space. In addition,

$$\underline{N}(\boldsymbol{x})\underline{N}(\boldsymbol{x})^T = \underline{M}_E(\boldsymbol{x}) \left[\underline{M}_E(\boldsymbol{x})^T \underline{M}_E(\boldsymbol{x})\right]^{-1} \underline{M}_E(\boldsymbol{x})^T.$$

That is, $\underline{N}(\boldsymbol{x})\underline{N}(\boldsymbol{x})^T$ is the projection matrix onto the columns of $\underline{M}_E(\boldsymbol{x})$.

(b) The columns of $\underline{N}(\boldsymbol{x})$ are orthonormal to each other.

(c) For $\boldsymbol{x} \in \underline{R}_d$, the column space of $\underline{N}(\boldsymbol{x})$ is normal to the (tangent) direction of $\underline{R}_d$ at $\boldsymbol{x}$.

(d) For $\boldsymbol{x} \in \underline{R}_d$, the smallest eigenvalue $\lambda_{\min}(\underline{M}(\boldsymbol{x})^T \underline{M}(\boldsymbol{x})) = \lambda_{\min}(\underline{M}(\boldsymbol{x})^T(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\underline{M}(\boldsymbol{x})) \geq \underline{\beta}_1^2 > 0$, and

$$\texttt{rank}(\underline{M}(\boldsymbol{x})) = \texttt{rank}\left((\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\underline{M}(\boldsymbol{x})\right) = q - d.$$

Moreover, all the nonzero singular values of $\underline{M}_E(\boldsymbol{x})$ are greater than $\min\left\{\underline{\beta}_1, 1\right\} > 0$, and

$$\texttt{rank}(\underline{N}(\boldsymbol{x})) = \texttt{rank}(\underline{M}_E(\boldsymbol{x})) = q + 1 - d.$$

Therefore, $\underline{R}_d$ is a d-dimensional submanifold that contains neither intersections nor endpoints on Ω_q. Namely, $\underline{R}_d$ is a finite union of connected and compact submanifolds on Ω_q.

(e) For all $\boldsymbol{x} \in \underline{R}_d$,

$$\left|\left|\left[\underline{M}(\boldsymbol{x})^T \left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \underline{M}(\boldsymbol{x})\right]^{-1}\right|\right|_2 \leq \frac{1}{\underline{\beta}_1^2} \quad \textit{and} \quad \left|\left|\left[\underline{J}(\boldsymbol{x})^T\right]^{-1}\right|\right|_2 \leq \max\left\{\frac{1}{\underline{\beta}_1}, 1\right\}.$$

(f) When $||\boldsymbol{x} - \boldsymbol{y}||_2$ *is sufficiently small and* $\boldsymbol{x}, \boldsymbol{y} \in \left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$,

$$\left|\left|\underline{N}(\boldsymbol{x})\underline{N}(\boldsymbol{x})^T - \underline{N}(\boldsymbol{y})\underline{N}(\boldsymbol{y})^T\right|\right|_{\max} \leq \underline{A}_0 \left(||f||_\infty^{(3)} + ||f||_\infty^{(4)}\right)^2 ||\boldsymbol{x} - \boldsymbol{y}||_2$$

for some constant $\underline{A}_0 > 0$.

(g) Assume that another directional density function g *also satisfies Assumptions 2.1–2.3 after the extension* $g(\boldsymbol{x}) \equiv g\left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||}\right)$ *in* $\mathbb{R}^{q+1} \setminus \{\boldsymbol{0}\}$, *and* $||f - g||_{\infty,3}^*$ *is sufficiently small. Then,*

$$\left|\left|\underline{N}_f(\boldsymbol{x})\underline{N}_f(\boldsymbol{x})^T - \underline{N}_g(\boldsymbol{x})\underline{N}_g(\boldsymbol{x})^T\right|\right|_{\max} \leq \underline{A}_1 \cdot ||f - g||_{\infty,3}^*$$

for some constant $\underline{A}_1 > 0$ *and any* $\boldsymbol{x} \in \underline{R}_d$, *where* $\underline{N}_f(\boldsymbol{x})$ *is the matrix defined in* (A.10) *with directional density* f.

(h) The reach of $\underline{R}_d$ *satisfies*

$$\texttt{reach}(\underline{R}_d) \geq \min\left\{\underline{\rho}/2, \frac{\min\left\{\underline{\beta}_1, 1\right\}^2}{\underline{A}_2\left(||f||_\infty^{(3)} + ||f||_\infty^{(4)}\right)}\right\}$$

for some constant $\underline{A}_2 > 0$.

Lemma A.2 is the directional analogue of Lemma A.1 to the directional data scenario; thus, its proof is similar to the proof of Lemma A.1.

Proof of Lemma A.2. The proofs of properties (a), (b), and (c) can be inherited from the corresponding ones in Lemma A.1 with mild modifications and are therefore omitted.

(d) We will prove that the $(q-d)$ nonzero singular values of $\underline{M}(\boldsymbol{x})$ and $\left(\boldsymbol{I}_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\underline{M}(\boldsymbol{x}) \in \mathbb{R}^{(q+1)\times(q-d)}$ are bounded away from 0. Recall that

$$\underline{M}(\boldsymbol{x}) = \underline{V}_d(\boldsymbol{x})\underline{\Lambda}_0(\boldsymbol{x}) + \sum_{i=1}^{d} \underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})$$

with

$$\begin{aligned}
\underline{\Lambda}_0(\boldsymbol{x}) &= \texttt{Diag}\left[\underline{\lambda}_{d+1}(\boldsymbol{x}), ..., \underline{\lambda}_q(\boldsymbol{x})\right] \\
\underline{\Lambda}_i(\boldsymbol{x}) &= \texttt{Diag}\left[\frac{1}{\underline{\lambda}_{d+1}(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}, ..., \frac{1}{\underline{\lambda}_q(\boldsymbol{x}) - \underline{\lambda}_i(\boldsymbol{x})}\right] \\
\underline{T}_i(\boldsymbol{x}) &= \left[\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla f(\boldsymbol{x})\right] \cdot \underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla^3 f(\boldsymbol{x})
\end{aligned}$$

for $i = 1, ..., d$. Under Assumption 2.2,

$$\begin{aligned}
&\left|\left|\left(\boldsymbol{I}_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\cdot\sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})\right|\right|_2 \\
&\leq \left|\left|\sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})\right|\right|_2 \quad \text{since } \left|\left|\boldsymbol{I}_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right|\right|_2 = 1 \\
&\leq \sum_{i=1}^{d}||\underline{T}_i(\boldsymbol{x})||_2 \cdot ||\underline{V}_d(\boldsymbol{x})||_2 \cdot \frac{1}{\underline{\beta}_0} \quad \text{by Assumption 2.2} \\
&\leq \sum_{i=1}^{d}\left|\left|\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla f(\boldsymbol{x})\right|\right|_2 \cdot \left|\left|\underline{\boldsymbol{v}}_i(\boldsymbol{x})^T\nabla^3 f(\boldsymbol{x})\right|\right|_2 \cdot \frac{1}{\underline{\beta}_0} \quad \text{since } ||\underline{V}_d(\boldsymbol{x})||_2 = 1 \\
&\leq \frac{d\,||\nabla f(\boldsymbol{x})||_2 \cdot q^{\frac{3}{2}}\,||\nabla^3 f(\boldsymbol{x})||_{\max}}{\underline{\beta}_0} \\
&\leq \underline{\beta}_0 - \underline{\beta}_1.
\end{aligned}$$

It shows that all the singular values of $\left(\boldsymbol{I}_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})$ or simply $\sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})$ are less than $\underline{\beta}_0 - \underline{\beta}_1$. Moreover, under Assumption 2.2 again, all the $(q-d)$ singular values of

$$\left(\boldsymbol{I}_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_0(\boldsymbol{x}) = \underline{V}_d(\boldsymbol{x})\underline{\Lambda}_0(\boldsymbol{x})$$

are at least $\underline{\beta}_0$.

By Theorem 3.3.16 in Horn and Johnson (1991), we know that all the singular values of $\underline{M}(\boldsymbol{x})$ and $\left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\underline{M}(\boldsymbol{x})$ are greater than

$$\sigma_i\left(\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_0(\boldsymbol{x})\right) - \sigma_1\left(\sum_{i=1}^{d}\underline{T}_i(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})\underline{\Lambda}_i(\boldsymbol{x})\right) \geq \underline{\beta}_0 - (\underline{\beta}_0 - \underline{\beta}_1) = \underline{\beta}_1 > 0,$$

where $\sigma_i(A), i = 1, ..., q-d$ are singular values of a matrix $A \in \mathbb{R}^{(q+1)\times(q-d)}$ in their descending order. Therefore, the minimum eigenvalue of $\underline{M}(\boldsymbol{x})^T\underline{M}(\boldsymbol{x})$ satisfies

$$\lambda_{\min}\left(\underline{M}(\boldsymbol{x})^T\underline{M}(\boldsymbol{x})\right) = \lambda_{\min}\left(\underline{M}(\boldsymbol{x})^T(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\underline{M}(\boldsymbol{x})\right) \geq \underline{\beta}_1^2 > 0. \tag{A.11}$$

Now, given $\underline{M}_E(\boldsymbol{x}) = [\underline{M}(\boldsymbol{x}), \boldsymbol{x}] \in \mathbb{R}^{(q+1)\times(q+1-d)}$ and $\boldsymbol{x} \in \Omega_q$, we know that

$$\underline{M}_E(\boldsymbol{x})^T\underline{M}_E(\boldsymbol{x}) = \begin{pmatrix} \underline{M}(\boldsymbol{x})^T\underline{M}(\boldsymbol{x}) & \underline{M}(\boldsymbol{x})^T\boldsymbol{x} \\ \boldsymbol{x}^T\underline{M}(\boldsymbol{x}) & 1 \end{pmatrix}.$$

If we denote the orthonormal eigenvectors of $\underline{M}(\boldsymbol{x})^T\underline{M}(\boldsymbol{x})$ by $\boldsymbol{v}_{\underline{M},1}(\boldsymbol{x}), ..., \boldsymbol{v}_{\underline{M},q-d}(\boldsymbol{x}) \in \mathbb{R}^{q-d}$, then

$$\begin{pmatrix} \boldsymbol{v}_{\underline{M},1}(\boldsymbol{x}) \\ 0 \end{pmatrix}, ..., \begin{pmatrix} \boldsymbol{v}_{\underline{M},q-d}(\boldsymbol{x}) \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{0} \\ 1 \end{pmatrix} \in \mathbb{R}^{q+1-d}$$

are the orthonormal eigenvectors of $\underline{M}_E(\boldsymbol{x})^T\underline{M}_E(\boldsymbol{x})$, whose eigenvalues are thus lower bounded by $\min\left\{\underline{\beta}_1, 1\right\}$ due to (A.11). Hence, $\texttt{rank}(\underline{N}(\boldsymbol{x})) = \texttt{rank}(\underline{M}_E(\boldsymbol{x})) = q + 1 - d$.

By the implicit function theorem and the extra constraint $\underline{R}_d \subset \Omega_q$, $\underline{R}_d$ is a d-dimensional submanifold on Ω_q. It also implies that $\underline{R}_d$ cannot have intersections, because otherwise the intersected points will violate the rank condition.

Finally, we argue by contradiction that $\underline{R}_d$ has no endpoints. Assume, on the contrary, that $\underline{R}_d$ has an end point $\boldsymbol{x}_0$. Our preceding argument has shown that $\underline{M}(\boldsymbol{x})$, the derivative of $\underline{V}_d(\boldsymbol{x})^T\nabla f(\boldsymbol{x})$, is bounded. In addition, $\boldsymbol{x}_0 \in \underline{R}_d$. However, this contradicts the implicit function theorem indicating that $\underline{R}_d$ is a d-dimensional submanifold on Ω_q, because at the end point $\boldsymbol{x}_0 \in \underline{R}_d$, there exists no local coordinate chart for $\underline{R}_d$ defined on an open set in $\underline{R}_d$. The results follow.

(e) By the proof of (d), we already know that all the $(q-d)$ nonzero singular values of $\underline{M}(\boldsymbol{x})$ and $\left(\boldsymbol{I}_D - \boldsymbol{x}\boldsymbol{x}^T\right)\underline{M}(\boldsymbol{x})$ are greater than $\underline{\beta}_1 > 0$. Also, all the $(q+1-d)$ nonzero singular values of $\underline{M}_E(\boldsymbol{x})$ are greater than $\min\left\{\underline{\beta}_1, 1\right\}$. Thus, the results follow easily from the argument of (e) in Lemma A.1.

Finally, the proofs of properties (f), (g), and (h) are essentially the same as the corresponding claims in Chen et al. (2015). We thus omit them. For (h), the reader should be aware that we have extended the directional density f from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$. In addition, it is the columns of $\underline{M}_E(\boldsymbol{x})$ that span the normal space of $\underline{R}_d$ in the ambient space, whose nonzero singular values are lower bounded by $\min\left\{\underline{\beta}_1, 1\right\}$. The proof of (h) can also be found in Theorem 3 of Chen (2022). □

A.4 Stability of Directional Density Ridge

A.4.1 Subspace Constrained Gradient Flows

This subsection is modified from Section 4 in Genovese et al. (2014) for directional densities and their ridges on Ω_q. A map $\varpi : \mathbb{R} \to \Omega_q$ is a subspace constrained gradient flow with the principal Riemannian gradient $\underline{G}_d$ if $\varpi(0) = \boldsymbol{x} \in \Omega_q$ and

$$\varpi'(t) = \underline{G}_d(\varpi(t)) = \underline{U}_d(\varpi(t)) \cdot \texttt{grad}\, f(\varpi(t)) = \underline{V}_d(\varpi(t))\underline{V}_d(\varpi(t))^T \nabla f(\varpi(t)), \tag{A.12}$$

where the last equality follows from (2.12). Given the definition of the directional density ridge $\underline{R}_d$ in (A.6), it consists of the destinations of the subspace constrained gradient flow ϖ, *i.e.*, $\boldsymbol{y} \in \underline{R}_d$ if $\lim_{t\to\infty} \varpi(t) = \boldsymbol{y}$ for some ϖ satisfying (A.12). It will be convenient to parametrize the subspace constrained gradient ascent path with ϖ by arc length. Let $s \equiv s(t)$ be the arc length from $\varpi(t)$ to $\varpi(\infty)$:

$$s(t) = \int_t^\infty ||\varpi'(u)||_2 \, du.$$

Denote the inverse of $s(t)$ by $t \equiv t(s)$. Note that

$$t'(s) = \frac{1}{s'(t)} = -\frac{1}{||\varpi'(t(s))||_2} = -\frac{1}{||\underline{G}_d\left(\varpi(t(s))\right)||_2}.$$

With $\gamma(s) = \varpi(t(s))$, we have that

$$\gamma'(s) = -\frac{\underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2}, \tag{A.13}$$

which is a reparametrization of (A.12) by arc length. Note that γ always lies on Ω_q because its velocity is within the tangent space $T_{\gamma(s)}$ for every $s \in [0, \infty)$. Lemma 2 in Chen et al. (2015) justifies the uniqueness of γ passing through any particular point $\boldsymbol{x} \in \left((\underline{R}_d \oplus \underline{\rho}) \setminus \underline{R}_d\right) \cap \Omega_q$ under Assumptions 2.1–2.3. The (reversed) subspace constrained gradient flow γ can be lifted onto the directional function f, as we may define

$$\xi(s) = f(\varpi(\infty)) - f(\varpi(t(s))) = f(\gamma(0)) - f(\gamma(s)). \tag{A.14}$$

Sometimes, we may add the subscript $\boldsymbol{x}$ to the curves $\varpi_{\boldsymbol{x}}, \gamma_{\boldsymbol{x}}, \xi_{\boldsymbol{x}}$ if we want to emphasize that ϖ, γ, ξ start from or pass through the specific point $\boldsymbol{x}$.

To analyze the behavior of the subspace constrained gradient flow ξ lifted on f, we need the derivative of the projection matrix $\underline{U}(\boldsymbol{x}) \equiv \underline{U}_d(\boldsymbol{x})$ along the path γ. Recall that $\underline{U}(\boldsymbol{x}) \equiv \underline{U}_d(\mathcal{H}f(\boldsymbol{x})) = \underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T$. The collection $\{\underline{U}(\boldsymbol{x}) : \boldsymbol{x} \in \Omega_q\}$ defines a matrix field: there is a matrix $\underline{U}(\boldsymbol{x})$ attached to each point $\boldsymbol{x}$. As mentioned earlier, there is a unique path γ and a unique $s > 0$ such that $\boldsymbol{x} = \gamma(s)$ for any $\boldsymbol{x} \in \left(\underline{R}_d \oplus \underline{\rho}\right) \setminus \underline{R}_d$. Define

$$\dot{\underline{U}}_{\gamma(s)} \equiv \lim_{\epsilon \to 0} \frac{\underline{U}(\mathcal{H}f(\gamma(s+\epsilon))) - \underline{U}(\mathcal{H}f(\gamma(s)))}{\epsilon} = \lim_{t \to 0} \frac{\underline{U}(\mathcal{H}f(\gamma(s)) + tE_s) - \underline{U}(\mathcal{H}f(\gamma(s)))}{t}, \tag{A.15}$$

where $E_s = \frac{d}{ds}\mathcal{H}f(\gamma(s)) = \bar{\nabla}_{\gamma'(s)}\mathcal{H}f(\gamma(s))$ with $\bar{\nabla}$ being the Riemannian connection on Ω_q. Under Assumptions 2.1–2.3, ξ has a quadratic-like behavior near the directional ridge $\underline{R}_d$, analogous to Lemma 3 in Genovese et al. (2014).

Lemma A.3. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. For all $\boldsymbol{x} \in \left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$, we have the following properties:*

(a) $\xi(0) = 0$, $\xi'(s) = ||\underline{G}_d(\gamma(s))||_2$, and $\xi'(0) = 0$. Thus, $\xi(s)$ is non-decreasing in s.

(b) The second derivative of ξ satisfies $\xi''(s) \geq \frac{\beta_0}{2}$.

(c) $\xi(s) \geq \frac{\beta_0}{4} ||\gamma(0) - \gamma(s)||_2^2$.

Proof of Lemma A.3. The proof is adapted from Lemma 3 in Genovese et al. (2014).

(a) The first property $\xi(0) = 0$ is obvious from the definition (A.14). Then,

$$\begin{aligned}\xi'(s) = -\texttt{grad}\, f(\gamma(s))^T \gamma'(s) &= \frac{\nabla f(\gamma(s))^T \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2} \quad \text{by (A.13)} \\ &= \frac{\nabla f(\gamma(s))^T \underline{V}_d(\gamma(s)) \underline{V}_d(\gamma(s))^T \nabla f(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2} \\ &= ||\underline{G}_d(\gamma(s))||_2 \,,\end{aligned}$$

since $\underline{V}_d(\gamma(s))^T \underline{V}_d(\gamma(s)) = \boldsymbol{I}_{q-d}$ for all $\gamma(s) \in \Omega_q$. By the definition of $\underline{R}_d$ in (A.6), $\underline{G}_d(\gamma(s)) = 0$ when $\gamma(s) \in \underline{R}_d$. Thus, $\xi'(0) = 0$ and $\xi(s)$ is non-decreasing in s.

(b) Note that

$$\begin{aligned}\left(\xi'(s)\right)^2 = ||\underline{G}_d(\gamma(s))||_2^2 &= \nabla f(\gamma(s))^T \underline{V}_d(\gamma(s)) \underline{V}_d(\gamma(s))^T \nabla f(\gamma(s)) \\ &= \left[\texttt{grad}\, f(\gamma(s))\right]^T \underline{U}(\gamma(s)) \left[\texttt{grad}\, f(\gamma(s))\right].\end{aligned}$$

Differentiating both sides of the equation, we have that

$$\begin{aligned}2\xi'(s)\xi''(s) = 2\gamma'(s)^T \mathcal{H} f(\gamma(s)) \underline{U}(\gamma(s)) \left[\texttt{grad}\, f(\gamma(s))\right] \\ + \left[\texttt{grad}\, f(\gamma(s))\right]^T \dot{\underline{U}}(\gamma(s)) \left[\texttt{grad}\, f(\gamma(s))\right].\end{aligned}$$

Since $\underline{U}(\gamma(s)) \cdot \underline{U}(\gamma(s)) = \underline{U}(\gamma(s))$ (idempotent), we have that $\dot{\underline{U}}(\gamma(s)) = \underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) + \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s))$, and hence the second term on the right-hand side of the above equation becomes

$$\begin{aligned}&\left[\texttt{grad}\, f(\gamma(s))\right]^T \dot{\underline{U}}_{\gamma(s)}(\gamma(s)) \left[\texttt{grad}\, f(\gamma(s))\right] \\ &= \nabla f(\gamma(s))^T \underline{U}(\gamma(s)) \dot{\underline{U}}_{\gamma(s)}(\gamma(s)) \nabla f(\gamma(s)) + \nabla f(\gamma(s))^T \dot{\underline{U}}_{\gamma(s)}(\gamma(s)) \underline{U}(\gamma(s)) \nabla f(\gamma(s)) \\ &= 2\nabla f(\gamma(s))^T \dot{\underline{U}}(\gamma(s)) \underline{G}_d(\gamma(s))\end{aligned}$$

Thus,

$$2\xi'(s)\xi''(s) = 2\gamma'(s)^T \mathcal{H} f(\gamma(s)) \underline{G}_d(\gamma(s)) + 2\nabla f(\gamma(s))^T \dot{\underline{U}}(\gamma(s)) \underline{G}_d(\gamma(s)).$$

By (a) and (A.13), we conclude that

$$\xi''(s) = -\frac{\underline{G}_d(\gamma(s))^T \mathcal{H}f(\gamma(s)) \cdot \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2^2} + \frac{\nabla f(\gamma(s))^T \dot{\underline{U}}(\gamma(s)) \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2}. \tag{A.16}$$

Now, we will bound the two terms in (A.16), respectively. As for the first term $-\frac{\underline{G}_d(\gamma(s))^T \mathcal{H}f(\gamma(s)) \cdot \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2^2}$, we notice that $\underline{G}_d(\gamma(s))$ is in the column space of $\underline{V}_d(\gamma(s))$. Hence,

$$\underline{G}_d(\gamma(s))^T \mathcal{H}f(\gamma(s)) \cdot \underline{G}_d(\gamma(s)) = \underline{G}_d(\gamma(s))^T \left[\underline{V}_d(\gamma(s)) \Lambda_{\underline{R}_d}(\gamma(s)) \underline{V}_d(\gamma(s))^T\right] \underline{G}_d(\gamma(s)),$$

where $\Lambda_{\underline{R}_d}(\gamma(s)) = \texttt{Diag}\left[\underline{\lambda}_{d+1}(\gamma(s)), ..., \underline{\lambda}_q(\gamma(s))\right]$. Therefore, from Assumption 2.2,

$$\begin{aligned}\frac{\underline{G}_d(\gamma(s))^T \mathcal{H}f(\gamma(s)) \cdot \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2^2} &= \frac{\underline{G}_d(\gamma(s))^T \left[\underline{V}_d(\gamma(s)) \Lambda_{\underline{R}_d}(\gamma(s)) \underline{V}_d(\gamma(s))^T\right] \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2^2} \\ &\leq \lambda_{\max}\left[\underline{V}_d(\gamma(s)) \Lambda_{\underline{R}_d}(\gamma(s)) \underline{V}_d(\gamma(s))^T\right] \leq -\underline{\beta}_0\end{aligned}$$

and consequently,

$$-\frac{\underline{G}_d(\gamma(s))^T \mathcal{H}f(\gamma(s)) \cdot \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2^2} \geq \underline{\beta}_0.$$

As for the second term $\frac{\nabla f(\gamma(s))^T \dot{\underline{U}}_{\gamma(s)} \underline{G}_d(\gamma(s))}{||\underline{G}_d(\gamma(s))||_2}$, we notice that $\underline{U}(\gamma(s)) + \underline{U}^{\perp}(\gamma(s)) = \boldsymbol{I}_{q+1}$, where $\underline{U}^{\perp} \equiv \underline{U}_d^{\perp}$, and $\underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s)) = \underline{G}_d(\gamma(s))$. Then,

$$\begin{aligned}&\nabla f(\gamma(s))^T \dot{\underline{U}}(\gamma(s)) \underline{G}_d(\gamma(s)) \\ &= \nabla f(\gamma(s))^T \underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) \underline{G}_d(\gamma(s)) + \nabla f(\gamma(s))^T \underline{U}^{\perp}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) G_d(\gamma(s)) \\ &= \nabla f(\gamma(s))^T \underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s)) \\ &\quad + \nabla f(\gamma(s))^T \underline{U}^{\perp}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s)).\end{aligned}$$

However, $\left|\nabla f(\gamma(s))^T \underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s))\right| = 0$. To see this, note that $\underline{U}(\gamma(s)) \cdot \underline{U}(\gamma(s)) = \underline{U}(\gamma(s))$ and it implies that

$$\begin{aligned}&\underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) + \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) = \dot{\underline{U}}(\gamma(s)) \\ \Longrightarrow\ &\underline{U}(\gamma(s)) \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) + \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)) = \dot{\underline{U}}(\gamma(s)) \underline{U}(\gamma(s)),\end{aligned}$$

showing that $\underline{U}(\gamma(s))\dot{\underline{U}}_{\gamma(s)}\underline{U}(\gamma(s)) = \mathbf{0}$. To bound $\nabla f(\gamma(s))^T \underline{U^\perp}(\gamma(s))\dot{\underline{U}}(\gamma(s))\underline{U}(\gamma(s)) \cdot \underline{G_d}(\gamma(s))$, we proceed as follows. As before, we let $E_s = \frac{d}{ds}\mathcal{H}f(\gamma(s)) = \bar{\nabla}\mathcal{H}f(\gamma(s)) \cdot \gamma'(s)$. Then, by the Davis-Kahan theorem (Lemma B.9),

$$\begin{aligned}
&\left|\nabla f(\gamma(s))^T \underline{U}^\perp(\gamma(s))\dot{\underline{U}}(\gamma(s))\underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s))\right| \\
&= \lim_{t\to 0} \frac{\left|\nabla f(\gamma(s))^T \underline{U}^\perp(\gamma(s)) \left[\underline{U}(\mathcal{H}f(\gamma(s)) + tE_s) - \underline{U}(\mathcal{H}f(\gamma(s)))\right]\underline{U}(\gamma(s)) \cdot \underline{G}_d(\gamma(s))\right|}{t} \\
&\leq \left|\left|\underline{U}^\perp(\gamma(s))\nabla f(\gamma(s))\right|\right|_2 \cdot \lim_{t\to 0} \frac{||\underline{U}(\mathcal{H}f(\gamma(s)) + tE_s) - \underline{U}(\mathcal{H}f(\gamma(s)))||_2}{t} \cdot ||\underline{G}_d(\gamma(s))||_2 \\
&\leq \frac{\sqrt{2}\left|\left|\underline{U}^\perp(\gamma(s))\nabla f(\gamma(s))\right|\right|_2 \cdot ||E_s||_F \cdot ||\underline{G}_d(\gamma(s))||_2}{\underline{\beta}_0}.
\end{aligned}$$

Note that $||E_s||_F \leq \left|\left|\bar{\nabla}\mathcal{H}f(\gamma(s))\right|\right|_{\max} \cdot ||\gamma'(s)||_2 \leq q^{\frac{3}{2}} ||\nabla^3 f(\gamma(s))||_{\max}$, because $||\gamma'(s)||_2 = 1$. Thus, from Assumption 2.3,

$$\frac{\left|\nabla f(\gamma(s))^T \dot{\underline{U}}(\gamma(s))\underline{G}_d(\gamma(s))\right|}{||\underline{G}_d(\gamma(s))||_2} \leq \frac{\sqrt{2}q^{\frac{3}{2}} ||\nabla^3 f(\gamma(s))||_{\max} \left|\left|\underline{U}^\perp(\gamma(s))\nabla f(\gamma(s))\right|\right|_2}{\underline{\beta}_0} \leq \frac{\underline{\beta}_0}{2}.$$

Therefore, $\xi''(s) \geq \underline{\beta}_0 - \frac{\underline{\beta}_0}{2} = \frac{\underline{\beta}_0}{2}$.

(c) For some $0 \leq \widetilde{s} \leq s$,

$$\xi(s) = \xi(0) + s\xi'(0) + \frac{s^2}{2}\xi''(\widetilde{s}) = \frac{s^2}{2}\xi''(\widetilde{s}) \geq \frac{\underline{\beta}_0 s^2}{4}$$

by (a) and (b). As γ is parametrized by arc length, we conclude that

$$\xi(s) - \xi(0) \geq \frac{\underline{\beta}_0}{4}s^2 \geq \frac{\underline{\beta}_0}{4}||\gamma(0) - \gamma(s)||_2^2.$$

The result follows. □

The statement (c) in Lemma A.3 is known as the quadratic growth condition in the optimization literature (Anitescu, 2000; Drusvyatskiy and Lewis, 2018). Under Assumptions 2.1–2.3, such a quadratic growth of the subspace constrained gradient flow ξ lifted onto the directional density f enables us to quantify the stability of directional ridges under small perturbations of the directional density and develop the linear convergence of the (directional) SCGA algorithms on Ω_q.

A.4.2 Proof of Theorem 2.2

We now show that if two directional densities f and $\widetilde{f}$ are close, their corresponding ridges $\underline{R}_d$ and $\underline{\widetilde{R}}_d$ are also close. We will use, for instance, $\underline{\widetilde{G}}_d$ and $\underline{\widetilde{U}}_d$, to refer to the principal (Riemannian) gradient and projection matrix with its columns as the eigenvectors corresponding to the smallest $q-d$ eigenvalues of the (Riemannian) Hessian $\mathcal{H}\widetilde{f}$ with the tangent space of Ω_q defined by $\widetilde{f}$.

Proof of Theorem 2.2. Our arguments are modified from the proof of Theorem 4 in Genovese et al. (2014) as well as Proposition 4 and Theorem 5 in Chen (2022).

(a) We write the spectral decompositions of $\mathcal{H}f$ and $\mathcal{H}\widetilde{f}$ as

$$\mathcal{H}f = V\Lambda V^T \quad \text{and} \quad \mathcal{H}\widetilde{f} = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^T.$$

By Weyl's Theorem (Theorem 4.3.1 in Horn and Johnson (2012)), we know that

$$|\lambda_j - \widetilde{\lambda}_j| \leq \left|\left|\mathcal{H}f - \mathcal{H}\widetilde{f}\right|\right|_2 \leq q \left|\left|f - \widetilde{f}\right|\right|_{\infty,2}^*,$$

where we recall that there are at most q nonzero eigenvalues of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ on Ω_q. Thus, $\widetilde{f}$ satisfies Assumption 2.2. Moreover, since Assumption 2.3 depends only on the first- and third-order derivatives of f, it holds for $\widetilde{f}$ when $\left|\left|f - \widetilde{f}\right|\right|_{\infty,3}^*$ is small enough.

(b) We present two methods based on two different flows to prove this statement and comment on their pros and cons in Remark A.1.

Method A: By the Davis-Kahan theorem (Lemma B.9 and (B.25)),

$$\begin{aligned}\left|\left|\underline{U}_d(\boldsymbol{x}) - \underline{\widetilde{U}}_d(\boldsymbol{x})\right|\right|_2 &= \left|\left|\underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T - \underline{\widetilde{V}}_d(\boldsymbol{x})\underline{\widetilde{V}}_d(\boldsymbol{x})^T\right|\right|_2 \\ &\leq \frac{\sqrt{2}\left|\left|\mathcal{H}f(\boldsymbol{x}) - \mathcal{H}\widetilde{f}(\boldsymbol{x})\right|\right|_F}{\underline{\beta}_0}\end{aligned}$$

$$\leq \frac{\sqrt{2}q \left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}}{\underline{\beta}_0}$$

for any $\boldsymbol{x} \in \Omega_q$. Then, given that $\left|\left|\widetilde{\underline{V}}_d(\boldsymbol{x})\widetilde{\underline{V}}_d(\boldsymbol{x})^T\right|\right|_2 = 1$,

$$\begin{aligned}
&\left|\left|\underline{G}_d(\boldsymbol{x}) - \widetilde{\underline{G}}_d(\boldsymbol{x})\right|\right|_2 \\
&= \left|\left|\underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T\nabla f(\boldsymbol{x}) - \widetilde{\underline{V}}_d(\boldsymbol{x})\widetilde{\underline{V}}_d(\boldsymbol{x})^T\nabla \widetilde{f}(\boldsymbol{x})\right|\right|_2 \\
&\leq \left|\left|\left[\underline{V}_d(\boldsymbol{x})\underline{V}_d(\boldsymbol{x})^T - \widetilde{\underline{V}}_d(\boldsymbol{x})\widetilde{\underline{V}}_d(\boldsymbol{x})^T\right]\nabla f(\boldsymbol{x})\right|\right|_2 + \left|\left|\widetilde{\underline{V}}_d(\boldsymbol{x})\widetilde{\underline{V}}_d(\boldsymbol{x})^T\left[\nabla f(\boldsymbol{x}) - \nabla \widetilde{f}(\boldsymbol{x})\right]\right|\right|_2 \\
&\leq \frac{\sqrt{2}q \left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}}{\underline{\beta}_0} \cdot ||\nabla f(\boldsymbol{x})||_2 + \sqrt{q}\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,1}.
\end{aligned}$$

Therefore, by the differentiability of f from Assumption 2.1 and the compactness of Ω_q, we obtain from the above calculations that

$$\sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\underline{G}_d(\boldsymbol{x}) - \widetilde{\underline{G}}_d(\boldsymbol{x})\right|\right|_2 \leq C_1\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}$$

for some constant $C_1 > 0$ that only depends on dimension q.

Now, let $\widetilde{\boldsymbol{x}} \in \widetilde{\underline{R}}_d$. Then, $\left|\left|\widetilde{\underline{G}}_d(\widetilde{\boldsymbol{x}})\right|\right|_2 = 0$, and $||\underline{G}_d(\widetilde{\boldsymbol{x}})||_2 \leq C_1\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}$. Let γ be the subspace constrained gradient ascent flow through $\widetilde{\boldsymbol{x}}$ as defined in Section A.4.1 so that $\gamma(s) = \widetilde{\boldsymbol{x}}$ for some s. Note that $\gamma(0) \in \underline{R}_d$. From property (a) of Lemma A.3, we have that $\xi'(s) = ||\underline{G}_d(\widetilde{\boldsymbol{x}})||_2$. Moreover, by Taylor's theorem,

$$C_1\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2} \geq ||\underline{G}_d(\widetilde{\boldsymbol{x}})||_2 = \xi'(s) = \xi'(0) + s\xi''(u)$$

for some u between 0 and s. Since $\xi'(0) = 0$, from property (b) of Lemma A.3,

$$C_1\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2} \geq s\xi''(u) \geq \frac{s\underline{\beta}_0}{2},$$

and consequently, $d_g(\gamma(0), \widetilde{\boldsymbol{x}}) \leq s \leq \frac{2C_1}{\underline{\beta}_0}\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}$, where $d_g(\boldsymbol{x}, \boldsymbol{y})$ denotes the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ on Ω_q. Therefore,

$$d_E(\widetilde{\boldsymbol{x}}, \underline{R}_d) \leq ||\gamma(0) - \widetilde{\boldsymbol{x}}||_2 \leq d_g(\gamma(0), \widetilde{\boldsymbol{x}}) \leq \frac{2C_1}{\underline{\beta}_0}\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}.$$

Now let $\boldsymbol{x} \in \underline{R}_d$. The same argument shows that $d_E(\boldsymbol{x}, \widetilde{\underline{R}}_d) \leq \frac{2C_1}{\underline{\beta}_0} \left|\left| f - \widetilde{f} \right|\right|^*_{\infty,2}$ for some constant $C_2 > 0$ because Assumptions 2.1–2.3 hold for $\widetilde{f}$.
As a result, $\texttt{Haus}(\underline{R}_d, \widetilde{\underline{R}}_d) \leq \frac{2C_1}{\underline{\beta}_0} \left|\left| f - \widetilde{f} \right|\right|^*_{\infty,2} = O\left(\left|\left| f - \widetilde{f} \right|\right|^*_{\infty,2} \right)$.

Method B: Since we are only required to bound the maximum Euclidean distance between $\underline{R}_d$ and $\widetilde{\underline{R}}_d$, *i.e.*, $\texttt{Haus}(\underline{R}_d, \widetilde{\underline{R}}_d)$, we may view $\underline{R}_d$ and $\widetilde{\underline{R}}_d$ as solution manifolds in $\mathbb{R}^{q+1}$ and tentatively ignore the manifold constraint $\underline{R}_d, \widetilde{\underline{R}}_d \subset \Omega_q$. Define $h(\boldsymbol{x}) = \left|\left| \underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x}) \right|\right|_2 = \sqrt{\nabla f(\boldsymbol{x})^T \underline{V}_d(\boldsymbol{x}) \underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})}$. Given that $\underline{M}(\boldsymbol{x}) = \nabla \left[\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x}) \right]^T$, the gradient of $h(\boldsymbol{x})$,

$$\nabla h(\boldsymbol{x}) = \frac{\nabla \left[\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x}) \right]^T \underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})}{||\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})||_2} = \frac{\underline{M}(\boldsymbol{x}) \underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})}{||\underline{V}_d(\boldsymbol{x})^T \nabla f(\boldsymbol{x})||_2}, \tag{A.17}$$

is a vector in $\mathbb{R}^{q+1}$. Let $\boldsymbol{z} \in \widetilde{\underline{R}}_d$. We define a flow $\phi_{\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}^{q+1}$ such that

$$\phi_{\boldsymbol{z}}(0) = \boldsymbol{z}, \quad \phi'_{\boldsymbol{z}}(t) = -\nabla h(\phi_{\boldsymbol{z}}(t)).$$

It can be argued by Theorem 7 in Chen (2022) that $\phi_{\boldsymbol{z}}(\infty) \in \underline{R}_d$ when $\boldsymbol{z} \in \underline{R}_d \oplus \delta_0$ for some small $\delta_0 > 0$. In addition, we can always choose $\left|\left| f - \widetilde{f} \right|\right|^*_{\infty,3}$ to be small enough so that $\widetilde{\underline{R}}_d \subset \underline{R}_d \oplus \delta_0$. By Theorem 3.39 in Irwin (2001), $\phi_{\boldsymbol{z}}(t)$ is uniquely defined because the gradient $\nabla h(\boldsymbol{z})$ is well-defined for all $\boldsymbol{z} \notin \underline{R}_d$. We can also reparametrize $\phi_{\boldsymbol{z}}(t)$ by arc length as:

$$\gamma_{\boldsymbol{z}}(0) = \boldsymbol{z}, \quad \gamma'_{\boldsymbol{z}}(s) = -\frac{\nabla h(\gamma_{\boldsymbol{z}}(s))}{||\nabla h(\gamma_{\boldsymbol{z}}(s))||_2}.$$

Let $\mathcal{S}_{\boldsymbol{z}} = \inf \{ s > 0 : \gamma_{\boldsymbol{z}}(s) \in \underline{R}_d \}$ be the terminal time/arc-length point and $\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}) \in \underline{R}_d$ be the destination of $\gamma_{\boldsymbol{z}}$ on $\underline{R}_d$. The above argument also demonstrates that the flows $\phi_{\boldsymbol{z}}$ or $\gamma_{\boldsymbol{z}}$ converge to the manifold $\underline{R}_d$ from the normal direction of $\underline{R}_d$, because we can write

$$\gamma'_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}) = -\frac{\nabla h(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))}{||\nabla h(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))||_2} = \sum_{k=d+1}^{q} \boldsymbol{a}_k \cdot \underline{\boldsymbol{m}}_k(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))$$

$$\text{with} \quad \boldsymbol{a}_k = -\frac{\boldsymbol{e}_{k-d}^T \underline{V}_d(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) \nabla f(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))}{||\underline{M}(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) \underline{V}_d(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) \nabla f(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))||_2}$$

and the column space of $\underline{M}(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}))$ spans the normal space of $\underline{R}_d$ at $\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}) \in \underline{R}_d$.

The goal now is to bound $\mathcal{S}_{\boldsymbol{z}}$ because its length must be greater than or equal to

$\left|\left|\boldsymbol{z} - \pi_{\underline{R}_d}(\boldsymbol{z})\right|\right|_2$. We then define $\vartheta_{\boldsymbol{z}}(s) = h(\gamma_{\boldsymbol{z}}(s)) - h(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) = h(\gamma_{\boldsymbol{z}}(s))$. Differentiating $\vartheta_{\boldsymbol{z}}(s)$ with respect to s leads to

$$\begin{aligned}
\vartheta'_{\boldsymbol{z}}(s) = \frac{d}{ds} h(\gamma_{\boldsymbol{z}}(s)) &= \left[\nabla h(\gamma_{\boldsymbol{z}}(s))\right]^T \gamma'_{\boldsymbol{z}}(s) \\
&= -\left|\left|\nabla h(\gamma_{\boldsymbol{z}}(s))\right|\right|_2 \\
&= -\frac{\left|\left|\underline{M}(\gamma_{\boldsymbol{z}}(s))\underline{V}_d(\gamma_{\boldsymbol{z}}(s))^T \nabla f(\gamma_{\boldsymbol{z}}(s))\right|\right|_2}{\left|\left|\underline{V}_d(\gamma_{\boldsymbol{z}}(s))^T \nabla f(\gamma_{\boldsymbol{z}}(s))\right|\right|_2} \\
&\leq -\sigma_{\min}\left(\underline{M}(\gamma_{\boldsymbol{z}}(s))\right) \leq -\widetilde{\beta}_1
\end{aligned} \tag{A.18}$$

by (d) in Lemma A.2. (Note that $0 < \widetilde{\beta}_1 \leq \underline{\beta}_1$ because the smallest singular value of $\underline{M}(\boldsymbol{x})$ is at least $\underline{\beta}_1$ and by the continuity of $\lambda_{\min}(\underline{M}(\boldsymbol{y})^T \underline{M}(\boldsymbol{y}))$, we can always choose $\delta_0 > 0$ such that $\sigma_{\min}(\underline{M}(\boldsymbol{y})) \geq \widetilde{\beta}_1$ for all $\boldsymbol{y} \in \underline{R}_d \oplus \delta_0$.) As $\boldsymbol{z} \in \widetilde{\underline{R}}_d$, by the proof of **Method A**, we know that

$$\begin{aligned}
C_1 \left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2} \geq \left|\left|\underline{G}_d(\boldsymbol{z})\right|\right| &= \left|\left|\underline{V}_d(\boldsymbol{z})^T \nabla f(\boldsymbol{z})\right|\right|_2 \\
&= h(\gamma_{\boldsymbol{z}}(0)) - h(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) \quad \text{since } h(\gamma_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}})) = 0 \\
&= \vartheta_{\boldsymbol{z}}(0) - \vartheta_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}) \quad \text{since } \vartheta_{\boldsymbol{z}}(\mathcal{S}_{\boldsymbol{z}}) = 0 \text{ and } \vartheta_{\boldsymbol{z}}(0) = h(\gamma_{\boldsymbol{z}}(0)) \\
&= -\mathcal{S}_{\boldsymbol{z}} \vartheta'_{\boldsymbol{z}}(\mathcal{S}^*_{\boldsymbol{z}}) \quad \text{by the mean value theorem} \\
&\geq \mathcal{S}_{\boldsymbol{z}} \widetilde{\beta}_1 \quad \text{by (A.18)},
\end{aligned}$$

where $\mathcal{S}^*_{\boldsymbol{z}}$ is some value between 0 and $\mathcal{S}_{\boldsymbol{z}}$. Hence, $\mathcal{S}_{\boldsymbol{z}} \leq \frac{C_1}{\widetilde{\beta}_1} \left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2} = O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}\right)$, which is independent of $\boldsymbol{z} \in \widetilde{\underline{R}}_d$. This implies that

$$\sup_{\boldsymbol{z} \in \widetilde{\underline{R}}_d} d_E(\boldsymbol{z}, \underline{R}_d) \leq \frac{C_1}{\widetilde{\beta}_1} \left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2} = O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}\right).$$

We can exchange the role of $\underline{R}_d$ and $\widetilde{\underline{R}}_d$ and apply the same argument to show that

$$\sup_{\boldsymbol{x} \in \underline{R}_d} d_E(\boldsymbol{x}, \widetilde{\underline{R}}_d) = O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}\right).$$

In total, this leads to the conclusion that $\texttt{Haus}(\underline{R}_d, \widetilde{\underline{R}}_d) = O\left(\left|\left|f - \widetilde{f}\right|\right|^*_{\infty,2}\right)$.

(c) By (h) in Lemma A.2, the reach of $\underline{R}_d$ has a lower bound, $\min\left\{\underline{\rho}/2, \frac{\min\{\underline{\beta}_1,1\}^2}{\underline{A}_2\left(||f||_\infty^{(3)}+||f||_\infty^{(4)}\right)}\right\}$. Note that $\underline{\rho}$ and $\underline{\beta}_1$ depend on the derivatives up to third order of f. Thus, the lower bound for the reach of $\widetilde{\underline{R}}_d$ will be identical to the one for $\underline{R}_d$ with an error rate $O\left(\left|\left|f-\widetilde{f}\right|\right|_{\infty,3}^*\right)$. □

Note that for the stability of directional ridges, one can relax Assumption 2.1 by requiring f to be β-Hölder with $\beta \geq 3$.

Remark A.1. *We apply two different methods to establish the stability theorem of directional density ridges.* **Method A** *utilizes the subspace constrained gradient flow constructed in Section A.4.1 and its quadratic behavior (Lemma A.3), while* **Method B** *defines a normal flow to the ridge $\underline{R}_d$ induced by the column space of $\underline{M}(\boldsymbol{x})$. Each of these two flows has its pros and cons. The subspace constrained gradient flow aligns more coherently with our directional SCMS algorithm (Algorithm 2) to identify the (estimated) directional ridge from data, because it relies only on the first and second-order derivatives of the (estimated) density f. Nevertheless, the subspace constrained gradient flow does not necessarily converge to $\underline{R}_d$ in the optimal direction, that is, the normal direction to $\underline{R}_d$. This can be seen from the explicit formula (A.9) of $\underline{M}(\boldsymbol{x})$, which spans the normal space of $\underline{R}_d$. The normal flow*

$$\phi_{\boldsymbol{z}}(0) = \boldsymbol{z}, \quad \phi'_{\boldsymbol{z}}(t) = -\frac{\underline{M}(\phi_{\boldsymbol{z}}(t))\underline{V}_d(\phi_{\boldsymbol{z}}(t))^T\nabla f(\phi_{\boldsymbol{z}}(t))}{||\underline{V}_d(\phi_{\boldsymbol{z}}(t))^T\nabla f(\phi_{\boldsymbol{z}}(t))||_2}$$

defined in **Method B**, *however, converges to $\underline{R}_d$ in its normal direction by construction. In general, the normal flow tends to the ridge $\underline{R}_d$ faster than the subspace constrained gradient flow, but it may be complicated to compute in any practical ridge-finding task due to its involvement with third-order derivatives of the (estimated) density f. Recently, Qiao and Polonik (2021) presented explicit formulae for finding density ridges via such a normal flow and its discrete gradient descent approximation. Additionally, they defined a smoothed version of the ridgeness function that also circumvents the computation of third-order derivatives of f.*

Appendix B

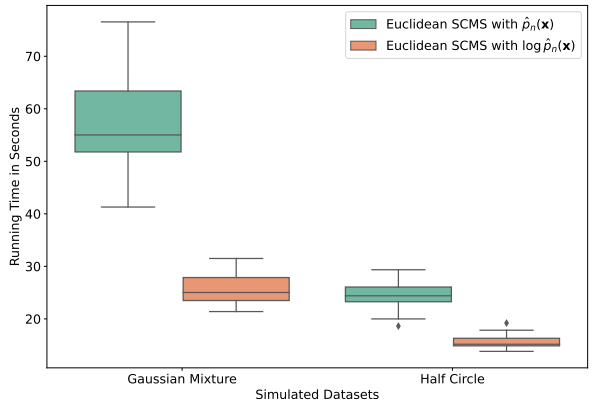
APPENDIX TO CHAPTER 3

B.1 Algorithmic Summaries of Euclidean and Directional SCMS Algorithms

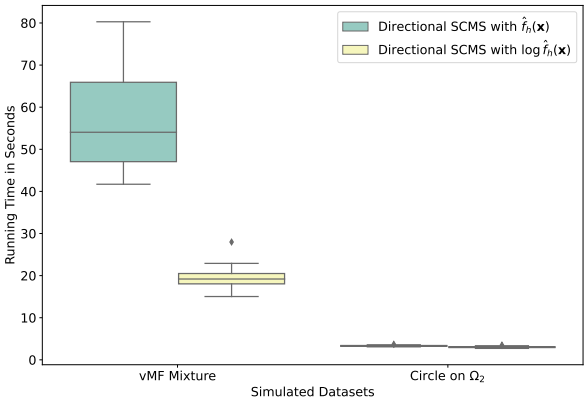
In this section, we provide algorithmic summaries of the Euclidean and directional SCMS algorithms for practical reference. Algorithm 1 describes each step of the Euclidean SCMS algorithm in detail. In our actual implementation of the algorithm, we replace the density estimator $\widehat{p}_n(\boldsymbol{x})$ with $\log \widehat{p}_n(\boldsymbol{x})$. To demonstrate that the (directional) SCMS algorithms under the log-density implementation give rise to a faster convergence process, we repeat our experiments in Sections 3.7.1 and 3.7.2 (*i.e.*, Figure 3.3 and Figure 3.4) 20 times for each simulated dataset with the (directional) SCMS algorithms under the original (estimated) density and the (estimated) log-density, respectively. The comparisons between their running times are shown in Figure B.1, in which the (directional) SCMS algorithms under the log-density implementation clearly outperform their counterparts with the original density in terms of the average elapsed time until convergence.

Additionally, when the observational data in practice are noisy, it is common to incorporate an extra denoising step before Step 2 of Algorithm 1 to remove observations in low-density areas and stabilize the (Euclidean) SCMS algorithm; see Genovese et al. (2014); Chen et al. (2015) for comparative studies that demonstrate the significance of denoising.

We summarize the directional SCMS algorithm in Algorithm 2. Note that in Step 2-1 of Algorithm 2, we compute the scaled versions $\frac{nh^2}{c_{L,q}(h)} \cdot \widehat{\underline{G}}_d(\boldsymbol{x})$ and $\frac{nh^2}{c_{L,q}(h)} \mathcal{H}\widehat{f}_h(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega_q$ because the estimated principal Riemannian gradient $\widehat{\underline{G}}_d(\boldsymbol{x})$ and Hessian $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ are often very small. The scaling stabilizes the numerical computation. The spectral decomposition is thus performed on the scaled Hessian estimator $\frac{nh^2}{c_{L,q}(h)} \mathcal{H}\widehat{f}_h(\boldsymbol{x})$, and the scaled principal



(a) Repeated experiments with the simulated Euclidean datasets in Figure 3.3.



(b) Repeated experiments with the simulated directional datasets in Figure 3.4.

Figure B.1: Running time comparisons between the (directional) SCMS algorithms with the original density and the log-density applied to our simulated datasets in Figure 3.3 and Figure 3.4.

Riemannian gradient estimator is calculated as

$$\begin{aligned}\frac{nh^2}{c_{L,q}(h)} \cdot \widehat{G}_d(\boldsymbol{x}) &= \widehat{\underline{V}}_d(\boldsymbol{x})\widehat{\underline{V}}_d(\boldsymbol{x})^T \left[\sum_{i=1}^{n} \left(\boldsymbol{x} \cdot \boldsymbol{x}^T\boldsymbol{X}_i - \boldsymbol{X}_i\right) L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)\right] \\ &= -\sum_{i=1}^{n} \widehat{\underline{V}}_d(\boldsymbol{x})\widehat{\underline{V}}_d(\boldsymbol{x})^T\boldsymbol{X}_i \cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right),\end{aligned}$$

where $\widehat{\underline{V}}_d(\boldsymbol{x}) = [\widehat{\underline{\boldsymbol{v}}}_{d+1}(\boldsymbol{x}), ..., \widehat{\underline{\boldsymbol{v}}}_q(\boldsymbol{x})]$ has its columns equal to the orthonormal eigenvectors associated with the d smallest eigenvalues of the scaled Hessian estimator $\frac{nh^2}{c_{L,q}(h)}\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ (or equivalently, $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$) inside the tangent space $T_{\boldsymbol{x}}$.

B.2 Proof of the DMS Q-function Ascent Result

This section verifies the ascent property of the auxiliary Q-function used to interpret the DMS iteration as a generalized EM step.

Proof of Theorem 3.2. As we take $\kappa_i = \frac{1}{h^2}$ and $\alpha_i = \frac{1}{n}$ for all $i = 1, ..., n$ in (3.13) and (3.12),

Algorithm 1 (Euclidean) Subspace Constrained Mean Shift (SCMS) Algorithm

Input:

- A data sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim p(\boldsymbol{x})$ in $\mathbb{R}^D$.
- The order d of the ridge, smoothing bandwidth $h > 0$, and tolerance level $\epsilon > 0$.
- A suitable mesh $\mathbb{M}_E \subset \mathbb{R}^D$ of initial points. By default, $\mathbb{M}_E = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$.

Step 1: Compute the density estimator $\widehat{p}_n(\boldsymbol{x}) = \frac{c_{k,D}}{nh^D} \sum_{i=1}^{n} k\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ on the mesh $\mathbb{M}_E$.

Step 2: For each initial point $\widehat{\boldsymbol{x}}^{(0)} \in \mathbb{M}_E$, iterate the following SCMS update until convergence:

while $\left|\left|\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \nabla \widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)})\right|\right|_2 > \epsilon$ **do**

Step 2-1: Compute the estimated Hessian matrix as:

$$\nabla\nabla\widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)}) = \frac{c_{k,D}}{nh^{D+2}} \sum_{i=1}^{n} \left[2\boldsymbol{I}_D \cdot k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right) + \frac{4}{h^2}(\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i)(\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i)^T \cdot k''\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right) \right].$$

Step 2-2: Perform the spectral decomposition on the Hessian $\nabla\nabla\widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)})$ and obtain that $\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)}) = \left[\widehat{\boldsymbol{v}}_{d+1}(\widehat{\boldsymbol{x}}^{(t)}), ..., \widehat{\boldsymbol{v}}_D(\widehat{\boldsymbol{x}}^{(t)})\right]$ whose columns are orthonormal eigenvectors associated with the smallest $(D-d)$ eigenvalues of $\nabla\nabla\widehat{p}_n(\widehat{\boldsymbol{x}}^{(t)})$.

Step 2-3: Update $\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \widehat{\boldsymbol{x}}^{(t)} + \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \left[\frac{\sum_{i=1}^{n} \boldsymbol{X}_i k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n} k'\left(\left|\left|\frac{\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \widehat{\boldsymbol{x}}^{(t)} \right]$.

end while

Output: An estimated d-ridge $\widehat{R}_d$ represented by the collection of resulting points.

the Q-function reduces to

$$Q(\boldsymbol{\mu}|\underline{\widehat{\boldsymbol{x}}}^{(t)}) = \sum_{i=1}^{n} \left[\log \frac{C_{q,L}(1/h^2)}{n} + \log L\left(\frac{1 - \boldsymbol{X}_i^T\boldsymbol{\mu}}{h^2}\right) \right] \cdot P(Z_1 = i|\boldsymbol{Y}_1, \underline{\widehat{\boldsymbol{x}}}^{(t)}), \tag{B.1}$$

Algorithm 2 Directional Subspace Constrained Mean Shift (SCMS) Algorithm

Input:

- A directional data sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim f(\boldsymbol{x})$ on Ω_q.
- The order d of the directional ridge, smoothing bandwidth $h > 0$, and tolerance level $\epsilon > 0$.
- A suitable mesh $\mathbb{M}_D \subset \Omega_q$ of initial points. By default, $\mathbb{M}_D = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$.

Step 1: Compute the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ on the mesh $\mathbb{M}_D$.

Step 2: For each $\widehat{\underline{\boldsymbol{x}}}^{(0)} \in \mathbb{M}_D$, iterate the following directional SCMS update until convergence:

while $\left|\left|\frac{nh^2}{c_{L,q}(h)} \cdot \widehat{\underline{G}}_d\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)\right|\right|_2 > \epsilon$ **do**:

Step 2-1: Compute the scaled version of the estimated Hessian matrix as:

$$\frac{nh^2}{c_{L,q}(h)}\mathcal{H}\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) = \left[\boldsymbol{I}_{q+1} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)^T\right]\left[\frac{1}{h^2}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \cdot L''\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right) + \sum_{i=1}^{n}\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}\boldsymbol{I}_{q+1} \cdot L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right]\left[\boldsymbol{I}_{q+1} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)^T\right].$$

Step 2-2: Perform the spectral decomposition on $\frac{nh^2}{c_{L,q}(h)}\mathcal{H}\widehat{f}_h\left(\widehat{\underline{\boldsymbol{x}}}^{(t)}\right)$ and compute $\underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)}) = \left[\underline{\widehat{\boldsymbol{v}}}_{d+1}(\widehat{\underline{\boldsymbol{x}}}^{(t)}), ..., \underline{\widehat{\boldsymbol{v}}}_q(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right]$, whose columns are orthonormal eigenvectors corresponding to the smallest $q-d$ eigenvalues inside the tangent space $T_{\widehat{\underline{\boldsymbol{x}}}^{(t)}}$.

Step 2-3: Update $\widehat{\underline{\boldsymbol{x}}}^{(t+1)} \leftarrow \widehat{\underline{\boldsymbol{x}}}^{(t)} - \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\left[\frac{\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\left|\left|\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right|\right|_2}\right]$.

