Statistical Divergences for Learning and Inference: Limit Laws and Non-Asymptotic Bounds

Lang Liu

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Zaid Harchaoui, Chair

Soumik Pal, Chair

Thomas S. Richardson

Kevin Jamieson

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical Divergences for Learning and Inference:
Limit Laws and Non-Asymptotic Bounds

Lang Liu

Co-Chairs of the Supervisory Committee:
Zaid Harchaoui
Department of Statistics

Soumik Pal
Department of Mathematics

Statistical divergences have been widely used in statistics and artificial intelligence to measure the dissimilarity between probability distributions. The applications range from generative modeling to statistical inference. Early works in statistics have focused on discrete and low-dimensional probability distributions. We choose to tackle problems emerging in modern applications of statistics and artificial intelligence in which the sample space is either discrete with a large alphabet (e.g., natural language processing) or continuous of high dimension (e.g., computer vision). This dissertation revisits statistical divergences in modern applications and addresses challenges arising from the complex nature of the data.

Chapter 2 studies the minimum Kullback-Leibler divergence estimation which is equivalent to the widely used maximum likelihood estimation. While the classical asymptotic theory is well established in a rather general setting, high-dimensional problems reveal several of its limitations. We develop finite-sample bounds characterizing the asymptotic behavior in a non-asymptotic fashion, allowing the dimension to grow with the sample size. Unlike previous work that relies heavily on the strong convexity of the objective function, we only assume the Hessian is lower bounded at optimum and allow it to gradually become degenerate. This is enabled by the notion of self-concordance originating from convex optimization.

Chapter 3 investigates the framework of divergence frontiers, a notion of trade-off curves built upon statistical divergences, for comparing generative models. These trade-off curves are analogous to operating characteristic curves in statistical decision theory. Due to the complex and high-dimensional nature of the input space, an effective approach used by practitioners to estimate divergence frontiers involves a quantization step followed by an estimation step. We establish non-asymptotic bounds on the sample complexity of this estimator. We also show how smoothed distribution estimators such as Good-Turing or Krichevsky-Trofimov can overcome the missing mass problem and lead to faster rate of convergence.

Chapters 4 and 5 explore the Schrödinger bridge problem—an information projection problem which projects a reference measure onto a linear subspace of probability distributions in terms of the Kullback-Leibler divergence. This problem is equivalent to the entropy-regularized optimal transport problem that recently attracted a huge attention from the statistics and machine learning communities. We develop limit laws and non-asymptotic bounds for its empirical estimators. Unlike the unregularized optimal transport, our results enjoy a parametric rate of convergence that does not suffer from the curse of dimensionality. We also propose statistical tests for testing homogeneity and independence based on the Schrödinger bridge problem.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

I want to express my sincere gratitude to my academic advisors, Zaid Harchaoui and Soumik Pal. I am grateful to Zaid for his continuing support and guidance since the very beginning of my PhD. His contagious enthusiasm for research, consistent pursuit of excellence, and steadfast belief in me have made this journey both fruitful and enjoyable. I was fortunate to start working with Soumik at the low point of my PhD. His enlightening insight into technical problems and remarkable persistence in meeting challenges have inspired me and reignited my passion for theory.

I thank the other members of my committee, Thomas Richardson, Kevin Jamieson, and Hannaneh Hajishirzi, for your concrete feedback and career advice. I am grateful to the other faculty of the Department of Statistics. In particular, I want to thank Emily Fox for her mentorship even before my PhD, Jon Wellner for being so generous with his time having inspiring conversations with me, Vincent Roulet and Carlos Cinelli for fruitful discussions, Fang Han for academic and career advice, Ema Perkovic for detailed feedback on my presentations, Sham Kakade for his recognition of my first presentation at ML-OPT, and Abel Rodriguez and Marina Meila for keeping the department and the PhD program operating smoothly. I also want to thank Joseph Salmon for his mentorship during his visit at UW.

I am grateful to the Institute for Foundations of Data Science, the Kantorovich Initiative, and the National AI Institute for Foundations of Machine Learning for creating a collaborative research environment. In particular, I would like to thank Maryam Fazel, Sewoong Oh, Ludwig Schmidt, Bamdad Hosseini, Dmitriy Drusvyatskiy, Lalit Jain, Yanqin Fan.

I want to thank Fabian Moerchen and Brandyn Kusenda for their mentorship during my internship at Amazon Music and thank Sen Zhao and Mahdi Milani Fard for their mentorship during my internship at Google Research. I am grateful to my undergraduate advisors Dong Li and Xuegong Zhang for guiding me in research. I would also like to thank my math teacher MingLiang Yang in high school for cultivating my interest in math.

I would like to thank my fellow students for sharing the academic life with me: Richard Guo, Gang Cheng, Kristof Glauninger, Wenyu Chen, Peter Gao, Jessica Perry, Sara Laplante, Christopher Aicher, Alex Tank, Samson Koelle, Alec Greaves-Tunnell, Corinne Jones, Kunhui Zhang, Xinyuan Dong, Yang Liu, Zhen Miao, Zhenman Yuan, Krishna Pillutla, John Thickstun, Meyer Scetbon, Zhaoqi Li, Hanyu Zhang, Zeyu Wei, Yuhao Wan, Nick Irons, Zhanlin Liu, Han Zhang, Yizhong Wang, Jillian Fisher, Ronak Mehta, Medha Agarwal, and Haoyang Wu. Special thanks to Krishna with whom I benefited a lot from the interactions. The working experiences with Jillian Fisher, Ronak Mehta, and Medha Agarwal have been fruitful and delightful. I am also grateful to Nick Foti for his mentorship and to Lijun Ding, Sean Welleck, and Tianxiao Shen for rewarding discussions.

I am grateful to the staff of the Department of Statistics: Ellen Reynolds, Kristine Chan, Tracy Pham, Asa Sourdiffe, and Vickie Graybeal.

I thank all my co-authors (in chronological order): Zaid Harchaoui, Joseph Salmon, Soumik Pal, Mahdi Milani Fard, Sen Zhao, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Carlos Cinelli.

I have been fortunate to be companied by wonderful friends during this journey: Xuanxuan Wu, Yiping Hu, Huangyu Xiao, Bowen Xiao, Xiaonan Sun, Yilun Xing, Yinsheng Wang and Yilin Song. I thank Xiaonan for her support and understanding during the hard times in my last year.

Lastly and most importantly, I thank my parents for their continual support. I can never go this far without you.

DEDICATION

To my parents

Chapter 1

INTRODUCTION

Statistical divergences, which quantify the discrepancy between probability distributions, are ubiquitous in statistics and machine learning. In statistical learning, they have been used to estimate, either parametrically or non-parametrically, the density from which the data are generated via minimizing the discrepancy between the data distribution and the model distribution. This includes the classical maximum likelihood estimation, the minimum distance estimation, and the modern adversarial generative modeling. In statistical inference, statistical divergences provide a principled way to design test statistics to determine whether two samples are coming from the same population known as the homogeneity (or two-sample) testing problem. In the same spirit, they can also be applied to the independence testing problem by reformulating the independence of two random elements into the equivalence of the joint distribution and the product of marginal distributions.

Early works in statistics focus on discrete and one-dimensional probability distributions where statistical divergences are defined via the probability mass function (Pearson, 1900; Bhattacharyya, 1946) and the cumulative distribution function, respectively (Cramér, 1928; von Mises, 1931; Kolmogorov, 1933; Smirnov, 1939; Anderson and Darling, 1954). They are not applicable in modern statistics and machine learning applications where the sample space is either discrete with a large alphabet or continuous of high dimension. For instance, when testing the independence of two documents (Gretton et al., 2007b), the observations are natural languages which live in a discrete space of nearly infinite size. When learning a generative model on images (Goodfellow et al., 2014), the observations are images consisting of hundreds of thousands of pixel values. This raises challenges both in their statistical analysis and computation.

This dissertation revisits statistical divergences in modern applications and addresses challenges arising from the complex nature of the data. In particular, we use the well-known Kullback-Leibler (KL) divergence to estimate model parameters, build trade-off curves for generative models, and study the Schrödinger problem. To facilitate the exposition of the results, we introduce in the following several concepts involving statistical divergences. First, we review in Section 1.1 the KL divergence and its extension which we call information divergences. We then in Section 1.2 explain how the KL divergence can be applied to parameter estimation which coincides with the celebrated maximum likelihood estimation. Next, in Section 1.3, we review two types of errors and trade-off curves in statistical decision theory which serves as the basis of modern trade-off curves for generative models. Finally, we consider in Section 1.4 a structured information projection problem with marginal constraints which leads to the so-called Schrödinger problem. It can be reformulated as an entropy-regularized optimal transport problem which is closely related to the optimal transport distance. Moreover, it induces the so-called Sinkhorn divergence. Both the optimal transport distance and the Sinkhorn divergence are statistical divergences adapted to the metric of the sample space. We conclude with a summary of the main contributions and an outline of the rest of the thesis.

1.1 Information Divergences

In their seminal work, Kullback and Leibler (1951) introduced the KL divergence (or the relative entropy) while studying generalizations of the Shannon entropy (Shannon, 1948).

Definition 1.1 (Kullback-Leibler divergence)**.** *Let $\mathcal{X}$ be some measurable space and $\mathcal{M}_1(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. For any $P, Q \in \mathcal{M}_1(\mathcal{X})$, the KL divergence between P and Q is defined as*

$$\mathrm{KL}(P\|Q) := \begin{cases} \int \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P & \text{if } P \ll Q \\ \infty & \text{otherwise.} \end{cases}$$

An important feature of the KL divergence is that it is indeed a divergence, i.e., a metric without the triangle inequality, which behaves similarly as the squared Euclidean distance.

Property 1.1 (Lemma 3.1 in Kullback and Leibler (1951))**.** *For any $P, Q \in \mathcal{M}_1(\mathcal{X})$, we have $\mathrm{KL}(P\|Q) \geq 0$ with equality iff $P = Q$ Q-a.s.*

Remark 1.2. *Note that the KL divergence is asymmetric. In fact, as noted by Kullback and Leibler (1951), its symmetrized version, i.e., $\mathrm{KL}(P\|Q) + \mathrm{KL}(Q\|P)$, has appeared earlier in Jeffreys (1946).*

Beyond information theory, the KL divergence has found its wide applications in statistics and probability. For instance, it can be used to estimate parameters of a statistical model which corresponds to the famous maximum likelihood estimation as demonstrated in Section 1.2. Moreover, in large deviation theory, it characterizes the rate at which the rare event probability converges to zero (see, e.g., Dembo and Zeitouni, 2009). In particular, according to Sanov's theorem, given a subset of distributions $\mathcal{P} \subset \mathcal{M}_1(\mathcal{X})$, the log of the probability that the empirical measure of an i.i.d. sample from $Q \in \mathcal{M}_1(\mathcal{X})$ belongs to $\mathcal{P}$ is asymptotically close to

$$\min_{P \in \mathcal{P}} \mathrm{KL}(P\|Q) \tag{1.1}$$

with a factor of $-n$, where n is the sample size. The above variational problem of the KL divergence is later studied by Csiszár (1975) which is referred to as the information projection (or I-projection) of Q onto $\mathcal{P}$.

A case of special interest to us is when $\mathcal{P} := \cap_{k=1}^K \mathcal{P}_k$ for some linear subsets $\{\mathcal{P}_k\}_{k=1}^K$ of $\mathcal{M}_1(\mathcal{X})$. Under these linear constraints, the solution (if exists) admits a special structure involving these constraints (Csiszár, 1975, Theorem 3.1). Furthermore, it can be obtain by the iterative proportional fitting procedure (IPFP)—projecting Q onto each linear subset iteratively—which can be dated back to Deming and Stephan (1940): for $t \geq 1$,

$$Q_t = \operatorname*{arg\,min}_{P \in \mathcal{P}_k} \mathrm{KL}(P\|Q_{t-1}) \quad \text{if } t \bmod K = k,$$

where $Q_0 := Q$. This variational problem and the IPFP form the backbone of the Schrödinger problem as we will see in Section 1.4.

Finally, we note that the KL divergence was generalized by Csiszár (1967) to a class of divergences known as the f-divergence.

Definition 1.2 (f-divergence)**.** *Let $f : (0, \infty) \to \mathbb{R}_+$ be a convex function with $f(1) = 0$. For any $P, Q \in \mathcal{M}_1(\mathcal{X})$, the f-divergence between P and Q is defined as*

$$D_f(P\|Q) := \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q$$

with the convention that $f(0) := \lim_{t\to 0_+} f(t)$ and $0f(p/0) = pf^*(0)$, where $f^*(0) \in [0, \infty]$ is the limit of $f^* : t \mapsto tf(1/t)$ at 0_+. We call f the generator to D_f and f^* the conjugate generator[1] *of f since $D_{f^*}(P\|Q) = D_f(Q\|P)$.*

Remark 1.3. *When P and Q are dominated by $\mu \in \mathcal{M}_1(\mathcal{X})$ with densities p and q, respectively, we have*

$$D_f(P\|Q) = \int_{q>0} f(p(x)/q(x))q(x)\mathrm{d}\mu(x) + f^*(0)P[q = 0]$$

with the agreement that the last term is taken to be zero if $P[q = 0] = 0$ no matter what value $f^*(0)$ takes (which could be infinity).

Example 1.4. *We illustrate a number of examples.*

(a) ***KL divergence****: $f(x) = x \log x - x + 1$.*

(b) ***Total variation****: $f(x) = |x - 1|/2$,*

$$\mathrm{TV}(P, Q) := \frac{1}{2}\int \left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right| \mathrm{d}Q.$$

(c) χ^2***-divergence****: $f(x) = (x - 1)^2$,*

$$\chi^2(P\|Q) := \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q.$$

[1] The conjugacy between f and f^* is unrelated to the usual Fenchel or Lagrange duality in convex analysis, but is related to the perspective transform (Rockafellar, 1970).

1.2 Parameter Estimation with the Kullback-Leibler Divergence

Let Z be a random element following some unknown distribution P. Consider a parametric family of distributions $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ which may or may not contain P. For simplicity of the presentation, we assume that P and P_θ are absolutely continuous w.r.t. some dominating measure with densities p and p_θ, respectively. We are interested in finding the parameter θ_* so that the model $P_{\theta_\star}$ best approximates the underlying distribution P. For this purpose, one strategy is to minimize the KL divergence, i.e.,

$$\theta_\star \in \arg\min_{\theta \in \Theta} \mathrm{KL}(P \| P_\theta). \tag{1.2}$$

Note that, when $P \ll P_\theta$,

$$\mathrm{KL}(P \| P_\theta) = \mathbb{E}\left[\log \frac{p(Z)}{p_\theta(Z)}\right] = \mathbb{E}[-\log p_\theta(Z)] + \mathbb{E}[\log p(Z)].$$

Consequently, minimizing $\mathrm{KL}(P \| P_\theta)$ over θ is equivalent to maximizing the expected log-likelihood $\mathbb{E}[\log p_\theta(Z)]$, which is known as the maximum likelihood method; see, e.g., (Casella and Berger, 2001).

Remark 1.5. *The KL minimization problem* (1.2) *is reminiscent of the information projection problem* (1.1)*. However, they are not exactly the same due to the asymmetry of the KL divergence. In fact, problems of the form* (1.2) *are often referred to as the* reverse information projection*.*

In many real applications, we do not have access to the underlying distribution P. Instead, we have an i.i.d. sample $\{Z_i\}_{i=1}^n$ from P. To learn the parameter $\theta_\star$ from the data, we maximize the data log-likelihood to obtain the *maximum likelihood estimator (MLE)*

$$\theta_n \in \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_\theta(Z_i).$$

Due to its flexibility and effectiveness, it has become a dominate approach in statistical inference and has wide applications in economics (Hendry and Nielsen, 2012), survey statistics (Chambers et al., 2012), and social sciences (Ward and Ahlquist, 2018).

To evaluate the performance of the estimator, we use the *excess KL divergence*, i.e.,

$$\mathcal{E}_{\mathrm{KL}}(\theta_n) := \mathrm{KL}(P \| P_{\theta_n}) - \inf_{\theta \in \Theta} \mathrm{KL}(P \| P_\theta) = \mathrm{KL}(P \| P_{\theta_n}) - \mathrm{KL}(P \| P_{\theta_\star}). \tag{1.3}$$

For instance, Rigollet (2012, Theorem 3.1) used the excess KL divergence in aggregation problems. The (symmetrized) KL divergence is also used in Gu (2013, Chapter 9) to quantify the convergence rate of penalized likelihood estimates.

The classical asymptotic theory of maximum likelihood estimation is well established in a rather general setting under the assumption that the parametric model is well-specified, i.e., the underlying data distribution belongs to the parametric family. We mention here, among many of them, the monographs by Ibragimov and Has' Minskii (1981); van der Vaart (2000); Geer et al. (2000). One of the main result is

$$\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star) \Rightarrow \mathcal{N}(0, I_d), \tag{1.4}$$

where $\Rightarrow$ represents weak convergence (or convergence in distribution), $H_\star := \nabla_\theta^2 \mathrm{KL}(P \| P_{\theta_\star})$ is the Hessian of the population objective, and d is the parameter dimension. Model misspecification has been considered in, e.g., Huber (1967), where the weak limit of $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star)$ becomes $\mathcal{N}_d(0, H_\star^{-1/2} G_\star H_\star^{-1/2})$. Here $G_\star := \mathbb{E}[\nabla_\theta \mathrm{KL}(P \| P_{\theta_\star}) \nabla_\theta \mathrm{KL}(P \| P_{\theta_\star})^\top]$ is the second moment of the gradient of the population objective.

The weak limit in (1.4) allows us to do statistical inference via the construction of a *confidence set* for $\theta_\star$ if we equip it with a consistent estimator H_n of $H_\star$. By Slutsky's lemma, we have $n(\theta_n - \theta_\star)^\top H_n(\theta_n - \theta_\star) \Rightarrow \chi_d^2$ and thus $\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \to 1 - \delta$, where

$$\mathcal{C}_n(\delta) := \left\{ \theta : n(\theta_n - \theta_\star)^\top H_n(\theta_n - \theta_\star) \le q_{\chi_d^2}(\delta) \right\}$$

and $q_{\chi_d^2}(\delta)$ is the upper δ-quantile of χ_d^2. That is, the probabiliy that $\theta_\star$ belongs to the confidence set $\mathcal{C}_n(\delta)$ is approximately $1 - \delta$ for large enough n. However, due to its asymptotic nature, it does not tell us how large n should be for this approximation to be accurate.

Another limitation of classical asymptotic theory is its asymptotic regime where $n \to \infty$ and the parameter dimension d is fixed. This is inapplicable in the modern context where the data are of rather high dimension involving a huge number of parameters. We establish non-asymptotic bounds for both the excess KL divergence and the estimator in Chapter 2. These bounds hold for any n and d such that $d = O(n)$ as $n \to \infty$. We also obtain confidence bounds for $\theta_\star$ in a non-asymptotic fashion and characterize the *critical sample size* that is enough to enter the asymptotic regime.

1.3 Two Types of Errors and Tradeoff Curves

In statistical hypothesis testing, the goal is to decide between the null hypothesis $\mathbf{H}_0$ and the alternative hypothesis $\mathbf{H}_1$ given a sample of observations (see, e.g., Casella and Berger, 2001). A typical procedure for such problems consists of the following procedure. First, we choose a quantity $T \in \mathbb{R}$ depending on $\mathbf{H}_0$ and $\mathbf{H}_1$ so that a large value of T favors $\mathbf{H}_1$. Second, we estimate T from the data to obtain a test statistic T_n. Finally, we select a threshold t_n and adopt the decision rule (or test) $\mathbb{1}\{T_n > t_n\}$, that is, we reject the null hypothesis if the test statistic exceeds the threshold.

The performance of a test is usually evaluated by two types of errors. A *type I error (or false positive)* occurs if the null hypothesis is wrongly rejected when it is true, and a *type II error (or false negative)* occurs if the null hypothesis fails to be rejected when the alternative hypothesis is true. The rates at which each of the two errors occur, i.e., the type I error rate $\mathbb{P}(T_n > t_n \mid \mathbf{H}_0)$ and the type II error rate $\mathbb{P}(T_n \le t_n \mid \mathbf{H}_1)$, constitute the *operating characteristics* of the test. Note that it is also common to use the *statistical power (or true positive rate)*, i.e., $1 - \mathbb{P}(T_n \le t_n \mid \mathbf{H}_1) = \mathbb{P}(T_n > t_n \mid \mathbf{H}_1)$ instead of the type II error rate. Often, one prescribes a *significance level* $\alpha \in (0, 1)$ and selects t_n to minimize the type II error rate under the constraint that the (asymptotic) type I error rate does not exceed α.

Example 1.6 (Neyman-Pearson test)**.** *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from some distribution μ. Consider two simple hypotheses $\mathbf{H}_0 : \mu = \mu_0$ and $\mathbf{H}_1 : \mu = \mu_1$, where μ_0 and μ_1 are two continuous distributions. In the Neyman-Pearson test, one chooses $T := \mathbb{E}_{X\sim\mu}[\log[\mu_1(X)/\mu_0(X)]]$, that is, the expected log likelihood ratio. A large value of*

T favors $\mathbf{H}_1$ since $T = \mathrm{KL}(\mu_1 \| \mu_0) \geq 0$ under $\mathbf{H}_1$ and $T = -\mathrm{KL}(\mu_0 \| \mu_1) \leq 0$ under $\mathbf{H}_0$. We can estimate T by $T_n := \frac{1}{n}\sum_{i=1}^n \log\left[\mu_1(X_i)/\mu_0(X_i)\right]$. The type I error rate and type II error rate are then given by, respectively,

$$\mathbb{P}_{\mu_0}\left(\frac{1}{n}\sum_{i=1}^n \log\left[\mu_1(X_i)/\mu_0(X_i)\right] > t_n\right) \quad \textit{and} \quad \mathbb{P}_{\mu_1}\left(\frac{1}{n}\sum_{i=1}^n \log\left[\mu_1(X_i)/\mu_0(X_i)\right] \leq t_n\right).$$

Instead of considering a fixed threshold t_n, the *receiver operating characteristic (ROC) curve* plots the statistical power versus the type I error rate for all possible choices of t (see, e.g., Pepe, 2000). By construction, it is an increasing function mapping from $[0, 1]$ to $[0, 1]$, where higher curves correspond to better tests. It displays the trade-offs between the type I error rate and the statistical power. It is particularly useful for comparing tests on different scales where comparisons based on the test statistics are not meaningful. The ROC curve was originally developed in electrical engineering and became increasingly popular in statistical machine learning (Cortes and Mohri, 2004; Clémençon and Vayatis, 2009; Flach, 2012).

In deep generative modeling, where the goal is to train generative models to generate artificial samples from the real data distribution, it is of great interest to develop quantitative evaluation tools to measure their statistical performance and diagnose where and why they fail (Salimans et al., 2016; Lopez-Paz and Oquab, 2017; Heusel et al., 2017; Sajjadi et al., 2018; Karras et al., 2019). There is a quality-diversity trade-off inherent to deep generative modeling. In particular, a good generative model must not only produce high-quality samples that are likely under the target distribution but also cover the target distribution with diverse samples.

To quantify this trade-off, similar notions of the type I error and type II error can be defined in the context of generative modeling (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Simon et al., 2019; Djolonga et al., 2020; Naeem et al., 2020; Pillutla et al., 2021). A type I error occurs if a generated sample point is unlikely under the real data distribution, i.e., it is unrealistic. A type II error occurs if a real data point is unlikely under the model distribution, i.e., it can rarely be generated by the model. Now a model with high type I error rate generates samples of low quality and a model with high type II error rate generates samples of low diversity. Choosing a reference measure R that interpolates between the data distribution P and model distribution Q, these two types of errors can be quantified by $\mathrm{KL}(Q\|R)$ and $\mathrm{KL}(P\|R)$, respectively. By sweeping through all adequate reference measures, we arrive at a curve, called the *divergence frontier* (Djolonga et al., 2020), that displays the quality-diversity trade-off of a generative model. This will be the main subject of Chapter 3.

1.4 The Schrödinger Problem and the Sinkhorn Divergence

We introduce the Schrödinger problem in modern terms. Given $\varepsilon \in \mathbb{R}_+$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, we assume that the following Markov transition density is well-defined:

$$p_\varepsilon(x, y) := \frac{1}{Z_\varepsilon(x)} \exp\left[-\frac{1}{\varepsilon} c(x, y)\right],$$

where $Z_\varepsilon(x)$ is the normalizing constant. For instance, when c is the quadratic cost, this is the transition density of the Brownian motion with diffusion ε considered in Schrödinger's lazy gas experiment (Schrödinger, 1932). Suppose that (W_0, W_1) is a pair of random vectors distributed according to this Markov transition kernel. Let P and Q be two probability measures on $\mathbb{R}^d$. Informally, the *(static) Schrödinger bridge* connecting P and Q at temperature ε is the joint distribution of (W_0, W_1) conditioned to have $W_0 \sim P$ and $W_1 \sim Q$. In continuum, when P and Q have densities w.r.t. the Lebesgue measure, it can be made precise as the solution to the following information projection problem (Föllmer, 1988; Léonard, 2012, 2014):

$$\min_{\nu \in \Pi(P,Q)} \mathrm{KL}(\nu \| R_\varepsilon), \tag{1.5}$$

where $\Pi(P, Q)$ is the set of couplings (joint distributions) with marginals P and Q and[2] $R_\varepsilon(x, y) := P(x)p_\varepsilon(x, y)$. We refer to (1.5) as the *Schrödinger problem.*

With this formulation, the Schrödinger bridge has a geometry interpretation—we first construct the reference measure R_ε by starting from the initial configuration P and jumping

[2] We follow the standard abuse of keeping the same notation for an absolutely continuous measure and its density.

according to the Markov transition kernel p_ε, and we then project R_ε onto the space of couplings $\Pi(P,Q)$ since R_ε may not belong to $\Pi(P,Q)$. Note that $\Pi(P,Q)$ is the intersection of two linear marginal constraints. This will be important in characterizing and computing the Schrödinger bridge in Chapter 4.

The Schrödinger problem is closely related to the optimal transport (OT) problem (see, e.g., Villani, 2021):

$$\min_{\nu\in\Pi(P,Q)} \int c(x,y)\mathrm{d}\nu(x,y), \tag{1.6}$$

where the goal is to find the optimal coupling (or transport plan) which minimizes the transport cost. When the cost is chosen to be $c(x,y) := \|x-y\|^p$ for $p \geq 1$, the OT cost induces a metric on $\mathcal{M}_1(\mathbb{R}^d)$ which is known as the Wasserstein-p distance (Santambrogio, 2015, Chapter 5). To see the connection, note that, for any $\nu \in \Pi(P,Q)$,

$$\mathrm{KL}(\nu\|R_\varepsilon) = \frac{1}{\varepsilon}\int c(x,y)\mathrm{d}\nu(x,y) + H(\nu) + \int \log Z_\varepsilon(x)\mathrm{d}P(x) - H(P),$$

where $H(\nu)$ is the differential entropy of ν defined as $H(\nu) := \int \log \nu(x,y)\mathrm{d}\nu(x,y)$ if ν is a density and infinity otherwise. As a result, the Schrödinger problem (1.5) is equivalent to

$$S_\varepsilon(P,Q) := \min_{\nu\in\Pi(P,Q)} \left[\int c(x,y)\mathrm{d}\nu(x,y) + \varepsilon H(\nu)\right] \tag{1.7}$$

which is known as the entropy-regularized optimal transport (EOT) problem. Whereas the OT problem usually admits a degenerate solution given by a transport map with zero-measure support (Santambrogio, 2015, Theorem 1.17), the entropy term in the EOT problem prevents such solutions from existing. Moreover, as $\varepsilon \to 0$, the minimum of the Schrödinger problem converges to the one of the OT problem and the minimizer (if exists) as well (Léonard, 2012, Theorem 3.3). In other words, the Schrödinger problem can be viewed as a *smooth* approximation to the OT problem which quantifies how close two distributions are. As opposed to the OT problem, the Schrödinger problem can not only be solved efficiently but also be estimated without suffering from the curse of dimensionality as shown in Chapters 4 and 5.

While S_ε is close to the OT distance when $\varepsilon \to 0$, it is not a statistical divergence for a fixed ε since $S_\varepsilon(P, Q)$ is not necessarily 0 when $P = Q$. A remedy to this issue is to consider the centered version

$$\bar{S}_\varepsilon(P, Q) := S_\varepsilon(P, Q) - \frac{1}{2} S_\varepsilon(P, P) - \frac{1}{2} S_\varepsilon(Q, Q).$$

This defines a semi-metric on the space of probability measures which is known as the *Sinkhorn divergence* (Feydy et al., 2019) and has been applied to two-sample testing (Ramdas et al., 2017) and generative modeling (Genevay et al., 2018). We will use it to measure independence in Chapter 5.

1.5 Contributions and Outline

In the remainder of this dissertation, we study the aforementioned problems in each of the four chapters. We conclude with a discussion on future research directions.

Non-asymptotic analysis of the maximum likelihood estimator. In Chapter 2 we present an excess KL divergence bound and a confidence bound for the maximum likelihood estimator (or the minimum KL divergence estimator). Our analysis allows the dimension to grow with the sample size and the model to be mis-specified. The complexity of the parameter space is measured by the *effective dimension* $d_\star$, i.e., the trace of the asymptotic covariance $H_\star^{-1/2} G_\star H_\star^{-1/2}$ of $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star)$, which can be much smaller than the parameter dimension in some regimes. Moreover, we obtain a novel non-asymptotic confidence bound for the estimator whose shape is adapted to the optimization landscape induced by the loss function. Along the way, we demonstrate how the effective dimension can be estimated from data and characterize its estimation accuracy. Compared to classical asymptotic theory, our results recover the limiting behavior in a non-asymptotic fashion whenever the sample size n grows linearly in the "dimension" $d + d_\star$. In other words, they provide a characterization of the *critical sample size* sufficient to enter the asymptotic regime. Compared to previous works in the non-asymptotic theory, our approach avoids strong global assumptions on the data likelihood by restricting ourselves to *generalized self-concordant* losses, a tool borrowed from convex analysis, while being more general than the generalized linear models.

This chapter is joint work with Zaid Harchaoui. A preliminary version was presented at the NeurIPS 2022 Workshop on Score-Based Methods (Liu and Harchaoui, 2022) and a longer version is under review. The part on detecting changes in model parameters is joint work with Joseph Salmon and Zaid Harchaoui which was published at ICASSP 2021 (Liu et al., 2021b). In a collaboration with Carlos Cinelli and Zaid Harchaoui that was published at COLT 2022 (Liu et al., 2022a), the results are extended to double machine learning/orthogonal statistical learning for semi-parametric models. In a collaboration with Jillian Fisher, Krishna Pillutla, Yejin Choi, and Zaid Harchaoui that is under review (Fisher et al., 2022), the techniques are applied to analyze influence functions.

Non-asymptotic analysis of the divergence frontiers. In Chapter 3 we focus on the notion of divergence frontiers which was recently proposed as an evaluation framework for generative models to quantify the quality-diversity trade-offs inherent to deep generative modeling. Due to the complex and high-dimensional nature of the input space (e.g., images or text), we conform to the recipe used by practitioners to estimate divergence frontiers: (a) jointly quantize the data and model distributions into disjoint groups, (b) estimate the quantized distributions from samples, and (c) compute the divergence frontier between the estimated distributions. We establish non-asymptotic bounds for the sample complexity of this estimator in the large-alphabet regime, i.e., the quantization level is large compared to the sample size. Along the way, we introduce frontier integrals which provide summary statistics of divergence frontiers. The frontier integral itself turns out to be a symmetric f-divergence. We discuss the choice of quantization level by balancing the two types of approximation errors arise from the computation. We also show how smoothed distribution estimators such as Good-Turing or Krichevsky-Trofimov can overcome the missing mass problem and lead to faster rates of convergence.

This chapter is joint work with Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. It was published at NeurIPS 2021 (Liu et al., 2021a) and a journal-length version is in preparation.

Asymptotics of discrete Schrödinger bridges. In Chapter 4 we turn our attention to the Schrödinger problem. In light of Schrödinger's lazy gas experiment, we propose the discrete Schrödinger bridge as an estimator of the Schrödinger bridge in continuum. We establish its asymptotic consistency as the sample size goes to infinity while the dimension kept fixed. We prove limiting Gaussian fluctuations for this convergence in the form of central limit theorems for integrated test functions. This suggests a parametric rate of convergence $O(n^{-1/2})$ that does not suffer from the curse of dimensionality. A result of independent interest is the limit law of a two-sample U-statistic of infinite order constructed from the permanent of the Gram matrix of a bivariate function. The proofs are based on a novel chaos decomposition of the discrete Schrödinger bridge consisting of functions of the empirical distributions as Taylor approximations in the space of measures. This is achieved by extending the Hoeffding decomposition from the classical theory of U-statistics. The kernels in the first order chaos are given by an alternating conditional expectation procedure induced by Markov operators associated with the Schrödinger bridge which is reminiscent of the IPFP. Finally, we design an efficient algorithm to compute the discrete Schrödinger bridge and apply it to homogeneity testing.

This chapter is joint work with Soumik Pal and Zaid Harchaoui. A preliminary version was presented at the NeurIPS 2021 Workshop on Optimal Transport and Machine Learning, and a journal-length version is under revision at Bernoulli (Harchaoui et al., 2022).

Entropy-regularized optimal transport independence criterion. In Chapter 5 we introduce an independence criterion based on the entropy-regularized optimal transport which uses the Sinkhorn divergence to quantify the discrepancy between the joint distribution and the product of marginals. The plug-in estimator can be used as a statistic to test for independence. We establish non-asymptotic bounds for our test statistic and study its statistical behavior under both the null hypothesis and the alternative hypothesis. The rate of convergence is $O(\sigma^{3d} n^{-1/2})$ where σ is some problem-specific constant. In other words, it enjoys a parametric rate of convergence where the dimension only appears in a problem-specified constant. A result of independent interest is a metric entropy bound for degenerate two-sample U-processes. The theoretical results involve tools from U-process theory and optimal transport theory. We propose an efficient algorithm based on random feature approximations and symbolic matrices to compute the test statistic in large-scale problems, which admits a quadratic time complexity and a linear space complexity. Finally, we show how to differentiate through the proposed independence criterion in a differentiable programming framework.

This chapter is joint work with Soumik Pal and Zaid Harchaoui. A major part of it was published at AISTATS 2022 (Liu et al., 2022c) and the implementational aspects for large-scale problems were presented at the International Conference on Computational Statistics in Summer 2022.

Other explorations. This dissertation does not include (1) the work on meta-learning with heterogeneous covariate spaces (Liu et al., 2022b) where the dissertation author is the primary author and conducted during an internship, (2) a collaborative work on spectral risk measures (Mehta et al., 2022).

Chapter 2

INFORMATION DIVERGENCES FOR ESTIMATION AND INFERENCE

2.1 Introduction

The problem of statistical inference on learned parameters is regaining the importance it deserves as machine learning and data science are increasingly impacting humanity and society through a large range of successful applications from transportation to healthcare. The classical asymptotic theory of minimum KL divergence estimation (or maximum likelihood estimation) is well established in a rather general setting under the assumption that the parametric model is well-specified, i.e., the underlying data distribution belongs to the parametric family. From this theory a Wald-type confidence set, which relies on the weighted difference between the estimator and the target parameter, can be constructed to quantify the statistical uncertainty of the estimator. The main tool is the local asymptotic normality (LAN) condition introduced by Le Cam (1960) (see also Geyer, 2013). We mention here, among many of them, the monographs Ibragimov and Has' Minskii (1981); van der Vaart (2000); Geer et al. (2000).

In many real problems, the parametric model is usually an approximation to the data distribution, so it is too restrictive to assume that the model is well-specified. To relax this restriction, model mis-specification has been considered by Huber (1967); see also Wakefield (2013); Dawid et al. (2016). Another limitation of classical asymptotic theory is its asymptotic regime where $n \to \infty$ and the parameter dimension d is fixed. This is inapplicable in the modern context where the data are of rather high dimension involving a huge number of parameters.