Step 2-4: Standardize $\widehat{\underline{\boldsymbol{x}}}^{(t+1)}$ as $\widehat{\underline{\boldsymbol{x}}}^{(t+1)} \leftarrow \frac{\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{\left|\left|\widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right|\right|_2}$.

end while

Output: An estimated directional d-ridge $\underline{\widehat{R}}_d$ represented by the collection of resulting points.

and the posterior $Z_1|\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}$ becomes

$$P(Z_1 = i|\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}) = \frac{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\sum_{\ell=1}^n L\left(\frac{1-\boldsymbol{X}_\ell^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}. \tag{B.2}$$

The differentiability and convexity of L indicates that

$$L(y) - L(x) \geq L'(x) \cdot (y - x), \tag{B.3}$$

where $L'(x)$ is replaced by any subgradient at non-differentiable points. As $0 < L(0) < \infty$, we have that

$$\begin{aligned}
&Q(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}|\widehat{\underline{\boldsymbol{x}}}^{(t)}) - Q(\widehat{\underline{\boldsymbol{x}}}^{(t)}|\widehat{\underline{\boldsymbol{x}}}^{(t)}) \\
&= \sum_{i=1}^n \left[\log L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{h^2}\right) - \log L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right] \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\sum_{\ell=1}^n L\left(\frac{1-\boldsymbol{X}_\ell^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)} \\
&\overset{\text{(i)}}{\geq} \sum_{i=1}^n \frac{1}{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{h^2}\right)} \cdot \left[L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{h^2}\right) - L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right] \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\sum_{\ell=1}^n L\left(\frac{1-\boldsymbol{X}_\ell^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)} \\
&\overset{\text{(ii)}}{\geq} \sum_{i=1}^n \frac{L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{h^2}\right)} \cdot \left[\frac{\boldsymbol{X}_i^T\left(\widehat{\underline{\boldsymbol{x}}}^{(t)} - \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right)}{h^2}\right] \cdot \frac{L\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)}{\sum_{\ell=1}^n L\left(\frac{1-\boldsymbol{X}_\ell^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)} \\
&\overset{\text{(iii)}}{\geq} \frac{L\left(\frac{2}{h^2}\right)}{nL(0)^2} \sum_{i=1}^n L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right) \boldsymbol{X}_i^T\left(\widehat{\underline{\boldsymbol{x}}}^{(t)} - \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right) \\
&\overset{\text{(iv)}}{=} \frac{L\left(\frac{2}{h^2}\right)}{nL(0)^2} \left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right|\right| \cdot \left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right)^T\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \\
&\overset{\text{(v)}}{=} \frac{L\left(\frac{2}{h^2}\right)}{2nL(0)^2} \left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right)\right|\right| \cdot \left|\left|\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right|\right|_2 \\
&\geq 0,
\end{aligned}$$

where (i) uses the fact that $\log y - \log x \geq \frac{y-x}{y}$ by the concavity of $\log(\cdot)$, (ii) leverages the inequality (B.3), (iii) applies the monotonicity of L, (iv) plugs in (3.16), and (v) utilizes the fact that $\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right)^T\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) = \frac{1}{2}\left|\left|\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right|\right|_2^2$ on Ω_q. The result follows. □

B.2.1 Proof of the Monotonicity of the DMS Objective

The proof of Theorem 3.3 relies on the following proposition, which equates the observed log-likelihood values in the EM framework with the density estimates defined by the DMS sequence.

Proposition B.1. *The observed log-likelihood* $\log P(\boldsymbol{Y}_1|\boldsymbol{\mu})$ *is given by*

$$\log P(\boldsymbol{Y}_1|\boldsymbol{\mu}) = \log\left[\sum_{\ell=1}^{n} \alpha_\ell \cdot C_{\kappa_\ell,q+1,L} \cdot L\left[\kappa_\ell\left(1-\boldsymbol{X}_\ell^T\boldsymbol{\mu}\right)\right]\right].$$

In particular, when $\alpha_\ell = \frac{1}{n}$ *and* $\kappa_\ell = \frac{1}{h^2}$ *for all* $\ell = 1, ..., n$ *(i.e., simplifying to the DMS case), it implies that* $\log P(\boldsymbol{Y}_1|\boldsymbol{\mu}) = \log \widehat{f}_h(\boldsymbol{\mu})$.

Proof of Proposition B.1. By (3.10), the observed log-likelihood in the EM framework of Section 3.3 with $\boldsymbol{Y}_1 = (0, ..., 0, 1)^T \in \Omega_q$ is given by

$$\begin{aligned}\log P(\boldsymbol{Y}_1|\boldsymbol{\mu}) &= \log P(\boldsymbol{Y}_1, Z_1|\boldsymbol{\mu}) - \log P(Z_1|\boldsymbol{Y}_1, \boldsymbol{\mu}) \\ &= \sum_{i=1}^{n} \mathbb{1}_{\{Z_1=i\}} \cdot \left\{\log\alpha_i + \log C_{q,L}(\kappa_i) + \log L\left[\kappa_i(1-\boldsymbol{X}_i^T\boldsymbol{\mu})\right]\right\} \\ &\quad - \sum_{i=1}^{n} \mathbb{1}_{\{Z_1=i\}} \cdot \log P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}).\end{aligned} \tag{B.4}$$

Taking the expectation over $Z_1|\left(\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right)$ on both sides of (B.4) yields that

$$\begin{aligned}\log P(\boldsymbol{Y}_1|\boldsymbol{\mu}) &= \sum_{i=1}^{n} P\left(Z_1 = i|\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \left\{\log\alpha_i + \log C_{q,L}(\kappa_i) + \log L\left[\kappa_i(1-\boldsymbol{X}_i^T\boldsymbol{\mu})\right]\right\} \\ &\quad - \sum_{i=1}^{n} P\left(Z_1 = i|\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \log P(Z_1 = i|\boldsymbol{Y}_1, \boldsymbol{\mu}).\end{aligned} \tag{B.5}$$

After plugging (3.12) into (B.5), we obtain that

$$\log P(\boldsymbol{Y}_1|\boldsymbol{\mu}) = \sum_{i=1}^{n} P\left(Z_1 = i|\boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \left[\log\alpha_i + \log C_{q,L}(\kappa_i) + \log L\left[\kappa_i(1-\boldsymbol{X}_i^T\boldsymbol{\mu})\right]\right.$$

$$
\begin{aligned}
&\quad - \log \frac{\alpha_i \cdot C_{q,L}(\kappa_i) \cdot L\left[\kappa_i(1 - \boldsymbol{X}_i^T \boldsymbol{\mu})\right]}{\sum_{\ell=1}^n \alpha_\ell \cdot C_{q,L}(\kappa_\ell) \cdot L\left[\kappa_\ell\left(1 - \boldsymbol{X}_\ell^T \boldsymbol{\mu}\right)\right]}\Bigg] \\
&= \sum_{i=1}^n P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \log \left[\sum_{\ell=1}^n \alpha_\ell \cdot C_{q,L}(\kappa_\ell) \cdot L\left[\kappa_\ell\left(1 - \boldsymbol{X}_\ell^T \boldsymbol{\mu}\right)\right]\right] \\
&= \log \left[\sum_{\ell=1}^n \alpha_\ell \cdot C_{q,L}(\kappa_\ell) \cdot L\left[\kappa_\ell\left(1 - \boldsymbol{X}_\ell^T \boldsymbol{\mu}\right)\right]\right].
\end{aligned}
$$

When $\alpha_\ell = \frac{1}{n}$ and $\kappa_\ell = \frac{1}{h^2}$ for all $\ell = 1, ..., n$ (*i.e.*, simplifying to the DMS form), the above equation indicates that

$$
\log P(\boldsymbol{Y}_1 | \boldsymbol{\mu}) = \log \left[\frac{C_{q,L}(1/h^2)}{n} \sum_{i=1}^n L\left(\frac{1 - \boldsymbol{X}_i^T \boldsymbol{\mu}}{h^2}\right)\right] + \log \frac{C_{q,L}(1/h^2)}{c_{L,q}(h)} = \log \widehat{f}_h(\boldsymbol{\mu}), \quad \text{(B.6)}
$$

where we recall from (2.5) that $c_{L,q}(h) = C_{q,L}(1/h^2)$ is the same normalizing constant for the directional kernel L. The results follow. □

Proof of Theorem 3.3. The difference of observed log-likelihoods (B.5) between two consecutive steps of our DMS (or generalized EM) iteration is calculated as

$$
\begin{aligned}
&\log P\left(\boldsymbol{Y}_1 | \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right) - \log P\left(\boldsymbol{Y}_1 | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \\
&\overset{\text{(i)}}{=} Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) - \sum_{i=1}^n P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \log P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right) \\
&\quad - Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) + \sum_{i=1}^n P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \log P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \\
&= Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) - Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) - \sum_{i=1}^n P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \cdot \log \frac{P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right)}{P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t)}\right)} \\
&\overset{\text{(ii)}}{\geq} Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) - Q\left(\widehat{\underline{\boldsymbol{x}}}^{(t)} | \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \underbrace{- \log \left[\sum_{i=1}^n P\left(Z_1 = i | \boldsymbol{Y}_1, \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right)\right]}_{=0} \\
&\overset{\text{(iii)}}{\geq} 0,
\end{aligned}
\quad \text{(B.7)}
$$

where (i) uses (3.13), (ii) applies Jensen's inequality, and (iii) utilizes Theorem 3.2. The inequality (ii) is strict except when $\underline{\widehat{\boldsymbol{x}}}^{(t+1)} = \underline{\widehat{\boldsymbol{x}}}^{(t)}$. Therefore, combining (B.6) with (B.7) show that

$$\log \widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t+1)}) - \log \widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)}) = \log P\left(\boldsymbol{Y}_1|\underline{\widehat{\boldsymbol{x}}}^{(t+1)}\right) - \log P\left(\boldsymbol{Y}_1|\underline{\widehat{\boldsymbol{x}}}^{(t)}\right) \geq 0.$$

As $\widehat{f}_h$ is differentiable on the compact set Ω_q under our assumptions on L, $\left\{\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$ is thus bounded from above and converges. □

B.3 Convergence Theory for Generalized EM Sequences and Proof of Theorem 3.4

This section collects the generalized EM convergence arguments used in the proof of Theorem 3.4. The goal is to connect monotonicity of the auxiliary function with convergence of the DMS iterates.

To establish the global convergence of (generalized) EM sequences toward stationary points (in most cases, local maxima), we adopt the regularity conditions formulated by Wu (1983) regarding the parameter space Ψ and the observed log-likelihood function $\ell(\phi)$:

Assumption EM1. *Ψ is a subset of a finite-dimensional Euclidean space.*

Assumption EM2. *For any $\ell(\phi_0) > -\infty$, the level set $\Psi_0 = \{\phi \in \Psi : \ell(\phi) \geq \ell(\phi_0)\}$ is compact, where ϕ_0 is the initial parameter for the (generalized) EM algorithm.*

Assumption EM3. *ℓ is continuous on Ψ and differentiable in the interior of Ψ.*

Let $\phi \to M(\phi)$ be the point-to-set map defined by a generalized EM iteration such that the Q-function[1] satisfies

$$Q(\phi'|\phi) \geq Q(\phi|\phi) \quad \text{for any } \phi' \in \Psi.$$

[1]The Q-function is defined as the expectation of the complete log-likelihood with respect to the posterior distribution of the latent variables given the observed data and current parameter values.

Denote by $\mathcal{S}$ the set of stationary points of the observed log-likelihood ℓ. Under Assumptions EM1–EM3, the limit points of a generalized EM sequence are guaranteed to be stationary points of the observed log-likelihood.

Lemma B.2 (Theorem 1 of Wu 1983)**.** *Let $\{\phi_p\}$ be a generalized EM sequence generated by $\phi_{p+1} \in M(\phi_p)$, and suppose that (i) M is a closed point-to-set map over the complement of $\mathcal{S}$; (ii) $\ell(\phi_{p+1}) > \ell(\phi_p)$ whenever $\phi_p \notin \mathcal{S}$. Then, every limit point of $\{\phi_p\}$ is a stationary point of ℓ, and $\ell(\phi_p)$ converges monotonically to $\ell^* = \ell(\phi^*)$ for some $\phi^* \in \mathcal{S}$.*

In practice, generalized EM algorithms are typically terminated when the (Euclidean) distance between two consecutive iteration points $||\phi_{p+1} - \phi_p||$ falls below a prescribed tolerance. However, this criterion does not strictly guarantee the convergence of the generalized EM sequence to a unique limit point in $\mathcal{S}$; see the discussion in Wu (1983) and a counterexample provided in Li et al. (2007). To ensure the convergence of a generalized EM sequence to a single stationary point in $\mathcal{S}$ (in most cases, a local maximum), we require an additional structural condition on $\ell(\phi)$:

Assumption EM4. *The set of local modes of the observed log-likelihood $\ell(\phi)$ is discrete, and each local mode is isolated from other critical points.*

We now verify that these conditions hold in the DMS setting.

Proof of Theorem 3.4. We verify in detail that Assumptions EM1–EM4 for the convergence of generalized EM sequences can be satisfied by our conditions (a,b) on directional KDE $\widehat{f}_h$ and kernel function L.

First, the estimated parameter $\boldsymbol{\mu}$ lies on Ω_q, which is a compact manifold in the ambient Euclidean space $\mathbb{R}^{q+1}$. Thus, Assumption EM1 holds.

Second, by Proposition B.1, the observed log-likelihood coincides with the logarithm of the directional KDE $\log \widehat{f}_h$. For any initialization $\boldsymbol{\mu}^{(0)} \in \Omega_q$ with $\log \widehat{f}_h(\boldsymbol{\mu}^{(0)}) > -\infty$, the constrained parameter space $\Psi_0 = \left\{\boldsymbol{\mu} \in \Omega_q : \log \widehat{f}_h(\boldsymbol{\mu}) \geq \log \widehat{f}_h(\boldsymbol{\mu}^{(0)})\right\}$ is a closed subset of the compact set Ω_q, and is therefore compact. Hence, Assumption EM2 is satisfied.

Third, under condition (b), $\widehat{f}_h$ is differentiable on Ω_q. Therefore, Assumption EM3 holds.

Finally, it is obvious that our condition (a) implies Assumption EM4. By Lemma B.2 and the discreteness of the set of local modes, the DMS sequence converges to a single stationary point. Since saddle points and local minima are unstable for the ascent iteration, the convergence is to a local mode for all initializations outside a set of Lebesgue measure zero. □

B.3.1 Proof of Theorem 3.5

B.3.2 Function Classes on Riemannian Manifolds

The key definitions in this subsection are modified from Section 2 in Zhang and Sra (2016).

Definition B.3 (Geodesic Concavity)**.** *A function $f : M \to \mathbb{R}$ is said to be* geodesically concave *(or g-concave) if for any $p, q \in M$, a geodesic γ such that $\gamma(0) = p$ and $\gamma(1) = q$, and $t \in [0, 1]$, it holds that*

$$f(\gamma(t)) \geq (1 - t)f(p) + tf(q).$$

Equivalently, it can be shown that there exists a tangent vector $g_p \in T_p(M)$ such that

$$f(q) \leq f(p) + \langle g_p, \texttt{Exp}_p^{-1}(q)\rangle_p, \tag{B.8}$$

where g_p is called a subgradient *of f at p, or the* gradient *if f is differentiable, and $\langle \cdot, \cdot \rangle_p$ denotes the inner product in the tangent space of p induced by the Riemannian metric.*

See, for instance, Section 1.2 in Bubeck (2015) for the definition of subgradients of convex functions.

Definition B.4 (Geodesically Strong Concavity)**.** *A function $f : M \to \mathbb{R}$ is said to be geodesically μ-strongly concave if for any $p, q \in M$,*

$$f(q) \leq f(p) + \langle g_p, \texttt{Exp}_p^{-1}(q)\rangle_p - \frac{\mu}{2} \cdot d^2(p, q), \tag{B.9}$$

where $d(p, q) = \sqrt{\langle \texttt{Exp}_p^{-1}(q), \texttt{Exp}_p^{-1}(q)\rangle_p} = \left|\left|\texttt{Exp}_p^{-1}(q)\right|\right|$.

Definition B.5 (Lipschitzness). *A function $f : M \to \mathbb{R}$ is said to be geodesically L_f-Lipschitz if for any $p, q \in M$,*

$$|f(p) - f(q)| \leq L_f \cdot d(p, q). \tag{B.10}$$

Definition B.6 (β-Smoothness). *A differentiable function $f : M \to \mathbb{R}$ is said to be geodesically β-smooth if its gradient is β-Lipschitz. That is, for any $p, q \in M$,*

$$||g_p - \Gamma_q^p(g_q)|| \leq \beta \cdot d(p, q), \tag{B.11}$$

where Γ_q^p is the parallel transport from q to p and $\beta > 0$ is a constant.

Before proving Theorem 3.5, we introduce the following two useful results. As pointed out in Zhang and Sra (2016), a main hurdle in analyzing non-asymptotic convergence of first-order methods on smooth manifolds is that the Euclidean law of cosines does not hold. Fortunately, there is a trigonometric distance bound stated below for Alexandrov space (Burago et al., 1992) with curvature bounded below.

Lemma B.7 (Lemma 5 in Zhang and Sra 2016; see also Bonnabel 2013). *If a, b, c are the sides (that is, side lengths) of a geodesic triangle in an Alexandrov space with sectional curvature lower bounded by κ, and A is the angle between sides b and c, then*

$$a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)} b^2 + c^2 - 2bc\cos(A). \tag{B.12}$$

The sketch proof of Lemma B.7 can be found in Lemma 5 of Zhang and Sra (2016). Note that $\kappa = 1$ on Ω_q. We inherit the notation in Zhang and Sra (2016) and denote $\frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)}$ by $\zeta(\kappa, c)$ for the curvature-dependent quantity from inequality (B.12). One can show by differentiating $\zeta(\kappa, c)$ with respect to c that $\zeta(\kappa, c)$ is strictly increasing and greater than 1 for any $c > 0$ and fixed $\kappa \neq 0$. With Lemma B.7 in hand, we are able to state a straightforward corollary indicating an important relation between two consecutive updates of a gradient ascent algorithm on Ω_q.

Corollary B.8. *For any point $\boldsymbol{x}, \boldsymbol{y}^{(t)}$ in a convex set on Ω_q, the update in (3.18) satisfies*

$$2\eta\langle \texttt{grad}\, f(\boldsymbol{y}^{(t)}), \texttt{Exp}^{-1}_{\boldsymbol{y}^{(t)}}(\boldsymbol{x})\rangle \leq d^2(\boldsymbol{y}^{(t)}, \boldsymbol{x}) - d^2(\boldsymbol{y}^{(t+1)}, \boldsymbol{x}) + \zeta(1, d(\boldsymbol{y}^{(t)}, \boldsymbol{x})) \cdot \eta^2 ||\texttt{grad}\, f(\boldsymbol{y}^{(t)})||_2^2,$$

recalling that $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\langle \texttt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y}), \texttt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y})\rangle} = ||\texttt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y})||_2$ on Ω_q.

The proof is similar to Corollary 8 in Zhang and Sra (2016) and thus omitted.

Proof of Theorem 3.5. (a) **Linear convergence of gradient ascent with** f: The proof of the linear convergence of the population-level gradient ascent algorithm (3.18) is similar to standard results in optimization theory, except that we now work in the manifold setting. Recall from (3.18) that the iterative formula reads $\boldsymbol{y}^{(t+1)} = \texttt{Exp}_{\boldsymbol{y}^{(t)}}\left(\eta \cdot \texttt{grad}\, f(\boldsymbol{y}^{(t)})\right)$ for $t = 0, 1, \ldots$. We begin by deriving the following three facts.

• *Fact 1*: Given Assumption M1, f is geodesically strongly concave in a small neighborhood of each point in $\underline{\mathcal{M}}$. In particular, when $0 < r_0 \leq \sqrt{2 - 2\cos\left[\frac{3\lambda_*}{2(q+1)^{\frac{3}{2}}||f||^*_{\infty,3}}\right]}$,

$$f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) - \langle \texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}^{-1}_{\underline{\boldsymbol{m}}_k}(\boldsymbol{y})\rangle \leq -\frac{\lambda_*}{4}\left|\left|\texttt{Exp}^{-1}_{\underline{\boldsymbol{m}}_k}(\boldsymbol{y})\right|\right|_2^2 \tag{B.13}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and any $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$.

• *Fact 2.* Given Assumptions D1 and M1, we know that $||\texttt{grad}\, f(\boldsymbol{x})||_2 \equiv ||\texttt{Tang}\,(\nabla f(\boldsymbol{x}))||_2 > 0$ and

$$f(\underline{\boldsymbol{m}}_k) - f\left(\texttt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)\,||f||^*_{\infty,3}}\texttt{grad}\, f(\boldsymbol{x})\right)\right) \geq 0$$

for all $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\} \setminus \{\underline{\boldsymbol{m}}_k\}$ and any $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$.

• *Fact 3.* Given Assumption D1, the directional density f is $(q+1)\,||f||^*_{\infty,3}$-smooth.

As for *Fact 1*, it follows from the differentiability of f guaranteed by Assumption D1 and the eigenvalue condition in Assumption M1. By Taylor's expansion on manifolds (Pennec,

2006) and Assumption M1,

$$\begin{aligned}
&f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) \\
&= \langle \texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \rangle + \frac{1}{2} \cdot \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y})^T \left[\mathcal{H} f(\underline{\boldsymbol{m}}_k) \right] \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) + o\left(\left|\left| \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \right|\right|_2^2 \right) \\
&\leq \langle \texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \rangle - \frac{\lambda_*}{2} \left|\left| \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \right|\right|_2^2 + \frac{(q+1)^{\frac{3}{2}} ||f||_{\infty,3}^*}{6} \cdot \left|\left| \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \right|\right|_2^3
\end{aligned} \tag{B.14}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$. Since $||\boldsymbol{y} - \underline{\boldsymbol{m}}_k||_2 \leq r_0$, the geodesic distance between $\boldsymbol{y}$ and $\underline{\boldsymbol{m}}_k$ satisfies $d_g(\boldsymbol{y}, \underline{\boldsymbol{m}}_k) = \left|\left| \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \right|\right|_2 = \arccos(\boldsymbol{y}^T \underline{\boldsymbol{m}}_k) \leq \frac{3\lambda_*}{2(q+1)^{\frac{3}{2}} ||f||_{\infty,3}^*}$. Plugging this result back into (B.14) yields that

$$f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) \leq \langle \texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \rangle - \frac{\lambda_*}{4} \left|\left| \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y}) \right|\right|_2^2.$$

For our purpose, it suffices to only prove (B.13) as above. One can shrink the upper bound of the convergence radius $r_0 > 0$ so that the geodesically strong concavity is valid for any pair of points within $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$. Indeed, the local strong concavity of f is a natural consequence of Morse Lemma (Lemma 3.11 in Banyaga and Hurtubise (2004)) given Assumption M1.

Fact 2 is an obvious result under the eigenvalue condition in Assumption M1 and the differentiability condition in Assumption D1. This is because $\underline{\boldsymbol{m}}_k$ is the unique local mode of f within the neighborhood $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and the geodesic distance between $\boldsymbol{x}$ and a one-step gradient ascent iteration from $\boldsymbol{x}$ with the step size $\frac{1}{(q+1)||f||_{\infty,3}^*}$ satisfies

$$\begin{aligned}
d\left(\texttt{Exp}_{\boldsymbol{x}} \left(\frac{1}{(q+1)\,||f||_{\infty,3}^*} \texttt{grad}\, f(\boldsymbol{x}) \right), \boldsymbol{x} \right) &= \frac{1}{(q+1)\,||f||_{\infty,3}^*} ||\texttt{grad}\, f(\boldsymbol{x})||_2 \\
&= \frac{1}{(q+1)\,||f||_{\infty,3}^*} \left|\left| \texttt{grad}\, f(\boldsymbol{x}) - \underbrace{\Gamma_{\underline{\boldsymbol{m}}_k}^{\boldsymbol{x}} \left(\texttt{grad}\, f(\underline{\boldsymbol{m}}_k) \right)}_{=0} \right|\right|_2 \\
&\leq \frac{1}{(q+1)\,||f||_{\infty,3}^*} ||\mathcal{H} f(\boldsymbol{x})||_2 \cdot \left|\left| \texttt{Exp}_{\boldsymbol{x}}^{-1}(\underline{\boldsymbol{m}}_k) \right|\right|_2 \\
&\leq \left|\left| \texttt{Exp}_{\boldsymbol{x}}^{-1}(\underline{\boldsymbol{m}}_k) \right|\right|_2 = d_g(\boldsymbol{x}, \underline{\boldsymbol{m}}_k),
\end{aligned}$$

where we use the fact that $||\mathcal{H}f(\boldsymbol{x})||_2 \leq (q+1)\,||\mathcal{H}f(\boldsymbol{x})||_{\max} \leq (q+1)\,||f||^*_{\infty,3}$ to deduce the last inequality. This shows that the one-step gradient ascent iteration $\texttt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)||f||^*_{\infty,3}} \cdot \texttt{grad}\, f(\boldsymbol{x})\right)$ on Ω_q will stay within the neighborhood $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ whenever $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\} \setminus \{\underline{\boldsymbol{m}}_k\}$. Therefore, $f(\underline{\boldsymbol{m}}_k) - f\left(\texttt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)||f||^*_{\infty,3}} \cdot \texttt{grad}\, f(\boldsymbol{x})\right)\right) \geq 0$.

As for *Fact 3*, note that $||\mathcal{H}f(\boldsymbol{x})||_{\max} \leq ||f||^*_{\infty,3}$ for all $\boldsymbol{x} \in \Omega_q$. Thus,

$$\begin{aligned}\left|\left|\texttt{grad}\, f(\boldsymbol{x}) - \Gamma_{\boldsymbol{y}}^{\boldsymbol{x}}\left(\texttt{grad}\, f(\boldsymbol{y})\right)\right|\right|_2 &= \left|\left|\left(\mathcal{H}f(\boldsymbol{x})\right)\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}^*)\right|\right|_2 \\ &\leq ||\mathcal{H}f(\boldsymbol{x})||_2 \cdot \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 \\ &\leq (q+1)\,||f||^*_{\infty,3}\left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2,\end{aligned}$$

where $\boldsymbol{x}^* \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{x}||_2 \leq ||\boldsymbol{y} - \boldsymbol{x}||_2\}$, and we use the fact that $||A||_2 \leq \sqrt{mn}||A||_{\max}$ for any $A \in \mathbb{R}^{m\times n}$. See Section 3.3 in Genovese et al. (2014) or Section 5.6 in Horn and Johnson (2012) for detailed relations between different types of matrix norms.

With *Fact 1* and *Fact 3*, we have that

$$\begin{aligned}&-\frac{(q+1)\,||f||^*_{\infty,3}}{2}\left|\left|\texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 \leq f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) - \langle\texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y})\rangle, \\ &f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) - \langle\texttt{grad}\, f(\underline{\boldsymbol{m}}_k), \texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y})\rangle \leq -\frac{\lambda_*}{4}\left|\left|\texttt{Exp}_{\underline{\boldsymbol{m}}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 < 0\end{aligned} \tag{B.15}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and $\underline{\boldsymbol{m}}_k \in \underline{\mathcal{M}}$.

Hence, given a point $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and using *Fact 2*,

$$\begin{aligned}&f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) \\ &\leq f(\boldsymbol{y}) - f(\underline{\boldsymbol{m}}_k) + f(\underline{\boldsymbol{m}}_k) - f\left(\texttt{Exp}_{\boldsymbol{y}}\left(\frac{1}{(q+1)\,||f||^*_{\infty,3}} \cdot \texttt{grad}\, f(\boldsymbol{y})\right)\right) \\ &= -\left[f\left(\texttt{Exp}_{\boldsymbol{y}}\left(\frac{1}{(q+1)\,||f||^*_{\infty,3}} \cdot \texttt{grad}\, f(\boldsymbol{y})\right)\right) - f(\boldsymbol{y})\right] \\ &\leq -\left[\langle\texttt{grad}\, f(\boldsymbol{y}), \frac{1}{(q+1)\,||f||^*_{\infty,3}}\texttt{grad}\, f(\boldsymbol{y})\rangle - \frac{(q+1)\,||f||^*_{\infty,3}}{2}\left|\left|\texttt{Exp}_{\boldsymbol{y}}^{-1}\left(\frac{1}{(q+1)\,||f||^*_{\infty,3}} \cdot \texttt{grad}\, f(\boldsymbol{y})\right)\right|\right|_2^2\right] \\ &= -\frac{1}{2(q+1)\,||f||^*_{\infty,3}}\,||\texttt{grad}\, f(\boldsymbol{y})||_2^2,\end{aligned}$$

where we apply the first inequality in (B.15) to obtain the fourth line. Thus, for any $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$,

$$||\texttt{grad}\, f(\boldsymbol{x})||_2^2 \leq 2(q+1)\, ||f||_{\infty,3}^* \left[f(\underline{\boldsymbol{m}}_k) - f(\boldsymbol{x})\right]. \tag{B.16}$$

With $\boldsymbol{y}^{(0)} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ and Corollary B.8, we deduce that

$$\begin{aligned}
d^2(\boldsymbol{y}^{(t+1)}, \underline{\boldsymbol{m}}_k) &\leq d^2(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k) - 2\eta \langle \texttt{grad}\, f(\boldsymbol{y}^{(t)}), \texttt{Exp}_{\boldsymbol{y}^{(t)}}^{-1}(\underline{\boldsymbol{m}}_k)\rangle + \zeta(1, d(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k)) \cdot \eta^2 \left|\left|\texttt{grad}\, f(\boldsymbol{y}^{(t)})\right|\right|_2^2 \\
&\overset{\text{(i)}}{\leq} d^2(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k) + 2\eta \left[f(\boldsymbol{y}^{(t)}) - f(\underline{\boldsymbol{m}}_k) - \frac{\lambda_*}{4} d^2(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k)\right] \\
&\quad + \zeta(1, r_0) \cdot \eta^2 \cdot 2(q+1)\, ||f||_{\infty,3}^* \left[f(\underline{\boldsymbol{m}}_k) - f(\boldsymbol{y}^{(t)})\right] \\
&= \left(1 - \frac{\eta\lambda_*}{2}\right) d^2(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k) - 2\eta \left[1 - \zeta(1, r_0)(q+1)\, ||f||_{\infty,3}^* \, \eta\right] \cdot \underbrace{\left[f(\underline{\boldsymbol{m}}_k) - f(\boldsymbol{y}^{(t)})\right]}_{\geq 0} \\
&\leq \left(1 - \frac{\eta\lambda_*}{2}\right) d^2(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k)
\end{aligned}$$

whenever $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)||f||_{\infty,3}^* \cdot \zeta(1,r_0)}\right\}$, where we use the second inequality of (B.15), the monotonicity of $\zeta(1, c)$ with respect to c, and (B.16) to obtain (i). By telescoping, we conclude that when $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)||f||_{\infty,3}^* \cdot \zeta(1,r_0)}\right\}$ and $\boldsymbol{y}^{(0)} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$,

$$d(\boldsymbol{y}^{(t)}, \underline{\boldsymbol{m}}_k) = \left|\left|\texttt{Exp}_{\boldsymbol{y}^{(t)}}^{-1}(\underline{\boldsymbol{m}}_k)\right|\right|_2 \leq \left(1 - \frac{\eta\lambda_*}{2}\right)^{\frac{t}{2}} \cdot d(\boldsymbol{y}^{(0)}, \underline{\boldsymbol{m}}_k) = \left(1 - \frac{\eta\lambda_*}{2}\right)^{\frac{t}{2}} \left|\left|\texttt{Exp}_{\boldsymbol{y}^{(0)}}^{-1}(\underline{\boldsymbol{m}}_k)\right|\right|_2.$$

The result follows.

(b) **Linear convergence of gradient ascent with $\widehat{f}_h$**: The proof here is partially adapted from the proof of Theorem 2 in Balakrishnan et al. (2017). By (2.29) and the continuity of the exponential map, when h is sufficiently small and $\frac{nh^{q+2}}{|\log h|}$ is sufficiently large, we have that for any $\delta \in (0, 1)$,

$$\begin{aligned}
d\left(\texttt{Exp}_{\boldsymbol{x}}(\eta \cdot \texttt{grad}\, \widehat{f}_h(\boldsymbol{x})), \texttt{Exp}_{\boldsymbol{x}}(\eta \cdot \texttt{grad}\, f(\boldsymbol{x}))\right) &\leq \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x} \in \Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} \\
&\equiv \epsilon_{n,h} \\
&\leq (1 - \Upsilon) \cdot \arccos\left(1 - \frac{r_0^2}{2}\right)
\end{aligned} \tag{B.17}$$

with probability at least $1-\delta$, where $\bar{C}_4$ is some constant independent of $\boldsymbol{x} \in \Omega_q$, and $\epsilon_{n,h} = \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$.
We now claim that $d(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k) \leq \arccos\left(1-\frac{r_0^2}{2}\right)$ and

$$d(\widehat{\boldsymbol{y}}^{(t+1)}, \underline{\boldsymbol{m}}_k) \leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k) + \epsilon_{n,h} \tag{B.18}$$

for any fixed $t = 0, 1, 2, \ldots$ with probability at least $1-\delta$. We will prove this claim by induction on the iteration number. Recall that

$$\widehat{\boldsymbol{y}}^{(t+1)} = \texttt{Exp}_{\widehat{\boldsymbol{y}}^{(t)}}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}^{(t)})\right).$$

Then with $t=0$, we have that

$$\begin{aligned}
&d(\widehat{\boldsymbol{y}}^{(1)}, \underline{\boldsymbol{m}}_k) \\
&= d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}^{(0)}}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}^{(0)})\right), \underline{\boldsymbol{m}}_k\right) \\
&\leq d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}^{(0)}}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}^{(0)})\right), \underline{\boldsymbol{m}}_k\right) + d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}^{(0)}}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}^{(0)})\right), \texttt{Exp}_{\widehat{\boldsymbol{y}}^{(0)}}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}^{(0)})\right)\right) \\
&\leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k) + \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} \\
&= \Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k) + \epsilon_{n,h},
\end{aligned}$$

where the first inequality follows by the triangle inequality, the second one is from our result in (a), whereas the third equality is by (B.17). The triangle inequality is valid in this context because a geodesic measures the minimal distance between two points on Ω_q. In addition, the bound in (B.17) and our initialization $\widehat{\boldsymbol{y}}^{(0)} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ ensure that $d(\widehat{\boldsymbol{y}}^{(1)}, \underline{\boldsymbol{m}}_k) \leq \arccos\left(1-\frac{r_0^2}{2}\right)$. In the induction from $t \to t+1$, suppose that $d(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k) \leq \arccos\left(1-\frac{r_0^2}{2}\right)$ and the claim (B.18) holds at step t. With the fact proved in (a) that

$$d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}^{(t)}}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}^{(t)})\right), \underline{\boldsymbol{m}}_k\right) \leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k),$$

the same argument implies that the claim (B.18) holds for step $t+1$ and $d(\widehat{\boldsymbol{y}}^{(t+1)}, \underline{\boldsymbol{m}}_k) \leq \arccos\left(1-\frac{r_0^2}{2}\right)$. The claim (B.18) is thus proved.

As a result, $\widehat{\boldsymbol{y}}^{(t)}$ always lies within $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \underline{\boldsymbol{m}}_k||_2 \leq r_0\}$ for all $t = 0, 1, \ldots$. Now, with this claim and $\Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}} < 1$, we iterate it to show that

$$\begin{aligned}
d(\widehat{\boldsymbol{y}}^{(t)}, \underline{\boldsymbol{m}}_k) &\leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(t-1)}, \underline{\boldsymbol{m}}_k) + \epsilon_{n,h} \\
&\leq \Upsilon \cdot \left[\Upsilon \cdot d(\widehat{\boldsymbol{y}}^{(t-2)}, \underline{\boldsymbol{m}}_k) + \epsilon_{n,h}\right] + \epsilon_{n,h} \\
&\leq \Upsilon^t \cdot d(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k) + \left\{\sum_{k=0}^{t-1} \Upsilon^k\right\} \cdot \epsilon_{n,h} \\
&\leq \Upsilon^t \cdot d(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k) + \frac{\epsilon_{n,h}}{1 - \Upsilon} \\
&\leq \Upsilon^t \cdot d(\widehat{\boldsymbol{y}}^{(0)}, \underline{\boldsymbol{m}}_k) + O(h^2) + O_P\left(\frac{|\log h|}{nh^{q+2}}\right),
\end{aligned}$$

where the fourth inequality follows by summing the geometric series, and the last one follows from our notation that $\epsilon_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$. It completes the proof. □

B.4 *Proofs of Lemma 3.1, Proposition 3.7, and Theorem 3.10*

A useful interpretation of Lemma 3.1 is that $\widehat{g}_n(\boldsymbol{x})$ diverges to infinity at the rate

$$O\left(h^{-2}\right) + O_P\left(\sqrt{\frac{1}{nh^{D+4}}}\right) = O\left(n^{\frac{2}{D+4}}\right) + O_P(1)$$

if we select the bandwidth $h_{\text{opt}} \asymp O\left(n^{-\frac{1}{D+4}}\right)$ to minimize the asymptotic mean integrated square error (Chen, 2017), where "$\asymp$" stands for the asymptotic equivalence.

Proof of Lemma 3.1. Note that

$$h^2\widehat{g}_n(\boldsymbol{x}) = \mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right] + h^2\widehat{g}_n(\boldsymbol{x}) - \mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]. \tag{B.19}$$

Given the differentiability of p guaranteed by Assumption 2.1, the expectation of $h^2\widehat{g}_n(\boldsymbol{x})$ is given by:

$$\begin{aligned}
\mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right] &= \frac{2c_{k,D}}{h^D}\int_{\mathbb{R}^D} -k'\left(\left|\left|\frac{\boldsymbol{x} - \boldsymbol{y}}{h}\right|\right|_2^2\right) \cdot p(\boldsymbol{y})d\boldsymbol{y} \\
&= 2c_{k,D}\int_{\mathbb{R}^D} -k'\left(||\boldsymbol{u}||_2^2\right) \cdot p(\boldsymbol{x} + h\boldsymbol{u})d\boldsymbol{u} \quad \text{by } \boldsymbol{u} = \frac{\boldsymbol{y} - \boldsymbol{x}}{h}
\end{aligned}$$

$$= 2c_{k,D} \int_{\mathbb{R}^D} -k'\left(||\boldsymbol{u}||_2^2\right)\left[p(\boldsymbol{x}) + h\nabla p(\boldsymbol{x})^T\boldsymbol{u} + O(h^2)\right] d\boldsymbol{u}$$
$$= 2c_{k,D} \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} -k'\left(||\boldsymbol{u}||_2^2\right) d\boldsymbol{u} + O(h^2).$$

By Assumption E1, the dominating constant $-2c_{k,D} \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right) d\boldsymbol{u}$ is finite and therefore,

$$\mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right] = -2c_{k,D} \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right) d\boldsymbol{u} + O\left(h^2\right) = O(1) + O(h^2). \tag{B.20}$$

In addition, we calculate the variance of $\widehat{g}_n(\boldsymbol{x})$ as

$$\begin{aligned}
&\mathrm{Var}\left[h^2\widehat{g}_n(\boldsymbol{x})\right] \\
&= \frac{4c_{k,D}^2}{nh^{2D}} \cdot \mathrm{Var}\left[k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_1}{h}\right|\right|_2^2\right)\right] \\
&= \frac{4c_{k,D}^2}{nh^{2D}} \int_{\mathbb{R}^D} k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{y}}{h}\right|\right|_2^2\right)^2 p(\boldsymbol{y})\, d\boldsymbol{y} - \frac{1}{n}\left\{\mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]\right\}^2 \\
&= \frac{4c_{k,D}^2}{nh^{D}} \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right)^2 \cdot p(\boldsymbol{x}+h\boldsymbol{u}) d\boldsymbol{u} - \frac{1}{n}\left[O(1) + O\left(h^2\right)\right]^2 \qquad \text{by } \boldsymbol{u} = \frac{\boldsymbol{y}-\boldsymbol{x}}{h} \\
&= \frac{4c_{k,D}^2}{nh^{D}} \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right)^2 \left[p(\boldsymbol{x}) + h\nabla p(\boldsymbol{x})^T\boldsymbol{u} + O(h^2)\right] d\boldsymbol{u} + o\left(\frac{1}{nh^D}\right) \\
&= \frac{4c_{k,D}^2 \cdot p(\boldsymbol{x})}{nh^{D}} \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right)^2 d\boldsymbol{u} + o\left(\frac{1}{nh^D}\right),
\end{aligned}$$

Again, by Assumption E1, the dominating constant $4c_{k,D}^2 \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right)^2 d\boldsymbol{u}$ is finite. Thus, by the central limit theorem,

$$\begin{aligned}
h^2\widehat{g}_n(\boldsymbol{x}) - \mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right] &= \sqrt{\mathrm{Var}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]} \cdot \frac{h^2\widehat{g}_n(\boldsymbol{x}) - \mathbb{E}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]}{\sqrt{\mathrm{Var}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]}} \\
&= \sqrt{\mathrm{Var}\left[h^2\widehat{g}_n(\boldsymbol{x})\right]} \cdot \boldsymbol{Z}_n(\boldsymbol{x}) \\
&= O_P\left(\sqrt{\frac{1}{nh^D}}\right),
\end{aligned} \tag{B.21}$$

where $\boldsymbol{Z}_n(\boldsymbol{x}) \xrightarrow{d} \mathcal{N}_D(\boldsymbol{0}, \boldsymbol{I}_D)$. Combining (B.19), (B.20), and (B.21), we conclude that

$$h^2\widehat{g}_n(\boldsymbol{x}) = -2c_{k,D} \cdot p(\boldsymbol{x}) \int_{\mathbb{R}^D} k'\left(||\boldsymbol{u}||_2^2\right) + O\left(h^2\right) + O_P\left(\sqrt{\frac{1}{nh^D}}\right)$$

$$= O(1) + O\left(h^2\right) + O_P\left(\sqrt{\frac{1}{nh^D}}\right)$$

for any $\boldsymbol{x} \in \mathbb{R}^D$ as $nh^D \to \infty$ and $h \to 0$. □

Remark B.1. *Previous work on the mean shift algorithm Li et al. (2007); Carreira-Perpiñán (2007); Ghassabeh (2015); Arias-Castro et al. (2016) have already justified that the algorithm converges to a local mode of the KDE $\widehat{p}_n$ when its local modes are isolated and the algorithm starts within a small neighborhood of each of these estimated local modes. Lemma 3.1 here provides a (probabilistic) perspective on the linear convergence of the mean shift algorithm. It is well-known that the set of true local modes of p can be approximated by the set of estimated modes defined by $\widehat{p}_n$ (Chen et al., 2016). Moreover, around the true local modes of the density p, one can argue that $\widehat{p}_n$ is strongly concave and has a Lipschitz gradient with probability tending to 1 by the uniform consistency of $\nabla\widehat{p}_n$ and $\nabla\nabla\widehat{p}_n$ as $h \to 0$ and $\frac{nh^{D+4}}{|\log h|} \to \infty$; see the uniform bounds (2.27). Hence, by some standard results in optimization theory (*e.g.*, Chapter 3 in Bubeck (2015)), a sample-based gradient ascent algorithm with objective function $\widehat{p}_n$ converges linearly to an estimated local mode when initialized in its neighborhood as long as its step size is below some threshold value. Finally, recall that (i) the mean shift algorithm is a special variant of the sample-based gradient ascent method with an adaptive step size $\eta_{n,h}^{(t)}$ by (3.3) and (ii) $\eta_{n,h}^{(t)}$ can be sufficiently small but bounded away from 0 when nh^D is large and h is small by Lemma 3.1; see also Remark 3.2. Therefore, the mean shift algorithm will converge linearly with high probability from a sufficiently small neighborhood of the local modes of $\widehat{p}_n$ when the sample size n is sufficiently large and h is chosen to be small.*

Proof of Proposition 3.7. (a) We first derive the following fact about the objective function p.

- *Fact 1.* Given Assumption 2.1, p is $D\,||p||_\infty^{(2)}$-smooth, that is, ∇p is $D\,||p||_\infty^{(2)}$-Lipschitz. This fact follows from the differentiability of p ensured by Assumption 2.1 and Taylor's

theorem:

$$||\nabla p(\boldsymbol{x}) - \nabla p(\boldsymbol{y})||_2 \leq ||\nabla \nabla p(\widetilde{\boldsymbol{y}})||_2 \cdot ||\boldsymbol{x} - \boldsymbol{y}||_2 \leq D\,||p||_\infty^{(2)} \cdot ||\boldsymbol{x} - \boldsymbol{y}||_2$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$, where $\widetilde{\boldsymbol{y}}$ is within a $||\boldsymbol{x} - \boldsymbol{y}||_2$-neighborhood of $\boldsymbol{y}$. Moreover,

$$\begin{aligned}
\left|p(\boldsymbol{y}) - p(\boldsymbol{x}) - \nabla p(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})\right| &= \int_0^1 \nabla p(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{x})dt - \nabla p(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \\
&\leq \int_0^1 ||\nabla p(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - p(\boldsymbol{x})||_2 \cdot ||\boldsymbol{y} - \boldsymbol{x}||_2\, dt \\
&\leq \frac{D\,||p||_\infty^{(2)}}{2} \cdot ||\boldsymbol{y} - \boldsymbol{x}||_2^2 .
\end{aligned} \tag{B.22}$$

When $0 < \eta < \frac{2}{D||p||_\infty^{(2)}}$, we have that

$$\begin{aligned}
&p(\boldsymbol{x}^{(t+1)}) - p(\boldsymbol{x}^{(t)}) \\
&= p\left(\boldsymbol{x}^{(t)} + \eta \cdot V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right) - p(\boldsymbol{x}^{(t)}) \\
&\geq \nabla p(\boldsymbol{x}^{(t)})^T \cdot \eta V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)}) - \frac{D\,||p||_\infty^{(2)}}{2}\eta^2 \left|\left|V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \quad \text{by (B.22)} \\
&= \eta\left(1 - \frac{D\,||p||_\infty^{(2)}\,\eta}{2}\right)\left|\left|V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \geq 0,
\end{aligned} \tag{B.23}$$

showing that the objective function p is non-decreasing along $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$. Given the boundedness of p guaranteed by Assumption 2.1, we know that the sequence $\left\{p(\boldsymbol{x}^{(t)})\right\}_{t=0}^{\infty}$ is bounded. Thus, $\left\{p(\boldsymbol{x}^{(t)})\right\}_{t=0}^{\infty}$ also converges.

(b) From (a), we know that when $0 < \eta < \frac{2}{D||p||_\infty^{(2)}}$,

$$\begin{aligned}
p(\boldsymbol{x}^{(t+1)}) - p(\boldsymbol{x}^{(t)}) &\geq \eta\left(1 - \frac{D\,||p||_\infty^{(2)}\,\eta}{2}\right)\left|\left|V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \\
&= \left(\frac{2 - D\,||p||_\infty^{(2)}\,\eta}{2\eta}\right)\left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 .
\end{aligned}$$

Since the sequence $\left\{p(\boldsymbol{x}^{(t)})\right\}_{t=0}^{\infty}$ converges as $t \to \infty$, it follows that

$$\lim_{t\to\infty} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 = 0 \quad \text{and} \quad \lim_{t\to\infty} \left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right|\right|_2 = 0.$$

(c) Given Assumption 2.2 and the fact that $r_1 < \frac{\rho}{2}$, we know that

$$\boldsymbol{x} \in R_d \oplus r_1 \text{ and } \left|\left|V_d(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right|\right|_2 = 0 \quad \text{if and only if} \quad \boldsymbol{x} \in R_d.$$

Let $\boldsymbol{z}^{(t)} = \pi_{R_d}(\boldsymbol{x}^{(t)})$ denote the projection of $\boldsymbol{x}^{(t)}$ in the SCGA sequence onto the ridge R_d. Since $r_1 \leq \texttt{reach}(R_d)$ by (h) of Lemma A.1, $\boldsymbol{z}^{(t)}$ is well-defined when $\boldsymbol{x}^{(t)} \in R_d \oplus r_1$. Given that the definition of $M(\boldsymbol{z}^{(t)}) = \nabla \left[V_d(\boldsymbol{z}^{(t)})^T \nabla p(\boldsymbol{z}^{(t)})\right]^T \in \mathbb{R}^{D\times(D-d)}$ in (A.3), we know that

$$\begin{aligned}
&\left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \\
&= \left|\left| V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)}) - \underbrace{V_d(\boldsymbol{z}^{(t)})^T \nabla p(\boldsymbol{z}^{(t)})}_{=0} \right|\right|_2 \\
&= \left|\left| \int_0^1 M\left(\boldsymbol{z}^{(t)} + \epsilon(\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\right)^T (\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\, d\epsilon \right|\right|_2 \qquad \text{by Taylor's theorem} \\
&= \left|\left| M(\boldsymbol{z}^{(t)})^T (\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}) + \int_0^1 \left[M\left(\boldsymbol{z}^{(t)} + \epsilon(\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\right) - M(\boldsymbol{z}^{(t)})\right]^T (\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\, d\epsilon \right|\right|_2 \\
&\geq \left|\left|M(\boldsymbol{z}^{(t)})^T (\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\right|\right|_2 - \frac{1}{2} \sup_{\epsilon\in[0,1]} \left|\left| \nabla M\left(\boldsymbol{z}^{(t)} + \epsilon(\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\right)^T \right|\right|_2 \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}\right|\right|_2^2 \\
&\overset{\text{(i)}}{\geq} \beta_1 \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}\right|\right|_2 - \frac{B_4(p)}{2} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}\right|\right|_2^2 \\
&= d_E(\boldsymbol{x}^{(t)}, R_d) \cdot \left(\beta_1 - \frac{B_4(p)}{2} d_E(\boldsymbol{x}^{(t)}, R_d)\right) \\
&\geq \frac{\beta_1}{2} \cdot d_E(\boldsymbol{x}^{(t)}, R_d)
\end{aligned} \tag{B.24}$$

when $\boldsymbol{x}^{(t)} \in R_d \oplus r_1$, where we leverage (e) of Lemma A.1 to obtain the inequality (i). More specifically, $M(\boldsymbol{z}^{(t)})^T$ is a full row rank matrix by (d) of Lemma A.1 and $\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}$ lies within the row space of $M(\boldsymbol{z}^{(t)})^T$ because $\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}$ is normal to R_d at $\boldsymbol{z}^{(t)}$. Since the nonzero singular values of $M(\boldsymbol{z}^{(t)})$ are lower bounded by $\beta_1 > 0$, it follows that $\left|\left|M(\boldsymbol{z}^{(t)})^T (\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)})\right|\right|_2 \geq \beta_1 \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{z}^{(t)}\right|\right|_2$. From the above derivation, we also know that $B_4(p) > 0$ is indeed the

supremum norm of $\nabla M(\boldsymbol{x})^T = \nabla\nabla\left[V_d(\boldsymbol{x})^T\nabla p(\boldsymbol{x})\right]^T$ over the line segment connecting $\boldsymbol{x}^{(t)}$ and $\boldsymbol{z}^{(t)}$, which depends on the uniform functional norm $||p||^*_{\infty,4}$ of the partial derivatives of p up to the fourth order. The result follows from (b). □

The following Davis-Kahan theorem (Davis and Kahan, 1970) is one of the most notable theorems in matrix perturbation theory. We present the theorem in a version adapted from von Luxburg (2007); Genovese et al. (2014). Other useful variants of the Davis-Kahan theorem can be found in Yu et al. (2014).

Lemma B.9 (Davis-Kahan). *Let H and $\widetilde{H}$ be two symmetric matrices in $\mathbb{R}^{D\times D}$ whose spectra (Definition 1.1.4 in Horn and Johnson 2012) are $\sigma(H)$ and $\sigma(\widetilde{H})$, and let $S_1 \subset \mathbb{R}$ be an interval. Denote by $\sigma_{S_1}(H)$ the set of eigenvalues of H that are contained in S_1, and by V_1 the matrix whose columns are the corresponding (unit) eigenvectors associated with $\sigma_{S_1}(H)$ (more formally, V_1 is the image of the spectral projection induced by $\sigma_{S_1}(H)$). Denote by $\sigma_{S_1}(\widetilde{H})$ and $\widetilde{V}_1$ the analogous quantities for $\widetilde{H}$. If*

$$\delta := \inf\left\{|\lambda - \widetilde{\lambda}| : \lambda \in \sigma_{S_1}(H), \widetilde{\lambda} \in \sigma(\widetilde{H}) \setminus \sigma_{S_1}(\widetilde{H})\right\} > 0,$$

then the distance $d(V_1, \widetilde{V}_1) := \left|\left|\sin\boldsymbol{\Theta}(V_1, \widetilde{V}_1)\right|\right|$ between two subspaces is bounded by

$$d(V_1, \widetilde{V}_1) \leq \frac{\left|\left|H - \widetilde{H}\right|\right|}{\delta}$$

for any orthogonally invariant norm $||\cdot||$, such as the Frobenius norm $||\cdot||_F$ and the L_2-operator norm $||\cdot||_2$, where $\boldsymbol{\Theta}(V_1, \widetilde{V}_1)$ is a diagonal matrix with the ascending principal angles between the column spaces of V_1 and $\widetilde{V}_1$ on the diagonal.

Note that when we take the Frobenius norm in Lemma B.9, $\left|\left|\sin\boldsymbol{\Theta}(V_1, \widetilde{V}_1)\right|\right|_F = \frac{||V_1V_1^T - \widetilde{V}_1\widetilde{V}_1^T||_F}{\sqrt{2}}$ by straightforward algebra. Consequently, we will utilize the following in-

equality from the Davis-Kahan theorem in our subsequent proofs:

$$\left|\left|V_1V_1^T-\widetilde{V}_1\widetilde{V}_1^T\right|\right|_2\leq\left|\left|V_1V_1^T-\widetilde{V}_1\widetilde{V}_1^T\right|\right|_F=\sqrt{2}\left|\left|\sin\boldsymbol{\Theta}(V_1,\widetilde{V}_1)\right|\right|_F\leq\frac{\sqrt{2}\left|\left|H-\widetilde{H}\right|\right|_F}{\delta}$$
$$\leq\frac{\sqrt{2}D\left|\left|H-\widetilde{H}\right|\right|_{\max}}{\delta}$$
$$\leq\frac{\sqrt{2}D\left|\left|H-\widetilde{H}\right|\right|_2}{\delta}. \tag{B.25}$$

Proof of Theorem 3.10. The entire proof is inspired by some standard results in optimization theory. However, the objective function p is no longer strongly concave, and we focus on the SCGA iteration instead of the standard gradient ascent method. We first recall the following two facts.

- *Fact 1.* Given Assumption 2.1, p is $D\left|\left|p\right|\right|_\infty^{(2)}$-smooth, that is, ∇p is $D\left|\left|p\right|\right|_\infty^{(2)}$-Lipschitz.
- *Fact 2.* Given Assumptions 2.1–2.3, we know that $\left|\left|V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2>0$ for any $\boldsymbol{x}^{(t)}\in\mathrm{Ball}_D(\boldsymbol{x}^*,r_2)\setminus R_d$ and

$$p(\boldsymbol{x}^*)-p\left(\boldsymbol{x}^{(t)}+\frac{1}{D\left|\left|p\right|\right|_\infty^{(2)}}\cdot V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right)\geq 0$$

for any $\boldsymbol{x}^{(t)}\in\mathrm{Ball}_D(\boldsymbol{x}^*,r_2)$.

Fact 1 has been proved in Proposition 3.7, implying that the objective function sequence $\left\{p(\boldsymbol{x}^{(t)})\right\}_{t=0}^\infty$ is non-decreasing when $\eta<\frac{2}{D\left|\left|p\right|\right|_\infty^{(2)}}$. *Fact 2* is a natural corollary by Proposition 3.7, because $\boldsymbol{x}^{(t)}\in\mathrm{Ball}_D(\boldsymbol{x}^*,r_2)$ and $p\left(\boldsymbol{x}^{(t)}+\frac{1}{D\left|\left|p\right|\right|_\infty^{(2)}}\cdot V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\nabla p(\boldsymbol{x}^{(t)})\right)$ is the objective function value after one-step SCGA iteration with step size $\frac{1}{D\left|\left|p\right|\right|_\infty^{(2)}}$. The iteration will move $\boldsymbol{x}^{(t)}$ towards the ridge R_d. With the help of these two facts, we start the proofs of (a-d).

(a) We first show that the following claim: for all $t\geq 0$ and the initial point $\boldsymbol{x}^{(0)}\in\mathrm{Ball}_D(\boldsymbol{x}^*,r_2)$,

$$p(\boldsymbol{x}^*)-p(\boldsymbol{x}^{(t)})\leq\nabla p(\boldsymbol{x}^{(t)})^TV_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\left(\boldsymbol{x}^*-\boldsymbol{x}^{(t)}\right)-\frac{\beta_0}{4}\left|\left|\boldsymbol{x}^*-\boldsymbol{x}^{(t)}\right|\right|_2^2+\epsilon_t, \tag{B.26}$$

where $\epsilon_t = \left(D \left|\left|p\right|\right|_\infty^{(2)} \beta_2^2 \rho + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \right) \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 = o\left(\left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^2 \right)$. By the differentiability of p guaranteed by Assumption 2.1 and Taylor's theorem, we have that

$$
\begin{aligned}
& p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)}) \\
& \leq \nabla p(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) + \frac{1}{2} (\boldsymbol{x}^* - \boldsymbol{x}^{(t)})^T \nabla \nabla p(\boldsymbol{x}^{(t)}) (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 \\
& \overset{\text{(i)}}{=} \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) + \nabla p(\boldsymbol{x}^{(t)})^T U_d^\perp(\boldsymbol{x}^{(t)}) (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \\
& \quad + \frac{1}{2} (\boldsymbol{x}^* - \boldsymbol{x}^{(t)})^T \left(V_\diamond(\boldsymbol{x}^{(t)}), V_d(\boldsymbol{x}^{(t)}) \right) \begin{pmatrix} \lambda_1(\boldsymbol{x}^{(t)}) & & \\ & \ddots & \\ & & \lambda_D(\boldsymbol{x}^{(t)}) \end{pmatrix} \begin{pmatrix} V_\diamond(\boldsymbol{x}^{(t)})^T \\ V_d(\boldsymbol{x}^{(t)})^T \end{pmatrix} (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \\
& \quad + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 \\
& \overset{\text{(ii)}}{\leq} \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) + \frac{\beta_0}{4} \left|\left| \boldsymbol{x}^{(t)} - \boldsymbol{x}^* \right|\right|_2^2 \\
& \quad + \frac{\lambda_1(\boldsymbol{x}^{(t)})}{2} \left|\left| U_d^\perp(\boldsymbol{x}^{(t)}) (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \right|\right|_2^2 - \frac{\beta_0}{2} \left|\left| V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \right|\right|_2^2 \\
& \quad + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 \\
& \overset{\text{(iii)}}{=} \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) + \frac{\beta_0}{4} \left|\left| \boldsymbol{x}^{(t)} - \boldsymbol{x}^* \right|\right|_2^2 \\
& \quad + \frac{\left(\lambda_1(\boldsymbol{x}^{(t)}) + \beta_0\right)}{2} \left|\left| U_d^\perp(\boldsymbol{x}^{(t)}) (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \right|\right|_2^2 - \frac{\beta_0}{2} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^2 + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 \\
& \overset{\text{(iv)}}{\leq} \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) - \frac{\beta_0}{4} \left|\left| \boldsymbol{x}^{(t)} - \boldsymbol{x}^* \right|\right|_2^2 \\
& \quad + \frac{\left(\lambda_1(\boldsymbol{x}^{(t)}) + \beta_0\right)}{2} \cdot \beta_2^2 \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^4 + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3 \\
& \overset{\text{(v)}}{\leq} \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) - \frac{\beta_0}{4} \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^2 \\
& \quad + \left(D \left|\left|p\right|\right|_\infty^{(2)} \beta_2^2 \rho + \frac{D^{\frac{3}{2}} \left|\left|p\right|\right|_\infty^{(3)}}{6} \right) \left|\left| \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \right|\right|_2^3,
\end{aligned}
$$

where we use the equality $\boldsymbol{I}_D = V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T + U_d^\perp(\boldsymbol{x}^{(t)})$ in (i) and (iii), leverage Assumptions 2.2 and A4 to obtain that $\lambda_D(\boldsymbol{x}^{(t)}) \leq \cdots \leq \lambda_{d+1}(\boldsymbol{x}^{(t)}) \leq -\beta_0$ and $\nabla p(\boldsymbol{x}^{(t)})^T U_d^\perp(\boldsymbol{x}^{(t)}) (\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \leq \frac{\beta_0}{4} \left|\left| \boldsymbol{x}^{(t)} - \boldsymbol{x}^* \right|\right|_2^2$ in (ii), apply the quadratic bound

on $\left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})\right|\right|_2$ from Assumption A4 to obtain (iv), and use the fact that $\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \rho$ in (v). We also use the fact that $\max\left\{\beta_0, |\lambda_1(\boldsymbol{x}^{(t)})|\right\} \leq D\,||p||_\infty^{(2)}$ and $\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \rho$ in (v). Claim (B.26) thus follows.

Given *Fact 2* and any $\boldsymbol{x}^{(t)} \in \text{Ball}_D(\boldsymbol{x}^*, r_2)$,

$$
\begin{aligned}
& p(\boldsymbol{x}^{(t)}) - p(\boldsymbol{x}^*) \\
& \leq p(\boldsymbol{x}^{(t)}) - p(\boldsymbol{x}^*) + p(\boldsymbol{x}^*) - p\left(\boldsymbol{x}^{(t)} + \frac{1}{D\,||p||_\infty^{(2)}} \cdot V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right) \\
& = -\left[p\left(\boldsymbol{x}^{(t)} + \frac{1}{D\,||p||_\infty^{(2)}} \cdot V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right) - p(\boldsymbol{x}^{(t)})\right] \\
& \leq -\left[\nabla p(\boldsymbol{x}^{(t)})^T \frac{1}{D\,||p||_\infty^{(2)}} V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)}) - \frac{1}{2D\,||p||_\infty^{(2)}} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2\right] \\
& = -\frac{1}{2D\,||p||_\infty^{(2)}} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2,
\end{aligned}
$$

where we apply (B.22) to obtain the inequality. It implies that

$$
\left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \leq 2D\,||p||_\infty^{(2)} \cdot \left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right]. \tag{B.27}
$$

Therefore,

$$
\begin{aligned}
& \left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^*\right|\right|_2^2 \\
& = \left|\left|\boldsymbol{x}^{(t)} + \eta V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)}) - \boldsymbol{x}^*\right|\right|_2^2 \\
& = \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2^2 + 2\eta \nabla p(\boldsymbol{x}^{(t)})^T V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \left(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right) + \eta^2 \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \\
& \overset{\text{(i)}}{\leq} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2^2 + 2\eta \Bigg[p(\boldsymbol{x}^{(t)}) - p(\boldsymbol{x}^*) - \frac{\beta_0}{4} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2 \\
& \quad + \left(D\,||p||_\infty^{(2)} \beta_2^2 \rho + \frac{D^{\frac{3}{2}}\,||p||_\infty^{(3)}}{6}\right) \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^3\Bigg] + \eta^2 \cdot 2D\,||p||_\infty^{(2)} \cdot \left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right] \\
& \overset{\text{(ii)}}{\leq} \left(1 - \frac{\beta_0 \eta}{4}\right) \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2^2 - 2\eta \left(1 - \eta D\,||p||_\infty^{(2)}\right) \underbrace{\left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right]}_{\geq 0} \\
& \leq \left(1 - \frac{\beta_0 \eta}{4}\right) \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2^2
\end{aligned}
$$

whenever $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$, where we apply (B.26) and (B.27) in (i) and use the choice of $r_2 > 0$ to argue that

$$\begin{aligned}
&\left(D\,||p||_\infty^{(2)}\,\beta_2^2\rho + \frac{D^{\frac{3}{2}}\,||p||_\infty^{(3)}}{6}\right)\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^3 \\
&\leq \left(D\,||p||_\infty^{(2)}\,\beta_2^2\rho + \frac{D^{\frac{3}{2}}\,||p||_\infty^{(3)}}{6}\right) r_2\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2 \\
&\leq \frac{\beta_0}{8}\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2
\end{aligned}$$

in (ii). By telescoping, we conclude that when $0 < \eta \leq \min\left\{\frac{4}{\beta_0}, \frac{1}{D||p||_\infty^{(2)}}\right\}$ and $\boldsymbol{x}^{(0)} \in \text{Ball}_D(\boldsymbol{x}^*, r_2)$,

$$\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq \left(1 - \frac{\beta_0\eta}{4}\right)^{\frac{t}{2}}\left|\left|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right|\right|_2.$$

The result follows.

(b) The result follows easily from (a) and the inequality $d_E(\boldsymbol{x}^{(t)}, R_d) \leq \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2$ for all $t \geq 0$.

(c) The proof here is partially inspired by the proof of Theorem 2 in Balakrishnan et al. (2017). We write the spectral decompositions of $\nabla\nabla p(\boldsymbol{x})$ and $\nabla\nabla\widehat{p}_n(\boldsymbol{x})$ as

$$\nabla\nabla p(\boldsymbol{x}) = V(\boldsymbol{x})\Lambda(\boldsymbol{x})V(\boldsymbol{x})^T \quad \text{and} \quad \nabla\nabla\widehat{p}_n(\boldsymbol{x}) = \widehat{V}(\boldsymbol{x})\widehat{\Lambda}(\boldsymbol{x})\widehat{V}(\boldsymbol{x})^T,$$

where $\Lambda(\boldsymbol{x}) = \texttt{Diag}\left[\lambda_1(\boldsymbol{x}), ..., \lambda_D(\boldsymbol{x})\right]$ and $\widehat{\Lambda}(\boldsymbol{x}) = \texttt{Diag}\left[\widehat{\lambda}_1(\boldsymbol{x}), ..., \widehat{\lambda}_D(\boldsymbol{x})\right]$. By Weyl's theorem (Theorem 4.3.1 in Horn and Johnson 2012) and uniform bounds (2.27), we know that for any $j = 1, ..., D$,

$$\begin{aligned}
|\lambda_j(\boldsymbol{x}) - \widehat{\lambda}_j(\boldsymbol{x})| &\leq \max_j\left|\lambda_j\left(\nabla\nabla p(\boldsymbol{x}) - \nabla\nabla\widehat{p}_n(\boldsymbol{x})\right)\right| \\
&= \left|\left|\nabla\nabla p(\boldsymbol{x}) - \nabla\nabla\widehat{p}_n(\boldsymbol{x})\right|\right|_2 \\
&\leq D\,||p - \widehat{p}_n||_\infty^{(2)} \\
&= O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+4}}}\right).
\end{aligned}$$

Thus, $\widehat{p}_n$ satisfies the first two inequalities in Assumption 2.2 when h is sufficiently small and $\frac{nh^{D+4}}{|\log h|}$ is sufficiently large. According to the Davis-Kahan theorem (Lemma B.9 here) and uniform bounds (2.27),

$$
\begin{aligned}
&\left|\left|G_d(\boldsymbol{y}) - \widehat{G}_d(\boldsymbol{y})\right|\right|_2 \\
&= \left|\left|V_d(\boldsymbol{y})V_d(\boldsymbol{y})^T\nabla p(\boldsymbol{y}) - \widehat{V}_d(\boldsymbol{y})\widehat{V}_d(\boldsymbol{y})^T\nabla\widehat{p}_n(\boldsymbol{y})\right|\right|_2 \\
&\leq \left|\left|\left[V_d(\boldsymbol{y})V_d(\boldsymbol{y})^T - \widehat{V}_d(\boldsymbol{y})\widehat{V}_d(\boldsymbol{y})^T\right]\nabla p(\boldsymbol{y})\right|\right|_2 + \left|\left|\widehat{V}_d(\boldsymbol{y})\widehat{V}_d(\boldsymbol{y})^T\left[\nabla p(\boldsymbol{y}) - \nabla\widehat{p}_n(\boldsymbol{y})\right]\right|\right|_2 \\
&\overset{\text{(i)}}{\leq} \frac{\sqrt{2}D\,||\nabla\nabla p(\boldsymbol{y}) - \nabla\nabla\widehat{p}_n(\boldsymbol{y})||_{\max}\cdot||\nabla p(\boldsymbol{y})||_2}{\beta_0} + ||\nabla p(\boldsymbol{y}) - \nabla\widehat{p}_n(\boldsymbol{y})||_2 \\
&\leq \frac{\sqrt{2}D\,||p - \widehat{p}_n||_\infty^{(2)}\cdot\sqrt{D}\,||p||_\infty^{(1)}}{\beta_0} + \sqrt{D}\,||p - \widehat{p}_n||_\infty^{(1)} \\
&\equiv \epsilon_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+4}}}\right)
\end{aligned}
$$

for any $\boldsymbol{y} \in R_d \oplus r_2$, where we use (B.25) and the fact that $\left|\left|\widehat{V}_d(\boldsymbol{y})\widehat{V}_d(\boldsymbol{y})^T\right|\right|_2 = 1$ to obtain (i). Hence, when $h \to 0$ and $\frac{nh^{D+4}}{|\log h|} \to \infty$,

$$
\left|\left|G_d(\boldsymbol{y}) - \widehat{G}_d(\boldsymbol{y})\right|\right|_2 \leq \epsilon_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{D+4}}}\right) \leq \frac{(1-\Upsilon)r_2}{\eta} \tag{B.28}
$$

with probability tending to 1.

We now claim that $\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq r_2$ and

$$
\left|\left|\widehat{\boldsymbol{x}}^{(t+1)} - \boldsymbol{x}^*\right|\right|_2 \leq \Upsilon\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2 + \eta\cdot\epsilon_{n,h} \tag{B.29}
$$

for all $t \geq 0$. We will prove this claim by induction on the iteration number. Note that when $t = 1$, we derive from triangle inequality that

$$
\begin{aligned}
\left|\left|\widehat{\boldsymbol{x}}^{(1)} - \boldsymbol{x}^*\right|\right|_2 &= \left|\left|\widehat{\boldsymbol{x}}^{(0)} + \eta\widehat{V}_d(\widehat{\boldsymbol{x}}^{(0)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(0)})^T\nabla\widehat{p}_n(\widehat{\boldsymbol{x}}^{(0)}) - \boldsymbol{x}^*\right|\right|_2 \\
&\leq \left|\left|\widehat{\boldsymbol{x}}^{(0)} + \eta V_d(\widehat{\boldsymbol{x}}^{(0)})V_d(\widehat{\boldsymbol{x}}^{(0)})^T\nabla p(\widehat{\boldsymbol{x}}^{(0)}) - \boldsymbol{x}^*\right|\right|_2 + \eta\left|\left|G_d(\widehat{\boldsymbol{x}}^{(0)}) - \widehat{G}_d(\widehat{\boldsymbol{x}}^{(0)})\right|\right| \\
&\leq \Upsilon\left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + \eta\cdot\epsilon_{n,h},
\end{aligned}
$$

where we apply the result in (a) to obtain the last inequality. Moreover, by the choice of $\widehat{\boldsymbol{x}}^{(0)}$ and (B.28), we are guaranteed that $\left|\left|\widehat{\boldsymbol{x}}^{(1)} - \boldsymbol{x}^*\right|\right|_2 \leq r_2$. In the induction from $t \mapsto t+1$, we suppose that $\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq r_2$ and the claim (B.29) holds at iteration t. The same argument then implies that the claim (B.29) holds for iteration $t+1$ and that $\left|\left|\widehat{\boldsymbol{x}}^{(t+1)} - \boldsymbol{x}^*\right|\right|_2 \leq r_2$. The claim (B.29) is thus proved.

Now, given that $\Upsilon = \sqrt{1 - \frac{\beta_0 \eta}{4}} < 1$, we iterate the claim (B.29) to show that

$$\begin{aligned}
\left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2 &\leq \Upsilon \left|\left|\widehat{\boldsymbol{x}}^{(t-1)} - \boldsymbol{x}^*\right|\right|_2 + \eta \cdot \epsilon_{n,h} \\
&\leq \Upsilon \left[\Upsilon \left|\left|\widehat{\boldsymbol{x}}^{(t-2)} - \boldsymbol{x}^*\right|\right|_2 + \eta \cdot \epsilon_{n,h}\right] + \eta \cdot \epsilon_{n,h} \\
&\leq \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + \left[\sum_{s=0}^{t-1} \Upsilon^s\right] \eta \cdot \epsilon_{n,h} \\
&\leq \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + \frac{\eta \cdot \epsilon_{n,h}}{1 - \Upsilon} \\
&= \Upsilon^t \left|\left|\widehat{\boldsymbol{x}}^{(0)} - \boldsymbol{x}^*\right|\right|_2 + O(h^2) + O_P\left(\frac{|\log h|}{nh^{D+4}}\right),
\end{aligned}$$

where the fourth inequality follows by summing the geometric series, and the last equality is due to our notation $\epsilon_{n,h} = O(h^2) + O_P\left(\frac{|\log h|}{nh^{D+4}}\right)$. It completes the proof.

(d) The result follows easily from (c) and the inequality $d_E(\widehat{\boldsymbol{x}}^{(t)}, R_d) \leq \left|\left|\widehat{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^*\right|\right|_2$ for all $t \geq 0$. □

B.5 Discussion on Assumption A4

In this section, we explore several avenues to derive Assumption A4 based on some potentially weaker assumptions. Recall from Section 3.5 that Assumption A4 imposes quadratic behavior on the residual vector $U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})$ and its inner product with the residual gradient, where $\beta_0 > 0$ is the constant defined in Assumption 2.2.

B.5.1 Self-Contractedness Assumption

Assumption A5 (Self-Contractedness)**.** *We assume that the SCGA sequence* $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ *satisfies*

$$\left|\left|\boldsymbol{x}^{(t_3)} - \boldsymbol{x}^{(t_2)}\right|\right|_2 \leq \left|\left|\boldsymbol{x}^{(t_3)} - \boldsymbol{x}^{(t_1)}\right|\right|_2 \quad \text{for all} \quad 0 \leq t_1 \leq t_2 \leq t_3.$$

Assumption A5 requires the SCGA sequence to move towards the ridge R_d along a relatively straight and shrinking path. As we proved in Proposition 3.7, the SCGA sequence $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ converges to R_d when the sequence is initialized near R_d and its step size is small, so Assumption A5 is indeed mild as long as the sequence $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ does not move erratically around R_d. More importantly, Proposition B.10 below shows that Assumption A5 can be implied by a subspace constrained version of the concavity assumption on the objective (density) function p.

Assumption A6 (Subspace Constrained Concavity)**.** *For any* $\boldsymbol{x}, \boldsymbol{y} \in R_d \oplus r_4$ *with* $r_4 > 0$ *being a constant radius, it holds that*

$$p(\boldsymbol{y}) - p(\boldsymbol{x}) \leq \nabla p(\boldsymbol{x})^T V_d(\boldsymbol{x}) V_d(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}).$$

Proposition B.10. *Assume Assumptions 2.1 and A6. Then, the SCGA sequence* $\{\boldsymbol{x}^{(t)}\}_{t=0}^{\infty}$ *defined in* (3.23) *with step size* $0 < \eta \leq \frac{1}{D||p||_{\infty}^{(2)}}$ *and initial point* $\boldsymbol{x}^{(0)} \in R_d \oplus r_4$ *is self-contracted.*

Notice that the density function (3.24) satisfies the "subspace constrained concavity" Assumption A6 around a small neighborhood of its ridge R_1. Moreover, it is intuitive to verify that Assumption A6 is weaker than our established "subspace constrained strong concavity" in Theorem 3.10; see also Remark 3.3.

Proof of Proposition B.10. The proof is inspired by Lemma 14 in Gupta et al. (2021). We show the self-contractedness for $s_3 = t$ as follows, where $t > 0$ is arbitrary. For all $s < t$ and $0 < \eta \leq \frac{1}{D||p||_{\infty}^{(2)}}$ with $\boldsymbol{x}^{(0)} \in R_d \oplus r_4$, we calculate that

$$\left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(t)}\right|\right|_2^2$$

$$
\begin{aligned}
&= \left|\left|\boldsymbol{x}^{(s)} + \eta \cdot V_d(\boldsymbol{x}^{(s)}) V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)}) - \boldsymbol{x}^{(t)}\right|\right|_2^2 \\
&= \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 + 2\eta \nabla p(\boldsymbol{x}^{(s)})^T V_d(\boldsymbol{x}^{(s)}) V_d(\boldsymbol{x}^{(s)})^T (\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}) \\
&\quad + \eta^2 \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&\overset{\text{(i)}}{\leq} \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 + 2\eta \left[p(\boldsymbol{x}^{(s)}) - p(\boldsymbol{x}^{(t)})\right] + \eta^2 \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&\overset{\text{(ii)}}{\leq} \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 + 2\eta \left[p(\boldsymbol{x}^{(s)}) - p(\boldsymbol{x}^{(s+1)})\right] + \eta^2 \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&\overset{\text{(iii)}}{\leq} \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 - 2\eta \cdot \eta \left(1 - \frac{\eta D \left|\left|p\right|\right|_\infty^{(2)}}{2}\right) \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&\quad + \eta^2 \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&= \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 + \eta^2 \cdot \left(\eta D \left|\left|p\right|\right|_\infty^{(2)} - 1\right) \left|\left|V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2^2 \\
&\leq \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2,
\end{aligned}
$$

where we apply Assumption A6 in inequality (i), use the ascending property of p from part (a) of Proposition 3.7 to argue that $p(\boldsymbol{x}^{(t)}) \geq p(\boldsymbol{x}^{(s+1)})$ in inequality (ii), and leverage the inequality (B.23) guaranteed by Assumption 2.1 to obtain (iii). The self-contractedness of the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$ thus follows. □

Under Assumption A5, the following lemma shows that the existing Assumptions 2.1–2.3 in the literature (Genovese et al., 2014; Chen et al., 2015) are nearly sufficient to imply the quadratic behavior of the residual vector $U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})$ along the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$. In other words, Assumption A4 and the linear convergence of the SCGA algorithm hold without any extra assumption.

Lemma B.11. *Assume Assumption A5 throughout the lemma.*

(a) The total length of the SCGA trajectory is of the linear order, i.e.,

$$
\sum_{s=t}^{\infty} \left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right|\right|_2 \leq 2^{10D^2} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \quad \text{for any } t \geq 0.
$$

(b) We further assume Assumptions 2.1 and 2.2. Then,

$$
\nabla p(\boldsymbol{x}^{(t)})^T U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})
$$

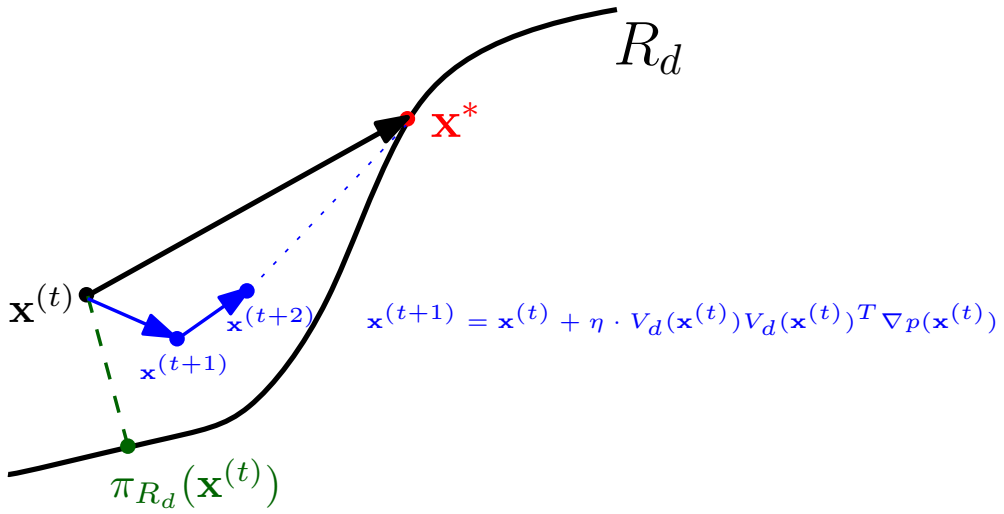


Figure B.2: Decomposition of the vector $\boldsymbol{x}^* - \boldsymbol{x}^{(t)}$ into the summation $\sum_{s=t}^{\infty} \left(\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right)$ of subspace constrained gradient ascent iterative vectors.

$$\leq \frac{2^{10D^2+1} D^{\frac{5}{2}} \left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)}) \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}}{\beta_0} \cdot \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2$$

and

$$\left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)}) \left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)\right|\right|_2 \leq \frac{2^{10D^2+1} D^{\frac{5}{2}} \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}}{\beta_0} \cdot \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2$$

for any $\boldsymbol{x}^{(t)} \in Ball_D(\boldsymbol{x}^*, r_5)$ with some radius $0 < r_5 < \rho$, where we recall that $\rho > 0$ is the effective radius in Assumption 2.2 under which the underlying density p has an eigengap $\beta_0 > 0$ between the d-th and $(d+1)$-th eigenvalues of its Hessian matrix $\nabla\nabla p$.

Proof of Lemma B.11. (a) This result follows directly from Theorem 15 of Gupta et al. (2021). Note that although their results are stated for the standard gradient descent path, the associated proof only utilizes the self-contractedness property of the iterative path. Thus, their proofs are applicable to our SCGA setting under Assumption A5.

(b) We first decompose the vector $\boldsymbol{x}^* - \boldsymbol{x}^{(t)}$ into an infinite sum of SCGA iterations $\boldsymbol{x}^{(s+1)} =$

$\boldsymbol{x}^{(s)} + \eta \cdot V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})$ for $s \geq t$ and obtain that

$$\begin{aligned}
&U_d^{\perp}(\boldsymbol{x}^{(t)}) \left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right) \\
&= U_d^{\perp}(\boldsymbol{x}^{(t)}) \cdot \sum_{s=t}^{\infty} \left(\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right) \\
&= \sum_{s=t}^{\infty} U_d^{\perp}(\boldsymbol{x}^{(t)}) \cdot \eta V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)}) \\
&\overset{\text{(i)}}{=} \sum_{s=t}^{\infty} U_d^{\perp}(\boldsymbol{x}^{(t)}) \cdot \eta \left[V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T - V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\right] V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})
\end{aligned} \tag{B.30}$$

for any $\boldsymbol{x}^{(t)} \in R_d \oplus \rho$, where we leverage the orthogonality between $U_d^{\perp}(\boldsymbol{x}^{(t)})$ and $V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T$ and the idempotence of $V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T$ for all $s \geq t$ in (ii). See also Figure B.2 for a graphical illustration of the decomposition. By the Davis-Kahan theorem (Lemma B.9 and (B.25) here) and Assumptions 2.1 and 2.2, we deduce that for all $s \geq t$,

$$\begin{aligned}
&\left|\left|V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T - V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\right|\right|_2 \\
&\leq \frac{\sqrt{2}D \left|\left|\nabla\nabla p(\boldsymbol{x}^{(s)}) - \nabla\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2}{\beta_0} \\
&\overset{\text{(i)}}{\leq} \frac{\sqrt{2}D \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0} + \frac{\sqrt{2}D^3 \left|\left|p\right|\right|_{\infty}^{(4)}}{\beta_0} \cdot \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 \\
&\overset{\text{(ii)}}{\leq} \frac{2D^{\frac{5}{2}} \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0},
\end{aligned}$$

where we use Taylor's theorem in (i), apply Assumption A5, and possibly shrink the radius $r_5 > 0$ so that $\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \frac{(2-\sqrt{2})\left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}}{\sqrt{2}D\left|\left|p\right|\right|_{\infty}^{(4)}}$ in (ii). Hence, by (B.30) and the fact that $\left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})\right|\right|_2 = 1$, we obtain that

$$\begin{aligned}
&\left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)}) \left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)\right|\right|_2 \\
&\leq \sup_{s \geq t} \left|\left|V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T - V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\right|\right|_2 \cdot \sum_{s=t}^{\infty} \left|\left|\eta V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2 \\
&\leq \frac{2D^{\frac{5}{2}} \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0} \cdot 2^{10D^2} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \\
&= \frac{2^{10D^2+1} D^{\frac{5}{2}} \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0},
\end{aligned}$$

implying the second bound in Assumption A4 with $\beta_2 = \frac{2^{10D^2+1}D^{\frac{5}{2}}||p||_\infty^{(3)}}{\beta_0}$. In addition,

$$\begin{aligned}
&\nabla p(\boldsymbol{x}^{(t)})^T U_d^\perp(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \\
&\leq \left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)\right|\right|_2 \\
&\leq \frac{2^{10D^2+1}D^{\frac{5}{2}}\left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0}.
\end{aligned}$$

The results follow. □

According to part (b) of Lemma B.11, Assumption A4 holds with $\beta_2 = \frac{2^{10D^2+1}D^{\frac{5}{2}}||p||_\infty^{(3)}}{\beta_0}$ whenever

$$2^{10D^2+1}D^{\frac{5}{2}}\left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \leq \frac{\beta_0^2}{4}. \tag{B.31}$$

The choice of $\beta_2 > 0$ is valid under Assumption 2.1. More importantly, (B.31) is essentially the same as the first inequality in Assumption 2.3. Compared with the corresponding inequality in Assumption 2.3, the upper bound in (B.31) for $\left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}$ around the ridge R_d is only shrunk by a dimension-dependent factor $\frac{1}{D\cdot 2^{10D^2+2}}$. As Assumption 2.3 and (B.31) are local, this adjustment does not induce too much extra strictness on the underlying density p.

B.5.2 Subspace Constrained Polyak-Łojasiewicz Inequality Assumption

We have demonstrated in Appendix B.5.1 that the crucial Assumption A4 is valid under the self-contractedness Assumption A5 on the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^\infty$. Consequently, the linear convergence of the SCGA algorithm can be established by slightly modifying the common Assumptions 2.1–2.3 in ridge estimation. Nevertheless, the self-contractedness property of the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^\infty$ does not always hold in practice, and it may only be implied by the subspace constrained concavity Assumption A6 as proved in Proposition B.10.

Given that the underlying density function p or its estimator $\widehat{p}_n$ may not satisfy the subspace constrained concavity assumption in many practical applications of SCGA and SCMS algorithms, we present another approach to deduce Assumption A4 based on the well-known Polyak-Łojasiewicz inequality (Polyak, 1963; Lojasiewicz, 1963). Given any SCGA

sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$ with limiting point $\boldsymbol{x}^* \in R_d$ and step size $0 < \eta < \frac{2}{D||p||_{\infty}^{(2)}}$, we consider the following assumption:

Assumption A7 (Subspace Constrained Polyak-Łojasiewicz Inequality)**.** *For all $t \geq 0$, there exists a constant $\beta_3 > 0$ such that*

$$\frac{1}{2}\left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \geq \beta_3 \left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right].$$

As with the standard Polyak-Łojasiewicz inequality, some objective functions satisfy Assumption A7 but fail to be concave in the subspace constrained sense of Assumption A6; see Charles and Papailiopoulos (2018); Fazel et al. (2018) and Equation (36) in Chen (2022). In this respect, Assumption A7 incorporates additional SCGA sequences that satisfy Assumption A4 and converge linearly to the ridge R_d. However, as Assumption A7 does not imply Assumption A5 or A6, it should not be regarded as a more general condition. Furthermore, unlike the standard gradient ascent/descent method (Theorem 2 in Karimi et al. 2016), the error bound condition (*i.e.*, Equation (B.24) here) does not imply Assumption A7, indicating a challenge in validating this assumption in practice.

Despite these disadvantages, the subspace constrained Polyak-Łojasiewicz inequality condition does give rise to a concise proof for the linear convergence of the objective function value $\left\{p(\boldsymbol{x}^{(t)})\right\}_{t=0}^{\infty}$ along the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$.

Proposition B.12. *Assume Assumptions 2.1 and A7. Then, for any SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^{\infty}$ with step size $0 < \eta < \min\left\{\frac{1}{D||p||_{\infty}^{(2)}}, \frac{1}{\beta_3}\right\}$, we have that*

$$p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)}) \leq \left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(0)})\right] \cdot (1 - \eta\beta_3)^t.$$

Proof of Proposition B.12. The proof is inspired by Theorem 1 in Karimi et al. (2016). From (B.23) and Assumption A7, we know that

$$p(\boldsymbol{x}^{(t+1)}) - p(\boldsymbol{x}^{(t)}) \geq \eta\left(1 - \frac{D\,||p||_{\infty}^{(2)}\,\eta}{2}\right)\left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2 \geq \eta\beta_3\left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right]$$

for all $t \geq 0$ when $0 < \eta < \frac{1}{D||p||_\infty^{(2)}}$. By some rearrangements, we conclude that

$$p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t+1)}) \leq (1 - \eta\beta_3) \left[p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})\right].$$

The final display follows from telescoping. □

More importantly, Assumption A7 controls the total length of the SCGA path to be of the linear order and implies the quadratic behavior of residual vectors required by Assumption A4.

Lemma B.13. *Assume Assumptions 2.1 and A7 throughout the lemma.*

(a) The total length of the SCGA trajectory is of the linear order, i.e.,

$$\sum_{s=t}^{\infty} \left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right|\right|_2 \leq \frac{4D\,||p||_\infty^{(2)}}{\beta_3} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \quad \text{for any } t \geq 0.$$

(b) We further assume Assumption 2.2. Then,

$$\begin{aligned}
&\nabla p(\boldsymbol{x}^{(t)})^T U_d^\perp(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \\
&\leq \frac{32D^{\frac{9}{2}} \left(||p||_\infty^{(2)}\right)^2 \left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0\beta_3^2},
\end{aligned}$$

and

$$\left|\left|U_d^\perp(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)})\right|\right|_2 \leq \frac{32D^{\frac{9}{2}} \left(||p||_\infty^{(2)}\right)^2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0\beta_3^2}$$

for any $\boldsymbol{x}^{(t)} \in Ball_D(\boldsymbol{x}^*, r_6)$ with some radius $0 < r_6 < \rho$, where we recall that $\rho > 0$ is the effective radius in Assumption 2.2 under which the underlying density p has an eigengap $\beta_0 > 0$ between the d-th and $(d+1)$-th eigenvalues of its Hessian matrix $\nabla\nabla p$.

Proof of Lemma B.13. (a) This part of the proof is inspired by the arguments in Theorem 9 of Gupta et al. (2021). Based on the proof of part (a) in Proposition 3.7 under Assumption 2.1, we know from (B.23) that

$$p(\boldsymbol{x}^{(t+1)}) - p(\boldsymbol{x}^{(t)}) \geq \eta \left(1 - \frac{D\,||p||_\infty^{(2)}\,\eta}{2}\right) \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2$$

$$\geq \frac{\eta}{2} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2$$

when $0 < \eta < \frac{1}{D||p||_\infty^{(2)}}$. Using this inequality and Assumption A7, we derive that

$$\begin{aligned}
\sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t+1)})} &= \sqrt{p(\boldsymbol{x}^*) - p\left(\boldsymbol{x}^{(t)} + \eta \cdot V_d(\boldsymbol{x}^{(t)}) V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right)} \\
&\leq \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)}) - \frac{\eta}{2} ||V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})||_2^2} \\
&\overset{\text{(i)}}{\leq} \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})} - \frac{\eta \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2^2}{4\sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})}} \\
&\overset{\text{(ii)}}{\leq} \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})} - \frac{\eta\sqrt{2\beta_3}}{4} \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2,
\end{aligned}$$

where we use the inequality $\sqrt{a-b} \leq \sqrt{a} - \frac{b}{2\sqrt{a}}$ to obtain (i) and apply Assumption A7 in inequality (ii). Since $\left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right|\right|_2 = \eta \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2$, some rearrangement of the above inequality suggests that

$$\sqrt{\frac{\beta_3}{8}} \left|\left|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})} - \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t+1)})}.$$

Therefore,

$$\begin{aligned}
\sum_{s=t}^{\infty} \left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right|\right|_2 &\leq \sqrt{\frac{8}{\beta_3}} \sum_{s=t}^{\infty} \left[\sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(s)})} - \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(s+1)})}\right] \\
&= \sqrt{\frac{8}{\beta_3}} \cdot \sqrt{p(\boldsymbol{x}^*) - p(\boldsymbol{x}^{(t)})} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{8}{\beta_3}} \cdot \sqrt{\frac{1}{2\beta_3}} \cdot \left|\left|V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \quad \text{by Assumption A4} \\
&= \frac{2}{\beta_3} \left|\left| V_d(\boldsymbol{x}^{(t)})^T \nabla p(\boldsymbol{x}^{(t)}) - \underbrace{V_d(\boldsymbol{x}^*)^T \nabla p(\boldsymbol{x}^*)}_{=0} \right|\right|_2 \\
&\overset{\text{(ii)}}{\leq} \frac{2}{\beta_3} \left|\left| \sup_{\epsilon \in [0,1]} M(\boldsymbol{x}^* + \epsilon(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*))^T \left(\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right) \right|\right|_2 \\
&\overset{\text{(iii)}}{\leq} \frac{2}{\beta_3} \left(D\,||p||_\infty^{(2)} + \beta_0 - \beta_1\right) \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2 \\
&\leq \frac{4D\,||p||_\infty^{(2)}}{\beta_3} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2,
\end{aligned}$$

where we leverage Assumption A7 again in (i). In addition, to obtain inequalities (ii) and (iii), we recall from the proof of part (d) of Lemma A.1 that $M(\boldsymbol{x}) = \nabla \left[V_d(\boldsymbol{x})^T \nabla p(\boldsymbol{x})\right]^T = V_d(\boldsymbol{x})\Lambda_0(\boldsymbol{x}) + \sum_{i=1}^d T_i(\boldsymbol{x}) V_d(\boldsymbol{x})\Lambda_i(\boldsymbol{x})$, in which the singular values of $V_d(\boldsymbol{x})\Lambda_0(\boldsymbol{x})$ are bounded by $D\,||p||_\infty^{(2)}$ and the singular values of $\sum_{i=1}^d T_i(\boldsymbol{x}) V_d(\boldsymbol{x})\Lambda_i(\boldsymbol{x})$ are bounded by $\beta_0 - \beta_1 \leq D\,||p||_\infty^{(2)}$. The result thus follows.

(b) This part of the proof is analogous to our arguments in part (b) of Lemma B.11, except that the SCGA sequence $\left\{\boldsymbol{x}^{(t)}\right\}_{t=0}^\infty$ is no longer self-contracted. For completeness, we still repeat some arguments and highlight the differences here. By the Davis-Kahan theorem (Lemma B.9 and (B.25) here) and Assumptions 2.1 and 2.2, we have that for all $s \geq t$,

$$
\begin{aligned}
&\left|\left|V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T - V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\right|\right|_2 \\
&\leq \frac{\sqrt{2}D\left|\left|\nabla\nabla p(\boldsymbol{x}^{(s)}) - \nabla\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2}{\beta_0} \\
&\leq \frac{\sqrt{2}D\left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0} + \frac{\sqrt{2}D^3\,||p||_\infty^{(4)}}{\beta_0}\left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2^2 \\
&\overset{\text{(ii)}}{\leq} \frac{4\sqrt{2}D^2\,||p||_\infty^{(2)}\left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0\beta_3} + \frac{16\sqrt{2}D^5\,||p||_\infty^{(4)}\left(||p||_\infty^{(2)}\right)^2}{\beta_0\beta_3^2}\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2^2 \\
&\overset{\text{(ii)}}{\leq} \frac{8D^{\frac{7}{2}}\,||p||_\infty^{(2)}\left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}\left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0\beta_3},
\end{aligned}
$$

where we possibly shrink the radius $r_6 > 0$ so that $\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \leq \frac{(\sqrt{2}-1)\beta_3\left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}}{4D^{\frac{3}{2}}||p||_\infty^{(2)}||p||_\infty^{(4)}}$ to obtain inequality (ii). Notice also that, since $\left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2$ may not hold without the self-contractedness property, we use a looser bound

$$
\left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2 \leq \sum_{s=t}^\infty \left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right|\right|_2 \leq \frac{4D\,||p||_\infty^{(2)}}{\beta_3}\left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2
$$

from (a) to derive inequality (i). Therefore, by (B.30) and the fact that

$$
\left|\left|\eta V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T\nabla p(\boldsymbol{x}^{(s)})\right|\right|_2 = \left|\left|\boldsymbol{x}^{(s+1)} - \boldsymbol{x}^{(s)}\right|\right|_2,
$$

we obtain that

$$
\left|\left|U_d^\perp(\boldsymbol{x}^{(t)})\left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)\right|\right|_2
$$

$$
\begin{aligned}
&\leq \sup_{s\geq t} \left|\left|V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T - V_d(\boldsymbol{x}^{(t)})V_d(\boldsymbol{x}^{(t)})^T\right|\right|_2 \cdot \sum_{s=t}^{\infty} \left|\left|\eta V_d(\boldsymbol{x}^{(s)})V_d(\boldsymbol{x}^{(s)})^T \nabla p(\boldsymbol{x}^{(s)})\right|\right|_2 \\
&\leq \frac{8D^{\frac{7}{2}} \left|\left|p\right|\right|_\infty^{(2)} \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^{(s)} - \boldsymbol{x}^{(t)}\right|\right|_2}{\beta_0\beta_3} \cdot \frac{4D \left|\left|p\right|\right|_\infty^{(2)}}{\beta_3} \left|\left|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right|\right|_2 \\
&= \frac{32D^{\frac{9}{2}} \left(\left|\left|p\right|\right|_\infty^{(2)}\right)^2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0\beta_3^2},
\end{aligned}
$$

which implies the second bound in Assumption A4 with $\beta_2 = \frac{32D^{\frac{9}{2}}\left(\left|\left|p\right|\right|_\infty^{(2)}\right)^2 \left|\left|p\right|\right|_\infty^{(3)}}{\beta_0\beta_3^2}$. Finally,

$$
\begin{aligned}
&\nabla p(\boldsymbol{x}^{(t)})^T U_d^{\perp}(\boldsymbol{x}^{(t)})(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}) \\
&\leq \left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})\left(\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right)\right|\right|_2 \\
&\leq \frac{32D^{\frac{9}{2}} \left(\left|\left|p\right|\right|_\infty^{(2)}\right)^2 \left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max} \left|\left|\boldsymbol{x}^* - \boldsymbol{x}^{(t)}\right|\right|_2^2}{\beta_0\beta_3^2}.
\end{aligned}
$$

The results follow. □

The results in part (b) of Lemma B.13 also imply Assumption A4 with $\beta_2 = \frac{32D^{\frac{9}{2}}\left(\left|\left|p\right|\right|_\infty^{(2)}\right)^2 \left|\left|p\right|\right|_\infty^{(3)}}{\beta_0\beta_3^2}$ whenever

$$
\frac{32D^{\frac{9}{2}} \left(\left|\left|p\right|\right|_\infty^{(2)}\right)^2 \left|\left|U_d^{\perp}(\boldsymbol{x}^{(t)})\nabla p(\boldsymbol{x}^{(t)})\right|\right|_2 \left|\left|\nabla^3 p(\boldsymbol{x}^{(t)})\right|\right|_{\max}}{\beta_3^2} \leq \frac{\beta_0^2}{4}. \tag{B.32}
$$

Once again, the choice of β_2 is feasible under Assumption 2.1, and the upper bound (B.32) can be viewed as a variant of the first inequality in Assumption 2.3. From this perspective, Assumption A7 also leads to an alternative route to Assumption A4 and the linear convergence of the SCGA algorithm.

Remark B.2. *Note that the results in Proposition B.13 can be generalized to the directional or arbitrary manifold cases under Assumptions 2.1–2.3. First, the subspace constrained Polyak-Łojasiewicz inequality for the SCGA sequence* $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ *on* Ω_q *or an arbitrary manifold can be modified as:*

$$
\frac{1}{2}\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2 \geq \underline{\beta}_3 \left[f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)})\right] \quad \textit{for some } \underline{\beta}_3 > 0 \textit{ and any } t \geq 0,
$$

where f is the objective (density) function. Based on the proof of (a) in Proposition 3.13 and our arguments in (a) of Lemma B.13, it follows that the total length of the SCGA trajectory on Ω_q or an arbitrary manifold is of the linear order, i.e.,

$$\sum_{s=t}^{\infty} d_g\left(\underline{\boldsymbol{x}}^{(s+1)}, \underline{\boldsymbol{x}}^{(s)}\right) \leq \frac{4q\,||f||_{\infty}^{(2)}}{\underline{\beta}_3} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^{*}\right) \quad \text{for any } t \geq 0.$$

Second, to establish the quadratic bounds for $\left\langle \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)})\texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{*})\right\rangle$ *and* $\left|\left|\underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)})\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{*})\right|\right|_2$*, one can follow the arguments in the proof of (b) in Lemma B.13 and leverage the two facts:*

1. *The tangent vector* $\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{*})$ *can be decomposed into an infinite sum of SCGA updates (3.31) on* Ω_q *or an arbitrary manifold as:*

$$\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{*}) = \sum_{s=t}^{\infty} \Gamma_{\underline{\boldsymbol{x}}^{(s)}}^{\underline{\boldsymbol{x}}^{(t)}}\left(\texttt{Exp}_{\underline{\boldsymbol{x}}^{(s)}}^{-1}(\underline{\boldsymbol{x}}^{(s+1)})\right).$$

See Figure B.3 for a graphical illustration. This equation is valid because parallel transports preserve inner products and are linear.

2. *Under Assumptions 2.1 and 2.2, we know that*

$$\begin{aligned}
&\left|\left|\underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \cdot \sum_{s=t}^{\infty} \Gamma_{\underline{\boldsymbol{x}}^{(s)}}^{\underline{\boldsymbol{x}}^{(t)}}\left(\texttt{Exp}_{\underline{\boldsymbol{x}}^{(s)}}^{-1}(\underline{\boldsymbol{x}}^{(s+1)})\right)\right|\right|_2 \\
&= \left|\left| \sum_{s=t}^{\infty} \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \left[\Gamma_{\underline{\boldsymbol{x}}^{(s)}}^{\underline{\boldsymbol{x}}^{(t)}}\left(\underline{\eta} \underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T \underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(s)})\right)\right.\right. \\
&\qquad\qquad \left.\left. - \underline{\eta} \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(s)})\right]\right|\right|_2 \\
&\leq \sum_{s=t}^{\infty} \underline{A}\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T - \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T\right|\right|_2 \cdot \left|\left|\underline{\boldsymbol{x}}^{(s)} - \underline{\boldsymbol{x}}^{(t)}\right|\right|_2 \\
&\leq \underline{A}\left|\left|\underline{\boldsymbol{x}}^{*} - \underline{\boldsymbol{x}}^{(t)}\right|\right|_2 \cdot \sum_{s=t}^{\infty} \left|\left|\underline{\boldsymbol{x}}^{(s)} - \underline{\boldsymbol{x}}^{(t)}\right|\right|_2 \\
&= O\left(\left|\left|\underline{\boldsymbol{x}}^{*} - \underline{\boldsymbol{x}}^{(t)}\right|\right|_2^2\right)
\end{aligned}$$

for some constant $\underline{A} > 0$*, where we leverage the fact that the vector field*

$$X(\gamma(t)) = \underline{V}_d(\gamma(t))\underline{V}_d(\gamma(t))^T \underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(s)})$$

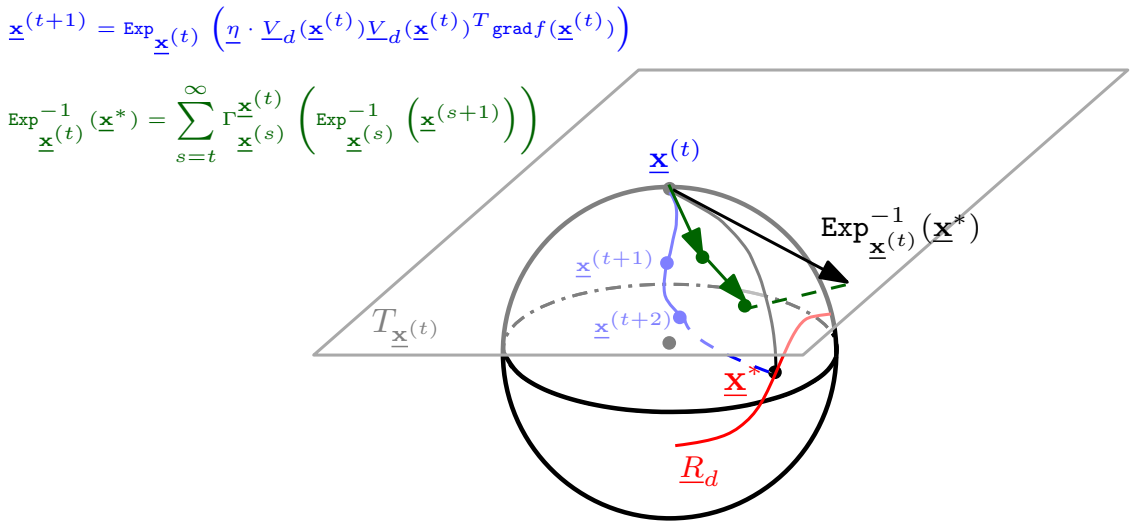


Figure B.3: Decomposition of the vector $\mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*)$ within the tangent space $T_{\underline{\boldsymbol{x}}^{(t)}}$ into the summation $\sum_{s=t}^{\infty} \Gamma^{\underline{\boldsymbol{x}}^{(t)}}_{\underline{\boldsymbol{x}}^{(s)}}\left(\mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(s)}}(\underline{\boldsymbol{x}}^{(s+1)})\right)$ of parallel transported SCGA iterative vectors. Here, the blue curves on Ω_q are iterative paths of the SCGA algorithm, while the green vectors are tangent vectors $\mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(s)}}(\underline{\boldsymbol{x}}^{(s+1)}) \in T_{\underline{\boldsymbol{x}}^{(s)}}$ after being parallel transported to $T_{\underline{\boldsymbol{x}}^{(t)}}$.

with $\gamma(0) = \underline{\boldsymbol{x}}^{(s)}$ *and* $\gamma(1) = \underline{\boldsymbol{x}}^{(t)}$ *has its variation* $||X(\gamma(t_1)) - X(\gamma(t_2))||_2$ *bounded by*

$$\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})\underline{V}_d(\underline{\boldsymbol{x}}^{(s)})^T - \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T\right|\right|_2 = O\left(\left|\left|\underline{\boldsymbol{x}}^{(s)} - \underline{\boldsymbol{x}}^{(t)}\right|\right|_2\right)$$

according to the Davis-Kahan theorem for any $0 \leq t_1, t_2 \leq 1$. *However, we are not sure if the self-contractedness condition can also be adaptive to the directional or general manifold cases, given that the arguments in Theorem 15 of Gupta et al. (2021) are based on the Euclidean geometry.*

B.6 Other Technical Concepts of Differential Geometry on Ω_q

• **Taylor's Theorem on** Ω_q. Given a smooth function f on Ω_q, its Taylor's expansion is often written as (Pennec, 2006):

$$f(\mathtt{Exp}_{\boldsymbol{x}}(\boldsymbol{v})) = f(\boldsymbol{x}) + \langle \mathtt{grad}\, f(\boldsymbol{x}), \boldsymbol{v}\rangle + \frac{1}{2}\boldsymbol{v}^T \mathcal{H} f(\boldsymbol{x})\boldsymbol{v} + o\left(||\boldsymbol{v}||_2^2\right) \tag{B.33}$$

for any $\boldsymbol{v} \in T_{\boldsymbol{x}}$, where $\texttt{Exp}_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \Omega_q$ is the *exponential map* at $\boldsymbol{x} \in \Omega_q$. One may replace the exponential map with a more general concept called the *retractions* on an arbitrary manifold; see Section 4.1 and Proposition 5.5.5 in Absil et al. (2008).

• **Parallel Transport**. When comparing vectors in two different tangent spaces $T_{\boldsymbol{x}}, T_{\boldsymbol{y}}$ on Ω_q, we leverage the notion of *parallel transport* $\Gamma_{\boldsymbol{x}}^{\boldsymbol{y}} : T_{\boldsymbol{x}} \to T_{\boldsymbol{y}}$ to transport vectors from one tangent space to another along a geodesic. In addition, $\Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{v})$ is a tangent vector in $T_{\boldsymbol{y}}$ after being parallel transported from $\boldsymbol{v} \in T_{\boldsymbol{x}}$ along a geodesic (or great circle) on Ω_q. The parallel transport mapping $\Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}$ is a linear isometry along any smooth curve on Ω_q, *i.e.*, $\langle \Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{u}), \Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{v}) \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle$ for any $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{x}}$; see Proposition 5.5 in Lee (2018) or Proposition 1 in Section 4-4 of do Carmo (2016).

• **Sectional Curvature**. *Sectional curvature* is the Gaussian curvature of a two-dimensional submanifold formed as the image of a two-dimensional subspace of a tangent space after exponential mapping; see Section 3-2 in do Carmo (2016) for a detailed discussion of the Gaussian curvature. It is known that a two-dimensional submanifold with positive, zero, or negative sectional curvature is locally isometric to a two-dimensional sphere, a Euclidean plane, or a hyperbolic plane with the same Gaussian curvature (Zhang and Sra, 2016).

• **Geodesically Strong Concavity**. A function $f : \Omega_q \to \mathbb{R}$ is said to be *geodesically concave* if for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega_q$, it holds that

$$f(\varphi(t)) \geq (1-t) f(\boldsymbol{x}) + f(\boldsymbol{y})$$

for any $t \in [0,1]$, where $\varphi : [0,1] \to \Omega_q$ is a geodesic with $\varphi(0) = \boldsymbol{x}$ and $\varphi(1) = \boldsymbol{y}$. When f is differentiable, an equivalent statement of the geodesic concavity is that (Theorem 11.17 in Boumal 2020):

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \leq \left\langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \right\rangle.$$

A function $f : \Omega_q \to \mathbb{R}$ is said to be *geodesically μ_g-strongly concave* if for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega_q$, it holds that

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \left\langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \right\rangle - \frac{\mu_g}{2} \cdot d_g(\boldsymbol{x}, \boldsymbol{y})^2.$$

B.7 Proofs of Proposition 3.12, Proposition 3.13, and Theorem 3.15

This section adapts the Euclidean arguments to the sphere. The proofs replace Euclidean line segments by geodesics, Euclidean distances by geodesic distances, and ordinary Hessian bounds by their Riemannian analogues.

Proof of Proposition 3.12. (a) The sequence $\left\{\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$ is bounded if the kernel L is non-increasing with $L(0) < \infty$. Hence, it suffices to show that it is non-decreasing. The convexity and differentiability of kernel L imply that

$$L(x_2) - L(x_1) \geq L'(x_1) \cdot (x_2 - x_1) \tag{B.34}$$

for all $x_1, x_2 \in [0, \infty)$. Then, with $\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{L,q}(h)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ and the iterative formula (3.28) in the main paper, we derive that

$$\begin{aligned}
&\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}) - \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \\
&= \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} \left[L\left(\frac{1 - \boldsymbol{X}_i^T \widehat{\underline{\boldsymbol{x}}}^{(t+1)}}{h^2}\right) - L\left(\frac{1 - \boldsymbol{X}_i^T \widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right) \right] \\
&\geq \frac{c_{L,q}(h)}{nh^2} \sum_{i=1}^{n} L'\left(\frac{1 - \boldsymbol{X}_i^T \widehat{\underline{\boldsymbol{x}}}^{(t)}}{h^2}\right) \boldsymbol{X}_i^T \left(\widehat{\underline{\boldsymbol{x}}}^{(t)} - \widehat{\underline{\boldsymbol{x}}}^{(t+1)}\right) \\
&= \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \left(\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right) \\
&= \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \left[\frac{\underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)}}{\left|\left| \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)} \right|\right|_2} - \widehat{\underline{\boldsymbol{x}}}^{(t)} \right] \\
&\overset{\text{(i)}}{=} \frac{\left|\left| \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2}{\sqrt{\left|\left| \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2 + \left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2}} \\
&\quad + \frac{\left[\left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2 - \sqrt{\left|\left| \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2 + \left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2} \right] \cdot \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \widehat{\underline{\boldsymbol{x}}}^{(t)}}{\sqrt{\left|\left| \underline{\widehat{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2 + \left|\left| \nabla \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) \right|\right|_2^2}}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{=} \frac{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}{\sqrt{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}} \\
&\quad\times\frac{\left[\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2+\sqrt{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}-\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\underline{\widehat{\boldsymbol{x}}}^{(t)}\right]}{\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2+\sqrt{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}} \\
&\overset{\text{(iii)}}{\geq} \frac{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}{\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2+\sqrt{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}} \\
&\overset{\text{(iv)}}{\geq} \frac{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}{\left(1+\sqrt{2}\right)\cdot\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2} \\
&\geq 0,
\end{aligned}
$$

where we use the orthogonality between $\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T$ and $\underline{\widehat{\boldsymbol{x}}}^{(t)}$ in (i), multiply $\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2+\sqrt{\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2+\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2^2}$ to both the numerators and denominators of the two summands to obtain (ii), leverage the fact that $\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\geq\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\underline{\widehat{\boldsymbol{x}}}^{(t)}$ in (iii), use the inequality $\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\leq\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2$ in (iv). It thus completes the proof of (a).

(b) Our derivation in (a) already shows that

$$
\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\leq\sqrt{\left(1+\sqrt{2}\right)\cdot\left|\left|\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\cdot\left[\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t+1)})-\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right]}.
$$

Notice that, on the one hand, the differentiability of kernel L and the compactness of Ω_q imply that $\left|\left|\nabla\widehat{f}_h(\boldsymbol{x})\right|\right|_2\leq B_{h,L}$ for all $\boldsymbol{x}\in\Omega_q$, where $B_{h,L}>0$ only depends on the bandwidth h and kernel L. On the other hand, our argument in (a) already proves the convergence of $\left\{\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right\}_{t=0}^{\infty}$. Therefore,

$$
\left|\left|\underline{\widehat{V}}_d(\underline{\widehat{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right|\right|_2\leq\sqrt{\left(1+\sqrt{2}\right)B_{h,L}\cdot\left[\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t+1)})-\widehat{f}_h(\underline{\widehat{\boldsymbol{x}}}^{(t)})\right]}\to 0
$$

as $t \to \infty$. The result follows.

(c) Given the iterative formula (3.28) in the main paper, we deduce that

$$
\begin{aligned}
&\left\|\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right\|_2^2 \\
&= \left\| \frac{\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)}}{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2 \cdot \widehat{\underline{\boldsymbol{x}}}^{(t)}\right\|_2} - \widehat{\underline{\boldsymbol{x}}}^{(t)} \right\|_2^2 \\
&\overset{\text{(i)}}{=} \frac{\left\| \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)}) + \left[\left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2 - \sqrt{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}\right] \widehat{\underline{\boldsymbol{x}}}^{(t)} \right\|_2^2}{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2} \\
&\overset{\text{(ii)}}{=} \frac{2\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2} \\
&\quad + \underbrace{\frac{2\left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 - 2\left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2 \cdot \sqrt{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}}{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}}_{\leq 0} \\
&\leq \frac{2\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2},
\end{aligned}
$$

where we leverage the orthogonality between $\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})$ and $\widehat{\underline{\boldsymbol{x}}}^{(t)}$ to obtain (i) and (ii). Under the assumption that the kernel L is strictly decreasing and (twice) continuously differentiable, we know that $\left\|\nabla\widehat{f}_h(\boldsymbol{x})\right\|_2$ is lower bounded away from 0 on Ω_q. Therefore, with the result in (b), the above calculation indicates that

$$
\left\|\widehat{\underline{\boldsymbol{x}}}^{(t+1)} - \widehat{\underline{\boldsymbol{x}}}^{(t)}\right\|_2 \leq \frac{\sqrt{2}\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2}{\sqrt{\left\|\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(t)})^T\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2 + \left\|\nabla\widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(t)})\right\|_2^2}} \to 0
$$

as $t \to \infty$. The result follows. □

Remark B.3. *The conditions imposed on the kernel L in Proposition 3.12 are satisfied by some commonly used kernels, such as the von Mises kernel $L(r) = e^{-r}$. However, they can be further relaxed. On the one hand, it is sufficient to assume that the kernel L is twice continuously differentiable except for finitely many points on $[0, \infty)$. On the other hand, as long as the kernel L satisfies $C_{L,q} = -\frac{\int_0^\infty L'(r) r^{\frac{q}{2}-1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} > 0$ and the true directional density f is positive almost everywhere on Ω_q, Lemma 3.6 demonstrates that $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 \to \infty$ with probability tending to 1 when $h \to 0$ and $nh^q \to \infty$. Therefore, our upper bound on $\left|\left|\underline{\widehat{\boldsymbol{x}}}^{(t+1)} - \underline{\widehat{\boldsymbol{x}}}^{(t)}\right|\right|_2$ in our proof of (c) will be asymptotically valid for all $t \geq 0$, even without the strictly decreasing property of the kernel L. Under such a relaxation, our conclusions in Proposition 3.12 are applicable to directional SCMS algorithms with other kernels that have bounded supports on $[0, \infty)$.*

Proof of Proposition 3.13. The proof is similar to our arguments in Proposition 3.7. For completeness, we still delineate the detailed steps because the proof requires some nontrivial techniques, such as parallel transports and line integrals, on general Riemannian manifolds.