The non-asymptotic viewpoint has been fruitful to address higher dimensional problems—

the results are developed for fixed n so that it also captures the asymptotic regime where d grows with n. Early works in this line of research focus on specific models such as Gaussian models (Beran, 1996; Beran and Dumbgen, 1998; Laurent and Massart, 2000; Baraud, 2004), ridge regression (Hsu et al., 2012), logistic regression (Bach, 2010), and robust M-estimation (Zhou et al., 2018; Chen and Zhou, 2020); see Bach (2021) for a survey. Spokoiny (2012) addressed the non-asymptotic regime in full generality in a spirit similar to the classical LAN theory. The approach of Spokoiny (2012) relies on heavy empirical process machinery and requires strong global assumptions on the deviation of the empirical risk process. More recently, Ostrovskii and Bach (2021) focused on risk bounds, specializing their discussion to linear models with (pseudo) self-concordant losses and obtained a more transparent analysis under neater assumptions.

A critical tool arose from this line of research is the so-called *Dikin ellipsoid*, a geometric object identified in the theory of convex optimization (Nesterov and Nemirovskii, 1994; Ben-Tal and Nemirovski, 2001; Boyd and Vandenberghe, 2004; Tunçel and Nemirovski, 2010; Bubeck and Lee, 2016; Bubeck and Eldan, 2019). The Dikin ellipsoid corresponds to the distance measured by the Euclidean distance weighted by the Hessian matrix at the optimum. This weighted Euclidean distance is adapted to the geometry near the target parameter and thus leads to sharper bounds which do not depend on the minimum eigenvalue of the Hessian. This important property has been used fruitfully in various problems of learning theory and mathematical statistics (Zhang and Lin, 2015; Yang and Mohri, 2016; Faury et al., 2020).

The remainder of this chapter is organized as follows. In Section 2.2 we recall several definitions preliminary to our results. In Section 2.3 we introduce the framework of minimum KL divergence estimation and the Wald-type confidence set from classical asymptotic theory. In Section 2.4 we establish non-asymptotic bounds to characterize this confidences set, whose size is controlled by the *effective dimension*, in a non-asymptotic fashion. Our results hold for a general class of models encompassing generalized linear models characterized by the notion of *generalized self-concordance*. Along the way, we show how the effective dimension can be estimated from data and establish its estimation accuracy. This is a novel result and

Table 2.1: Examples of generalized linear models.

Model	**Data**	**Parameter**	**Conditional probability**
Linear	$X \in \mathbb{R}^d, Y \in \mathbb{R}$	$\theta \in \mathbb{R}^d$	$\propto \exp\left(-(y-\theta^\top x)^2/2\sigma^2\right)$
Poisson	$X \in \mathbb{R}^d, Y \in \mathbb{N}$	$\theta \in \mathbb{R}^d$	$\propto \exp(y\theta^\top x)/y!$
Logistic	$X \in \mathbb{R}^d, Y \in \{-1,1\}$	$\theta \in \mathbb{R}^d$	$= (1+\exp(-y\theta^\top x))^{-1}$
Softmax	$X \in \mathbb{R}^p, Y \in [K]$	$(w_k)_{k=1}^K \subset \mathbb{R}^p$	$\propto \exp(w_y^\top x)$

is of independent interest. We apply our results to compare Rao's score test, the likelihood ratio test, and the Wald test for goodness-of-fit testing in Section 2.5. Finally, in Section 2.6, we illustrate the interest of our results on synthetic data.

2.2 *Preliminaries*

2.2.1 Generalized linear models

As an extension of linear models, generalized linear models (GLM) introduced by Nelder and Wedderburn (1972) provide a class of models that are of broader applicability and maintain desirable statistical properties of linear models; see, e.g., also Wakefield (2013, Chapter 6.3). Given a pair of independent and response variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, a GLM assumes that the condition distribution $Y \mid X$ follows an exponential family. To be more concrete,

$$p_\theta(y \mid x) = \frac{\exp\left[\langle \theta, t(x,y)\rangle + h(x,y)\right]}{\int \exp\left[\langle \theta, t(x,\bar{y})\rangle + h(x,\bar{y})\right] \mathrm{d}\mu(\bar{y})} \mathrm{d}\mu(y), \tag{2.1}$$

where μ is a dominating measure on $\mathcal{Y}$, $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, and $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

To show the broad applicability of GLMs, we give below several popular examples and summarize them in Table 2.1.

(a) **Linear regression.** The conditional probability is $p_\theta(y \mid x) \propto \exp\left(-(y-\theta^\top x)^2/2\sigma^2\right)$ for $y \in \mathbb{R}$. This can be rewritten in the form of (2.1) with $t(x,y) := xy/\sigma^2$ and

$h(x,y) := -y^2/2\sigma^2$.

(b) **Poisson regression.** The conditional probability is $p_\theta(y \mid x) \propto \exp(y\theta^\top x)/y!$ for $y \in \mathbb{N}$. This can be rewritten in the form of (2.1) with $t(x,y) := xy$ and $h(x,y) := -\log y!$.

(c) **Logistic regression.** The conditional probability is $p_\theta(y \mid x) = (1+\exp(-y\theta^\top x))^{-1}$ for $y \in \{-1,1\}$. This can be rewritten in the form of (2.1) with $t(x,y) := x\mathbb{1}\{y=1\}$ and $h(x,y) \equiv 0$.

(d) **Softmax regression.** The conditional probability is $p_\theta(y \mid x) \propto \exp(w_y^\top x)$ for $y \in [K] := \{1,\ldots,K\}$ and $\{w_k\}_{k=1}^K \subset \mathbb{R}^p$. This can be rewritten in the form of (2.1) with $\theta^\top := (w_1^\top,\ldots,w_K^\top) \in \mathbb{R}^{Kp}$, $t(x,y)^\top := (0_p^\top,\ldots,0_p^\top,x^\top,\ldots,0_p^\top) \in \mathbb{R}^{Kp}$ whose elements from $(y-1)\tau+1$ to $y\tau$ are given by $x^\top$ and 0 elsewhere, and $h(x,y) \equiv 0$.

2.2.2 *Self-concordance*

We will use the notion of *self-concordance* from convex optimization in our analysis. Self-concordance originated from the analysis of the interior-point and Newton-type convex optimization methods (Nesterov and Nemirovskii, 1994). It is later modified by Bach (2010), which we call the *pseudo self-concordance*, to derive finite-sample bounds for the generalization properties of the logistic regression. This is later extended by Ostrovskii and Bach (2021) to a larger class of models. Recently, Sun and Tran-Dinh (2019) proposed the *generalized self-concordance* which unifies these two notions. In a different line of research, pseudo self-concordance has also been utilized to analyze logistic bandits (Faury et al., 2020; Abeille et al., 2021; Jun et al., 2021; Mason et al., 2022).

For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define $\mathrm{D}f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x+tu)|_{t=0}$, $\mathrm{D}^2 f(x)[u,v] := \mathrm{D}(\mathrm{D}f(x)[u])[v]$ for $x,u,v \in \mathbb{R}^d$, and $\mathrm{D}^3 f(x)[u,v,w]$ similarly.

Definition 2.1 (Generalized self-concordance)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be open and $f : \mathcal{X} \to \mathbb{R}$ be a closed convex function. For $R > 0$ and $\nu > 0$, we say f is (R,ν)-generalized self-concordant on $\mathcal{X}$ if*

$$\left|\mathrm{D}^3 f(x)[u,u,v]\right| \le R \left\|v\right\|_{\nabla^2 f(x)}^{\nu-2} \left\|v\right\|_2^{3-\nu} \left\|u\right\|_{\nabla^2 f(x)}^2, \quad \textit{for all } x \in \mathcal{X}, u \in \mathbb{R}^d,$$

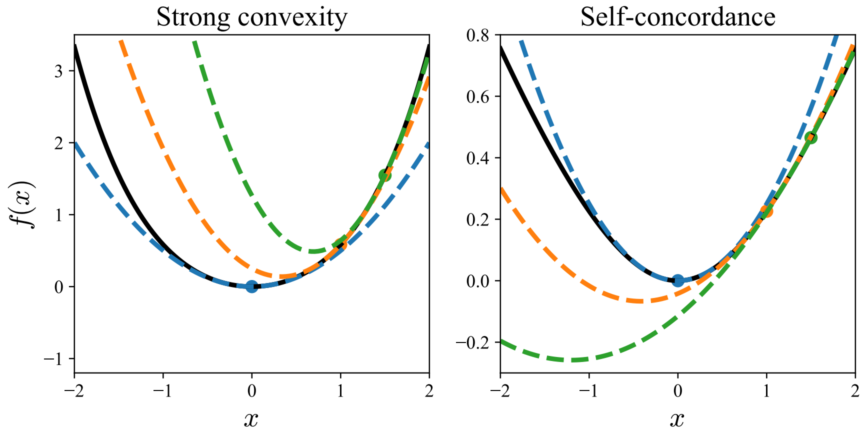

Figure 2.1: Strong convexity versus self-concordance. Black curve: objective function; colored dot: reference point; colored dashed curve: quadratic approximation at the corresponding reference point.

where $\|u\|^2_{\nabla^2 f(x)} := u^\top \nabla^2 f(x) u$.

Remark 2.1. *When* $\nu = 2$ *and* $\nu = 3$*, this definition recovers the pseudo self-concordance and the standard self-concordance, respectively. For our purposes, we focus on the case when* $\nu \in [2, 3]$.

In contrast to strong convexity which imposes a gross lower bound on the Hessian, generalized self-concordance specifies the rate at which the Hessian can vary, leading to a finer control on the Hessian; see Figure 2.1 for a illustration. This property is characterized by the following proposition. Let

$$d_\nu(x,y) := \begin{cases} R\,\|y-x\|_2 & \text{if } \nu = 2 \\ (\nu/2-1)R\,\|y-x\|_2^{3-\nu}\,\|y-x\|_x^{\nu-2} & \text{if } \nu > 2 \end{cases}$$

and

$$\omega_\nu(\tau) := \begin{cases} e^\tau & \text{if } \nu = 2 \\ (1-\tau)^{-2/(\nu-2)} & \text{if } \nu > 2. \end{cases}$$

Proposition 2.1 (Proposition 8 in Sun and Tran-Dinh (2019))**.** *Assume that f is (R,ν)-generalized self-concordant. For any $x, y \in dom(f)$, we have*

$$\frac{1}{\omega_\nu(d_\nu(x,y))}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \omega_\nu(d_\nu(x,y))\nabla^2 f(x),$$

where it holds if $d_\nu(x,y) < 1$ for the case $\nu > 2$.

One useful property of generalized self-concordant functions is that the sum of generalized self-concordant functions is also generalized self-concordant.

Proposition 2.2 (Proposition 1 in Sun and Tran-Dinh (2019))**.** *Let $\{f_i\}_{i=1}^n$ be a set of generalized self-concordant functions on $\mathcal{X}$ with parameters $\{(R_i,\nu)\}_{i=1}^n$. Then, for any $\{a_i\} \subset \mathbb{R}_+$, the function $f := \sum_{i=1}^n a_i f_i$ is (R,ν)-generalized self-concordant on $\mathcal{X}$ with $R := \max_{i\in[n]} a_i^{1-\nu/2} R_i$.*

Another useful property is that the local distance between the minimizer of f and x only depends on the geometry at x. It can be used to localize the maximum likelihood estimator as shown in Proposition 2.10. The proof is inspired by Ostrovskii and Bach (2021, Proposition B.4) and deferred to Appendix A.3. Let $\lambda_{\min} := \lambda_{\min}(H(x))$ and

$$R_\nu := \begin{cases} \lambda_{\min}^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2 - 1)\lambda_{\min}^{(\nu-3)/2} R & \text{if } \nu \in (2, 3]. \end{cases}$$

Proposition 2.3. *There exists $K_\nu \in (0, 1/2]$ such that, whenever $R_\nu \left\|\nabla f(x)\right\|_{\nabla^2 f(x)^{-1}} \le K_\nu$, the function f has a unique minimizer $\bar{x}$ and*

$$\left\|\bar{x} - x\right\|_{\nabla^2 f(x)} \le 4 \left\|\nabla f(x)\right\|_{\nabla^2 f(x)^{-1}}.$$

2.2.3 Sub-Gaussian random vectors

The exposition of this section largely follows Vershynin (2018). We start by recalling the definition of sub-Gaussian random variables.

Definition 2.2 (Sub-Gaussian random variables)**.** *Let X be a mean-zero random variable. We say X is sub-Gaussian with parameter σ^2, denoted by subG(σ^2), if $\mathbb{E}[\exp(X^2/\sigma^2)] \leq 2$. The sub-Gaussian norm of X is defined by*

$$\|X\|_{\psi_2} := \inf\left\{\sigma : \mathbb{E}[\exp(X^2/\sigma^2)] \leq 2\right\}.$$

It is well-known that the tail of a sub-Gaussian random variable is as least as light as a Gaussian. Concretely, there exists an absolute constant $c > 0$ such that, for every $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2).$$

The notion of a sub-Gaussian random vector is simply requiring all the one-dimensional projections to be sub-Gaussian.

Definition 2.3 (Sub-Gaussian random vectors)**.** *Let $S \in \mathbb{R}^d$ be a mean-zero random vector. We say S is sub-Gaussian if $\langle S, s\rangle$ is sub-Gaussian for every $s \in \mathbb{R}^d$. Moreover, we define the sub-Gaussian norm of S as*

$$\|S\|_{\psi_2} := \sup_{\|s\|_2=1} \|\langle S, s\rangle\|_{\psi_2}.$$

We call a random vector $S \in \mathbb{R}^d$ isotropic if $\mathbb{E}[S] = 0$ and $\mathbb{E}[SS^\top] = I_d$. The following theorem provides a tail bound for quadratic forms of isotropic sub-Gaussian random vectors.

Theorem 2.4 (Theorem A.1 in Ostrovskii and Bach (2021))**.** *Let $S \in \mathbb{R}^d$ be an isotropic random vector with $\|S\|_{\psi_2} \leq K$, and let $J \in \mathbb{R}^{d\times d}$ be positive semi-definite. Then,*

$$\mathbb{P}\left(\|S\|_J^2 - \operatorname{Tr}(J) \geq t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2\|J\|_2^2}, \frac{t}{K\|J\|_\infty}\right\}\right), \quad \textit{for all } t \geq 0,$$

where $\|S\|_J^2 := X^\top J X$ and c is an absolute constant. In other words, with probability at least $1-\delta$, we have

$$\|S\|_J^2 - \operatorname{Tr}(J) \lesssim K^2\left[\|J\|_2\sqrt{\log(e/\delta)} + \|J\|_\infty \log(1/\delta)\right],$$

where $\lesssim$ omits an absolute constant.

2.2.4 Matrix Bernstein condition

The exposition of this section largely follows Wainwright (2019). Before we introduce the matrix Bernstein condition, let us review the Bernstein condition for random variables. The Bernstein condition is a way to characterize the tail behavior of heavy tail random variables.

Definition 2.4 (Bernstein condition)**.** *Let X be a mean-zero random variable with variance σ^2. We say that X satisfies a Bernstein condition with parameter $b > 0$ if, for all $j \geq 2$,*

$$\left|\mathbb{E}[X^j]\right| \leq \frac{1}{2} j! b^{j-2} \sigma^2.$$

When X satisfies the Bernstein condition, its tail probability can be controlled by the Bernstein inequality

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/(2(\sigma^2 + bt))), \quad \text{for all } t \geq 0.$$

The matrix Bernstein inequality is an extension of the Bernstein inequality to random matrices. A key challenge in establishing such type of results is that matrices do not necessarily commute. Thanks to recent developments in random matrix theory, we have the following matrix Bernstein inequality. Let $\|J\| := \sqrt{\lambda_{\max}(J^\top J)}$ be the spectral norm for a squared matrix. Note that, when J is symmetric, we have $\|J\| = \max\{\lambda_{\max}(J), \lambda_{\min}(J)\}$. When J is random, we define $\mathbb{V}\mathrm{ar}(J) := \mathbb{E}[JJ^\top] - \mathbb{E}[J]\,\mathbb{E}[J]^\top$.

Definition 2.5 (Matrix Bernstein condition)**.** *Let $H \in \mathbb{R}^{d \times d}$ be a zero-mean symmetric random matrix. We say H satisfies a matrix Bernstein condition with parameter $b > 0$ if, for all $j \geq 3$,*

$$\mathbb{E}[H^j] \preceq \frac{1}{2} j! b^{j-2}\,\mathbb{V}\mathrm{ar}(H).$$

Theorem 2.5 (Theorem 6.17 in Wainwright (2019))**.** *Let $\{H_i\}_{i=1}^n$ be a sequence of independent zero-mean symmetric random matrices that satisfy the matrix Bernstein condition with* *parameter $b > 0$. Then, for all $t > 0$, it holds that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} H_i\right\| \geq t\right) \leq 2\,\mathrm{Rank}\left(\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(H_i)\right) \exp\left\{-\frac{nt^2}{2(\sigma^2 + bt)}\right\},$$

where $\sigma^2 := \frac{1}{n}\left\|\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(H_i)\right\|$.

2.3 Minimum Kullback-Leibler Divergence Estimation

Let $\mathcal{Z}$ be a measurable space. Let $Z \in \mathcal{Z}$ be a random element following some unknown distribution P. Consider a parametric family of distributions $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ which may or may not contain P. For simplicity of the presentation, we assume that P and P_θ are absolutely continuous w.r.t. some dominating measure with densities p and p_θ, respectively. We are interested in finding the parameter θ_* so that the model $P_{\theta_\star}$ best approximates the underlying distribution P.

For this purpose, one strategy is to minimize the KL divergence $\mathrm{KL}(P\|P_\theta)$ over $\theta \in \Theta$. Note that

$$\mathrm{KL}(P\|P_\theta) = \mathbb{E}\left[\log\frac{p(Z)}{p_\theta(Z)}\right] = \mathbb{E}[-\log p_\theta(Z)] + \mathbb{E}[\log p(Z)].$$

Consequently, minimizing $\mathrm{KL}(P\|P_\theta)$ over θ is equivalent to maximizing the expected log-likelihood $\mathbb{E}[\log p_\theta(Z)]$, which is known as the maximum likelihood method (Fisher, 1922). More broadly, the minimum KL divergence method fits into the framework of *statistical learning*—one selects a *loss function* $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and obtains $\theta_\star$ by minimizing the *population risk* $L(\theta) := \mathbb{E}[\ell(\theta; Z)]$. To maintain full generality, we work with the statistical learning framework while keeping in mind that we can recover the minimum KL divergence method by choosing $\ell(\theta; Z) = -\log p_\theta(Z)$. Throught out this chapter, we assume that

$$\theta_\star = \arg\min_{\theta\in\Theta} L(\theta)$$

uniquely exists and satisfies $\theta_\star \in \mathrm{int}(\Theta)$, $\nabla_\theta L(\theta_\star) = 0$, and $\nabla_\theta^2 L(\theta_\star) \succ 0$.

Consistent loss function. We focus on loss functions that are consistent in the following sense.

Assumption 2.1. *When the model is* well-specified*, i.e., there exists $\theta_0 \in \Theta$ such that $\mathbb{P} = P_{\theta_0}$, it holds that $\theta_\star = \theta_0$. We say such a loss function is* consistent*.*

In the statistics literature, such loss functions are known as proper scoring rules (Dawid et al., 2016). Due to the property of the KL divergence, the loss function $\ell(\theta; Z) = -\log p_\theta(Z)$ corresponding to the minimum KL divergence estimation (or maximum likelihood estimation) is consistent. We give here another popular choice of consistent loss functions.

Example 2.2 (Score matching estimation)**.** *One important example of consistent loss functions appears in* score matching *(Hyvärinen, 2005). Assume that $\mathcal{Z} = \mathbb{R}^p$. Let $p_\theta(z) = q_\theta(z)/\Lambda(\theta)$ where $\Lambda(\theta)$ is an unknown normalizing constant. We can choose the loss*

$$\ell(\theta; z) := \Delta_z \log q_\theta(z) + \frac{1}{2} \left\| \nabla_z \log q_\theta(z) \right\|^2 + const.$$

Here $\Delta := \sum_{k=1}^p \partial^2/\partial z_k^2$ is the Laplace operator. Since (Hyvärinen, 2005, Theorem 1)

$$L(\theta) = \frac{1}{2} \mathbb{E}\left[\| \nabla_z q_\theta(z) - \nabla_z p(z) \|^2 \right],$$

it follows that when $p = p_{\theta_0}$ we have $\theta_0 \in \arg\min_{\theta\in\Theta} L(\theta)$. In fact, when $q_\theta > 0$ and there is no θ such that $p_\theta =_{a.s.} p_{\theta_0}$, the true parameter θ_0 is the unique minimizer of L (Hyvärinen, 2005, Theorem 2).

Empirical risk minimization. In many real applications, we do not have access to the underlying distribution P. Instead, we have an i.i.d. sample $\{Z_i\}_{i=1}^n$ from P. To learn the parameter $\theta_\star$ from the data, we minimize the empirical risk to obtain the *empirical risk minimizer*

$$\theta_n \in \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \ell(\theta; Z_i).$$

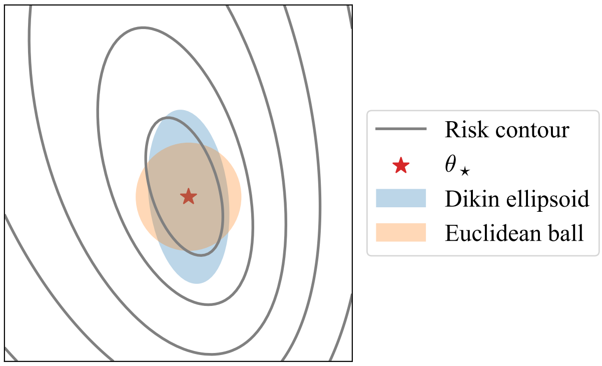

Figure 2.2: Dikin ellipsoid and Euclidean ball.

In Section 2.4, we will prove that, with high probability, the estimator θ_n exists and is unique under a generalized self-concordance assumption. To evaluate the performance of the estimator, we use the *excess risk*, i.e., $L(\theta_n) - L(\theta_\star)$ which is the difference between the population risk of θ_n and the best risk we can achieve. One of the main results in Section 2.4 is a non-asymptotic bound for the excess risk.

Remark 2.3. *When $L(\theta) = \mathbb{E}[-\log p_\theta(Z)]$ is the expected negative log-likelihood, the excess risk exactly equals the excess KL divergence defined in* (1.3).

Confidence set. In statistical inference, it is of great interest to quantify the uncertainty in the estimator θ_n. In classical asymptotic theory, this is achieved by constructing an asymptotic confidence set. We review here the commonly used *Wald confidence set*, assuming the model is well-specified. From classical asymptotic theory, we know that $n(\theta_n - \theta_\star)^\top H_n(\theta_n)(\theta_n - \theta_\star) \to_d \chi_d^2$, where $H_n(\theta) := \nabla^2 L_n(\theta)$. Hence, one may consider a confidence set $\{\theta : n(\theta_n - \theta)^\top H_n(\theta_n)(\theta_n - \theta) \le q_{\chi_d^2}(\delta)\}$ where $q_{\chi_d^2}(\delta)$ is the upper δ-quantile of χ_d^2. This confidence set enjoys two merits: 1) its shape is an ellipsoid (known as the *Dikin ellipsoid*) which is adapted to the optimization landscape induced by the population risk; 2) it is asymptotically valid, i.e., its coverage is exactly $1 - \delta$ as $n \to \infty$. However, due to its asymptotic nature, it is unclear how large n should be in order for it to be accurate.

Non-asymptotic theory usually focuses on developing finite-sample bounds for the *excess risk*, i.e., $\mathbb{P}(L(\theta_n) - L(\theta_\star) \le C_n(\delta)) \ge 1 - \delta$. To obtain a confidence set, one may assume that the population risk is twice continuously differentiable and λ-strongly convex. Consequently, we have $\lambda \left\| \theta_n - \theta_\star \right\|_2^2 / 2 \le L(\theta_n) - L(\theta_\star)$ and thus we can consider the confidence set $\mathcal{C}_{\text{finite},n}(\delta) := \{\theta : \|\theta_n - \theta\|_2^2 \le 2C_n(\delta)/\lambda\}$. Since it is originated from a finite-sample bound, it is valid for fixed n, i.e., $\mathbb{P}(\theta_\star \in \mathcal{C}_{\text{finite},n}(\delta)) \ge 1 - \delta$; however, it is usually conservative, meaning that the coverage is strictly larger than $1 - \delta$. Another drawback is that its shape is a Euclidean ball which remains the same no matter which loss function is chosen. We illustrate this phenomenon in Figure 2.2. Note that a similar observation has also been made in the bandit literature (Faury et al., 2020).

We are interested in developing non-asymptotic confidence sets. However, instead of using excess risk bounds and strong convexity, we construct in Section 2.4 the Wald confidence set in a non-asymptotic fashion, under a generalized self-concordance condition. This confidence set has the same shape as its asymptotic counterparts while maintaining validity for fixed n, which is achieved by characterizing the critical sample size enough to enter the asymptotic regime.

Effective dimension. A quantity that plays a central role in our analysis is the *effective dimension*. Define $G(\theta) := \mathbb{E}[\nabla_\theta \ell(\theta; Z) \nabla_\theta \ell(\theta; Z)^\top]$ and $H(\theta) := \mathbb{E}[\nabla_\theta^2 \ell(\theta; Z)]$.

Definition 2.6 (Effective dimension)**.** *Let* $\Omega(\theta) := H(\theta)^{-1/2} G(\theta) H(\theta)^{-1/2}$. *We define the effective dimension to be*

$$d_\star := \operatorname{Tr}(\Omega(\theta_\star)) = \operatorname{Tr}\left\{ H(\theta_\star)^{-1/2} G(\theta_\star) H(\theta_\star)^{-1/2} \right\}. \tag{2.2}$$

The effective dimension appears recently in non-asymptotic analyses of (penalized) M-estimation (see, e.g., Spokoiny, 2017; Ostrovskii and Bach, 2021). It provides a characterization of the complexity of the parameter space Θ that is adapted to both the data distribution and the loss function. When the model is well-specified, it can be shown that $H(\theta_\star) = G(\theta_\star)$ and thus $d_\star = d$. When the model is mis-specified, it can be much smaller than d depending on the spectra of $H(\theta_\star)$ and $G(\theta_\star)$; see Section 2.4.3 for a thorough discussion. The effective dimension is also closely connected to classical asymptotic theory of M-estimation under model misspecification. According to Huber (1967, Section 4), under suitable regularity conditions, $\sqrt{n}(\theta_n - \theta_\star)$ is asymptotically normal with mean 0 and covariance $H(\theta_\star)^{-1}G(\theta_\star)H(\theta_\star)^{-1}$. This implies that $H(\theta_\star)^{1/2}(\theta_n - \theta_\star)$ has asymptotic covariance $H(\theta_\star)^{-1/2}G(\theta_\star)H(\theta_\star)^{-1/2}$. Hence, the effective dimension is simply the trace of the limiting covariance matrix of $H(\theta_\star)^{1/2}(\theta_n - \theta_\star)$.

Dikin ellipsoid. Our analysis is local to a Dikin ellipsoid of the parameter $\theta_\star$ defined as

$$\Theta_r(\theta_\star) := \left\{\theta \in \Theta : \|\theta - \theta_\star\|_{H(\theta_\star)} < r\right\}, \tag{2.3}$$

where, unlike Euclidean balls, the radius is quantified by the Hessian-weighted Euclidean distance $\|\theta - \theta_\star\|_{H(\theta_\star)} := \sqrt{(\theta - \theta_\star)^\top H(\theta_\star)(\theta - \theta_\star)}$. As illustrated in Figure 2.2, while the Euclidean ball is agnostic to the risk function, the Dikin ellipsoid is adapted to the geometry of the underlying optimization landscape around $\theta_\star$. The Dikin ellipsoid was, up to our knowledge, first put to use in machine learning by Abernethy et al. (2008) in the context of sequential allocation of experiments and multi-armed bandits. The key observation is that, within the Dikin ellipsoid, the variation of the Hessian can be easily controlled. More recently, it is used to obtain finite-sample analysis of the maximum likelihood estimator (Spokoiny, 2012; Ostrovskii and Bach, 2021). Our results and proof techniques also rely on this observation. We show how to leverage this observation to obtain risk bounds and confidence sets for a broad class of statistical models under a generalized self-concordance assumption owing to the use of the matrix Bernstein inequality. For instance, we obtain confidence bounds for parameter estimation using score matching and generalized linear models under possible model misspecification as provided in Section 2.5.

2.4 Non-Asymptotic Analysis

We provide non-asymptotic analysis for the empirical risk minimization. We give in Section 2.4.1 the assumptions required by our analysis. In Section 2.4.2 we state our main results which include an excess risk bound and a confidence bound for the estimator. We sketch their main proof ideas in Section 2.4.4.

2.4.1 Assumptions

Notation. We denote by $S(\theta; z) := \nabla_\theta \ell(\theta; z)$ the gradient of the loss at z and $H(\theta; z) := \nabla_\theta^2 \ell(\theta; z)$ the Hessian at z. Their population versions are $S(\theta) := \mathbb{E}[S(\theta; Z)]$ and $H(\theta) := \mathbb{E}[H(\theta; Z)]$, respectively. We assume standard regularity assumptions so that $S(\theta) = \nabla_\theta L(\theta)$ and $H(\theta) = \nabla_\theta^2 L(\theta)$. Note that the two optimality conditions then read $S(\theta_\star) = 0$ and $H(\theta_\star) \succ 0$. It follows that $\lambda_\star := \lambda_{\min}(H(\theta_\star)) > 0$. Furthermore, we let $G(\theta; z) := S(\theta; z)S(\theta; z)^\top$ and $G(\theta) := \mathbb{E}[S(\theta; Z)S(\theta; Z)^\top]$ be the autocorrelation matrices of the gradient. We define their empirical quantities as $\ell_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$, $S_n(\theta) := \frac{1}{n}\sum_{i=1}^n S(\theta; Z_i)$, $H_n(\theta) := \frac{1}{n}\sum_{i=1}^n H(\theta; Z_i)$, and $G_n(\theta) := \frac{1}{n}\sum_{i=1}^n G(\theta; Z_i)$.

Our key assumption is that the loss function is globally generalized self-concordant. Recall its definition in Definition 2.1.

Assumption 2.2 (Pseudo self-concordance)**.** *For any $z \in \mathcal{Z}$, the loss $\ell(\cdot; z)$ is (R, ν) generalized self-concordant on Θ with $R > 0$ and $\nu \in [2, 3)$.*

Remark 2.4. *According to Proposition 2.2, both the empirical risk ℓ_n and the population risk are generalized self-concordant on Θ with the same order ν.*

We then make local distributional assumptions on the gradient and the Hessian. In order to control the empirical gradient $S_n(\theta)$, we assume that the standardized gradient at $\theta_\star$ is a sub-Gaussian random vector defined in Definition 2.3.

Assumption 2.3 (Sub-Gaussian gradient)**.** *There exists a constant $K_1 > 0$ such that the standardized gradient $G(\theta_\star)^{-1/2}[S(\theta_\star; Z) - S(\theta_\star)]$ is sub-Gaussian with parameter K_1.*

Remark 2.5. *When the loss function is of the form $\ell(\theta; z) = \ell(y, \theta^\top x)$, we have $S(\theta; Z) = \ell'(Y, \theta^\top X)X$. As a result, Assumption 2.3 holds true if (i) $\ell'(Y, \theta_\star^\top X)$ is sub-Gaussian and X is bounded or (ii) $\ell'(Y, \theta_\star^\top X)$ is bounded and X is sub-Gaussian. For least squares with $\ell(y, \theta^\top x) = \frac{1}{2}(y - \theta^\top x)^2$, the derivative $\ell'(Y, \theta_\star^\top X) = \theta_\star^\top X - Y$ is the negative residual. Assumption 2.3 is guaranteed if the residual is sub-Gaussian and X is bounded. For logistic regression with $\ell(y, \theta^\top x) = -\log \sigma(y \cdot \theta^\top x)$ where $\sigma(u) = (1 + e^{-u})^{-1}$, the derivative $\ell'(Y, \theta_\star^\top X) = [\sigma(Y \cdot \theta_\star^\top X) - 1]Y \in [-1, 1]$ is bounded. Thus, Assumption 2.3 is guaranteed if X is sub-Gaussian.*

In order to control the empirical Hessian $H_n(\theta)$, we assume that the standardized Hessian satisfies the matrix Bernstein condition (Definition 2.5) in a neighborhood of $\theta_\star$.

Assumption 2.4 (Matrix Bernstein of Hessian)**.** *There exist constants $K_2, r > 0$ such that, for any $\theta \in \Theta_r(\theta_\star)$, the standardized Hessian*

$$H(\theta)^{-1/2} H(\theta; Z) H(\theta)^{-1/2} - I_d$$

satisfies the matrix Bernstein condition with parameter K_2. Moreover,

$$\sigma_H^2 := \left\| \mathbb{V}\mathrm{ar}\left(H(\theta_\star)^{-1/2} H(\theta_\star; Z) H(\theta_\star)^{-1/2} \right) \right\| < \infty.$$

2.4.2 Main results

We now give the simplified versions of our main theorems. We use C_ν to represent a constant depending only on ν that may change from line to line; and $C_{K_1,\nu}$ similarly. We use $\lesssim$ and $\gtrsim$ to hide constants depending only on K_1, K_2, σ_H, ν. Recall that $\lambda_\star := \lambda_{\min}(H(\theta_\star))$ and the effective dimension $d_\star$ from Definition 2.6. The precise versions and proofs can be found in Appendix A.1.

Theorem 2.6. *Under Assumptions 2.2, 2.3, and 2.4 with $r = 0$, whenever*

$$n \gtrsim \log(2d/\delta) + \lambda_\star^{-1} \left[R^2 d_\star \log(e/\delta) \right]^{1/(3-\nu)}, \tag{2.4}$$

the empirical risk minimizer θ_n uniquely exists and satisfies, with probability at least $1-\delta$,

$$\|\theta_n - \theta_\star\|^2_{H(\theta_\star)} \lesssim \frac{d_\star + \log(e/\delta)\,\|\Omega(\theta_\star)\|_2}{n}. \tag{2.5}$$

Moreover, it holds that

$$L(\theta_n) - L(\theta_\star) \lesssim \frac{d_\star + \log(e/\delta)\,\|\Omega(\theta_\star)\|_2}{n}. \tag{2.6}$$

Remark 2.6. *When the loss $\ell(\theta; z) = -\log p_\theta(z)$ is the negative log-likelihood, the bound in (2.6) also holds for the excess KL divergence defined in (1.3).*

Remark 2.7. *Note that $\|\Omega(\theta_\star)\|_2$ is (usually much) smaller than $d_\star = \mathrm{Tr}(\Omega(\theta_\star))$. In fact, when the model is well-specified $\|\Omega(\theta_\star)\|_2 = 1$ and $d_\star = d$. Hence, the leading term in (2.5) is $d_\star/n$, which matches the misspecifed Cramér-Rao lower bound (e.g., Fortunati et al., 2016, Thm. 1) up to a constant factor.*

Remark 2.8. *The results in Theorem 2.6 can be extended to semi-parametric models as demonstrated in Liu et al. (2022a).*

With a local matrix Bernstein condition, we can replace $H(\theta_\star)$ by $H_n(\theta_n)$ in (2.5) and obtain a non-asymptotic version of the Wald confidence set.

Theorem 2.7. *Suppose Assumptions 2.2, 2.3, and 2.4 with $r = C_\nu \lambda_\star^{(3-\nu)/2}/R$ hold true. Let*

$$\mathcal{C}_n(\delta) := \left\{\theta \in \Theta : \|\theta - \theta_n\|^2_{H_n(\theta_n)} \le C_{K_1,\nu} \frac{d_\star + \log(e/\delta)\,\|\Omega(\theta_\star)\|_2}{n}\right\}. \tag{2.7}$$

Then we have $\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \ge 1-\delta$ whenever n satisfies

$$n \gtrsim \log\frac{2d}{\delta} + d\log n + \lambda_\star^{-1}\left[R^2 d_\star \log\frac{e}{\delta}\right]^{\frac{1}{3-\nu}}. \tag{2.8}$$

Theorem 2.7 suggests that the tail probability of $\|\theta_n - \theta_\star\|^2_{H_n(\theta_n)}$ is governed by a χ^2 distribution with $d_\star$ degrees of freedom, which coincides with the asymptotic result. In fact, according to Huber (1967, Section 4), under suitable regularity assumptions, it holds that $\sqrt{n}H_n(\theta_n)^{1/2}(\theta_n - \theta_\star) \to_d W \sim \mathcal{N}(0, \Omega(\theta_\star))$ which implies that

$$n(\theta_n - \theta_\star)^\top H_n(\theta_n)(\theta_n - \theta_\star) \to_d W^\top W.$$

This induces an asymptotic confidence set with a similar form of (2.7) whose radius is given by $O(\mathbb{E}[W^\top W]/n) = O(d_\star/n)$. Our result characterizes the critical sample size enough to enter the asymptotic regime.

When the model is misspecified, the effective dimension $d_\star$ is unknown and thus we cannot construct the confidence set in Theorem 2.7. Alternatively, we use the following empirical counterpart

$$d_n := \operatorname{Tr}\left(H_n(\theta_n)^{-1/2} G_n(\theta_n) H_n(\theta_n)^{-1/2}\right).$$

The next result implies that we do not lose much if we replace $d_\star$ by d_n. This result is novel and of independent interest since one also needs to estimate $d_\star$ in order to construct asymptotic confidence sets under model misspecification.

Assumption 2.3'. *There exist constants $K_1, r > 0$ such that, for any $\theta \in \Theta_r(\theta_\star)$, we have $\left\|G(\theta)^{-1/2} S(\theta; Z)\right\|_{\psi_2} \le K_1$.*

Assumption 2.5. *There exists $r > 0$ such that $M := \mathbb{E}[M(Z)] < \infty$ where*

$$M(z) := \sup_{\theta_1 \neq \theta_2 \in \Theta_r(\theta_\star)} \frac{\left\|G(\theta_\star)^{-1/2}[G(\theta_1; z) - G(\theta_2; z)]G(\theta_\star)^{-1/2}\right\|_2}{\|\theta_1 - \theta_2\|_{H(\theta_\star)}}.$$

Remark 2.9. *Assumption 2.5 is a Lipschitz-type condition for $G(\theta; z)$. This assumption was previously used by (Mei et al., 2018, Assumption 3) to analyze non-convex risk landscapes.*

Proposition 2.8. *Let $\nu \in [2, 3)$. Under Asms. 2.2, 2.3', 2.4 and 2.5 with $r = C_\nu \lambda_\star^{(\nu-3)/2}/R$, it holds that*

$$\frac{1}{C_\nu} d_\star \le d_n \le C_\nu d_\star,$$

with probability at least $1 - \delta$, whenever n is large enough (see Appendix A.1 for the precise condition).

Remark 2.10. *The precise version of Proposition 2.8 in Appendix A.1 implies that d_n is a consistent estimator of d.*

Table 2.2: Comparison between the effective dimension $d_\star$ and the parameter dimension d in different regimes of eigendecays of $G(\theta_\star)$ and $H(\theta_\star)$ assuming they share the same eigenvectors.

	Eigendecay		**Dimension Dependency**		**Ratio**
	$G(\theta_\star)$	$H(\theta_\star)$	$d_\star$	d	$d_\star/d$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$d^{(\beta-\alpha+1)\vee 0}$	d	$d^{(\beta-\alpha)\vee(-1)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$d^{1-\alpha}e^{\nu d}$	d	$d^{-\alpha}e^{\nu d}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	1	d	d^{-1}
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	d if $\mu = \nu$	d	1 if $\mu = \nu$
			1 if $\mu > \nu$		d^{-1} if $\mu > \nu$
			$e^{(\nu-\mu)d}$ if $\mu < \nu$		$d^{-1}e^{(\nu-\mu)d}$ if $\mu < \nu$

With Proposition 2.8 at hand, we can obtain non-asymptotic confidence sets involving d_n, which can be computed from data.

Corollary 2.9. *Suppose the same assumptions in Proposition 2.8 hold true. Let*

$$\mathcal{C}_n'(\delta) := \left\{\theta \in \Theta : \|\theta - \theta_\star\|^2_{H_n(\theta_n)} \le C_{K_1,\nu} \log(e/\delta)\frac{d_n}{n}\right\}.$$

Then we have $\mathbb{P}(\theta_\star \in \mathcal{C}_n'(\delta)) \ge 1 - \delta$ whenever n satisfies the same condition as in Proposition 2.8.

2.4.3 Discussion

Fisher information and model misspecification. When the model is well-specified, the autocorrelation matrix $G(\theta)$ coincides with the well-known Fisher information $\mathcal{I}(\theta) := \mathbb{E}_{Z\sim P_\theta}[S(\theta; Z)S(\theta; Z)^\top]$ at $\theta_\star$. The Fisher information plays a central role in mathematical statistics and, in particular, M-estimation; see (Pennington and Worah, 2018; Kunstner et al., 2019; Ash et al., 2021; Soen and Sun, 2021) for recent developments in this line of research. It quantifies the amount of information a random variable carries about the model parameter. Under a well-specified model, it also coincides with the Hessian matrix $H(\theta)$ at the optimum which captures the local curvature of the population risk. When the model is misspecified, the Fisher information deviates from the Hessian matrix. In the asymptotic regime, this discrepancy is reflected in the limiting covariance of the weighted M-estimator which admits a sandwich form $H(\theta_\star)^{-1/2}G(\theta_\star)H(\theta_\star)^{-1/2}$; see, e.g., (Huber, 1967, Section 4).

Effective dimension. The counterpart of the sandwich covariance in the non-asymptotic regime is the effective dimension $d_\star$; see, e.g., (Spokoiny, 2017; Ostrovskii and Bach, 2021). Our bounds also enjoy the same merit—its dimension dependency is via the effective dimension. When the model is well-specified, the effective dimension reduces to d, recovering the same rate of convergence $O(d/n)$ as in classical linear regression; see, e.g., (Bach, 2021, Proposition 3.5). When the model is misspecified, the effective dimension provides a characterization of the problem complexity which is adapted to both the data distribution and the loss function via the matrix $H(\theta_\star)^{-1/2}G(\theta_\star)H(\theta_\star)^{-1/2}$. To gain a better understanding on the effective dimension $d_\star$, we compare it with d in Table 2.2 under different regimes of eigendecay, assuming that $G(\theta_\star)$ and $H(\theta_\star)$ share the same eigenvectors. It is clear that, when the spectrum of $G(\theta_\star)$ decays faster than the one of $H(\theta_\star)$, the dimension dependency can be better than $O(d)$. In fact, it can be as good as $O(1)$ when the spectrum of $G(\theta_\star)$ and $H(\theta_\star)$ decay exponentially and polynomially, respectively.

Comparison to classical asymptotic theory. Classical asymptotic theory of M-estimation is usually based on two assumptions: (a) the model is well-specified and (b) the sample size n is much larger than the parameter dimension d. These assumptions prevent it from being applicable to many real applications where the parametric family is only an approximation to the unknown data distribution and the data is of high dimension involving a large amount of parameters. On the contrary, our results do not require a well-specified model and the dimension dependency is replaced by the effective dimension $d_\star$ which captures the complexity of the parameter space. Moreover, they are of non-asymptotic nature—they hold true for any n as long as it exceeds some constant factor of $d_\star$. This allows the number of parameters to potentially grow with the same size.

Comparison to recent non-asymptotic theory. Recently, Spokoiny (2012) achieved a breakthrough on finite-sample analysis of parametric M-estimation. Although being fully general, their results require strong global assumptions on the deviation of the empirical risk process and are built upon advanced tools from empirical process theory. Restricting ourselves to generalized self-concordant losses, we are able to provide a more transparent analysis with neater assumptions only in a neighborhood of the optimum parameter $\theta_\star$. Moreover, our results maintain some generality, covering several interesting examples in statistical machine learning as provided in Section 2.5.1.

Ostrovskii and Bach (2021) also considered self-concordant losses for M-estimation. However, their results are limited to generalized linear models whose loss is (pseudo) self-concordant and admits the form $\ell(\theta; Z) := \ell(Y, \theta^\top X)$. While sharing the same rate $O(d_\star/n)$, our results are more general than theirs in two aspects. First, the loss need not be of the form $\ell(Y, \theta^\top X)$, encompassing the score matching loss in Example 2.13 below. Second, we go beyond pseudo self-concordance via the notion of generalized self-concordance. Moreover, they focus on bounding the excess risk rather than providing confidence sets, and they do not study the estimation of $d_\star$.

Regularization. Our results can also be applied to regularized empirical risk minimization by including the regularization term in the loss function. Let θ_n^λ and $\theta_\star^\lambda$ be the minimizers of the *regularized* empirical and population risk, respectively. Let

$$d_\star^\lambda := \operatorname{Tr}\left((H(\theta_\star)^\lambda)^{-1/2} G(\theta_\star)^\lambda (H(\theta_\star)^\lambda)^{-1/2}\right).$$

where $H(\theta_\star)^\lambda$ and $G(\theta_\star)^\lambda$ are the regularized Hessian and the autocorrelation matrix of the regularized gradient at $\theta_\star^\lambda$, respectively. Then our results characterize the concentration of θ_n^λ around $\theta_\star^\lambda$:

$$\left\|\theta_n^\lambda - \theta_\star^\lambda\right\|_{H(\theta_\star)^\lambda}^2 \leq O(d_\star^\lambda/n).$$

This result coincides with Spokoiny (2017, Theorem 2.1). If the goal is to estimate the unregularized population risk minimizer $\theta_\star$, then we need to pay an additional error $\left\|\theta_\star^\lambda - \theta_\star\right\|_{H(\theta_\star)^\lambda}^2$ which is referred to as the modeling bias (Spokoiny, 2017, Section 2.5). One can invoke a so-called *source condition* to bound the modeling bias and a *capacity condition* to bound $d_\star^\lambda$. An optimal value of λ can be obtained by balancing between these two terms (see, e.g., Marteau-Ferey et al., 2019).

2.4.4 Proof sketches

We give the proof sketch of Theorems 2.6 and 2.7 here. We start with showing the existence and uniqueness of θ_n. The next result shows that θ_n exists and is unique whenever the quadratic form $S_n(\theta_\star)^\top H_n^{-1}(\theta_\star) S_n(\theta_\star)$ is small. Note that this quantity is also known as the Rao's score statistic for goodness-of-fit testing. This result also localizes the empirical risk minimizer to a neighborhood of the optimal parameter $\theta_\star$.

Proposition 2.10. *Under Assumption 2.2, whenever*

$$\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} \leq C_\nu[\lambda_{\min}(H_n(\theta_\star))]^{(3-\nu)/2}/(Rn^{\nu/2-1}),$$

the estimator θ_n uniquely exists and satisfies

$$\|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \leq 4\,\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}\,.$$

The main tool used in the proof of Proposition 2.10 is a strong convexity type result for generalized self-concordant functions recalled in Appendix A.3. In order to apply Proposition 2.10, we need to a bound for $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$ which is summarized in the following proposition.

Proposition 2.11. *Under Assumptions 2.3 and 2.4 with $r = 0$, it holds that, with probability at least $1 - \delta$,*

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \lesssim \frac{d_\star + \log(e/\delta) \|\Omega(\theta_\star)\|_2}{n}$$

whenever $n \gtrsim \log(2d/\delta)$.

The proof of Proposition 2.11 consists of two steps: (a) lower bound $H_n(\theta_\star)$ by $H(\theta_\star)$ up to a constant using the Bernstein inequality and (b) upper bound $\|S_n(\theta_\star)\|_{H^{-1}(\theta_\star)}$ using a concentration inequality for isotropic random vectors, where the tools are recalled in Theorem 2.4 and Theorem 2.5. Combining them completes the proof.

Note that Proposition 2.11 implies that $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$ can be arbitrarily small and thus satisfies the requirement in Proposition 2.10 for sufficiently large n. This not only proves the existence and uniqueness of the empirical risk minimizer θ_n but also provides an upper bound for $\|\theta_n - \theta_\star\|_{H_n(\theta_\star)}$ through $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$. Now using the result from step (a) above gives the bound (2.5). To prove the excess risk bound (2.6), we use the Taylor expansion

$$L(\theta_n) - L(\theta_\star) = (\theta_n - \theta_\star)^\top S(\theta_\star) + \frac{1}{2}\|\theta_n - \theta\|^2_{H(\bar\theta)}$$

for some $\bar\theta \in \mathrm{Conv}\{\theta_n, \theta_\star\}$. The first order optimality condition implies that $S(\theta_\star) = 0$ and thus $L(\theta_n) - L(\theta_\star) \le \frac{1}{2}\|\theta_n - \theta\|^2_{H(\bar\theta)}$. Moreover, due to Proposition 2.1, we can upper bound $H(\bar\theta)$ by $H(\theta_\star)$ paying a factor of $e^{R\|\bar\theta - \theta_\star\|}$. Combining it with (2.5) leads to (2.6).

To prove Theorem 2.7, it remains to upper bound $\|\theta_n - \theta_\star\|_{H_n(\theta_n)}$ by $\|\theta_n - \theta_\star\|_{H(\theta_\star)}$ up to a constant factor. This can be achieved by the following result.

Proposition 2.12. *Under Assumptions 2.2 and 2.4 with $r = C_\nu \lambda_\star^{(\nu-3)/2}/R$, it holds that, with probability at least $1 - \delta$,*

$$\frac{1}{2C_\nu} H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2} C_\nu H(\theta_\star), \quad \textit{for all } \theta \in \Theta_r(\theta_\star),$$

whenever whenever $n \gtrsim \{\log(2d/\delta) + d(\nu/2 - 1)\log n\}$.

Remark 2.11. *Our proof techniques developed here can also be harnessed to analyze influence functions. This has been done in a follow-up work (Fisher et al., 2022).*

2.5 Applications

We give several examples whose loss function is generalized self-concordant so that our results can be applied. We also provide non-asymptotic analysis for Rao's score test, the likelihood ratio test, and the Wald test in goodness-of-fit testing. All the proofs and derivations are deferred to Appendix A.2.

2.5.1 Examples

Example 2.12 (Generalized linear models)**.** *Let $Z := (X, Y)$ be a pair of input and output, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Let $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ and μ be a measure on $\mathcal{Y}$. Consider the statistical model*

$$p_\theta(y \mid x) \sim \frac{\exp(\theta^\top t(x,y) + h(x,y))}{\int \exp(\theta^\top t(x,\bar{y}) + h(x,\bar{y}))\mathrm{d}\mu(\bar{y})}\mathrm{d}\mu(y)$$

with $\|t(X,Y)\|_2 \leq_{a.s.} M$. It induces the loss function

$$\ell(\theta; z) := -\theta^\top t(x,y) - h(x,y) + \log \int \exp(\theta^\top t(x,\bar{y}) + h(x,\bar{y}))\mathrm{d}\mu(\bar{y}),$$

which is generalized self-concordant for $\nu = 2$ and $R = 2M$. Moreover, this model satisfies Assumptions 2.3, 2.4, 2.5 and 2.3'.

Example 2.13 (Score matching with exponential families)**.** *Assume that $\mathbb{Z} = \mathbb{R}^p$. Consider an exponential family on $\mathbb{R}^d$ with densities*

$$\log p_\theta(z) = \theta^\top t(z) + h(z) - \Lambda(\theta).$$

The non-normalized density q_θ then reads $\log q_\theta(z) = \theta^\top t(z) + h(z)$. As a result, the score matching loss becomes

$$\ell(\theta; z) = \frac{1}{2}\theta^\top A(z)\theta - b(z)^\top \theta + c(z) + const,$$

where $A(z) := \sum_{k=1}^p \frac{\partial t(z)}{\partial z_k}\big(\frac{\partial t(z)}{\partial z_k}\big)^\top$ is p.s.d, $b(z) := \sum_{k=1}^p \Big[\frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial h(z)}{\partial z_k}\frac{\partial t(z)}{\partial z_k}\Big]$, and $c(z) := \sum_{k=1}^p \Big[\frac{\partial^2 h(z)}{\partial z_k^2} + \big(\frac{\partial h(z)}{\partial z_k}\big)^2\Big]$. Therefore, the score matching loss $\ell(\theta; z)$ is convex. Moreover, since the third derivatives of $\ell(\cdot; z)$ is zero, the score matching loss is generalized self-concordant for all $\nu \geq 2$ and $R \geq 0$.

2.5.2 Rao's score test and its relatives

We discuss how our results can be applied to analyze three classical goodness-of-fit tests. In this subsection, we will assume that the model is well-specified. Due to Assumption 2.1, we will use $\theta_\star$ to denote the true parameter of $\mathbb{P}$ and reserve θ_0 for the parameter under the null hypothesis.

Given a subset $\Theta_0 \subset \Theta$, a goodness-of-fit testing problem is to test the hypotheses

$$\mathbf{H}_0 : \theta_\star \in \Theta_0 \leftrightarrow \mathbf{H}_1 : \theta_\star \notin \Theta_0.$$

We focus on a simple null hypothesis where $\Theta_0 := \{\theta_0\}$ is a singleton. A statistical test consists of a test statistic $T := T(Z_1, \dots, Z_n)$ and a prescribed critical value t, and we reject the null hypothesis if $T > t$. The performance is quantified by the *type I error rate* $\mathbb{P}(T > t \mid \mathbf{H}_0)$ and *statistical power* $\mathbb{P}(T > t \mid \mathbf{H}_1)$. Classical goodness-of-fit tests include Rao's score test, the likelihood ratio test (LRT), and the Wald test. Their test statistics are $T_{\mathrm{Rao}} := \|S_n(\theta_0)\|^2_{H_n^{-1}(\theta_0)}$, $T_{\mathrm{LR}} := 2[\ell_n(\theta_0) - \ell_n(\theta_n)]$, and $T_{\mathrm{Wald}} := \|\theta_n - \theta_0\|^2_{H_n(\theta_n)}$, respectively.

Our approach can be applied to analyze the type I error rate of these tests as summarized in the following proposition.

Proposition 2.13 (Type I error rate)**.** *Suppose that Assumptions 2.3 and 2.4 with $r = 0$ hold true. Under $\mathbf{H}_0$, we have, with probability at least $1 - \delta$,*

$$T_{Rao} \lesssim \log{(e/\delta)} \frac{d}{n}$$

whenever $n \gtrsim \log{(2d/\delta)}$. Furthermore, if Assumptions 2.2 to 2.4 with $r = C_\nu \lambda_\star^{(\nu-3)/2}/R$ hold true, we have, with probability at least $1 - \delta$,

$$T_{LR} \lesssim \log{(e/\delta)} \frac{d}{n} \quad \textit{and} \quad T_{Wald} \lesssim \log{(e/\delta)} \frac{d}{n}$$

whenever n satisfies (2.8).

This result implies that the three test statistics all scale as $O(d/n)$ under the null hypothesis. Consequently, for a fixed significance level $\alpha \in (0, 1)$, we can choose the critical value

$t = t_n(\alpha) = O(d/n)$ so that their type I error rates are below α. With this choice, we can then characterize the statistical powers of these tests under alternative hypotheses $\theta_\star \neq \theta_0$ where $\theta_\star$ may depend on n. Recall $\Omega(\theta) := G(\theta)^{1/2} H(\theta)^{-1} G(\theta)^{1/2}$ and let $h(\tau) := \min\{\tau^2, \tau\}$.

Proposition 2.14 (Statistical power). *Let $\theta_\star \neq \theta_0$. The following statements are true for sufficiently large n.*

(a) Suppose that Assumptions 2.2 to 2.4 hold true with $r = 0$. When $\theta_\star - \theta_0 = O(n^{-1/2})$ and $\tau_n := t_n(\alpha)/4 - \|S(\theta_0)\|^2_{H(\theta_0)^{-1}} - \mathrm{Tr}(\Omega(\theta_0))/n > 0$, we have

$$\mathbb{P}(T_{Rao} > t_n(\alpha)) \leq 2de^{-C_{K_2,\sigma_H} n} + e^{-C_{K_1} h(n\tau_n/\|\Omega(\theta_0)\|_2)}.$$

When $\theta_* - \theta_n = \omega(n^{-1/2})$, we have

$$\mathbb{P}(T_{Rao} > t_n(\alpha)) \geq 1 - 2de^{-C_{K_2,\sigma_H} n} - e^{-C_{K_1} n\bar{\tau}_n/\|\Omega(\theta_0)\|_2},$$

where $\bar{\tau}_n = \Theta(\|\theta_\star - \theta_n\|^2)$.

(b) Suppose that the assumptions in Theorem 2.7 hold true. When $\theta_\star - \theta_0 = O(n^{-1/2})$ and $\tau'_n := t_n(\alpha)/384 - \|\theta_\star - \theta_0\|^2_{H(\theta_\star)}/64 - d/n > 0$, we have

$$\mathbb{P}(T_{LR} > t_n(\alpha)) \leq e^{-C_{K_1} h(n\tau'_n/\|\Omega(\theta_\star)\|_2)} + e^{-C_{K_1,\nu}(\lambda_\star n)^{3-\nu}/(R^2 d)}.$$

When $\theta_* - \theta_n = \omega(n^{-1/2})$, we have

$$\mathbb{P}(T_{LR} > t_n(\alpha)) \geq 1 - e^{-C_{K_1} \frac{n\bar{\tau}'_n}{\|\Omega(\theta_\star)\|_2}} - e^{-\frac{C_{K_1,\nu}(\lambda_\star n)^{3-\nu}}{R^2 d}},$$

where $\bar{\tau}'_n = \Theta(\|\theta_\star - \theta_n\|^2)$.

(c) The same statements replacing T_{LR} by T_{Wald}.

According to Proposition 2.14, when $\theta_\star - \theta_0 = O(n^{-1/2})$, the powers of the three tests are asymptotically upper bounded; when $\theta_\star - \theta_0 = \omega(n^{-1/2})$, the power of Rao's score test tends to one at rate $O(e^{-n\|\theta_\star - \theta_0\|^2})$ and the ones of the other two tests tend to one at rate $O(e^{-n\|\theta_\star - \theta_0\|^2 \wedge n^{3-\nu}})$.

2.5.3 Score-based change detection

Rao's score test can also be used to detect changes in model parameters. To be more concrete, we consider a well-specified model with the true parameter $\theta_\star$. Under abnormal circumstances, this true value may not remain the same for all observations. Hence, we allow a potential parameter change in the model—the model parameter $\theta = \theta_k$ may evolve over time, i.e., $Z_k \sim P_{\theta_k}$. A time point $\tau \in [n-1] := \{1, \dots, n-1\}$ is called a *changepoint* if there exists $\Delta \neq 0$ such that $\theta_k = \theta_\star$ for $k \leq \tau$ and $\theta_k = \theta_\star + \Delta$ for $k > \tau$. We say that there is a jump (or change) in the data sequence if such a changepoint exists. We aim to determine if there exists a jump in this sequence, which we formalize as a hypothesis testing problem.

(P0) Testing the presence of a jump

$$\begin{aligned} &\mathbf{H}_0 : \theta_k = \theta_\star \text{ for all } k = 1, \dots, n \\ &\mathbf{H}_1 : \text{after some time } \tau\text{, } \theta_k \text{ jumps from } \theta_\star \text{ to } \theta_\star + \Delta. \end{aligned}$$

Likelihood score and score-based testing. Let $\mathbb{1}\{\cdot\}$ be the indicator function. Given $\tau \in [n-1]$ and $1 \leq s \leq t \leq n$, we define the partial log-likelihood under the alternative as

$$\ell_{s:t}(\theta, \Delta; \tau) := \sum_{k=s}^{t} \ell(\theta + \Delta \mathbb{1}\{k > \tau\}; Z_k).$$

We will write $\ell_{s:t}(\theta, \Delta)$ for short if there is no confusion. Under the null, we denote by $\ell_{s:t}(\theta) := \ell_{s:t}(\theta, 0; n)$ the partial log-likelihood. The *score function* w.r.t. θ is defined as $S_{s:t}(\theta) := \nabla_\theta \ell_{s:t}(\theta)$, and the *Hessian* w.r.t. θ is denoted by $H_{s:t}(\theta) := \nabla_\theta^2 \ell_{s:t}(\theta)$.

Let us design a test for Problem (P0). We start with the case when the changepoint τ is fixed. A standard choice is the *generalized score statistic* given by

$$R_n(\tau) := S_{\tau+1:n}^\top(\theta_n) H_n(\theta_n; \tau)^{-1} S_{\tau+1:n}(\theta_n), \tag{2.9}$$

where $H_n(\theta_n; \tau)$ is the *partial observed information* w.r.t. Δ (Wakefield, 2013, Chapter 2.9)

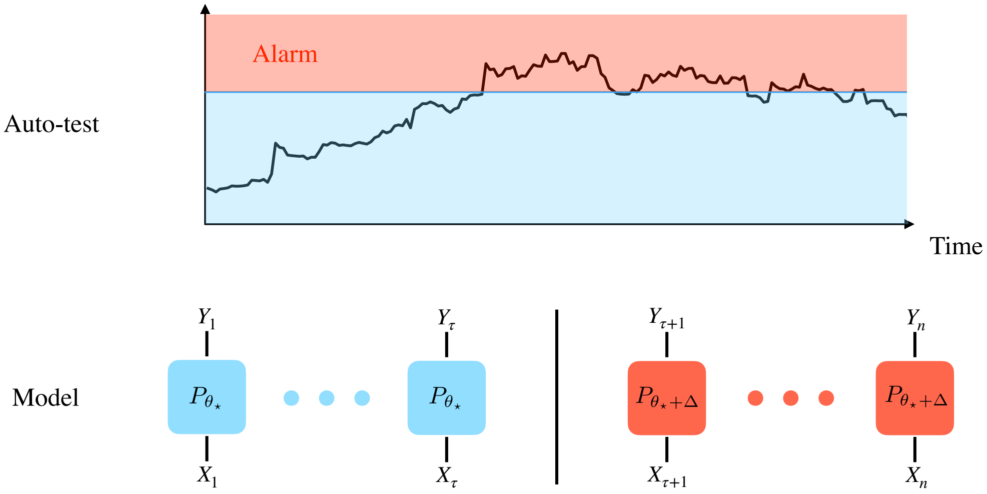

Figure 2.3: Illustration of detecting changes in model parameters.

defined as

$$H_n(\theta_n;\tau) := H_{\tau+1:n}(\theta_n) - H_{\tau+1:n}(\theta_n)^\top H_{1:n}(\theta_n)^{-1} H_{\tau+1:n}(\theta_n). \tag{2.10}$$

To adapt to an unknown changepoint τ, a natural statistic is $R_{\text{lin}} := \max_{\tau\in[n-1]} R_n(\tau)$. And, given a significance level α, the decision rule reads $\psi_{\text{lin}}(\alpha) := \mathbb{1}\{R_{\text{lin}} > H_{\text{lin}}(\alpha)\}$, where $H_{\text{lin}}(\alpha)$ is a prescribed threshold. We call R_{lin} the *linear statistic* and ψ_{lin} the *linear test.*

Sparse alternatives. There are cases when the jump only happens in a small subset of components of $\theta_\star$. The linear test, which is built assuming the jump is large, may fail to detect such small jumps. Therefore, we also consider *sparse alternatives.*

(P1) Testing the presence of a small jump:

$$\begin{aligned}
&\mathbf{H}_0 : \theta_k = \theta_\star \text{ for all } k = 1, \ldots, n\\
&\mathbf{H}_1 : \text{after some time } \tau,\ \theta_k \text{ jumps from } \theta_\star \text{ to } \theta_\star + \Delta,\\
&\qquad \text{where } \Delta \text{ has at most } M \text{ nonzero entries.}
\end{aligned}$$

Here M is referred to as the *maximum cardinality*, which is set to be much smaller than d, the dimension of θ. We denote by T the changed components, i.e., $\Delta_T \neq 0$ and $\Delta_{[d]\setminus T} = 0$.

Given a fixed T, we consider the *truncated statistic*

$$R_n(\tau, T) = S_{\tau+1:n}^\top(\theta_n)_T \left[H_n(\theta_n;\tau)_{T,T}\right]^{-1} S_{\tau+1:n}(\theta_n)_T.$$

Let $\mathcal{T}_m$ be the collection of all subsets of size m of $[d]$. To adapt to unknown T, we use

$$R_n(\tau, M; \alpha) := \max_{m\in[M]} \max_{T\in\mathcal{T}_m} H_m(\alpha)^{-1} R_n(\tau, T), \tag{2.11}$$

where we use a different threshold $H_m(\alpha)$ for each $m \in [M]$. Finally, since τ is also unknown, we propose $R_{\text{scan}}(\alpha) := \max_{\tau\in[n-1]} R_n(\tau, M; \alpha)$, with the decision rule $\psi_{\text{scan}}(\alpha) := \mathbb{1}\{R_{\text{scan}}(\alpha) > 1\}$. We call $R_{\text{scan}}(\alpha)$ the *scan statistic* and ψ_{scan} the *scan test.*

To combine the respective strengths of these two tests, we consider the test

$$\psi_{\text{auto}}(\alpha) := \max\{\psi_{\text{lin}}(\alpha_l), \psi_{\text{scan}}(\alpha_s)\}, \tag{2.12}$$

with $\alpha_l + \alpha_s = \alpha$, and we refer to it as the *auto-test.* The choice of α_l and α_s should be based on prior knowledge regarding how likely the jump is small. We illustrate how to detect changes in model parameters with *auto-test* in Fig. 2.3. Statistical properties of the auto-test and choices of the thresholds can be found in Liu et al. (2021b).

Differentiable programming. A naïve implementation of the auto-test statistic involves materializing and inverting the Hessian matrix $H_n(\theta_n;\tau) \in \mathbb{R}^{d\times d}$ with $O(nd^2 + d^3)$ time and $O(d^2)$ space. This approach does not scale to modern applications in deep learning with dense Hessians and large n, d. Instead, we rely on iterative algorithms to approximately minimize the quadratic

$$f_n(u) := \frac{1}{2} u^\top H_n(\theta_n;\tau) u + u^\top S_{\tau+1:n}(\theta_n).$$

Indeed, the unique minimizer $u_\star$ of f_n satisfies $0 = \nabla f_n(u_\star) = H_n(\theta_n;\tau)u_\star + S_{\tau+1:n}(\theta_n)$ and thus $u_\star = H_n(\theta_n;\tau)^{-1} S_{\tau+1:n}(\theta_n)$ as desired. Modern automatic differentiation software supports the efficient computation of the Hessian-vector product $u \mapsto H_n(\theta_n;\tau)u$ without materializing the Hessian. Several iterative algorithms that can achieve this are the conjugate gradient method, stochastic gradient descent, the LiSSA algorithm (Agarwal et al., 2017), and a low-rank approximation (Schioppa et al., 2022).

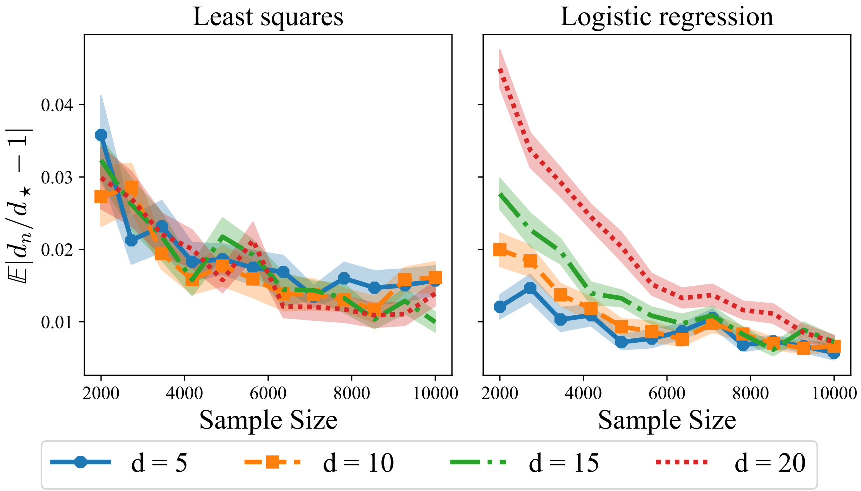

Figure 2.4: Absolute error of the empirical effective dimension. **(Left)**: least squares; **(Right)**: logistic regression.

2.6 Simulation Study

We run simulation study to illustrate our theoretical results. We start by demonstrating the consistency of d_n and the shape of the Wald confidence set defined in Corollary 2.9, i.e.,

$$\mathcal{C}_n(\delta) = \left\{ \theta \in \Theta : \|\theta - \theta_n\|^2_{H_n(\theta_n)} \leq C_{K_1,\nu} \frac{d_n}{n} \log\left(e/\delta\right) \right\}. \tag{2.13}$$

Note that the oracle Wald confidence set should be constructed from $\|\theta_n - \theta_\star\|_{H_\star}$ and $d_\star$; however, Corollary 2.9 suggests we can replace $H_\star$ and $d_\star$ by $H_n(\theta_n)$ and d_n without losing too much. To empirically verify our theoretical results, we calibrate the Wald confidence set in Corollary 2.9 with the threshold from the oracle Wald confidence set and compare its coverage with the one calibrated by the multiplier bootstrap—a popular resampling-based approach for calibration. In all the experiments, we generate n i.i.d. pairs by sampling X and then sampling $Y \mid X$. The code to reproduce the experiments is available online (confset, 2022).

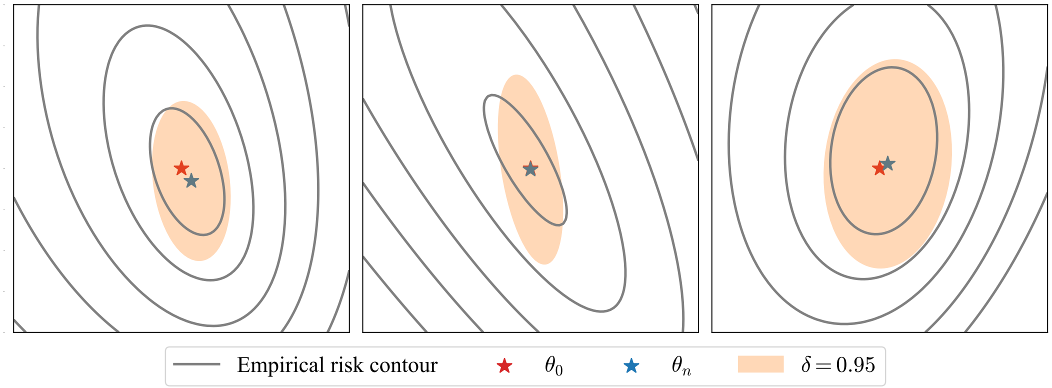

Figure 2.5: Confidence set in (2.13) under a logistic regression model. **Left:** $\Sigma = (2, 0; 0, 1)$; **Middle:** $\Sigma = (2, 1; 1, 1)$; **Right:** $\Sigma = (2, -1; -1, 1)$.

2.6.1 Numerical illustrations

Approximation of the effective dimension. By Proposition 2.8, we know that d_n is a consistent estimator of $d_\star$. We verify it with simulations. We consider two models. For least squares, the data are generated from $X \sim \mathcal{N}(0, I_d)$ and $Y|X \sim \mathcal{N}(\mathbf{1}^\top X, 1)$. For logistic regression, the data are generated from $X \sim \mathcal{N}(0, I_d)$ and $Y \mid X \sim p(Y \mid X) = \sigma(Y\,\mathbf{1}^\top X)$ for $Y \in \{-1, 1\}$ where $\sigma(u) := (1+e^{-u})^{-1}$. We then estimate $d_\star = d$ (since the model is well-specified) by d_n and quantify its estimation error by $\mathbb{E}\,|d_n/d_\star - 1|$. We vary $n \in [2000, 10000]$ and $d \in \{5, 10, 15, 20\}$, and give the plots in Figure 2.4. For a fixed d, the absolute error decays to zero as the sample size increases as predicted by Proposition 2.8. For a fixed n, the absolute error raises as the dimension becomes larger in logistic regression, but it remains similar in least squares.