(a) We first derive the following property of the objective function f supported on Ω_q, which is a counterpart of *Fact 1* in the proof of Proposition 3.7.

- *Property 1.* Given Assumption 2.1, the function f is $q\,||f||_\infty^{(2)}$-smooth on Ω_q, that is, $\texttt{grad}\, f$ is $q\,||f||_\infty^{(2)}$-Lipschitz.

This property follows easily from the differentiability of f guaranteed by Assumption 2.1 and Theorem 4.34 in Lee (2018) that

$$\begin{aligned} ||\texttt{grad}\, f(\boldsymbol{y}) - \Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}\left(\texttt{grad}\, f(\boldsymbol{x})\right)||_2 &\leq ||\mathcal{H}f(\widetilde{\boldsymbol{y}})||_2 \cdot \left|\left|\texttt{Exp}_{\boldsymbol{y}}^{-1}(\boldsymbol{x})\right|\right|_2 \\ &\leq q\,||f||_\infty^{(2)} \cdot \left|\left|\texttt{Exp}_{\boldsymbol{y}}^{-1}(\boldsymbol{x})\right|\right|_2 \end{aligned} \tag{B.35}$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega_q$, where $\widetilde{\boldsymbol{y}}$ lies on the geodesic curve $\varphi : [0,1] \to \Omega_q$ with $\varphi(0) = \boldsymbol{x}, \varphi(1) = \boldsymbol{y}$,

and $\varphi'(0) = \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$. Then,

$$
\begin{aligned}
&\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\rangle\right| \\
&\overset{\text{(i)}}{=} \left|\int_0^1 \langle \texttt{grad}\, f(\varphi(t)), \varphi'(t)\rangle dt - \langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\rangle\right| \\
&\overset{\text{(ii)}}{=} \left|\int_0^1 \left\langle \Gamma_{\varphi(t)}^{\boldsymbol{x}} \left(\texttt{grad}\, f(\varphi(t))\right), \Gamma_{\varphi(t)}^{\boldsymbol{x}} \left(\varphi'(t)\right)\right\rangle dt - \langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\rangle\right| \\
&\overset{\text{(iii)}}{=} \left|\int_0^1 \left\langle \Gamma_{\varphi(t)}^{\boldsymbol{x}} \left(\texttt{grad}\, f(\varphi(t))\right) - \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right\rangle\right| \\
&\leq \int_0^1 \left|\left|\Gamma_{\varphi(t)}^{\boldsymbol{x}} \left(\texttt{grad}\, f(\varphi(t))\right) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_2 \cdot \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 dt \\
&\overset{\text{(iv)}}{\leq} q\, ||f||_\infty^{(2)} \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 \int_0^1 d_g(\varphi(t), \boldsymbol{x})\, dt \\
&\leq q\, ||f||_\infty^{(2)} \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 \int_0^1 \left|\left|t \cdot \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 dt \\
&= \frac{q\, ||f||_\infty^{(2)}}{2} \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2^2,
\end{aligned}
\tag{B.36}
$$

where the equality (i) follows from the fundamental theorem for line integrals (Theorem 11.39 in Lee 2012), equality (ii) utilizes the isometric property of parallel transports, and inequality (iv) follows from (B.35). Moreover, since the velocity of the geodesic φ is always constant, we deduce that $\Gamma_{\varphi(t)}^{\boldsymbol{x}} \left(\varphi'(t)\right) = \varphi'(0) = \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$ and the equality (iii) follows. We will make use of the following direction of the inequality (B.36):

$$
f(\boldsymbol{y}) - f(\boldsymbol{x}) - \left\langle \texttt{grad}\, f(\boldsymbol{x}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right\rangle \geq -\frac{q\, ||f||_\infty^{(2)}}{2} \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2^2. \tag{B.37}
$$

Moreover, when $0 < \underline{\eta} < \frac{2}{q||f||_\infty^{(2)}}$,

$$
\begin{aligned}
&f(\underline{\boldsymbol{x}}^{(t+1)}) - f(\underline{\boldsymbol{x}}^{(t)}) \\
&= f\left(\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}} \left(\underline{\eta} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)\right) - f(\underline{\boldsymbol{x}}^{(t)}) \\
&\geq \left\langle \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}),\, \underline{\eta}\, \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right\rangle \\
&\quad - \frac{q\, ||f||_\infty^{(2)}}{2} \cdot \underline{\eta}^2 \left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2 \\
&= \underline{\eta} \left(1 - \frac{q\, ||f||_\infty^{(2)}\, \underline{\eta}}{2}\right) \left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2 \geq 0,
\end{aligned}
$$

showing that the objective function f is non-decreasing along the SCGA path $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ on Ω_q. Given the compactness of Ω_q and the differentiability of f, we know that the sequence $\left\{f(\underline{\boldsymbol{x}}^{(t)})\right\}_{t=0}^{\infty}$ is bounded. Thus, it converges.

(b) From (a), we know that when $0 < \underline{\eta} < \frac{2}{q||f||_{\infty}^{(2)}}$,

$$\begin{aligned} f(\underline{\boldsymbol{x}}^{(t+1)}) - f(\underline{\boldsymbol{x}}^{(t)}) &\geq \underline{\eta}\left(1 - \frac{q\,||f||_{\infty}^{(2)}\,\underline{\eta}}{2}\right)\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2 \\ &= \left(\frac{2 - q\,||f||_{\infty}^{(2)}\,\underline{\eta}}{2\underline{\eta}}\right) d_g\left(\underline{\boldsymbol{x}}^{(t+1)}, \underline{\boldsymbol{x}}^{(t)}\right)^2. \end{aligned}$$

Since the sequence $\left\{f(\underline{\boldsymbol{x}}^{(t)})\right\}_{t=0}^{\infty}$ converges, it follows that

$$\lim_{t\to\infty}\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2 = 0 \quad \text{and} \quad \lim_{t\to\infty} d_g\left(\underline{\boldsymbol{x}}^{(t+1)}, \underline{\boldsymbol{x}}^{(t)}\right) = 0.$$

Recall from (2.6) that $\left|\left|\underline{\boldsymbol{x}}^{(t)} - \underline{\boldsymbol{x}}^{(t+1)}\right|\right|_2 = 2\sin\left(\frac{d_g\left(\underline{\boldsymbol{x}}^{(t+1)}, \underline{\boldsymbol{x}}^{(t)}\right)}{2}\right)$, so $\lim_{t\to\infty}\left|\left|\underline{\boldsymbol{x}}^{(t)} - \underline{\boldsymbol{x}}^{(t+1)}\right|\right|_2 = 0$ as well.

(c) Given Assumption 2.2 and the fact that $\underline{r}_1 < \underline{\rho}/2$, we know that

$$\boldsymbol{x} \in (\underline{R}_d \oplus \underline{r}_1) \cap \Omega_q \text{ and } \left|\left|\underline{V}_d(\boldsymbol{x})^T \mathtt{grad}\, f(\boldsymbol{x})\right|\right|_2 = 0 \quad \text{if and only if} \quad \boldsymbol{x} \in \underline{R}_d.$$

Let $\underline{\boldsymbol{z}}^{(t)} = \pi_{\underline{R}_d}(\underline{\boldsymbol{x}}^{(t)})$ be the projection of $\underline{\boldsymbol{x}}^{(t)} \in \Omega_q$ in the SCGA sequence onto the directional ridge $\underline{R}_d$. Since $\underline{r}_1 \leq \mathtt{reach}(\underline{R}_d)$ by (h) of Lemma A.2, $\underline{\boldsymbol{z}}^{(t)}$ is well-defined when $\underline{\boldsymbol{x}}^{(t)} \in (\underline{R}_d \oplus \underline{r}_1) \cap \Omega_q$. Recall from (A.8) that the column space of

$$\begin{aligned} &\left(\boldsymbol{I}_{q+1} - \boldsymbol{z}^{(t)}\left(\boldsymbol{z}^{(t)}\right)^T\right)\underline{M}(\underline{\boldsymbol{z}}^{(t)}) \\ &= \left(\boldsymbol{I}_{q+1} - \boldsymbol{z}^{(t)}\left(\boldsymbol{z}^{(t)}\right)\right)^T \nabla\left[\underline{V}_d(\underline{\boldsymbol{z}}^{(t)})^T \nabla f(\underline{\boldsymbol{z}}^{(t)})\right]^T \in \mathbb{R}^{(q+1)\times(q-d)} \end{aligned}$$

coincides with the normal space of $\underline{R}_d$ within the tangent space $T_{\underline{\boldsymbol{z}}^{(t)}}$. We define a geodesic $\varphi : [0,1] \to \Omega_q$ with $\varphi(0) = \underline{\boldsymbol{z}}^{(t)}, \varphi(1) = \underline{\boldsymbol{x}}^{(t)}, \varphi'(0) = \mathtt{Exp}^{-1}_{\underline{\boldsymbol{z}}^{(t)}}(\underline{\boldsymbol{x}}^{(t)})$ and calculate that

$$\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \nabla f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2$$

$$
\begin{aligned}
&= \left|\left| \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \nabla f(\underline{\boldsymbol{x}}^{(t)}) - \underbrace{\underline{V}_d(\underline{\boldsymbol{z}}^{(t)})^T \nabla f(\underline{\boldsymbol{z}}^{(t)})}_{=0} \right|\right|_2 \\
&= \left|\left| \int_0^1 \left\langle \nabla \left[\underline{V}_d(\underline{\boldsymbol{z}}^{(t)})^T \nabla f(\underline{\boldsymbol{z}}^{(t)}) \right]^T, \varphi'(\epsilon) \right\rangle d\epsilon \right|\right|_2 \quad \text{by Theorem 11.39 in Lee (2012)} \\
&\overset{\text{(i)}}{=} \left|\left| \langle \underline{M}(\varphi(0)), \varphi'(0) \rangle + \int_0^1 \left[\langle \underline{M}(\varphi(\epsilon)), \varphi'(\epsilon) \rangle - \left\langle \Gamma_{\varphi(0)}^{\varphi(\epsilon)} \left(\underline{M}(\varphi(0)) \right), \Gamma_{\varphi(0)}^{\varphi(\epsilon)} \left(\varphi'(0) \right) \right\rangle \right] d\epsilon \right|\right|_2 \\
&\overset{\text{(ii)}}{\geq} \left|\left| \underline{M}(\underline{\boldsymbol{z}}^{(t)})^T \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 - \left|\left| \int_0^1 \left\langle \underline{M}(\varphi(\epsilon)) - \Gamma_{\varphi(0)}^{\varphi(\epsilon)} \left(\underline{M}(\varphi(0)) \right), \varphi'(\epsilon) \right\rangle d\epsilon \right|\right| \\
&\geq \left|\left| \underline{M}(\underline{\boldsymbol{z}}^{(t)})^T \left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \cdot \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 \\
&\quad - \int_0^1 \left|\left| \underline{M}(\varphi(\epsilon)) - \Gamma_{\varphi(0)}^{\varphi(\epsilon)} \left(\underline{M}(\varphi(0)) \right) \right|\right|_2 \cdot ||\varphi'(\epsilon)||_2 \, d\epsilon \\
&\geq \left|\left| \left[\left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \underline{M}(\underline{\boldsymbol{z}}^{(t)}) \right]^T \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 \\
&\quad - \left|\left| \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 \int_0^1 \sup_{\epsilon \in [0,1]} \left|\left| \bar{\nabla} \underline{M} \left(\varphi(\epsilon) \right) \right|\right|_2 \cdot \epsilon \cdot \left|\left| \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 d\epsilon \\
&\quad \text{by Theorem 4.34 in Lee (2018)} \\
&\overset{\text{(iii)}}{\geq} \underline{\beta}_1 \cdot d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right) - \frac{\underline{B}_4(f)}{2} \cdot d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right)^2 \\
&= d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right) \left(\underline{\beta}_1 - \frac{\underline{B}_4(f)}{2} \cdot d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right) \right) \\
&\overset{\text{(iv)}}{\geq} \frac{\underline{\beta}_1}{2} \cdot d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right)
\end{aligned}
$$

where we utilize the isometric properties of parallel transports in (i), note that the velocity of the geodesic is constant, *i.e.*, $\varphi'(0) = \varphi'(\epsilon)$ for any $\epsilon \in [0,1]$ to obtain (ii), leverage (d) of Lemma A.2 to deduce (iii), and use the fact that $d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right) \leq \frac{\underline{\beta}_1}{\underline{B}_4(f)}$ when $\underline{\boldsymbol{x}}^{(t)} \in (\underline{R}_d \oplus \underline{r}_1) \cap \Omega_q$ in the inequality (iv). In particular, for inequality (iii), $\left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \underline{M}(\underline{\boldsymbol{z}}^{(t)})$ is a full column rank matrix and $\mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)})$ lies within the column space of $\left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \underline{M}(\underline{\boldsymbol{z}}^{(t)})$. Since the nonzero singular values of $\left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \underline{M}(\underline{\boldsymbol{z}}^{(t)})$ are lower bounded by $\underline{\beta}_1 > 0$, it follows that

$$
\left|\left| \left[\left(\boldsymbol{I}_{q+1} - \underline{\boldsymbol{z}}^{(t)} \left(\underline{\boldsymbol{z}}^{(t)} \right)^T \right) \underline{M}(\underline{\boldsymbol{z}}^{(t)}) \right]^T \mathtt{Exp}_{\underline{\boldsymbol{z}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2 \geq \underline{\beta}_1 d_g \left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)} \right).
$$

In addition, we know that $\underline{B}_4(f) > 0$ comes from the supremum norm of $\bar{\nabla}\underline{M}(\boldsymbol{x})$ over the geodesic connecting $\underline{\boldsymbol{x}}^{(t)}$ and $\underline{\boldsymbol{z}}^{(t)}$ with $\bar{\nabla}$ being the Riemannian connection, which in turn depends on the uniform functional norm $||f||^*_{\infty,4}$ of the partial derivatives of f up to the fourth order. By (b), we deduce that

$$\lim_{t\to\infty} d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{z}}^{(t)}\right) = d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{R}_d\right) = 0.$$

The results follow. □

With Lemma B.7, we state a straightforward corollary indicating an important relation between two consecutive points in the SCGA sequence $\left\{\underline{\boldsymbol{x}}^{(t)}\right\}_{t=0}^{\infty}$ on Ω_q defined by (3.31):

$$\underline{\boldsymbol{x}}^{(t+1)} = \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\underline{\eta}\cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right). \tag{B.38}$$

Corollary B.14. *For any point $\boldsymbol{y}, \underline{\boldsymbol{x}}^{(t)}$ in a geodesically convex set on Ω_q, the update in (B.38) satisfies*

$$\begin{aligned}
&2\underline{\eta}\left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T\nabla f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\boldsymbol{y})\right\rangle \\
&\le d_g\left(\underline{\boldsymbol{x}}^{(t)}, \boldsymbol{y}\right)^2 - d_g\left(\underline{\boldsymbol{x}}^{(t+1)}, \boldsymbol{y}\right)^2 + \zeta\left(1, d_g(\underline{\boldsymbol{x}}^{(t)}, \boldsymbol{y})\right)\cdot \underline{\eta}^2 \left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2,
\end{aligned}$$

where $d_g(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\langle \mathtt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y}), \mathtt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y})\rangle} = ||\mathtt{Exp}^{-1}_{\boldsymbol{x}}(\boldsymbol{y})||_2$ is the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ on Ω_q.

Proof of Corollary B.14. Recall that the (population) SCGA iterative formula on Ω_q is given by $\underline{\boldsymbol{x}}^{(t+1)} = \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\underline{\eta}\cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)$. Note that for the geodesic triangle $\triangle \underline{\boldsymbol{x}}^{(t)}\underline{\boldsymbol{x}}^{(t+1)}\boldsymbol{y}$ with $\boldsymbol{y}\in\Omega_q$, we have that

$$d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^{(t+1)}) = \underline{\eta}\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2 = \underline{\eta}\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2$$

and

$$\begin{aligned}
&d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^{(t+1)})\cdot d_g(\underline{\boldsymbol{x}}^{(t)}, \boldsymbol{y})\cdot \cos\left(\angle \underline{\boldsymbol{x}}^{(t+1)}\underline{\boldsymbol{x}}^{(t)}\boldsymbol{y}\right) \\
&= \underline{\eta}\left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\boldsymbol{y})\right\rangle.
\end{aligned}$$

By letting $a = \overline{\underline{\boldsymbol{x}}^{(t+1)}\boldsymbol{y}}, b = \overline{\underline{\boldsymbol{x}}^{(t+1)}\underline{\boldsymbol{x}}^{(t)}}, c = \overline{\underline{\boldsymbol{x}}^{(t)}\boldsymbol{y}}$, and $A = \angle\underline{\boldsymbol{x}}^{(t+1)}\underline{\boldsymbol{x}}^{(t)}\boldsymbol{y}$ in Lemma B.7, we obtain that

$$d_g(\underline{\boldsymbol{x}}^{(t+1)}, \boldsymbol{y})^2 \leq \zeta\left(1, d_g(\underline{\boldsymbol{x}}^{(t)}, \boldsymbol{y})\right) \cdot \underline{\eta}^2 \left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2^2 + d_g(\underline{\boldsymbol{x}}^{(t)}, \boldsymbol{y})^2 - 2\underline{\eta}\left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \texttt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\boldsymbol{y})\right\rangle.$$

Some rearrangements will yield the final display. □

Note that $\left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$ in Assumptions 2.2 and 2.3 is a geodesically convex set, where the minimal geodesic between two points in the set $\left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$ always lies within the set. Hence, Corollary B.14 is applicable to our interested SCGA algorithm initialized within $\left(\underline{R}_d \oplus \underline{\rho}\right) \cap \Omega_q$.

Proof of Theorem 3.15. The proof is similar to our argument in Theorem 3.10, except that the objective function f is supported on a nonlinear manifold Ω_q here. The key arguments are credited to Corollary B.14. We first recall the following two properties.

- *Property 1.* Given Assumption 2.1, the function f is $q\,||f||_\infty^{(2)}$-smooth on Ω_q, that is, $\texttt{grad}\, f$ is $q\,||f||_\infty^{(2)}$-Lipschitz.
- *Property 2.* Given Assumptions 2.1–2.3, we know that $\left|\left|\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right|\right|_2 > 0$ for any $\underline{\boldsymbol{x}}^{(t)} \in \left(\text{Ball}_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q\right) \setminus \underline{R}_d$ and

$$f(\underline{\boldsymbol{x}}^*) - f\left(\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\frac{1}{q\,||f||_\infty^{(2)}} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)\right) \geq 0$$

for any $\underline{\boldsymbol{x}}^{(t)} \in \text{Ball}_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$ with $\underline{\boldsymbol{x}}^* \in \underline{R}_d$.

Property 1 has been established in the proof of Proposition 3.13, indicating that the objective function sequence $\left\{f(\underline{\boldsymbol{x}}^{(t)})\right\}_{t=0}^\infty$ is non-decreasing when $0 < \underline{\eta} < \frac{2}{q||f||_\infty^{(2)}}$. *Property 2* is a natural corollary by Proposition 3.13, because $\underline{\boldsymbol{x}}^{(t)} \in \text{Ball}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$ and

$$f\left(\texttt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\frac{1}{q\,||f||_\infty^{(2)}} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)\right)$$

is the objective function value after one-step SCGA iteration on Ω_q with step size $\frac{1}{q||f||_\infty^{(2)}}$. The iteration will move $\underline{\boldsymbol{x}}^{(t)}$ closer to the directional ridge $\underline{R}_d$. With the help of these two

properties, we start the proofs of (a-d).

(a) We first prove the following claim using Lemma B.2: for all $t \geq 0$ and $\underline{\boldsymbol{x}}^{(0)} \in \mathrm{Ball}_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$,

$$f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)}) \leq \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle - \frac{\underline{\beta}_0}{4} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2 + \underline{\epsilon}_t, \tag{B.39}$$

where $\underline{\epsilon}_t = \left[2q\, ||f||^{(2)}_\infty\, \underline{\beta}_2^2 \arcsin\left(\underline{\rho}/2\right) + \frac{q^{\frac{3}{2}} ||f||^{(3)}_\infty}{6} \right] = o\left(d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2\right)$. By the differentiability of f ensured by Assumption 2.1 and Taylor's theorem on Ω_q, we deduce that

$$\begin{aligned}
&f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)}) \\
&\leq \left\langle \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle + \frac{1}{2} \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*)^T \left[\mathcal{H} f(\underline{\boldsymbol{x}}^{(t)})\right] \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \\
&\quad + \frac{q^{\frac{3}{2}} ||f||^{(3)}_\infty}{6} \cdot \left|\left| \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right|\right|_2^3 \\
&\overset{\text{(i)}}{=} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle + \left\langle \underline{U}_d^\perp(\underline{\boldsymbol{x}}^{(t)}) \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle \\
&\quad + \frac{1}{2} \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*)^T \left(\underline{V}_\diamond(\underline{\boldsymbol{x}}^{(t)}), \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\right) \begin{pmatrix} 0 & & \\ & \underline{\lambda}_1(\underline{\boldsymbol{x}}^{(t)}) & \\ & & \ddots \\ & & & \underline{\lambda}_q(\underline{\boldsymbol{x}}^{(t)}) \end{pmatrix} \begin{pmatrix} \underline{V}_\diamond(\underline{\boldsymbol{x}}^{(t)}) \\ \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \end{pmatrix} \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \\
&\quad + \frac{q^{\frac{3}{2}} ||f||^{(3)}_\infty}{6} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \\
&\overset{\text{(ii)}}{\leq} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle + \frac{\underline{\beta}_0}{4} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2 \\
&\quad + \frac{\max\left\{0, \underline{\lambda}_1(\underline{\boldsymbol{x}}^{(t)})\right\}}{2} \left|\left| \underline{U}_d^\perp(\underline{\boldsymbol{x}}^{(t)}) \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right|\right|_2^2 - \frac{\underline{\beta}_0}{2} \left|\left| \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right|\right|_2^2 \\
&\quad + \frac{q^{\frac{3}{2}} ||f||^{(3)}_\infty}{6} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \\
&\overset{\text{(iii)}}{\leq} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})\underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle + \frac{\underline{\beta}_0}{4} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2 \\
&\quad + \frac{\left(\underline{\beta}_0 + \max\left\{0, \underline{\lambda}_1(\underline{\boldsymbol{x}}^{(t)})\right\}\right)}{2} \left|\left| \underline{U}_d^\perp(\underline{\boldsymbol{x}}^{(t)}) \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right|\right|_2^2 - \frac{\underline{\beta}_0}{2} \left|\left| \mathtt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right|\right|_2^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{q^{\frac{3}{2}} ||f||_{\infty}^{(3)}}{6} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \\
&\overset{\text{(iv)}}{\leq} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \right\rangle - \frac{\underline{\beta}_0}{4} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2 \\
&\quad + \frac{\left(\underline{\beta}_0 + \max\left\{0, \underline{\lambda}_1(\underline{\boldsymbol{x}}^{(t)})\right\}\right)}{2} \cdot \underline{\beta}_2^2 \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^4 + \frac{q^{\frac{3}{2}} ||f||_{\infty}^{(3)}}{6} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \\
&\overset{\text{(v)}}{\leq} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \right\rangle - \frac{\underline{\beta}_0}{4} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2 \\
&\quad + \left[2q\, ||f||_{\infty}^{(2)}\, \underline{\beta}_2^2 \arcsin\left(\underline{\rho}/2\right) + \frac{q^{\frac{3}{2}} ||f||_{\infty}^{(3)}}{6} \right] d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3,
\end{aligned}
$$

where we leverage the equality $\boldsymbol{I}_{q+1} = \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T + \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)})$ in (i) and (iii), use Assumptions 2.2 and A4 to obtain $\underline{\lambda}_q(\underline{\boldsymbol{x}}^{(t)}) \leq \cdots \leq \underline{\lambda}_{d+1}(\underline{\boldsymbol{x}}^{(t)}) < -\underline{\beta}_0$ and

$$
\left\langle \underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*) \right\rangle \leq \frac{\underline{\beta}_0}{4} \cdot d_g\left(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*\right)^2
$$

in (ii), apply the quadratic bound for $\left|\left|\underline{U}_d^{\perp}(\underline{\boldsymbol{x}}^{(t)}) \mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}^{-1}(\underline{\boldsymbol{x}}^*)\right|\right|_2$ in Assumption A4 to obtain (iv), and leverage the facts that $\max\left\{\underline{\beta}_0, 0, \underline{\lambda}_1(\underline{\boldsymbol{x}}^{(t)})\right\} \leq q\, ||f||_{\infty}^{(2)}$ and $d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)}) \leq 2 \arcsin\left(\underline{\rho}/2\right)$ when $\left|\left|\underline{\boldsymbol{x}}^{(t)} - \underline{\boldsymbol{x}}^*\right|\right|_2 \leq \underline{\rho}$ in (v); recall (2.6). Our claim (B.39) is thus proved.

In addition, given *Property 2* and any $\underline{\boldsymbol{x}}^{(t)} \in \underline{R}_d \oplus \underline{r}_2$, we derive that

$$
\begin{aligned}
&f(\underline{\boldsymbol{x}}^{(t)}) - f(\underline{\boldsymbol{x}}^*) \\
&\leq f(\underline{\boldsymbol{x}}^{(t)}) - f(\underline{\boldsymbol{x}}^*) + f(\underline{\boldsymbol{x}}^*) - f\left(\mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\frac{1}{q\, ||f||_{\infty}^{(2)}} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)\right) \\
&= -\left[f\left(\mathtt{Exp}_{\underline{\boldsymbol{x}}^{(t)}}\left(\frac{1}{q\, ||f||_{\infty}^{(2)}} \cdot \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)})\right)\right) - f(\underline{\boldsymbol{x}}^{(t)}) \right] \\
&\leq -\left[\left\langle \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \frac{1}{q\, ||f||_{\infty}^{(2)}} \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}) \right\rangle \right. \\
&\qquad \left. - \frac{q\, ||f||_{\infty}^{(2)}}{2} \cdot \left|\left| \frac{1}{q\, ||f||_{\infty}^{(2)}} \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2^2 \right] \\
&= -\frac{1}{2q\, ||f||_{\infty}^{(2)}} \left|\left| \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2^2,
\end{aligned}
$$

where we apply (B.36) to obtain the inequality. This indicates that

$$
\left|\left| \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \mathtt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2^2 \leq 2q\, ||f||_{\infty}^{(2)} \left[f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)}) \right] \tag{B.40}
$$

for any $\underline{\boldsymbol{x}}^{(t)} \in \underline{R}_d \oplus \underline{r}_3$. Therefore, by Corollary B.14, we obtain that

$$
\begin{aligned}
& d_g(\underline{\boldsymbol{x}}^{(t+1)}, \underline{\boldsymbol{x}}^*) \\
& \overset{\text{(i)}}{\leq} d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*) - 2\underline{\eta} \left\langle \underline{V}_d(\underline{\boldsymbol{x}}^{(t)}) \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}), \texttt{Exp}^{-1}_{\underline{\boldsymbol{x}}^{(t)}}(\underline{\boldsymbol{x}}^*) \right\rangle \\
& \quad + \zeta\left(1, \underline{\rho}\right) \cdot \underline{\eta}^2 \left|\left| \underline{V}_d(\underline{\boldsymbol{x}}^{(t)})^T \texttt{grad}\, f(\underline{\boldsymbol{x}}^{(t)}) \right|\right|_2^2 \\
& \overset{\text{(ii)}}{\leq} d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*)^2 + 2\underline{\eta} \Bigg[f(\underline{\boldsymbol{x}}^{(t)}) - f(\underline{\boldsymbol{x}}^*) - \frac{\underline{\beta}_0}{4} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2 \\
& \qquad\qquad + \left(2q \left|\left| f \right|\right|_\infty^{(2)} \underline{\beta}_2^2 \arcsin\left(\underline{\rho}/2\right) + \frac{q^{\frac{3}{2}} \left|\left| f \right|\right|_\infty^{(3)}}{6} \right) d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \Bigg] \\
& \quad + \zeta\left(1, \underline{\rho}\right) \cdot \underline{\eta}^2 \cdot 2q \left|\left| f \right|\right|_\infty^{(2)} \left[f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)}) \right] \\
& \overset{\text{(iii)}}{\leq} \left(1 - \frac{\underline{\beta}_0 \underline{\eta}}{4} \right) \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2 - 2\underline{\eta} \left[1 - \underline{\eta} \cdot \zeta\left(1, \underline{\rho}\right) \cdot q \left|\left| f \right|\right|_\infty^{(2)} \right] \cdot \underbrace{\left[f(\underline{\boldsymbol{x}}^*) - f(\underline{\boldsymbol{x}}^{(t)}) \right]}_{\geq 0} \\
& \leq \left(1 - \frac{\underline{\beta}_0 \underline{\eta}}{4} \right) \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2
\end{aligned}
$$

whenever $0 < \underline{\eta} \leq \min\left\{ \frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta\left(1, \underline{\rho}\right)} \right\}$, where we utilize Corollary B.14 and the monotonicity of $\zeta(1, c)$ with respect to c in (i), apply (B.39) and (B.40) to obtain (ii), and use the choice of $\underline{r}_2$ to argue that

$$
\begin{aligned}
& \left(2q \left|\left| f \right|\right|_\infty^{(2)} \underline{\beta}_2^2 \arcsin\left(\underline{\rho}/2\right) + \frac{q^{\frac{3}{2}} \left|\left| f \right|\right|_\infty^{(3)}}{6} \right) d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^3 \\
& \leq \left(2q \left|\left| f \right|\right|_\infty^{(2)} \underline{\beta}_2^2 \arcsin\left(\underline{\rho}/2\right) + \frac{q^{\frac{3}{2}} \left|\left| f \right|\right|_\infty^{(3)}}{6} \right) d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2 \cdot 2\arcsin\left(\underline{r}_2/2\right) \\
& \leq \frac{\underline{\beta}_0}{8} \cdot d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)})^2
\end{aligned}
$$

in (iii). By telescoping, we conclude that when $0 < \underline{\eta} \leq \min\left\{ \frac{4}{\underline{\beta}_0}, \frac{1}{q||f||_\infty^{(2)} \cdot \zeta\left(1, \underline{\rho}\right)} \right\}$ and $\underline{\boldsymbol{x}}^{(0)} \in \underline{R}_d \oplus \underline{r}_2$,

$$
d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(t)}) \leq \left(1 - \frac{\underline{\beta}_0 \underline{\eta}}{4} \right)^{\frac{t}{2}} d_g(\underline{\boldsymbol{x}}^*, \underline{\boldsymbol{x}}^{(0)}).
$$

The result follows.

(b) The result follows directly from (a) and the fact that $d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{R}_d) \leq d_g(\underline{\boldsymbol{x}}^{(t)}, \underline{\boldsymbol{x}}^*)$ for all $t \geq 0$.

(c) The proof is logically similar to the proof of (c) in Theorem 3.10. We write the spectral decompositions of $\mathcal{H}f(\boldsymbol{x})$ and $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ as:

$$\mathcal{H}f(\boldsymbol{x}) = \underline{V}(\boldsymbol{x})\underline{\Lambda}(\boldsymbol{x})\underline{V}(\boldsymbol{x})^T \quad \text{and} \quad \mathcal{H}\widehat{f}_h(\boldsymbol{x}) = \underline{\widehat{V}}(\boldsymbol{x})\underline{\widehat{\Lambda}}(\boldsymbol{x})\underline{\widehat{V}}(\boldsymbol{x})^T.$$

By Weyl's theorem (Theorem 4.3.1 in Horn and Johnson 2012) and uniform bounds (2.29),

$$\begin{aligned}
|\underline{\lambda}_j(\boldsymbol{x}) - \underline{\widehat{\lambda}}_j(\boldsymbol{x})| &\leq \left|\left|\mathcal{H}f(\boldsymbol{x}) - \mathcal{H}\widehat{f}_h(\boldsymbol{x})\right|\right|_2 \\
&\leq q \left|\left|f(\boldsymbol{x}) - \widehat{f}_h(\boldsymbol{x})\right|\right|_\infty^{(2)} \\
&= O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right).
\end{aligned}$$

Thus, $\widehat{f}_h$ will satisfy Assumption 2.2 with high probability when h is sufficiently small and $\frac{nh^{q+4}}{|\log h|}$ is sufficiently large. According to the Davis-Kahan theorem (Lemma B.9 here), uniform bounds (2.29), and the continuity of the exponential maps, we have that

$$\begin{aligned}
&d_g\left(\texttt{Exp}_{\boldsymbol{y}}\left(\underline{\eta} \cdot \underline{\widehat{V}}_d(\boldsymbol{y})\underline{\widehat{V}}_d(\boldsymbol{y})^T \texttt{grad}\, \widehat{f}_h(\boldsymbol{y})\right), \texttt{Exp}_{\boldsymbol{y}}\left(\underline{\eta} \cdot \underline{V}_d(\boldsymbol{y})\underline{V}_d(\boldsymbol{y})^T \texttt{grad}\, f(\boldsymbol{y})\right)\right) \\
&\leq \underline{\eta} C_3 \left|\left|\underline{\widehat{V}}_d(\boldsymbol{y})\underline{\widehat{V}}_d(\boldsymbol{y})^T \texttt{grad}\, \widehat{f}_h(\boldsymbol{y}) - \underline{V}_d(\boldsymbol{y})\underline{V}_d(\boldsymbol{y})^T \texttt{grad}\, f(\boldsymbol{y})\right|\right|_2 \\
&\leq \underline{\eta} C_3 \left|\left|\underline{\widehat{V}}_d(\boldsymbol{y})\underline{\widehat{V}}_d(\boldsymbol{y})^T \left[\texttt{grad}\, \widehat{f}_h(\boldsymbol{y}) - \texttt{grad}\, f(\boldsymbol{y})\right]\right|\right|_2 \\
&\quad + \underline{\eta} C_3 \left|\left|\left[\underline{\widehat{V}}_d(\boldsymbol{y})\underline{\widehat{V}}_d(\boldsymbol{y})^T - \underline{V}_d(\boldsymbol{y})\underline{V}_d(\boldsymbol{y})^T\right] \texttt{grad}\, f(\boldsymbol{y})\right|\right|_2 \\
&\overset{\text{(i)}}{\leq} \underline{\eta} C_3 \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{y}) - \texttt{grad}\, f(\boldsymbol{y})\right|\right|_2 + \underline{\eta} C_3 \cdot \frac{\left|\left|\mathcal{H}f(\boldsymbol{y}) - \mathcal{H}\widehat{f}_h(\boldsymbol{y})\right|\right|_2 \cdot ||\texttt{grad}\, f(\boldsymbol{y})||_2}{\underline{\beta}_0} \\
&\leq \underline{\eta} C_3 \sqrt{q} \left|\left|\widehat{f}_h - f\right|\right|_\infty^{(1)} + \underline{\eta} C_3 \cdot \frac{q \left|\left|\widehat{f}_h - f\right|\right|_\infty^{(2)} \sqrt{q+1}\, ||f||_\infty^{(1)}}{\underline{\beta}_0} \\
&\equiv \underline{\epsilon}_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right)
\end{aligned}$$

for any $\boldsymbol{y} \in \mathrm{Ball}_{q+1}(\underline{\boldsymbol{x}}^*, \underline{r}_2) \cap \Omega_q$, where we utilize the Davis-Kahan theorem and $\left|\left|\widehat{\underline{V}}_d(\boldsymbol{y})\widehat{\underline{V}}_d(\boldsymbol{y})^T\right|\right|_2 = 1$ in (i). Hence, when $h \to 0$ and $\frac{nh^{q+4}}{|\log h|} \to \infty$,

$$\begin{aligned} & d_g\left(\mathtt{Exp}_{\boldsymbol{y}}\left(\underline{\eta} \cdot \widehat{\underline{V}}_d(\boldsymbol{y})\widehat{\underline{V}}_d(\boldsymbol{y})^T \mathtt{grad}\, \widehat{f}_h(\boldsymbol{y})\right), \mathtt{Exp}_{\boldsymbol{y}}\left(\underline{\eta} \cdot \underline{V}_d(\boldsymbol{y})\underline{V}_d(\boldsymbol{y})^T \mathtt{grad}\, f(\boldsymbol{y})\right)\right) \\ & \leq \underline{\epsilon}_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right) \leq (1-\Upsilon) \cdot 2\arcsin\left(\underline{r}_2/2\right) \end{aligned} \tag{B.41}$$

with probability tending to 1.

We now claim that $d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*) \leq 2\arcsin\left(\underline{r}_2/2\right)$ and

$$d_g(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}, \underline{\boldsymbol{x}}^*) \leq \Upsilon \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*) + \underline{\epsilon}_{n,h} \tag{B.42}$$

for all $t \geq 0$. We again prove this claim by induction on the iteration number. Note that when $t = 1$, we derive that

$$\begin{aligned} & d_g(\widehat{\underline{\boldsymbol{x}}}^{(1)}, \underline{\boldsymbol{x}}^*) \\ & = d_g\left(\mathtt{Exp}_{\widehat{\underline{\boldsymbol{x}}}^{(0)}}\left(\underline{\eta} \cdot \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})^T \mathtt{grad}\, \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(0)})\right), \underline{\boldsymbol{x}}^*\right) \\ & \overset{\text{(i)}}{\leq} d_g\left(\mathtt{Exp}_{\widehat{\underline{\boldsymbol{x}}}^{(0)}}\left(\underline{\eta} \cdot \underline{V}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})\underline{V}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})^T \mathtt{grad}\, f(\widehat{\underline{\boldsymbol{x}}}^{(0)})\right), \underline{\boldsymbol{x}}^*\right) \\ & \quad + d_g\left(\mathtt{Exp}_{\widehat{\underline{\boldsymbol{x}}}^{(0)}}\left(\underline{\eta} \cdot \widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})\widehat{\underline{V}}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})^T \mathtt{grad}\, \widehat{f}_h(\widehat{\underline{\boldsymbol{x}}}^{(0)})\right), \mathtt{Exp}_{\widehat{\underline{\boldsymbol{x}}}^{(0)}}\left(\underline{\eta} \cdot \underline{V}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})\underline{V}_d(\widehat{\underline{\boldsymbol{x}}}^{(0)})^T \mathtt{grad}\, f(\widehat{\underline{\boldsymbol{x}}}^{(0)})\right)\right) \\ & \overset{\text{(ii)}}{\leq} \Upsilon \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + \underline{\epsilon}_{n,h}, \end{aligned}$$

where we apply the triangle inequality in (i) and leverage the result in (a) and (B.41) to obtain (ii). The triangle inequality is valid in this context because the geodesic measures the minimal distance between two points on Ω_q. Moreover, by the choice of $\widehat{\underline{\boldsymbol{x}}}^{(0)}$ and (B.41), we are ensured that $d_g(\widehat{\underline{\boldsymbol{x}}}^{(1)}, \underline{\boldsymbol{x}}^*) \leq 2\arcsin\left(\underline{r}_2/2\right)$. In the induction from $t \mapsto t+1$, we suppose that $d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*) \leq 2\arcsin\left(\underline{r}_2/2\right)$ and the claim (B.42) holds at iteration t. The same argument then implies that the claim (B.42) holds for iteration $t+1$ and that $d_g(\widehat{\underline{\boldsymbol{x}}}^{(t+1)}, \underline{\boldsymbol{x}}^*) \leq 2\arcsin\left(\underline{r}_2/2\right)$. The claim (B.42) is thus verified.

Now, given that $\Upsilon = \sqrt{1 - \frac{\underline{\beta}_0 \underline{\eta}}{4}} < 1$, we iterate the claim (B.42) to show that

$$d_g(\widehat{\underline{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*) \leq \Upsilon \cdot d_g(\widehat{\underline{\boldsymbol{x}}}^{(t-1)}, \underline{\boldsymbol{x}}^*) + \underline{\epsilon}_{n,h}$$

$$
\begin{aligned}
&\leq \Upsilon \left[\Upsilon \cdot d_g(\underline{\widehat{\boldsymbol{x}}}^{(t-2)}, \underline{\boldsymbol{x}}^*) + \underline{\epsilon}_{n,h} \right] + \underline{\epsilon}_{n,h} \\
&\leq \Upsilon^t \cdot d_g(\underline{\widehat{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + \left[\sum_{s=0}^{t-1} \Upsilon^s \right] \underline{\epsilon}_{n,h} \\
&\leq \Upsilon^t \cdot d_g(\underline{\widehat{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + \frac{\underline{\epsilon}_{n,h}}{1-\Upsilon} \\
&= \Upsilon^t \cdot d_g(\underline{\widehat{\boldsymbol{x}}}^{(0)}, \underline{\boldsymbol{x}}^*) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}} \right),
\end{aligned}
$$

where the fourth inequality follows by summing the geometric series, and the last equality is due to our notation $\underline{\epsilon}_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}} \right)$. It completes the proof.

(d) The result follows directly from (c) and the inequality $d_g(\underline{\widehat{\boldsymbol{x}}}^{(t)}, \underline{R}_d) \leq d_g(\underline{\widehat{\boldsymbol{x}}}^{(t)}, \underline{\boldsymbol{x}}^*)$ for all $t \geq 0$. □

Appendix C

APPENDIX TO CHAPTER 4

C.1 Tuning Parameter Selection

We describe the choices of two important tuning parameters used in our directional ridge estimation procedure in Section 4.5: the smoothing bandwidth for the directional KDE and the density threshold used for denoising. Throughout this section, the observed galaxy/QSO locations on a fixed spherical slice are represented by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \Omega_2$.

• **Smoothing Bandwidth Parameter h:** The bandwidth h is a central tuning parameter for the directional KDE in (2.4) and, consequently, for the DirSCMS filament estimates. For the catalog construction, we use the rule-of-thumb selector from Proposition 2 of García-Portugués (2013), together with the estimator of the von Mises-Fisher concentration parameter from Banerjee et al. (2005). This choice is simple, computationally stable across hundreds of spherical slices, and effective in the entire filament detection pipeline.

Specifically, the smoothing bandwidth is chosen as:

$$h_{\mathrm{ROT}} = A_0 \cdot \left[\frac{8 \sinh^2(\widehat{\kappa})}{n\widehat{\kappa} \cdot [(1 + 4\widehat{\kappa}^2)\sinh(2\widehat{\kappa}) - 2\widehat{\kappa}\cosh(2\widehat{\kappa})]} \right]^{\frac{1}{6}}, \tag{C.1}$$

where

$$\widehat{\kappa} = \frac{\bar{R}(3 - \bar{R}^2)}{1 - \bar{R}^2}, \qquad \bar{R} = \frac{||\sum_{i=1}^{n} \boldsymbol{X}_i||_2}{n}.$$

Here, n is the number of observations in the current spherical slice and A_0 is a multiplicative hyperparameter controlling the overall amount of smoothing. The exponent $\frac{1}{6}$ reflects the AMISE bandwidth rate for density estimation on Ω_2, whose intrinsic dimension is two.

The role of A_0 is to adapt the rule-of-thumb bandwidth to the filament detection task. A smaller value of A_0 produces an undersmoothed density estimate $\widehat{f}_h$ and may generate spurious filamentary features, whereas a larger value may oversmooth the density field and

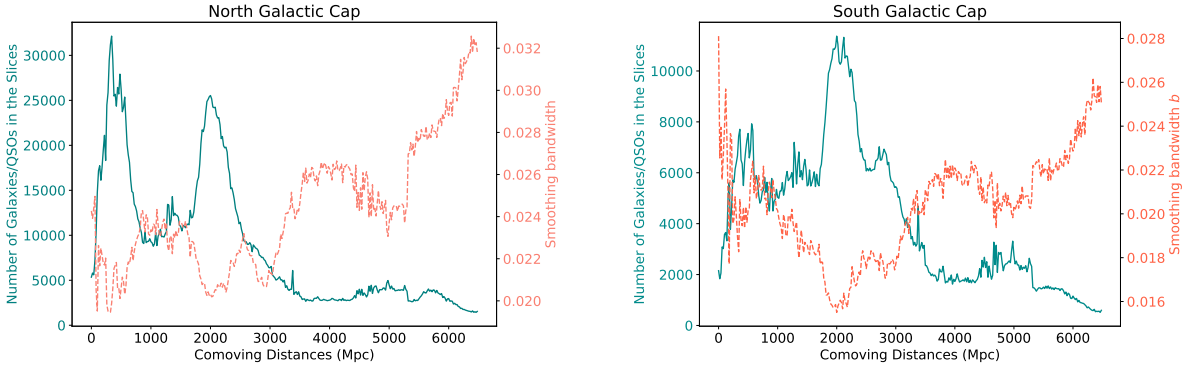


Figure C.1: Number of observations and corresponding smoothing bandwidth $\widehat{h}$, selected by (C.1), in each spherical slice across comoving distance (or redshift) on the North and South Galactic Caps.

remove genuine structures. When $A_0 = 1$ and the true density field is a von Mises-Fisher distribution on Ω_2 (Mardia and Jupp, 2000), the rule-of-thumb bandwidth minimizes the dominant term of the mean integrated squared error

$$\mathrm{MISE}(h) \equiv \mathbb{E}\left[\int_{\Omega_2} \left(\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \omega_2(d\boldsymbol{x})\right],$$

where f denotes the underlying galaxy/QSO density field on Ω_2. In astronomical survey data, the true density field would not be von Mises-Fisher distributed. Thus, we use the rule primarily as a stable baseline and set $A_0 = 0.25$ for the SDSS-IV catalog. The resulting bandwidth $\widehat{h}$ adapts to the number of observations in each slice. On the one hand, slices with more observations receive smaller bandwidths, allowing the directional KDE and DirSCMS algorithm to retain finer filamentary structure. On the other hand, sparser slices receive larger bandwidths and retain only more dominant structures. This behavior is illustrated in Figure C.1 for the North and South Galactic Caps.

- **Denoising Threshold τ:** The available data $\mathcal{D} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ in a spherical slice may contain noisy observations and outliers. Such observations can create small low-density bumps in the estimated density field on Ω_2, which in turn may lead to spurious filaments

and nodes. To suppress these artifacts, we include a denoising step after estimating the galaxy/QSO density field by (2.4). Specifically, before applying the DirSCMS algorithm, we remove observations whose estimated density values fall below a threshold τ.

Following Chen et al. (2015), we set τ using the root mean square variation of the estimated density values, with an additional quantile safeguard:

$$\tau = \min \left\{ Q_{0.2}, \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_h(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_h(\boldsymbol{X}_i)\right)^2} \right\}, \tag{C.2}$$

where Q_α denotes the lower-α quantile of the directional KDE values $\{\widehat{f}_h(\boldsymbol{X}_1), \ldots, \widehat{f}_h(\boldsymbol{X}_n)\}$ evaluated on the original sample $\mathcal{D}$. The quantile safeguard ensures that τ is no larger than the 20th percentile of the estimated density values. Consequently, the denoising step removes only observations whose estimated densities are small relative to the slice-specific density distribution, while retaining a sufficiently large sample for filament detection.

Appendix D

APPENDIX TO CHAPTER 5

D.1 Implementation of the Proposed Debiasing Method

Our proposed debiasing method consists of three main estimation steps: (i) computing the Lasso pilot estimate (5.2), (ii) estimating the propensity scores $\pi_i = \pi(X_i) = \mathrm{P}(R_i = 1|X_i), i = 1, ..., n$, and (iii) solving the debiasing programs (5.3) and (5.7). We explain the implementation details and strategies for selecting two tuning parameters $\lambda > 0$ in (5.2) and $\gamma > 0$ in (5.3) as follows.

1. Lasso pilot estimate: We use the complete-case (or equivalently, observed) data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ with $R_i = 1$ for $i = 1, ..., n$ to construct the Lasso pilot estimate $\widehat{\beta}$. To select the regularization parameter $\lambda > 0$ in a data-adaptive way, we adopt the scaled Lasso (Sun and Zhang, 2012) with its universal regularization parameter $\lambda_0 = \sqrt{\frac{2 \log d}{n}}$ as the initialization. Specifically, it provides an iterative algorithm to obtain the Lasso estimate $\widehat{\beta}$ and a consistent estimator $\widehat{\sigma}_\epsilon$ of the noise level σ_ϵ from the following jointly convex optimization problem:

$$\left(\widehat{\beta}(\widetilde{\lambda}), \widehat{\sigma}_\epsilon(\widetilde{\lambda})\right) = \underset{\beta \in \mathbb{R}^d, \sigma_\epsilon > 0}{\arg\min} \left[\frac{1}{2n\sigma_\epsilon} \sum_{i=1}^{n} R_i \left(Y_i - X_i^T \beta\right)^2 + \frac{\sigma_\epsilon}{2} + \widetilde{\lambda} ||\beta||_1 \right],$$

where the regularization parameter $\widetilde{\lambda} > 0$ is updated iteratively as well. We also utilize this estimated noise level $\widehat{\sigma}_\epsilon$ to construct our asymptotic $1 - \tau$ confidence interval (5.5).

2. Propensity score estimation: When comparing our debiasing method with other existing high-dimensional inference methods, we utilize the Lasso-type logistic regression (D.4) in Section D.3 to estimate the propensity scores $\pi_i = \mathrm{P}(R_i = 1|X_i), i = 1, ..., n$. The regularization parameter ζ is selected from 40 logarithmically spaced values within the interval $\left[0.1\sqrt{\frac{\log d}{n}}, 300\sqrt{\frac{\log d}{n}}\right]$ via 5-fold cross-validation. However, as demonstrated

in Section 5.4.4 through simulation studies, our debiasing method can perform even better when the propensity scores are estimated more accurately by other nonparametric methods, reflecting the theoretical results in Section 5.3.

3. Solving the debiasing program: When our primal debiasing program is strictly feasible, strong duality holds by Slater's condition (Theorem 3.2.6 in Renegar 2001), and the debiasing weights $\widehat{w}_i(x), i = 1, ..., n$ can be estimated from either the primal program (5.3) or the dual program (5.7) by the identity (5.8) in Proposition 5.1. We implement our debiasing method in both Python and R. In particular, we solve the primal program (5.3) using the Python package "`CVXPY`" (Diamond and Boyd, 2016; Agrawal et al., 2018) or the R package "`CVXR`" (Fu et al., 2020). For the dual program (5.7), we formulate and implement the classical coordinate descent algorithm (Wright, 2015) as:

$$\left[\widehat{\ell}(x)\right]_j \leftarrow \frac{\mathcal{S}_{\frac{\gamma}{n}}\left(-\frac{1}{2n}\sum_{i=1}^n \widehat{\pi}_i \cdot \left(\sum_{k\neq j} X_{ij}X_{ik}\left[\widehat{\ell}(x)\right]_k\right) - x_j\right)}{\frac{1}{2n}\sum_{i=1}^n \widehat{\pi}_i X_{ij}^2} \quad \text{for} \quad j = 1, ..., d, \qquad \text{(D.1)}$$

where $\mathcal{S}_{\frac{\gamma}{n}}(u) = \text{sign}(u)\cdot\left(|u| - \frac{\gamma}{n}\right)_+$ is the soft-thresholding operator, and the above procedure iterates cyclically from $j = 1$ to d until convergence.

The tuning parameter $\gamma > 0$ of our debiasing program is selected from 41 equally spaced values between 0 and $n\,||x||_\infty$ via 5-fold cross-validation according to the dual problem (5.7), where $x \in \mathbb{R}^d$ is the current query point. Here, the upper bound for the candidate values of $\gamma > 0$ is set to $n\,||x||_\infty$ because the optimal solution $\widehat{w}(x)$ for the primal problem (5.3) becomes 0 when $\frac{\gamma}{n} \geq ||x||_\infty$. There are three different criteria for selecting the final value of $\gamma > 0$ from the cross-validation results.

(i) The first criterion selects the value of $\gamma > 0$ that minimizes the average cross-validated risk of the dual debiasing program (5.7), which is known as the "min-CV" rule.

(ii) The second criterion selects a value of $\gamma > 0$ that is smaller than the risk-minimizing choice and whose average cross-validated risk is at least one standard error (calculated across the five folds) away from the minimum; see Section 3.4.3 in Breiman et al. (1984)

and Chen and Yang (2021) for details. We call this criterion the "1SE" rule. Its rationale is to improve inference by sacrificing some statistical efficiency (*i.e.*, potentially increasing the estimated asymptotic variance) in exchange for less bias.

(iii) The third criterion chooses the smallest value of $\gamma > 0$ for which the primal program is feasible for all five folds of the data. We call this criterion the "min-feas" rule, which leads to the most conservative confidence interval for our debiasing method.

We recommend choosing the final value of γ between the values selected by the "min-feas" and "1SE" rules, depending on how conservative the subsequent inference needs to be; recall from Section 5.4.2 for our simulation results.

D.2 Sufficient Conditions for Assumption 5.6

We present several sufficient conditions in Lemma D.1 under which Assumption 5.6 on the sparse approximation to the population dual solution $\ell_0(x) \in \mathbb{R}^d$ in Definition 5.3 holds and illuminate these sufficient conditions through some motivating examples.

Lemma D.1 (Sufficient conditions for the sparse approximation to $\ell_0(x)$)**.** *Under Assumption 5.4, we let* $\left[\mathbb{E}(RXX^T)\right]^{-1} := M = (\boldsymbol{m}_1, ..., \boldsymbol{m}_d) \in \mathbb{R}^{d\times d}$*, where* $\boldsymbol{m}_k, k = 1, ..., d$ *are columns of the matrix* $M \in \mathbb{R}^{d\times d}$*. We also denote by* $M_{S,T} \in \mathbb{R}^{|S|\times|T|}$ *the submatrix of* M *indexed by* $S, T \subset \{1, ..., d\}$ *and let* $\sigma_{\min}(M_{S,T})$ *be its smallest singular value.*

(a) If each column of M *has at most* $m \geq 1$ *nonzero entries and the query vector* $x \in \mathbb{R}^d$ *has at most* $s_x \geq 1$ *nonzero entries, then Assumption 5.6 holds with* $s_\ell(x) = m \cdot s_x$ *and* $r_\ell = 0$ *for all* $n \geq 1$*.*

(b) Suppose that each column $\boldsymbol{m}_k$ *of* M *lies in* $\mathcal{C}_1(J_k, \vartheta)$ *for some* $J_k \subset \{1, ..., d\}$ *and* $\vartheta \in [0, \infty)$*. If the query vector* $x \in \mathbb{R}^d$ *has support* $\text{support}(x) = S_x$ *of size at most* $s_x > 0$ *and an index set* $J \subset \{1, ..., d\}$ *satisfies* $\sigma_{\min}(M_{J,S_x}) > 0$*, then Assumption 5.6*

holds with $s_\ell(x) = |J| \cdot R_\ell(n)$ *and* $r_\ell = O\left(\frac{(1+\vartheta)\cdot K_{J,x}}{\sqrt{R_\ell(n)}}\right)$ *for some* $R_\ell(n) \to \infty$ *as* $n \to \infty$, *where* $K_{J,x} = \max_{k\in S_x} \frac{\sqrt{s_x}\left|\left|M_{J_k,k}\right|\right|_1}{\sigma_{\min}(M_{J,S_x})}$.

(c) *Suppose that each column* $\boldsymbol{m}_k$ *of* M *lies in* $\mathcal{C}_1(J_k, \vartheta)$ *for some* $J_k \subset \{1, ..., d\}$ *and* $\vartheta \in [0, \infty)$. *If the query vector* $x \in \mathbb{R}^d$ *lies in* $\mathcal{C}_1(J, \vartheta_x)$ *with an index set* $J \subset \{1, ..., d\}$ *satisfying* $\kappa_{\min}(J, \vartheta_x) > 0$, *then Assumption 5.6 holds with* $s_\ell(x) = |J| \cdot R_\ell(n)$ *and* $r_\ell = O\left(\frac{(1+\vartheta)K_{J,x}}{\sqrt{R_\ell(n)}}\right)$ *for some* $R_\ell(n) \to \infty$ *as* $n \to \infty$, *where* $K_{J,x} = (1 + \vartheta_x) \max_{1\le k\le d} \frac{\sqrt{|J|}\cdot\left|\left|M_{J_k,k}\right|\right|_1}{\kappa_{\min}(J,\vartheta_x)}$ *and* $\kappa_{\min}(J, \vartheta_x) = \inf_{u\in\mathcal{C}_1(J,\vartheta_x)} \frac{\sqrt{|J|}\cdot||M_J u||_2}{||u_J||_1}$ *is the compatibility constant of the submatrix* M_J.

Proof of Lemma D.1. The proof of Lemma D.1 is inspired and modified from Lemma 4 in Giessing and Wang (2023).

(a) Let $S_x = \text{support}(x) \subset \{1, ..., d\}$ for the fixed query vector $x \in \mathbb{R}^d$. Given that $\max_{1\le k\le d} ||M_k||_0 \le m$ and $\ell_0(x) = -2\left[\mathbb{E}\left(RXX^T\right)\right]^{-1} x \in \mathbb{R}^d$, we compute that

$$||\ell_0(x)||_0 = ||Mx||_0 = \left|\left|\sum_{k=1}^d \boldsymbol{m}_k x_k\right|\right|_0 = \left|\left|\sum_{k\in S_x} \boldsymbol{m}_k x_k\right|\right|_0 \le \sum_{k\in S_x} ||\boldsymbol{m}_k||_0 \le m \cdot s_x.$$

Thus, $\ell_0(x)$ by itself is sparse and we can take $\widetilde{\ell}(x) = \ell_0(x)$ in Assumption 5.6 with $s_\ell(x) = m \cdot s_x$ and $r_\ell = 0$ for all $n \ge 1$.

(b) Since each column of M lies in $\mathcal{C}_1(J_k, \vartheta)$ for some $J_k \subset \{1, ..., d\}$ and $\vartheta \in [0, \infty)$, we know that $\left|\left|M_{J_k^c,k}\right|\right|_1 \le \vartheta\, ||M_{J_k,k}||_1$ for all $k \in S_x$. Moreover, $\sigma_{\min}(M_{J,S_x}) = \inf_{\text{support}(u)=S_x} \frac{||M_J u||_2}{||u||_2} > 0$ by our assumption. We therefore have that

$$\begin{aligned}
||[\ell_0(x)]_{J^c}||_1 &= 2\, ||[Mx]_{J^c}||_1 \\
&= 2\left|\left|\sum_{k=1}^d M_{J^c,k}\cdot x_k\right|\right|_1 \\
&\le 2\sum_{k\in S_x} ||M_{J^c,k}||_1\, |x_k|
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \sum_{k \in S_x} \left(\left|\left| M_{J^c \cap J_k^c, k} \right|\right|_1 + ||M_{J^c \cap J_k, k}||_1 \right) |x_k| \\
&\leq 2 \sum_{k \in S_x} \left(\left|\left| M_{J_k^c, k} \right|\right|_1 + ||M_{J_k, k}||_1 \right) |x_k| \\
&\leq 2(1+\vartheta) \sum_{k \in S_x} ||M_{J_k, k}||_1 \, |x_k| \\
&\leq 2(1+\vartheta)\sqrt{s_x} \cdot \max_{k \in S_x} ||M_{J_k, k}||_1 \, ||x||_2 \\
&\leq 2(1+\vartheta)\sqrt{s_x} \cdot \max_{k \in S_x} ||M_{J_k, k}||_1 \, ||(Mx)_J||_2 \sup_{\text{support}(u) = S_x} \frac{||u||_2}{||(Mu)_J||_2} \\
&\leq \frac{2(1+\vartheta)\sqrt{s_x} \max_{k \in S_x} ||M_{J_k, k}||_1}{\sigma_{\min}(M_{J, S_x})} \cdot ||(Mx)_J||_1 \\
&\equiv (1+\vartheta) K_{J,x} \, ||[\ell_0(x)]_J||_1 \,,
\end{aligned}
$$

showing that $\ell_0(x) \in \mathcal{C}_1\left(J, (1+\vartheta)K_{J,x}\right)$. By Theorem 2.5 in Foucart and Rauhut (2013), there exists a $s_\ell(x)$-sparse approximation $\widetilde{\ell}(x) \in \mathbb{R}^d$ to the population dual solution $\ell_0(x)$ such that

$$
\left|\left| \widetilde{\ell}(x) - \ell_0(x) \right|\right|_2 \leq \frac{1}{2\sqrt{s_\ell(x)}} \, ||\ell_0(x)||_1 \,.
$$

Combining with $\ell_0(x) \in \mathcal{C}_1\left(J, (1+\vartheta)K_{J,x}\right)$, we have that

$$
\begin{aligned}
\left|\left| \widetilde{\ell}(x) - \ell_0(x) \right|\right|_2 &\leq \frac{1}{2\sqrt{s_\ell(x)}} \left[||[\ell_0(x)]_J||_1 + ||[\ell_0(x)]_{J^c}||_1 \right] \\
&\leq \frac{1 + (1+\vartheta)K_{J,x}}{2\sqrt{s_\ell(x)}} \cdot ||[\ell_0(x)]_J||_1 \\
&\leq \frac{1 + (1+\vartheta)K_{J,x}}{2} \sqrt{\frac{|J|}{s_\ell(x)}} \cdot ||\ell_0(x)||_2 \,.
\end{aligned}
$$

Thus, when $s_\ell(x) = |J| \cdot R_\ell(n)$, it follows that

$$
\left|\left| \widetilde{\ell}(x) - \ell_0(x) \right|\right|_2 \leq \frac{1 + (1+\vartheta)K_{J,x}}{2\sqrt{R_\ell(n)}} \cdot ||\ell_0(x)||_2 \,,
$$

so we take $r_\ell = O\left(\frac{(1+\vartheta)K_{J,x}}{\sqrt{R_\ell(n)}} \right)$ to fulfill Assumption 5.6.

(c) Since $x \in \mathcal{C}_1(J, \vartheta_x)$, it follows that $||x_{J^c}||_1 \leq \vartheta_x ||x_J||_1$. Now, we compute that

$$\begin{aligned}
||[\ell_0(x)]_{J^c}||_1 &= 2\,||(Mx)_{J^c}||_1 \\
&= 2\left|\left|\sum_{k=1}^{d} M_{J^c,k} x_k\right|\right|_1 \\
&\leq 2 \max_{1\leq k\leq d} ||M_{J^c,k}||_1\, ||x||_1 \\
&\leq 2(1+\vartheta_x) \max_{1\leq k\leq d} ||M_{J^c,k}||_1\, ||x_J||_1 \\
&\leq 2(1+\vartheta_x) \max_{1\leq k\leq d} ||M_{J^c,k}||_1\, ||(Mx)_J||_2 \sup_{u\in\mathcal{C}_1(J,\vartheta_x)} \frac{||u_J||_1}{||(Mu)_J||_2} \\
&\leq 2(1+\vartheta)(1+\vartheta_x) \max_{1\leq k\leq d} ||M_{J_k,k}||_1\, ||(Mx)_J||_2 \sup_{u\in\mathcal{C}_1(J,\vartheta_x)} \frac{||u_J||_1}{||M_J u||_2} \\
&= 2(1+\vartheta)(1+\vartheta_x) \max_{1\leq k\leq d} \frac{\sqrt{|J|}\cdot||M_{J_k,k}||_1}{\kappa_{\min}(J,\vartheta_x)} \cdot ||(Mx)_J||_1 \\
&\equiv (1+\vartheta) K_{J,x}\, ||[\ell_0(x)]_J||_1\,,
\end{aligned}$$

showing that $\ell_0(x) \in \mathcal{C}_1\left(J, (1+\vartheta)K_{J,x}\right)$. Therefore, as in the proof of Part (b), we conclude that Assumption 5.6 holds with $s_\ell(x) = |J| \cdot R_\ell(n)$ and $r_\ell = O\left(\frac{(1+\vartheta)K_{J,x}}{\sqrt{R_\ell(n)}}\right)$. □

Remark D.1. *The constants $\sigma_{\min}(M_{J,S_x})$ and $\kappa_{\min}(J,\vartheta_x)$ in (b) and (c) of Lemma D.1 are well-known in the statistical literature. Sun and Zhang (2012, 2013) introduced the eigenvalue version of $\sigma_{\min}(M_{J,S_x}) > 0$ to study least-squares and precision matrix estimation after the scaled Lasso selection. The compatibility constant $\kappa_{\min}(J,\vartheta_x)$ appears even more commonly in the literature (see Section 6.2.2 in Bühlmann and van De Geer 2011), given that it is also related to our restricted eigenvalue condition (5.17) as:*

$$\kappa_{\min}(J,\vartheta_x) = \inf_{u\in\mathcal{C}_1(J,\vartheta_x)} \frac{\sqrt{|J|}\cdot||M_J u||_2}{||u_J||_1} \geq \inf_{u\in\mathcal{C}_1(J,\vartheta_x)} \frac{||M_J u||_2}{||u||_2}.$$

Now, we illustrate these sufficient conditions in Lemma D.1 through the linear regression model (5.1) under both MCAR and MAR data settings and connect the sparsity structure of $\left[\mathbb{E}(RXX^T)\right]^{-1} \in \mathbb{R}^{d\times d}$ with the well-studied covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ of the covariate vector $X \in \mathbb{R}^d$ in the statistical literature.

Example D.1 (Linear regression model under the MCAR mechanism)**.** *Recall the high-dimensional sparse linear regression model* (5.1) *in Assumption 5.1 as:*

$$Y = X^T\beta + \epsilon \quad \textit{with} \quad \mathbb{E}(\epsilon|X) = 0 \quad \textit{and} \quad \mathbb{E}\left(\epsilon^2|X\right) = \sigma_\epsilon^2.$$

Furthermore, we suppose, without loss of generality, that the covariate vector $X \in \mathbb{R}^d$ *is centered in this example,* i.e.*,* $\mathbb{E}(X) = 0$*. In order to study the sparsity structure of*

$$\ell_0(x) = -2\left[\mathbb{E}(RXX^T)\right]^{-1} x := -2Mx,$$

it suffices to exploit the (weak) sparsity of $M = \left[\mathbb{E}(RXX^T)\right]^{-1} \in \mathbb{R}^{d\times d}$ *by Lemma D.1 when the query point* $x \in \mathbb{R}^d$ *is sparse or lies in a cone of dominant coordinates* $\mathcal{C}_1(J, \vartheta_x)$*. In particular, under the MCAR setting, we know that* $R \perp\!\!\!\perp (Y, X)$*, and the propensity score satisfies* $\pi(X, Y) = \mathrm{P}(R = 1|X, Y) = \mathrm{P}(R = 1)$ *so that*

$$M = \left[\mathbb{E}(RXX^T)\right]^{-1} = \frac{1}{\mathrm{P}(R=1)} \cdot \Sigma^{-1}$$

will be sparse if and only if the precision matrix Σ^{-1} *is sparse. Examples of sparse precision matrices include the autoregressive model of order* $q \geq 1$ *(whose precision matrix is banded; see,* e.g.*, Section 9.2 in Lindsey 2004) and solutions to the graphical Lasso model (Friedman et al., 2008) or other* L_1*-minimization programs (Cai et al., 2016). In addition, as the sufficient conditions in Lemma D.1(b,c) only require each column of* Σ^{-1} *to lie in* $\mathcal{C}_1(J, \vartheta)$ *for some* $J \subset \{1, ..., d\}$*, any moving-average process of order* $q \geq 1$ *(whose precision matrix is Toeplitz; recall* Σ^{ar} *in Section 5.4.1) and other short-range dependent processes (Rosenblatt, 2015) will have their precision matrices fulfilling the requirements as well.*

Example D.2 (Linear regression model under the MAR mechanism)**.** *Under the same linear model setup in Example D.1, we consider the MAR setting* $R \perp\!\!\!\perp Y|X$*, in which the propensity score becomes* $\pi(X) = \mathrm{P}(R = 1|X)$*. We consider several scenarios that relate the sparsity structures of* $M = \left[\mathbb{E}(RXX^T)\right]^{-1} \in \mathbb{R}^{d\times d}$ *to the covariance matrix* $\Sigma = \mathbb{E}(XX^T) \in \mathbb{R}^{d\times d}$*.*

1. *Suppose that the propensity score* $\pi(X)$ *depends on at most* $1 \leq s_\pi \ll d$ *covariates in* $X \in \mathbb{R}^d$*,* i.e.*, without loss of generality,*

$$\pi(x) = \mathrm{P}\left(R = 1|X = (x_1, ..., x_d)^T\right) = \pi(x_1, ..., x_{s_\pi}). \tag{D.2}$$

We further assume that the first few covariates $X^{(1)}, ..., X^{(d_1)}$ *with* $d_1 \geq s_\pi$ *are independent of the rest covariates* $X^{(d_1+1)}, ..., X^{(d_1+d_2)}$ *in the covariate vector* $X = \left(X^{(1)}, ..., X^{(d)}\right)^T \in \mathbb{R}^d$ *with* $d = d_1 + d_2$. *Then, the covariance matrix* Σ *is block diagonal as* $\Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix}$, *where* $\Sigma_1 \in \mathbb{R}^{d_1 \times d_1}$ *and* $\Sigma_2 \in \mathbb{R}^{d_2 \times d_2}$. *Under the sparsity assumption* (D.2) *on the propensity score* $\pi(X)$, *we know that*

$$\mathbb{E}(RXX^T) = \mathbb{E}\left(\pi(X)XX^T\right) = \begin{pmatrix} \widetilde{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\Sigma}_2 \end{pmatrix} \quad \text{and} \quad M = \left[\mathbb{E}(RXX^T)\right]^{-1} = \begin{pmatrix} \widetilde{\Sigma}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \widetilde{\Sigma}_2^{-1} \end{pmatrix}$$

for some matrices $\widetilde{\Sigma}_1^{-1}$ *and* Σ_2^{-1}. *Thus, the sufficient condition (a) in Lemma D.1 holds.*

2. *More generally, the complete-case Gram matrix* $\mathbb{E}(RXX^T) = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix}$ *has its inverse as:*

$$\begin{aligned} M &= \left[\mathbb{E}(RXX^T)\right]^{-1} \\ &= \begin{pmatrix} \widetilde{\Sigma}_{11}^{-1} + \widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1}\widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1} & -\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1} \\ -\left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1}\widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1} & \left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1} \end{pmatrix}. \end{aligned}$$

If the entries of $\widetilde{\Sigma}_{21} = \widetilde{\Sigma}_{12}^T$ *are small in absolute value, as can occur when* $\Sigma = \mathbb{E}(XX^T)$ *is diagonally dominant (such as for banded, Toeplitz, or other sparse covariance matrices; recall* Σ^{cs} *and* Σ^{ar} *in Section 5.4.1) together with* (D.2), *then the off-diagonal block matrices of* $M = \left[\mathbb{E}(RXX^T)\right]^{-1}$ *given by*

$$-\left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1}\widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1} = -\left[\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\left(\widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{21}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\right)^{-1}\right]^T$$

are also small in their absolute values compared with the diagonal block matrices of M *and therefore, the sufficient conditions (b) and (c) in Lemma D.1 apply.*

D.3 Generalized Linear Model with Lasso Regularization for Propensity Score Estimation

Recall that the consistency and asymptotic normality of our proposed debiasing method hold as long as the propensity scores are consistently estimated at the sample covariate vectors $X_i, i = 1, ..., n$ under the mild rate condition (5.21). While it is not necessary to estimate the propensity scores via parametric models for our debiasing method (see Section 5.4.4 for our simulation results), the generalized linear model serves as a baseline approach and can approximate many complex nonparametric models by including sufficiently many well-designed covariates. As a result, the covariate vector is often high-dimensional with $d \gg n$ in practice, and we consider estimating the propensity score $\pi(x) = \mathrm{P}(R = 1|X = x)$ under the MAR assumption (Assumption 5.1) using the data $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0, 1\}$ with the Lasso-type generalized linear regression defined by

$$\begin{aligned}\widehat{\theta} &= \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} R_i \log\left[\frac{\varphi(X_i^T\theta)}{1-\varphi(X_i^T\theta)}\right] - \frac{1}{n}\sum_{i=1}^{n} \log\left[1-\varphi(X_i^T\theta)\right] + \zeta\,||\theta||_1 \right\} \\ &:= \underset{\theta\in\mathbb{R}^d}{\arg\min} \left[\mathcal{L}_\varphi(\theta) + \zeta\,||\theta||_1\right],\end{aligned} \tag{D.3}$$

where $\mathcal{L}_\varphi : \mathbb{R}^d \to \mathbb{R}$ is the negative log-likelihood function associated with a known link function $\varphi : \mathbb{R} \to (0, 1)$ and $\zeta > 0$ is a regularization parameter. In particular, when the logistic link function $\varphi_{\log}(u) = \frac{1}{1+e^{-u}}$ is used, the regression problem (D.3) reduces to the following Lasso-type logistic regression as:

$$\widehat{\theta} = \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \left[R_i X_i^T\theta - \log\left(1+e^{X_i^T\theta}\right)\right] + \zeta\,||\theta||_1 \right\}. \tag{D.4}$$

With the estimate $\widehat{\theta}$ from the Lasso-type generalized linear regression (D.3), it is also feasible to derive an explicit formula for $r_\pi = r_\pi(n, \delta)$ in Assumption 5.5 for the estimated propensity scores $\widehat{\pi}_i, i = 1, ..., n$ under the following regularity conditions.

Assumption D.1 (Sparse generalized linear propensity score model)**.** *Under Assumption 5.1, the missingness variable $R \in \{0, 1\}$ follows the model*

$$R|X \sim \text{Bernoulli}\left(\varphi(X^T\theta_0)\right), \tag{D.5}$$

where $\varphi : \mathbb{R} \to (0,1)$ *is a known link function and* $\theta_0 \in \mathbb{R}^d$ *is a high-dimensional sparse vector with support* S_θ *and* $|S_\theta| = ||\theta_0||_0 = s_\theta \ll d$.

Assumption D.2 ($\mathcal{C}_1(S_\theta, 3)$-restricted eigenvalue condition on $\mathbb{E}(RXX^T)$)**.** *The* $\mathcal{C}_1(S_\theta, 3)$*-restricted minimum eigenvalue of the complete-case population Gram matrix* $\mathbb{E}\left(RXX^T\right) \in \mathbb{R}^{d\times d}$ *is bounded away from zero,* i.e.*, there exists an absolute constant* $\rho > 0$ *such that*

$$\mathbb{E}\left[R(X^T v)^2\right] > \rho\, ||v||_2^2 \quad \text{for any} \quad v \in \mathcal{C}_1(S_\theta, 3),$$

where S_θ *is the support of the* s_θ*-sparse vector* $\theta_0 \in \mathbb{R}^d$.

Assumption D.3 (Regular link function)**.** *The link function* φ *satisfies the following regularity conditions.*

(a) $\varphi : \mathbb{R} \to [0,1]$ *is twice differentiable, nondecreasing, and* L_φ*-Lipschitz on* $\mathbb{R}$ *as well as concave on* $\mathbb{R}^+$ *for some absolute constant* $L_\varphi > 0$*; also, for any* $u \in \mathbb{R}$*, it holds that* $\varphi(u) + \varphi(-u) = 1$.

(b) There exists an absolute constant $A_\varphi > 0$ *such that* $\frac{\varphi'(u)}{1-\varphi(u)} \le A_\varphi$ *for all* $u \ge 0$.

(c) For $\mathcal{L}_\varphi(\theta)$ *defined in* (D.3)*, there exists an absolute constant* $A_H > 1$ *such that the Hessian matrix* $\nabla\nabla\mathcal{L}_\varphi(\theta)$ *can be expressed as* $\nabla\nabla\mathcal{L}_\varphi(\theta) = \frac{1}{n}\sum_{i=1}^n H(\theta; R_i, X_i) X_i X_i^T$ *for some* $H(\theta; R_i, X_i) > 0$ *satisfying*

$$|\log H(\theta + b; R_i, X_i) - \log H(\theta; R_i, X_i)| \le A_H\left(|X_i^T\theta|^2 + |X_i^T b|^2 + |X_i^T b|\right)$$

for any $\theta, b \in \mathbb{R}^d$ *and* $i = 1, ..., n$.

The sparse generalized linear regression setup in Assumption D.1 is commonly imposed in the statistical literature to handle high-dimensional data (van de Geer, 2008; Huang and Zhang, 2012; Cai et al., 2023). Assumption D.2 is the restricted eigenvalue condition over the cone $\mathcal{C}_1(S_\theta, 3)$; recall (5.17). In Assumption D.3, the constants $A_\varphi, A_H > 0$ depend only on the link function φ. A similar condition is also imposed in Cai et al. (2023) to establish

the consistency of $\widehat{\theta}$ in (D.3) and derive the asymptotic normality of their debiased estimator for $\widehat{\theta}$. One can verify by direct calculation and Example 8 in Huang and Zhang (2012) that the logistic link function $\varphi_{\log}(u) = \frac{1}{1+e^{-u}}$ satisfies Assumption D.3. Furthermore, Section 3 of the supplement in Cai et al. (2023) also verifies Assumption D.3 for the probit link function $\varphi_{\text{probit}}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt$, the class of generalized logistic functions, as well as the CDFs of Student's t_ν-distributions with any $\nu \in \mathbb{N}$ in their latent variable model (Example 3 in their supplement).

Remark D.2. *Assumption D.3(b) implies that* $\max\left\{\frac{\varphi'(u)}{\varphi(u)}, \frac{\varphi'(u)}{1-\varphi(u)}\right\} \le A_\varphi$ *for the same constant* $A_\varphi > 0$. *This is because if* $u \ge 0$, *then by Assumption D.3(a),* $\varphi(u) \ge \frac{1}{2}$ *and consequently,*

$$\frac{\varphi'(u)}{\varphi(u)} \le \frac{\varphi'(u)}{1-\varphi(u)} \le A_\varphi.$$

On the other hand, if $u < 0$, *then Assumption D.3(a) also shows that* $\varphi(u) \le \frac{1}{2}$ *and therefore,*

$$\frac{\varphi'(u)}{1-\varphi(u)} \le \frac{\varphi'(u)}{\varphi(u)} = \frac{\varphi'(-u)}{1-\varphi(-u)} \le A_\varphi,$$

where we use the condition that $\varphi(x) + \varphi(-x) = 1$ *and* $\varphi'(x) - \varphi'(-x) = 0$ *in Assumption D.3(a).*

As for Assumption D.3(c), some algebra shows that

$$H(\theta; R_i, X_i) = -\frac{R_i\left[\varphi''(X_i^T\theta)\varphi(X_i^T\theta) - \left(\varphi'(X_i^T\theta)\right)^2\right]}{\left[\varphi(X_i^T\theta)\right]^2} + \frac{(1-R_i)\left[\varphi''(X_i^T\theta)\left(1-\varphi(X_i^T\theta)\right) + \left(\varphi'(X_i^T\theta)\right)^2\right]}{\left[1-\varphi(X_i^T\theta)\right]^2}.$$

Hence, a sufficient condition for $H(\theta; R_i, X_i) > 0$ *is that*

$$\left[\varphi'(X_i^T\theta)\right]^2 > \max\left\{\varphi''(X_i^T\theta)\varphi(X_i^T\theta), \varphi''(X_i^T\theta)\left[\varphi(X_i^T\theta) - 1\right]\right\}, \tag{D.6}$$

which is also equivalent to $\left[\varphi'(X_i^T\theta)\right]^2 > \varphi''(X_i^T\theta)\varphi(X_i^T\theta)$ *for all* $X_i^T\theta < 0$. *This condition can be easily verified for the logistic link function* $\varphi_{\log}(u) = \frac{1}{1+e^{-u}}$ *and the probit link function* $\varphi_{\text{probit}}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt$.

Under the above regularity conditions, we derive the explicit formula for $r_\pi = r_\pi(n, \delta)$ as follows.

Proposition D.2. *Let $\delta \in (0,1)$ and $A_1, A_2, A_3, A_4 > 0$ be some absolute constants. Suppose that Assumptions 5.2, D.1, D.2, and D.3 hold as well as*

$$d \geq s_\theta + 2, \quad A_1 \phi_{s_\theta} \log\left(\frac{\phi_{s_\theta}}{\rho}\right) \sqrt{\frac{\phi_1 s_\theta \log d}{n}} \leq \frac{\rho}{8}, \quad \text{and} \quad A_2 \exp\left[-\frac{n\rho^2}{\phi_{s_\theta}^2 \log^2\left(\frac{\phi_{s_\theta}}{\rho}\right)} - \phi_1 s_\theta \log d\right] \leq \delta.$$

If $\zeta = A_3 A_\varphi \sqrt{\frac{\phi_1 \log(d/\delta)}{n}}$, then the rate of consistency $r_\pi = r_\pi(n, \delta)$ in Assumption 5.5 can be taken as:

$$r_\pi(n, \delta) = \frac{A_4 L_\varphi A_\varphi}{\rho} \sqrt{\frac{\phi_1 \phi_{s_\theta} s_\theta \log(d/\delta)}{n} \left[\log\left(\frac{n}{\delta}\right) + s_\theta \log\left(\frac{ed}{s_\theta}\right)\right]}.$$

This proposition suggests that the rate of convergence for r_π is of the order $O_P\left(\frac{s_\theta \log d}{\sqrt{n}}\right)$, provided that all other model parameters $L_\varphi, A_\varphi, \rho, \phi_1, \phi_{s_\theta} > 0$ are independent of d and n with $\delta = \frac{1}{d}$ and $d > n$.

D.3.1 Other Examples for Propensity Score Estimation

• **Single-index model with unknown link function.** As a preliminary step for more complex high-dimensional semiparametric models, we first consider a simple univariate semiparametric model

$$\pi(z) = \mathbb{E}(R_i | Z_i = z) = F_0(z), \quad \text{and} \quad \pi_i = \pi(Z_i), \quad i = 1, \ldots, n,$$

where $F_0 \in \Lambda = \left\{F : \mathbb{R} \to (0,1) : \int_{-\infty}^{\infty} \left[F''(x)\right]^2 dx < \infty\right\}$ and $z = g_0(x) \in \mathbb{R}$ for some transformation function $g_0 : \mathbb{R}^d \to \mathbb{R}$. Assume tentatively that g_0 is given so that $Z_i = g_0(X_i), i = 1, ..., n$. The unknown function F_0 can be estimated by solving the penalized maximum likelihood estimator

$$\widehat{F} = \arg\max_{F \in \Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \log\left[R_i F(Z_i) + (1 - R_i)\big(1 - F(Z_i)\big)\right] - \lambda^2 \int_{-\infty}^{\infty} \left[F''(x)\right]^2 dx\right],$$

and the propensity score estimates are $\widehat{\pi}_i = \widehat{F}(Z_i)$, $i = 1, \ldots, n$.

If $\lambda \asymp n^{-\frac{2}{5}}$ and F_0 is bounded away from zero and one and its first derivative F_0' is bounded away from zero over a compact set, we have

$$\max_{1 \le i \le n} |\widehat{\pi}_i - \pi_i| \le \sup_{z \in \mathbb{R}} |\widehat{F}(z) - F_0(z)| = O_P\left(n^{-\frac{3}{10}}\right).$$

For detailed derivations and precise regularity conditions, see Lemma 11.14 on Page 235 in van de Geer (2000). Hence, Assumption 5.5 holds with $r_\pi = O\left(n^{-\frac{3}{10}}\right)$, and the rate condition (5.21) is satisfied if $s_{\max}\sqrt{\log d} = o\left(n^{\frac{3}{10}}\right)$ for those high-dimensional settings.

• **High-dimensional series estimator.** Consider the following generalization of the example from (D.5):

$$\pi(x) = \mathbb{E}(R_i | X_i = x) = g\big(\alpha_0^T \Psi(x)\big), \quad \text{and} \quad \pi_i = \pi(X_i), \quad i = 1, \ldots, n,$$

where $g : \mathbb{R} \to [0, 1]$ is a known L_g-Lipschitz continuous link function, $\Psi(x) = \{\psi_k(x)\}_{k=1}^K$ is a set of K known basis functions, possibly high-dimensional with $K \gg n$, $\alpha_0 \in \mathbb{R}^K$ is an unknown sparse parameter vector, and $x \in \mathcal{X}$. Under this model, Section 5.1 in Chakrabortty et al. (2019) suggests estimating the propensity scores at the sample points $X_1, ..., X_n$ via

$$\widehat{\pi}_i = g\big(\widehat{\alpha}^T \Psi(X_i)\big), \quad i = 1, \ldots, n,$$

where $\widehat{\alpha}$ denotes an ℓ_1-norm consistent estimate of α_0.

If $\mathrm{P}\left(\|\widehat{\alpha} - \alpha_0\|_1 > a_n\right) \le q_n$ for some $a_n, q_n = o(1)$ and $a_n \ge 0$, $q_n \in [0, 1]$, then

$$\max_{1 \le i \le n} |\widehat{\pi}_i - \pi_i| \le \sup_{x \in \mathcal{X}} \left|g\big(\widehat{\alpha}^T \Psi(x)\big) - g\big(\alpha_0^T \Psi(x)\big)\right| = O_P(a_n).$$

For detailed derivations and precise regularity conditions, see Theorem B.1 in Chakrabortty et al. (2019). Furthermore, according to Appendix B.1 in Chakrabortty et al. (2019), it is reasonable to assume that $a_n \propto s_\pi \sqrt{\frac{\log K}{n}}$ and $q_n = o\left(\frac{1}{nK}\right)$, where s_π is the (small) number of basis functions of the series representation of π_i (*i.e.*, sparsity of α_0). Hence, Assumption 5.5 holds with $r_\pi = O\left(s_\pi \sqrt{\frac{\log K}{n}}\right)$ and the rate condition (5.21) is satisfied if $s_{\max}^2 s_\pi^2 (\log d)(\log K) = o(n)$.

• **High-dimensional single-index kernel smoother.** Consider the following semiparametric single-index model

$$\pi(x) = \mathbb{E}(R_i|X_i = x) = g_0(\gamma_0^T x), \quad \text{and} \quad \pi_i = \pi(X_i), \quad i = 1, \ldots, n,$$

where $g_0 : \mathbb{R} \to [0, 1]$ is some unknown link function and $\gamma_0 \in \mathbb{R}^d$ is an unknown direction parameter (identifiable only up to scalar multiples).

We consider estimating the π_i's via a one-dimensional Nadaraya–Watson-type kernel smoother over the estimated scores $\{\widehat{\gamma}^T X_i\}_{i=1}^n$ as:

$$\widehat{\pi}_i \equiv \widehat{\pi}(\widehat{\gamma}^T X_i) \equiv \frac{\sum_{j=1}^n R_j \cdot K\left(\frac{\widehat{\gamma}^T X_j - \widehat{\gamma}^T X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{\widehat{\gamma}^T X_j - \widehat{\gamma}^T X_i}{h}\right)}, \quad i = 1, \ldots, n,$$

where $K : \mathbb{R} \to [0, \infty)$ is a (symmetric) kernel function, $h > 0$ is a smoothing bandwidth parameter, and $\widehat{\gamma}$ is an estimate of γ_0. The estimate $\widehat{\gamma}$ can be computed via any standard method in the literature on signal recovery for single-index models (possibly under additional assumptions such as elliptically distributed covariates). More details can be found in Section 5.2 in Chakrabortty et al. (2019) and references therein.

If $\mathrm{P}\left(\|\widehat{\gamma} - \widetilde{\gamma}_0\|_1 > a_n\right) \le q_n$ for some $a_n, q_n = o(1)$ for some direction $\widetilde{\gamma}_0 \propto \gamma_0$, and $a_n \ge 0$, $q_n \in [0, 1]$, then for all $t > 0$, with probability at least $1 - 27(nd)^{-t^2} - 6q_n$,

$$|\widehat{\pi}_i - \pi_i| \lesssim t\sqrt{\frac{\log(nd)}{nh}} + h^2 + a_n + \frac{a_n^2}{h^2} + \frac{\log(nd)}{nh}.$$

For detailed derivations and precise regularity conditions, see Theorem B.3 in Chakrabortty et al. (2019). Furthermore, according to Section 5.2 in Chakrabortty et al. (2019), there exist estimators $\widehat{\gamma}$ such that $a_n \propto s_\pi\sqrt{\frac{\log d}{n}}$ and $q_n = o\left(\frac{1}{nd}\right)$, where s_π is the sparsity of the direction $\widetilde{\gamma}_0 \propto \gamma_0$. If $\log(nd) = o(nh)$, then Assumption 3.2 in Chakrabortty et al. (2019) holds with $v_{n,\pi} = O\left(\sqrt{\frac{\log(nd)}{nh}}\right)$, $w_{n,\pi} = O\left(h^2 + s_\pi\sqrt{\frac{\log d}{n}} + \frac{s_\pi^2 \log d}{nh^2} + \frac{\log(nd)}{nh}\right)$, and $q_{n,\pi} = o\left(\frac{1}{nd}\right)$, which in turn implies our Assumption 5.5. The rate condition (5.21) is therefore satisfied if $\left(v_{n,\pi}\sqrt{\log n} + w_{n,\pi}\right) s_{\max}\sqrt{\log d} = o(1)$.

• **Partially linear single-index model.** Consider the following semiparametric model

$$\pi(x) = \mathbb{E}(R_i|X_i = x) = F\left(u^T\alpha_0 + \gamma_0(v)\right), \quad x = (u, v), \quad \text{and} \quad \pi_i = \pi(X_i), \quad i = 1, \ldots, n,$$

where the link function $F : \mathbb{R} \to (0,1)$ is known and differentiable, $\alpha_0 \in \mathbb{R}^{d_\pi}$, and $\gamma_0 \in \mathcal{G}$ with

$$\mathcal{G} = \left\{ \gamma : [0,1] \to \mathbb{R}, \int \left[\gamma^{(m)}(v)\right]^2 dv < \infty \right\},$$

for some integer $m \geq 1$. To estimate $g_0(u,v) = u^T\alpha_0 + \gamma_0(v)$, we use the penalized maximum likelihood estimator $\widehat{g}(u,v) = u^T\widehat{\alpha} + \widehat{\gamma}(v)$ defined as

$$\widehat{g} = \arg\max_{g \in \Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \log\left[R_i F(g(X_i)) + (1-R_i)\big(1 - F(g(X_i))\big)\right] - \lambda^2 \int_{-\infty}^{\infty} \left[\gamma^{(m)}(v)\right]^2 dv \right],$$

where $\Lambda = \{g : \mathbb{R}^{d_\pi} \times [0,1] \to \mathbb{R}, g(u,v) = u^T\beta + \gamma(v), \beta \in \mathbb{R}^{d_\pi}, \gamma \in \mathcal{G}\}$.

If $\lambda \asymp n^{-\frac{m}{2m+1}}$, $F \circ g$ is bounded away from zero and one, $\sup_{x \in \mathbb{R}} |F'(x)| < \infty$, and the dimension d_π of the linear component is fixed, then

$$\frac{1}{n} \sum_{i=1}^{n} \left[F(\widehat{g}(X_i)) - F(g_0(X_i))\right]^2 = O_P\big(n^{-\frac{2m}{2m+1}}\big).$$

For detailed derivations and precise regularity conditions, see Lemma 11.4 on Page 206 in van de Geer (2000). Since $\frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}_i - \pi_i| \leq \left(\frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}_i - \pi_i|^2\right)^{1/2}$, the alternative condition to Assumption 5.5, $\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta$, holds with $r_\pi = O\left(n^{-\frac{m}{2m+1}}\right)$.

Notice that a more careful analysis of the metric entropy numbers will allow one to extend this result to high-dimensional partially linear single-index models, provided that the high-dimensional regression vector is sparse. Similarly, it is possible to generalize this model to allow for a multivariate nuisance function $\gamma_0 : \mathbb{R}^p \to \mathbb{R}$; see Page 19 in van de Geer (2000).

• **High-dimensional generalized sparse additive model.** Consider the following high-dimensional generalized sparse additive model as:

$$\pi(x) = \mathbb{E}(R_i | X_i = x) = g^{-1}\left(\sum_{j=1}^{d} f_j^*(x_j)\right), \quad \text{and} \quad \pi_i = \pi(X_i) \quad i = 1, \ldots, n,$$

where $g : \mathbb{R} \to \mathbb{R}$ is a known link function whose inverse is L_g-Lipschitz (*e.g.*, $g(u) = \log\left(\frac{u}{1-u}\right)$), x_j is the jth entry of vector $x \in \mathbb{R}^d$, and $f_1^*, \ldots, f_d^* \in \mathcal{F}$ for some suitable function class $\mathcal{F}$. For concreteness, we assume that

$$\mathcal{F} = \left\{ f : [0,1] \to \mathbb{R}, \int \left|f^{(m)}(x)\right|^r dx < \infty \right\},$$

for some $r \geq 1$ and $1 \leq m \leq 2$. Furthermore, we assume that there exists an absolute constant $M > 0$ such that $\sum_{j=1}^{d} \mathbf{1}(f_j^* \neq 0) \leq M$; see Section 4.1 in Haris et al. (2022) for details and definitions.

The unknown functions $f_j^*, j = 1, ..., d$ can be estimated by solving the doubly penalized M-estimation problem

$$\left(\widehat{f}_1, ..., \widehat{f}_d\right) = \underset{f_1,...,f_d \in \mathcal{F}}{\arg\min} \left[-\frac{1}{n} \sum_{i=1}^{n} L\left(R_i, \sum_{j=1}^{d} f_j(X_{ij})\right) + \lambda^2 \sum_{j=1}^{d} P_{st}(f_j) + \frac{\lambda}{n} \sum_{j=1}^{d} \sum_{i=1}^{n} f_j^2(X_{ij}) \right],$$

where $L(\cdot, \cdot)$ is a loss function (*e.g.*, cross-entropy loss), $\lambda > 0$ is a regularization parameter, and P_{st} is a structure-inducing penalty; see Section 2.1 in Haris et al. (2022) for detailed examples. Define the estimator of $\pi_i = \pi(X_i)$ by

$$\widehat{\pi}_i = g^{-1}\left(\sum_{j=1}^{d} \widehat{f}_j(X_{ij})\right), \quad i = 1, \ldots, n.$$

If $\lambda \asymp n^{-\frac{1}{2}}$ and under the above regularity conditions on $\mathcal{F}$, then, with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^{n} |\widehat{\pi}_i - \pi(X_i)| \overset{\text{(i)}}{\leq} L_g \left(\frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{d} \left\{ \widehat{f}_j(X_{ij}) - f_j^*(X_{ij}) \right\} \right]^2 \right)^{1/2} \lesssim M \left(\frac{\log(d/\delta)}{n} \right)^{1/4},$$

where (i) follows from the L_g-Lipschitz continuity of g^{-1} and from Hölder's inequality between the empirical ℓ_1- and ℓ_2-norms. For detailed derivations and precise regularity conditions, see Proposition 3 in Tan and Zhang (2019). Hence, the alternative condition to Assumption 5.5 as $\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta$ holds with $r_\pi = O\left((\log d)^{\frac{1}{4}} n^{-\frac{1}{4}}\right)$.

D.4 Additional Simulation Results

We have carried out simulations in Section 5.4 to compare our proposed debiasing inference method with standard methods in the literature under various data-generating processes and MAR settings. Here, we provide the complete simulation results presented in the chapter. Additionally, we conduct comparative experiments under the simpler MCAR setting for completeness and to illustrate Example D.1.

D.4.1 Full Comparative Simulation Results

Here is an outline for all the comparative simulation results in this section.

- Figure D.1 displays comparative plots of the three evaluation metrics and a normality assessment under the Toeplitz (or autoregressive) covariance design with Gaussian $\mathcal{N}(0,1)$ noise for a selected simulation setting.

- Table D.1 presents the simulation results under the circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0,1)$, and the MAR setting (5.23) in Section 5.4.1.

- Table D.2 presents the simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, Gaussian noise $\mathcal{N}(0,1)$, and the MAR setting (5.23).

- Table D.3 presents the simulation results under the circulant symmetric covariance matrix Σ^{cs}, Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$-distributed noise, and the MAR setting (5.23).

- Table D.4 presents the simulation results under the circulant symmetric covariance matrix Σ^{cs}, t_2-distributed noise, and the MAR setting (5.23).

- Table D.5 presents the simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$-distributed noise, and the MAR setting (5.23).

- Table D.6 presents the simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, t_2-distributed noise, and the MAR setting (5.23).

- Figure D.3 displays the simulation results for our debiasing method with dense β_0^{de} and sparse $x^{(2)}$ when the propensity scores are estimated by various machine learning methods under the new MAR setting (5.24).

- Table D.7 presents the simulation results for our proposed debiasing method when the propensity scores are estimated by various machine learning methods under the

circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0,1)$, and the new MAR setting (5.24) in Section 5.4.4.

- Figure D.2 supplements the "QQ-plots" corresponding to the asymptotic normality plots for our proposed debiasing method in Figure D.3 and Figure 5.5 when the propensity scores are estimated by various machine learning methods under the circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0,1)$, and the new MAR setting (5.24).

D.4.2 Simulations Under the MCAR Setting

We follow the comparative simulation designs in Section 5.4.1 but generate the missingness indicators $R_i, i = 1, ..., n$, under an MCAR mechanism with $\mathrm{P}(R_i = 1) = 0.7$. Under MCAR, the inverses of both the circulant symmetric matrix Σ^{cs} and the Toeplitz matrix Σ^{ar} satisfy our sufficient conditions for the sparse approximation of the population dual solution $\ell_0(x)$ in Lemma D.1, as illustrated in Example D.1. The simulation results under Gaussian $\mathcal{N}(0,1)$ noise are shown in Table D.8 and Table D.9, and the conclusions are similar to the main findings from the MAR simulation studies in Section 5.4.2. We omit results for other noise distributions because they exhibit patterns similar to those under Gaussian noise, and we have demonstrated in Section 5.4.3 that our proposed debiasing method is robust to other heavy-tailed noise distributions.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0.1}^{(a)}$																		
Avg-Bias	0.033	0.035	0.028	0.040	0.038	0.032	0.249	0.198	0.315	0.032	0.036	0.027	0.039	0.038	0.188	0.042	0.039	0.038
Avg-Cov	0.962	0.962	0.967	0.899	0.882	0.904	0.962	0.958	0.940	0.954	0.906	0.953	0.910	0.899	0.482	0.853	0.898	0.914
Avg-Len	0.176	0.184	0.145	0.161	0.152	0.133	0.369	0.305	0.638	0.162	0.156	0.133	0.166	0.154	10.887	0.151	0.161	0.168
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0.1}^{(b)}$																		
Avg-Bias	0.146	0.154	0.108	0.181	0.158	0.136	0.098	0.104	0.207	0.184	0.193	0.143	0.161	0.152	0.222	0.224	0.191	0.172
Avg-Cov	0.995	0.992	0.996	0.891	0.855	0.913	1.00	0.998	0.996	0.922	0.838	0.919	0.357	0.906	0.497	0.860	0.931	0.969
Avg-Len	0.981	0.998	0.844	0.757	0.593	0.583	1.253	1.042	1.784	0.791	0.665	0.610	0.185	0.627	7.805	0.819	0.871	0.908
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0.1}^{(c)}$																		
Avg-Bias	0.044	0.049	0.036	0.054	0.049	0.041	0.068	0.057	0.088	0.049	0.050	0.037	0.048	0.052	0.379	0.065	0.059	0.055
Avg-Cov	0.991	0.989	0.996	0.891	0.868	0.908	0.987	0.983	0.985	0.930	0.897	0.938	0.898	0.914	0.466	0.857	0.913	0.949
Avg-Len	0.284	0.292	0.237	0.223	0.190	0.173	0.404	0.351	0.527	0.222	0.200	0.172	0.197	0.220	7.729	0.241	0.256	0.267
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.042	0.042	0.037	0.069	0.056	0.051	NA	NA	NA	0.035	0.042	0.030	0.062	0.038	0.286	0.039	0.039	0.042
Avg-Cov	0.985	0.995	0.990	0.807	0.842	0.839	NA	NA	NA	0.946	0.898	0.948	0.862	0.914	0.497	0.929	0.937	0.938
Avg-Len	0.262	0.312	0.236	0.215	0.202	0.172	NA	NA	NA	0.175	0.170	0.144	0.218	0.169	14.284	0.174	0.183	0.201
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.171	0.186	0.130	0.344	0.215	0.226	NA	NA	NA	0.252	0.297	0.184	0.211	0.167	0.368	0.217	0.207	0.183
Avg-Cov	1.00	1.00	1.00	0.775	0.851	0.821	NA	NA	NA	0.844	0.704	0.887	0.344	0.902	0.477	0.921	0.940	0.990
Avg-Len	1.659	1.860	1.514	1.008	0.789	0.755	NA	NA	NA	0.931	0.788	0.726	0.241	0.690	9.101	0.939	0.989	1.089
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.054	0.056	0.046	0.103	0.073	0.069	NA	NA	NA	0.061	0.067	0.046	0.063	0.057	0.506	0.067	0.065	0.060
Avg-Cov	1.00	1.00	1.00	0.765	0.826	0.806	NA	NA	NA	0.935	0.842	0.928	0.904	0.900	0.480	0.904	0.927	0.965
Avg-Len	0.467	0.533	0.414	0.297	0.253	0.224	NA	NA	NA	0.266	0.239	0.205	0.258	0.243	10.243	0.277	0.291	0.321
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0.100}^{(a)}$																		
Avg-Bias	0.025	0.026	0.022	0.032	0.032	0.027	0.740	0.556	0.739	0.000	0.001	0.000	0.034	0.032	0.101	0.031	0.033	0.035
Avg-Cov	0.991	0.992	0.990	0.952	0.945	0.950	0.960	0.963	0.937	1.00	0.992	1.00	0.947	0.941	0.491	0.949	0.944	0.944
Avg-Len	0.176	0.185	0.145	0.162	0.152	0.133	6.590	1.483	1.177	0.000	0.001	0.000	0.163	0.151	12.879	0.151	0.161	0.168
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0.100}^{(b)}$																		
Avg-Bias	0.102	0.102	0.090	0.169	0.132	0.128	0.093	0.096	0.084	0.186	0.193	0.191	0.139	0.143	0.109	0.192	0.174	0.161
Avg-Cov	0.999	1.00	0.999	0.922	0.909	0.929	1.00	0.997	0.997	0.000	0.050	0.004	0.364	0.917	0.516	0.901	0.948	0.972
Avg-Len	0.982	0.999	0.844	0.758	0.594	0.583	1.256	1.073	1.728	0.006	0.057	0.021	0.184	0.623	6.645	0.819	0.870	0.910
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0.100}^{(c)}$																		
Avg-Bias	0.031	0.031	0.028	0.047	0.042	0.038	0.122	0.561	0.316	0.039	0.042	0.040	0.041	0.049	0.084	0.053	0.052	0.051
Avg-Cov	0.996	0.996	0.999	0.947	0.925	0.942	0.996	0.987	0.987	0.000	0.019	0.000	0.938	0.920	0.487	0.929	0.948	0.966
Avg-Len	0.285	0.293	0.237	0.223	0.190	0.173	1.567	1.857	28.765	0.000	0.008	0.002	0.195	0.217	8.732	0.241	0.256	0.268
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.043	0.042	0.037	0.078	0.060	0.056	NA	NA	NA	0.036	0.043	0.030	0.068	0.058	0.388	0.052	0.045	0.041
Avg-Cov	1.00	1.00	1.00	0.800	0.864	0.827	NA	NA	NA	0.949	0.880	0.945	0.868	0.721	0.511	0.788	0.881	0.934
Avg-Len	0.541	0.630	0.449	0.232	0.221	0.183	NA	NA	NA	0.175	0.169	0.144	0.235	0.160	15.971	0.162	0.171	0.186
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.165	0.168	0.145	0.892	0.436	0.590	NA	NA	NA	0.800	0.799	0.620	0.276	0.629	0.623	1.259	1.062	0.762
Avg-Cov	1.00	1.00	1.00	0.067	0.501	0.156	NA	NA	NA	0.082	0.040	0.074	0.265	0.055	0.477	0.000	0.009	0.117
Avg-Len	3.992	4.368	3.449	1.088	0.861	0.806	NA	NA	NA	0.863	0.730	0.669	0.263	0.646	12.970	0.878	0.925	1.004
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.078	0.063	0.082	0.250	0.156	0.170	NA	NA	NA	0.150	0.149	0.110	0.099	0.268	0.582	0.364	0.303	0.216
Avg-Cov	1.00	1.00	1.00	0.106	0.393	0.170	NA	NA	NA	0.334	0.286	0.398	0.785	0.009	0.462	0.000	0.017	0.166
Avg-Len	1.090	1.211	0.907	0.320	0.276	0.239	NA	NA	NA	0.244	0.220	0.189	0.291	0.228	11.660	0.259	0.273	0.296
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.036	0.036	0.031	0.043	0.043	0.032	NA	NA	NA	0.003	0.008	0.002	0.040	0.016	0.076	0.018	0.017	0.020
Avg-Cov	1.00	1.00	1.00	0.936	0.937	0.939	NA	NA	NA	0.889	0.534	0.870	0.944	0.000	0.491	0.000	0.000	0.961
Avg-Len	1.768	2.028	1.367	0.203	0.197	0.154	NA	NA	NA	0.010	0.015	0.008	0.190	0.000	7.220	0.000	0.006	0.104

Table D.1: Simulation results under the circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0, 1)$, and the MAR setting (5.23). Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "DL-vdG", "R-Proj", or "SSL-AIPW" and thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																	
Avg-Bias	0.073	0.073	0.061	0.073	0.073	0.061	0.579	0.481	0.608	0.073	0.074	0.060	0.073	0.074	0.216	0.074	0.075	0.089
Avg-Cov	0.871	0.870	0.894	0.900	0.893	0.917	0.953	0.950	0.950	0.946	0.944	0.956	0.910	0.919	0.496	0.889	0.913	0.939
Avg-Len	0.279	0.281	0.247	0.308	0.305	0.266	1.196	1.539	1.917	0.354	0.352	0.301	0.306	0.322	40.120	0.296	0.326	0.413
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																	
Avg-Bias	0.274	0.276	0.152	0.267	0.218	0.153	0.297	0.219	0.372	0.428	0.257	0.178	0.254	0.225	0.172	0.468	0.405	0.358
Avg-Cov	0.757	0.762	0.927	0.571	0.427	0.559	0.943	0.800	0.945	0.781	0.748	0.824	0.630	0.957	0.486	0.845	0.915	0.987
Avg-Len	0.828	0.842	0.671	0.528	0.306	0.301	1.008	0.617	0.833	1.157	0.758	0.593	0.576	1.190	16.077	1.651	1.819	2.306
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																	
Avg-Bias	0.091	0.092	0.071	0.093	0.091	0.073	0.118	0.181	0.265	0.098	0.095	0.075	0.095	0.092	0.226	0.103	0.096	0.101
Avg-Cov	0.870	0.863	0.906	0.847	0.799	0.863	0.948	0.935	0.925	0.912	0.903	0.921	0.852	0.953	0.499	0.874	0.926	0.961
Avg-Len	0.340	0.341	0.300	0.329	0.295	0.275	0.611	0.593	1.490	0.416	0.388	0.335	0.339	0.457	28.389	0.390	0.429	0.544
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																	
Avg-Bias	0.057	0.057	0.047	0.068	0.065	0.055	NA	NA	NA	0.048	0.049	0.041	0.066	0.053	0.188	0.056	0.060	0.070
Avg-Cov	0.982	0.990	0.994	0.910	0.913	0.927	NA	NA	NA	0.948	0.948	0.952	0.921	0.908	0.493	0.934	0.942	0.956
Avg-Len	0.353	0.386	0.319	0.290	0.287	0.248	NA	NA	NA	0.237	0.237	0.202	0.290	0.217	58.341	0.255	0.287	0.348
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																	
Avg-Bias	0.235	0.241	0.126	0.239	0.199	0.137	NA	NA	NA	0.376	0.244	0.177	0.223	0.214	0.154	0.386	0.336	0.310
Avg-Cov	0.987	0.987	0.999	0.577	0.429	0.577	NA	NA	NA	0.698	0.693	0.750	0.668	0.873	0.492	0.852	0.939	0.981
Avg-Len	1.367	1.457	1.071	0.497	0.288	0.280	NA	NA	NA	0.996	0.650	0.519	0.546	0.802	17.531	1.426	1.603	1.943
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																	
Avg-Bias	0.082	0.083	0.065	0.090	0.085	0.070	NA	NA	NA	0.086	0.083	0.067	0.085	0.088	0.193	0.091	0.085	0.087
Avg-Cov	0.988	0.992	0.998	0.836	0.807	0.860	NA	NA	NA	0.918	0.906	0.921	0.856	0.852	0.460	0.872	0.933	0.962
Avg-Len	0.521	0.555	0.456	0.310	0.278	0.256	NA	NA	NA	0.375	0.349	0.301	0.321	0.309	29.895	0.336	0.378	0.459
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																	
Avg-Bias	0.057	0.057	0.054	0.076	0.076	0.068	0.597	0.926	0.733	0.000	0.000	0.000	0.078	0.096	0.264	0.080	0.091	0.101
Avg-Cov	0.984	0.986	0.978	0.948	0.943	0.939	0.971	0.970	0.958	1.000	0.999	1.000	0.950	0.953	0.479	0.947	0.947	0.943
Avg-Len	0.348	0.350	0.312	0.367	0.364	0.322	1.667	11.133	8.075	0.000	0.000	0.000	0.385	0.463	38.258	0.373	0.428	0.480
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																	
Avg-Bias	0.173	0.173	0.130	0.224	0.190	0.167	0.309	0.239	0.279	0.260	0.223	0.200	0.231	0.304	0.186	0.399	0.372	0.390
Avg-Cov	0.978	0.981	0.990	0.782	0.529	0.604	0.989	0.888	0.954	0.251	0.349	0.391	0.827	0.976	0.507	0.960	0.986	0.992
Avg-Len	1.055	1.072	0.856	0.629	0.365	0.364	1.304	0.792	1.097	0.411	0.376	0.331	0.724	1.718	40.498	2.084	2.392	2.681
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																	
Avg-Bias	0.072	0.072	0.067	0.094	0.088	0.082	0.421	0.244	0.394	0.059	0.058	0.056	0.096	0.128	0.238	0.103	0.107	0.115
Avg-Cov	0.973	0.973	0.969	0.892	0.884	0.890	0.962	0.946	0.963	0.302	0.341	0.377	0.923	0.968	0.523	0.943	0.960	0.969
Avg-Len	0.431	0.432	0.381	0.392	0.352	0.333	2.658	0.774	1.953	0.136	0.146	0.143	0.427	0.660	57.514	0.492	0.564	0.633
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																	
Avg-Bias	0.050	0.050	0.040	0.069	0.063	0.054	NA	NA	NA	0.044	0.044	0.037	0.068	0.044	0.134	0.044	0.045	0.047
Avg-Cov	1.000	1.000	1.000	0.850	0.890	0.873	NA	NA	NA	0.955	0.952	0.960	0.854	0.855	0.494	0.906	0.932	0.941
Avg-Len	0.718	0.750	0.568	0.244	0.242	0.203	NA	NA	NA	0.214	0.213	0.182	0.236	0.158	31.767	0.180	0.208	0.222
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																	
Avg-Bias	0.166	0.170	0.089	0.167	0.126	0.090	NA	NA	NA	0.343	0.218	0.159	0.159	0.156	0.119	0.408	0.295	0.254
Avg-Cov	1.000	1.000	1.000	0.682	0.555	0.684	NA	NA	NA	0.626	0.583	0.613	0.742	0.851	0.487	0.664	0.866	0.945
Avg-Len	3.585	3.681	2.527	0.417	0.242	0.229	NA	NA	NA	0.728	0.460	0.359	0.447	0.582	12.082	1.004	1.161	1.238
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																	
Avg-Bias	0.068	0.068	0.054	0.138	0.105	0.103	NA	NA	NA	0.064	0.061	0.050	0.114	0.063	0.146	0.075	0.062	0.059
Avg-Cov	1.000	1.000	1.000	0.458	0.579	0.528	NA	NA	NA	0.887	0.880	0.886	0.622	0.852	0.482	0.793	0.906	0.941
Avg-Len	1.289	1.321	1.020	0.260	0.234	0.210	NA	NA	NA	0.252	0.235	0.203	0.264	0.224	20.412	0.237	0.274	0.292
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																	
Avg-Bias	0.027	0.027	0.020	0.037	0.035	0.029	NA	NA	NA	0.001	0.003	0.001	0.033	0.005	0.023	0.006	0.007	0.009
Avg-Cov	1.000	1.000	1.000	0.946	0.955	0.951	NA	NA	NA	0.892	0.630	0.889	0.950	0.000	0.519	0.896	0.903	0.960
Avg-Len	2.545	2.584	1.879	0.183	0.182	0.142	NA	NA	NA	0.006	0.007	0.005	0.165	0.000	7.839	0.030	0.034	0.044