Shape of the Wald confidence set. Note that the Wald confidence set in (2.13) is an ellipsoid whose shape is determined by the empirical Hessian $H_n(\theta_n)$ and thus can effectively handles the local curvature of the empirical risk. We illustrate this feature on a logistic regression example. We generate data from $X \sim \mathcal{N}(0, \Sigma)$ with different Σ's and $Y \mid X \sim$

Table 2.3: Coverage of the oracle and bootstrap confidence sets.

Model	**Confidence set**	$\delta = 0.95$	$\delta = 0.9$	$\delta = 0.85$	$\delta = 0.8$	$\delta = 0.75$
Well-specified LS	Oracle	0.957	0.908	0.868	0.792	0.770
	Bootstrap	0.947	0.908	0.855	0.791	0.735
Misspecified LS	Oracle	0.972	0.916	0.882	0.841	0.764
	Bootstrap	0.968	0.924	0.865	0.779	0.727
Well-specified LR	Oracle	0.961	0.915	0.868	0.809	0.776
	Bootstrap	0.938	0.885	0.826	0.781	0.706

$p(Y \mid X) = \sigma(Y\theta_0^\top X)$ for $Y \in \{-1, 1\}$ where $\theta_0 = (-1, 2)^\top$. We then construct the confidence set with $d_\star = d$. As shown in Figure 2.5, the shape of the confidence set varies with Σ and captures the curvature of the empirical risk at θ_0.

2.6.2 Calibration

We investigate two calibration schemes. Inspired by the setting in Chen and Zhou (2020, Sec. 5.1), we generate $n = 100$ i.i.d. observations from three models with true parameter θ_0 whose elements are equally spaced between $[0, 1]$—1) *well-specified least squares* with $X \sim \mathcal{N}(0, I_d)$ and $Y \mid X \sim \mathcal{N}(\theta_0^\top X, 1)$, 2) *misspecified least squares* with $X \sim \mathcal{N}(0, I_d)$ and $Y \mid X \sim \theta_0^\top X + t_{3.5}$, and 3) *well-specified logistic regression* with $X \sim \mathcal{N}(0, I_d)$ and $Y \mid X \sim p(Y \mid X) = \sigma(Y\theta_0^\top X)$ for $Y \in \{-1, 1\}$. For each $\delta \in \{0.95, 0.9, 0.85, 0.8, 0.75\}$, we construct a confidence set using either *oracle calibration* or *multiplier bootstrap*. We repeat the whole process for 1000 times and report the coverage of each confidence set in Table 2.3.

Oracle calibration. According to Theorem 2.6, if we have access to $H_\star$ and $d_\star$, we can construct a confidence set of the form $\mathcal{C}_\star(\delta) := \{\theta : \|\theta_n - \theta\|_{H_\star} \le d_\star/n + c_n(\delta)\}$. Now Corollary 2.9 suggests that $H_\star$ and $d_\star$ can be accurately estimated by $H_n(\theta_n)$ and d_n, respectively, leading the confidence set $\mathcal{C}_n(\delta) := \{\theta : \|\theta_n - \theta\|_{H_n(\theta_n)} \le d_n/n + c_n(\delta)\}$. To calibrate $\mathcal{C}_n(\delta)$, we use the data generating distribution to estimate $c_n(\delta)$ so that $\mathbb{P}(\theta_\star \in \mathcal{C}_\star(\delta)) \approx 1 - \delta$, and then plug it into $\mathcal{C}_n(\delta)$. We call it the *oracle confidence set*. As shown in Table 2.3, its coverage is very close to the prescribed confidence level in the well-specified case and it tends to be more conservative in the misspecified case.

Multiplier bootstrap. To further evaluate the oracle calibration, we compare its coverage with the one calibrated by the multiplier bootstrap (e.g., Chen and Zhou, 2020)—a popular resampling-based calibration approach that is widely used in practice. We construct a *bootstrap confidence set* with $B = 2000$ bootstrap samples in the following steps. For each $b \in \{1, \dots, B\}$, we 1) generate weights $\{W_i^b\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1)$, 2) compute the bootstrap estimator

$$\theta_n^b = \arg\min_\theta \left[L_n^b(\theta) := \frac{1}{n} \sum_{i=1}^n W_i^b \ell(\theta; Z_i) \right],$$

3) compute the bootstrap statistic $T_{\text{Wald}}^b := \left\|\theta_n^b - \theta_n\right\|_{H_n^b(\theta_n^b)}^2$ where $H_n^b(\theta) := \nabla_\theta^2 L_n^b(\theta)$. Finally, we compare $\|\theta_n - \theta_0\|_{H_n(\theta_n)}^2$ with the upper δ quantile of $\{T_{\text{Wald}}^b\}_{b=1}^B$ to decide if the confidence set covers the true parameter. It is clear that the bootstrap confidence set performs similarly as the oracle confidence set in least squares, but it is more liberal in logistic regression.

Chapter 3

INFORMATION DIVERGENCES FOR COMPARING DISTRIBUTIONS

3.1 Introduction

Deep generative models have recently taken a giant leap forward in their ability to model complex, high-dimensional distributions. Recent advances are able to produce incredibly detailed and realistic images (Kingma and Dhariwal, 2018; Razavi et al., 2019; Karras et al., 2020), strikingly consistent and coherent text (Radford et al., 2019; Zellers et al., 2019; Brown et al., 2020), and music of near-human quality (Dhariwal et al., 2020). The advances in these models, particularly in the image domain, have been spurred by the development of quantitative evaluation tools which enable a large-scale comparison of models, as well as diagnosing of where and why a generative model fails (Salimans et al., 2016; Lopez-Paz and Oquab, 2017; Heusel et al., 2017; Binkowski et al., 2018; Sajjadi et al., 2018; Karras et al., 2019).

Divergence frontiers were recently proposed by Djolonga et al. (2020) to quantify the trade-offs between quality and diversity in generative modeling with modern deep neural networks (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Simon et al., 2019; Naeem et al., 2020; Pillutla et al., 2021). In particular, a good generative model must not only produce high-quality samples that are likely under the target distribution but also cover the target distribution with diverse samples.

While the framework of divergence frontiers is mathematically elegant and empirically successful (Kynkäänniemi et al., 2019; Pillutla et al., 2021), its statistical properties are not well understood. The recipe taken by practitioners to estimate divergence frontiers from data for large generative models usually involves two approximations: (a) joint quantization of the model distribution and the target distribution into discrete distributions with quantization level k, and (b) statistical estimation of the divergence frontiers based on the empirical estimators of the quantized distributions.

Djolonga et al. (2020) argue that the quantization in the first approximation often introduces a positive bias, making the distributions appear closer than they really are; while a small sample size can result in a pessimistic estimate of the divergence frontiers. The latter effect is due to the missing mass of the samples, causing the two distributions to appear farther than they really are because the samples do not cover some parts of the distributions especially when the support size is large. The first consideration favors a large k, while the second favors a small k.

We are interested in answering the following questions in this chapter: (a) Given two distributions, how many samples are needed to achieve a desired estimation accuracy, or in other words, what is the sample complexity of the estimation procedure? (b) Given a sample size budget, how to choose the quantization level to balance the errors induced by the two approximations? (c) Can we have estimators better than the naïve empirical estimator?

The remainder of this section is organized as follows. In Section 3.2 we recall the problem of KL estimation and the missing mass problem. We review in Section 3.3 the definition of divergence frontiers and propose a novel statistical summary. We establish in Section 3.4 non-asymptotic bounds for the estimation of divergence frontiers as well as frontier integrals which characterizes the sample complexity. We discuss the choice of the quantization level by balancing the errors induced by the two approximations. We show in Section 3.5 how smoothed distribution estimators, such as the add-constant estimator and the Good-Turing estimator, improve the estimation accuracy. We also generalize our results to a large class of f-divergences satisfying some regularity assumptions. Finally, we demonstrate in Section 3.6, through simulations on synthetic data as well as generative adversarial networks on images and transformer-based language models on text, that our bounds exhibit the correct dependence of the estimation error on the sample size n and the support size k.

3.2 Preliminaries

3.2.1 Estimation of KL divergence in the large-alphabet regime

A closely related problem is the estimation of KL divergence between two discrete distributions (Cai et al., 2006; Zhang and Grabchak, 2014; Bu et al., 2018; Han et al., 2020). To be more concrete, let P and Q be two distributions supported on a common alphabet $[k] := \{1, \ldots, k\}$ such that $P \ll Q$. We denote by $\mathcal{D}([k])$ the collection of such pairs of distributions. Given two independent i.i.d. samples $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$, a natural way to estimate $\mathrm{KL}(P\|Q)$ is based on the empirical distributions P_n and Q_m, where $P_n(a) = N_a/n := |\{i : X_i = a\}|/n$ and $Q_m(a) = M_a/m := |\{j : Y_j = a\}|/m$ for each $a \in [k]$. However, the minimax quadratic risk of this type of estimators over the set $\mathcal{D}([k])$ is infinite for all $k \geq 2$ (Bu et al., 2018, Theorem 1). The main challenge is that when Q has a long tail, its tail masses contribute significantly to the KL divergence but requires a large amount of observations to estimate accurately. In other words, it is highly likely that some masses are missing in the sample. This phenomenon is especially prominent in the large-alphabet regime, i.e., $k \to \infty$.

This challenge can be addressed in two steps. First, we restrict the class of distributions so that the mass ratio between P and Q is bounded, i.e., we consider

$$\mathcal{D}([k], C_k) := \left\{(P, Q) : |P| = |Q| = k, \frac{P(a)}{Q(a)} \leq C_k, \forall a \in [k]\right\}.$$

Second, we smooth the empirical distribution Q_n so that there is no missing mass. A popular choice is the technique called *add-constant smoothing* (Krichevsky and Trofimov, 1981; Braess and Sauer, 2004). It adds a small constant to the counts of the alphabet and normalize these pseudo counts to form a distribution. Precisely, the add-b estimator of Q is defined as

$$Q_{m,b}(a) := (M_a + b)/(m + kb), \quad \text{for each } a \in [k]. \tag{3.1}$$

Now, the so-called *augmented plug-in estimator* $\mathrm{KL}(P_n\|Q_{m,b})$ can achieve the following worst-case quadratic risk. In fact, as shown by Bu et al. (2018, Theorem 3), this risk is minimax optimal up to a logarithmic factor $\log k$.

Theorem 3.1 (Theorem 2 in Bu et al. (2018))**.** *For any $k \geq 1$, $n \geq k$ and $m \geq 10kC_k$, we have*

$$\sup_{(P,Q)\in\mathcal{D}([k],C_k)} \mathbb{E}\left[(\mathrm{KL}(P_n\|Q_{m,b}) - \mathrm{KL}(P\|Q))^2\right] \asymp \left(\frac{k}{n} + \frac{kC_k}{m}\right)^2 + \frac{\log^2 C_k}{n} + \frac{C_k}{m}.$$

3.2.2 The Good-Turing estimator and the missing mass

The missing mass problem is of interest in many applications involving sampling from a large alphabet, e.g., species in a population and words in a language corpus. The study of this problem can be dated back to Turing's work on solving the Enigma cypher during World War II, which later developed by Good (1953) in the context of estimating the population frequency of species. To be more precise, consider an i.i.d. sample $\{X_i\}_{i=1}^n$ drawn from a large alphabet $[k]$. Let N_a be the number of times symbol $a \in [k]$ appears in the sample and $\varphi_r := \sum_{a=1}^k \mathbb{1}\{N_a = r\}$ be the number of distinct symbols which appear exactly $r \in \mathbb{N}$ times in the sample. It is clear that

$$\sum_{r=1}^{\infty} r\varphi_r = n.$$

Let $P_r := \sum_{a=1}^k P(a)\mathbb{1}\{N_a = r\}$ be the masses of the symbols which appear exactly r times in the sample. Note that it is a random variable since it depends on the sample. In particular, the quantity P_0 is called the *missing mass*. The basic Good-Turing estimator estimates P_r by $(r+1)\varphi_{r+1}/n$.

The Good-Turing estimator induces an estimator of the distribution P. Take a symbol $a \in [k]$. By definition, it appears in the sample exactly N_a times. Intuitively, it makes sense to assume that the symbols appearing exactly N_a (there are $\varphi_{N_a} \geq 1$ such symbols) times in the sample share the same mass, i.e., $P(a) \approx P_{N_a}/\varphi_{N_a}$. Hence, the Good-Turing distribution estimator is defined as

$$P_{n,\mathrm{GT}}(a) = \frac{(N_a+1)\varphi_{N_a+1}}{n\varphi_{N_a}}, \quad \text{for each } a \in [k].$$

This estimator has been widely studied in language modeling (Katz, 1987; Church and Gale, 1991; Chen and Goodman, 1999) and in theory (McAllester and Schapire, 2000; Orlitsky et al., 2003; Orlitsky and Suresh, 2015). An inspiring result coming from this line of work is that the missing mass itself concentrates around its expectation (McAllester and Ortiz, 2003) which decays as $O(k/n)$ (Berend and Kontorovich, 2012).

3.3 Divergence Frontiers and Frontier Integrals

We introduce the notion of divergence frontiers for comparing two distributions in Section 3.3.1. We also propose a statistical summary of the divergence frontiers in Section 3.3.2.

3.3.1 Evaluating generative models via divergence frontiers

Let $\mathcal{X}$ be a measurable space in which the data live. Consider a generative model $Q \in \mathcal{M}_1(\mathcal{X})$ which attempts to model the target distribution $P \in \mathcal{M}_1(\mathcal{X})$. It has been argued in Sajjadi et al. (2018) and Kynkäänniemi et al. (2019) that one must consider two types of costs to evaluate Q with respect to P: (a) a type I cost (loss in precision), which is the mass of Q that has low or zero probability mass under P, and (b) a type II cost (loss in recall), which is the mass of P that Q does not adequately capture.

Suppose P and Q are uniform distributions on their supports, and R is uniform on the union of their supports. Then, the type I cost is the mass of $\mathrm{Supp}(Q) \setminus \mathrm{Supp}(P)$, or equivalently, the mass of $\mathrm{Supp}(R) \setminus \mathrm{Supp}(P)$. We measure it using the surrogate $\mathrm{KL}(Q\|R)$, which is large if there exists $x \in \mathcal{X}$ such that $Q(x)$ is large but $R(x)$ is small. Likewise, the type II cost is measured by $\mathrm{KL}(P\|R)$. When P and Q are not constrained to be uniform, it is not clear what the measure R should be. Djolonga et al. (2020) propose to vary R over all possible probability measures and consider the Pareto frontier of the multi-objective optimization $\min_R \big(\mathrm{KL}(P\|R), \mathrm{KL}(Q\|R)\big)$. This leads to a curve called the *divergence frontier*, and is reminiscent of the precision-recall curve in binary classification. See Cortes and Mohri (2005); Clémençon and Vayatis (2009); Flach (2012) and references therein on trade-off curves in machine learning.

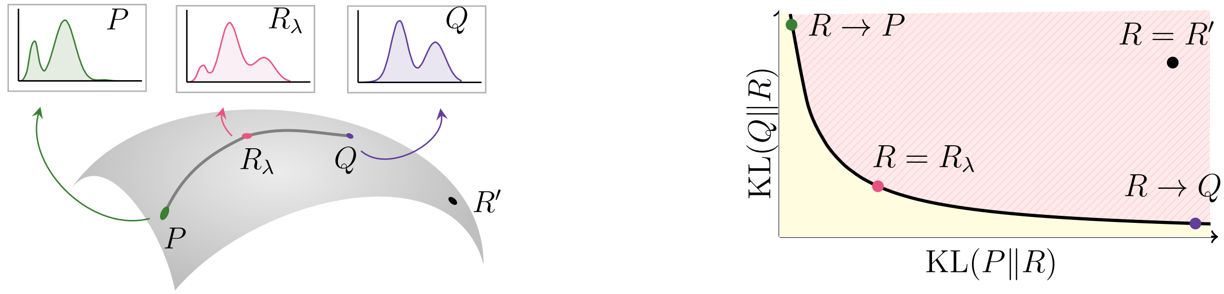

Figure 3.1: **Left**: Comparing two distributions P and Q. Here, $R_\lambda = \lambda P + (1-\lambda)Q$ is the interpolation between P and Q for $\lambda \in (0,1)$ and R' denotes some arbitrary distribution. **Right**: The corresponding divergence frontier (black curve) between P and Q. The interpolations R_λ for $\lambda \in (0,1)$ make up the frontier, while all other distributions such as R' must lie above the frontier.

Formally, the divergence frontier of probability measures P and Q is defined as

$$\begin{aligned}\mathcal{F}(P,Q) := \Big\{ \big(\mathrm{KL}(P\|R), \mathrm{KL}(Q\|R)\big) \,:\, \nexists R' \in \mathcal{M}_1(\mathcal{X}) \text{ such that} \\ \mathrm{KL}(P\|R') < \mathrm{KL}(P\|R) \text{ and } \mathrm{KL}(Q\|R') < \mathrm{KL}(Q\|R) \Big\}\,.\end{aligned} \tag{3.2}$$

It admits the closed-form expression (Djolonga et al., 2020, Propositions 1 and 2)

$$\mathcal{F}(P,Q) = \Big\{ \big(\mathrm{KL}(P\|R_\lambda), \mathrm{KL}(Q\|R_\lambda)\big) \,:\, \lambda \in (0,1) \Big\}\,,$$

where

$$R_\lambda := \operatorname*{arg\,min}_{R \in \mathcal{M}_1(\mathcal{X})} \big\{ \lambda\,\mathrm{KL}(P\|R) + (1-\lambda)\,\mathrm{KL}(Q\|R) \big\} = \lambda P + (1-\lambda) Q. \tag{3.3}$$

Intuitively, each point on the divergence frontier compares the two individual distributions against a linear mixture of the two. By sweeping through mixtures, the curve interpolates between measurements of the two types of costs. See Figure 3.1 for an illustration.

In practical applications, P is usually a complex, high-dimensional distribution which could either be discrete, as in natural language processing, or continuous, as in computer vision. Likewise, Q is often a deep generative model such as GPT-3 (Brown et al., 2020) for text and variants of GANs (Goodfellow et al., 2014) for images. It is infeasible to compute the divergence frontier $\mathcal{F}(P, Q)$ directly because we only have samples from P and the integrals or sums over Q are intractable. Therefore, the recipe used by practitioners (Sajjadi et al., 2018; Djolonga et al., 2020; Pillutla et al., 2021) has been to (a) jointly quantize P and Q over a partition $\mathcal{S} = \{S_t\}_{t=1}^k$ of $\mathcal{X}$ to obtain discrete distributions $P_{\mathcal{S}} = (P(S_t))_{t=1}^k$ and $Q_{\mathcal{S}} = (Q(S_t))_{t=1}^k$, (b) estimate the quantized distributions from samples to get $\hat{P}_{\mathcal{S}}$ and $\hat{Q}_{\mathcal{S}}$, and (c) compute $\mathcal{F}(\hat{P}_{\mathcal{S}}, \hat{Q}_{\mathcal{S}})$. In practice, the best quantization schemes are data-dependent transformations such as k-means clustering or lattice-type quantization of dense representations of images or text (Sablayrolles et al., 2019).

3.3.2 Statistical summary of divergence frontiers

In the minimax theory of hypothesis testing, where the goal is also to study two types of errors (yet different from the ones considered here), it is common to theoretically analyze their linear combination; see, e.g., Ingster and Suslina (2003, Sec. 1.2) and Cai et al. (2011, Thm. 7). In the same spirit, we consider a linear combination of the two costs, quantified by the KL divergences,

$$\mathcal{L}_\lambda(P, Q) := \lambda \, \mathrm{KL}(P \| R_\lambda) + (1 - \lambda) \, \mathrm{KL}(Q \| R_\lambda). \tag{3.4}$$

Recall from (3.3) that R_λ is exactly the minimizer of the linearized objective $\lambda \, \mathrm{KL}(P \| R) + (1 - \lambda) \, \mathrm{KL}(Q \| R)$. This linear combination $\mathcal{L}_\lambda$ is also known as the λ-skew Jensen-Shannon Divergence (Nielsen and Bhatia, 2013).

The linearized cost $\mathcal{L}_\lambda$ depends on the choice of the interpolation parameter λ. To remove this dependency, we define a novel statistical summary, called the *frontier integral*, as

$$\mathrm{FI}(P, Q) := 2 \int_0^1 \mathcal{L}_\lambda(P, Q) \, \mathrm{d}\lambda \, . \tag{3.5}$$

We can interpret the frontier integral as the average linearized cost over $\lambda \in (0, 1)$. As shown in Section 3.6, it can also be used to evaluate generative models which is more convenient than the divergence frontier when comparing a large number of models.

The frontier integral is always bounded in $[0, 1]$ thanks to the factor 2 in front of the integral. Moreover, it is a symmetric f-divergence. We summarize these properties below.

Property 3.1. *Let P and Q be dominated by some probability measure μ with densities p and q, respectively. Then,*

$$\mathrm{FI}(P, Q) = \int_{\mathcal{X}} \mathbb{1}\{p(x) \neq q(x)\} \left(\frac{p(x) + q(x)}{2} - \frac{p(x)q(x)}{p(x) - q(x)} \log \frac{p(x)}{q(x)} \right) \mathrm{d}\mu(x) \, , \tag{3.6}$$

with the convention $0 \log 0 = 0$. *Moreover,* FI *is an f-divergence generated by the convex function*

$$f_{\mathrm{FI}}(t) = \frac{t+1}{2} - \frac{t}{t-1} \log t \, ,$$

with the understanding that $f_{\mathrm{FI}}(1) = \lim_{t \to 1} f_{\mathrm{FI}}(t) = 0$.

Proof. Let $\bar{\lambda} := 1 - \lambda$. By Tonelli's theorem, we have $\mathrm{FI}(P, Q) = 2 \int_{\mathcal{X}} h(p(x), q(x)) \mathrm{d}\mu(x)$, where

$$h(p, q) = \int_0^1 \left(\lambda p \log p + \bar{\lambda} q \log q - (\lambda p + \bar{\lambda} q) \log(\lambda p + \bar{\lambda} q) \right) \mathrm{d}\lambda.$$

When $p = q$, the integrand is 0 and thus $h(p, q) = 0$. If $q = 0$, then the second term inside the integral is 0, while the rest of the terms is

$$\int_0^1 \lambda p \log \frac{1}{\lambda} \mathrm{d}\lambda = \frac{p}{4} \, .$$

Finally, when $p \neq q$ are both non-zero, we evaluate the integral to get

$$h(p, q) = \frac{p}{2} \log p + \frac{q}{2} \log q - \frac{2p^2 \log p - p^2 - 2q^2 \log q + q^2}{4(p - q)} \, ,$$

Rearranging the expression completes the proof. □

Property 3.2. *The frontier integral satisfies the following properties:*

(a) $\mathrm{FI}(P, Q) = \mathrm{FI}(Q, P)$.

(b) $0 \leq \mathrm{FI}(P, Q) \leq 1$ *with* $\mathrm{FI}(P, Q) = 0$ *if and only if* $P = Q$.

Proof. The first part follows from the closed form expression in Property 3.1. For the second part, we get the upper bound as

$$\mathrm{FI}(P, Q) \le \int_{\mathcal{X}} \frac{p(x) + q(x)}{2} \mathrm{d}\mu(x) = 1 \,.$$

We have $\mathrm{FI}(P, Q) \ge 0$ with $\mathrm{FI}(P, P) = 0$ since FI is an f-divergence. Further, since f_{FI} is strictly convex at 1, we get that $\mathrm{FI}(P, Q) = 0$ only if $P = Q$. □

Although the frontier integral in (3.5) involves an integral w.r.t. λ, thanks to Property 3.1, when we have access to P and Q, computing $\mathrm{FI}(P, Q)$ is computationally no worse than computing $\big(\mathrm{KL}(P \| R_\lambda), \mathrm{KL}(Q \| R_\lambda)\big)$ which is a point on the divergence frontier. In practice, when we have samples from P and Q, it can be estimated using the same recipe as the divergence frontier, i.e., $\mathrm{FI}(\hat{P}_{\mathcal{S}}, \hat{Q}_{\mathcal{S}})$.

3.4 Non-Asymptotic Analysis

This section is devoted to deriving the rate of convergence for the overall error in estimating the frontier integral. We decompose the overall estimation error into two components: the statistical error of estimating the quantized distribution and the quantization error. We control the statistical error in Section 3.4.1. The strategy is to use a different treatment for the masses that appear in the sample and the ones that never appear (i.e., the missing mass). We obtain a high probability bound as well as a bound for its expectation, leading to a rate of convergence in both the small-alphabet and large-alphabet regimes. These results carry over to the divergence frontiers as well. The quantization error is discussed in Section 3.4.2. We construct a distribution-dependent quantization scheme whose error is at most $O(k^{-1})$ where k is the quantization level. In Section 3.4.3 we combine these two bounds to obtain the sample complexity of estimating frontier integrals, shedding light on the optimal choice of the quantization level.

For $P, Q \in \mathcal{M}_1(\mathcal{X})$, let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two i.i.d. samples from P and Q, respectively, and denote by P_n and Q_n the respective empirical measures. Note that our results hold for two samples with different sizes, and the same size is assumed here for simplicity of the presentation. We use $\lesssim$ and $\gtrsim$ to represent $\leq$ and $\geq$ omitting an absolute constant factor. The precise statements and proofs can be found in Appendix B.

3.4.1 Statistical error

We focus on in this section distributions P and Q supported on a countable alphabet. Here P and Q should be understood as the quantized distributions in the estimation pipeline of frontier integrals. We are interested in deriving a non-asymptotic bound for the absolute error of the empirical estimator, i.e., $\left|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)\right|$. A natural strategy is to exploit the smoothness properties of FI, giving a naïve upper bound $O(L\sqrt{k/n})$ where $L = \log 1/p_*$ with $p_* = \min_{a \in \mathrm{Supp}(P)} P(a)$ reflecting the smoothness of FI. The dependency on p_* requires P to have a finite support and a short tail. However, in many real-world applications, the distributions can either be supported on a countable set or have long tails (Chen and Goodman, 1999; Wang et al., 2017). By considering the *missing mass* in the sample, we are able to obtain a high probability bound that is independent of p_*. Define

$$\alpha_n(P) := \sum_{a \in \mathcal{X}} \sqrt{\frac{P(a)}{n}} \quad \text{and} \quad \beta_n(P) := \mathbb{E}\left[\sum_{a : P_n(a) = 0} P(a) \max\left\{1, \log \frac{1}{P(a)}\right\}\right].$$

Theorem 3.2. *Let* $k = \max\{|Supp(P)|, |Supp(Q)|\} \in \mathbb{N} \cup \{\infty\}$*. For any* $\delta \in (0,1)$*, it holds that, with probability at least* $1 - \delta$*,*

$$|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)| \lesssim \left(\sqrt{\frac{\log(1/\delta)}{n}} + \alpha_n(P) + \alpha_n(Q)\right) \log n + \beta_n(P) + \beta_n(Q), \quad (3.7)$$

Furthermore, if the support size $k < \infty$*, then* $\alpha_n(P) \leq \sqrt{k/n}$ *and* $\beta_n(P) \leq k \log n / n$*. In particular, with probability at least* $1 - \delta$*,*

$$|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)| \lesssim \left[\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{k}{n}} + \frac{k}{n}\right] \log n. \quad (3.8)$$

There are several merits to Theorem 3.2. First, (3.7) holds for any distributions with a countable support. Second, it does not depend on p_* and is adapted to the tail behavior of P and Q. For instance, if P is defined as $P(a) \propto a^{-2}$ for $a \in [k]$, then $\alpha_n(P) \propto (\log k)/\sqrt{n}$, which is much better than $\sqrt{k/n}$ in (3.8) in terms of the dependency on k. This phenomenon is also demonstrated empirically in Section 3.6. Third, it captures a parametric rate of convergence, i.e., $O(n^{-1/2})$, up to a logarithmic factor. In fact, as discussed in Section 3.2.1, this rate is not improvable in a related problem of estimating $\mathrm{KL}(P\|Q)$, even with the assumption that P/Q is bounded. The bound in (3.8) is a distribution-free bound, assuming k is finite. Note that it also gives an upper bound on the sample complexity by setting the right hand side of (3.8) to be ϵ and solving for n, this is roughly $O((\sqrt{\log 1/\delta} + \sqrt{k})^2/\epsilon^2)$.

The proof of Theorem 3.2 relies on two new results: (a) a concentration bound around $\mathbb{E}[\mathrm{FI}(P_n, Q_n)]$, which can be obtained by McDiarmid's inequality, and (b) an upper bound for the expected statistical error, i.e.,

$$\begin{aligned}\mathbb{E}\,|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)| &\lesssim [\alpha_n(P) + \alpha_n(Q)] \log n + \beta_n(P) + \beta_n(Q) \\ &\lesssim (\sqrt{k/n} + k/n) \log n, \quad \text{if } k < \infty\,.\end{aligned} \tag{3.9}$$

The concentration bound gives the term $\sqrt{n^{-1}\log{(1/\delta)}}$. The expected statistical error bound is achieved by splitting the masses of P and Q into two parts: one that appears in the sample and one that never appears. The first part can be controlled by a Lipschitz-like property of the frontier integral, leading to the term $\alpha_n(P)+\alpha_n(Q)$, and the second part, $\beta_n(P)+\beta_n(Q)$, falls into the missing mass framework. In addition, the rate k/n for β_n shown here matches the rate for the missing mass.

While Theorem 3.2 establishes the consistency of the frontier integral, it is also of great interest to know whether the divergence frontier itself can be consistently estimated. In fact, similar bounds hold for the worst-case error of $\mathcal{F}(P_n, Q_n)$.

Corollary 3.3. *Under the same assumptions as in Theorem 3.2, for any $\lambda_0 \in (0,1)$, the bounds in (3.7) and (3.8) hold for*

$$\sup_{\lambda \in [\lambda_0, 1-\lambda_0]} \left\| \big(\mathrm{KL}(P_n \| R_{\lambda,n}), \mathrm{KL}(Q_n \| R_{\lambda,n})\big) - \big(\mathrm{KL}(P\|R_\lambda), \mathrm{KL}(Q\|R_\lambda)\big) \right\|_1,$$

where $R_{\lambda,n} := \lambda P_n + (1-\lambda)Q_n$, with an additional factor of $1/\lambda_0$. In particular, if λ_0 is chosen as $\lambda_n = o(1)$ and $\lambda_n = \omega(\sqrt{k/n}\log n)$, then the expected worst-case error above converges to zero at rate $O(\lambda_n^{-1}\sqrt{k/n}\log n)$.

The truncation in Corollary 3.3 is necessary without imposing additional assumptions, since $\mathrm{KL}(P \| R_\lambda)$ is close to $\mathrm{KL}(P \| Q)$ for small λ and it is known that the minimax quadratic risk of estimating the KL divergence over all distributions with k bins is always infinity (Bu et al., 2018, Theorem 3).

3.4.2 Quantization error

Recall from Section 3.3 that computing the divergence frontiers in practice usually involves a quantization step. Since every quantization will inherently introduce a positive bias in the estimation procedure, it is desirable to control the error, which we call the quantization error, induced by this step. We show that there exists a quantization scheme with error proportional to the inverse of its level.

We say $\mathcal{S}$ is a partition of $\mathcal{X}$ if $\{S_i\}_{i=1}^k$ are mutually disjoint and $\cup_{i=1}^k S_i = \mathcal{X}$. The quantization of P associated with $\mathcal{S}$ is defined as a distribution $P_\mathcal{S}$ on k bins satisfying

$$P_\mathcal{S}(i) = P(S_i) \quad \text{for each } i \in [k].$$

The quantization error of $\mathcal{S}$ is the difference $|\mathrm{FI}(P_\mathcal{S}, Q_\mathcal{S}) - \mathrm{FI}(P, Q)|$. It can be shown that there exists a distribution-dependent partition whose quantization error is no larger than the inverse of its level.

Proposition 3.4. *For any $k \ge 1$, we have*

$$\sup_{P,Q} \inf_{|\mathcal{S}| \le 2k} |\mathrm{FI}(P, Q) - \mathrm{FI}(P_\mathcal{S}, Q_\mathcal{S})| \le k^{-1}.$$

Moreover, there exists $\mathcal{S}_\star := \mathcal{S}_\star(P, Q)$ with $|\mathcal{S}_\star| = k$ such that

$$|\mathrm{FI}(P, Q) - \mathrm{FI}(P_{\mathcal{S}_\star}, Q_{\mathcal{S}_\star})| \lesssim k^{-1}. \tag{3.10}$$

We call $\mathcal{S}_\star$ the oracle quantization.

The key idea behind the construction of this oracle quantization is to partition $\mathcal{X}$ according to the value of the generator f_{FI} in Property 3.1 at the likelihood ratio $\mathrm{d}P(x)/\mathrm{d}Q(x)$

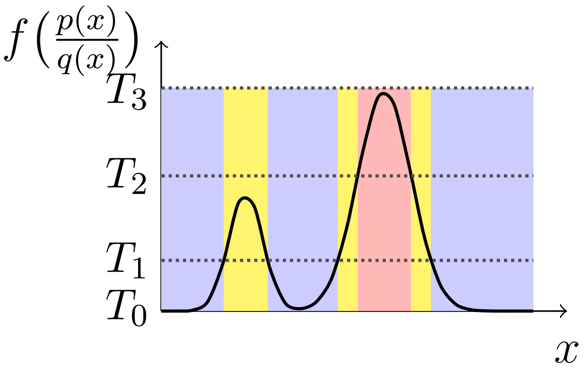

Figure 3.2: Oracle quantization into 3 bins: blue, yellow and red. Bin i is given by the set $\{x : f(\mathrm{d}P(x)/\mathrm{d}Q(x)) \in [T_{i-1}, T_i)\}$.

which is visualized in Figure 3.2. To be more concrete, we focus here the set $\mathcal{X}_1 := \{x \in \mathcal{X} : \mathrm{d}P(x)/\mathrm{d}Q(x) \le 1\}$. Since FI is symmetric, its complement can be handled similarly. Recall from Property 3.1 that $f_{\mathrm{FI}} \in [0, f_{\mathrm{FI}}(0)]$ on $[0, 1]$. Thus, we can select $k - 1$ cutoff points $0 < T_1 < \cdots < T_{k-1} < f_{\mathrm{FI}}(0)$ and partition $\mathcal{X}_1$ into

$$\mathcal{S}_{\star,s} := \left\{ x \in \mathcal{X} : T_{s-1} \le f_{\mathrm{FI}}\left(\frac{\mathrm{d}P(x)}{\mathrm{d}Q(x)}\right) < T_s \right\}, \quad \text{for } s \in [k],$$

where $T_0 := 0$ and $T_{k+1} := f(0)$. For instance, one reasonable choice is to set $T_s = s f_{\mathrm{FI}}(0)/k$. On each $\mathcal{S}_{\star,s}$, the frontier integral $\mathrm{FI}(P, Q)$ can be controlled by

$$\sum_{s=1}^{k} T_{s-1} Q(\mathcal{S}_{\star,s}) \le \int_{\mathcal{S}_{\star,s}} f_{\mathrm{FI}}\left(\frac{\mathrm{d}P(x)}{\mathrm{d}Q(x)}\right) \mathrm{d}Q(x) \le \sum_{s=1}^{k} T_s Q(\mathcal{S}_{\star,s}).$$

On the other hand, since f_{FI} is non-increasing on $(0, 1]$, the term in $\mathrm{FI}(P_{\mathcal{S}_\star}, Q_{\mathcal{S}_\star})$ associated with $\mathcal{S}_{\star,s}$ can be controlled by

$$T_{s-1} Q(\mathcal{S}_{\star,s}) \le f_{\mathrm{FI}}\left(\frac{P(\mathcal{S}_{\star,s})}{Q(\mathcal{S}_{\star,s})}\right) Q(\mathcal{S}_{\star,s}) \le T_s Q(\mathcal{S}_{\star,s}).$$

Hence, the quantization error $|\mathrm{FI}(P_{\mathcal{S}_\star}, Q_{\mathcal{S}_\star}) - \mathrm{FI}(P, Q)|$ is small as long as $T_s - T_{s-1}$ is small. It scales as $O(1/k)$ for the choice of $T_s = s f_{\mathrm{FI}}(0)/k$.