Table D.2: Simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, Gaussian noise $\mathcal{N}(0, 1)$, and the MAR setting (5.23). Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "R-Proj" or "SSL-AIPW" and are thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0.1}^{(a)}$																		
Avg-Bias	0.033	0.036	0.028	0.039	0.038	0.031	0.070	0.063	0.492	0.034	0.039	0.028	0.039	0.039	0.189	0.043	0.040	0.039
Avg-Cov	0.969	0.961	0.964	0.905	0.881	0.924	0.969	0.969	0.928	0.945	0.898	0.948	0.911	0.895	0.485	0.859	0.901	0.911
Avg-Len	0.176	0.185	0.145	0.162	0.152	0.133	0.334	0.329	0.926	0.162	0.156	0.133	0.166	0.154	6.343	0.151	0.161	0.167
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0.1}^{(b)}$																		
Avg-Bias	0.148	0.159	0.108	0.183	0.163	0.133	0.093	0.207	0.088	0.178	0.193	0.143	0.160	0.148	0.223	0.227	0.198	0.180
Avg-Cov	0.994	0.991	0.997	0.892	0.851	0.917	0.998	0.999	0.996	0.918	0.821	0.909	0.372	0.914	0.511	0.846	0.917	0.958
Avg-Len	0.982	0.998	0.844	0.758	0.595	0.582	1.250	1.304	1.729	0.792	0.666	0.610	0.185	0.628	6.032	0.821	0.873	0.908
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0.1}^{(c)}$																		
Avg-Bias	0.045	0.049	0.035	0.053	0.050	0.040	0.056	0.078	0.083	0.050	0.053	0.038	0.049	0.053	0.378	0.063	0.058	0.055
Avg-Cov	0.988	0.981	0.993	0.897	0.859	0.906	0.992	0.981	0.980	0.926	0.858	0.928	0.890	0.906	0.472	0.868	0.918	0.945
Avg-Len	0.285	0.293	0.236	0.223	0.190	0.173	0.387	0.362	0.543	0.223	0.199	0.171	0.197	0.219	7.248	0.241	0.256	0.267
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.041	0.041	0.036	0.068	0.055	0.050	NA	NA	NA	0.036	0.043	0.029	0.060	0.039	0.288	0.041	0.041	0.045
Avg-Cov	0.983	0.995	0.991	0.827	0.867	0.854	NA	NA	NA	0.945	0.888	0.944	0.875	0.912	0.491	0.910	0.922	0.930
Avg-Len	0.262	0.312	0.236	0.215	0.202	0.172	NA	NA	NA	0.175	0.170	0.144	0.216	0.170	10.974	0.173	0.183	0.201
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.179	0.195	0.130	0.352	0.230	0.225	NA	NA	NA	0.244	0.284	0.184	0.216	0.168	0.364	0.226	0.217	0.194
Avg-Cov	1.00	1.00	1.00	0.766	0.831	0.828	NA	NA	NA	0.861	0.716	0.877	0.342	0.903	0.505	0.904	0.933	0.978
Avg-Len	1.659	1.860	1.513	1.008	0.790	0.754	NA	NA	NA	0.935	0.791	0.727	0.240	0.692	8.013	0.941	0.992	1.091
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.054	0.056	0.045	0.102	0.072	0.068	NA	NA	NA	0.061	0.070	0.047	0.063	0.059	0.505	0.070	0.068	0.065
Avg-Cov	0.998	0.999	0.999	0.765	0.835	0.815	NA	NA	NA	0.912	0.828	0.922	0.907	0.900	0.490	0.890	0.909	0.956
Avg-Len	0.467	0.532	0.413	0.297	0.252	0.224	NA	NA	NA	0.267	0.238	0.205	0.258	0.244	9.533	0.277	0.291	0.321
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0.100}^{(a)}$																		
Avg-Bias	0.025	0.025	0.021	0.032	0.032	0.026	0.430	1.097	0.407	0.000	0.001	0.000	0.033	0.030	0.100	0.033	0.034	0.036
Avg-Cov	0.988	0.993	0.986	0.958	0.941	0.948	0.960	0.958	0.930	1.00	0.994	1.00	0.953	0.963	0.502	0.952	0.951	0.948
Avg-Len	0.176	0.185	0.145	0.162	0.152	0.133	1.138	14.143	0.509	0.000	0.001	0.000	0.163	0.151	10.364	0.151	0.161	0.168
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0.100}^{(b)}$																		
Avg-Bias	0.105	0.105	0.087	0.170	0.134	0.127	0.134	0.108	0.226	0.191	0.192	0.190	0.137	0.140	0.111	0.200	0.180	0.168
Avg-Cov	0.996	0.997	0.999	0.924	0.910	0.932	0.998	0.999	0.995	0.001	0.044	0.005	0.352	0.923	0.512	0.900	0.935	0.964
Avg-Len	0.983	0.998	0.844	0.758	0.595	0.583	1.638	52.948	2.199	0.017	0.051	0.020	0.184	0.625	6.294	0.820	0.873	0.910
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0.100}^{(c)}$																		
Avg-Bias	0.033	0.032	0.027	0.049	0.042	0.037	0.433	0.159	0.350	0.040	0.041	0.040	0.040	0.046	0.085	0.055	0.053	0.052
Avg-Cov	0.997	0.996	0.997	0.930	0.918	0.932	0.998	0.989	0.984	0.001	0.015	0.000	0.932	0.938	0.511	0.915	0.934	0.953
Avg-Len	0.285	0.292	0.236	0.223	0.190	0.173	5.491	1.625	0.909	0.001	0.006	0.002	0.195	0.218	7.371	0.241	0.256	0.267
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.034	0.035	0.029	0.047	0.042	0.036	NA	NA	NA	0.034	0.039	0.027	0.044	0.036	0.260	0.036	0.036	0.036
Avg-Cov	0.976	0.989	0.983	0.880	0.902	0.898	NA	NA	NA	0.942	0.889	0.946	0.905	0.904	0.483	0.897	0.927	0.933
Avg-Len	0.197	0.234	0.180	0.174	0.164	0.141	NA	NA	NA	0.161	0.156	0.133	0.178	0.153	7.641	0.152	0.161	0.168
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.148	0.161	0.109	0.230	0.171	0.157	NA	NA	NA	0.184	0.220	0.148	0.168	0.140	0.287	0.211	0.190	0.177
Avg-Cov	1.00	1.00	1.00	0.852	0.870	0.883	NA	NA	NA	0.897	0.768	0.894	0.373	0.908	0.497	0.887	0.937	0.963
Avg-Len	1.214	1.355	1.122	0.815	0.642	0.620	NA	NA	NA	0.791	0.665	0.609	0.197	0.625	6.076	0.823	0.876	0.911
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.044	0.048	0.036	0.067	0.053	0.047	NA	NA	NA	0.050	0.056	0.039	0.050	0.052	0.438	0.067	0.061	0.058
Avg-Cov	0.996	0.997	0.998	0.849	0.870	0.882	NA	NA	NA	0.925	0.832	0.921	0.895	0.907	0.468	0.854	0.924	0.941
Avg-Len	0.344	0.391	0.309	0.240	0.205	0.184	NA	NA	NA	0.222	0.199	0.171	0.211	0.219	7.349	0.242	0.257	0.268
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.036	0.035	0.030	0.042	0.042	0.031	NA	NA	NA	0.002	0.007	0.002	0.039	0.016	0.073	0.018	0.017	0.021
Avg-Cov	1.00	1.00	1.00	0.942	0.936	0.950	NA	NA	NA	0.890	0.557	0.880	0.948	0.000	0.485	0.000	0.000	0.950
Avg-Len	1.761	2.029	1.367	0.203	0.197	0.154	NA	NA	NA	0.010	0.015	0.008	0.190	0.000	6.109	0.000	0.006	0.103

Table D.3: Simulation results under the circulant symmetric covariance matrix Σ^{cs}, Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$-distributed noise, and the MAR setting (5.23). Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "R-Proj" or "SSL-AIPW" and are thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																		
Avg-Bias	0.099	0.103	0.086	0.118	0.112	0.098	0.189	0.198	0.452	0.098	0.110	0.085	0.135	0.108	0.329	0.125	0.117	0.110
Avg-Cov	0.948	0.957	0.962	0.888	0.890	0.902	0.948	0.945	0.923	0.957	0.911	0.947	0.546	0.916	0.518	0.868	0.919	0.939
Avg-Len	0.516	0.529	0.439	0.474	0.444	0.403	0.858	1.144	1.629	0.480	0.462	0.403	0.239	0.465	27.737	0.471	0.501	0.527
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																		
Avg-Bias	0.183	0.188	0.142	0.220	0.198	0.170	0.177	0.177	0.471	0.218	0.228	0.176	0.233	0.185	0.312	0.259	0.227	0.206
Avg-Cov	0.990	0.990	0.995	0.889	0.862	0.917	0.999	0.993	0.993	0.924	0.851	0.913	0.313	0.917	0.483	0.874	0.930	0.957
Avg-Len	1.194	1.205	1.049	0.910	0.753	0.726	1.487	1.301	2.202	0.935	0.814	0.741	0.233	0.787	19.773	0.965	1.026	1.075
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																		
Avg-Bias	0.110	0.114	0.091	0.129	0.120	0.103	0.168	0.174	0.298	0.117	0.123	0.098	0.141	0.119	0.446	0.140	0.129	0.121
Avg-Cov	0.982	0.978	0.983	0.889	0.884	0.904	0.970	0.960	0.960	0.934	0.892	0.944	0.575	0.912	0.511	0.865	0.918	0.945
Avg-Len	0.627	0.640	0.537	0.522	0.474	0.435	0.905	0.873	1.366	0.527	0.499	0.434	0.267	0.502	21.064	0.537	0.572	0.601
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.128	0.126	0.110	0.206	0.169	0.152	NA	NA	NA	0.106	0.118	0.092	0.167	0.121	0.429	0.121	0.122	0.128
Avg-Cov	0.983	0.996	0.993	0.805	0.851	0.847	NA	NA	NA	0.953	0.895	0.953	0.579	0.920	0.488	0.928	0.935	0.943
Avg-Len	0.769	0.893	0.714	0.631	0.590	0.522	NA	NA	NA	0.520	0.503	0.437	0.308	0.509	34.504	0.540	0.570	0.626
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.224	0.234	0.178	0.437	0.293	0.293	NA	NA	NA	0.309	0.342	0.228	0.305	0.210	0.450	0.257	0.249	0.233
Avg-Cov	0.999	1.00	1.00	0.755	0.820	0.813	NA	NA	NA	0.838	0.756	0.887	0.339	0.894	0.506	0.913	0.943	0.968
Avg-Len	1.969	2.201	1.851	1.211	1.002	0.940	NA	NA	NA	1.087	0.955	0.876	0.302	0.865	27.198	1.107	1.167	1.283
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.148	0.147	0.125	0.265	0.202	0.185	NA	NA	NA	0.146	0.158	0.123	0.184	0.139	0.582	0.143	0.141	0.144
Avg-Cov	0.996	1.00	0.997	0.713	0.794	0.785	NA	NA	NA	0.921	0.862	0.931	0.585	0.892	0.512	0.903	0.931	0.951
Avg-Len	0.983	1.122	0.909	0.695	0.630	0.564	NA	NA	NA	0.624	0.592	0.516	0.345	0.552	33.809	0.617	0.651	0.715
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																		
Avg-Bias	0.075	0.074	0.067	0.094	0.093	0.082	0.810	0.566	0.628	0.000	0.003	0.000	0.081	0.099	0.297	0.097	0.102	0.106
Avg-Cov	0.986	0.985	0.988	0.943	0.943	0.949	0.959	0.957	0.940	1.00	0.993	1.00	0.778	0.951	0.528	0.955	0.953	0.957
Avg-Len	0.516	0.529	0.440	0.473	0.443	0.403	5.314	2.157	2.126	0.000	0.004	0.000	0.236	0.467	25.547	0.471	0.502	0.522
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																		
Avg-Bias	0.133	0.130	0.115	0.207	0.173	0.159	0.174	0.196	0.396	0.185	0.195	0.188	0.188	0.181	0.233	0.227	0.211	0.204
Avg-Cov	0.998	0.997	0.996	0.928	0.912	0.927	0.999	0.996	0.995	0.000	0.028	0.001	0.316	0.920	0.509	0.908	0.951	0.972
Avg-Len	1.193	1.203	1.049	0.909	0.752	0.726	1.757	2.160	2.170	0.004	0.045	0.011	0.231	0.788	18.631	0.966	1.028	1.070
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																		
Avg-Bias	0.082	0.080	0.071	0.109	0.103	0.089	0.271	0.523	0.269	0.040	0.042	0.040	0.089	0.109	0.276	0.114	0.117	0.120
Avg-Cov	0.997	0.996	0.995	0.944	0.936	0.947	0.974	0.969	0.963	0.000	0.000	0.000	0.805	0.933	0.487	0.943	0.954	0.963
Avg-Len	0.627	0.639	0.538	0.522	0.473	0.436	7.329	6.314	1.360	0.001	0.004	0.001	0.263	0.505	27.951	0.538	0.573	0.596
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.127	0.124	0.114	0.231	0.179	0.173	NA	NA	NA	0.108	0.125	0.095	0.173	0.132	0.487	0.167	0.143	0.132
Avg-Cov	1.00	1.00	1.00	0.805	0.871	0.846	NA	NA	NA	0.951	0.883	0.952	0.588	0.846	0.510	0.774	0.854	0.908
Avg-Len	1.585	1.803	1.360	0.680	0.644	0.556	NA	NA	NA	0.520	0.503	0.436	0.329	0.477	32.541	0.505	0.533	0.577
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.263	0.234	0.274	1.053	0.599	0.738	NA	NA	NA	0.887	0.861	0.708	0.363	0.641	0.666	1.335	1.120	0.803
Avg-Cov	1.00	1.00	1.00	0.078	0.447	0.147	NA	NA	NA	0.109	0.090	0.103	0.269	0.174	0.498	0.008	0.032	0.192
Avg-Len	4.620	5.043	4.105	1.306	1.093	1.003	NA	NA	NA	1.018	0.891	0.810	0.327	0.809	26.129	1.035	1.091	1.182
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.214	0.179	0.224	0.519	0.370	0.381	NA	NA	NA	0.284	0.277	0.234	0.195	0.312	0.639	0.563	0.456	0.326
Avg-Cov	1.00	1.00	1.00	0.182	0.431	0.265	NA	NA	NA	0.522	0.485	0.568	0.583	0.341	0.488	0.035	0.149	0.462
Avg-Len	2.169	2.416	1.875	0.749	0.687	0.602	NA	NA	NA	0.574	0.546	0.474	0.374	0.516	28.206	0.576	0.608	0.658
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.108	0.105	0.092	0.129	0.125	0.098	NA	NA	NA	0.008	0.020	0.007	0.087	0.031	0.210	0.053	0.051	0.060
Avg-Cov	1.00	1.00	1.00	0.948	0.941	0.941	NA	NA	NA	0.880	0.585	0.885	0.825	0.000	0.461	0.000	0.000	0.961
Avg-Len	5.186	5.798	4.133	0.596	0.573	0.468	NA	NA	NA	0.029	0.042	0.024	0.279	0.000	19.908	0.000	0.018	0.308

Table D.4: Simulation results under the circulant symmetric covariance matrix Σ^{cs}, t_2-distributed noise, and the MAR setting (5.23). Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "R-Proj" and "SSL-AIPW" and are thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																		
Avg-Bias	0.071	0.072	0.060	0.070	0.071	0.060	0.545	0.738	0.250	0.073	0.074	0.061	0.078	0.076	0.215	0.074	0.076	0.088
Avg-Cov	0.887	0.888	0.907	0.915	0.913	0.933	0.939	0.946	0.944	0.945	0.940	0.947	0.886	0.902	0.490	0.881	0.914	0.942
Avg-Len	0.279	0.281	0.247	0.308	0.305	0.266	0.967	1.735	2.550	0.353	0.351	0.300	0.307	0.319	91.508	0.295	0.325	0.412
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																		
Avg-Bias	0.287	0.290	0.157	0.278	0.227	0.160	0.205	0.236	0.377	0.411	0.246	0.176	0.249	0.217	0.164	0.457	0.394	0.363
Avg-Cov	0.747	0.746	0.909	0.557	0.426	0.552	0.937	0.774	0.953	0.781	0.782	0.815	0.642	0.967	0.464	0.845	0.924	0.982
Avg-Len	0.828	0.842	0.672	0.528	0.305	0.301	0.996	0.625	1.368	1.166	0.761	0.593	0.574	1.189	52.275	1.648	1.816	2.302
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																		
Avg-Bias	0.094	0.095	0.074	0.098	0.095	0.076	0.318	0.127	0.251	0.097	0.095	0.080	0.102	0.093	0.218	0.102	0.096	0.103
Avg-Cov	0.857	0.856	0.893	0.826	0.794	0.851	0.940	0.927	0.931	0.917	0.895	0.915	0.823	0.946	0.511	0.866	0.911	0.956
Avg-Len	0.341	0.342	0.300	0.329	0.296	0.276	0.765	0.569	0.879	0.415	0.387	0.334	0.340	0.456	27.226	0.389	0.429	0.543
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.057	0.058	0.048	0.067	0.066	0.056	NA	NA	NA	0.050	0.051	0.043	0.067	0.055	0.197	0.059	0.064	0.075
Avg-Cov	0.987	0.992	0.991	0.922	0.919	0.935	NA	NA	NA	0.942	0.930	0.936	0.925	0.888	0.509	0.916	0.929	0.937
Avg-Len	0.353	0.387	0.319	0.290	0.287	0.248	NA	NA	NA	0.237	0.237	0.202	0.290	0.216	23.525	0.255	0.286	0.347
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.233	0.240	0.128	0.237	0.197	0.138	NA	NA	NA	0.377	0.247	0.179	0.229	0.214	0.154	0.390	0.341	0.328
Avg-Cov	0.987	0.994	1.00	0.584	0.433	0.587	NA	NA	NA	0.683	0.688	0.757	0.648	0.866	0.478	0.866	0.941	0.981
Avg-Len	1.366	1.456	1.072	0.497	0.287	0.281	NA	NA	NA	0.999	0.652	0.519	0.545	0.802	19.405	1.423	1.599	1.939
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.081	0.082	0.066	0.089	0.085	0.071	NA	NA	NA	0.088	0.085	0.071	0.090	0.087	0.196	0.090	0.086	0.092
Avg-Cov	0.991	0.995	0.994	0.830	0.804	0.849	NA	NA	NA	0.909	0.894	0.899	0.850	0.824	0.495	0.857	0.918	0.949
Avg-Len	0.522	0.556	0.456	0.310	0.278	0.257	NA	NA	NA	0.373	0.348	0.301	0.322	0.309	30.022	0.336	0.377	0.458
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																		
Avg-Bias	0.056	0.055	0.053	0.074	0.073	0.066	0.700	1.163	0.562	0.000	0.000	0.000	0.077	0.090	0.268	0.074	0.085	0.097
Avg-Cov	0.986	0.984	0.980	0.957	0.948	0.947	0.952	0.948	0.948	1.00	0.999	1.00	0.952	0.961	0.503	0.953	0.951	0.953
Avg-Len	0.347	0.350	0.312	0.367	0.363	0.322	3.483	15.279	1.778	0.000	0.000	0.000	0.384	0.461	140.571	0.373	0.427	0.479
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																		
Avg-Bias	0.171	0.171	0.125	0.220	0.186	0.165	0.175	0.333	0.218	0.250	0.215	0.198	0.245	0.304	0.188	0.378	0.352	0.371
Avg-Cov	0.978	0.980	0.985	0.777	0.561	0.621	0.993	0.893	0.975	0.237	0.352	0.371	0.800	0.975	0.490	0.954	0.986	0.997
Avg-Len	1.054	1.071	0.857	0.629	0.363	0.364	1.287	0.803	1.125	0.384	0.370	0.329	0.723	1.716	73.685	2.081	2.386	2.676
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																		
Avg-Bias	0.069	0.068	0.064	0.090	0.085	0.077	0.129	0.212	0.345	0.056	0.055	0.053	0.101	0.118	0.242	0.095	0.098	0.108
Avg-Cov	0.974	0.977	0.973	0.911	0.904	0.904	0.984	0.966	0.959	0.302	0.330	0.363	0.912	0.973	0.482	0.954	0.970	0.977
Avg-Len	0.431	0.432	0.382	0.392	0.352	0.333	0.793	1.870	1.138	0.134	0.142	0.135	0.426	0.659	39.055	0.491	0.563	0.631
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.047	0.048	0.040	0.065	0.059	0.052	NA	NA	NA	0.045	0.045	0.038	0.065	0.046	0.138	0.046	0.048	0.049
Avg-Cov	1.00	1.00	1.00	0.876	0.909	0.893	NA	NA	NA	0.934	0.929	0.948	0.868	0.837	0.504	0.891	0.922	0.932
Avg-Len	0.717	0.752	0.570	0.244	0.242	0.203	NA	NA	NA	0.214	0.213	0.182	0.237	0.157	60.803	0.179	0.207	0.221
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.174	0.176	0.094	0.169	0.129	0.095	NA	NA	NA	0.340	0.214	0.159	0.156	0.153	0.116	0.403	0.296	0.257
Avg-Cov	1.00	1.00	1.00	0.673	0.548	0.651	NA	NA	NA	0.632	0.575	0.614	0.738	0.870	0.490	0.647	0.872	0.943
Avg-Len	3.580	3.675	2.529	0.417	0.242	0.230	NA	NA	NA	0.731	0.462	0.359	0.446	0.582	36.161	1.002	1.158	1.236
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.066	0.066	0.054	0.133	0.100	0.101	NA	NA	NA	0.064	0.061	0.052	0.111	0.062	0.144	0.074	0.063	0.060
Avg-Cov	1.00	1.00	1.00	0.509	0.628	0.538	NA	NA	NA	0.873	0.867	0.883	0.637	0.844	0.506	0.799	0.904	0.938
Avg-Len	1.290	1.324	1.020	0.260	0.234	0.210	NA	NA	NA	0.251	0.234	0.203	0.265	0.224	18.848	0.236	0.273	0.292
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.028	0.028	0.021	0.038	0.036	0.029	NA	NA	NA	0.002	0.003	0.001	0.034	0.005	0.024	0.006	0.007	0.009
Avg-Cov	1.00	1.00	1.00	0.945	0.962	0.954	NA	NA	NA	0.859	0.653	0.885	0.947	0.000	0.502	0.891	0.895	0.953
Avg-Len	2.542	2.592	1.889	0.183	0.182	0.142	NA	NA	NA	0.006	0.007	0.005	0.165	0.000	25.538	0.030	0.034	0.044

Table D.5: Simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$-distributed noise, and the MAR setting (5.23). Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "R-Proj" or "SSL-AIPW" and are thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																		
Avg-Bias	0.213	0.211	0.186	0.218	0.218	0.191	0.497	0.569	0.939	0.228	0.227	0.195	0.233	0.244	0.636	0.233	0.235	0.267
Avg-Cov	0.879	0.886	0.907	0.910	0.917	0.920	0.944	0.939	0.956	0.939	0.933	0.949	0.839	0.912	0.517	0.897	0.911	0.937
Avg-Len	0.816	0.820	0.749	0.903	0.894	0.808	2.284	2.156	3.056	1.068	1.060	0.918	0.796	1.052	74.236	0.926	1.020	1.294
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																		
Avg-Bias	0.382	0.380	0.263	0.423	0.373	0.291	0.409	0.369	0.391	0.556	0.399	0.328	0.349	0.336	0.520	0.557	0.488	0.485
Avg-Cov	0.802	0.814	0.884	0.678	0.614	0.729	0.950	0.887	0.961	0.779	0.815	0.843	0.660	0.937	0.505	0.861	0.944	0.988
Avg-Len	1.234	1.241	1.049	1.097	0.842	0.845	2.024	1.589	2.093	1.519	1.277	1.103	0.847	1.583	68.001	2.110	2.324	2.948
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																		
Avg-Bias	0.230	0.229	0.196	0.239	0.236	0.204	0.326	0.428	0.395	0.260	0.256	0.216	0.246	0.255	0.620	0.262	0.258	0.286
Avg-Cov	0.874	0.874	0.895	0.888	0.874	0.895	0.950	0.942	0.956	0.929	0.925	0.938	0.808	0.915	0.517	0.882	0.921	0.947
Avg-Len	0.873	0.875	0.793	0.925	0.884	0.822	1.648	1.723	2.003	1.129	1.097	0.955	0.795	1.110	104.853	1.014	1.116	1.416
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.166	0.167	0.150	0.200	0.194	0.174	NA	NA	NA	0.162	0.159	0.133	0.180	0.182	0.599	0.187	0.199	0.227
Avg-Cov	0.985	0.993	0.994	0.918	0.928	0.938	NA	NA	NA	0.942	0.939	0.953	0.917	0.889	0.495	0.919	0.941	0.953
Avg-Len	1.031	1.126	0.963	0.851	0.841	0.752	NA	NA	NA	0.723	0.722	0.621	0.743	0.716	147.268	0.799	0.901	1.093
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.320	0.321	0.219	0.351	0.317	0.252	NA	NA	NA	0.474	0.366	0.294	0.296	0.289	0.493	0.458	0.409	0.407
Avg-Cov	0.982	0.986	0.996	0.740	0.664	0.778	NA	NA	NA	0.729	0.742	0.790	0.717	0.863	0.472	0.888	0.958	0.988
Avg-Len	1.894	2.012	1.562	1.034	0.792	0.787	NA	NA	NA	1.345	1.104	0.964	0.797	1.073	105.389	1.821	2.050	2.487
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.199	0.200	0.174	0.225	0.218	0.195	NA	NA	NA	0.227	0.219	0.187	0.208	0.214	0.579	0.226	0.226	0.245
Avg-Cov	0.983	0.992	0.988	0.880	0.876	0.896	NA	NA	NA	0.925	0.915	0.936	0.860	0.829	0.486	0.897	0.936	0.960
Avg-Len	1.190	1.284	1.082	0.872	0.831	0.765	NA	NA	NA	0.970	0.948	0.836	0.744	0.755	189.113	0.874	0.986	1.196
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																		
Avg-Bias	0.167	0.169	0.156	0.224	0.224	0.194	0.818	0.814	0.862	0.000	0.001	0.000	0.212	0.353	0.914	0.247	0.282	0.318
Avg-Cov	0.985	0.982	0.982	0.943	0.936	0.943	0.957	0.953	0.955	1.00	0.997	1.00	0.946	0.940	0.500	0.942	0.945	0.940
Avg-Len	1.017	1.021	0.943	1.076	1.065	0.977	3.865	3.571	3.892	0.000	0.001	0.000	0.987	1.507	183.936	1.170	1.344	1.506
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																		
Avg-Bias	0.246	0.247	0.215	0.345	0.301	0.277	0.413	0.560	0.513	0.268	0.253	0.240	0.306	0.484	0.732	0.483	0.465	0.507
Avg-Cov	0.976	0.979	0.978	0.867	0.811	0.863	0.989	0.949	0.960	0.239	0.325	0.327	0.843	0.963	0.495	0.968	0.992	0.996
Avg-Len	1.563	1.570	1.332	1.307	1.003	1.022	2.625	2.187	2.728	0.481	0.512	0.454	1.058	2.279	81.276	2.665	3.058	3.428
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																		
Avg-Bias	0.175	0.176	0.165	0.238	0.229	0.209	0.406	0.495	0.521	0.071	0.070	0.066	0.223	0.368	0.880	0.259	0.285	0.319
Avg-Cov	0.982	0.978	0.982	0.931	0.928	0.934	0.965	0.959	0.948	0.139	0.181	0.171	0.931	0.936	0.516	0.950	0.957	0.960
Avg-Len	1.094	1.096	1.002	1.103	1.052	0.994	2.628	2.263	2.614	0.140	0.170	0.144	0.988	1.593	143.525	1.280	1.470	1.647
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.147	0.147	0.126	0.204	0.185	0.169	NA	NA	NA	0.143	0.138	0.119	0.143	0.143	0.394	0.141	0.144	0.145
Avg-Cov	1.00	1.00	1.00	0.873	0.903	0.869	NA	NA	NA	0.945	0.946	0.950	0.921	0.830	0.495	0.894	0.933	0.942
Avg-Len	2.091	2.176	1.710	0.715	0.708	0.616	NA	NA	NA	0.650	0.650	0.561	0.607	0.513	386.621	0.563	0.650	0.695
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.235	0.233	0.167	0.385	0.272	0.284	NA	NA	NA	0.420	0.301	0.235	0.215	0.207	0.327	0.448	0.339	0.302
Avg-Cov	1.00	1.00	1.00	0.631	0.686	0.623	NA	NA	NA	0.689	0.716	0.763	0.763	0.846	0.494	0.733	0.904	0.967
Avg-Len	4.691	4.776	3.413	0.868	0.667	0.644	NA	NA	NA	0.985	0.797	0.684	0.650	0.772	41.916	1.283	1.483	1.582
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.202	0.200	0.168	0.380	0.310	0.314	NA	NA	NA	0.166	0.159	0.133	0.164	0.151	0.385	0.168	0.160	0.157
Avg-Cov	0.999	0.999	1.00	0.492	0.608	0.514	NA	NA	NA	0.920	0.913	0.925	0.855	0.836	0.522	0.863	0.927	0.946
Avg-Len	2.638	2.722	2.106	0.732	0.700	0.626	NA	NA	NA	0.686	0.670	0.583	0.608	0.541	64.975	0.616	0.711	0.760
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.079	0.079	0.064	0.108	0.102	0.091	NA	NA	NA	0.005	0.009	0.004	0.042	0.017	0.071	0.020	0.022	0.027
Avg-Cov	1.00	1.00	1.00	0.949	0.963	0.943	NA	NA	NA	0.866	0.657	0.881	1.00	0.000	0.515	0.908	0.913	0.957
Avg-Len	7.400	7.480	5.642	0.537	0.533	0.430	NA	NA	NA	0.017	0.021	0.015	0.435	0.000	30.483	0.096	0.107	0.139

Table D.6: Simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, t_2-distributed noise, and the MAR setting in Section 5.4.1. Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "R-Proj" or "SSL-AIPW" and are thus omitted.

	Oracle (1SE)	Oracle (min-feas)	LR (1SE)	LR (min-feas)	NB (1SE)	NB (min-feas)	RF (1SE)	RF (min-feas)	SVM (1SE)	SVM (min-feas)	NN (1SE)	NN (min-feas)
Avg-MAE	0	0	0.334	0.334	0.121	0.121	0.158	0.158	0.321	0.321	**0.106**	**0.106**
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$											
Avg-Bias	0.034	0.033	0.048	0.047	0.035	0.034	0.037	0.036	0.045	0.044	0.034	**0.033**
Avg-Cov	0.902	0.928	0.815	0.824	0.899	0.924	0.885	0.918	0.828	0.846	0.901	**0.931**
Avg-Len	0.140	0.150	0.157	0.160	0.139	0.151	0.145	0.153	0.154	0.158	**0.134**	0.139
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$											
Avg-Bias	0.172	0.144	0.234	0.226	0.174	0.142	0.181	0.157	0.221	0.211	0.173	**0.138**
Avg-Cov	0.904	0.974	0.871	0.887	0.899	0.977	0.904	**0.961**	0.875	0.899	0.902	0.978
Avg-Len	0.747	0.800	0.842	0.858	0.746	0.807	0.776	0.820	0.826	0.843	**0.744**	0.810
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$											
Avg-Bias	0.050	0.045	0.068	0.066	0.051	0.045	0.053	0.049	0.064	0.062	0.050	**0.044**
Avg-Cov	0.913	0.966	0.861	0.879	0.909	0.968	0.911	**0.959**	0.873	0.889	0.912	0.973
Avg-Len	0.218	0.233	0.245	0.250	**0.217**	0.235	0.226	0.239	0.241	0.246	**0.217**	0.236
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$											
Avg-Bias	0.035	0.037	0.050	0.053	**0.035**	0.038	0.036	0.038	0.047	0.050	0.036	0.038
Avg-Cov	0.924	0.945	0.845	0.850	0.924	0.940	0.937	**0.945**	0.864	0.867	0.923	0.940
Avg-Len	0.159	0.179	0.176	0.191	**0.160**	0.182	0.164	0.182	0.172	0.187	**0.160**	0.181
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$											
Avg-Bias	0.184	0.150	0.305	0.303	0.180	0.147	0.188	0.160	0.276	0.271	0.181	**0.146**
Avg-Cov	0.940	0.991	0.778	0.853	0.951	0.995	0.938	0.987	0.819	0.889	**0.950**	0.992
Avg-Len	0.853	0.960	0.939	1.019	0.857	0.974	0.877	0.973	0.923	1.002	**0.855**	0.969
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$											
Avg-Bias	0.056	0.051	0.085	0.088	0.054	**0.051**	0.055	0.052	0.078	0.079	0.055	**0.051**
Avg-Cov	0.922	0.972	0.813	0.837	0.923	0.976	0.931	**0.967**	0.836	0.866	0.920	0.968
Avg-Len	0.249	0.280	0.274	0.297	0.250	0.284	0.256	0.284	0.269	0.292	**0.249**	0.282
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$											
Avg-Bias	0.0307	0.033	0.033	0.034	**0.031**	0.033	**0.031**	0.032	0.032	0.033	**0.031**	0.034
Avg-Cov	0.946	0.939	0.939	0.936	**0.948**	0.938	0.945	0.944	0.941	0.938	0.946	0.938
Avg-Len	0.1485	0.161	0.158	0.161	**0.149**	0.162	0.150	0.160	0.155	0.159	**0.149**	0.164
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$											
Avg-Bias	0.164	0.139	0.162	0.155	0.163	0.137	0.162	0.138	0.155	0.149	0.162	**0.132**
Avg-Cov	0.945	0.991	0.967	0.976	0.947	0.991	0.947	0.989	0.968	0.978	**0.950**	0.993
Avg-Len	0.794	0.859	0.843	0.862	**0.795**	0.867	0.804	0.858	0.829	0.849	0.796	0.875
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$											
Avg-Bias	0.047	0.045	0.049	0.048	0.047	0.046	0.047	**0.045**	0.048	0.047	0.047	**0.045**
Avg-Cov	0.956	0.973	0.966	0.967	**0.957**	0.974	0.963	0.974	0.966	0.967	0.959	0.975
Avg-Len	0.232	0.250	0.246	0.251	**0.232**	0.253	0.234	0.250	0.242	0.247	**0.232**	0.255
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$											
Avg-Bias	0.030	0.031	0.041	0.041	**0.030**	0.031	0.031	0.032	0.038	0.038	**0.030**	0.031
Avg-Cov	0.938	0.949	0.879	0.883	0.940	**0.949**	0.944	0.945	0.898	0.903	0.938	0.948
Avg-Len	0.1392	0.149	0.158	0.161	**0.139**	0.150	0.145	0.153	0.155	0.158	**0.139**	0.151
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$											
Avg-Bias	0.166	0.145	0.193	0.192	0.165	**0.139**	0.157	0.141	0.180	0.177	0.166	**0.139**
Avg-Cov	0.934	0.972	0.927	0.932	0.933	0.976	**0.945**	0.973	0.936	**0.945**	0.935	0.979
Avg-Len	0.745	0.798	0.845	0.860	0.744	0.804	0.775	0.818	0.829	0.845	**0.742**	0.806
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$											
Avg-Bias	0.050	0.046	0.055	0.056	0.050	**0.045**	0.047	**0.045**	0.053	0.052	0.050	**0.045**
Avg-Cov	0.915	0.960	0.931	0.932	0.918	0.970	**0.943**	0.971	0.932	0.937	0.917	0.967
Avg-Len	0.217	0.232	0.246	0.251	0.217	0.234	0.226	0.238	0.242	0.246	**0.216**	0.235

Table D.7: Simulation results for our proposed debiasing method when the propensity scores are estimated by various machine learning methods under the circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0, 1)$, and the new MAR setting (5.24). Note that the "Avg-MAE" for propensity score estimation is independent of the choices of β_0 and x, so we place it on the top row. We also highlight the best results in terms of the evaluation metrics for each simulation design, excluding those with the oracle propensity scores, in boldface.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL-AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																		
Avg-Bias	0.033	0.033	0.028	0.039	0.036	0.032	0.104	0.283	0.315	0.031	0.032	0.027	0.040	0.039	0.183	0.042	0.039	0.039
Avg-Cov	0.957	0.957	0.967	0.890	0.904	0.904	0.969	0.955	0.940	0.957	0.952	0.953	0.895	0.879	0.513	0.848	0.905	0.912
Avg-Len	0.173	0.173	0.145	0.159	0.155	0.133	0.379	0.376	0.638	0.160	0.158	0.133	0.164	0.151	9.534	0.149	0.161	0.164
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																		
Avg-Bias	0.146	0.146	0.108	0.178	0.157	0.136	0.096	0.202	0.207	0.181	0.177	0.143	0.147	0.152	0.213	0.228	0.186	0.180
Avg-Cov	0.995	0.995	0.996	0.891	0.881	0.913	0.998	0.997	0.996	0.923	0.866	0.919	0.375	0.895	0.474	0.852	0.941	0.949
Avg-Len	0.970	0.971	0.844	0.742	0.616	0.583	1.255	2.215	1.784	0.776	0.680	0.610	0.175	0.611	4.591	0.808	0.872	0.888
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																		
Avg-Bias	0.044	0.044	0.036	0.053	0.047	0.041	0.059	0.104	0.088	0.047	0.045	0.037	0.047	0.053	0.369	0.063	0.055	0.054
Avg-Cov	0.986	0.987	0.996	0.890	0.889	0.908	0.989	0.985	0.985	0.940	0.937	0.938	0.893	0.890	0.472	0.856	0.937	0.942
Avg-Len	0.280	0.281	0.237	0.218	0.195	0.173	0.407	0.347	0.527	0.218	0.201	0.172	0.193	0.215	6.362	0.237	0.256	0.260
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																		
Avg-Bias	0.042	0.042	0.037	0.066	0.056	0.051	NA	NA	NA	0.034	0.035	0.030	0.065	0.038	0.283	0.037	0.038	0.041
Avg-Cov	0.986	0.991	0.990	0.812	0.860	0.839	NA	NA	NA	0.953	0.943	0.948	0.830	0.914	0.530	0.924	0.939	0.950
Avg-Len	0.258	0.286	0.236	0.209	0.204	0.172	NA	NA	NA	0.173	0.171	0.144	0.215	0.167	10.778	0.171	0.180	0.196
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																		
Avg-Bias	0.168	0.168	0.130	0.327	0.218	0.226	NA	NA	NA	0.239	0.249	0.184	0.187	0.170	0.347	0.223	0.214	0.193
Avg-Cov	1.00	1.00	1.00	0.774	0.862	0.821	NA	NA	NA	0.852	0.800	0.887	0.393	0.886	0.443	0.905	0.932	0.967
Avg-Len	1.640	1.759	1.514	0.977	0.811	0.755	NA	NA	NA	0.913	0.807	0.726	0.228	0.674	6.303	0.922	0.975	1.057
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																		
Avg-Bias	0.052	0.052	0.046	0.095	0.072	0.069	NA	NA	NA	0.059	0.059	0.046	0.062	0.058	0.489	0.066	0.064	0.061
Avg-Cov	0.999	1.00	1.00	0.792	0.864	0.806	NA	NA	NA	0.927	0.903	0.928	0.900	0.895	0.475	0.901	0.926	0.962
Avg-Len	0.461	0.498	0.414	0.287	0.256	0.224	NA	NA	NA	0.261	0.240	0.205	0.253	0.238	9.531	0.270	0.286	0.310
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																		
Avg-Bias	0.026	0.026	0.022	0.032	0.031	0.027	0.320	1.289	0.739	0.00	0.00	0.00	0.034	0.031	0.101	0.032	0.034	0.035
Avg-Cov	0.990	0.990	0.990	0.942	0.941	0.950	0.972	0.963	0.937	1.00	0.999	1.00	0.941	0.944	0.503	0.934	0.944	0.944
Avg-Len	0.173	0.173	0.145	0.159	0.155	0.133	2.341	5.584	1.177	0.00	0.00	0.00	0.161	0.149	11.692	0.149	0.161	0.165
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																		
Avg-Bias	0.102	0.102	0.090	0.165	0.134	0.128	0.122	0.196	0.084	0.186	0.196	0.191	0.128	0.143	0.101	0.200	0.175	0.170
Avg-Cov	0.999	0.999	0.999	0.923	0.921	0.929	0.998	0.997	0.997	0.00	0.040	0.004	0.364	0.906	0.512	0.897	0.957	0.965
Avg-Len	0.971	0.972	0.844	0.742	0.616	0.583	1.343	2.516	1.728	0.008	0.057	0.021	0.175	0.608	3.763	0.807	0.870	0.891
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																		
Avg-Bias	0.032	0.032	0.028	0.046	0.041	0.038	0.333	0.369	0.316	0.040	0.042	0.040	0.041	0.048	0.096	0.054	0.052	0.051
Avg-Cov	0.998	0.998	0.999	0.934	0.923	0.942	0.991	0.990	0.987	0.00	0.015	0.00	0.936	0.913	0.483	0.923	0.947	0.952
Avg-Len	0.281	0.281	0.237	0.218	0.195	0.173	2.153	1.071	28.765	0.001	0.008	0.002	0.192	0.213	7.919	0.237	0.255	0.261
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																		
Avg-Bias	0.042	0.042	0.037	0.075	0.061	0.056	NA	NA	NA	0.035	0.036	0.030	0.073	0.056	0.379	0.050	0.041	0.038
Avg-Cov	1.00	1.00	1.00	0.799	0.866	0.827	NA	NA	NA	0.942	0.932	0.945	0.825	0.729	0.514	0.813	0.900	0.938
Avg-Len	0.530	0.560	0.449	0.227	0.221	0.183	NA	NA	NA	0.173	0.171	0.144	0.232	0.157	11.657	0.160	0.171	0.181
$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																		
Avg-Bias	0.159	0.159	0.145	0.855	0.488	0.590	NA	NA	NA	0.787	0.716	0.620	0.235	0.617	0.576	1.244	0.982	0.755
Avg-Cov	1.00	1.00	1.00	0.067	0.413	0.156	NA	NA	NA	0.080	0.066	0.074	0.306	0.054	0.450	0.00	0.040	0.110
Avg-Len	3.939	4.073	3.449	1.060	0.880	0.806	NA	NA	NA	0.846	0.745	0.669	0.249	0.630	7.242	0.864	0.925	0.979
$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																		
Avg-Bias	0.076	0.075	0.082	0.239	0.166	0.170	NA	NA	NA	0.145	0.127	0.110	0.112	0.261	0.543	0.358	0.278	0.212
Avg-Cov	1.00	1.00	1.00	0.100	0.324	0.170	NA	NA	NA	0.363	0.399	0.398	0.676	0.013	0.448	0.00	0.061	0.161
Avg-Len	1.072	1.114	0.907	0.312	0.278	0.239	NA	NA	NA	0.239	0.221	0.189	0.285	0.223	7.460	0.253	0.271	0.287
$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																		
Avg-Bias	0.035	0.035	0.031	0.042	0.040	0.032	NA	NA	NA	0.002	0.005	0.002	0.037	0.015	0.075	0.017	0.017	0.021
Avg-Cov	1.00	1.00	1.00	0.940	0.945	0.939	NA	NA	NA	0.870	0.647	0.870	0.948	0.00	0.500	0.00	0.00	0.939
Avg-Len	1.724	1.752	1.367	0.199	0.194	0.154	NA	NA	NA	0.010	0.011	0.008	0.189	0.00	8.699	0.00	0.006	0.100

Table D.8: Simulation results under the circulant symmetric covariance matrix Σ^{cs}, Gaussian noise $\mathcal{N}(0, 1)$, and the MCAR setting. Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "DL-vdG", "R-Proj", and "SSL-AIPW" and thus omitted.

	DL-Jav (CC)	DL-Jav (IPW)	DL-Jav (Oracle)	DL-vdG (CC)	DL-vdG (IPW)	DL-vdG (Oracle)	R-Proj (CC)	R-Proj (IPW)	R-Proj (Oracle)	Refit (CC)	Refit (IPW)	Refit (Oracle)	DDR	SAS	SSL AIPW	**Debias** (min-CV)	**Debias** (1SE)	**Debias** (min-feas)
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{sp} = \beta_{0,1}^{(a)}$																	
Avg-Bias	0.072	0.072	0.061	0.073	0.073	0.061	0.275	0.444	0.608	0.073	0.074	0.060	0.074	0.075	0.223	0.074	0.076	0.089
Avg-Cov	0.885	0.884	0.894	0.912	0.906	0.917	0.963	0.955	0.950	0.948	0.944	0.956	0.916	0.905	0.536	0.883	0.914	0.929
Avg-Len	0.282	0.282	0.247	0.312	0.309	0.266	0.885	1.413	1.917	0.360	0.358	0.301	0.311	0.327	40.764	0.301	0.332	0.420
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{de} = \beta_{0,1}^{(b)}$																	
Avg-Bias	0.299	0.298	0.152	0.289	0.221	0.153	0.204	0.200	0.372	0.454	0.264	0.178	0.266	0.233	0.170	0.489	0.430	0.404
Avg-Cov	0.740	0.748	0.927	0.552	0.389	0.559	0.941	0.759	0.945	0.772	0.751	0.824	0.650	0.963	0.485	0.850	0.932	0.983
Avg-Len	0.847	0.862	0.671	0.562	0.302	0.301	1.004	0.588	0.833	1.242	0.775	0.593	0.616	1.237	14.396	1.760	1.940	2.459
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(1)}$ **so that** $m(x^{(1)}) = x^{(1)T}\beta_0^{pd} = \beta_{0,1}^{(c)}$																	
Avg-Bias	0.091	0.091	0.071	0.094	0.090	0.073	0.137	0.174	0.265	0.097	0.095	0.075	0.101	0.094	0.230	0.105	0.098	0.106
Avg-Cov	0.868	0.867	0.906	0.845	0.817	0.863	0.954	0.920	0.925	0.911	0.900	0.921	0.824	0.941	0.511	0.867	0.929	0.964
Avg-Len	0.345	0.345	0.300	0.334	0.298	0.275	0.659	0.647	1.490	0.426	0.395	0.335	0.346	0.468	32.172	0.400	0.440	0.558
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{sp}$																	
Avg-Bias	0.057	0.057	0.047	0.068	0.066	0.055	NA	NA	NA	0.050	0.051	0.041	0.070	0.054	0.198	0.058	0.062	0.074
Avg-Cov	0.986	0.995	0.994	0.911	0.923	0.927	NA	NA	NA	0.935	0.937	0.952	0.917	0.900	0.495	0.925	0.933	0.942
Avg-Len	0.357	0.391	0.319	0.294	0.291	0.248	NA	NA	NA	0.241	0.241	0.202	0.295	0.221	51.810	0.260	0.292	0.354
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{de}$																	
Avg-Bias	0.244	0.246	0.126	0.241	0.197	0.137	NA	NA	NA	0.396	0.256	0.177	0.232	0.221	0.157	0.407	0.361	0.350
Avg-Cov	0.974	0.980	0.999	0.602	0.414	0.577	NA	NA	NA	0.709	0.694	0.750	0.683	0.867	0.495	0.862	0.946	0.977
Avg-Len	1.399	1.496	1.071	0.529	0.284	0.280	NA	NA	NA	1.067	0.662	0.519	0.584	0.836	17.159	1.519	1.708	2.071
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(2)}$ **so that** $m(x^{(2)}) = x^{(2)T}\beta_0^{pd}$																	
Avg-Bias	0.083	0.083	0.065	0.091	0.085	0.070	NA	NA	NA	0.087	0.084	0.067	0.089	0.087	0.197	0.091	0.087	0.091
Avg-Cov	0.990	0.994	0.998	0.824	0.806	0.860	NA	NA	NA	0.915	0.908	0.921	0.863	0.843	0.517	0.867	0.914	0.951
Avg-Len	0.529	0.564	0.456	0.315	0.280	0.256	NA	NA	NA	0.382	0.354	0.301	0.328	0.316	29.117	0.345	0.388	0.470
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{sp} = \beta_{0,100}^{(a)}$																	
Avg-Bias	0.057	0.057	0.054	0.076	0.076	0.068	0.421	1.217	0.733	0.000	0.000	0.000	0.081	0.094	0.269	0.077	0.088	0.097
Avg-Cov	0.985	0.985	0.978	0.943	0.943	0.939	0.972	0.967	0.958	1.000	1.000	1.000	0.947	0.956	0.498	0.953	0.956	0.955
Avg-Len	0.352	0.352	0.312	0.372	0.368	0.322	3.970	8.027	8.075	0.000	0.000	0.000	0.390	0.472	49.953	0.380	0.436	0.489
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{de} = \beta_{0,100}^{(b)}$																	
Avg-Bias	0.177	0.176	0.130	0.225	0.189	0.167	0.179	0.221	0.279	0.255	0.217	0.200	0.254	0.325	0.181	0.399	0.371	0.392
Avg-Cov	0.978	0.978	0.990	0.782	0.528	0.604	0.988	0.889	0.954	0.229	0.337	0.391	0.804	0.967	0.492	0.969	0.993	0.995
Avg-Len	1.078	1.097	0.856	0.669	0.359	0.364	1.288	1.680	1.097	0.388	0.353	0.331	0.773	1.791	26.954	2.223	2.550	2.858
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(3)}$ **so that** $m(x^{(3)}) = x^{(3)T}\beta_0^{pd} = \beta_{0,100}^{(c)}$																	
Avg-Bias	0.071	0.071	0.067	0.093	0.086	0.082	0.302	0.209	0.394	0.059	0.057	0.056	0.100	0.130	0.239	0.103	0.107	0.117
Avg-Cov	0.976	0.977	0.969	0.892	0.864	0.890	0.984	0.970	0.963	0.303	0.346	0.377	0.926	0.962	0.508	0.948	0.972	0.980
Avg-Len	0.436	0.437	0.381	0.398	0.354	0.333	2.462	3.289	1.953	0.143	0.149	0.143	0.434	0.677	46.809	0.505	0.579	0.649
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{sp}$																	
Avg-Bias	0.049	0.049	0.040	0.069	0.061	0.054	NA	NA	NA	0.045	0.045	0.037	0.069	0.045	0.138	0.045	0.046	0.048
Avg-Cov	1.000	1.000	1.000	0.857	0.901	0.873	NA	NA	NA	0.944	0.938	0.960	0.835	0.844	0.530	0.890	0.924	0.931
Avg-Len	0.736	0.766	0.568	0.248	0.246	0.203	NA	NA	NA	0.218	0.217	0.182	0.240	0.160	23.997	0.183	0.212	0.226
	$d = 1000$, $n = 900$, **dense** β_0^{de}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{de}$																	
Avg-Bias	0.180	0.181	0.089	0.173	0.123	0.090	NA	NA	NA	0.376	0.228	0.159	0.168	0.163	0.119	0.436	0.323	0.285
Avg-Cov	1.000	1.000	1.000	0.694	0.568	0.684	NA	NA	NA	0.616	0.560	0.613	0.760	0.844	0.488	0.650	0.867	0.944
Avg-Len	3.697	3.793	2.527	0.446	0.240	0.229	NA	NA	NA	0.782	0.470	0.359	0.479	0.606	9.202	1.070	1.238	1.320
	$d = 1000$, $n = 900$, **pseudo-dense** β_0^{pd}, **and** $x^{(4)}$ **so that** $m(x^{(4)}) = x^{(4)T}\beta_0^{pd}$																	
Avg-Bias	0.066	0.066	0.054	0.140	0.103	0.103	NA	NA	NA	0.067	0.063	0.050	0.115	0.064	0.146	0.078	0.065	0.062
Avg-Cov	1.000	1.000	1.000	0.474	0.625	0.528	NA	NA	NA	0.880	0.871	0.886	0.631	0.837	0.540	0.794	0.918	0.946
Avg-Len	1.320	1.351	1.020	0.266	0.236	0.210	NA	NA	NA	0.258	0.239	0.203	0.270	0.230	23.784	0.243	0.281	0.300
	$d = 1000$, $n = 900$, **sparse** β_0^{sp}, **and** $x^{(5)}$ **so that** $m(x^{(5)}) = x^{(5)T}\beta_0^{sp}$																	
Avg-Bias	0.028	0.028	0.020	0.039	0.036	0.029	NA	NA	NA	0.002	0.003	0.001	0.033	0.005	0.025	0.007	0.007	0.009
Avg-Cov	1.000	1.000	1.000	0.956	0.968	0.951	NA	NA	NA	0.865	0.625	0.889	0.950	0.000	0.488	0.904	0.906	0.960
Avg-Len	2.623	2.650	1.879	0.188	0.186	0.142	NA	NA	NA	0.006	0.007	0.005	0.167	0.000	7.529	0.031	0.035	0.045

Table D.9: Simulation results under the Toeplitz (or autoregressive) covariance matrix Σ^{ar}, Gaussian noise $\mathcal{N}(0, 1)$, and the MCAR setting. Nearly all the standard errors for the above average quantities are less than 0.01 except for those associated with "DL-vdG", "R-Proj", and "SSL-AIPW" and thus omitted.

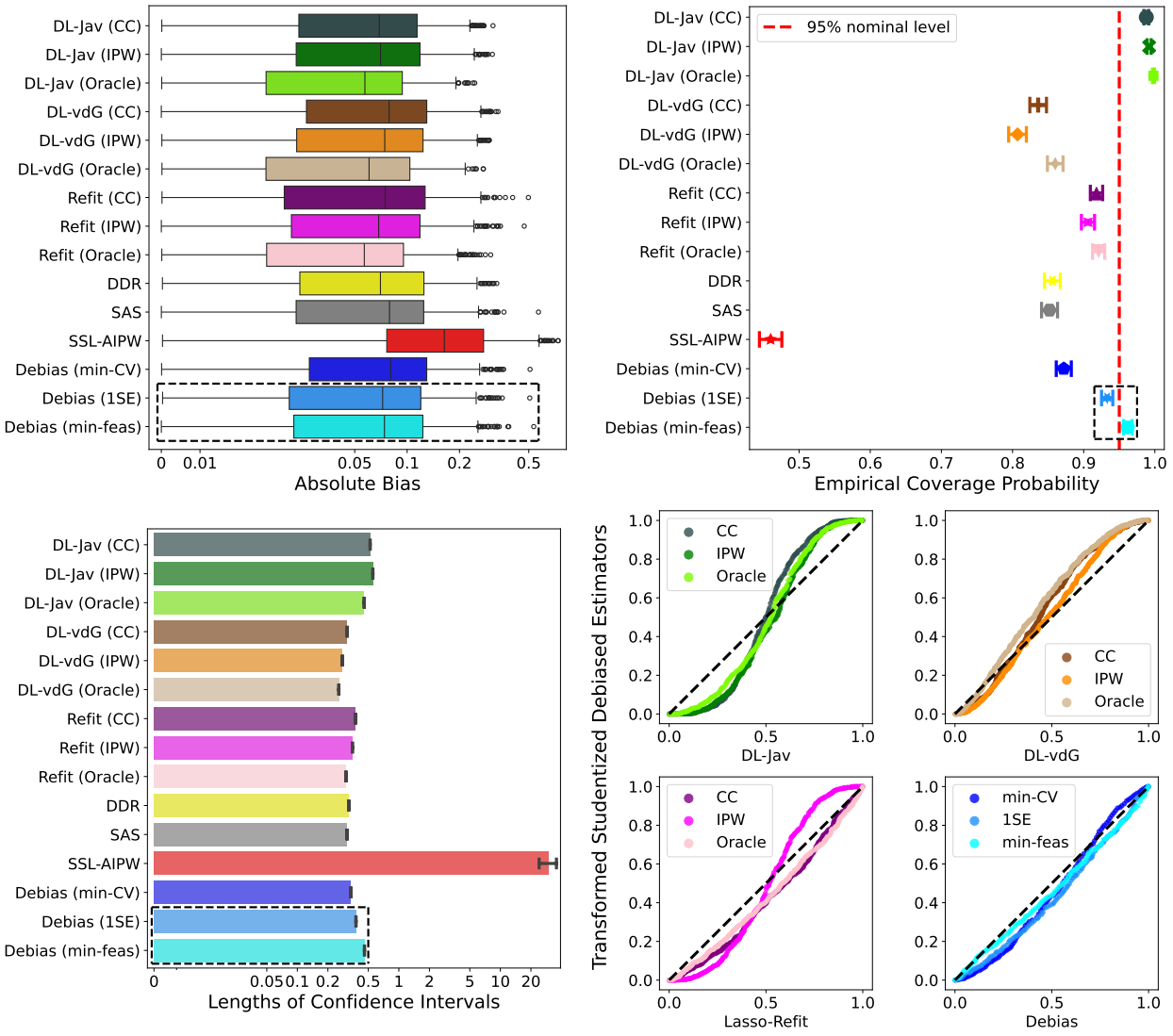
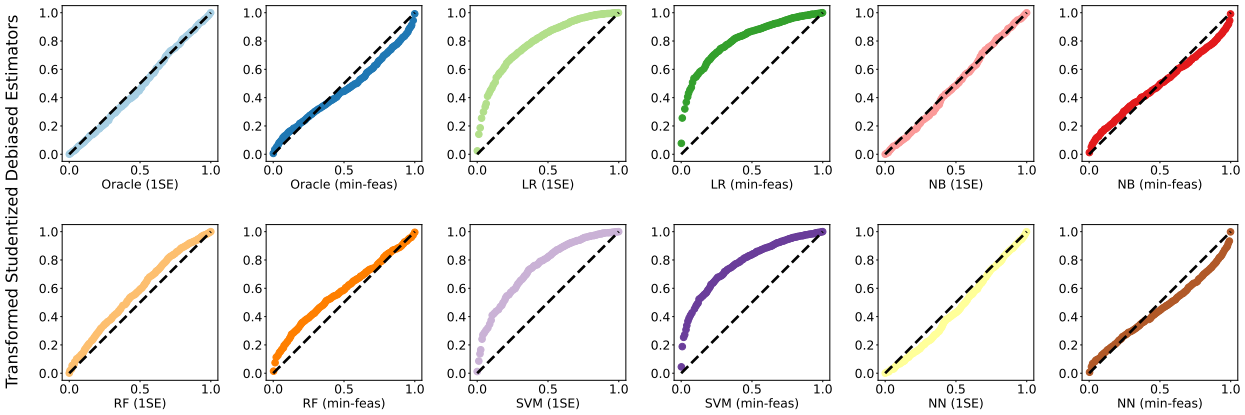
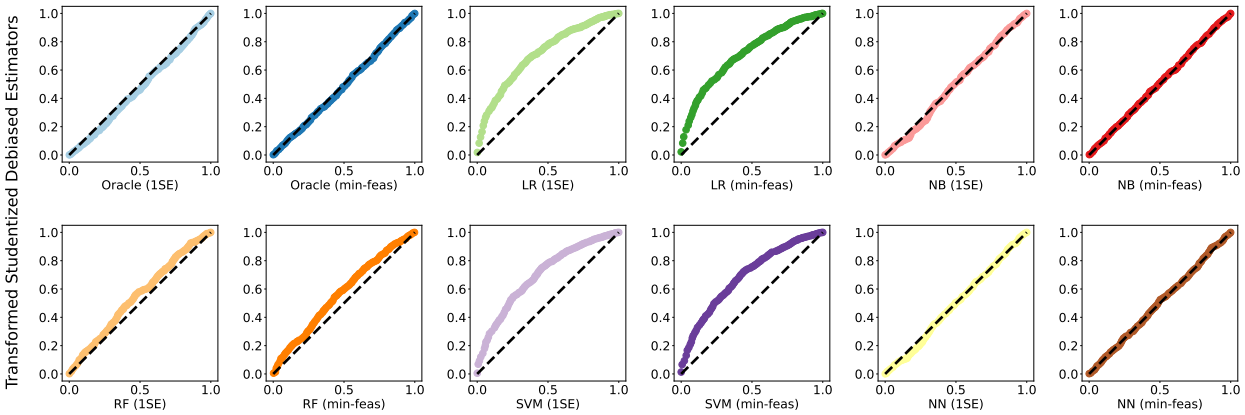


Figure D.1: Simulation results with pseudo-dense β_0^{pd} and sparse $x^{(2)}$ for the Toeplitz (or autoregressive) covariance matrix Σ^{ar} under Gaussian noise $\mathcal{N}(0,1)$ and the MAR setting (5.23) in Section 5.4.1. **Top Left:** Boxplots of the absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Bottom Left:** Average lengths of confidence intervals with standard error bars. **Bottom Right:** Four comparative "QQ-plots" of the transformed studentized debiased estimators $\Phi\left(\sqrt{n}\cdot\widehat{\sigma}_n^{-1}(x)\left[\widehat{m}(x)-m_0(x)\right]\right)$ obtained from different debiasing methods, where $\widehat{\sigma}_n(x)^2$ is the estimated (asymptotic) variance of $\widehat{m}(x)$ by each method and $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$. We also highlight the recommended rules "1SE" and "min-feas" for our proposed debiasing method via dashed rectangles in the first three panels.