3.4.3 Sample complexity for estimating frontier integrals

Combining the bound in Theorem 3.2 with the bound in Proposition 3.4 leads to the following bound for the total estimation error.

Theorem 3.5. *Assume that $\mathcal{S}_k$ is a partition of $\mathcal{X}$ such that $|\mathcal{S}_k| = k \geq 2$. Then, with probability at least $1-\delta$, the total error $|\mathrm{FI}(P_{\mathcal{S}_k,n}, Q_{\mathcal{S}_k,n}) - \mathrm{FI}(P,Q)|$ is upper bounded by (up to a constant factor)*

$$\left(\sqrt{\frac{\log(1/\delta)}{n}} + \alpha_n(P) + \alpha_n(Q)\right)\log n + \beta_n(P) + \beta_n(Q) + \left|\mathrm{FI}(P,Q) - \mathrm{FI}(P_{\mathcal{S}_k}, Q_{\mathcal{S}_k})\right|. \tag{3.11}$$

Moreover, if the quantization error satisfies the bound in (3.10) and $k < \infty$, we have, with probability at least $1-\delta$,

$$|\mathrm{FI}(P_{\mathcal{S}_k,n}, Q_{\mathcal{S}_k,n}) - \mathrm{FI}(P,Q)| \lesssim \left[\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{k}{n}} + \frac{k}{n}\right]\log n + \frac{1}{k}. \tag{3.12}$$

Based on the bound in (3.12), a good choice of k is $\Theta(n^{1/3})$ which balances between the two types of errors. We illustrate in Section 3.6 that this choice works well in practice. This balancing is enabled by the existence of a good quantizer with a distribution-free bound in (3.10). In practice, this suggests a data-dependent quantizer using nonparametric density estimators. However, directions such as kernel density estimation (Hulle, 1999; Meinicke and Ritter, 2002; Hegde et al., 2004) and nearest-neighbor methods (Alamgir et al., 2014) have not met empirical success, as they suffer from the curse of dimensionality common in nonparametric estimation. In particular, Wang et al. (2005); Silva and Narayanan (2007, 2010) propose quantized divergence estimators but only prove asymptotic consistency, and little progress has been made since then. On the other hand, modern data-dependent quantization techniques based on deep neural networks can successfully estimate properties of the density from high dimensional data (Sablayrolles et al., 2019; Hämäläinen et al., 2020). Theoretical results for those techniques could complement our analysis. We leverage these powerful methods to scale our approach on real data in Section 3.6.

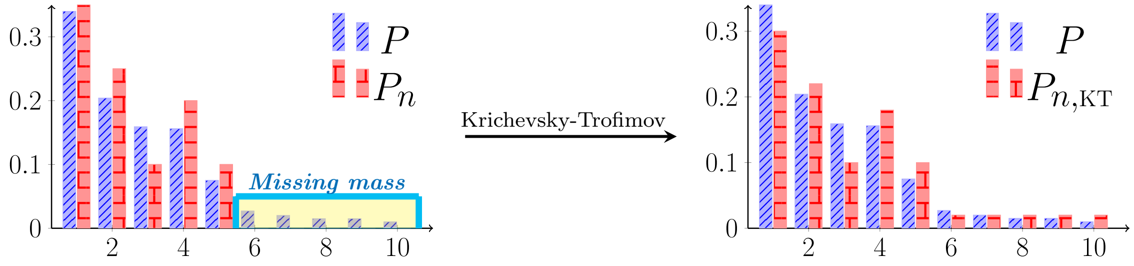

Figure 3.3: The empirical estimator with missing mass and the Krichevsky-Trofimov estimator.

3.5 Towards Better Estimators and General f-Divergences

In Section 3.5.1 we investigate the estimation error of the frontier integral estimated by smoothed distribution estimators. We show that the upper bound for the add-constant estimator improves the one for the empirical estimator, especially in the large-alphabet regime. In Section 3.5.2 we extend our results to general f-divergences satisfying some regularity conditions. The proofs are deferred to Appendix B.

3.5.1 Smoothed distribution estimators

When the support size k is large, the empirical estimator usually performs poorly due to the missing mass phenomenon. To overcome this challenge, practitioners often use more sophisticated distribution estimators such as add-constant estimators (Krichevsky and Trofimov, 1981; Braess and Sauer, 2004) and the Good-Turing estimator (Good, 1953; Orlitsky and Suresh, 2015) as we have seen in Section 3.2. We focus on the add-constant estimators defined in (3.1) and state here its estimation error when it is applied to estimate the frontier integral from data. We investigate and compare the performance of various distribution estimators in Section 3.6.

For notational simplicity, we assume that P and Q are supported on a common finite alphabet with size $k < \infty$. Note that this is true for the quantized distributions $P_{\mathcal{S}}$ and $Q_{\mathcal{S}}$.

Thanks to the smoothing, there is no mass missing in the add-constant estimator. This effect is illustrated for the Krichevsky-Trofimov (add-1/2) estimator in Figure 3.3. As a result, we can directly utilize the smoothness properties of the frontier integral to get the following bound. Note that both $P_{\mathcal{S}}$ and $Q_{\mathcal{S}}$ are estimated by the add-constant estimators. This is different from the augmented plug-in estimator for the KL divergence in Section 3.2.1 since KL is asymmetric but the frontier integral is symmetric.

Theorem 3.6. *Assume that $\mathcal{S}_k$ is a partition of $\mathcal{X}$ such that $|\mathcal{S}_k| = k \in [2, \infty)$. Then, with probability at least $1 - \delta$, the total error $|\mathrm{FI}(P_{\mathcal{S}_k,n,b}, Q_{\mathcal{S}_k,n,b}) - \mathrm{FI}(P, Q)|$ is upper bounded by (up to a constant factor)*

$$\begin{aligned}&\left(\frac{n(\sqrt{\log(1/\delta)/n} + \alpha_n(P) + \alpha_n(Q))}{n + bk} + \gamma_{n,k}(P) + \gamma_{n,k}(Q)\right) \log(n/b + k) \\ &\quad + |\mathrm{FI}(P, Q) - \mathrm{FI}(P_{\mathcal{S}_k}, Q_{\mathcal{S}_k})|,\end{aligned} \tag{3.13}$$

where $\gamma_{n,k}(P) = (n+bk)^{-1}bk \sum_{a\in\mathcal{X}} |P(a) - 1/k|$. Moreover, if the quantization error satisfies the bound in (3.10), it can be further upper bounded by

$$\frac{\sqrt{nk} + bk}{n + bk} \log(n/b + k) + \frac{1}{k}. \tag{3.14}$$

Let us compare the bounds in Theorem 3.6 with the ones in Theorem 3.5. For the distribution-dependent bound, the term $\alpha_n(P)$ in (3.11) is improved by a factor $n/(n + bk)$ in (3.13). The missing mass term $\beta_n(P)$ is replaced by $\gamma_{n,k}(P)$ which is the total variation distance between P and the uniform distribution on $[k]$ with a factor $bk/(n+bk)$. The improvements in both two terms are most significant when k/n is large. As for the distribution-free bound, when k/n is small, the bound in (3.14) scales the same as the one in (3.12); when k/n is large (i.e., bounded away from 0 or diverging), it scales as $O(\log n + \log(k/n) + k^{-1})$ while the one in (3.12) scales as $O(k \log n/n + k^{-1})$. Given the improvement, it would be an interesting venue for future work to consider adaptive estimators in the spirit of Goldenshluger and Lepski (2009).

3.5.2 Generalization to f-divergences

Estimation of the χ^2 divergence is useful for variational inference (Dieng et al., 2017) and GAN training (Mao et al., 2017; Tao et al., 2018). More generally, estimating f-divergences from samples is a fundamental problem in machine learning and statistics (Nguyen et al., 2010; Im et al., 2018; Chen et al., 2018; Rubenstein et al., 2019). The same two-step procedure used to estimate frontier integrals can be applied to estimate general f-divergences as well. Our previous results can be extend to general f-divergences as long as they satisfy some regularity conditions.

We start by recalling f-divergences defined in Definition 1.2. Let $f : (0, \infty) \to \mathbb{R}$ be a nonnegative and convex function with $f(1) = 0$. Let $P, Q \in \mathcal{P}(\mathcal{X})$ be dominated by some measure $\mu \in \mathcal{P}(\mathcal{X})$ with densities p and q, respectively. The f-divergence generated by f is defined as

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x)\,,$$

with the convention that $f(0) = f(0^+)$ and $0f(p/0) = pf^*(0)$, where $f^*(0) = f^*(0^+) \in [0, \infty]$ for $f^*(t) = tf(1/t)$. We call f^* the conjugate generator to f. The function f^* also generates an f-divergence, which is referred to as the *conjugate divergence* to D_f since $D_{f^*}(P\|Q) = D_f(Q\|P)$. In particular, the generator of the frontier integral is f_{FI} in Property 3.1 whose conjugate generator is also f_{FI}.

We then state and discuss the regularity assumptions required to extend the results in Section 3.4 to f-divergences. We use the convention that all higher order derivatives of f and f^* at 0 are defined as the corresponding limits as $x \to 0^+$ (if they exist).

Assumption 3.1. *The generator f is twice continuously differentiable with $f'(1) = 0$. Moreover,*

(A1) *We have $C_0 := f(0) < \infty$ and $C_0^* := f^*(0) < \infty$.*

(A2) *There exist constants $C_1, C_1^* < \infty$ such that for every $t \in (0, 1)$, we have,*

$$|f'(t)| \le C_1 \left(1 \vee \log 1/t\right), \quad \textit{and,} \quad |(f^*)'(t)| \le C_1^* \left(1 \vee \log 1/t\right)\,.$$

(A3) *There exist constants $C_2, C_2^* < \infty$ such that for every $t \in (0, \infty)$, we have,*

$$\frac{t}{2} f''(t) \le C_2 \,, \quad \textit{and,} \quad \frac{t}{2}(f^*)''(t) \le C_2^* \,.$$

Assumption **(A1)** ensures boundedness of the f-divergence. Indeed, $f(0) = \infty$ leads to $D_f(P\|Q) = \infty$ if there exists an atom $a \in \mathcal{X}$ such that $P(a) = 0$ but $Q(a) \neq 0$. This happens, for instance, with the reverse KL divergence whose generator is $f(t) = -\log t + t - 1$. By symmetry, $f^*(0) = \infty$ leads to a case where $D_f(P\|Q) = \infty$ if there exists an atom $a \in \mathcal{X}$ such that $Q(a) = 0$ but $P(a) \neq 0$, as in the (forward) KL divergence.

Since f' is monotonic nondecreasing and $f'(1) = 0$, we have that $f'(0) \le 0$ (with strict inequality if f is strictly convex at 1). In fact, $f'(0) = -\infty$ for most of the commonly used divergences such as the KL divergence, the Jensen-Shannon divergence, and etc. Assumption **(A2)** requires $f'(t)$ to behave as $\log 1/t$ when $t \to 0$. Analogously for $(f^*)'$.

Likewise, we have that $f''(0) = \infty$ and $f''(\infty) = 0$ for most of the commonly used divergences. Assumption **(A3)** imposes additional constraints on the rates of these limits. Namely, f'' should diverge no faster than $1/t$ as $t \to 0$ and f'' should converge to 0 at least as fast as $1/t^2$ as $t \to \infty$. We can summarize the implied asymptotics of f'' as

$$f''(t) = \begin{cases} \Omega(1/t) \,, & \text{if } t \to 0 \,, \\ O(1/t^2) \,, & \text{if } t \to \infty \,. \end{cases}$$

In particular, both the linearized cost $\mathcal{L}_\lambda$ in (3.4) and the frontier integral FI in (3.5) satisfy these assumptions. In Appendix B.2, we consider other f-divergences, e.g., the interpolated KL divergence, and verify or falsify these assumptions.

Proposition 3.7. *The linearized cost $\mathcal{L}_\lambda$ satisfies Assumption 3.1 with*

$$C_0 = \bar{\lambda} \log \frac{1}{\lambda}, \quad C_0^* = \lambda \log \frac{1}{\lambda}, \quad C_1 = \lambda, \quad C_1^* = \bar{\lambda}, \quad C_2 = \frac{\lambda}{2}, \quad C_2^* = \frac{\bar{\lambda}}{2},$$

where $\bar{\lambda} := 1 - \lambda$. Moreover, the frontier integral FI *satisfies Assumption 3.1 with*

$$C_0 = C_0^* = \frac{1}{2}, \quad C_1 = C_1^* = 4, \quad C_2 = C_2^* = \frac{1}{2}.$$

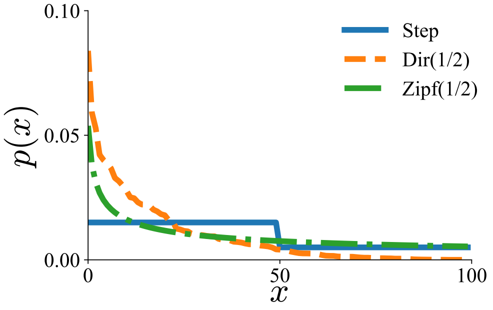

Figure 3.4: Tail decay of the Zipf(1/2), the Step, and the Dir(1/2).

The quantization error bound in Proposition 3.4 holds for all f-divergences which satisfy Assumption **(A1)**. The statistical bounds in Theorem 3.2 and its counterpart for add-constant estimators also hold for f-divergences satisfy Assumption 3.1. In the Appendix, we prove all the results for general f-divergences, recovering all the results in Section 3.4 as special cases due to Proposition 3.7.

3.6 Experiments

We investigate the empirical behavior of the divergence frontier and the frontier integral on both synthetic and real data. Our main findings are: (a) the statistical error bound approximately reveals the rate of convergence of the empirical estimator; (b) the smoothed distribution estimators improve the estimation accuracy; (c) the quantization level suggested by the theory works well empirically. In all the plots, we visualize the average absolute error computed from 100 repetitions with shaded region denoting one standard deviation around the mean. The results for the divergence frontier and the frontier integral are almost identical. We focus on the latter here. Results for the divergence frontier can be founded in Liu et al. (2021a, Appendix G). The code to reproduce the experiments is available online (df, 2021).

3.6.1 Experimental setup

We work with synthetic data in the case when $k = |\mathcal{X}| < \infty$ as well as real image and text data.

Synthetic data. Following the experimental settings in Orlitsky and Suresh (2015), we consider three types of distributions: (a) the Zipf(r) distribution with $r \in \{0, 1, 2\}$ where $P(i) \propto i^{-r}$. Note that Zipf(r) is regularly varying with index $-r$; see, e.g., Shorack (2000, Appendix B); (b) the Step distribution where $P(i) = 1/2$ for the first half bins and $P(i) = 3/2$ for the second half bins; (c) the Dirichlet distribution Dir(α) with $\alpha \in \{\mathbf{1}/2, \mathbf{1}\}$; see Figure 3.4 for an illustration. In total, there are 6 different distributions, giving 21 different pairs of (P, Q). For each pair (P, Q), we generate i.i.d. samples of size n from each of them, and estimate the frontier integral from these samples.

Real data. We consider two domains: images and text. For the image domain, we train a StyleGAN2 (Karras et al., 2020) on the CIFAR-10 dataset (Krizhevsky and Hinton, 2009) using the publicly available code[1] with default hyper-parameters. To evaluate the frontier integrals, we use the test set of 10k images as the target distribution P and we sample 10k images from the generative model as the model distribution Q. For the text domain, we fine-tune a pretrained GPT-2 (Radford et al., 2019) model with 124M parameters (i.e., GPT-2 small) on the Wikitext-103 dataset (Merity et al., 2017). We use the open-source HuggingFace Transformers library (Wolf et al., 2020) for training, and generate 10k 500-token completions using top-p sampling and 100-token prefixes.

We take the following steps to compute the frontier integral. First, we represent each image/text by its features (Heusel et al., 2017; Sajjadi et al., 2018; Kynkäänniemi et al., 2019). Second, we learn a low-dimensional feature embedding which maintains the neighborhood structure of the data while encouraging the features to be uniformly distributed on the unit sphere (Sablayrolles et al., 2019). Third, we quantize these embeddings on a uniform

[1] https://github.com/NVlabs/stylegan2-ada-pytorch.

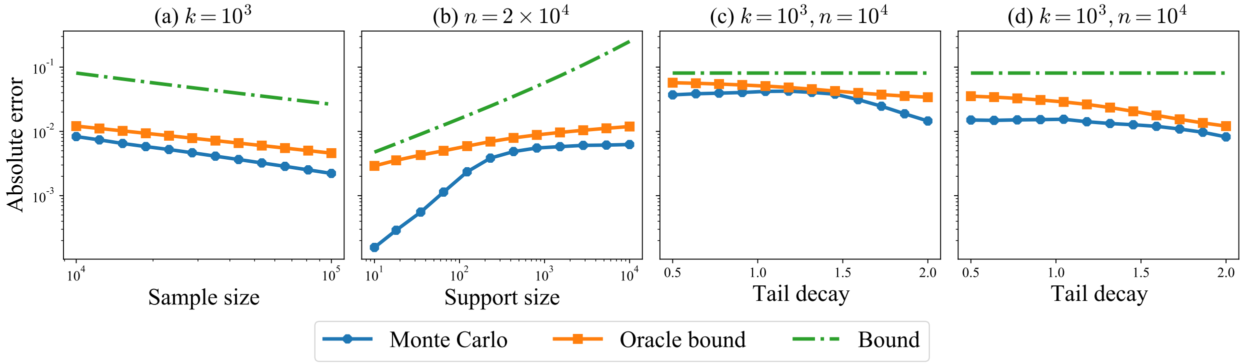

Figure 3.5: Statistical error of the estimated frontier integral on synthetic data. **(a)**: Zipf(2) and Zipf(2) with $k = 10^3$; **(b)**: Zipf(2) and Zipf(2) with $n = 2 \times 10^4$; **(c)**: Dir($\mathbf{1}$) and Zipf(r) with $k = 10^3$ and $n = 10^4$; **(d)**: Zipf(2) and Zipf(r) with $k = 10^3$ and $n = 10^4$. The bounds are scaled by 100.

lattice with k bins. For each support size k, this gives us quantized distributions $P_{\mathcal{S}_k}$ and $Q_{\mathcal{S}_k}$. Finally, we sample n i.i.d. observations from each of these distributions and consider the empirical distributions $P_{\mathcal{S}_k,n}$ and $Q_{\mathcal{S}_k,n}$ as well as the smoothed distribution estimators computed from these samples.

Performance metric. We are interested in the estimation of the frontier integral $\mathrm{FI}(P, Q)$ using estimators $\mathrm{FI}(P_n, Q_n)$ for the empirical estimator as well as the smoothed distribution estimator. We measure the quality of estimation using the absolute error, which is defined as $|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)|$. For the real data, we measure the error of estimating $\mathrm{FI}(P_{\mathcal{S}_k}, Q_{\mathcal{S}_k})$ by $\mathrm{FI}(P_{\mathcal{S}_k,n}, Q_{\mathcal{S}_k,n})$ with $|\mathrm{FI}(P_{\mathcal{S}_k,n}, Q_{\mathcal{S}_k,n}) - \mathrm{FI}(P_{\mathcal{S}_k}, Q_{\mathcal{S}_k})|$.

3.6.2 Tightness of the statistical bound

In order to verify the validity of the theory in practically relevant settings, we investigate the tightness of the statistical error bounds in Theorem 3.2 with respect to the sample size n and the support size k.

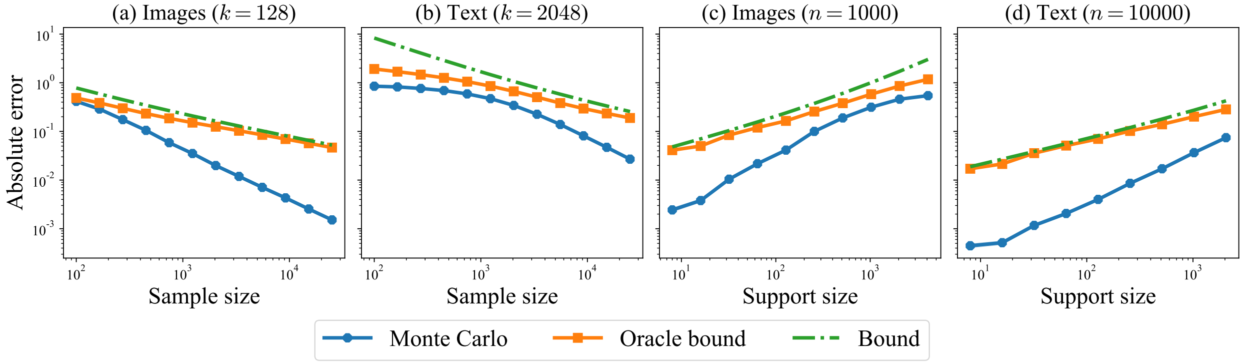

Figure 3.6: Statistical error of the estimated frontier integral on real data. **(a)**: Image data (CIFAR-10) with $k = 128$; **(b)**: Text data (WikiText-103) with $k = 2048$; **(c)**: Image data (CIFAR-10) with $n = 1000$; **(d)**: Text data (WikiText-103) with $n = 10000$. The bounds are scaled by 30.

We estimate the expected absolute error $\mathbb{E}\,|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)|$ from a Monte Carlo estimate using 100 random trials. We compare it with the following bounds in[2] Theorem 3.2:

(a) **Bound**: the distribution independent bound $(\sqrt{k/n} + k/n)\log n$.

(b) **Oracle Bound**: the distribution dependent bound $(\alpha_n(P) + \alpha_n(Q))\log n + \beta_n(P) + \beta_n(Q)$. We assume that the quantities α_n and β_n defined in Theorem 3.2 are known.

We fix k, plot each of these quantities in a log-log plot with varying n and compare their *slopes*.[3] We then repeat the experiment with n fixed and k varying. We often scale the bounds by a constant for easier visual comparison of the slopes; this only changes the intercept and leaves the slope unchanged.

Theorem 3.2 is tight for synthetic data. Figure 3.5 gives the Monte Carlo estimate and the bounds of the statistical error for various synthetic data distributions. In Figure 3.5(a), we observe that the bound has approximately the same slope as the Monte Carlo estimate,

[2] Specifically, we use the expected bounds in (3.9), from which Theorem 3.2 is derived.

[3] A log-log plot of the function $f(x) = cx^{\gamma}$ is a straight line with slope γ and intercept $\log c$. The slope thus captures the *degree* of a polynomial rate.

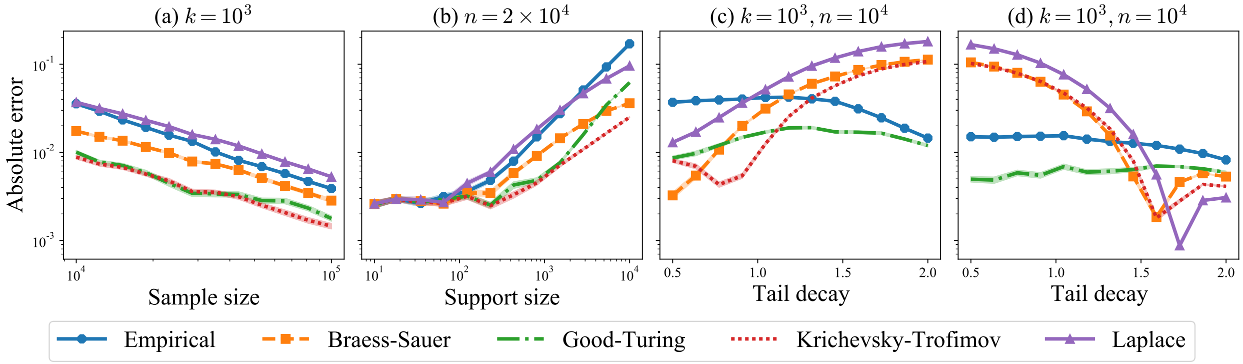

Figure 3.7: Statistical error of the estimated frontier integral with smoothed distribution estimators on synthetic data. **(a)**: Zipf(0) and Dir($\mathbf{1}/2$) with $k = 10^3$; **(b)**: Zipf(0) and Dir($\mathbf{1}/2$) with $n = 2 \times 10^4$; **(c)**: Dir($\mathbf{1}$) and Zipf(r) with $k = 10^3$ and $n = 10^4$; **(d)**: Zipf(2) and Zipf(r) with $k = 10^3$ and $n = 10^4$.

while the oracle bound has a slightly worse slope. In Figure 3.5(b), we observe that the oracle bound captures the correct rate for $k > 300$, while the distribution-independent bound captures the correct rate at small k. For the right two plots, both bounds capture the right rate over a wide range of tail decay. The oracle bound is tighter for fast decay, where the distribution-independent bounds on $\alpha_n(Q)$ and $\beta_n(Q)$ can be very pessimistic.

Theorem 3.2 is somewhat tight for real data. Figure 3.6 contains the analogous plot for real data, where the observations are similar. In Figure 3.6(b), we see that the oracle bound captures the right rate for small sample sizes where $k/n > 1$. However, for large n, the distribution-independent bound is better at matching the slope of the Monte Carlo estimate. The same is true for Figure 3.6(c), where the oracle bound is better for large k. For parts (a) and (d), however, both bounds do not capture the right slope of the Monte Carlo estimate; Theorem 3.2 is not a tight upper bound in this case. That being said, it is still a valid upper bound on the estimation error.

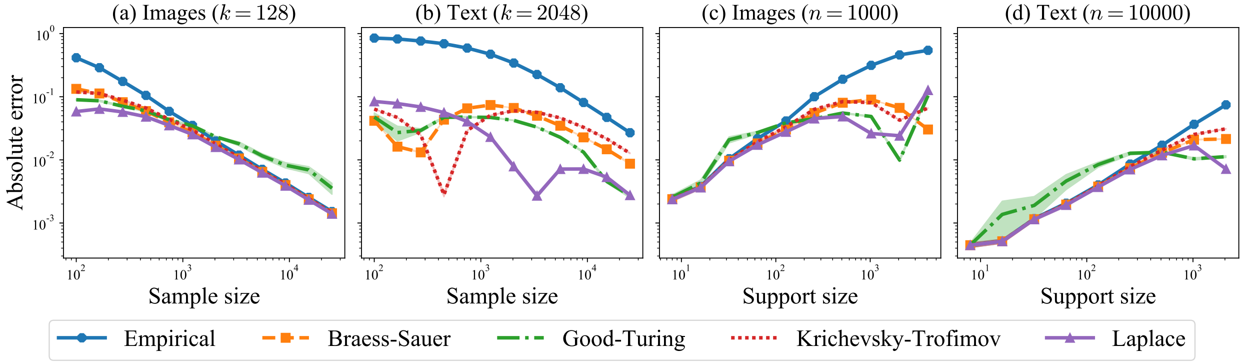

Figure 3.8: Statistical error of the estimated frontier integral with smoothed distribution estimators on real data. **(a)**: Image data (CIFAR-10) with $k = 128$; **(b)**: Text data (WikiText-103) with $k = 2048$; **(c)**: Image data (CIFAR-10) with $n = 1000$; **(d)**: Text data (WikiText-103) with $n = 10000$. The bounds are scaled by 15.

3.6.3 Effect of smoothed distribution estimators

We now show that smoothed estimators can lead to improved estimation over the empirical estimator and thus improved sample complexity as shown in Theorem 3.6. This is practically significant in the context of generative models, since one can have an equally good estimate of the divergence frontier with fewer samples using smoothed estimators (Sajjadi et al., 2018; Djolonga et al., 2020).

Concretely, we compare the Monte Carlo estimates of the absolute error $\mathbb{E}\,|\mathrm{FI}(P_n, Q_n) - \mathrm{FI}(P, Q)|$ for the empirical estimator (denoted "Empirical") as well as smoothed estimators. We consider 4 smoothed estimators as in Orlitsky and Suresh (2015): the (modified) *Good-Turing* estimator, as well as three add-constant estimators: the *Laplace*, *Krichevsky-Trofimov* and *Braess-Sauer* estimators.

Smoothed estimators are more efficient than the empirical estimator. We compare the smoothed estimators to the empirical one in Figure 3.7 on synthetic data and Figure 3.8 on real data. In general, the smoothed distribution estimators reduce the abso-

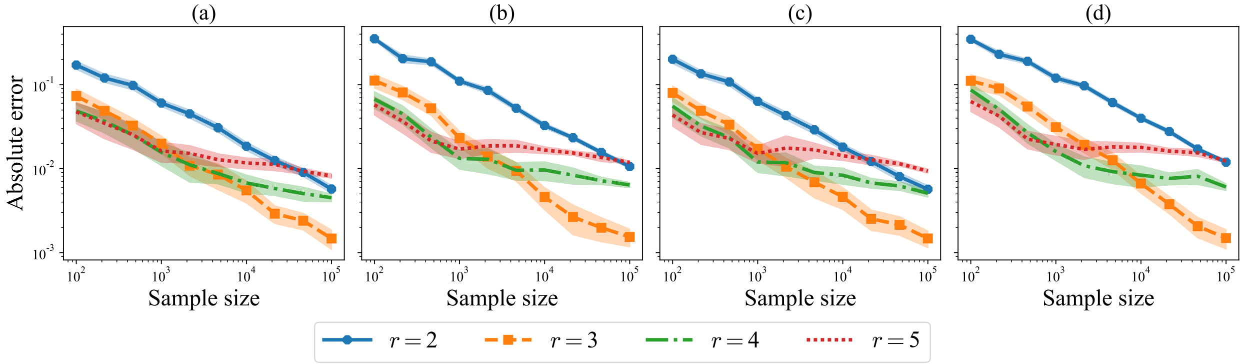

Figure 3.9: Total error of the estimated frontier integral with quantization level $k \propto n^{1/r}$ on 2-dimensional continuous data. **(a)**: $\mathcal{N}(0, I_2)$ and $\mathcal{N}(1, I_2)$; **(b)**: $\mathcal{N}(0, I_2)$ and $\mathcal{N}(0, 5I_2)$; **(c)**: $t_4(0, I_2)$ and $t_4(1, I_2)$ (multivariate t-distribution with 4 degrees of freedom); **(d)**: $t_4(0, I_2)$ and $t_4(0, 5I_2)$.

lute error. For parts (a) and (b) of Figure 3.7, the Good-Turing and the Krichevsky-Trofimov estimators have the best absolute error. For parts (c) and (d), the Good-Turing estimator is adapted to various regimes of tail-decay, outperforming the empirical estimator. The Krichevsky-Trofimov and Braess-Sauer estimators, on the other hand, exhibit small absolute error for particular decay regimes. The results are similar for real data in Figure 3.8.

Practical guidance on choosing a smoothed estimator. While the smoothed estimators offer a marked improvement when k/n is large (that is, close to 1), the best estimator is problem-dependent. As a rule of thumb, we suggest the Krichevsky-Trofimov estimator which works well in the large k/n regime but is still competitive when k/n is small.

3.6.4 Quantization error

Next, we study the effect of the quantization level k on the total error. We consider a simple 2-dimensional synthetic setting where the distributions P, Q are either multivariate normal distributions or t-distributions. We use data-driven quantization with k-means to obtain a quantization $\mathcal{S}_k$: each component of the partition is the region corresponding to one cluster. Finally, we plot the expected absolute error $\mathbb{E}\,|\mathrm{FI}(P, Q) - \mathrm{FI}(P_{\mathcal{S}_k,n}, Q_{\mathcal{S}_k,n})|$, where the $\mathrm{FI}(P, Q)$ is computed using numerical integration and the expectation is estimated with Monte Carlo simulations.

The choice $k = \Theta(n^{1/3})$ works the best. We compare $k = n^{1/r}$ for $r = 2, 3, 4, 5$ in Figure 3.9. For small n, the orders $r \geq 3$ all perform similarly, but $r = 3$ clearly outperforms other choices for $n \geq 10^4$. While our theory does not directly apply for data-dependent partitioning schemes, the choice $k = \Theta(n^{1/3})$ suggested by Theorem 3.5 nevertheless works well in practice. This gives a convenient rule of thumb for practical application of divergence frontiers.

Chapter 4

OPTIMAL TRANSPORT DISTANCES FOR TESTING HOMOGENEITY

4.1 Introduction

In 1932, Schrödinger (Schrödinger, 1932) considered the following lazy gas experiment; see, e.g., Chen et al. (2021) for a review. Image n indistinguishable particles in $\mathbb{R}^d$ moving independently as Brownian motion at temperature ε. At time $t = 0$, we observe that the empirical distribution of their initial locations approximately equals some density p. At time $t = 1$, we observe that the empirical distribution of their terminal locations approximately equals another density q, which differs significantly from what it should be by the law of large numbers, i.e.,

$$q(y) \neq \int \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2\varepsilon}\right) p(x)\mathrm{d}x.$$

It is clear that this situation is unlikely to happen. Schrödinger then inquires for, among all unlikely ways in which this could happen, the most likely path for each particle. As Föllmer (1988) shows, the paths are determined by first solving for the (static) Schrödinger bridge (which is introduced in Section 1.4) and then connecting the two end points by a Brownian bridge with diffusion ε.

Although Schrödinger's lazy gas experiment is typically defined in the dynamic setting for Brownian motion, its static counterpart, Schrödinger bridges (Föllmer, 1988; Léonard, 2012), can be defined more generally. In continuum, the Schrödinger bridge can be made precise as the solution to the entropy-regularized optimal transport (EOT) between two densities p and q (Léonard, 2014), where the entropy is given by the negative differential entropy. Recently, Schrödinger bridges have been used in score-based generative modeling (De Bortoli et al.,

2021) and Markov chain Monte Carlo (Bernton et al., 2019).

In its entropy-regularized form, the Schrödinger bridge problem is closely related to the EOT between two discrete distributions (Cuturi, 2013; Ferradans et al., 2014), where the entropy is given by the negative Shannon entropy. This discrete EOT is particularly attractive both from a computational viewpoint (Cuturi, 2013) and from a statistical viewpoint (Rigollet and Weed, 2018). When we only have access to i.i.d. samples from p and q, one may use the solution to the discrete EOT between the empirical distributions to estimate the Schrödinger bridge. However, it remains largely unclear if this estimation is consistent. Existing works either focus on the case when both p and q are discrete (Bigot et al., 2019; Klatt et al., 2020) or is limited to the regularized cost rather than the solution (Genevay et al., 2019; Mena and Weed, 2019).

The remainder of this chapter is organized as follows. In Section 4.2 we review the Schrödinger bridge problem and its connection to the entropy-regularized optimal transport problem. In Section 4.3 we introduce as an estimator of the Schrödinger bridge in continuum the so-called discrete Schrödinger bridge which recovers Schrödinger's original discrete set-up as the Schrödinger bridge connecting two empirical distributions. We show that it is the solution to a modified discrete EOT problem. In Section 4.4 we demonstrate how to apply the Schrödinger bridge to homogeneity testing. In Section 4.5 we prove its convergence towards the Schrödinger bridge in continuum as well as limiting Gaussian fluctuations for this convergence. We also derive the second order Gaussian chaos limit in Appendix C.5. In Section 4.6 we outline the proof sketches of our results. Finally, in Section 4.7, we compare the Schrödinger bridge based test with other alternatives on both synthetic and real data.

4.2 Schrödinger Bridge and Entropy-Regularized Optimal Transport

We review the Schrödinger problem and its connection to the information projection (or I-projection). We show that, through the lens of the KL divergence, the Schrödinger problem and the discrete entropy-regularized optimal transport problem considered by Cuturi (2013) can be written in a unified framework.

4.2.1 Schrödinger bridge in continuum

Given $\varepsilon \in \mathbb{R}_+$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, we assume that the following Markov transition density is well-defined:

$$p_\varepsilon(x, y) := \frac{1}{Z_\varepsilon(x)} \exp\left[-\frac{1}{\varepsilon} c(x, y)\right],$$

where $Z_\varepsilon(x)$ is the normalizing constant. For instance, when c is the quadratic cost, this is the transition density of Brownian motion with diffusion ε considered in Schrödinger's lazy gas experiment. Suppose that (W_0, W_1) is a pair of random vectors distributed according to this Markov transition kernel. Let P and Q be two probability measures on $\mathbb{R}^d$. Informally, the *(static) Schrödinger bridge* connecting P and Q at temperature ε is the joint distribution of (W_0, W_1) conditioned to have $W_0 \sim P$ and $W_1 \sim Q$. In continuum, when P and Q have densities w.r.t. the Lebesgue measure, it can be made precise as the solution to the following entropy-regularized optimal transport (EOT) problem (Föllmer, 1988; Léonard, 2012, 2014):

$$\min_{\nu \in \Pi(P,Q)} \left[\int c(x, y) \mathrm{d}\nu(x, y) + \varepsilon H(\nu)\right], \tag{4.1}$$

where $\Pi(P, Q)$ is the set of couplings with marginals P and Q, and $H(\nu)$ is the entropy of ν defined as $H(\nu) := \int \log \nu(x, y) \mathrm{d}\nu(x, y)$ if ν is a density[1] and infinity otherwise. We refer to (4.1) as the *Schrödinger problem.*

When $\varepsilon = 0$, the Schrödinger problem reduces to the optimal transport (OT) problem. Whereas the latter usually admits a degenerate solution given by a transport map with zero-measure support (Santambrogio, 2015, Theorem 1.17), the entropy term in the former prevents such solutions from existing. Moreover, as $\varepsilon \to 0$, the minimum of the Schrödinger problem converges to the one of the OT problem and the minimizer (if exists) as well (Léonard, 2012, Theorem 3.3). In other words, the Schrödinger problem can be viewed as a *smooth* approximation to the OT problem which quantifies how close two distributions are. As shown in Section 4.4, it can be used to test for homogeneity.

[1] We follow the standard abuse of keeping the same notation for an absolutely continuous measure and its density.

4.2.2 Characterization of the Schrödinger bridge

The Schrödinger problem (4.1) admits an alternative form that are connected to the KL divergence. With a probability measure R_ε defined as $R_\varepsilon(x, y) := P(x)p_\varepsilon(x, y)$, we have

$$\mathrm{KL}(\nu \| R_\varepsilon) = \frac{1}{\varepsilon} \int c(x, y) \mathrm{d}\nu(x, y) + H(\nu) + \int \log Z_\varepsilon(x) \mathrm{d}P(x) - H(P)$$

for all $\nu \in \Pi(P, Q)$. Thus, the minimizer of the problem (4.1) is the same as the one of

$$\min_{\nu \in \Pi(P,Q)} \mathrm{KL}(\nu \| R_\varepsilon).$$

Therefore, the Schrödinger bridge has the following geometry interpretation—it is the *I-projection* (Csiszár, 1975) of the reference measure R_ε onto the set of couplings $\Pi(P, Q)$.

Since $\mathrm{KL}(\cdot \| R_\varepsilon)$ is strictly convex, the Schrödinger problem (4.1) has a unique solution μ_ε (if it exists). Using results on I-projections developed by Csiszár (1975), it can be shown that, when there exists $\nu \in \Pi(P, Q)$ such that $\mathrm{KL}(\nu \| R_\varepsilon) < \infty$, the solution μ_ε exists and admits the following expression (Rüschendorf and Thomsen, 1993, Theorem 3): there exists two measurable functions a_ε and b_ε, to be called the *Schrödinger potentials*, such that

$$\frac{\mathrm{d}\mu_\varepsilon}{\mathrm{d}(P \otimes Q)}(x, y) = \xi(x, y) := \exp\left\{ -\frac{1}{\varepsilon}[c(x, y) - a_\varepsilon(x) - b_\varepsilon(y)] \right\}. \tag{4.2}$$

Note that $\mu_\varepsilon \in \Pi(P, Q)$. This implies

$$\int \xi(x, y) \mathrm{d}P(x) = 1 \ Q\text{-a.s.} \quad \text{and} \quad \int \xi(x, y) \mathrm{d}Q(y) = 1 \ P\text{-a.s.} \tag{4.3}$$

We assume throughout this chapter that the Schrödinger bridge μ_ε exists.

4.2.3 Connection to Cuturi's entropy-regularized optimal transport

Cuturi (2013) considered a discrete EOT problem between two discrete measures P and Q

$$\min_{\nu \in \Pi(P,Q)} \left[\int c(x, y) \mathrm{d}\nu(x, y) + \varepsilon \, \mathrm{Ent}(\nu) \right], \tag{4.4}$$

where, for a discrete measure ν, $\mathrm{Ent}(\nu)$ is the negative Shannon entropy of ν defined as $\sum_{a \in \mathrm{Supp}(\nu)} \nu(a) \log \nu(a)$; see also (Ferradans et al., 2014). This problem was initially introduced as an approximation to the OT problem between two discrete measures which can be solved efficiently using the Sinkhorn algorithm (Sinkhorn, 1967). It has now been popular in machine learning due to other advantages, for instance, it fits into a differentiable programming framework (Genevay et al., 2018; Salimans et al., 2018; Sanjabi et al., 2018). We will give a thorough discussion on this topic in Chapter 5. Due to its popularity, we refer to (4.4) as the *discrete EOT problem* and its solution the *discrete EOT plan* which also satisfies the property (4.2).

The discrete EOT problem can be viewed as a discrete counterpart of the Schrödinger problem through the lens of the KL divergence. Consider the problem

$$\min_{\nu\in\Pi(P,Q)}\left[\int c(x,y)\mathrm{d}\nu(x,y)+\varepsilon\,\mathrm{KL}(\nu\|P\otimes Q)\right]. \tag{4.5}$$

When P and Q are densities, we have

$$\mathrm{KL}(\nu\|P\otimes Q)=\begin{cases}\int\nu(x,y)\log\frac{\nu(x,y)}{P(x)Q(y)}\mathrm{d}x\mathrm{d}y=H(\nu)-H(P)-H(Q) & \text{if } \nu \text{ has a density}\\ \infty & \text{otherwise.}\end{cases}$$

Consequently, the solution to the Schrödinger problem is the same as the one to (4.5). Analogously, when P and Q are discrete measures, the solution to the discrete EOT problem is the same as the one to (4.5). Hence, the problem (4.5) unifies the Schrödinger problem and the discrete EOT problem. We call it the *EOT problem* and, for convenience, the solution to it the *Schrödinger bridge* even if P and Q are not densities. We focus on this problem in the remainder of this chapter.

4.3 Estimating the Schrödinger Bridge

We propose an empirical estimator of the Schrödinger bridge, called the *discrete Schrödinger bridge*, when we have i.i.d. samples from the two marginal distributions P and Q. This estimator is based on a convex combination of all Monge couplings between the two empirical marginal measures. It recovers Schrödinger's original discrete set-up as the Schrödinger bridge connecting the two empirical marginal measures. Finally, we show that it is the solution to a discrete EOT problem which is slightly different from the one considered by Cuturi (2013).

4.3.1 Discrete Schrödinger bridge

Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two independent i.i.d. samples from two distributions P and Q on $\mathbb{R}^d$ with $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ being the empirical measures, respectively. Since the EOT problem (4.5) can be viewed as a smooth approximation to the OT problem between P and Q. It is natural to construct an estimator of the Schrödinger bridge μ_ε by approximating the optimal transport plan between P_n and Q_n, i.e., the solution to

$$\min_{\nu \in \Pi(P_n, Q_n)} \sum_{i=1}^n \sum_{j=1}^n c(X_i, Y_j)\nu(X_i, Y_j). \tag{4.6}$$

We first recall some results regarding the empirical OT problem (4.6). Let $\mathcal{S}_n$ be the set of permutations on $[n] := \{1, \dots, n\}$. Every $\sigma = (\sigma_1, \dots, \sigma_n) \in \mathcal{S}_n$ can be viewed as a matching between these two sets of random vectors, i.e., X_i is matched to Y_{σ_i} for each $i \in [n]$. It induces a Monge coupling $M_\sigma := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ which belongs to $\Pi(P_n, Q_n)$. Hence, the empirical OT problem (4.6) is a convex relaxation of the following optimal matching problem:

$$\min_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^n c(X_i, Y_{\sigma_i}) = \min_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n \sum_{j=1}^n c(X_i, Y_j) M_\sigma(X_i, Y_j). \tag{4.7}$$

According to Peyré and Cuturi (2019, Proposition 2.1), the two problems (4.6) and (4.7) share the same minimum. Moreover, the problem (4.7) always has a solution $\sigma^\star$ whose associated Monge map $M_{\sigma^\star}$ solves the problem (4.6). Thus, a natural way to construct a smooth approximation to the solution $M_{\sigma^\star}$ is to consider a convex combination of all Monge maps, i.e., consider $\sum_{\sigma \in \mathcal{S}_n} \gamma(\sigma) M_\sigma$ for some probability distribution γ over $\mathcal{S}_n$.

The empirical estimator we study is based on a particular distribution γ_ε defined as follows. For every $\sigma \in \mathcal{S}_n$, its associated cost is given by $\sum_{i=1}^n c(X_i, Y_{\sigma_i})$. Since the objective is to minimize the cost, we assign each permutation σ the (random) weight $w(\sigma) := \exp(-\sum_{i=1}^n c(X_i, Y_{\sigma_i})/\varepsilon)$ so that a permutation with a large cost gets an exponentially small weight. Now we obtain a Gibbs measure on $\mathcal{S}_n$, i.e.,

$$\gamma_\varepsilon(\sigma) := \frac{w(\sigma)}{\sum_{\tau \in \mathcal{S}_n} w(\tau)} = \frac{\exp\left(-\sum_{i=1}^n c(X_i, Y_{\sigma_i})/\varepsilon\right)}{\sum_{\tau \in \mathcal{S}_n} \exp\left(-\sum_{i=1}^n c(X_i, Y_{\tau_i})/\varepsilon\right)} = \frac{\xi^{\otimes}(X, Y_\sigma)}{\sum_{\tau \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\tau)}, \tag{4.8}$$

where ξ is defined in (4.2) and $\xi^{\otimes}(X, Y_\sigma) := \prod_{i=1}^n \xi(X_i, Y_{\sigma_i})$. This leads to the estimator

$$\mu_{\varepsilon,n} := \sum_{\sigma \in \mathcal{S}_n} \gamma_\varepsilon(\sigma) M_\sigma = \frac{\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n} M_\sigma \xi^{\otimes}(X, Y_\sigma)}{\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma)}. \tag{4.9}$$

It can be shown that $\mu_{\varepsilon,n}$ recovers Schrödinger's original discrete set-up as the Schrödinger bridge connecting P_n and Q_n at temperature ε. To see this, consider a realization $X_i = x_i$ and $Y_i = y_i$ for $i \in [n]$. Then P_n and Q_n are (nonrandom) discrete distributions supported on n categories. Imagine n independent particles $\{W^i\}_{i=1}^n$, starting from positions $W_0^i = x_i$, $i \in [n]$, and making jumps according to the Markov transition kernel $p_\varepsilon(x_i, \cdot)$, $i \in [n]$. Let $L_n(1) := \frac{1}{n}\sum_{i=1}^n \delta_{W_1^i}$ be the empirical distribution of their terminal locations and $L_n(0,1) := \frac{1}{n}\sum_{i=1}^n \delta_{(W_0^i, W_1^i)}$ be the joint empirical distribution at two time points. According to Pal and Wong (2020, Section 3.2), the law of $L_n(0,1)$ given $L_n(1) = Q_n$ is exactly given by $\mu_{\varepsilon,n}$ (with $X_i = x_i$ and $Y_i = y_i$, $i \in [n]$). In other words, for each permutation $\sigma \in \mathcal{S}_n$,

$$\mathbb{P}\left(L_n(0,1) = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_{\sigma_i})} \,\middle|\, L_n(1) = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \right) = \gamma_\varepsilon(\sigma).$$

We refer to $\mu_{\varepsilon,n}$ as the *discrete Schrödinger bridge.*

4.3.2 Reformulation as a discrete entropy-regularized optimal transport

It turns out that the discrete Schrödinger bridge $\mu_{\varepsilon,n}$ is the solution to a discrete EOT problem that is different from the one of Cuturi (2013).

Lemma 4.1. *Let $\mathcal{M}_1(\mathcal{S}_n)$ be the set of probability measures on $\mathcal{S}_n$. We have*

$$\gamma_\varepsilon = \operatorname*{arg\,min}_{\gamma \in \mathcal{M}_1(\mathcal{S}_n)} \left[\sum_{i=1}^n \sum_{j=1}^n c(X_i, Y_j)\nu_\gamma(X_i, Y_j) + \frac{\varepsilon}{n}\operatorname{Ent}(\gamma) \right], \tag{4.10}$$

where $\nu_\gamma := \sum_{\sigma\in\mathcal{S}_n} \gamma(\sigma) M_\sigma \in \Pi(P_n, Q_n)$. In particular, $\mu_{\varepsilon,n} = \nu_{\gamma_\varepsilon}$.

Proof. We claim that minimizing (4.10) is equivalent to minimizing $\mathrm{KL}(\gamma \| \gamma_\varepsilon)$ which is uniquely minimized at $\gamma = \gamma_\varepsilon$. In fact,

$$\begin{aligned}\mathrm{KL}(\gamma\|\gamma_\varepsilon) &= \sum_{\sigma\in\mathcal{S}_n}\gamma(\sigma)\log\frac{\gamma(\sigma)}{\gamma_\varepsilon(\sigma)} = \sum_{\sigma\in\mathcal{S}_n}\gamma(\sigma)\log\left(\frac{\gamma(\sigma)\sum_{\tau\in\mathcal{S}_n}w(\tau)}{w(\sigma)}\right)\\ &= \mathrm{Ent}(\gamma) + \log\left[\sum_{\tau\in\mathcal{S}_n}w(\tau)\right]\sum_{\sigma\in\mathcal{S}_n}\gamma(\sigma) + \frac{1}{\varepsilon}\sum_{\sigma\in\mathcal{S}_n}c(X,Y_\sigma)\gamma(\sigma)\\ &= \frac{n}{\varepsilon}\sum_{i=1}^n\sum_{j=1}^n c(X_i,Y_j)\nu_\gamma(X_i,Y_j) + \mathrm{Ent}(\gamma) + \log\sum_{\tau\in\mathcal{S}_n}w(\tau),\end{aligned}$$

and thus the claim follows. □

Due to Birkhoff (1946), every doubly stochastic matrix can be written as a convex combination of permutation matrices. As a result, every coupling $M \in \Pi(P_n, Q_n)$ can be expressed as $M = \sum_{\sigma\in\mathcal{S}_n}\gamma_M(\sigma)M_\sigma$ for some $\gamma_M \in \mathcal{M}_1(\mathcal{S}_n)$. Note that such convex combinations are generally not unique. Hence, the problem (4.10) without the regularization term admits the same minimum as the empirical OT problem, i.e.,

$$\min_{\gamma\in\mathcal{M}_1(\mathcal{S}_n)}\sum_{i=1}^n\sum_{j=1}^n c(X_i,Y_j)\nu_\gamma(X_i,Y_j) = \min_{\nu\in\Pi(P_n,Q_n)}\sum_{i=1}^n\sum_{j=1}^n c(X_i,Y_j)\nu(X_i,Y_j).$$

This suggests that the problem (4.10) adds an entropy regularization term to the empirical OT problem in the space of probability measures on permutations.

4.3.3 Relationship to the plug-in estimator

Another estimator one may consider is based on the minimizer of the EOT problem (4.5) with P_n and Q_n plugged in, i.e.,

$$\min_{\nu\in\Pi(P_n,Q_n)}\left[\int c(x,y)\mathrm{d}\nu(x,y) + \varepsilon\,\mathrm{KL}(\nu\|P_n\otimes Q_n)\right]. \tag{4.11}$$

We refer to it as the *discrete EOT plan.* This problem is the discrete EOT problem (4.4) specialized to the empirical measures P_n and Q_n which adds an entropy regularization term in the original space of couplings. Even though there is a rich literature on the statistical properties of (4.11), they mainly focus on the minimum rather than the minimizer, i.e., the discrete EOT plan (Bigot et al., 2019; Genevay et al., 2019; Mena and Weed, 2019). Klatt et al. (2020) prove a Gaussian limit for the discrete EOT plan but is limited to the case when P and Q are discrete with finite supports. We will give a thorough discussion on this topic in Chapter 5.

The relationship between the discrete Schrödinger bridge and the discrete EOT plan remains unclear as of today. However, they are connected through the lens of matrix balancing; see (Beichl and Sullivan, 1999) and references therein. To see this, we define an $n \times n$ matrix K with (i, j)-th element being $K_{ij} = \exp(-c(X_i, Y_j)/\varepsilon)$. Let $|K|$ denote the permanent of K, i.e.,

$$|K| = \sum_{\sigma \in \mathcal{S}_n} \prod_{i=1}^{n} K_{i\sigma_i} = \sum_{\sigma \in \mathcal{S}_n} \exp\left(-\sum_{i=1}^{n} c(X_i, Y_{\sigma_i})/\varepsilon\right),$$

which is exactly the denominator in the second expression of γ_ε in (4.8). Define a matrix $A_{\text{dsb}} \in \mathbb{R}_+^{n\times n}$ by, for each $i, j \in [n]$,

$$(A_{\text{dsb}})_{i,j} = n\mu_{\varepsilon,n}(X_i, Y_j) = \sum_{\sigma:\sigma_i = j} \gamma_\varepsilon(\sigma) = \frac{\sum_{\sigma:\sigma_i=j} \exp(-\sum_{i=1}^{n} c(X_i, Y_{\sigma_i})/\varepsilon)}{\sum_{\sigma\in\mathcal{S}_n} \exp(-\sum_{i=1}^{n} c(X_i, Y_{\sigma_i})/\varepsilon)}.$$

That is, A_{dsb} is the probability matrix associated with the discrete Schrödinger bridge $\mu_{\varepsilon,n}$ scaled by n. A little bit of algebra omitted here shows that the numerator of $(A_{\text{dsb}})_{i,j}$ is exactly given by $\exp(-c(X_i, Y_j)/\varepsilon)\,|K^{ij}|$, where K^{ij} is the minor of K obtained by deleting the i-th row and j-th column. Hence, we have $(A_{\text{dsb}})_{i,j} = K_{ij}\,|K^{ij}|\,/\,|K|$. The matrix A_{dsb} is doubly stochastic and called the *matrix balance* of K (Beichl and Sullivan, 1999, Section 3). In the same spirit, we can define a doubly stochastic matrix associated with the discrete EOT plan scaled by n. This matrix is known as the *Sinkhorn balance* of K (Beichl and Sullivan, 1999, Section 4). In Section 4.1, Beichl and Sullivan (1999) shows that the Sinkhorn balance of a 0-1 matrix approximates the matrix balance of it. However, a more in-depth investigation on the relationship of these two objects is needed. As an illustration, we visualize the discrete Schrödinger bridge and the discrete EOT plan as heatmaps in Figure 4.1.

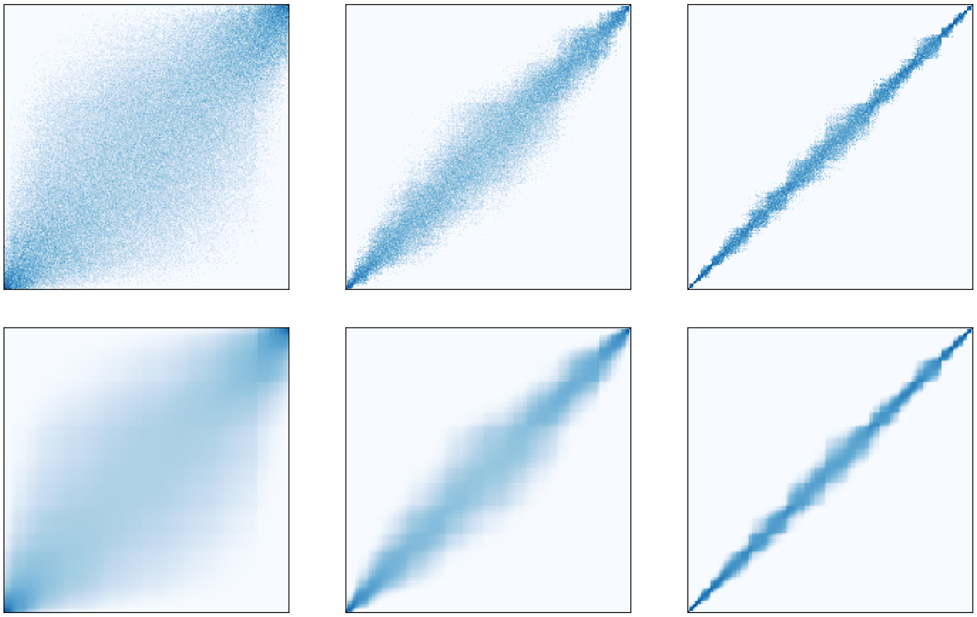

Figure 4.1: Heatmaps of the discrete Schrödinger bridge (**top**) and the discrete EOT plan (**bottom**) with $n = 200$ at decreasing values of ε (**from left to right**). The two marginals are both normal distributions with mean 0 and standard deviation 0.05. The cost function is chosen as the quadratic cost. The observations are ordered so that the optimal Monge coupling is the diagonal line.

4.3.4 Relationship to the optimal transport plan

When the regularization parameter is chosen to be $\varepsilon = \varepsilon_n = o(1)$, it is desirable that the discrete Schrödinger bridge $\mu_n := \mu_{\varepsilon_n,n}$ converges to the optimal transport plan. We show in this section that it is true at rate $O(n^{-2/d}(\log n)^{2/d})$. By the Kantorovich duality (Villani, 2009, Theorem 5.10), there exists a pair of functions (ϕ, ψ), known as the *Kantorovich potentials*, such that

$$D[y \mid x] := c(x, y) - \phi(x) - \psi(y) \geq 0, \quad \text{for all } x, y \in \mathbb{R}^d. \tag{4.12}$$

Assumption 4.1. *We make the following assumptions.*

(a) There exist $\alpha > 2$ and $\gamma > 0$ such that $\mathbb{E}[\exp(\gamma \|Z\|^{\alpha})] < \infty$ for $Z \sim P$ and $Z \sim Q$.

(b) For $Y \sim Q$, each of its coordinates is sub-Gaussian with parameter K.

(c) The unique solution to the OT problem is $\mu_\star := (id, T_\star)\sharp P$ where $T_\star$ is a Monge map; see (Villani, 2009, Theorem 10.38) for sufficient conditions on c.

(d) There exists $L, L' > 0$ such that, P-a.s.,

$$L \|y - T_\star(x)\|^2 \leq D[y \mid x] \leq L' \|y - T_\star(x)\|^2 . \tag{4.13}$$

Remark 4.1. *Due to the existence of the Monge map $T_\star$, we have $D[y \mid x] = 0$ iff $y = T_\star(x)$. This implies that $T_\star(x)$ is the unique minimizer of the function $D[\cdot \mid x]$. By Taylor's theorem, $D[\cdot \mid x]$ is quadratic in a neighborhood of $T_\star(x)$, which justifies the quadratic approximation in (4.13).*

The next result shows that μ_n converges to $\mu_\star$ in W_2^2 at rate $O(n^{-2/d} \log n)$ as $n \to \infty$. The proof is deferred to Appendix C.1.

Proposition 4.2. *Let $d > 4$ and $\varepsilon_n = n^{-2/d}$. Under Assumption 4.1, it holds for all large enough n that, with probability at least $1 - \delta$,*

$$\mathsf{W}_2^2(\mu_n, \mu_\star) \lesssim C \left[\left(\frac{\log (1/\delta)}{n} \right)^{2/d} + \frac{\log n}{n^{2/d}} \right],$$

where C is a problem-specific constant.

A closely related problem is estimating the optimal transport map $T_\star$. Hütter and Rigollet (2021) considered this problem in the minimax estimation framework. Assuming that $T_\star$ is α-smooth for some $\alpha \geq 1$, they showed that the minimax estimation error measured by the $\mathbf{L}^2(P)$ distance is at rate $O(n^{-2\alpha/(2\alpha-2+d)})$, which can be achieved up to a logarithmic factor. This rate is reminiscent of the standard nonparametric minimax estimation rate. Let $\hat{T}$ be the minimax near-optimal estimator constructed in their paper. Since $\hat{T}$ also defines a coupling $(id, \hat{T})\sharp P$, it follows that

$$\mathsf{W}_2^2\big((id, \hat{T})\sharp P, \mu_\star\big) \leq \int \left\| \hat{T}(x) - T_\star(x) \right\|^2 \mathrm{d}P(x) \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} \log^2(n).$$

Hence, our estimator achieves the same rate with $\alpha = 1$.

4.4 Schrödinger Bridge for Homogeneity Testing

We demonstrate how the Schrödinger bridge can be used to test homogeneity of distributions.

4.4.1 Two-sample homogeneity testing

Given two independent i.i.d. samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ from distributions P and Q, respectively, we are interested in determining whether they come from the same distribution. This can be formulated as a two-sample homogeneity testing problem:

$$\mathbf{H}_0 : P = Q \leftrightarrow \mathbf{H}_1 : P \neq Q. \tag{4.14}$$

That is, we test the null hypothesis that they come from the same distribution against the alternative hypothesis that they do not.

A typical procedure for solving such problems consists of the following steps. First, we choose a functional $T : \mathcal{M}_1(\mathbb{R}^d) \times \mathcal{M}_1(\mathbb{R}^d) \to \mathbb{R}$ so that $T(P, Q)$ quantifies the distance between P and Q. Second, we estimate $T(P, Q)$ from the data to obtain a test statistic T_n. Since the statistic T_n approximates the metric $T(P, Q)$, the larger it is the less likely $\mathbf{H}_0$ is true. Finally, we choose a threshold t_n and adopt the decision rule (or test) $\mathbb{1}\{T_n > t_n\}$, that is, we reject the null if the test statistic exceeds the threshold. The performance of a test can be measured by two quantities: the *type I error rate* $\mathbb{P}(T_n > t_n \mid \mathbf{H}_0)$, i.e., the probability of rejecting the null given that the null is true, and the *statistical power* $\mathbb{P}(T_n > t_n \mid \mathbf{H}_1)$, i.e., the probability of rejecting the null given that the null is not true.

4.4.2 Centered Schrödinger bridge cost

Since the Schrödinger bridge μ_ε is a smoothed approximation to the optimal transport plan, its cost of transport $T_\varepsilon(P, Q) := \int c \mathrm{d}\mu_\varepsilon$ can be used to measure the distance between P and Q. In fact, it is known as the Sinkhorn distance (Cuturi, 2013) when P and Q are discrete which satisfies all distance axioms except for the coincidence axiom, i.e., $T_\varepsilon(P, Q) = 0$ iff

$P = Q$. Hence, we use the centered version

$$\overline{T}_\varepsilon(P, Q) := T_\varepsilon(P, Q) - \frac{1}{2}T_\varepsilon(P, P) - \frac{1}{2}T_\varepsilon(Q, Q), \tag{4.15}$$

which we refer to as the *centered Schrödinger bridge cost.* Note that this centering trick also appears in (Ramdas et al., 2017, Section 3.3) to relate the EOT cost to the *energy distance.* The centered Schrödinger bridge cost is symmetric and equals zero if $P = Q$. Moreover, it has the following property which justifies its use as a probability metric. The proof is deferred to Appendix C.2.

Proposition 4.3. *The centered Schrödinger bridge cost $\overline{T}_\varepsilon(P, Q)$ is continuous in $\varepsilon \in (0, \infty)$. Moreover, if c is bounded and continuous, then*

$$\overline{T}_\infty(P, Q) := \lim_{\varepsilon \uparrow \infty} \overline{T}_\varepsilon(P, Q) = \int c \mathrm{d}(P \otimes Q) - \frac{1}{2}\int c \mathrm{d}(P \otimes P) - \frac{1}{2}\int c \mathrm{d}(Q \otimes Q). \tag{4.16}$$

Remark 4.2. *The limit at $\varepsilon = \infty$ in (4.16) is half the* energy distance *w.r.t. c introduced by Székely and Rizzo (2004) and generalized by Lyons (2013). Moreover, under appropriate assumptions, the limit at $\varepsilon = 0$, i.e., $\lim_{\varepsilon \downarrow 0} \overline{T}_\varepsilon(P, Q)$, is exactly the OT distance between P and Q; see Léonard (2012, Theorem 3.3) for the continuous case and Peyré and Cuturi (2019, Proposition 4.1) for the discrete case. Hence, the centered Schrödinger bridge cost interpolates between the OT distance and the energy distance.*

4.4.3 Schrödinger bridge statistic

We show in Section 4.5 that the discrete Schrödinger bridge $\mu_{\varepsilon,n}$ is a consistent estimator of μ_ε, so it is natural to estimate $\overline{T}_\varepsilon(P, Q)$ by

$$\overline{T}_{\varepsilon,n} := \overline{T}_\varepsilon(P_n, Q_n) = T_\varepsilon(P_n, Q_n) - \frac{1}{2}T_\varepsilon(P_n, P_n) - \frac{1}{2}T_\varepsilon(Q_n, Q_n), \tag{4.17}$$

where $T_\varepsilon(\nu_1, \nu_2)$ is the transport cost of the discrete Schrödinger bridge connecting ν_1 and ν_2 for $\nu_1, \nu_2 \in \{P_n, Q_n\}$. For instance,

$$T_\varepsilon(P_n, Q_n) := \int c \, \mathrm{d}\mu_{\varepsilon,n} = \sum_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} c(X_i, Y_{\sigma_i})\gamma_\varepsilon(\sigma), \tag{4.18}$$

Algorithm 1 Gibbs sampling for the Schrödinger bridge statistic

```
**Input:** samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, functions $c$, burn-in $B$ and number of iterations $L$.
**Initialization:** $\sigma^{(0)} \leftarrow \mathrm{id}$.
**for** $t = 0, \ldots, L-1$ **do**
    Randomly select $i \neq j \in [n]$.
    Compute $r \leftarrow \exp\left\{[c(X_i, Y_{\sigma_i^{(t)}}) + c(X_j, Y_{\sigma_j^{(t)}}) - c(X_i, Y_{\sigma_j^{(t)}}) - c(X_j, Y_{\sigma_i^{(t)}})]/\varepsilon\right\}$.
    Generate $a \sim \mathrm{Bern}(r/(1+r))$.
    **if** a = 1 **then**
        Obtain $\sigma^{(t+1)}$ from $\sigma^{(t)}$ by swapping the entries $\sigma_i^{(t)}$ and $\sigma_j^{(t)}$.
    **else**
        Set $\sigma^{(t+1)} \leftarrow \sigma^{(t)}$.
    **end if**
**end for**
**Output:** $\frac{1}{L-B} \sum_{t=B+1}^{L} \frac{1}{n} c(X, Y_{\sigma^{(t)}})$.
```

where γ_ε is defined in (4.8). We refer to $\overline{T}_{\varepsilon,n}$ as the *centered Schrödinger bridge statistic.*

Since the space of permutations $\mathcal{S}_n$ is prohibitively large, it is infeasible to compute the Schrödinger bridge statistic exactly. We adopt here a Gibbs sampling approach to sample from the distribution γ_ε and estimate the Schrödinger bridge statistic by the empirical average on the sample. We mention here among many others the monograph by Wakefield (2013, Chapter 3) for a review on Gibbs sampling. The procedure is summarize in Algorithm 1. Let $\sigma^{(t)}$ be the current sample. We choose the proposal distribution as

$$g(\sigma \mid \sigma^{(t)}) = \begin{cases} \frac{2}{n(n-1)} & \text{if } \sigma \text{ and } \sigma^{(t)} \text{ differ in exactly two indices} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, we randomly select $i \neq j \in [n]$ and swap $\sigma_i^{(t)}$ and $\sigma_j^{(t)}$ to obtain a candidate

$\sigma^{(t+1)}$. We then accept the candidate $\sigma^{(t+1)}$ with probability

$$r := \frac{\gamma_\varepsilon(\sigma^{(t+1)})}{\gamma_\varepsilon(\sigma^{(t)}) + \gamma_\varepsilon(\sigma^{(t+1)})} = \exp\left\{\frac{1}{\varepsilon}\left[c(X_i, Y_{\sigma_i^{(t)}}) + c(X_j, Y_{\sigma_j^{(t)}}) - c(X_i, Y_{\sigma_j^{(t)}}) - c(X_j, Y_{\sigma_i^{(t)}})\right]\right\}.$$

Hence, for all $\sigma \neq \sigma'$, the transition probability reads

$$h(\sigma' \mid \sigma) = \begin{cases} \frac{2}{n(n-1)} \frac{q_\varepsilon^*(\sigma')}{q_\varepsilon^*(\sigma) + q_\varepsilon^*(\sigma')} & \text{if } \sigma' \text{ and } \sigma \text{ differ in exactly two indices} \\ 0 & \text{otherwise.} \end{cases}$$

It satisfies the detailed balance equation $h(\sigma' \mid \sigma)\gamma_\varepsilon(\sigma) = h(\sigma \mid \sigma')\gamma_\varepsilon(\sigma')$, and thus the samples generated from Algorithm 1 can be used to approximate the distribution γ_ε as well as the Schrödinger bridge statistic.

4.5 Asymptotics of the Discrete Schrödinger Bridge

In this section, we summarize asymptotic properties of the discrete Schrödinger bridge $\mu_{\varepsilon,n}$ in (4.9). Given a probability measure ν and an integer $p \geq 1$, let $\mathbf{L}^p(\nu)$ be the space of functions that have finite p-th norm under ν and $\mathbf{L}_0^p(\nu)$ be the subset of $\mathbf{L}^p(\nu)$ whose expectation under ν is zero. We follow the standard abuse of keeping the same notation for an absolutely continuous measure and its density.

We express our results in their full generality. Let P and Q be two probability measures on $\mathbb{R}^d$. Let $\mu \in \Pi(P, Q)$ such that μ has density ξ w.r.t. $P \otimes Q$ satisfying

$$\int \xi(x, y) \mathrm{d}P(x) = 1 \ Q\text{-a.s.} \quad \text{and} \quad \int \xi(x, y) \mathrm{d}Q(y) = 1 \ P\text{-a.s.} \tag{4.19}$$

Given two independent i.i.d. samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ from P and Q, respectively, we consider the random measure

$$\mu_n := \frac{\frac{1}{n!}\sum_{\sigma \in \mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})} \xi^\otimes(X, Y_\sigma)}{\frac{1}{n!}\sum_{\sigma \in \mathcal{S}_n} \xi^\otimes(X, Y_\sigma)}, \tag{4.20}$$

where $\xi^\otimes(X, Y_\sigma) := \prod_{i=1}^n \xi(X_i, Y_{\sigma_i})$. As a special case, when both P and Q are densities (or discrete measures), the Schrödinger bridge (if exists) μ_ε satisfies (4.19) with $\xi(x, y) = \exp(-(c(x, y) - a_\varepsilon(x) - b_\varepsilon(y))/\varepsilon)$, and μ_n coincides with the discrete Schrödinger bridge.

4.5.1 Consistency of the measure

Our first result shows that μ_n is a consistent estimator of μ. Let us start by defining two operators on $\mathbf{L}^2(P)$ and $\mathbf{L}^2(Q)$ induced by μ whose validity follows from Jensen's inequality.

Definition 4.1. *Define a linear operator $\mathcal{A} : \mathbf{L}^2(P) \to \mathbf{L}^2(Q)$ and its adjoint $\mathcal{A}^* : \mathbf{L}^2(Q) \to \mathbf{L}^2(P)$ by*

$$(\mathcal{A}f)(y) := \int f(x)\xi(x,y)\mathrm{d}P(x) \quad \textit{and} \quad (\mathcal{A}^*g)(x) := \int g(y)\xi(x,y)\mathrm{d}Q(y). \tag{4.21}$$

We call $A : (x,y) \mapsto \xi(x,y)$ the kernel of $\mathcal{A}$ and $A^* : (y,x) \mapsto \xi(x,y)$ the kernel of $\mathcal{A}^*$.