(a) Simulation results with dense β_0^{de} and sparse $x^{(2)}$.



(b) Simulation results with sparse β_0^{sp} and dense $x^{(4)}$.

Figure D.2: "QQ-plots" of the transformed studentized debiased estimators $\Phi\left(\sqrt{n}\cdot\widehat{\sigma}_n^{-1}(x)\left[\widehat{m}(x)-m_0(x)\right]\right)$ obtained from our debiasing method when the propensity scores are estimated by various machine learning methods under the new MAR setting (5.24), where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$.

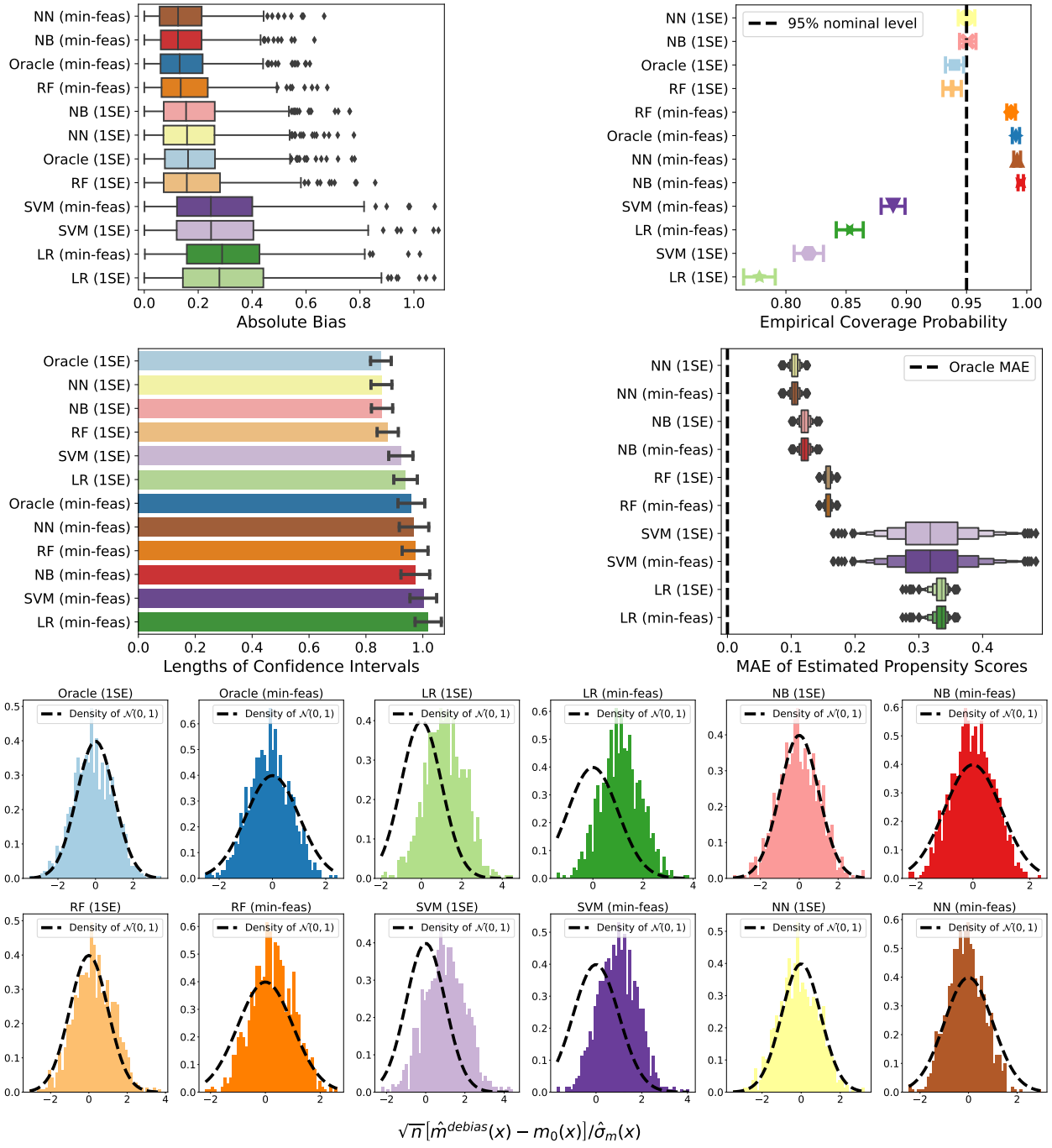


Figure D.3: Simulation results for our debiasing method with dense β_0^{de} and sparse $x^{(2)}$ when the propensity scores are estimated by various machine learning methods under the new MAR setting (5.24). **Top Left:** Boxplots of the absolute bias. **Top Right:** Coverage probabilities with standard error bars. **Middle Left:** Average lengths of confidence intervals with standard deviation bars. **Middle Right:** Letter-valued plots of the MAE of estimated propensity scores. **Bottom Two Rows:** Histograms of the studentized debiased estimators under different methods. We also order the performance metrics according to their means in the first four panels.

Appendix E

APPENDIX TO CHAPTER 6

This appendix collects additional technical results, implementation details, and supplementary analyses for Chapter 6. We examine what can be learned when treatment variation degenerates for some covariate values, provide proofs of selected identification and asymptotic results, discuss practical estimation of nuisance quantities, and present additional IllustrisTNG analyses across redshifts and tilting parameters.

E.1 Nonparametric Bounds Under Zero Treatment Variations

In this section, we study nonparametric bounds on the dose-response curve $m(t)$ and its derivative function $\theta(t)$ under the additive confounding model (6.4) when the intrinsic treatment variation has zero variance. Specifically,

$$Y^{(t)} = \bar{m}(t) + \eta(\boldsymbol{X}) + \epsilon \quad \text{and} \quad T = f(\boldsymbol{X}, E), \tag{E.1}$$

where $E \in \mathbb{R}$ is an independent treatment variation random variable with $\mathbb{E}(E) = 0$. Zero treatment variation means that $\mathrm{Var}(E) = 0$. Deriving analogous nonparametric bounds under more general structural models is left for future work.

We begin with an illustrative example demonstrating how zero treatment variation for certain covariate levels $\boldsymbol{x} \in \mathcal{X}$ can lead to ambiguities in defining the corresponding dose-response curves.

Example E.1 (Necessity of Assumption 6.3)**.** *Suppose that the positivity condition in Assumption 6.1 holds, and let* $T = f(\boldsymbol{X}, E) = X_1$ *almost surely so that* $\mathrm{Var}(E) = \mathrm{Var}(T|\boldsymbol{X} = \boldsymbol{x}) = 0$ *for any* $X_1 = x_1$*. Here,* X_1 *is the first component of* $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$*, and* $E \in \mathbb{R}$ *is an independent treatment variation random variable. We further assume that* $\mathbb{E}(X_1) = 0$ *and*

consider two equivalent conditional mean outcome functions:

$$\mu_1(T, \boldsymbol{X}) \equiv T + 2X_1 \quad \textit{and} \quad \mu_2(T, \boldsymbol{X}) \equiv 2T + X_1,$$

both of which are well-defined under the positivity condition in Assumption 6.1 and equal to $3X_1$ *almost surely. While they agree on the support* $\{(t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X} : t = f(\boldsymbol{x}) = x_1\}$*, these two conditional mean outcome functions lead to two distinct covariate-adjusted functions (or equivalently, dose-response curves):*

$$m_1(t) = \mathbb{E}\left[Y_1^{(t)}\right] = \mathbb{E}\left[\mu_1(t, \boldsymbol{X})\right] = t \quad \textit{and} \quad m_2(t) = \mathbb{E}\left[Y_2^{(t)}\right] = \mathbb{E}\left[\mu_2(t, \boldsymbol{X})\right] = 2t,$$

whose derivatives are different as well.

E.1.1 Nonparametric Bound on $m(t)$

Under the additive confounding model (E.1), Example E.1 above demonstrates that when $\mathrm{Var}(E) = 0$, the conditional mean outcome function $\mu(t, \boldsymbol{x})$ is identifiable only on the lower-dimensional surface $\{(t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X} : t = f(\boldsymbol{x}, 0) \equiv f(\boldsymbol{x})\}$, where

$$\mu\left(f(\boldsymbol{x}), \boldsymbol{x}\right) = \bar{m}\left(f(\boldsymbol{x})\right) + \eta(\boldsymbol{x}) = m\left(f(\boldsymbol{x})\right) + \eta(\boldsymbol{x}) \tag{E.2}$$

when $\mathbb{E}\left[\eta(\boldsymbol{X})\right] = 0$ by Proposition 6.2.

Since $\mathrm{Var}(E) = 0$, the relation $T = f(\boldsymbol{X})$ can always be estimated from the observed data. Hence, we assume that the function f is known for every $\boldsymbol{x} \in \mathcal{X}$ in subsequent analyses. To obtain nonparametric bounds for all possible values of $m(t)$, we impose the following assumption on the magnitude of the random effect function $\eta : \mathcal{X} \to \mathbb{R}$ in (E.1); see Manski (1990) for related discussions.

Assumption E.1 (Bounded random effect)**.** *Let* $L_f(t) = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = t\}$ *be a level set of the function* $f : \mathcal{X} \to \mathbb{R}$ *at* $t \in \mathcal{T}$*. There exists a constant* $\rho_1 > 0$ *such that*

$$\rho_1 \geq \max\left\{\sup_{t \in \mathcal{T}} \sup_{\boldsymbol{x} \in L_f(t)} |\eta(\boldsymbol{x})|, \; \frac{\sup_{t \in \mathcal{T}} \sup_{\boldsymbol{x} \in L_f(t)} \mu(f(\boldsymbol{x}), \boldsymbol{x}) - \inf_{t \in \mathcal{T}} \inf_{\boldsymbol{x} \in L_f(t)} \mu(f(\boldsymbol{x}), \boldsymbol{x})}{2}\right\}.$$

By (E.2) and the first lower bound on $\rho_1 \geq \sup_{t\in\mathcal{T}} \sup_{\boldsymbol{x}\in L_f(t)} |\eta(\boldsymbol{x})|$ in Assumption E.1, we know that

$$|\mu(f(\boldsymbol{x}), \boldsymbol{x}) - m(t)| = |\eta(\boldsymbol{x})| \leq \rho_1$$

for any $\boldsymbol{x} \in L_f(t)$. This also implies that

$$\begin{aligned} m(t) &\in \bigcap_{\boldsymbol{x}\in L_f(t)} [\mu(f(\boldsymbol{x}), \boldsymbol{x}) - \rho_1,\ \mu(f(\boldsymbol{x}), \boldsymbol{x}) + \rho_1] \\ &= \left[\sup_{\boldsymbol{x}\in L_f(t)} \mu(f(\boldsymbol{x}), \boldsymbol{x}) - \rho_1,\ \inf_{\boldsymbol{x}\in L_f(t)} \mu(f(\boldsymbol{x}), \boldsymbol{x}) + \rho_1 \right], \end{aligned} \tag{E.3}$$

which is the nonparametric bound on $m(t)$ that contains all the possible values of $m(t)$ for any fixed $t \in \mathcal{T}$ when $\text{Var}(E) = 0$. Notice that this bound (E.3) is well-defined and nonempty under the second lower bound on ρ_1 in Assumption E.1.

E.1.2 Nonparametric Bound on $\theta(t)$

Since the functions $\mu(f(\boldsymbol{x}), \boldsymbol{x})$ and $f(\boldsymbol{x})$ are identifiable even when $\text{Var}(E) = 0$, their derivatives with respect to $\boldsymbol{x}$ are also identifiable. Hence, we assume that the gradients $\nabla_{\boldsymbol{x}}\mu(f(\boldsymbol{x}), \boldsymbol{x})$ and $\nabla f(\boldsymbol{x})$ are known for every $\boldsymbol{x} \in \mathcal{X}$ in the following analysis. The nonparametric bound on the dose-response curve $m(t)$ relies on an upper bound on the random effect function $\eta : \mathcal{X} \to \mathbb{R}$ (recall Assumption E.1), while the nonparametric bound on the derivative function $\theta(t)$ requires some constraints on the gradient $\nabla\eta(\boldsymbol{x})$ of the random effect function as follows.

Assumption E.2 (Constraints on the random effect gradient)**.** *Let $L_f(t) = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = t\}$ be a level set of the function $f : \mathcal{X} \to \mathbb{R}$ at $t \in \mathcal{T}$. There exist constants $\rho_2, \rho_3 > 0$ such that*

$$\rho_3 \leq \inf_{t\in\mathcal{T}} \inf_{\boldsymbol{x}\in L_f(t)} ||\nabla\eta(\boldsymbol{x})||_{\min} \leq \sup_{t\in\mathcal{T}} \sup_{\boldsymbol{x}\in L_f(t)} ||\nabla\eta(\boldsymbol{x})||_{\max} \leq \rho_2$$

and

$$\sup_{t\in\mathcal{T}} \sup_{\boldsymbol{x}\in L_f(t)} \max_{j=1,\ldots,d} \left[\frac{v_j(\boldsymbol{x}) - \text{sign}(g_j(\boldsymbol{x})) \cdot \rho_2}{g_j(\boldsymbol{x})} \right] \leq \inf_{t\in\mathcal{T}} \inf_{\boldsymbol{x}\in L_f(t)} \min_{j=1,\ldots,d} \left[\frac{v_j(\boldsymbol{x}) + \text{sign}(g_j(\boldsymbol{x})) \cdot \rho_2}{g_j(\boldsymbol{x})} \right],$$

where $||\nabla\eta(\boldsymbol{x})||_{\min} = \min_{j=1,...,d}\left|\frac{\partial}{\partial x_j}\eta(\boldsymbol{x})\right|$, $||\nabla\eta(\boldsymbol{x})||_{\max} = \max_{j=1,...,d}\left|\frac{\partial}{\partial x_j}\eta(\boldsymbol{x})\right|$, $v_j(\boldsymbol{x}) = \frac{\partial}{\partial x_j}\mu(f(\boldsymbol{x}),\boldsymbol{x})$, *and* $g_j(\boldsymbol{x}) = \frac{\partial}{\partial x_j}f(\boldsymbol{x})$ *for* $j = 1, ..., d$.

The first inequality in Assumption E.2 ensures that we can derive a finite and meaningful nonparametric bound on $\theta(t)$ for any $t \in \mathcal{T}$. In particular, a direct calculation shows that

$$\begin{aligned}\nabla_{\boldsymbol{x}}\mu(f(\boldsymbol{x}),\boldsymbol{x}) &= \nabla_{\boldsymbol{x}}m(f(\boldsymbol{x})) + \nabla\eta(\boldsymbol{x})\\ &= m'(f(\boldsymbol{x}))\cdot\nabla f(\boldsymbol{x}) + \nabla\eta(\boldsymbol{x})\\ &= \theta(t)\cdot\nabla f(\boldsymbol{x}) + \nabla\eta(\boldsymbol{x})\end{aligned}$$

for all $\boldsymbol{x} \in L_f(t)$. This indicates that

$$\left|\frac{\partial}{\partial x_j}\mu(f(\boldsymbol{x}),\boldsymbol{x}) - \theta(t)\cdot\frac{\partial}{\partial x_j}f(\boldsymbol{x})\right| = |v_j(\boldsymbol{x}) - \theta(t)\cdot g_j(\boldsymbol{x})| = \left|\frac{\partial}{\partial x_j}\eta(\boldsymbol{x})\right| \le \rho_2$$

for $j = 1, ..., d$ and therefore, a nonparametric bound on $\theta(t)$ is given by

$$\theta(t) \in \bigcap_{j=1}^{d}\left[\frac{v_j(\boldsymbol{x}) - \text{sign}(g_j(\boldsymbol{x}))\cdot\rho_2}{g_j(\boldsymbol{x})}, \frac{v_j(\boldsymbol{x}) + \text{sign}(g_j(\boldsymbol{x}))\cdot\rho_2}{g_j(\boldsymbol{x})}\right]. \tag{E.4}$$

This bound is well-defined and nonempty under the second inequality in Assumption E.2. More importantly, the nonparametric bound (E.4) highlights the pivotal role of the covariate effect function f of $\boldsymbol{X}$ on treatment T in bounding the possible values of $\theta(t)$. Specifically, as the variation in f increases (*i.e.*, the magnitude of $\nabla f(\boldsymbol{x})$ becomes larger), the nonparametric bound (E.4) becomes tighter, providing more precise constraints on $\theta(t)$.

E.2 Proofs of Results in the Chapter

E.2.1 Proof of Proposition 6.1

Proof of Proposition 6.1. We first study the identification of $\theta(t)$. The equality (6.2) is a natural consequence of Assumptions 6.2 and 6.4:

$$\theta(t) \overset{(*)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)}|\boldsymbol{X}\right]\Big|T = t\right] \overset{(**)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y^{(t)}|T = t, \boldsymbol{X}\right]\Big|T = t\right] \overset{(***)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}\mathbb{E}\left[Y|T = t, \boldsymbol{X}\right]\Big|T = t\right],$$

where (*) follows from Assumption 6.4, (**) is by Assumption 6.2(b), and (***) is due to Assumption 6.2(a).

As for the identification of $m(t)$, we apply the fundamental theorem of calculus and argue that $m(t) = m(T) + \int_T^t \theta(\widetilde{t})\, d\widetilde{t}$. Taking the expectation over T yields that

$$m(t) = \mathbb{E}\left[m(T) + \int_T^t \theta(\widetilde{t})\, d\widetilde{t}\right] = \mathbb{E}\left\{Y + \int_T^t \mathbb{E}\left[\frac{\partial}{\partial t}\mu(T, \boldsymbol{X})\Big| T = \widetilde{t}\right] d\widetilde{t}\right\},$$

where the second equality follows from $\mathbb{E}[m(T)] = \mathbb{E}(Y)$ by Assumption 6.4. Here, the connectedness of $\mathcal{T}$ ensures that the integration of $\theta(t)$ is only over the region where it is identifiable. When $\mathcal{T}$ has multiple connected components, the integral formula (6.3) as well as the observations should be restricted to the connected component in which the point of interest $t \in \mathcal{T}$ lies. □

E.2.2 Proof of Proposition 6.2

Proof of Proposition 6.2. All the results follow from some simple calculations and the fact that

$$\mu(t, \boldsymbol{x}) = \mathbb{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x}) = \bar{m}(t) + \eta(\boldsymbol{x})$$

under the additive confounding model (6.4).

(a) Given that $\mathbb{E}[\eta(\boldsymbol{X})] = 0$, we have that

$$\mathbb{E}[\mu(t, \boldsymbol{X})] = \mathbb{E}[\mathbb{E}(Y|T = t, \boldsymbol{X})] = \mathbb{E}[\bar{m}(t) + \eta(\boldsymbol{X})] = \bar{m}(t).$$

(b) First, we notice that $m(t) = \mathbb{E}\left[Y^{(t)}\right] = \bar{m}(t) + \mathbb{E}[\eta(\boldsymbol{X})]$ and $\theta(t) = \bar{m}'(t)$ under model (6.4). Then,

$$\frac{d}{dt}\mathbb{E}[\mu(t, \boldsymbol{X})|T = t] = \frac{d}{dt}\{\bar{m}(t) + \mathbb{E}[\eta(\boldsymbol{X})|T = t]\} = \bar{m}'(t) + \frac{d}{dt}\mathbb{E}(\eta(\boldsymbol{X})|T = t) \neq \theta(t)$$

when $\frac{d}{dt}\mathbb{E}[\eta(\boldsymbol{X})|T = t] \neq 0$.

(c) From (b), we know that

$$\mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{X})\right] = \mathbb{E}\left[\bar{m}'(t)\right] = \bar{m}'(t) = \theta(t).$$

In addition, we have that

$$\theta_C(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t,\boldsymbol{X})\Big|T=t\right] = \mathbb{E}\left[\bar{m}'(t)|T=t\right] = \bar{m}'(t) = \theta(t).$$

(d) On the one hand, we know that

$$\mathbb{E}(Y) = \mathbb{E}\left[\mu(T,\boldsymbol{X})\right] = \mathbb{E}\left[\bar{m}(T)+\eta(\boldsymbol{X})\right] = \mathbb{E}\left[\bar{m}(T)\right] + \mathbb{E}\left[\eta(\boldsymbol{X})\right].$$

On the other hand, we also have that

$$\mathbb{E}\left[m(T)\right] = \mathbb{E}\left\{\bar{m}(T)+\mathbb{E}\left[\eta(\boldsymbol{X})\right]\right\} = \mathbb{E}\left[\bar{m}(T)\right] + \mathbb{E}\left[\eta(\boldsymbol{X})\right].$$

The proof is thus completed. □

E.3 Practical Considerations for Estimating $\bar{m}_\hbar(t_0)$

This section delineates the practical aspects involved in implementing our proposed estimators of the causal estimand $t_0 \mapsto \bar{m}_\hbar(t_0)$ for $\hbar > 0$ and its limiting curve $t_0 \mapsto \bar{m}_0(t_0) = \lim_{\hbar\to 0}\bar{m}_\hbar(t_0)$ introduced in Section 6.8.1.

E.3.1 Proposed Estimators With Cross-Fitting

To mitigate overfitting and ensure valid inference, we implement our proposed estimators (6.29) and (6.31) for $\bar{m}_\hbar(t_0)$ in Section 6.8.1 using the following cross-fitting procedure.

1. Data partitioning: The observed data $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ are randomly divided into L disjoint folds of approximately equal sizes. Let $I_\ell, \ell = 1, ..., L$ denote the fold index sets with $\cup_{\ell=1}^L I_\ell = \{1, ..., n\}$.

2. Nuisance function estimation: For each fold I_ℓ, we estimate the nuisance functions $\mu, p_{T|\boldsymbol{X}}$ using only observations *outside* I_ℓ; see Section E.3.2 for details about the estimation methods. The estimated nuisance functions are denoted by $\widehat{\mu}^{(\ell)}$ and $\widehat{p}^{(\ell)}_{T|\boldsymbol{X}}$ for $\ell = 1, ..., L$.

Here, $\mu(t, \boldsymbol{x})$ is the conditional mean outcome function and $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ is the conditional treatment density of T given $\boldsymbol{X} = \boldsymbol{x}$.

3. Final estimators with cross-fitting: The cross-fitted one-step (6.29) and RA (6.31) estimators of $\widehat{m}_\hbar(t_0)$ for $\hbar > 0$ are given by

$$\begin{aligned}\widehat{m}_\hbar(t_0) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{\frac{\widehat{q}^{(\ell)}_{\hbar,t_0}(T_i|\boldsymbol{X}_i)}{\widehat{p}^{(\ell)}_{T|\boldsymbol{X}}(T_i|\boldsymbol{X}_i)}\left[Y_i - \int_{\mathcal{T}}\widehat{\mu}^{(\ell)}(t_1, \boldsymbol{X}_i)\cdot\widehat{q}^{(\ell)}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\,dt_1\right]\right. \\ \left. + \int_{\mathcal{T}}\widehat{\mu}^{(\ell)}(t_1, \boldsymbol{X}_i)\cdot\widehat{q}^{(\ell)}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\,dt_1\right\}, \qquad \text{(E.5)} \\ \widehat{m}_{\hbar,\mathrm{RA}}(t_0) = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\int_{\mathcal{T}}\widehat{\mu}^{(\ell)}(t_1, \boldsymbol{X}_i)\cdot\widehat{q}^{(\ell)}_{\hbar,t_0}(t_1|\boldsymbol{X}_i)\,dt_1,\end{aligned}$$

where we recall from (6.30) that

$$\widehat{q}^{(\ell)}_{\hbar,t_0}(t|\boldsymbol{x}) = \frac{\eta(t; t_0, \hbar)\cdot\widehat{p}^{(\ell)}_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}}\eta(u; t_0, \hbar)\cdot\widehat{p}^{(\ell)}_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}$$

is completely determined by the estimated conditional density function $\widehat{p}^{(\ell)}_{T|\boldsymbol{X}}$ for $\ell = 1, ..., L$.

E.3.2 Nuisance Function Estimation

The implementation of our proposed estimators (6.29) and (6.31) requires estimating two nuisance functions: (i) the conditional mean outcome function $\mu(t, \boldsymbol{x}) = \mathbb{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$ and (ii) the conditional treatment density $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$. Our numerical experiments in Section 5.4 utilize the following methods.

- **Estimation of $\mu(t, \boldsymbol{x})$:** We employ a fully connected neural network model with one hidden layer of size 20 and the rectified linear unit (ReLU) activation function $u \mapsto \max\{0, u\}$. Other machine learning models can be used without modifications. Practically, the neural network model is implemented via the `scikit-learn` package (Pedregosa et al., 2011).

- **Estimation of $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$:** As in Section A.3 of Zhang and Chen (2025), we consider two approaches for estimating the conditional density function $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$.

1. *Method 1 (Residual-based Kernel Density Estimation; RKDE):* Assume an additive treatment model $T = g_{\boldsymbol{X}}(\boldsymbol{X}) + E$, where E is mean-zero and independent of $\boldsymbol{X}$. We first estimate the regression function $g_{\boldsymbol{X}}$ via any machine learning method on the data $\{(T_i, \boldsymbol{X}_i)\}_{i=1}^n$ with cross-fitting. In the actual implementation, we use a neural network model with one hidden layer of size 10 and ReLU activation function $u \mapsto \max\{0, u\}$. We then apply kernel density estimation (KDE) to the residuals and construct an estimator of $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ as follows:

$$\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{x}) = \frac{1}{nh_e} \sum_{i=1}^{n} K_e \left[\frac{t - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{x}) - (T_i - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{X}_i))}{h_e} \right],$$

   where the kernel function K_e can be different from the one in Section 6.7.1 and $h_e > 0$ is a smoothing bandwidth parameter. Although the residuals $\{T_1 - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{X}_1), ..., T_n - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{X}_n)\}$ for KDE are not i.i.d., we notice that this dependence has negligible empirical effects on $\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{x})$. Furthermore, this approach is better adapted to the covariate-dependent support $\mathcal{T}(\boldsymbol{x})$ than the following RKSO method and is more robust under positivity violations. Finally, when the noise E is heteroskedastic, we can estimate $\mathrm{Var}(E|\boldsymbol{X})$ and standardize the residuals before applying KDE. In our experiments, we use the Epanechnikov kernel $K_e(u) = \frac{3}{4}(1 - u^2)\,\mathbb{1}_{\{|u| \le 1\}}$ and Silverman's rule-of-thumb bandwidth as $\widehat{h}_e = \left(\frac{4}{3}\right)^{\frac{1}{5}} \widehat{\sigma}_e n^{-\frac{1}{5}}$, where $\widehat{\sigma}_e$ is the sample standard deviation of $\{T_1 - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{X}_1), ..., T_n - \widehat{g}_{\boldsymbol{X}}(\boldsymbol{X}_n)\}$.

2. *Method 2 (Regression on Kernel-Smoothed Outcomes; RKSO):* We estimate the conditional treatment density $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ via kernel-smoothed regression $g(t, \boldsymbol{x}) = \mathbb{E}\left[K_r\left(\frac{T-t}{b}\right) \middle| \boldsymbol{X} = \boldsymbol{x}\right]$, by regressing kernel-smoothed outcomes $\left\{K_r\left(\frac{T_i-t}{b}\right)\right\}_{i=1}^n$ against the covariate vectors $\{\boldsymbol{X}_i\}_{i=1}^n$ via any machine learning method. This approach is also called the MultiGPS estimator in Colangelo and Lee (2025). Here, the kernel function $K_r : \mathbb{R} \to [0, \infty)$ can be different from the one in Section 6.7.1 and $b > 0$ is a smoothing bandwidth parameter. As shown in Section A.3 of Zhang and Chen (2025), the fitted kernel-smoothed regression function $\widehat{g}(t, \boldsymbol{x})$ consistently estimates

$p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ when the regression function $\widehat{g}(t, \boldsymbol{x})$ itself is consistent. However, we notice in our experiments that this RKSO method may not respect the true conditional support $\mathcal{T}(\boldsymbol{x})$ of $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$, leading to potentially larger bias than the RKDE approach when estimating the limiting curve $t_0 \mapsto \bar{m}_0(t_0)$ defined by the NFTP. In the actual implementation, we kernelize each outcome observation via the Epanechnikov kernel $K_r(u) = \frac{3}{4}(1-u^2) \cdot \mathbb{1}_{\{|u|\leq 1\}}$ and fit the regression function through a fully connected neural network with one hidden layer of size 10 using `scikit-learn`. The bandwidth b is chosen via Silverman's rule of thumb as $\widehat{b} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \widehat{\sigma}_T n^{-\frac{1}{5}}$, where $\widehat{\sigma}_T$ is the sample standard deviation of $\{T_1, ..., T_n\}$.

E.3.3 Numerical Stabilization for the Riemann Sum Approximation (6.32)

When $t \mapsto \eta(t; t_0, \hbar) \propto \exp\left(-\left|\frac{t-t_0}{\hbar}\right|^{\gamma}\right)$, its values decay rapidly away from t_0 for small $\hbar > 0$. In regions where $t_0 \notin \mathcal{T}(\boldsymbol{x})$ for a given $\boldsymbol{x} \in \mathcal{X}$, the naive Riemann sum approximations of (6.32) may suffer from numerical underflow, causing the estimators to fail to capture the projection term of the projected g-formula (6.21) under positivity violations. To improve numerical stability, we adopt a log-sum-exp trick for calculating the Riemann sums in (6.32) as follows.

1. For a given covariate vector $\boldsymbol{x} \in \mathcal{X}$, we compute the (estimated) tilted intervention density in log-space at each grid point $\widetilde{t}_j$ as:

$$w_j = \log\left[\widehat{p}_{T|\boldsymbol{X}}(\widetilde{t}_j|\boldsymbol{x}) \cdot \eta(\widetilde{t}_j; t_0, \hbar)\right] = \log \widehat{p}_{T|\boldsymbol{X}}(\widetilde{t}_j|\boldsymbol{x}) - \left|\frac{\widetilde{t}_j - t_0}{\hbar}\right|^{\gamma} + C_1(\hbar)$$

   for $j = 1, ..., N$, where $C_1(\hbar) > 0$ is a normalizing constant that is independent of $\boldsymbol{x}$ and t_0.

2. Let $w_{\max} = \max_{j=1,...,N} w_j$. The Riemann sums in the numerator and denominator of (6.32) are then calculated as follows:

$$\sum_{j=1}^{N} \widehat{\mu}(\widetilde{t}_j, \boldsymbol{X}) \cdot \exp\left(w_j - w_{\max}\right) \quad \text{and} \quad \sum_{j=1}^{N} \exp\left(w_j - w_{\max}\right),$$

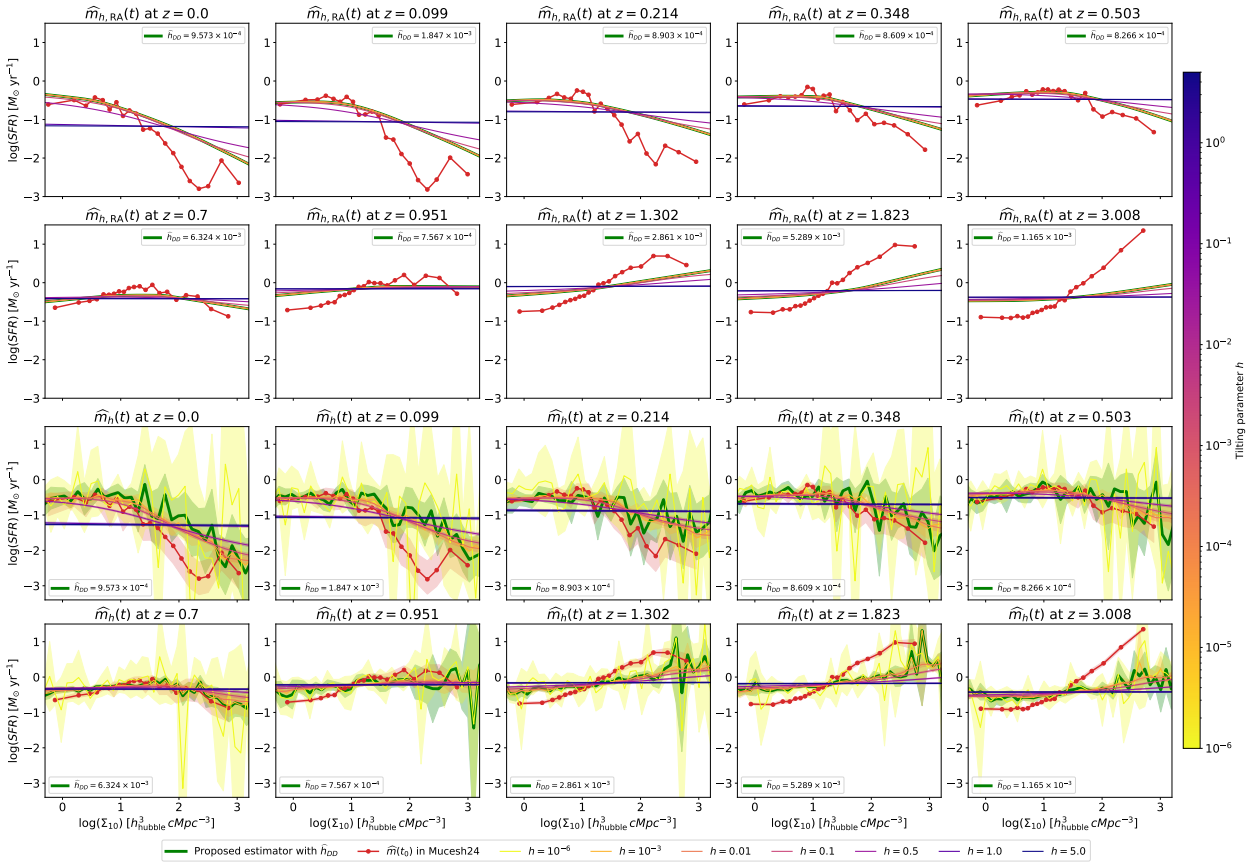


Figure E.1: Tilted intervention causal curves estimated by our proposed RA and one-step estimators under various tilting parameters, alongside the dose-response curve estimated by Mucesh et al. (2024) on the IllustrisTNG galaxy data across different redshift values. The first two rows display the results by our RA estimator (6.31) under varying values of $\hbar$, while the last two rows present the results from our one-step estimator (6.29).

respectively. Here, subtracting the maximum log-weights rescales $w_j, j = 1, ..., N$ and avoids arithmetic underflow after exponentiating.

E.4 Causal Analysis of the IllustrisTNG Data Under Different Redshifts z and Tilting Parameters $\hbar$

Under the same empirical setup as Section 6.9, we plot the tilted intervention causal curves $t_0 \mapsto \bar{m}_\hbar(t_0)$ estimated by our RA (6.31) and one-step (6.29) estimators for multiple tilting

parameter values $\hbar > 0$ across redshifts ranging from $z = 0$ to $z = 3$ in Figure E.1. As discussed in Section 6.9.2, the estimated causal effect of local environment on galactic SFR is negative at low redshifts and gradually flattens as redshift increases. Compared with the estimated dose-response curves by Mucesh et al. (2024), the tilted intervention curves estimated by our proposed RA (6.31) and one-step (6.29) estimators do not display a strong reversal effect at high redshifts $z > 1$. Instead, the curves remain relatively flat across different values of $\hbar$.

E.5 Proofs of Theorem 6.8 and Proposition 6.9

Proof of Theorem 6.8. For any fixed $\boldsymbol{x} \in \mathcal{X}$ that satisfies Assumptions 6.9, it suffices to show that for every $\epsilon > 0$,

$$Q_{\hbar,t_0}(A_\epsilon | \boldsymbol{X} = \boldsymbol{x}) \to 0 \quad \text{as} \quad \hbar \to 0,$$

where we define $A_\epsilon := \mathcal{T}(\boldsymbol{x}) \cap \{t \in \mathcal{T} : |t - t_0^\star(\boldsymbol{x})| \geq \epsilon\}$.

The claim is immediate if A_ϵ is empty. Otherwise, we let $\zeta(\boldsymbol{x}) = \inf_{t \in \mathcal{T}(\boldsymbol{x})} |t - t_0| = |t_0^\star(\boldsymbol{x}) - t_0|$ under Assumption 6.9. Then, there exists $\rho_\epsilon > 0$ such that

$$|t - t_0| \geq \zeta(\boldsymbol{x}) + \rho_\epsilon \quad \text{for all } t \in A_\epsilon. \tag{E.6}$$

This is because if such $\rho_\epsilon > 0$ does not exist, then we can find a sequence $\{t_n\}_{n=1}^\infty \subset A_\epsilon$ with $|t_n - t_0| < \zeta(\boldsymbol{x}) + \frac{1}{n}$. Hence, $\{|t_n - t_0|\}_{n=1}^\infty$ is bounded, and given that $\mathcal{T}(\boldsymbol{x})$ is closed, there is a convergent subsequence $\{t_{n_k}\}_{k=1}^\infty$ with $\lim_{k\to\infty} t_{n_k} = \bar{t} \in \mathcal{T}(\boldsymbol{x})$. By uniqueness and $|t_{n_k} - t_0| < \zeta(\boldsymbol{x}) + \frac{1}{n_k}$ for all k, we know that $\bar{t} = t_0^\star(\boldsymbol{x})$. However, $|t_{n_k} - t_0^\star(\boldsymbol{x})| \geq \epsilon$ when $\{t_{n_k}\}_{k=1}^\infty \subset A_\epsilon$ so that $|\bar{t} - t_0^\star(\boldsymbol{x})| \geq \epsilon$ as we take $k \to \infty$, which is a contradiction. This implies that (E.6) holds and

$$A_\epsilon \subseteq \{t \in \mathcal{T}(\boldsymbol{x}) : |t - t_0| \geq \zeta(\boldsymbol{x}) + \rho_\epsilon\}. \tag{E.7}$$

On the other hand, we define $B_\epsilon := \left\{t \in \mathcal{T}(\boldsymbol{x}) : |t - t_0^\star(\boldsymbol{x})| \leq \frac{\rho_\epsilon}{2}\right\}$ and note that

$$B_\epsilon \subseteq \left\{t \in \mathcal{T}(\boldsymbol{x}) : |t - t_0| \leq \zeta(\boldsymbol{x}) + \frac{\rho_\epsilon}{2}\right\}. \tag{E.8}$$

Moreover, $t_0^\star(\boldsymbol{x}) \in \mathcal{T}(\boldsymbol{x})$ ensures that $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(B_\epsilon) > 0$. Now, we calculate that

$$\begin{aligned} Q_{\hbar,t_0}(A_\epsilon|\boldsymbol{X}=\boldsymbol{x}) &= \frac{\int_{A_\epsilon} \eta(t;t_0,\hbar)\,\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(dt)}{\int_{\mathcal{T}} \eta(u;t_0,\hbar)\,\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(du)} \\ &\le \frac{\int_{A_\epsilon} \eta(t;t_0,\hbar)\,\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(dt)}{\int_{B_\epsilon} \eta(u;t_0,\hbar)\,\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(du)} \\ &\le \frac{\sup_{t\in A_\epsilon} \eta(t;t_0,\hbar)}{\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(B_\epsilon)\cdot \inf_{t\in B_\epsilon} \eta(t;t_0,\hbar)} \\ &\overset{\text{(i)}}{\le} \frac{\sup_{t\in\mathcal{T}:|t-t_0|\ge\zeta(\boldsymbol{x})+\rho_\epsilon} \eta(t;t_0,\hbar)}{\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(B_\epsilon)\cdot \inf_{t\in\mathcal{T}:|t-t_0|\le\zeta(\boldsymbol{x})+\rho_\epsilon/2} \eta(t;t_0,\hbar)} \\ &\overset{\text{(ii)}}{\to} 0 \end{aligned}$$

as $\hbar \to 0$, where (i) leverages the previous inclusion results (E.7) and (E.8), while (ii) applies Condition 6.1. Since the limit is a point mass and $\epsilon > 0$ is arbitrary, the weak convergence $Q_{\hbar,t_0}(\cdot|\boldsymbol{X}=\boldsymbol{x}) \xrightarrow{d} \delta_{t_0^\star(\boldsymbol{x})}(\cdot)$ as $\hbar \to 0$ thus follows. □

Remark E.1 (Non-unique projections)**.** *If* $\Pi(t_0,\boldsymbol{x}) = \arg\min_{t\in\mathcal{T}(\boldsymbol{x})} |t-t_0|$ *is not a singleton, then the conclusion of Theorem 6.8 can be weakened to*

$$Q_{\hbar,t_0}\left(\{t\in\mathcal{T} : d(t,\Pi(t_0,\boldsymbol{x})) > \epsilon\}\,\middle|\,\boldsymbol{X}=\boldsymbol{x}\right) \to 0 \quad \text{as } \hbar\to 0$$

for every $\epsilon > 0$*, where* $d(t,\Pi(t_0,\boldsymbol{x})) = \inf_{t'\in\Pi(t_0,\boldsymbol{x})} |t-t'|$*. Specifically, we define*

$$A_\epsilon = \{t\in\mathcal{T} : d(t,\Pi(t_0,\boldsymbol{x})) > \epsilon\}$$

and $\zeta(\boldsymbol{x}) = \inf_{t\in\mathcal{T}(\boldsymbol{x})} |t-t_0|$*. Since* $\mathcal{T}(\boldsymbol{x})$ *is closed and* $\Pi(t_0,\boldsymbol{x})$ *consists of all the minimizers of* $t\mapsto|t-t_0|$*, there again exists* $\rho_\epsilon > 0$ *such that* (E.7) *holds. Let*

$$B_\epsilon = \left\{t\in\mathcal{T} : d(t,\Pi(t_0,\boldsymbol{x})) \le \frac{\rho_\epsilon}{2}\right\}.$$

Then, $\Pi(t_0,\boldsymbol{x}) \subset \mathcal{T}(\boldsymbol{x})$ *implies that* $\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(B_\epsilon) > 0$*, and every* $t\in B_\epsilon$ *satisfies* $|t-t_0| \le \zeta(\boldsymbol{x}) + \frac{\rho_\epsilon}{2}$*. Thus, we again have that*

$$Q_{\hbar,t_0}(A_\epsilon|\boldsymbol{X}=\boldsymbol{x}) \le \frac{\sup_{t\in\mathcal{T}:|t-t_0|\ge\zeta(\boldsymbol{x})+\rho_\epsilon} \eta(t;t_0,\hbar)}{\mathrm{P}_{T|\boldsymbol{X}=\boldsymbol{x}}(B_\epsilon)\cdot \inf_{t\in\mathcal{T}:|t-t_0|\le\zeta(\boldsymbol{x})+\frac{\rho_\epsilon}{2}} \eta(t;t_0,\hbar)} \to 0.$$

Proof of Proposition 6.9. By Theorem 6.8, we know that $Q_{\hbar,t_0}(\cdot|\boldsymbol{X}=\boldsymbol{x}) \xrightarrow{d} \delta_{t_0^\star(\boldsymbol{x})}(\cdot)$ for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$.

(a) Notice that, by the definition of stochastic interventions,

$$\mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right] = \int_{\mathcal{X}} \int_{\mathcal{T}} \mu^Y(t,\boldsymbol{x})\, Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}),$$

where $\mu^Y(t,\boldsymbol{x}) = \mathbb{E}\left[Y^{(t)}\middle|\boldsymbol{X}=\boldsymbol{x}\right]$. We first show that

$$\int_{\mathcal{T}} \mu^Y(t,\boldsymbol{x})\, Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x}) \to \mu^Y(t_0^\star(\boldsymbol{x}),\boldsymbol{x}) \qquad \text{as } \hbar \to 0. \tag{E.9}$$

Given $\epsilon > 0$, the support-relative continuity in Assumption 6.10(b) implies that we can find $r > 0$ so that $\left|\mu^Y(t,\boldsymbol{x}) - \mu^Y(t_0^\star(\boldsymbol{x}),\boldsymbol{x})\right| < \epsilon$ for $t \in \mathcal{T}(\boldsymbol{x})$ and $|t - t_0^\star(\boldsymbol{x})| < r$. Then,

$$\begin{aligned}
&\left|\int_{\mathcal{T}} \mu^Y(t,\boldsymbol{x})\, Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x}) - \mu^Y(t_0^\star(\boldsymbol{x}),\boldsymbol{x})\right| \\
&\le \epsilon + \underbrace{\int_{|t-t_0^\star(\boldsymbol{x})|\ge r} \left|\mu^Y(t,\boldsymbol{x})\right| Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x})}_{\textbf{Term I}} + \underbrace{\left|\mu^Y(t_0^\star(\boldsymbol{x}),\boldsymbol{x})\right| \cdot Q_{\hbar,t_0}(|t-t_0^\star(\boldsymbol{x})| \ge r|\boldsymbol{X}=\boldsymbol{x})}_{\textbf{Term II}}.
\end{aligned}$$

Now, for every $M > 0$, **Term I** in the above display satisfies

$$\begin{aligned}
\textbf{Term I} &\le M \cdot Q_{\hbar,t_0}(|t-t_0^\star(\boldsymbol{x})| \ge r|\boldsymbol{X}=\boldsymbol{x}) \\
&\quad + \int_{\mathcal{T}} \left|\mu^Y(t,\boldsymbol{x})\right| \cdot \mathbb{1}\left\{\left|\mu^Y(t,\boldsymbol{x})\right| > M\right\} Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x}),
\end{aligned}$$

where the first term vanishes as $\hbar \to 0$ by Theorem 6.8 and the second term can be made arbitrarily small by Assumption 6.11 when M is sufficiently large and $\hbar \to 0$. Furthermore, **Term II** converges to 0 as $\hbar \to 0$ by Theorem 6.8 again. Taking $\epsilon \to 0$ establishes (E.9).

Finally, by the second uniform integrability condition in Assumption 6.11, together with (E.9), we can apply the Vitali convergence theorem under $\mathrm{P}_{\boldsymbol{X}}$ and obtain that

$$\mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right] = \int_{\mathcal{X}} \int_{\mathcal{T}} \mu^Y(t,\boldsymbol{x})\, Q_{\hbar,t_0}(dt|\boldsymbol{X}=\boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \to \int_{\mathcal{X}} \mu^Y(t_0^\star(\boldsymbol{x}),\boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})$$

as $\hbar \to 0$.

(b) The same argument as in (a) applies with $\mu(t,\boldsymbol{x})$ in place of $\mu^Y(t,\boldsymbol{x})$. □

E.6 Proof of Theorem 6.10

The proof of Theorem 6.10 relies on the following important lemma, which controls the average distance between a draw from the tilted intervention and its nearest feasible treatment value, conditional on $\boldsymbol{X} = \boldsymbol{x} \notin \mathcal{X}_0(t_0)$.

Lemma E.1. *Suppose that Assumptions 6.9, 6.12, and 6.14 hold, as well as that the kernel function satisfies Condition 6.2. Then, as $\hbar \to 0$,*

$$\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} O\left(\hbar^\gamma\right) & \text{when } 0 < \gamma \leq 1 \text{ or } \gamma > 1,\ \omega > \gamma - 1, \\ O(\hbar^\gamma |\log \hbar|) & \text{when } \gamma > 1,\ \omega = \gamma - 1, \\ O(\hbar^{\omega+1}) & \text{when } \gamma > 1,\ 0 < \omega < \gamma - 1, \end{cases}$$

under the kernel-smoothed tilted intervention (6.26).

Proof of Lemma E.1. By the strict positivity and continuity of K, together with Condition 6.2(b), there exist constants $0 < \underline{c}_K \leq \bar{C}_K < \infty$ such that

$$\underline{c}_K \exp(-|u|^\gamma) \leq K(u) \leq \bar{C}_K \exp(-|u|^\gamma), \qquad u \in \mathbb{R}. \tag{E.10}$$

Fix $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$ on which Assumption 6.12 holds, and we write $\zeta(\boldsymbol{x}) = |t_0 - t_0^\star(\boldsymbol{x})| > 0$. By the definition of $t_0^\star(\boldsymbol{x})$ with $\boldsymbol{x} \notin \mathcal{X}_0(t_0)$, we know that $|t - t_0| - \zeta(\boldsymbol{x}) \geq 0$.

Without loss of generality, we prove the bound when $t_0 > t_0^\star(\boldsymbol{x})$, and the case where $t_0 < t_0^\star(\boldsymbol{x})$ follows by symmetry. By Assumption 6.12(b), there exist constants $\epsilon_0, p_{\min} > 0$ such that $p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \geq p_{\min} > 0$ for all $t \in [t_0^\star(\boldsymbol{x}) - \epsilon_0, t_0^\star(\boldsymbol{x})]$. Let

$$D_\hbar(\boldsymbol{x}) := \int_{\mathcal{T}(\boldsymbol{x})} K\left(\frac{t - t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\, dt$$

and

$$N_\hbar(\boldsymbol{x}) := \int_{\mathcal{T}(\boldsymbol{x})} |t - t_0^\star(\boldsymbol{x})|\, K\left(\frac{t - t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\, dt.$$

We thus have that $\mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] = \frac{N_\hbar(\boldsymbol{x})}{D_\hbar(\boldsymbol{x})}$.

• **Scenario 1:** $0 < \zeta := \zeta(\boldsymbol{x}) \le \hbar$. For all sufficiently small $\hbar$ with $\hbar \le \epsilon_0$, we calculate that

$$D_\hbar(\boldsymbol{x}) = \int_{\mathcal{T}(\boldsymbol{x})} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\, dt \overset{\text{(i)}}{\ge} p_{\min} \int_0^\hbar K\left(-\frac{\zeta+v}{\hbar}\right) dv \overset{\text{(ii)}}{\gtrsim} \hbar,$$

where (i) follows from the change of variables $t = t_0^\star(\boldsymbol{x}) - v$ on the one-sided support interval and leverages Assumption 6.12(b) and (ii) uses the facts that $0 < \frac{\zeta+v}{\hbar} \le 2$ for all $v \in (0, \hbar]$ and K is positive on $[-2, 2]$ by Condition 6.2. On the other hand,

$$\begin{aligned}
N_\hbar(\boldsymbol{x}) &= \int_{\mathcal{T}(\boldsymbol{x})} |t - t_0^\star(\boldsymbol{x})|\, K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\, dt \\
&\le \bar{C}_K p_{\max} \int_{\mathcal{T}(\boldsymbol{x})} |t - t_0^\star(\boldsymbol{x})| \exp\left(-\frac{|t-t_0|^\gamma}{\hbar^\gamma}\right) dt \\
&\overset{\text{(iii)}}{\le} \bar{C}_K p_{\max} C_g \int_{\mathcal{T}(\boldsymbol{x})} [|t-t_0| - \zeta(\boldsymbol{x})] \exp\left(-\frac{|t-t_0|^\gamma}{\hbar^\gamma}\right) dt \\
&\overset{\text{(iv)}}{\le} \bar{C}_K p_{\max} C_g \int_0^\infty v \exp\left(-\frac{(\zeta+v)^\gamma}{\hbar^\gamma}\right) dv \\
&\le \bar{C}_K p_{\max} C_g \int_0^\infty v \exp\left(-\frac{v^\gamma}{\hbar^\gamma}\right) dv \\
&\overset{\text{(v)}}{\lesssim} \hbar^2,
\end{aligned}$$

where (iii) follows from Assumption 6.12, (iv) uses the change of variable $t = t_0^\star(\boldsymbol{x}) - v$, and (v) applies $v = \hbar y$. Therefore,

$$\mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] \lesssim \hbar, \qquad 0 < \zeta(\boldsymbol{x}) \le \hbar. \tag{E.11}$$

• **Scenario 2:** $\zeta := \zeta(\boldsymbol{x}) > \hbar$. Suppose first that $\hbar^\gamma \zeta^{1-\gamma} \le \epsilon_0$. Again, by the change of variable $t = t_0^\star(\boldsymbol{x}) - v$ and using our arguments in **Scenario 1**, we have that

$$\begin{aligned}
D_\hbar(\boldsymbol{x}) &\ge p_{\min} \underline{c}_K \int_0^{\hbar^\gamma \zeta^{1-\gamma}} \exp\left[-\frac{(\zeta+v)^\gamma}{\hbar^\gamma}\right] dv \\
&= p_{\min} \underline{c}_K \exp\left(-\frac{\zeta^\gamma}{\hbar^\gamma}\right) \int_0^{\hbar^\gamma \zeta^{1-\gamma}} \exp\left[\frac{\zeta^\gamma - (\zeta+v)^\gamma}{\hbar^\gamma}\right] dv \\
&\overset{\text{(vi)}}{\ge} p_{\min} \underline{c}_K \exp\left(-\frac{\zeta^\gamma}{\hbar^\gamma}\right) \int_0^{\hbar^\gamma \zeta^{1-\gamma}} \exp\left(-\frac{C_\gamma \zeta^{\gamma-1} v}{\hbar^\gamma}\right) dv
\end{aligned}$$

$$\gtrsim \hbar^{\gamma}\zeta^{1-\gamma}\exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right),$$

where (vi) follows from the mean value theorem that $(\zeta+v)^{\gamma}-\zeta^{\gamma}\leq C_{\gamma}\zeta^{\gamma-1}v$ with $C_{\gamma}<\infty$ depending only on γ.

For the numerator, we similarly calculate that

$$\begin{aligned}
N_{\hbar}(\boldsymbol{x}) &\leq \bar{C}_K p_{\max} C_g \int_0^{\infty} v\exp\left(-\frac{(\zeta+v)^{\gamma}}{\hbar^{\gamma}}\right)dv \\
&\lesssim \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\int_0^{\infty} v\exp\left(\frac{\zeta^{\gamma}-(\zeta+v)^{\gamma}}{\hbar^{\gamma}}\right)dv \\
&\overset{\text{(vii)}}{\lesssim} \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\left\{\int_0^{\zeta} v\exp\left(-\frac{C'_{\gamma}\zeta^{\gamma-1}v}{\hbar^{\gamma}}\right)dv + \int_{\zeta}^{\infty} v\exp\left[-C''_{\gamma}\left(\frac{v}{\hbar}\right)^{\gamma}\right]dv\right\} \\
&\leq \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\left[\hbar^{2\gamma}\zeta^{2-2\gamma}\int_0^{\zeta^{\gamma}/\hbar^{\gamma}} y\exp\left(-C'_{\gamma}y\right)dy + \hbar^2\int_{\zeta/\hbar}^{\infty} y\exp\left(-C''_{\gamma}y^{\gamma}\right)dy\right] \\
&\leq \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\left[\hbar^{2\gamma}\zeta^{2-2\gamma}\int_0^{\zeta^{\gamma}/\hbar^{\gamma}} y\exp\left(-C'_{\gamma}y\right)dy + \hbar^2\int_{\zeta^{\gamma}/\hbar^{\gamma}}^{\infty} \frac{z^{\frac{2}{\gamma}-1}}{\gamma}\exp\left(-C''_{\gamma}z\right)dz\right] \\
&\overset{\text{(viii)}}{\lesssim} \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\hbar^{2\gamma}\zeta^{2-2\gamma},
\end{aligned}$$

where (vii) follows from the mean value theorem again with $\frac{(\zeta+v)^{\gamma}-\zeta^{\gamma}}{\hbar^{\gamma}}\geq\frac{C'_{\gamma}\zeta^{\gamma-1}v}{\hbar^{\gamma}}$ when $v\in[0,\zeta]$ and $\frac{(\zeta+v)^{\gamma}-\zeta^{\gamma}}{\hbar^{\gamma}}\geq C''_{\gamma}\left(\frac{v}{\hbar}\right)^{\gamma}$ when $v>\zeta$ for two constants $C'_{\gamma},C''_{\gamma}>0$ that depend only on γ, while (viii) uses the bound on an incomplete gamma tail. As a result, $\frac{N_{\hbar}(\boldsymbol{x})}{D_{\hbar}(\boldsymbol{x})}\lesssim\hbar^{\gamma}\zeta^{1-\gamma}$.