Assumption 4.2. *All the results stated below hold under the following assumptions.*

(a) $\xi \in \mathbf{L}^2(P \otimes Q)$. As a consequence (Bickel et al., 1998, Appendix A.4), the operator $\mathcal{A}$ is compact. Then the operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ admit eigenvalue decomposition $\mathcal{A}^*\mathcal{A}\alpha_k = s_k^2\alpha_k$ and $\mathcal{A}\mathcal{A}^*\beta_k = s_k^2\beta_k$ for all $k \geq 0$ with $s_0 = 1$, $\alpha_0 = \beta_0 = \mathbf{1}$ and $0 \leq s_k \leq 1$ for all $k \geq 0$. Moreover, it holds that $\mathcal{A}\alpha_k = s_k\beta_k$ and $\mathcal{A}^*\beta_k = s_k\alpha_k$; see (Gohberg et al., 1990, Chapter 6.1). We call $\{s_k\}_{k\geq 0}$ the singular values of $\mathcal{A}$ and $\mathcal{A}^*$, and call $\{\alpha_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq 0}$ the singular functions.

(b) The operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ have positive eigenvalue gap, i.e., *$s_k \leq s_1 < 1$ for all $k \geq 1$. By Jentzsch's Theorem (Rugh, 2010, Theorem 7.2), a sufficient condition is that ξ is bounded.*

Theorem 4.4. *As $n \to \infty$, μ_n converges weakly to μ, in probability. That is, for any $\delta > 0$, we have $\mathbb{P}(D(\mu_n, \mu) > \delta) \to 0$ as $n \to \infty$, where D is the Lévy-Prokhorov metric induced by weak convergence.*

Towards the proof of Theorem 4.4, a critical result is the limit law of the denominator in (4.20) which we denote as D_n. We state it here since it is of independent interest.

Theorem 4.5. *As $n \to \infty$, the denominator in (4.20) has the following limiting distribution:*

$$D_n \to_d D := \frac{1}{\sqrt{\prod_{k=1}^{\infty}(1-s_k^2)}} \exp\left\{\frac{1}{2}\sum_{k=1}^{\infty}\left[-\frac{s_k^2}{1-s_k^2}(U_k^2+V_k^2)+\frac{2s_k}{1-s_k^2}U_kV_k\right]\right\}, \qquad (4.22)$$

where $\{U_k\}_{k\geq 1}$ and $\{V_k\}_{k\geq 1}$ are independent standard normal random variables.

It is noteworthy that D_N is a two-sample U-statistic of infinite order—a generalization of classical U-statistics introduced by Halmos (1946) and Hoeffding (1948a), where the kernel of the U-statistic depends on the sample size. Infinite-order U-statistics were first considered by Halász and Székely (1976) as a special class of elementary symmetric polynomials of random variables; see also (Móri and Székely, 1982; van Es, 1986; van Es and Helmers, 1988; Major, 1999) in this line of research. The limiting distribution of general infinite-order U-statistics was obtained by Dynkin and Mandelbaum (1983, Theorem 1) using randomization of the sample size and multiple Wiener integrals. Theorem 4.5 extends previous work on one-sample infinite-order U-statistics to *two-sample* infinite-order U-statistics.

Another closely related topic is the asymptotics of random permanents; see the monograph (Rempała and Wesołowski, 2007) for a review. An elementary symmetric polynomial is the permanent of a random matrix with identical rows (Rempała and Wesołowski, 2005, Page 2). The limiting behavior of general random permanents has been studied in the case of i.i.d. entries (Rempała and Wesołowski, 1999) as well as independent columns (Rempała and Wesołowski, 2005), where the limit law is the exponential of a Gaussian distribution. The denominator D_N can be viewed as the permanent of the random matrix $(\xi(X_i, Y_j))_{N\times N}$ scaled by $N!$. Hence, Theorem 4.5 characterizes the asymptotic behavior of the permanent of a random matrix induced by a bivariate function whose rows and columns are dependent—the limit law is given by the exponential of a weighted sum of products of Gaussians.

4.5.2 Central limit theorem of the linear functional

To conduct more refined analysis on the convergence of μ_n, we let η be any function on $\mathbb{R}^d \times \mathbb{R}^d$ integrable under μ and investigate the convergence of

$$T_n := T_n(\eta) := \int \eta \mathrm{d}\mu_n = \frac{\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\frac{1}{n}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i})\xi^{\otimes}(X, Y_\sigma)}{\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\xi^{\otimes}(X, Y_\sigma)} \tag{4.23}$$

towards $\theta := \theta(\eta) := \int \eta \mathrm{d}\mu$. A particularly important example is when $\eta = c$ is the cost function and $\mu = \mu_\varepsilon$ is the Schrödinger bridge. In this case T_n is the cost of the discrete Schrödinger bridge and θ is the cost of the Schrödinger bridge, which are used for homogeneity testing in Section 4.4.

The estimator T_n is a rather complicated function of the two empirical distributions P_n and Q_n. Our next result shows that it can be well approximated by linear functions of the two distributions. We further make the following assumptions.

Assumption 4.3. *All the results stated below hold under the following additional assumptions: $\eta^2\xi \in \mathbf{L}^1(P\otimes Q)$ and $\eta\xi \in \mathbf{L}^2(P\otimes Q)$.*

We denote by $I_\nu : \mathbf{L}^2(\nu) \to \mathbf{L}^2(\nu)$ the identity operator on $\mathbf{L}^2(\nu)$, and, by convention, its kernel is given by the Dirac delta function. When the context is clear, we will write I for short. Define

$$\begin{aligned} \kappa_{1,0}(x) &:= \int [\eta(x,y) - \theta]\xi(x,y)\mathrm{d}Q(y) \\ \kappa_{0,1}(y) &:= \int [\eta(x,y) - \theta]\xi(x,y)\mathrm{d}P(x). \end{aligned} \tag{4.24}$$

Theorem 4.6. *It holds that $T_n - \theta = \mathcal{L}_n + o_p\big(1/\sqrt{n}\big)$, where*

$$\mathcal{L}_n := \frac{1}{n}\sum_{i=1}^n [(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_i)].$$

We call $\mathcal{L}_n$ the first order chaos *of T_n. In particular, we have $\sqrt{n}(T_n - \theta) \to_d \mathcal{N}(0, \varsigma^2)$, where*

$$\varsigma^2 := \int (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x)^2 \mathrm{d}P(x) + \int (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y)^2 \mathrm{d}Q(y).$$

The first order chaos $\mathcal{L}_n$ admits a more compact expression using the notion of *tensor product* of operators. Let $\mathcal{A}_1 \in \{\mathcal{A}, \mathcal{A}^*, I_P, I_Q\}$ be an operator mapping from $\mathbf{L}^2(\nu_1)$ to $\mathbf{L}^2(\gamma_1)$ with kernel A_1. And define $\mathcal{A}_2, A_2$ similarly. The tensor product $\mathcal{A}_1 \otimes \mathcal{A}_2 : \mathbf{L}^2(\nu_1 \otimes \nu_2) \to \mathbf{L}^2(\gamma_1 \otimes \gamma_2)$ is defined by, for all $f \in \mathbf{L}^2(\nu_1 \otimes \nu_2)$,

$$(\mathcal{A}_1 \otimes \mathcal{A}_2)f(v_1, v_2) = \iint f(v_1', v_2')A_1(v_1', v_1)A_2(v_2', v_2)\mathrm{d}\nu_1(v_1')\mathrm{d}\nu_2(v_2').$$

For instance, $I_P \otimes \mathcal{A} : \mathbf{L}^2(P \otimes P) \to \mathbf{L}^2(P \otimes Q)$ is defined by

$$\begin{aligned}(I_P \otimes \mathcal{A})f(v_1, v_2) &= \iint f(v_1', v_2')\delta_{v_1}(v_1')\xi(v_2', v_2)\mathrm{d}P(v_1')\mathrm{d}P(v_2') \\ &= \int f(v_1, v_2')\xi(v_2', v_2)\mathrm{d}P(v_2').\end{aligned}$$

In particular, when $f := f_1 \oplus f_2$, we have $(\mathcal{A}_1 \otimes \mathcal{A}_2)(f_1 \oplus f_2)(v_1, v_2) = \mathcal{A}_1 f_1(v_1) + \mathcal{A}_2 f_2(v_2)$. Finally, define the *swap* operator $\mathcal{T}$ by $\mathcal{T}f(u, v) = f(v, u)$ for any f on $\mathbb{R}^d \times \mathbb{R}^d$. It is clear that $\mathcal{T}(\mathcal{A}_1 \otimes \mathcal{A}_2) = (\mathcal{A}_2 \otimes \mathcal{A}_1)\mathcal{T}$ on $\mathbf{L}^2(\nu_1 \otimes \nu_2)$.

Definition 4.2. *Define the following operator on the space* $\mathbf{L}^2(P \otimes Q)$*:*

$$\mathcal{B} := \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*) = (\mathcal{A}^* \otimes \mathcal{A})\mathcal{T}.$$

Remark 4.3. *In terms of this new operator $\mathcal{B}$, the first order chaos of T_n can be alternatively expressed as:*

$$\mathcal{L}_n = \frac{1}{n}\sum_{i=1}^{n}(I + \mathcal{B})^{-1}(\kappa_{1,0} \oplus \kappa_{0,1})(X_i, Y_i).$$

Both expressions come from the following system of linear equations. Assume the first order chaos in Theorem 4.6 is given by $\frac{1}{n}\sum_{i=1}^{n}[f(X_i) + g(Y_i)]$, then f and g are (almost surely) solutions to:

$$\kappa_{1,0}(x) = f(x) + \mathcal{A}^*g(x) \quad \textit{and} \quad \kappa_{0,1}(y) = \mathcal{A}f(y) + g(y).$$

Remark 4.4. *When $\varsigma = 0$ in Theorem 4.6, we can further obtain the second order chaos of T_n. We give the full derivation in Appendix C.5.*

4.6 Proof Sketches

We outline the proof strategies of the results in Section 4.5. The complete proofs can be found in Appendix C. In Section 4.6.1 we prove a novel contiguity result that allows us to change the model to $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu$ based on the limiting distribution of the denominator in Theorem 4.5. This change-of-measure brings in a more natural analysis for μ_n and Theorem 4.4 then follows from the martingale convergence theorem. Next in Section 4.6.2 we derive the first order approximation of T_n and prove Theorem 4.6 by a variance bound of the remainder. We show that this approximation is the first order chaos of T_n under the change of measure μ. Essentially, we extend the classical Hoeffding decomposition for U-statistics to the case of paired samples, which, by itself is a new result in the literature. In Section 4.6.3 we derive the asymptotic distribution of the denominator and the variance bound of the remainder used in the previous two sections. Both of them are two-sample U-statistics of infinite order that have not been studied in the literature. We develop novel techniques for analyzing this type of U-statistics.

4.6.1 Consistency and contiguity

We prove the weak convergence of μ_n in Theorem 4.4. By definition, it suffices to show the convergence of $T_n := \int \eta \mathrm{d}\mu_n$ to $\theta := \int \eta \mathrm{d}\mu$ for any continuous bounded function $\eta : \mathbb{R}^d \to \mathbb{R}$. In fact, the convergence holds for all η that is integrable under μ.

Recall from (4.23) that T_n admits a complicated expression which is difficult to analyze. Thanks to Proposition 4.8 below, it is a simple conditional expectation under a change of measure—we assume that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from μ rather than $P \otimes Q$. Hence, it is natural to ask if there is a way to do analysis under the changed measure μ and carry the results over to the original measure $P \otimes Q$. The contiguity (van der Vaart, 2000, Chapter 6) is exactly a tool for such purposes. When $\xi \neq 1$, the law of the entire i.i.d. sequence $\{(X_i, Y_i)\}_{i=1}^n$ under the two measures $P \otimes Q$ and μ are singular as $n \to \infty$. However, since T_n is symmetric under permutations of X_i's and Y_i's separately, it is measurable w.r.t. the

σ-algebra generated by the pair of empirical measures (P_n, Q_n). Restricted to this σ-algebra, we show that the two measures are mutually contiguous in Theorem 4.7 below.

We first set-up a measure-theoretic framework. We use the term "under the model γ" to indicate that the sample $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \gamma$ and use $\mathbb{E}_\gamma$ to denote the expectation under this model. When $\gamma = P \otimes Q$, we write $\mathbb{E}$ for short. Let $\mathcal{F}_n$ denote the σ-algebra generated by $\{(X_i, Y_i)\}_{i=1}^n$. Let $\mathcal{G}_n$ denote the sub-σ-algebra of $\mathcal{F}_n$ generated by (P_n, Q_n). Let $R^n := (P \otimes Q)^n|_{\mathcal{G}_n}$ and $S^n := \mu^n|_{\mathcal{G}_n}$.

According to Le Cam's first lemma (van der Vaart, 2000, page 88), the contiguity holds true if the likelihood ratio $\mathrm{d}S^n/\mathrm{d}R^n$ converges weakly, under R^n, to a random variable that is almost surely positive. It turns out that $\mathrm{d}S^n/\mathrm{d}R^n$ is precisely D_n, i.e., the denominator of T_n, whose weak limit is almost surely positive according to Theorem 4.5.

Fact 4.5. *The likelihood ratio* $\mathrm{d}S^n/\mathrm{d}R^n$ *admits the following expression:*

$$\frac{\mathrm{d}S^n}{\mathrm{d}R^n} = D_n := \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma). \tag{4.25}$$

Proof. Note that the likelihood ratio of μ^n and $(P \otimes Q)^n$ is given by

$$f_n := \frac{\mathrm{d}\mu^n}{\mathrm{d}(P \otimes Q)^n} = \prod_{i=1}^n \xi(X_i, Y_i), \quad \text{on } \left(\mathbb{R}^d \times \mathbb{R}^d\right)^n. \tag{4.26}$$

Hence, by the property of conditional expectation,

$$\frac{\mathrm{d}S^n}{\mathrm{d}R^n} = \frac{\mathrm{d}\mu^n|_{\mathcal{G}_n}}{\mathrm{d}(P \otimes Q)^n|_{\mathcal{G}_n}} = \mathbb{E}\left[f_n \mid \mathcal{G}_n\right],$$

It follows from exchangeability under $P \otimes Q$ that $\mathbb{E}[f_n \mid \mathcal{G}_n] = \mathbb{E}[\xi^{\otimes}(X, Y_\sigma) \mid \mathcal{G}_n]$ for each $\sigma \in \mathcal{S}_n$. Hence,

$$\mathbb{E}\left[f_n \mid \mathcal{G}_n\right] = \mathbb{E}\left[\frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma) \,\middle|\, \mathcal{G}_n\right] = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma). \tag{4.27}$$

□

Theorem 4.7. *Under Assumption 4.2, the sequences* $(R^n, n \geq 1)$ *and* $(S^n, n \geq 1)$ *are mutually contiguous, i.e.,* $R^n \lhd\rhd S^n$. *Explicitly, for a sequence of events* $(A_n \in \mathcal{G}_n, n \geq 1)$, *we have* $\lim_{n\to\infty} S^n(A_n) = 0$ *iff* $\lim_{n\to\infty} R^n(A_n) = 0$.

Proof. According to Le Cam's first lemma (van der Vaart, 2000, page 88), $R^n \triangleleft S^n$ if and only if the following statement holds true: if D_n, under $P \otimes Q$, converges weakly to D, along a sub-sequence, then $P(D > 0) = 1$. This statement follows directly from Theorem 4.5 (whose proof is deferred to Section 4.6.3), so we have $R^n \triangleleft S^n$. By a standard computation, it can be shown that $\mathbb{E}[D] = 1$. Hence, it follows from Le Cam's first lemma again that $S^n \triangleleft R^n$, that is, R^n and S^n are mutually contiguous. □

With Theorem 4.7 at hand, we can work under the model μ. The next result rewrites T_n as a simple conditional expectation and verifies its consistency whose proof is deferred to Appendix C.2.

Proposition 4.8. *Suppose that* $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$. *For any* $\eta \in \mathbf{L}^1(\mu)$, *it holds that* $T_n = \mathbb{E}\left[\eta(X_1, Y_1) \mid \mathcal{G}_n\right]$. *In particular,* $\mathbb{E}[T_n] = \theta$ *for all* n *and* $\lim_{n\to\infty} T_n = \theta$ *almost surely.*

Proof of Theorem 4.4. As shown in Proposition 4.8, for any $\eta \in \mathbf{L}^1(\mu)$, $T_n := \int \eta \mathrm{d}\mu_n \rightarrow_{a.s.} \theta := \int \eta \mathrm{d}\mu$ under the model μ. In particular, Proposition 4.8 holds for any bounded continuous function η. Thus, except for a null set, the convergence in Proposition 4.8 holds for a countable collection of bounded continuous functions. By separability of $\mathbb{R}^d$, almost sure weak convergence follows (Varadarajan, 1958, Theorem 3.1) by choosing such a countable collection judiciously. This shows almost sure weak convergence under the model μ. Weak convergence in probability under the model $P \otimes Q$ now follows from Theorem 4.7. □

4.6.2 Limit law and chaos decomposition

This subsection is devoted to the limit law of T_n in Theorem 4.4. Following the standard strategy, our goal is to find the first order approximation $\mathcal{L}_n$ of T_n in the form of a sum of i.i.d. terms. Now, provided that the remainder $T_n - \theta - \mathcal{L}_n = o_p(n^{-1/2})$, it follows from the CLT and Slutsky's lemma that $\sqrt{N}(T_n - \theta)$ converges weakly to a normal distribution. However, there are two main challenges. First, the statistic T_n has a rather complicated expression involving a ratio of two infinite-order U-statistics. This prevents us from utilizing the Hoeffding decomposition to derive the first order approximation. Second, due to its complicated nature, it is extremely challenging to control the remainder—the variance computation for classical U-statistics does not apply here.

To address the first challenge, the key observation is that T_n admits a simple expression under the model μ as shown in Proposition 4.8. This allows us to obtain a linear approximation of T_n under the model μ which we call the first order chaos. Due to the contiguity result in Theorem 4.7, the first order chaos can be viewed as the first order approximation of T_n under $P \otimes Q$. As for the second challenge, we develop a novel approach to control the remainder using the spectral gap of the operators $\mathcal{A}$ and $\mathcal{A}^*$. Since this approach is also used to establish the limit law of D_n in Theorem 4.5, we discuss the treatment of D_n and the remainder together in Section 4.6.3.

First order approximation

We first give a informal derivation of the first order approximation $\mathcal{L}_n$ and prove the asymptotic normality of T_n. Recall from Proposition 4.8 that $T_n = \mathbb{E}_\mu[\eta(X_1, Y_1) \mid \mathcal{G}_n]$, where $\mathbb{E}_\mu[\cdot \mid \mathcal{G}_n]$ represents the conditional expectation under μ. Hence, in order to obtain the first order approximation of T_n, it is natural to approximate $\eta(X, Y) - \theta$ by some linear term $f(X) + g(Y)$ under $(X, Y) \sim \mu$ and then use

$$\mathbb{E}_\mu[f(X_1) + g(Y_1) \mid \mathcal{G}_n] = \frac{1}{n}\sum_{i=1}^{n}[f(X_i) + g(Y_i)]$$

as the first order approximation of T_n. A good linear approximation $f(X) + g(Y)$ should satisfy

$$\begin{aligned} \mathbb{E}_\mu[\eta(X,Y) - \theta \mid X] &= \mathbb{E}_\mu[f(X) + g(Y) \mid X] \\ \mathbb{E}_\mu[\eta(X,Y) - \theta \mid Y] &= \mathbb{E}_\mu[f(X) + g(Y) \mid Y]. \end{aligned} \tag{4.28}$$

Recall that $\frac{\mathrm{d}\mu}{\mathrm{d}(P\otimes Q)}(x, y) = \xi(x, y)$ and $\kappa_{1,0}$ from (4.24). It holds that

$$\mathbb{E}_\mu[\eta(X,Y) - \theta \mid X](x) = \int [\eta(x,y) - \theta]\xi(x,y)\mathrm{d}Q(y) = \kappa_{1,0}(x).$$

Similarly, we have $\mathbb{E}_\mu[\eta(X,Y)-\theta \mid Y](y) = \kappa_{0,1}(y)$. Moreover, by Definition 4.1, we obtain

$$\begin{aligned}\mathbb{E}_\mu[g(Y) \mid X](x) &= \int g(y)\xi(x,y)\mathrm{d}Q(y) = (\mathcal{A}^* g)(x) \\ \mathbb{E}_\mu[f(X) \mid Y](y) &= \int f(x)\xi(x,y)\mathrm{d}P(x) = (\mathcal{A} f)(y).\end{aligned} \tag{4.29}$$

As a result, the condition (4.28) becomes

$$\kappa_{1,0}(X) = f(X) + \mathcal{A}^* g(X) \quad \text{and} \quad \kappa_{0,1}(Y) = \mathcal{A} f(Y) + g(Y). \tag{4.30}$$

Formally, we can solve the linear system (4.30) to get

$$f = (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1}) \quad \text{and} \quad g = (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0}).$$

We will make this rigorous later in this section. This suggests that the first order approximation of T_n should be

$$\frac{1}{n}\sum_{i=1}^n \left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_i)\right]$$

which is exactly the first order chaos $\mathcal{L}_n$ in Theorem 4.6. In fact, the next result shows that, after subtracting $\mathcal{L}_n$ from $T_n - \theta$, the variance of the numerator is of order $O(n^{-2})$.

By some standard algebra, it can be shown that the remainder $T_n - \theta - \mathcal{L}_n = U_n/D_n$, where D_n is defined in (4.25) and

$$U_n := \frac{1}{n \cdot n!}\sum_{\sigma \in \mathcal{S}_n}\sum_{i=1}^n \tilde{\eta}(X_i, Y_{\sigma_i})\xi^{\otimes}(X, Y_\sigma) \tag{4.31}$$

with $\tilde{\eta}(x,y)$ defined as

$$\eta(x,y) - \theta - (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x) - (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y). \tag{4.32}$$

Proposition 4.9. *Under Assumptions 4.2 and 4.3, we have* $\mathbb{E}[U_n^2] = O(n^{-2})$.

Similar to D_n, the numerator U_n is also a two-sample U-statistic of infinite order. We again defer the proof of Proposition 4.9 to Section 4.6.3. Let us prove the main result.

Proof of Theorem 4.6. According to Theorem 4.5 and Proposition 4.9, we have $D_n = O_p(1)$ and $U_n = o_p(n^{-1/2})$. By Slutsky's Lemma, it holds that $T_n - \theta - \mathcal{L}_n = U_n/D_n = o_p(n^{-1/2})$. Now, the asymptotic normality follows from the standard Lindeberg CLT (Billingsley, 1995, Section 27). □

Chaos decomposition for paired samples

We then derive $\mathcal{L}_n$ rigorously as the first order chaos of T_n. The key technique used to analyze U-statistics is the Hoeffding decomposition. Given an independent sample, it decomposes the statistic into terms of increasing complexity by projecting the statistic onto orthogonal $\mathbf{L}^2$ subspaces spanned by subsets of the sample. However, under the model μ, X_i and Y_i are dependent for each $i \in [n]$. Moreover, the statistic is not explicit but expressed as a conditional expectation on $\mathcal{G}_n := \sigma(P_n, Q_n)$. To address these challenges, we consider what we call the chaos decomposition where each term in the expansion is a polynomial function of (P_n, Q_n). This decomposition can be computed using orthogonal projections in $\mathbf{L}^2(\mu^n)$.

We change throughout this section the measure so that $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mu$. Let $H_0 \subset \mathbf{L}^2(\mu^n)$ be the subspace spanned by constant functions and $H_1 \subset \mathbf{L}^2(\mu^n)$ be the subspace spanned by functions of the type

$$\sum_{i=1}^n [f_{1,0}(X_i) + f_{0,1}(Y_i)] \tag{4.33}$$

that is orthogonal to H_0. It is clear that the (orthogonal) projection of T_n onto H_0 is given by $\mathrm{Proj}_{H_0}(T_n) = \theta$. It turns out that $\mathcal{L}_n$ is the projection of T_n onto H_1 (see Appendix C.3 for the proof), which we refer to as the *first order chaos*. Note that the elements in $\mathbf{L}^2$ spaces are only defined up to zero-measure sets (or equivalent classes). For two elements f and g in $\mathbf{L}^2$, $f = g$ should be understood as f equals g up to equivalent classes.

Proposition 4.10. *Under Assumptions 4.2 and 4.3, the first order chaos of the statistic T_n is given by*

$$\mathcal{L}_n := \frac{1}{n} \sum_{i=1}^n [(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_i)].$$

We then derive a more compact expression of $\mathcal{L}_n$ using the operator $\mathcal{B}$ in Definition 4.2. We start by providing an identity regarding $\mathcal{A}$ and $\mathcal{B}$ in the following lemma. Given a probability measure ν, let $\mathbf{L}^2_0(\nu)$ be the subspace of $\mathbf{L}^2(\nu)$ consisting of mean-zero functions.

Lemma 4.11. *Under Assumption 4.2, for any $f \in \mathbf{L}_0^2(P)$ and $g \in \mathbf{L}_0^2(Q)$, it holds that*

$$(I + \mathcal{B})^{-1}(f \oplus g) = [(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)] \oplus [(I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f)]. \tag{4.34}$$

Corollary 4.12. *Under Assumptions 4.2 and 4.3, the first order chaos of T_n admits an alternative expression $\mathcal{L}_n = \frac{1}{n}\sum_{i=1}^n (I + \mathcal{B})^{-1}(\kappa_{1,0} \oplus \kappa_{0,1})(X_i, Y_i)$.*

Remark 4.6. *Note that the above expression of $\mathcal{L}_n$ is permutation symmetric, i.e., $\sum_{i=1}^n (I + \mathcal{B})^{-1}(\kappa_{1,0} \oplus \kappa_{0,1})(X_i, Y_i) = \sum_{i=1}^n (I + \mathcal{B})^{-1}(\kappa_{1,0} \oplus \kappa_{0,1})(X_i, Y_{\sigma_i})$ for all $\sigma \in \mathcal{S}_n$.*

Remark 4.7. *Another way to see this is: due to (4.30), $\kappa_{1,0} \oplus \kappa_{0,1} = f \oplus g + \mathcal{A}^* g \oplus \mathcal{A} f = (I + \mathcal{B})(f \oplus g)$.*

4.6.3 Analysis of the denominator and the remainder

Recall from Section 4.6.2 that the first order remainder $R_1 := T_n - \theta - \mathcal{L}_n = U_n / D_n$, where

$$U_n := \frac{1}{n \cdot n!} \sum_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n \tilde{\eta}(X_i, Y_{\sigma_i}) \xi^{\otimes}(X, Y_\sigma) \quad \text{and} \quad D_n := \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \xi^{\otimes}(X, Y_\sigma), \tag{4.35}$$

with $\tilde{\eta}(x, y)$ defined in (4.32). We will prove the limit law of D_n in Theorem 4.5 and the variance bound of U_n in Proposition 4.9. The strategy is to decompose D_n and U_n into orthogonal pieces using the Hoeffding decomposition, and then bound the higher order terms using the spectral gap of the operators $\mathcal{A}$ and $\mathcal{A}^*$. Note that both D_n and U_n can be viewed as *two-sample* U-statistics of infinite order. Techniques for U-statistics of fixed order and one-sample U-statistics of infinite order do not directly apply here. We develop new techniques for such U-statistics.

We work throughout this section with the original model assuming that $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P \otimes Q$ and use $\mathbb{E}$ to denote the expectation. We first derive the Hoeffding decomposition of D_n and U_n. The proof can be found in Appendix C.4.1. We denote $\tilde{\xi} := \xi - 1$ and $h := \tilde{\eta}\xi$.

Proposition 4.13. *The following Hoeffding decompositions hold:*

$$\begin{aligned} D_n &= 1 + \sum_{\substack{A,B\subset[n]\\|A|=|B|>0}} \frac{1}{n!} \sum_{\sigma\in\mathcal{S}_n:\sigma_A=B} \prod_{i\in A} \tilde{\xi}(X_i, Y_{\sigma_i}) \\ U_n &= \sum_{\substack{A,B\subset[n]\\|A|=|B|>0}} \frac{1}{n\cdot n!} \sum_{\sigma\in\mathcal{S}_n:\sigma_A=B} \sum_{i\in A} h(X_i, Y_{\sigma_i}) \prod_{j\in A\setminus\{i\}} \tilde{\xi}(X_j, Y_{\sigma_j}), \end{aligned} \tag{4.36}$$

where $\sigma_A := \{\sigma_i : i \in A\}$*. Moreover,*

$$\begin{aligned} \mathbb{E}[D_n^2] &= 1 + \sum_{r=1}^{n} \sum_{\sigma\in\mathcal{S}_r} \mathbb{E}\Big[\prod_{j=1}^{r} \tilde{\xi}(X_j, Y_j)\tilde{\xi}(X_j, Y_{\sigma_j})\Big] \\ \mathbb{E}[U_n^2] &= \frac{1}{n^2}\sum_{r=1}^{n} \frac{r}{r!} \sum_{\sigma\in\mathcal{S}_r} \sum_{i=1}^{r} \mathbb{E}\Big[h(X_1, Y_1)\prod_{j=2}^{r} \tilde{\xi}(X_j, Y_j) h(X_i, Y_{\sigma_i}) \prod_{j\in[r]\setminus\{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\Big]. \end{aligned}$$

We then bound the variance of D_n and U_n using the spectral gap of operators $\mathcal{A}$ and $\mathcal{A}^*$. Assumption 4.2 guarantees that such spectral gap does exist. We start with a contraction property; see Appendix C.4.2 for a proof. For $\nu_1, \nu_2 \in \mathcal{M}_1(\mathbb{R}^d)$, we let $\mathbf{L}^2_{0,0}(\nu_1 \otimes \nu_2)$ be the set of functions such that

$$\mathbb{E}[f(Z_1, Z_2) \mid Z_1] \overset{\text{a.s.}}{=} 0 \quad \text{and} \quad \mathbb{E}[f(Z_1, Z_2) \mid Z_2] \overset{\text{a.s.}}{=} 0$$

where $(Z_1, Z_2) \sim \nu_1 \otimes \nu_2$.

Lemma 4.14. *Recall* s_1 *from Assumption 4.2. For any* $f \in \mathbf{L}^2_{0,0}(P\otimes P)$*, we have* $(I_P\otimes\mathcal{A})f \in \mathbf{L}^2_{0,0}(P\otimes Q)$ *and* $\|(I_P \otimes \mathcal{A})f\|_{\mathbf{L}^2(P\otimes Q)} \le s_1 \|f\|_{\mathbf{L}^2(P\otimes P)}$*. Similar results hold for* $I_P\otimes\mathcal{A}^*$*,* $\mathcal{A}\otimes I_Q$ *and* $\mathcal{A}^* \otimes I_Q$*.*

According to Proposition 4.13, the key quantity in the variances of D_n and U_n is

$$\mathbb{E}\left[f(X_1, Y_1)\prod_{j=2}^{n} \tilde{\xi}(X_j, Y_j) f(X_i, Y_{\sigma_i}) \prod_{j\in[n]\setminus\{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right] \tag{4.37}$$

for some $f \in \mathbf{L}^2(P\otimes Q)$, where $f = \tilde{\xi} = \xi - 1$ for D_n and $f = h = \tilde{\eta}\xi$ for U_n. In order to control it, we decompose a permutation into disjoint cycles. By independence, the expectation then equals the product of expectations w.r.t. each cycle. We then simplify each expectation by iteratively integrating w.r.t. a single variable, while keeping the rest of them being fixed. This procedure brings about operators in Lemma 4.14. Applying their contraction property then gives the following bound for (4.37). We defer its proof to Appendix C.4.2.

Lemma 4.15. *Let* $\varsigma_0 := \|\tilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)}$ *and* $\varsigma_h := \|h\|_{\mathbf{L}^2(P\otimes Q)}$. *Under Assumption 4.2, for any* $n \in \mathbb{N}_+$, $\sigma \in \mathcal{S}_n$ *and* $i \in [n]$, *we have*

$$\mathbb{E}\left[h(X_1,Y_1)\prod_{j=2}^{n}\tilde{\xi}(X_j,Y_j)h(X_i,Y_{\sigma_i})\prod_{j\in[n]\setminus\{i\}}\tilde{\xi}(X_j,Y_{\sigma_j})\right] \le s_1^{2(n-\#\sigma)}\varsigma_h^2\varsigma_0^{2(\#\sigma-1)},$$

where $\#\sigma$ *is the number of cycles of the permutation* σ.

Now we are ready to prove Proposition 4.9.

Proof of Proposition 4.9. Recall from Proposition 4.13 that

$$\mathbb{E}[U_n^2] = \frac{1}{n^2}\sum_{r=1}^{n}\frac{r}{r!}\sum_{\sigma\in\mathcal{S}_r}\sum_{i=1}^{r}\mathbb{E}\left[h(X_1,Y_1)\prod_{j=2}^{r}\tilde{\xi}(X_j,Y_j)h(X_i,Y_{\sigma_i})\prod_{j\in[n]\setminus\{i\}}\tilde{\xi}(X_j,Y_{\sigma_j})\right].$$

By Lemma 4.15, we know

$$\mathbb{E}[U_n^2] \le \frac{1}{n^2}\sum_{r=1}^{n}\frac{r^2}{r!}\sum_{\sigma\in\mathcal{S}_r}s_1^{2(r-\#\sigma)}\varsigma_0^{2(\#\sigma-1)}\varsigma^2, \tag{4.38}$$

where $\varsigma_0 := \|\tilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)}$, $\varsigma := \|\tilde{\eta}\xi\|_{\mathbf{L}^2(P\otimes Q)}$, and $\#\sigma$ is the number of cycles of σ.

Now, let σ^* be a random permutation uniformly sampled from $\mathcal{S}_r$. It is well-known (Arratia et al., 2003, Chapter 1) that the moment generating function of $\#\sigma^*$ is given by $\mathbb{E}[u^{\#\sigma^*}] = \prod_{i=1}^{r}(1-\frac{1}{i}+\frac{u}{i})$. Thus,

$$\frac{r^2}{r!}\sum_{\sigma\in\mathcal{S}_r}s_1^{2(r-\#\sigma)}\varsigma_0^{2(\#\sigma-1)} = r^2\,\mathbb{E}\left[s_1^{2(r-\#\sigma^*)}\varsigma_0^{2(\#\sigma^*-1)}\right] = r^2 s_1^{2r}\varsigma_0^{-2}\prod_{i=1}^{r}\left(1-\frac{1}{i}+\frac{\varsigma_0^2}{s_1^2 i}\right).$$

Let $m := \lceil \varsigma_0^2/s_1^2 - 1\rceil$. Then, for every $r \ge m$,

$$\prod_{i=1}^{r}\left(1-\frac{1}{i}+\frac{\varsigma_0^2}{s_1^2 i}\right) \le \prod_{i=1}^{r}(1+m/i) = \frac{\prod_{i=1}^{r}(i+m)}{r!} = \frac{\prod_{i=r-m}^{r}(i+m)}{m!} \le \frac{(r+m)^m}{m!},$$

and thus

$$\sum_{r=m}^{n} \frac{r^2}{r!} \sum_{\sigma \in \mathcal{S}_r} s_1^{2(r-\#\sigma)} \varsigma_0^{2(\#\sigma-1)} \leq \sum_{r=m}^{n} \frac{1}{m!\varsigma_0^2} r^2 (r+m)^m s_1^{2r}$$

converges as $n \to \infty$. It follows from (4.38) that $\mathbb{E}[U_n^2] = O(N^{-2})$. □

With the same proof techniques, a similar result holds for D_n. Recall from Proposition 4.13 that $D_n = 1 + \sum_{r=1}^{n} D_{n,r}$ where

$$D_{n,r} := \frac{1}{n!} \sum_{|A|=|B|=r} \sum_{\sigma \in \mathcal{S}_n : \sigma_A = B} \prod_{i \in A} \tilde{\xi}(X_i, Y_{\sigma_i}). \tag{4.39}$$

Proposition 4.16. *Under Assumptions 4.2 and 4.3, we have, for any integer $R \geq 0$,*

$$\mathbb{E}\left[\left(D_n - 1 - \sum_{r=1}^{R} D_{n,r}\right)^2\right] \leq \sum_{r=R+1}^{n} \frac{1}{r!} \sum_{\sigma \in \mathcal{S}_r} s_1^{2(r-\#\sigma)} \varsigma_0^{2\#\sigma},$$

which can be arbitrarily small as $R \to \infty$.