Now, if $\hbar^{\gamma}\zeta^{1-\gamma}>\epsilon_0$, which can only occur when $0<\gamma<1$, then

$$\begin{aligned}
D_{\hbar}(\boldsymbol{x}) &\geq p_{\min}\underline{c}_K\exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\int_0^{\epsilon_0}\exp\left[\frac{\zeta^{\gamma}-(\zeta+v)^{\gamma}}{\hbar^{\gamma}}\right]dv \\
&\gtrsim \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\int_0^{\epsilon_0}\exp\left(-\frac{C_{\gamma}\zeta^{\gamma-1}v}{\hbar^{\gamma}}\right)dv \\
&\overset{\text{(iv)}}{\gtrsim} \exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right),
\end{aligned}$$

where (iv) follows from the fact that $\frac{C_{\gamma}\zeta^{\gamma-1}v}{\hbar^{\gamma}}\in[0,C_{\gamma}]$. Additionally,

$$N_{\hbar}(\boldsymbol{x})\lesssim\exp\left(-\frac{\zeta^{\gamma}}{\hbar^{\gamma}}\right)\int_0^{\infty} v\exp\left[\frac{\zeta^{\gamma}-(\zeta+v)^{\gamma}}{\hbar^{\gamma}}\right]dv$$

$$
\begin{aligned}
&\lesssim \exp\left(-\frac{\zeta^\gamma}{\hbar^\gamma}\right) \sup_{v\geq 0} v \exp\left[\frac{\zeta^\gamma - (\zeta+v)^\gamma}{\hbar^\gamma}\right] \\
&\lesssim \exp\left(-\frac{\zeta^\gamma}{\hbar^\gamma}\right) \begin{cases} \sup_{0\leq v\leq \zeta} v \exp\left(-\frac{C'_\gamma \zeta^{\gamma-1} v}{\hbar^\gamma}\right), \\ \sup_{v>\zeta} v \exp\left[-C''_\gamma \left(\frac{v}{\hbar}\right)^\gamma\right] \end{cases} \\
&\lesssim \hbar^\gamma \zeta^{1-\gamma} \exp\left(-\frac{\zeta^\gamma}{\hbar^\gamma}\right).
\end{aligned}
$$

As a result,

$$
\mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] \lesssim \hbar^\gamma \zeta(\boldsymbol{x})^{1-\gamma}, \qquad \zeta(\boldsymbol{x}) \geq \hbar. \tag{E.12}
$$

We now integrate (E.11) and (E.12) over $\boldsymbol{X}$ and derive that

$$
\begin{aligned}
&\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\
&\lesssim \hbar \cdot \mathrm{P}_{\boldsymbol{X}}(0 < \zeta(\boldsymbol{X}) \leq \hbar) + \hbar^\gamma \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right] \\
&\overset{\text{(x)}}{\lesssim} \hbar^{1+\omega} + \hbar^\gamma \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right],
\end{aligned}
$$

where (x) applies the margin condition in Assumption 6.14. To derive the rate of convergence for the second term $\hbar^\gamma \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right]$, we consider two cases for the value of $\gamma > 0$.

• **Case A:** $0 < \gamma \leq 1$. Since $\mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right] < \infty$ under Assumption 6.14, we have that

$$
\begin{aligned}
\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) &\lesssim \hbar^{1+\omega} + \hbar^\gamma \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right] \\
&\lesssim \hbar^{\omega+1} + \hbar^\gamma \asymp \hbar^\gamma.
\end{aligned}
$$

• **Case B:** $\gamma > 1$. Then, $\mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > \hbar\right\}\right] < \infty$ automatically satisfies, and we have that

$$
\begin{aligned}
&\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\
&\lesssim \hbar^{1+\omega} + \hbar^\gamma \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\hbar < \zeta(\boldsymbol{X}) \leq r_0\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \hbar^{\gamma} \cdot \mathbb{E}\left[\zeta(\boldsymbol{X})^{1-\gamma} \cdot \mathbb{1}\left\{\zeta(\boldsymbol{X}) > r_0\right\}\right] \\
&\overset{\text{(xi)}}{\lesssim} \hbar^{\omega+1} + \hbar^{\gamma} \sum_{j=0}^{J_\hbar} (2^j \hbar)^{1-\gamma} \cdot \mathrm{P}_{\boldsymbol{X}}\left(2^j\hbar < \zeta(\boldsymbol{X}) \le 2^{j+1}\hbar\right) + \hbar^{\gamma} \\
&\overset{\text{(xii)}}{\lesssim} \hbar^{\omega+1} + \hbar \sum_{j=0}^{J_\hbar} 2^{j(1-\gamma)} \cdot (2^{j+1}\hbar)^{\omega} + \hbar^{\gamma} \\
&\lesssim \hbar^{\omega+1}\left(1 + \sum_{j=1}^{J_\hbar} 2^{j(\omega+1-\gamma)}\right) + \hbar^{\gamma},
\end{aligned}
$$

where $r_0 > 0$ is defined in Assumption 6.14, (xi) applies the dyadic upper bound with $J_\hbar$ the largest integer such that $2^{J_\hbar+1}\hbar \le r_0$, and (xii) leverages Assumption 6.14.

Now, when $0 < \omega < \gamma - 1$, we know that $\sum_{j=1}^{J_\hbar} 2^{j(\omega+1-\gamma)}$ is bounded. When $\omega = \gamma - 1$, then $\sum_{j=1}^{J_\hbar} 2^{j(\omega+1-\gamma)} = J_\hbar \asymp |\log \hbar|$. When $\omega > \gamma - 1$, we know that $\sum_{j=1}^{J_\hbar} 2^{j(\omega+1-\gamma)} \asymp \left(\frac{r_0}{\hbar}\right)^{\omega+1-\gamma}$. As a result, we obtain that

$$
\begin{aligned}
&\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\
&\lesssim \hbar^{\omega+1}\left(1 + \sum_{j=1}^{J_\hbar} 2^{j(\omega+1-\gamma)}\right) + \hbar^{\gamma} \\
&\lesssim \begin{cases} \hbar^{\omega+1} + \hbar^{\gamma} & \text{when } 0 < \omega < \gamma - 1, \\ \hbar^{1+\omega}|\log \hbar| + \hbar^{\gamma} & \text{when } \omega = \gamma - 1, \\ \hbar^{\omega+1}\left[1 + \left(\frac{r_0}{\hbar}\right)^{\omega+1-\gamma}\right] + \hbar^{\gamma} & \text{when } \omega > \gamma - 1, \end{cases} \\
&\asymp \begin{cases} \hbar^{\omega+1} & \text{when } 0 < \omega < \gamma - 1, \\ \hbar^{\gamma}|\log \hbar| & \text{when } \omega = \gamma - 1, \\ \hbar^{\gamma} & \text{when } \omega > \gamma - 1. \end{cases}
\end{aligned}
$$

Summarizing all these cases, we conclude that

$$
\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}}\left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x}\right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} O\left(\hbar^{\gamma}\right) & \text{when } 0 < \gamma \le 1 \text{ or } \gamma > 1,\ \omega > \gamma - 1, \\ O(\hbar^{\gamma}|\log \hbar|) & \text{when } \gamma > 1,\ \omega = \gamma - 1, \\ O(\hbar^{\omega+1}) & \text{when } \gamma > 1,\ 0 < \omega < \gamma - 1, \end{cases}
$$

The result thus follows. □

Remark E.2. *The proof of Lemma E.1 also yields analogous bounds for the r-th moment with $r > 0$ being an integer. Under the same assumptions in Lemma E.1, and assuming in addition that $\mathbb{E}\left[\zeta(\boldsymbol{X})^{r(1-\gamma)} \cdot \mathbb{1}_{\{\zeta(\boldsymbol{X})>r_0\}}\right] < \infty$, we obtain the following rates as $\hbar \to 0$:*

$$\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}} |T - t_0^\star(\boldsymbol{x})|^r \, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} O(\hbar^{r\gamma}) & \text{when } 0 < \gamma \le 1 \text{ or } \gamma > 1,\ \omega > r(\gamma-1), \\ O(\hbar^{r\gamma}|\log\hbar|) & \text{when } \gamma > 1,\ \omega = r(\gamma-1), \\ O(\hbar^{r+\omega}) & \text{when } \gamma > 1,\ 0 < \omega < r(\gamma-1). \end{cases}$$

Proof of Theorem 6.10. By the definition of $\bar{m}_\hbar(t_0) = \mathbb{E}\left[Y^{(q_{\hbar,t_0})}\right]$ under (6.26) and Assumption 6.2, we know that

$$\begin{aligned} \bar{m}_\hbar(t_0) &= \int_{\mathcal{X}} \int_{\mathcal{T}(\boldsymbol{x})} \mu(t,\boldsymbol{x}) \left[\frac{\eta(t;t_0,\hbar)\cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} \eta(u;t_0,\hbar)\cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}\right] dt\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\ &= \underbrace{\int_{\mathcal{X}_0(t_0)} \int_{\mathcal{T}(\boldsymbol{x})} \mu(t,\boldsymbol{x}) \left[\frac{K\left(\frac{t-t_0}{\hbar}\right)\cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}\right] dt\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})}_{\textbf{Term I}} \\ &\quad + \underbrace{\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \int_{\mathcal{T}(\boldsymbol{x})} \mu(t,\boldsymbol{x}) \left[\frac{K\left(\frac{t-t_0}{\hbar}\right)\cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}\right] dt\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})}_{\textbf{Term II}}. \end{aligned}$$

• **Term I:** Fix $\boldsymbol{x} \in \mathcal{X}_0(t_0)$ on which Assumptions 6.12 and 6.13 hold. Under Assumption 6.13, there exists a neighborhood $(t_0-\epsilon_0, t_0+\epsilon_0) \subset \mathcal{T}(\boldsymbol{x})$. Let $g_{\boldsymbol{x}}(t) := \mu(t,\boldsymbol{x})\cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$. By the change of variables $t = t_0 + \hbar u$, we know that $\textbf{Term I} = \int_{\mathcal{X}_0(t_0)} \frac{A_\hbar(\boldsymbol{x})}{B_\hbar(\boldsymbol{x})}\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})$, where

$$\begin{aligned} A_\hbar(\boldsymbol{x}) &:= \int_{\mathbb{R}} g_{\boldsymbol{x}}(t_0+\hbar u) K(u) \mathbb{1}\{t_0+\hbar u \in \mathcal{T}(\boldsymbol{x})\}\, du, \\ B_\hbar(\boldsymbol{x}) &:= \int_{\mathbb{R}} p_{T|\boldsymbol{X}}(t_0+\hbar u) K(u) \mathbb{1}\{t_0+\hbar u \in \mathcal{T}(\boldsymbol{x})\}\, du. \end{aligned}$$

Now, Condition 6.2 implies that as $\hbar \to 0$, $\int_{|u|>\epsilon_0/\hbar} \hbar^j |u|^j K(u)\, du = o(\hbar^2)$ for $j = 0,1,2$. In particular, the contribution to $A_\hbar(\boldsymbol{x})$ from $|u| > \frac{\epsilon_0}{\hbar}$ is bounded by

$p_{\max} M_\mu(\boldsymbol{x}) \int_{|u|>\epsilon_0/\hbar} K(u)\, du$, whose integral over $\boldsymbol{x}$ is $o(\hbar^2)$ by Assumption 6.13(b). Additionally, the corresponding tail contribution to $B_\hbar(\boldsymbol{x})$ is bounded by $p_{\max} \int_{|u|>\epsilon_0/\hbar} K(u)\, du$. Hence, we can focus on $|u| \le \frac{\epsilon_0}{\hbar}$, and both $t \mapsto g_{\boldsymbol{x}}(t)$ and $t \mapsto p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ are twice continuously differentiable by Assumption 6.13(a). By Taylor's theorem,

$$g_{\boldsymbol{x}}(t_0 + \hbar u) = g_{\boldsymbol{x}}(t_0) + \hbar u \frac{\partial}{\partial t} g_{\boldsymbol{x}}(t_0) + \frac{\hbar^2 u^2}{2} \frac{\partial^2}{\partial t^2} g_{\boldsymbol{x}}(t_0) + \hbar^2 u^2 r_{g,\hbar}(u, \boldsymbol{x}),$$
$$p_{T|\boldsymbol{X}}(t_0 + \hbar u) = p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) + \hbar u \frac{\partial}{\partial t} p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) + \frac{\hbar^2 u^2}{2} \frac{\partial^2}{\partial t^2} p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) + \hbar^2 u^2 r_{p,\hbar}(u, \boldsymbol{x}),$$

where, for every fixed $(u, \boldsymbol{x})$, $r_{g,\hbar}(u, \boldsymbol{x}), r_{p,\hbar}(u, \boldsymbol{x}) \to 0$ as $\hbar \to 0$, and both remainders are uniformly bounded. Since $u^2 K(u)$ is integrable, dominated convergence, together with the moment conditions in Condition 6.2(a), yields

$$A_\hbar(\boldsymbol{x}) = g_{\boldsymbol{x}}(t_0) + \frac{\kappa_2 \hbar^2}{2} \frac{\partial^2}{\partial t^2} g_{\boldsymbol{x}}(t_0) + r_{A,\hbar}(\boldsymbol{x}), \quad B_\hbar(\boldsymbol{x}) = p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) + \frac{\kappa_2 \hbar^2}{2} \frac{\partial^2}{\partial t^2} p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) + r_{B,\hbar}(\boldsymbol{x}),$$

where

$$\int_{\mathcal{X}_0(t_0)} |r_{A,\hbar}(\boldsymbol{x})| d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = o(\hbar^2), \qquad \int_{\mathcal{X}_0(t_0)} |r_{B,\hbar}(\boldsymbol{x})| d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = o(\hbar^2).$$

Furthermore, for all sufficiently small $\hbar$, $B_\hbar(\boldsymbol{x}) \ge p_{\min} \int_{|u|\le 1} K(u)\, du =: c_B > 0$ uniformly for some constant $c_B > 0$. The bounded derivatives in Assumption 6.13(a) also give $B_\hbar(\boldsymbol{x}) - p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x}) = O(\hbar^2)$ uniformly. Hence,

$$\frac{A_\hbar(\boldsymbol{x})}{B_\hbar(\boldsymbol{x})} = \mu(t_0, \boldsymbol{x}) + \frac{\kappa_2 \hbar^2}{2} \left\{ \frac{\frac{\partial^2}{\partial t^2} g_{\boldsymbol{x}}(t_0)}{p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})} - \frac{g_{\boldsymbol{x}}(t_0) \frac{\partial^2}{\partial t^2} p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})}{p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})^2} \right\} + r_{I,\hbar}(\boldsymbol{x}),$$

where $\int_{\mathcal{X}_0(t_0)} |r_{I,\hbar}(\boldsymbol{x})| d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) = o(\hbar^2)$. Because $g_{\boldsymbol{x}}(t) = \mu(t, \boldsymbol{x}) \cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$,

$$\frac{\frac{\partial^2}{\partial t^2} g_{\boldsymbol{x}}(t_0)}{p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})} - \frac{g_{\boldsymbol{x}}(t_0) \frac{\partial^2}{\partial t^2} p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})}{p_{T|\boldsymbol{X}}(t_0|\boldsymbol{x})^2} = \frac{\partial^2}{\partial t^2} \mu(t_0, \boldsymbol{x}) + 2 \frac{\partial}{\partial t} \mu(t_0, \boldsymbol{x}) \frac{\partial}{\partial t} \log p_{T|\boldsymbol{X}}(t|\boldsymbol{x}).$$

Integrating over $\mathcal{X}_0(t_0)$ therefore gives that

$$\textbf{Term I} = \int_{\mathcal{X}_0(t_0)} \mu(t_0, \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) + \underbrace{\frac{\kappa_2 \hbar^2}{2} \int_{\mathcal{X}_0(t_0)} \left[\frac{\partial^2}{\partial t^2} \mu(t_0, \boldsymbol{x}) + 2 \frac{\partial}{\partial t} \mu(t, \boldsymbol{x}) \cdot \frac{\partial}{\partial t} \log p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \Big|_{t=t_0} \right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})}_{=\Delta_{\mathrm{int}}(\hbar)} + o(\hbar^2),$$

where $\kappa_2 = \int_{\mathbb{R}} u^2 K(u)\, du < \infty$.

• **Term II:** Under positivity at t_0, $t_0 \in \mathcal{T}(\boldsymbol{X})$ almost surely, so $\mathrm{P}_{\boldsymbol{X}}(\mathcal{X} \setminus \mathcal{X}_0(t_0)) = 0$ and **Term II** is equal to zero.

Now, we consider the general case when the positivity condition is violated and the kernel function satisfies Condition 6.2. For those $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_0(t_0)$, we know from Lemma E.1 that

$$
\begin{aligned}
&\left| \textbf{Term II} - \int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \right| \\
&= \left| \int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \int_{\mathcal{T}(\boldsymbol{x})} \mu(t, \boldsymbol{x}) \left[\frac{K\left(\frac{t-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\, du} \right] dt\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) - \int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \right| \\
&\leq \int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \int_{\mathcal{T}(\boldsymbol{x})} |\mu(t, \boldsymbol{x}) - \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})|\, Q_{\hbar,t_0}(dt|\boldsymbol{x}) d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\
&\overset{\text{(iii)}}{\leq} L_\mu \int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mathbb{E}_{Q_{\hbar,t_0}} \left[|T - t_0^\star(\boldsymbol{x})| \mid \boldsymbol{X} = \boldsymbol{x} \right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) \\
&\overset{\text{(iv)}}{=} \begin{cases} O\left(\hbar^\gamma\right) & \text{when } 0 < \gamma \leq 1 \text{ or } \gamma > 1,\ \omega > \gamma - 1, \\ O(\hbar^\gamma |\log \hbar|) & \text{when } \gamma > 1,\ \omega = \gamma - 1, \\ O(\hbar^{\omega+1}) & \text{when } \gamma > 1,\ 0 < \omega < \gamma - 1, \end{cases},
\end{aligned}
$$

where (iii) follows from Assumption 6.13(c) and (iv) applies Lemma E.1.

In summary,

$$
\bar{m}_\hbar(t_0) = \int_{\mathcal{X}_0(t_0)} \mu(t_0, \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) + \underbrace{\int_{\mathcal{X}\setminus\mathcal{X}_0(t_0)} \mu(t_0^\star(\boldsymbol{x}), \boldsymbol{x})\, d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x})}_{\textbf{Projection Term}} + \Delta_{\text{int}}(\hbar) + \Delta_{\text{proj}}(\hbar),
$$

where

$$
\Delta_{\text{int}}(\hbar) = \frac{\kappa_2 \hbar^2}{2} \int_{\mathcal{X}_0(t_0)} \left[\frac{\partial^2}{\partial t^2} \mu(t_0, \boldsymbol{x}) + 2 \frac{\partial}{\partial t} \mu(t, \boldsymbol{x}) \cdot \frac{\partial}{\partial t} \log p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \Big|_{t=t_0} \right] d\mathrm{P}_{\boldsymbol{X}}(\boldsymbol{x}) + o(\hbar^2),
$$

and

$$
|\Delta_{\text{proj}}(\hbar)| = \begin{cases} O\left(\hbar^\gamma\right) & \text{when } 0 < \gamma \leq 1 \text{ or } \gamma > 1,\ \omega > \gamma - 1, \\ O(\hbar^\gamma |\log \hbar|) & \text{when } \gamma > 1,\ \omega = \gamma - 1, \\ O(\hbar^{\omega+1}) & \text{when } \gamma > 1,\ 0 < \omega < \gamma - 1, \end{cases}
$$

where $\gamma > 0$. The result thus follows. □

E.7 Proof of Proposition 6.12

The proof of Proposition 6.12 relies on the following critical lemma, which establishes an $\hbar$-dependent lower bound for the variance of the efficient influence function of our target parameter $\bar{m}_\hbar(t)$ for any fixed tilting parameter $\hbar > 0$ and an upper bound for a local least favorable submodel supported on an interior set $\mathcal{A}_0(t_0)$ for the target level t_0 defined in Assumption 6.15.

Lemma E.2. *Suppose that Assumption 6.15 and Condition 6.2 hold. Then, there exists $\hbar_0 > 0$ such that, for every $0 < \hbar \le \hbar_0$,*

$$V_{\mathrm{eff},\bar{m}_\hbar}(t_0) \ge \frac{\sigma^2_{\min} \cdot p_{\min} \cdot \mathrm{P}_{\boldsymbol{X}}(\mathcal{A}_0(t_0)) \cdot \int_{-1}^{1} K^2(u)\,du}{4\hbar \cdot p^2_{\max}} \gtrsim \frac{1}{\hbar},$$

where $\mathcal{A}_0(t_0)$ is a measurable set defined in Assumption 6.15(a). Furthermore, we have

$$\mathbb{E}\left[\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})} \cdot \mathrm{Var}\left(Y|T,\boldsymbol{X}\right) \cdot \mathbb{1}_{\{\boldsymbol{X}\in\mathcal{A}_0(t_0)\}}\right] \le \frac{2B^2_{\max} \cdot p_{\max} \int_{-\infty}^{\infty} K^2(u)\,du}{\hbar \cdot p^2_{\min}\left[\int_{-1}^{1} K(v)\,dv\right]^2} \lesssim \frac{1}{\hbar}.$$

Proof of Lemma E.2. Recall from Proposition 6.11 that the nonparametric efficiency bound for $\bar{m}_\hbar(t_0)$ for any fixed $\hbar > 0$ is given by

$$\mathbb{E}\left\{\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left(\mathrm{Var}\left(Y|T,\boldsymbol{X}\right) + \left[\mu(T,\boldsymbol{X}) - \mathbb{E}_{Q_{\hbar,t_0}}\left(\mu(T,\boldsymbol{X})|\boldsymbol{X}\right)\right]^2\right)\right\} + \mathrm{Var}\left\{\mathbb{E}_{Q_{\hbar,t_0}}\left[\mu(T,\boldsymbol{X})|\boldsymbol{X}\right]\right\}.$$

We first derive the uniform upper and lower bounds for $\int_{\mathcal{T}} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt$ on $\mathcal{A}_0(t_0)$. By Assumption 6.15(a), for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$,

$$\begin{aligned}
\int_{\mathcal{T}} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt &\ge p_{\min} \int_{|t-t_0|\le\epsilon_1} K\left(\frac{t-t_0}{\hbar}\right) dt \\
&= p_{\min}\hbar \int_{|u|\le\epsilon_1/\hbar} K(u)\,du \\
&\ge p_{\min}\hbar \int_{-1}^{1} K(u)\,du
\end{aligned}$$

when $\hbar \le \epsilon_1$. For the upper bound, we split the integral and derive that

$$\int_{\mathcal{T}} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt = \int_{|t-t_0|\le\epsilon_1} K\left(\frac{t-t_0}{\hbar}\right) dt + \int_{|t-t_0|>\epsilon_1} K\left(\frac{t-t_0}{\hbar}\right) dt$$

$$\leq p_{\max}\hbar \int_{|u|\leq \epsilon_1/\hbar} K(u)\,du + \sup_{|u|>\epsilon_1/\hbar} K(u)$$
$$\leq p_{\max}\hbar + o(\hbar),$$

where we use the exponential tail bound in Condition 6.2 to argue that $\sup_{|u|>\epsilon_1/\hbar} K(u) = o(\hbar)$. Thus, there exists $\hbar_0 \in (0, \epsilon_1)$ such that for every $0 < \hbar \leq \hbar_0$,

$$\int_{\mathcal{T}} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt \leq 2p_{\max}\hbar$$

for $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$. Similarly, we can derive that

$$\int_{\mathcal{T}} K^2\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt \leq 2p_{\max}\hbar \int_{\mathbb{R}} K^2(u)\,du$$

for every $0 < \hbar \leq \hbar_0$ and $\mathrm{P}_{\boldsymbol{X}}$-almost every $\boldsymbol{x} \in \mathcal{A}_0(t_0)$.

- **Lower bound:** Note that

$$\mathbb{E}\left\{\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[\mu(T,\boldsymbol{X}) - \mathbb{E}_{Q_{\hbar,t_0}}\left(\mu(T,\boldsymbol{X})|\boldsymbol{X}\right)\right]^2\right\} \geq 0 \quad \text{and} \quad \mathrm{Var}\left\{\mathbb{E}_{Q_{\hbar,t_0}}\left[\mu(T,\boldsymbol{X})|\boldsymbol{X}\right]\right\} \geq 0.$$

Hence, the nonparametric efficiency bound for $\bar{m}_\hbar(t_0)$ is lower bounded by

$$V_{\mathrm{eff},\bar{m}_\hbar}(t_0) \geq \mathbb{E}\left[\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})} \cdot \mathrm{Var}\left(Y|T,\boldsymbol{X}\right)\right].$$

By Assumption 6.15(b), we may restrict the expectation to $\{\boldsymbol{X} \in \mathcal{A}_0(t_0),\ |T-t_0| < \epsilon_1\}$ and obtain that

$$\begin{aligned}
V_{\mathrm{eff},\bar{m}_\hbar}(t_0) &\geq \sigma^2_{\min} \cdot \mathbb{E}\left[\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})} \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right] \\
&\geq \sigma^2_{\min} \cdot \mathbb{E}\left[\frac{\int_{|t-t_0|\leq\epsilon_1} K^2\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt}{\left[\int_{\mathcal{T}} K\left(\frac{t-t_0}{\hbar}\right) p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\,dt\right]^2} \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right] \\
&\geq \frac{p_{\min}}{p^2_{\max}} \cdot \mathrm{P}_{\boldsymbol{X}}(\mathcal{A}_0(t_0)) \cdot \mathbb{E}\left\{\frac{\int_{|t_1-t_0|\leq\epsilon_1} K^2\left(\frac{t_1-t_0}{\hbar}\right)\,dt_1}{\left[\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\,du\right]^2}\,\middle|\,\mathcal{A}_0(t_0)\right\} \\
&\geq \frac{\sigma^2_{\min} \cdot p_{\min} \cdot \mathrm{P}_{\boldsymbol{X}}(\mathcal{A}_0(t_0)) \cdot \int_{-1}^{1} K^2(u)\,du}{4\hbar p^2_{\max}},
\end{aligned}$$

where the last inequality uses our bounds for the integrated kernel function above.

- **Upper bound:** By Assumption 6.15(b), $|Y| \leq B_{\max}$ almost surely, and hence,

$$\operatorname{Var}(Y|T, \boldsymbol{X}) \leq \mathbb{E}\left(Y^2|T, \boldsymbol{X}\right) \leq B_{\max}^2 \quad \text{almost surely.}$$

Furthermore, we know that

$$\begin{aligned}
&\mathbb{E}\left[\frac{q_{\hbar,t_0}^2(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}^2(T|\boldsymbol{X})} \cdot \operatorname{Var}(Y|T, \boldsymbol{X}) \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right] \\
&\leq B_{\max}^2 \cdot \mathbb{E}\left\{\frac{\int_{\mathcal{T}(\boldsymbol{X})} K^2\left(\frac{t_1-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\, dt_1}{\left[\int_{\mathcal{T}(\boldsymbol{X})} K\left(\frac{u-t_0}{\hbar}\right) \cdot p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\, du\right]^2} \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right\} \\
&\leq \frac{B_{\max}^2 \cdot p_{\max}}{p_{\min}^2} \cdot \mathbb{E}\left\{\frac{\int_{t_0+u\hbar \in \mathcal{T}(\boldsymbol{X})} K^2(u)\, du}{\hbar\left[\int_{t_0+v\hbar \in \mathcal{T}(\boldsymbol{X})} K(v)\, dv\right]^2} \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right\} \\
&\overset{\text{(i)}}{\leq} \frac{2B_{\max}^2 \cdot p_{\max} \int_{-\infty}^{\infty} K^2(u)\, du}{\hbar \cdot p_{\min}^2 \left[\int_{-\epsilon_1/\hbar}^{\epsilon_1/\hbar} K(v)\, dv\right]^2} \\
&\leq \frac{2B_{\max}^2 \cdot p_{\max} \int_{-\infty}^{\infty} K^2(u)\, du}{\hbar \cdot p_{\min}^2 \left[\int_{-1}^{1} K(v)\, dv\right]^2}
\end{aligned}$$

when $\hbar < \epsilon_1$, where (i) follows from our previous calculations on the integrated kernel functions restricted to $\boldsymbol{x} \in \mathcal{A}_0(t_0)$ satisfying Assumption 6.15(a). The result thus follows. □

Proof of Proposition 6.12. The proof relies on Lemma E.2 and Le Cam's two-point method (Tsybakov, 2008). Specifically, the calculations in Lemma E.2 imply that there exist constants $0 < C_- < C_+ < \infty$ and $\hbar_0 > 0$ so that

$$\frac{C_-}{\hbar} \leq \mathbb{E}\left[\frac{q_{\hbar,t_0}^2(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}^2(T|\boldsymbol{X})} \cdot \operatorname{Var}(Y|T, \boldsymbol{X}) \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right] \leq \frac{C_+}{\hbar}$$

with $C_- = \frac{\sigma_{\min}^2 \cdot p_{\min} \cdot \mathrm{P}_{\boldsymbol{X}}(\mathcal{A}_0(t_0)) \cdot \int_{-1}^{1} K^2(u)\, du}{4p_{\max}^2}$ and $C_+ = \frac{2B_{\max}^2 \cdot p_{\max} \int_{\mathbb{R}} K^2(u)\, du}{p_{\min}^2 \left[\int_{-1}^{1} K(v)\, dv\right]^2}$

Let $\mathrm{P}_0, \mathrm{P}_1$ be the null and alternative distributions in $\mathcal{P}$ such that $\mathrm{P}_0 = \otimes_{i=1}^n P_{0i}$ and $\mathrm{P}_1 = \otimes_{i=1}^n P_{1i}$. Here, we define the density functions of P_{0i} and P_{1i} by

$$p_0(y, t, \boldsymbol{x}) = p(y|t, \boldsymbol{x}) \cdot p_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \cdot p(\boldsymbol{x}),$$

$$p_1(y,t,\boldsymbol{x}) = \left\{p(y|t,\boldsymbol{x})\left[1+\epsilon_n\cdot\phi_{\hbar,t_0}(y,t,\boldsymbol{x})\right]\right\}p_{T|\boldsymbol{X}}(t|\boldsymbol{x})\cdot p(\boldsymbol{x}),$$

where $\epsilon_n \in \mathbb{R}$ is sufficiently small and $\phi_{\hbar,t_0}(y,t,\boldsymbol{x}) = \frac{q_{\hbar,t_0}(t|\boldsymbol{x})}{p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}\left[y-\mu(t,\boldsymbol{x})\right]\cdot\mathbb{1}\left\{\boldsymbol{x}\in\mathcal{A}_0(t_0)\right\}$ under Assumption 6.15. Notice that we can also replace $[y-\mu(t,\boldsymbol{x})]$ with any bounded, mean-zero conditional score function in $\phi_{\hbar,t_0}(y,t,\boldsymbol{x})$ for p_1 to be an eligible density function. We denote our target parameter $\bar{m}_\hbar(t_0)$ under distributions P_0 and P_1 by $\bar{m}_\hbar(\mathrm{P}_0;t_0)$ and $\bar{m}_\hbar(\mathrm{P}_1;t_0)$, respectively.

By Le Cam's two-point method (Theorem 2.2(ii) in Tsybakov 2008), we know that if the Hellinger distance $H^2(\mathrm{P}_0,\mathrm{P}_1)\leq\alpha<2$ and $\bar{m}_\hbar(\mathrm{P}_1;t_0)-\bar{m}_\hbar(\mathrm{P}_0;t_0)\geq s>0$, then

$$\inf_{\widehat{m}_\hbar(t_0)}\sup_{\mathrm{P}\in\mathcal{P}}\mathbb{E}_{\mathrm{P}}\left[\widehat{m}_\hbar(t_0)-\bar{m}_\hbar(t_0)\right]^2 \geq \frac{s^2}{2}\left[\frac{1-\sqrt{\alpha(1-\alpha/4)}}{2}\right].$$

Thus, it remains to find a lower bound for the functional separation $\bar{m}_\hbar(\mathrm{P}_1;t_0)-\bar{m}_\hbar(\mathrm{P}_0;t_0)$ as s and an upper bound for the Hellinger distance $H^2(\mathrm{P}_0,\mathrm{P}_1)$ as α.

• **Lower bound for $\bar{m}_\hbar(\mathrm{P}_1;t_0)-\bar{m}_\hbar(\mathrm{P}_0;t_0)$:** Since $\mathbb{E}\left[\phi_{\hbar,t_0}(Y,T,\boldsymbol{X})\right]=0$, we know that $p_0(t|\boldsymbol{x})=p_1(t|\boldsymbol{x})=p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ and $p_0(\boldsymbol{x})=p_1(\boldsymbol{x})=p(\boldsymbol{x})$ for any $(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}$. Now, we evaluate the functional separation between $\bar{m}_\hbar(t_0):=\bar{m}_\hbar(\mathrm{P};t_0)$ under distributions $\mathrm{P}_0,\mathrm{P}_1$ and the kernel-smoothed tilted intervention (6.26) as:

$$\begin{aligned}
\bar{m}_\hbar(\mathrm{P}_1;t_0)-\bar{m}_\hbar(\mathrm{P}_0;t_0) &= \int_{\mathcal{X}}\int_{\mathcal{T}}\int_{\mathcal{Y}}\frac{y\cdot K\left(\frac{t_1-t_0}{\hbar}\right)\cdot\left[p_1(y,t_1,\boldsymbol{x})-p_0(y,t_1,\boldsymbol{x})\right]}{\int_{\mathcal{T}}K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}\,dydt_1d\boldsymbol{x}\\
&= \epsilon_n\int_{\mathcal{X}}\int_{\mathcal{T}}\int_{\mathcal{Y}}\frac{y\cdot K\left(\frac{t_1-t_0}{\hbar}\right)\cdot\phi_{\hbar,t_0}(y,t_1,\boldsymbol{x})\cdot p(y,t_1,\boldsymbol{x})}{\int_{\mathcal{T}}K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{x})\,du}\,dydt_1d\boldsymbol{x}\\
&= \epsilon_n\int_{\mathcal{X}}\int_{\mathcal{T}}\int_{\mathcal{Y}}\frac{y\cdot q^2_{\hbar,t_0}(t_1|\boldsymbol{x})\cdot\left[y-\mu(t_1,\boldsymbol{x})\right]\cdot p(y,t_1,\boldsymbol{x})}{p^2_{T|\boldsymbol{X}}(t_1|\boldsymbol{x})}\,dydt_1d\boldsymbol{x}\\
&= \epsilon_n\int_{\mathcal{X}}\int_{\mathcal{T}}\int_{\mathcal{Y}}\frac{q^2_{\hbar,t_0}(t_1|\boldsymbol{x})\cdot\left[y-\mu(t_1,\boldsymbol{x})\right]^2\cdot p(y,t_1,\boldsymbol{x})}{p^2_{T|\boldsymbol{X}}(t_1|\boldsymbol{x})}\,dydt_1d\boldsymbol{x}\\
&= \epsilon_n\cdot\mathbb{E}\left[\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\cdot\mathrm{Var}\left(Y|T,\boldsymbol{X}\right)\right]\\
&\geq \frac{\epsilon_n\cdot C_-}{\hbar},
\end{aligned}$$

where the last inequality follows from our calculations in the proof of Lemma E.2.

• **Upper bound for $H^2(\mathrm{P}_0, \mathrm{P}_1)$:** By Eq. (2.27) in Tsybakov (2008), we know that $H^2(\mathrm{P}_0, \mathrm{P}_1) \leq \chi^2(\mathrm{P}_0, \mathrm{P}_1)$. Furthermore, for product measures $\mathrm{P}_0 = \otimes_{i=1}^n P_{0i}$ and $\mathrm{P}_1 = \otimes_{i=1}^n P_{1i}$, we know from Section 2.4 of Tsybakov (2008) that

$$\chi^2(\mathrm{P}_0, \mathrm{P}_1) = \prod_{i=1}^n \left[1 + \chi^2(p_{0i}, p_{1i})\right] - 1.$$

Each of these χ^2 divergences $\chi^2(p_{0i}, p_{1i})$ can be bounded above as:

$$\begin{aligned}
\chi^2(p_{0i}, p_{1i}) &= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \frac{p_1^2(y, t_1, \boldsymbol{x})}{p_0(y, t_1, \boldsymbol{x})} \, dy dt_1 d\boldsymbol{x} - 1 \\
&= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} p(y, t_1, \boldsymbol{x}) \left[1 + \epsilon_n \cdot \phi_{\hbar, t_0}(y, t_1, \boldsymbol{x})\right]^2 dy dt_1 d\boldsymbol{x} - 1 \\
&\overset{\text{(i)}}{=} \epsilon_n^2 \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \phi_{\hbar, t_0}^2(y, t_1, \boldsymbol{x}) \cdot p(y, t_1, \boldsymbol{x}) \, dy dt_1 d\boldsymbol{x} \\
&\overset{\text{(ii)}}{=} \epsilon_n^2 \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \frac{q_{\hbar, t_0}^2(t_1 | \boldsymbol{x}) \cdot \left[y - \mu(t_1, \boldsymbol{x})\right]^2 \cdot p(y, t_1, \boldsymbol{x}) \cdot \mathbb{1}\left\{\boldsymbol{x} \in \mathcal{A}_0(t_0)\right\}}{p_{T|\boldsymbol{X}}^2(t_1 | \boldsymbol{x})} \, dy dt_1 d\boldsymbol{x} \\
&= \epsilon_n^2 \cdot \mathbb{E}\left[\frac{q_{\hbar, t_0}^2(T | \boldsymbol{X})}{p_{T|\boldsymbol{X}}^2(T | \boldsymbol{X})} \cdot \mathrm{Var}\left(Y | T, \boldsymbol{X}\right) \cdot \mathbb{1}\left\{\boldsymbol{X} \in \mathcal{A}_0(t_0)\right\}\right] \\
&\overset{\text{(iii)}}{\leq} \epsilon_n^2 \cdot \frac{C_+}{\hbar},
\end{aligned}$$

where (i) uses the facts that $\int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} p(y, t_1, \boldsymbol{x}) \, dy dt_1 d\boldsymbol{x} = 1$ and $\int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \phi_{\hbar, t_0}(y, t_1, \boldsymbol{x}) \cdot p(y, t_1, \boldsymbol{x}) \, dy dt_1 d\boldsymbol{x} = 0$, (ii) applies our calculations in the previous display, and (iii) leverages Lemma E.2. Thus,

$$\begin{aligned}
\log\left\{\prod_{i=1}^n \left[1 + \chi^2(p_{0i}, p_{1i})\right]\right\} &= \sum_{i=1}^n \log\left[1 + \chi^2(p_{0i}, p_{1i})\right] \\
&\leq n \cdot \log\left(1 + \epsilon_n^2 \cdot \frac{C_+}{\hbar}\right) \\
&\leq n\epsilon_n^2 \cdot \frac{C_+}{\hbar},
\end{aligned}$$

where we use the inequality $\log(1 + x) \leq x$ for all $x \geq 0$ to obtain the last inequality. Thus, $\chi^2(\mathrm{P}_0, \mathrm{P}_1) = \prod_{i=1}^n \left[1 + \chi^2(p_{0i}, p_{1i})\right] - 1 \leq \alpha$ as long as

$$n\epsilon_n^2 \cdot \frac{C_+}{\hbar} \leq \log(1 + \alpha)$$

for some $\alpha < 2$, which holds when $\epsilon_n^2 \leq \frac{\log(1+\alpha)\cdot\hbar}{nC_+}$. Hence, we set $\epsilon_n = \sqrt{\frac{\log(1+\alpha)\cdot\hbar}{2nC_+}}$ so that

$$\begin{aligned}\bar{m}_\hbar(\mathrm{P}_1;t_0) - \bar{m}_\hbar(\mathrm{P}_0;t_0) &\geq \frac{\epsilon_n \cdot C_-}{\hbar} \\ &= \frac{C_-}{\hbar}\sqrt{\frac{\log(1+\alpha)\cdot\hbar}{2nC_+}} \\ &= C_-\sqrt{\frac{\log(1+\alpha)}{2n\hbar C_+}}.\end{aligned}$$

Plugging these results back to the minimax lower bound, we obtain that

$$\inf_{\widehat{m}_\hbar(t_0)} \sup_{\mathrm{P}\in\mathcal{P}} \mathbb{E}_{\mathrm{P}}\left[\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0)\right]^2 \geq \frac{C_-^2\log(1+\alpha)}{2C_+n\hbar}\left[\frac{1-\sqrt{\alpha(1-\alpha/4)}}{2}\right].$$

The result thus follows. □

E.8 Proofs of Theorem 6.13 and Corollary 6.14

Proof of Theorem 6.13. Under Assumption 6.2, we know from (6.29) that

$$\begin{aligned}&\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0) \\ &= \mathbb{P}_n\left\{\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y - \int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right] + \int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right\} \\ &\quad - \mathbb{E}\left[\int_{\mathcal{T}}\mu(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right] \\ &= \underbrace{(\mathbb{P}_n - \mathrm{P})\,\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})}_{\textbf{Term I}} + \underbrace{\mathrm{P}\left[\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}}) - \Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}})\right]}_{\textbf{Term II}} \\ &\quad + \underbrace{(\mathbb{P}_n - \mathrm{P})\left[\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}}) - \Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})\right]}_{\textbf{Term III}},\end{aligned}$$

where $\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}}) := \frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y - \int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right] + \int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1$.

For any fixed $\hbar > 0$, we know from Proposition 6.11 that $\mathrm{Var}\left[\sqrt{n}\,(\mathbb{P}_n - \mathrm{P})\,\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})\right] = V_{\mathrm{eff},\bar{m}_\hbar}(t_0)$ with μ replaced by $\widetilde{\mu}$ in the expression. As a result,

$$\textbf{Term I} = (\mathbb{P}_n - \mathrm{P})\,\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}}) = O_P\left(\sqrt{\frac{V_{\mathrm{eff},\bar{m}_\hbar}(t_0)}{n}}\right).$$

It remains to show that the remainder terms are of order $o_P\left(\sqrt{\frac{V_{\text{eff},\bar{m}_\hbar}(t_0)}{n}}\right)$ for any fixed $t_0 \in \mathcal{T}$ in Section E.8.1 and Section E.8.2. Finally, we shall also derive the asymptotic normality of $\widehat{m}_\hbar(t_0)$ (or equivalently, **Term I**) in Section E.8.3.

*E.8.1 Analysis of **Term II** for $\widehat{m}_\hbar(t_0)$*

Since $\mathrm{P}\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\mu,p_{T|\boldsymbol{X}}) = \bar{m}_\hbar(t_0)$, we can calculate that

Term II

$$
\begin{aligned}
&= \mathbb{E}\left\{\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y - \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]\right\} \\
&\quad + \mathbb{E}\left[\int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1 - \int_{\mathcal{T}} \mu(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right] \\
&= \mathbb{E}\left\{\frac{K\left(\frac{T-t_0}{\hbar}\right)}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du}\left[\mu(T,\boldsymbol{X}) - \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]\right\} \\
&\quad + \mathbb{E}\left\{\int_{\mathcal{T}} \left[\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X}) - \mu(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\right]dt_1\right\} \\
&= \mathbb{E}\left\{\left[\frac{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du} - 1\right]\left[\int_{\mathcal{T}} \mu(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1 - \int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]\right\}.
\end{aligned}
$$

Since $\sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2} = o_P(1)$ and $\sup_{t\in\mathcal{T}} ||\widehat{\mu}(t,\boldsymbol{X}) - \widetilde{\mu}(t,\boldsymbol{X})||_{L_2} = o_P(1)$, we know that $\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ satisfies Assumptions 6.15(a) and that $\widehat{\mu}(t,\boldsymbol{x})$ is bounded above with probability tending to 1 as $\hbar \to 0$ and $n\hbar \to \infty$. Hence, for any fixed $t_0 \in \mathcal{T}$ and $\boldsymbol{X} \in \mathcal{X}$,

$$
\begin{aligned}
&|q_{\hbar,t_0}(t_1|\boldsymbol{X}) - \widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})| \\
&= \left|\frac{K\left(\frac{t_1-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du} - \frac{K\left(\frac{t_1-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du}\right| \\
&= \left|\frac{K\left(\frac{t_1-t_0}{\hbar}\right)\left[p_{T|\boldsymbol{X}}(t_1|\boldsymbol{X}) - \widehat{p}_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\right]}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du} + \frac{K\left(\frac{t_1-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\left[\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\right]du}{\left[\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du\right]\left[\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du\right]}\right| \\
&\lesssim \sup_{t\in\mathcal{T}}\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|.
\end{aligned}
$$

Therefore, we can upper bound **Term II** as:

Term II

$$
\begin{aligned}
&= \mathbb{E}\Bigg\{ \frac{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\left[p_{T|\boldsymbol{X}}(u|\boldsymbol{X}) - \widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\right] du}{\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\, du} \\
&\qquad \times \left[\int_{\mathcal{T}} \mu(t_1,\boldsymbol{X})\left[q_{\hbar,t_0}(t_1|\boldsymbol{X}) - \widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\right] dt_1 + \int_{\mathcal{T}} \left[\mu(t_1,\boldsymbol{X}) - \widehat{\mu}(t_1,\boldsymbol{X})\right]\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1\right]\Bigg\} \\
&\overset{\text{(i)}}{\lesssim} \sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2} \left[\sup_{t\in\mathcal{T}} ||\widehat{\mu}(t,\boldsymbol{X}) - \mu(t,\boldsymbol{X})||_{L_2} + \sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2}\right] \\
&= o_P\left(\sqrt{\frac{V_{\text{eff},\bar{m}_\hbar}(t_0)}{n}}\right)
\end{aligned}
$$

by our assumption, in which (i) implicitly uses the Cauchy-Schwarz inequality.

E.8.2 Analysis of ***Term III*** *for* $\widehat{m}_\hbar(t_0)$

First, we know that

$$
\begin{aligned}
&\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}},\widehat{q}_{\hbar,t_0}) - \Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}},q_{\hbar,t_0}) \\
&= \underbrace{\left[\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})} - \frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right] Y}_{\textbf{Term IIIa}} + \underbrace{\int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1 - \int_{\mathcal{T}} \widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1}_{\textbf{Term IIIb}} \\
&\quad + \underbrace{\frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\int_{\mathcal{T}} \widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1 - \frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\int_{\mathcal{T}} \widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\, dt_1}_{\textbf{Term IIIc}}.
\end{aligned}
$$

We then compute the upper bounds for the second moments of these three terms as follows.

For **Term IIIa**, we have that

$$
\begin{aligned}
&\mathbb{E}\left[(\textbf{Term IIIa})^2\right] \\
&= \mathbb{E}\left\{\left[\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})\cdot p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}{q_{\hbar,t_0}(T|\boldsymbol{X})\cdot\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})} - 1\right]^2 \frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\cdot Y^2\right\} \\
&= \mathbb{E}\left\{\frac{\left(\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\left[p_{T|\boldsymbol{X}}(u|\boldsymbol{X}) - \widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\right] du\right)^2}{\left[\int_{\mathcal{T}} K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\, du\right]^2}\cdot\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\cdot\mathbb{E}(Y^2|T,\boldsymbol{X})\right\} \\
&\overset{\text{(ii)}}{\lesssim} \mathbb{E}\left\{\frac{\left[\int_{\mathbb{R}} K(v)\left[p_{T|\boldsymbol{X}}(t_0+v\hbar|\boldsymbol{X}) - \widehat{p}_{T|\boldsymbol{X}}(t_0+v\hbar|\boldsymbol{X})\right]\mathbb{1}_{\left\{t_0+v\hbar\in\mathcal{T}(\boldsymbol{X})\cap\widehat{\mathcal{T}}(\boldsymbol{X})\right\}} dv\right]^2}{\left[\int_{\mathbb{R}} K(v)\cdot\widehat{p}_{T|\boldsymbol{X}}(t_0+v\hbar|\boldsymbol{X})\cdot\mathbb{1}_{\left\{t_0+v\hbar\in\widehat{\mathcal{T}}(\boldsymbol{X})\right\}} dv\right]^2}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\times \mathbb{E}\left\{\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\cdot \mathrm{Var}(Y|T,\boldsymbol{X})\right\}\\
&\overset{\text{(iii)}}{\lesssim} \sup_{t\in\mathcal{T}}\left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X})-p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|^2_{L_2}\cdot V_{\text{eff},\bar{m}_\hbar}(t_0)\\
&= o_P\left(V_{\text{eff},\bar{m}_\hbar}(t_0)\right),
\end{aligned}
$$

where (ii) applies a change of variables while (iii) utilizes our boundedness condition on K and Assumption 6.15.

Similarly, for **Term IIIb**, we know that

$$
\begin{aligned}
&\mathbb{E}\left[(\textbf{Term IIIb})^2\right]\\
&=\mathbb{E}\left\{\left[\int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1-\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]^2\right\}\\
&=\mathbb{E}\left\{\left[\frac{\int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot K\left(\frac{t_1-t_0}{\hbar}\right)\cdot\widehat{p}_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\,dt_1}{\int_{\mathcal{T}}K\left(\frac{u-t_0}{\hbar}\right)\widehat{p}_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du}-\frac{\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot K\left(\frac{t_1-t_0}{\hbar}\right)\cdot p_{T|\boldsymbol{X}}(t_1|\boldsymbol{X})\,dt_1}{\int_{\mathcal{T}}K\left(\frac{u-t_0}{\hbar}\right)p_{T|\boldsymbol{X}}(u|\boldsymbol{X})\,du}\right]^2\right\}\\
&=\mathbb{E}\Bigg\{\Bigg[\frac{\int_{\mathbb{R}}\widehat{\mu}(t_0+u\hbar,\boldsymbol{X})\cdot K(u)\cdot\widehat{p}_{T|\boldsymbol{X}}(t_0+u\hbar|\boldsymbol{X})\cdot\mathbb{1}_{\{t_0+u\hbar\in\mathcal{T}(\boldsymbol{X})\}}du}{\int_{\mathbb{R}}K(v)\cdot\widehat{p}_{T|\boldsymbol{X}}(t_0+v\hbar|\boldsymbol{X})\cdot\mathbb{1}_{\{t_0+v\hbar\in\mathcal{T}(\boldsymbol{X})\}}dv}\\
&\qquad-\frac{\int_{\mathbb{R}}\widetilde{\mu}(t_0+u\hbar,\boldsymbol{X})\cdot K(u)\cdot p_{T|\boldsymbol{X}}(t_0+u\hbar|\boldsymbol{X})\cdot\mathbb{1}_{\{t_0+u\hbar\in\mathcal{T}(\boldsymbol{X})\}}du}{\int_{\mathbb{R}}K(v)\cdot p_{T|\boldsymbol{X}}(t_0+v\hbar|\boldsymbol{X})\cdot\mathbb{1}_{\{t_0+v\hbar\in\mathcal{T}(\boldsymbol{X})\}}dv}\Bigg]^2\Bigg\}\\
&\lesssim\left[\sup_{t\in\mathcal{T}}\left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X})-p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2}+\sup_{t\in\mathcal{T}}\left|\left|\widehat{\mu}(t,\boldsymbol{X})-\widetilde{\mu}(t,\boldsymbol{X})\right|\right|_{L_2}\right]^2\\
&=o_P(1).
\end{aligned}
$$

Finally, for **Term IIIc**, we have that

$$
\begin{aligned}
&\mathbb{E}\left[(\textbf{Term IIIc})^2\right]\\
&=\mathbb{E}\left\{\left[\frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1-\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]^2\right\}\\
&=\mathbb{E}\left\{\frac{q^2_{\hbar,t_0}(T|\boldsymbol{X})}{p^2_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[\int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1-\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]^2\right\}\\
&\quad+\mathbb{E}\left\{\left[\frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}-\frac{\widehat{q}_{\hbar,t_0}(T|\boldsymbol{X})}{\widehat{p}_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\right]^2\left[\int_{\mathcal{T}}\widehat{\mu}(t_1,\boldsymbol{X})\cdot\widehat{q}_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]^2\right\}
\end{aligned}
$$

$$\overset{\text{(iv)}}{\lesssim} \left[\sup_{t\in\mathcal{T}} \left|\left|\widehat{p}_{T|\boldsymbol{X}}(t|\boldsymbol{X}) - p_{T|\boldsymbol{X}}(t|\boldsymbol{X})\right|\right|_{L_2} + \sup_{t\in\mathcal{T}} ||\widehat{\mu}(t,\boldsymbol{X}) - \widetilde{\mu}(t,\boldsymbol{X})||_{L_2}\right]^2 V_{\text{eff},\bar{m}_\hbar}(t_0)$$
$$= o_P\left(V_{\text{eff},\bar{m}_\hbar}(t_0)\right),$$

where (iv) follows from our previous calculations for **Term IIIa** and **Term IIIb**.

Therefore, by Markov's inequality, we conclude that

$$\begin{aligned}\textbf{Term III} &= (\mathbb{P}_n - \mathrm{P})\left[\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widehat{\mu},\widehat{p}_{T|\boldsymbol{X}}) - \Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})\right] \\ &= O_P\left(\sqrt{\frac{\mathbb{E}\left[(\textbf{Term III})^2\right]}{n}}\right) = o_P\left(\sqrt{\frac{V_{\text{eff},\bar{m}_\hbar}(t_0)}{n}}\right).\end{aligned}$$

E.8.3 Asymptotic Normality of $\widehat{m}_\hbar(t_0)$

To establish the asymptotic normality of $\widehat{m}_\hbar(t_0)$, it suffices to verify the Lyapunov or Lindeberg condition as in Theorem 4 of Schindl et al. (2024). Specifically, we have already shown at the beginning of the proof and subsequent subsections that

$$\begin{aligned}&\sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\left[\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0)\right] \\ &= \sqrt{n}\,(\mathbb{P}_n - \mathrm{P})\,\frac{\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}},q_{\hbar,t_0})}{\sqrt{V_{\text{eff},\bar{m}_\hbar}(t_0)}} + o_P\left(\sqrt{\frac{1}{n\cdot V_{\text{eff},\bar{m}_\hbar}(t_0)}}\right) \\ &= \sqrt{n}\,(\mathbb{P}_n - \mathrm{P})\,\frac{\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}},q_{\hbar,t_0})}{\sqrt{V_{\text{eff},\bar{m}_\hbar}(t_0)}} + o_P(1)\end{aligned}$$

with $\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}}) = \frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y - \int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right] + \int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1$. It suffices to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>\epsilon\, s_n\right\}}\right] = 0,$$

where $s_n^2 = \sum_{i=1}^n \mathrm{Var}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right] = n\cdot V_{\text{eff},\bar{m}_\hbar}(t_0)$.

Notice that

$$\mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>\epsilon\, s_n\right\}}\right]$$

$$\leq \mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>C_1\epsilon\sqrt{n\cdot V_{\mathrm{eff},\bar{m}_\hbar}(t_0)}\right\}}\right],$$

where the constant $C_1 > 0$ is specified in the lower bound for $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)$ in Lemma E.2. Additionally, we can bound $\left|\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}},q_{\hbar,t_0})\right|$ as:

$$\begin{aligned}
&\left|\Phi_{\hbar,t_0}(Y,T,\boldsymbol{X};\widetilde{\mu},p_{T|\boldsymbol{X}})\right| \\
&\leq \left|\frac{q_{\hbar,t_0}(T|\boldsymbol{X})}{p_{T|\boldsymbol{X}}(T|\boldsymbol{X})}\left[Y-\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right]\right| + \left|\int_{\mathcal{T}}\widetilde{\mu}(t_1,\boldsymbol{X})\cdot q_{\hbar,t_0}(t_1|\boldsymbol{X})\,dt_1\right| \\
&\leq C_2\cdot V_{\mathrm{eff},\bar{m}_\hbar}(t_0)+C_3,
\end{aligned}$$

with probability 1 by Assumption 6.15 for some absolute constants $C_2, C_3 > 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>\epsilon\, s_n\right\}}\right] \\
&\leq \mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{C_2\cdot V_{\mathrm{eff},\bar{m}_\hbar}(t_0)+C_3>C_1\epsilon\sqrt{\frac{n}{\hbar}}\right\}}\right]
\end{aligned}$$

and

$$\mathbb{1}_{\left\{C_2\cdot V_{\mathrm{eff},\bar{m}_\hbar}(t_0)+C_3>C_1\epsilon\sqrt{n\cdot V_{\mathrm{eff},\bar{m}_\hbar}(t_0)}\right\}}\to 0$$

as $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)\to\infty$ when $n\to\infty, \hbar\to 0$, and $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)=o(n)$. Finally, we also know that

$$\begin{aligned}
&\frac{1}{s_n^2}\sum_{i=1}^n \Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>\epsilon\, s_n\right\}} \\
&\leq \frac{1}{s_n^2}\sum_{i=1}^n \Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2
\end{aligned}$$

and $\mathbb{E}\left[\frac{1}{s_n^2}\sum_{i=1}^n \Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\right] = 1 < \infty$. Thus, by the dominated convergence theorem, we establish that

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\left[\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})^2\,\mathbb{1}_{\left\{\left|\Phi_{\hbar,t_0}(Y_i,T_i,\boldsymbol{X}_i;\widetilde{\mu},p_{T|\boldsymbol{X}})\right|>\epsilon\, s_n\right\}}\right]=0$$

as $n\to\infty$ and $\hbar\to 0$ with $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)=o(n)$.

Hence, the Lindeberg condition holds, and we have that

$$\sqrt{\frac{n}{V_{\mathrm{eff},\bar{m}_\hbar}(t_0)}}\left[\widehat{m}_\hbar(t_0)-\bar{m}_\hbar(t_0)\right]\xrightarrow{d}\mathcal{N}(0,1) \tag{E.13}$$

as $n\to\infty$ and $\hbar\to 0$ with $V_{\mathrm{eff},\bar{m}_\hbar}(t_0)=o(n)$. The proof of Theorem 6.13 is thus completed.

□

Proof of Corollary 6.14. Under the assumptions in Theorem 6.10 and Theorem 6.13, we know from Theorem 6.10 that the bias term of $\bar{m}_\hbar(t_0)$ as $\hbar \to 0$ consists of two parts: $\Delta_{\text{int}}(\hbar)$ and $\Delta_{\text{proj}}(\hbar)$. We also know that $\sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\,[\widehat{m}_\hbar(t_0) - \bar{m}_\hbar(t_0)] \asymp \sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\,[\widehat{m}_\hbar(t_0) - \bar{m}_0(t_0)]$ as long as

$$\sqrt{\frac{n}{V_{\text{eff},\bar{m}_\hbar}(t_0)}}\,[\Delta_{\text{int}}(\hbar) + \Delta_{\text{proj}}(\hbar)] \to 0.$$

The result thus follows from Theorem 6.13 and our conditions in the corollary statement. □