Finally, we establish the limiting distribution of D_n. For any integer $R \geq 1$, the finite sum $1 + \sum_{r=1}^{R} D_{n,r}$ is a two-sample U-statistic of order R whose asymptotic distribution, given by Gaussian chaoses (i.e., Hermite polynomials of independent Gaussians), can be obtained using standard argument in the literature; see, e.g., (Serfling, 1980a, Chapter 5.5.2). Note that the variance of the remainder $D_n - 1 - \sum_{r=1}^{R} D_{n,r}$ can be arbitrarily small as $R \to \infty$ by Proposition 4.16. Now Theorem 4.5 follows from expanding D in terms of Hermite polynomials. The full proof can be found in Appendix C.4.3.

4.7 Experiments

In this section, we apply the (centered) Schrödinger bridge (SCB) statistic to test for homogeneity on both synthetic and real data. We compare its type I error rate and statistical power with the centered discrete EOT considered by Ramdas et al. (2017) and the maximum mean discrepancy (MMD) proposed by Gretton et al. (2012). For comparison purposes, all the thresholds are determined by permutation test: we 1) randomly permute the pooled

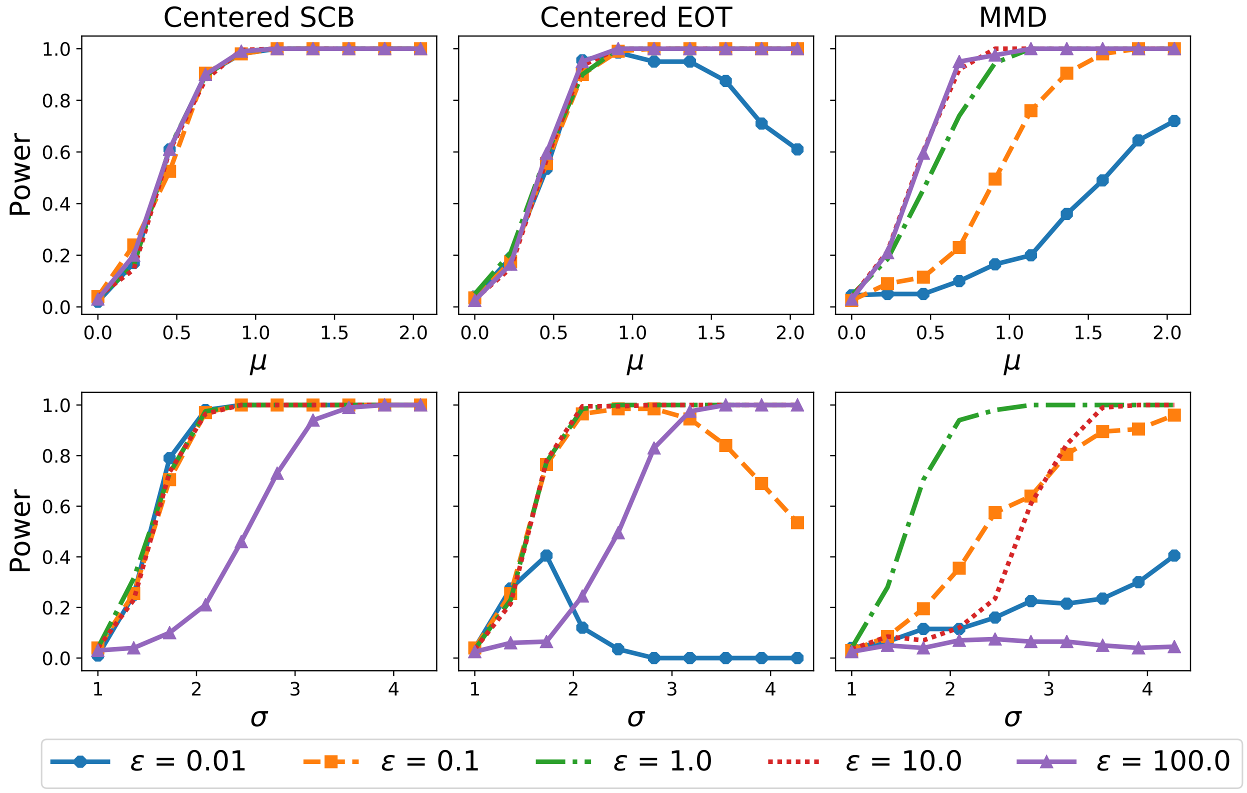

Figure 4.2: Statistical power versus parameter for a pair of Gaussian distributions. **Top**: $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$; **Bottom**: $\mathcal{N}(0,1)$ and $\mathcal{N}(0,\sigma^2)$.

sample $(X_1,\ldots,X_n,Y_1,\ldots,Y_n)$ and split them equally into two subsets, 2) compute the test statistic for the permuted sample, and 3) repeat previous two steps for 500 times and choose the threshold as the upper 5-percentile of these statistics. This procedure guarantees that the type I error rates of the three tests are all close to 0.05. The code to reproduce the experiments is available online (dsb, 2022).

4.7.1 Synthetic data

Settings. We consider 4 different pairs of distributions: 1) $\mathcal{N}(0,1)$ v.s. $\mathcal{N}(\mu,1)$, 2) $\mathcal{N}(0,1)$ v.s. $\mathcal{N}(0,\sigma^2)$, 3) $\mathrm{VM}(0,1)$ v.s. $\mathrm{VM}(\mu,1)$, and 4) $\mathrm{VM}(0,1)$ v.s. $\mathrm{VM}(0,\kappa)$, where $\mathrm{VM}(\mu,\kappa)$ is

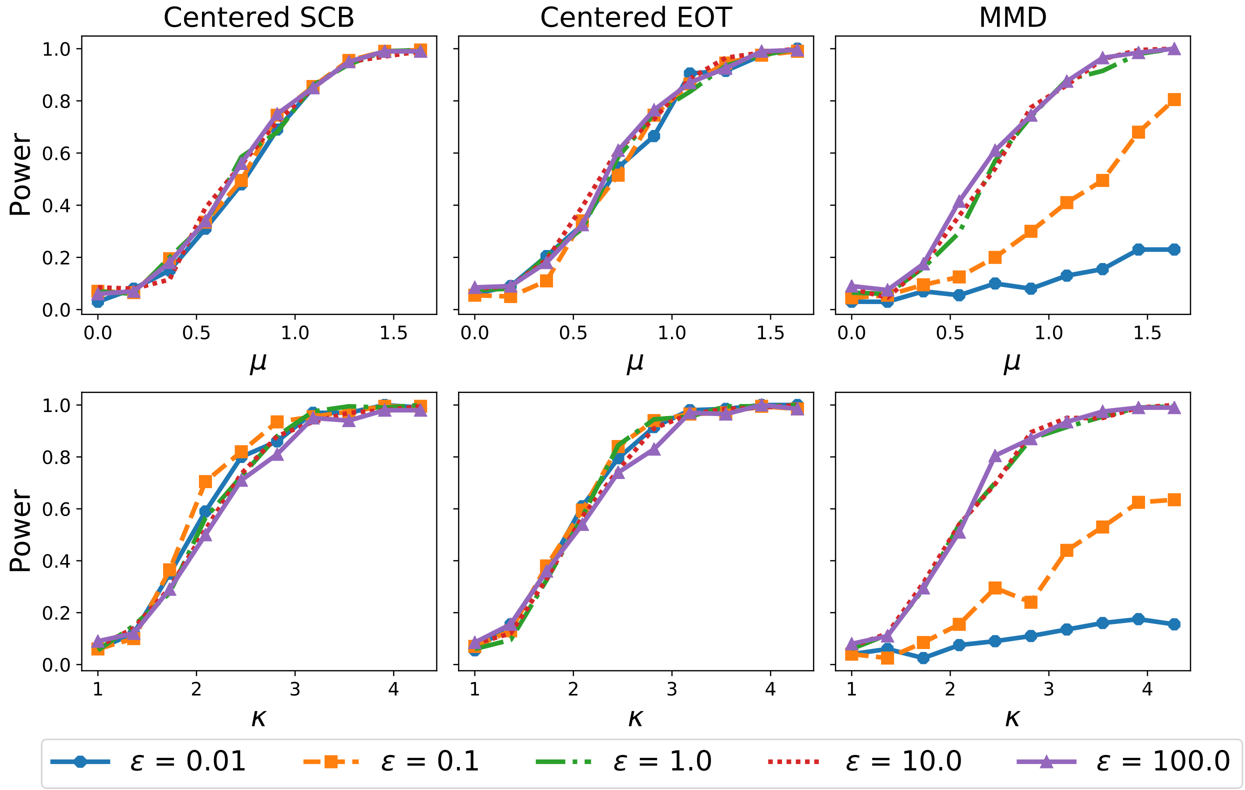

Figure 4.3: Statistical power versus parameter for a pair of Gaussian distributions. **Top**: $\mathrm{VM}(0, 1)$ and $\mathrm{VM}(\mu, 1)$; **Bottom**: $\mathrm{VM}(0, 1)$ and $\mathrm{VM}(0, \kappa)$.

the von Mises distribution with location μ and concentration κ. For each pair of distributions (P, Q), we independently generate $n = 50$ i.i.d. observations from each of the distributions. Then we perform the three tests and store their decisions. For the SCB test and EOT test, we use the quadratic cost and set $\varepsilon \in \{0.01, 0.1, 1, 10, 100\}$. For the MMD test, we use the RBF kernel $k(x, x') = \exp(-\|x - x'\|^2 / \varepsilon)$ and set $\varepsilon \in \{0.01, 0.1, 1, 10, 100\}$. We repeat the whole procedure 200 times and compute the rejection frequency. We plot the rejection frequency as we vary the parameter (e.g., μ in the first pair). When $P = Q$, the rejection frequency is an estimate of the type I error rate; when $P \neq Q$, it is an estimate of the statistical power.

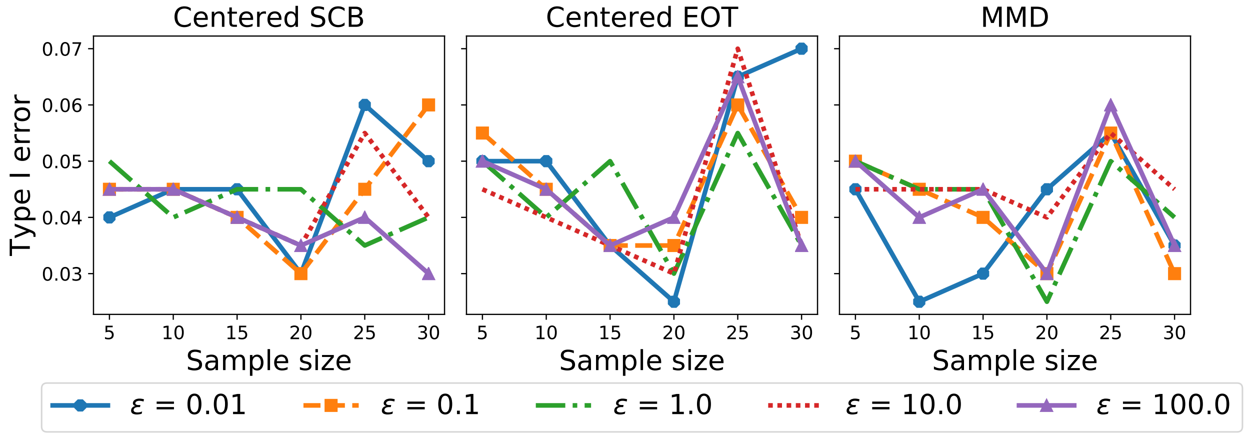

Figure 4.4: Type I error rate versus sample size for digits 3 and 3.

Normal distribution. The results for normal distributions are in Figure 4.2. When the two distributions differ in mean, the SCB test demonstrates similar performance across different values of ε. The EOT test shows a similar behavior except for $\varepsilon = 0.01$: the statistical power increases in the beginning and then decreases as μ increases. This decline is due to the computational instability of the Sinkhorn algorithm used to compute the EOT when ε is relatively small. As for the MMD test, its performance largely depends on the parameter in the RBF kernel. The three tests perform analogously with their best parameter. When the two distributions differ in variance, most of the findings are the same. The parameter $\varepsilon = 100$ gives significantly worse performance and the instability issue in the EOT test is more prominent.

Von Mises distribution. The results for von Mises distributions are in Figure 4.3. The SCB test and EOT test performs similarly without the instability issue. The performance of the MMD test heavily depends on the parameter ε, and its statistical power with the best parameter is close to the ones of the SCB test and the EOT test.

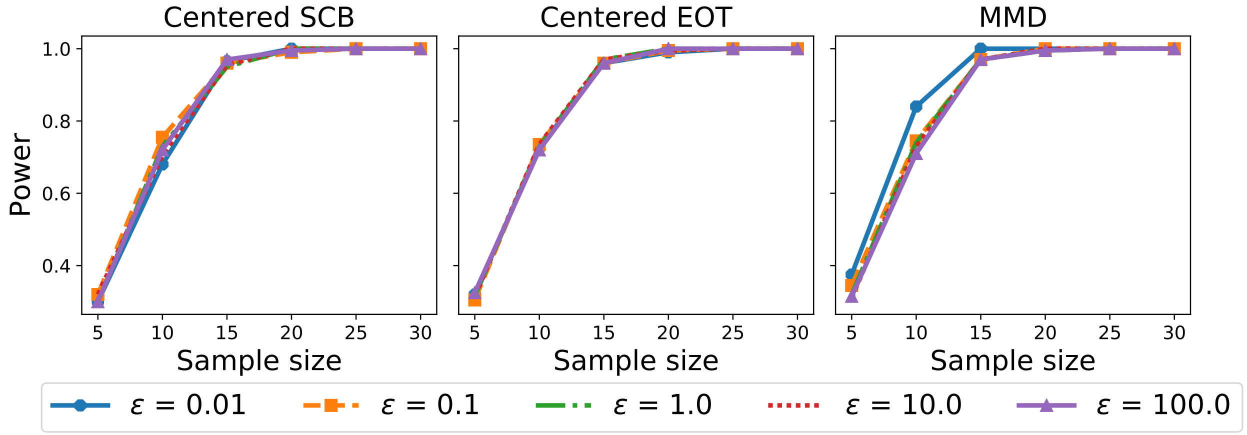

Figure 4.5: Statistical power versus sample size for digits 3 and 5.

4.7.2 Real data

Settings. We compare the three tests on the MNIST dataset (LeCun et al., 1998). Given two digits m_1 and m_2, we randomly sample $n \in \{5, 10, 15, 20, 25, 30\}$ images from each of the two classes. For the SCB test and EOT test, we define the cost on images as follows: for each image M, we normalize it and view it as a discrete distribution; the cost between two images is then chosen as the Wasserstein-2 distance between the corresponding discrete distributions. Again, we set the regularization parameter $\varepsilon \in \{0.01, 0.1, 1, 10, 100\}$. For the MMD test, we use $k(M_1, M_2) := \exp(-c(M_1, M_2)/\varepsilon)$ as the kernel on images, where c is the cost defined above. We repeat the whole procedure 200 times and plot the rejection frequency as we vary the sample size.

Results. The results for $m_1 = m_2 = 3$ is shown in Figure 4.4. The type I error rate of all the tests are close to 0.05 with different parameters. The results for $m_1 = 3$ and $m_2 = 5$ is presented in Figure 4.5. All the tests performs similarly with the MMD test with $\varepsilon = 0.01$ being slightly better. All the tests achieve power 1 with a relatively small sample size, and their performance is robust to the value of parameters considered in the experiments.

Chapter 5

OPTIMAL TRANSPORT DISTANCES FOR MEASURING INDEPENDENCE

5.1 Introduction

Statistical independence measures have been widely used in machine learning and statistics, ranging from independent component analysis (Bach and Jordan, 2002; Gretton et al., 2005) to causal inference (Pfister et al., 2018; Chakraborty and Zhang, 2019), and recently in self-supervised learning (Li et al., 2021) and representation learning (Ozair et al., 2019). Classical dependence measures such as Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ (Hoeffding, 1948b; Kruskal, 1958; Lehmann, 1966) focus on real-valued one dimensional random variables and thus are not suitable for high dimensional data; see also (Schweizer and Wolff, 1981; Nikitin, 1995). Modern dependence measures designed for high-dimensional applications rely heavily on statistical divergences to compare the joint distribution and the product of marginals.

One popular approach to compare distributions is to embed them into reproducing kernel Hilbert spaces (Gretton et al., 2007a, 2012), leading to the Hilbert-Schmidt independence criterion (HSIC) and the associated independence test (Gretton et al., 2005, 2007b). Several extensions of HSIC are available, such as a relative dependency measure (Bounliphone et al., 2015) and a joint independence measure among multiple random elements (Pfister et al., 2018). Another approach is to compare distributions defined on Euclidean spaces via their characteristic functions or the energy distance (Székely and Rizzo, 2004), leading to the distance covariance (dCov) of Székely et al. (2007). It was later generalized to metric spaces of negative type by Lyons (2013). In fact, in their most general form, HSIC and dCov are equivalent as shown by Sejdinovic et al. (2013). Their corresponding empirical estimators all admit a U-statistics expression, and enjoy a convergence rate that is independent of the dimension. These results can be established using tools from U-statistics theory (see, e.g., Serfling, 1980b).

A different line of research explored optimal transport to measure dependence. The Wasserstein distance naturally defines a dependence measure when it is used to quantify the dissimilarity between the joint distribution and the product of marginals (see, e.g., Cifarelli and Regazzini, 2017). The normalized version—the so-called Wasserstein correlation coefficient—has recently gained attention in Mordant and Segers (2021); Nies et al. (2021); Wiesel (2021). Following the classical rank-based tests such as Pearson's ρ, optimal transport is also used to define multivariate ranks and the subsequent independence tests (Shi et al., 2020; Deb and Sen, 2021). However, these tests can suffer from the curse of dimensionality (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019; Lei, 2020) or high computational complexity (Peyré and Cuturi, 2019), limiting their practical usefulness.

A remedy to this challenge is to use the entropy regularized formulation of optimal transport. This is particularly attractive from both a computational viewpoint (Cuturi, 2013) and a statistical viewpoint (Rigollet and Weed, 2018). Moreover the empirical counterpart of entropy regularized optimal transport enjoys as an estimator a parametric rate of convergence and thus appears to overcome the curse of dimensionality (Genevay et al., 2019; Mena and Weed, 2019). The centered version, the Sinkhorn divergence (Feydy et al., 2019), defines a semi-metric on probability measures which metrizes weak convergence. Ramdas et al. (2017) used it for two-sample testing and Genevay et al. (2018) for generative modeling; see also Salimans et al. (2018); Sanjabi et al. (2018).

The independence criterion we propose uses entropy regularized optimal transport to compare the joint distribution and the product of marginals. The empirical counterpart involves a product of two empirical measures, leading to a two-sample U-process on paired samples. The resulting U-process requires a sophisticated analysis of its statistical behavior; common tools from empirical processes are ineffective here. Using the decoupling technique (Peña and Giné, 1999) and duality theory (Peyré and Cuturi, 2019), we prove a rate of convergence roughly $O(\sigma^{3d}n^{-1/2})$, where n is the sample size, d is the ambient dimension, and σ is the sub-Gaussian parameter, recovering previous results for two sample statistics.

The remainder of this chapter is organized as follows. In Section 5.2 we recall several definitions preliminary to our results. In Section 5.3 we review the entropy-regularized optimal transport problem with a focus on its computational aspects. In Section 5.4 we introduce the entropy-regularized optimal transport independence criterion (ETIC) and discuss its key properties. We propose the Tensor Sinkhorn algorithm with a random feature approximation to compute ETIC, which admits a quadratic scaling in time and space. We also show how to approximate ETIC using random features, and how to differentiate through ETIC in a framework of differentiable programming. In Section 5.5, we give our main theoretical results, i.e., non-asymptotic bounds, characterizing the statistical behavior of the empirical estimator of ETIC under both the null and alternative hypotheses. These results, derived from U-process theory and optimal transport theory tools, extend previous ones from a single measure to tensor products of measures. In Section 5.6, we compare the empirical behavior of ETIC with HSIC on both synthetic and real data.

5.2 Preliminaries

We introduce the independence testing problem and explain its connection to the homogeneity testing problem. We then define a few existing independence tests which we use in our experiments. Finally, we briefly recall the notion of metric entropy from the empirical process theory which is used in our proofs.

5.2.1 Independence testing

Let $Z := (X, Y)$ be a pair of random vectors from a distribution μ_{XY} on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with marginals μ_X and μ_Y. Given i.i.d. copies $\{Z_i := (X_i, Y_i)\}_{i=1}^n$ of Z, we are interested in determining whether X and Y are independent or not. This can be formulated as an independence testing problem:

$$\mathbf{H}_0 : X \perp\!\!\!\perp Y \leftrightarrow \mathbf{H}_1 : X \not\!\perp\!\!\!\perp Y. \tag{5.1}$$

That is, we test the null hypothesis that X and Y are independent against the alternative hypothesis that they are not.

The procedure for solving this independence testing problem is essentially the same as the one for homogeneity testing in Section 4.4.1. The only difference is that, rather than quantifying the distance between the marginals μ_X and μ_Y, we choose an independence criterion T such that $T(X, Y)$ measures the dependence between X and Y, that is, the "more dependent" X and Y are, the larger $T(X, Y)$ should be. To be more precise, we give a definition of a valid independence criterion.

Definition 5.1. *Let $\mathcal{P}$ be a set of distributions on $\mathcal{X} \times \mathcal{Y}$. We say an independence criterion T is valid if, for any $(X, Y) \sim \mu_{XY} \in \mathcal{P}$, it holds that $T(X, Y) \geq 0$ and $T(X, Y) = 0$ iff $X \perp\!\!\!\perp Y$.*

The independence testing problem (5.1) is closely related to the homogeneity testing problem. To see this connection, we note that $X \perp\!\!\!\perp Y$ if and only if $\mu_{XY} = \mu_X \otimes \mu_Y$. Consequently, we can rewrite the problem (5.1) in its equivalent form:

$$\mathbf{H}_0 : \mu_{XY} = \mu_X \otimes \mu_Y \leftrightarrow \mathbf{H}_1 : \mu_{XY} \neq \mu_X \otimes \mu_Y. \tag{5.2}$$

In other words, determining the independence between X and Y is equivalent to determining the equality between the joint distribution μ_{XY} and the product of marginals $\mu_X \otimes \mu_Y$, where both of them are measures on $\mathcal{Z}$. Now, a natural way to measure the independence between X and Y is to quantify the distance between μ_{XY} and $\mu_X \otimes \mu_Y$.

Remark 5.1. *The independence testing problem is not equivalent to the homogeneity testing problem. In homogeneity testing, we have two independent i.i.d. samples from the two distributions of interest. However, in the equivalent formulation (5.2) of the independence testing problem, the two distributions of interest are μ_{XY} and $\mu_X \otimes \mu_Y$ with samples $\{(X_i, Y_i)\}_{i=1}^n$*

and $\{(X_i, Y_j)\}_{i \neq j}$, respectively. The second sample is not an i.i.d. sample and the two samples are not independent. This imposes challenges in the statistical analysis as we will see in Section 5.5.

5.2.2 Kernel based independence criterion

A popular choice of independence criterion in high dimension is the Hilbert-Schmidt independence criterion proposed by Gretton et al. (2005) which we recall here. Let $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_2 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two psd kernels. Due to Steinwart and Christmann (2008, Lemma 4.6),

$$k((x, y), (x', y')) := k_1(x, x')k_2(y, y')$$

is a psd kernel on the product space $\mathcal{X} \times \mathcal{Y}$. The *Hilbert-Schmidt independence criterion (HSIC)* between X and Y is then defined as the MMD between μ_{XY} and $\mu_X \otimes \mu_Y$ with kernel k, i.e.,

$$\begin{aligned} \mathrm{HSIC}(X, Y) &:= \mathrm{HSIC}(X, Y)_{k_1,k_2} \\ &:= \mathrm{MMD}_k(\mu_{XY}, \mu_X \otimes \mu_Y) = \int k \mathrm{d}[(\mu_{XY} - \mu_X \otimes \mu_Y) \otimes (\mu_{XY} - \mu_X \otimes \mu_Y)]. \end{aligned}$$

Following Smola et al. (2007, Section 2.3), it can be expanded as

$$\begin{aligned} \mathrm{HSIC}(X, Y) = \mathbb{E}[k_1(X, X')k_2(Y, Y')] + \mathbb{E}[k_1(X, X')]\,\mathbb{E}[k_2(Y, Y')] \\ - 2\,\mathbb{E}[\mathbb{E}[k_1(X, X') \mid X]\,\mathbb{E}[k_2(Y, Y') \mid Y]], \end{aligned} \tag{5.3}$$

where (X', Y') is an independent copy of (X, Y). When $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces, it is shown by Gretton et al. (2005, Theorem 6) that HSIC is a valid independence criterion on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ when both k_1 and k_2 are universal. Here universality is defined by Steinwart (2001, Definition 4) which we recall below; see also (Steinwart and Christmann, 2008, Chapter 4) for more background materials on universal kernels.

Definition 5.2. *A continuous kernel k on a compact metric space $\mathcal{X}$ is called universal if the space of all functions induced by k is dense in $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$, with respect to the infinity norm.*

Given an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ from μ_{XY}, we can estimate $\mathrm{HSIC}(X, Y)$ by

$$\frac{1}{n^2}\sum_{i,j=1}^{n} k_1(X_i, X_j)k_2(Y_i, Y_j) + \frac{1}{n^4}\sum_{i,j,s,t=1}^{n} k_1(X_i, X_j)k_2(Y_s, Y_t) - \frac{2}{n^3}\sum_{i,j,s=1}^{n} k_1(X_i, X_j)k_2(Y_i, Y_s),$$

We refer to it as the HSIC statistic. It can be shown that (Gretton et al., 2007b, Theorems 1 and 2) the HSIC statistic converges to $\mathrm{HSIC}(X, Y)$ at rate $O(n^{-1})$ and $O(n^{-1/2})$ under $\mathbf{H}_0$ and $\mathbf{H}_1$, respectively. This suggests that the properly calibrated HSIC-based test has power converging to one as $n \to \infty$.

5.2.3 Distance based independence criterion

Another widely used independence criterion is the distance covariance (dCov) introduced by Székely and Rizzo (2004) on Euclidean spaces and later generalized to semi-metric spaces of negative type (Lyons, 2013; Sejdinovic et al., 2013). Let $(\mathcal{X}, \rho_1)$ and $(\mathcal{Y}, \rho_2)$ be semi-metric spaces of negative type. The dCov of X and Y is defined to be

$$\mathrm{dCov}(X, Y) := \mathrm{dCov}(X, Y)_{\rho_1,\rho_2} := \int \rho_1\rho_2 \mathrm{d}[(\mu_{XY} - \mu_X \otimes \mu_Y) \otimes (\mu_{XY} - \mu_X \otimes \mu_Y)],$$

where $\rho_1\rho_2$ is viewed as a function on $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$. While this definition suggests that the distance covariance and the energy distance are closed related, it is not true that $\mathrm{dCov}_{\rho_1,\rho_2}(X, Y)$ equals $\mathrm{ED}_{-\rho_1\rho_2}(\mu_{XY}, \mu_X \otimes \mu_Y)$ since $-\rho_1\rho_2$ is not a semi-metric. However, due to Sejdinovic et al. (2013, Corollary 26), there exists a semi-metric $\tilde{\rho}$ such that $\mathrm{dCov}_{\rho_1,\rho_2}(X, Y) = \mathrm{ED}_{\tilde{\rho}}(\mu_{XY}, \mu_X \otimes \mu_Y)$. When $(\mathcal{X}, \rho_1)$ and $(\mathcal{Y}, \rho_2)$ are metric spaces of strong negative type, Lyons (2013, Theorem 3.20) shows that the distance covariance is a valid independence criterion.

The distance covariance and HSIC inherit the equivalence between the energy distance and MMD. To be more specific, there exist (Sejdinovic et al., 2013, Theorem 24) psd kernels k_1 and k_2 on $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that $\mathrm{dCov}_{\rho_1,\rho_2}(X, Y) = 4\,\mathrm{HSIC}_{k_1,k_2}(X, Y)$. For this reason, we only consider HSIC in our experiments.

5.2.4 Sub-Gaussian processes and metric entropy

The presentation of this section mainly follows Wainwright (2019, Section 5). The sub-Gaussian process is an extension of the sub-Gaussian random variable to stochastic processes. Intuitively, a sub-Gaussian process is a stochastic process with sub-Gaussian increments. We give its precise definition here. Let $(\mathcal{F}, \rho)$ be a metric space.

Definition 5.3. *Let $\{Z_f : f \in \mathcal{F}\}$ be a collection of mean-zero random variables. We call it a sub-Gaussian process with respect to ρ if*

$$\mathbb{E}[\exp(\lambda(Z_f - Z_{f'}))] \leq \exp[\lambda^2 \rho^2(f, f')/2], \quad \textit{for all } \lambda > 0.$$

A well-known result for sub-Gaussian processes is Dudley's entropy integral bound. Before we state it, let us review the notions of covering number and metric entropy. The *covering number* of a metric space is the number of balls of a fixed radius $\tau > 0$ required to cover it, which provides a way to measure the size of this space.

Definition 5.4. *Let $\{f^1, \ldots, f^N\} \subset \mathcal{F}$. We call it a τ-*cover *of $\mathcal{F}$ w.r.t. ρ if for each $f \in \mathcal{F}$ there exists $i \in [N]$ such that $\rho(f, f_i) \leq \tau$. The τ-*covering number*, denoted by $N(\tau, \mathcal{F}, \rho)$, is the cardinality of the smallest τ-cover.*

Typically, the covering number diverges as $\tau \to 0^+$, and the growth rate on a logarithmic scale is of interest to us. This is characterized by $\log N(\tau, \mathcal{F}, \rho)$ which is known as the *metric entropy*. Now we are ready to give Dudley's entropy integral bound.

Theorem 5.1 (Theorem 5.22 in Wainwright (2019))**.** *Let $\{Z_f : f \in \mathcal{F}\}$ be a mean-zero sub-Gaussian process w.r.t. ρ. Then we have*

$$\mathbb{E}\left[\sup_{f,f' \in \mathcal{F}} (Z_f - Z_{f'})\right] \leq 32 \int_0^D \sqrt{\log N(\tau, \mathcal{F}, \rho)} \mathrm{d}\tau,$$

where $D := \sup_{f,f' \in \mathcal{F}} \rho(f, f')$ is the diameter of $\mathcal{F}$.

5.2.5 Empirical processes

Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from some distribution P on $\mathcal{X}$. Denote $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the empirical measure. Given a collection $\mathcal{F}$ of real-valued measurable functions on $\mathcal{X}$, the $\mathcal{F}$-indexed *empirical process* $\mathbb{G}_n$ is defined as

$$f \mapsto \mathbb{G}_n f := \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n [f(X_i) - Pf],$$

where $Qf := \int f \mathrm{d}Q$ for a signed measure Q. Given a function f such that $Pf^2 < \infty$, it follows from the LLN and CLT that

$$P_n f \to_{a.s.} Pf \quad \text{and} \quad \mathbb{G}_n f \to_d \mathcal{N}(0, \mathbb{V}\mathrm{ar}_P(f)).$$

The empirical process theory aims to study these two convergences uniformly in f over $\mathcal{F}$ (see, e.g., van der Vaart and Wellner, 1996). The key quantity to control is

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n f(X_i) - Pf\right|. \tag{5.4}$$

One of the main approaches towards such results is based on the symmetrization trick reviewed below. The symmetrized empirical process is defined as

$$f \mapsto \mathbb{S}_n f := \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i),$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables that are independent of $\{X_i\}_{i=1}^n$.

Lemma 5.2 (Lemma 2.3.1 in van der Vaart and Wellner (1996))**.** *For every non-decreasing and convex $\Phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}[\Phi(\|P_n - P\|_{\mathcal{F}})] \le \mathbb{E}[\Phi(2\,\|\mathbb{S}_n\|_{\mathcal{F}})].$$

The symmetrized empirical process is, conditioned on $\{X_i\}_{i=1}^n$, a sub-Gaussian process, and thus Dudley's entropy integral bound can be applied. Take a realization $X_i = x_i$ for $i \in [n]$. We define the mean-zero random variable $X_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)$ and consider the stochastic process $\{X_f : f \in \mathcal{F}\}$. It can be shown that (Wainwright, 2019, Example 5.24) this process is sub-Gaussian w.r.t. the metric

$$\|f - g\|_{\mathbf{L}^2(P_n)} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(x_i) - g(x_i)]^2}.$$

Moreover, it holds that

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} |X_f| \leq 24 \int_0^{\sup_{f,g\in\mathcal{F}}\|f-g\|_{\mathbf{L}^2(P_n)}} \sqrt{\log N(\tau, \mathcal{F}, \|\cdot\|_{P_n})}\mathrm{d}\tau. \tag{5.5}$$

5.3 *Entropy-Regularized Optimal Transport*

Recall the EOT problem (4.5) from Section 4.2:

$$S_\varepsilon(P, Q) := \min_{\nu\in\Pi(P,Q)} \left[\int c(z, z')\mathrm{d}\nu(z, z') + \varepsilon\, \mathrm{KL}(\nu\|P \otimes Q)\right]. \tag{5.6}$$

Note that we have changed the notation (x, y) to (z, z') for the sake of presentation in this section where $z, z' \in \mathcal{Z}$. In the following, we will take a closer look at this problem.

5.3.1 Dual formulation

According to Genevay et al. (2016, Proposition 2.1), the EOT problem (5.6) admits a dual formulation

$$\sup_{f,g\in\mathcal{C}(\mathcal{Z})} \left[\int f(z)\mathrm{d}P(z) + \int g(z')\mathrm{d}Q(z') + \varepsilon - \varepsilon \int e^{[f(z)+g(z')-c(z,z')]/\varepsilon}\mathrm{d}P(z)\mathrm{d}Q(z')\right], \tag{5.7}$$

where $\mathcal{C}(\mathcal{Z})$ is the set of real-valued continuous functions on $\mathcal{Z}$. Let $(f_\varepsilon, g_\varepsilon)$ be a solution pair to (5.7). Even though it is not unique, their sum $f_\varepsilon + g_\varepsilon$ is unique. Moreover, the coupling μ_ε defined via $\mathrm{d}\mu_\varepsilon(z, z') = \xi_\varepsilon(z, z')\mathrm{d}P(z)\mathrm{d}Q(z')$ is the unique solution to the EOT problem, where $\xi_\varepsilon(z, z') := \exp\{[c(z, z') - f_\varepsilon(z) - g_\varepsilon(z')]/\varepsilon\}$ satisfies

$$\int \xi_\varepsilon(z, z')\mathrm{d}P(z) \overset{\text{a.s.}}{=} 1 \quad \text{and} \quad \int \xi_\varepsilon(z, z')\mathrm{d}Q(z') \overset{\text{a.s.}}{=} 1. \tag{5.8}$$

Note that the coupling μ_ε is exactly the Schrödinger bridge defined in Section 4.2.2.

5.3.2 Sinkhorn algorithm

When P and Q are discrete measures supported on $\{z_i\}_{i=1}^s$ and $\{z'_j\}_{j=1}^t$, respectively, we can efficiently solve the EOT problem (5.6) by the so-called *Sinkhorn algorithm.* Before we introduce it, let us give a matrix formulation of the EOT problem between two discrete measures. Since P and Q have finite support, they can be represented by probability vectors $p \in \Delta_s$ and $q \in \Delta_t$, respectively, where $\Delta_k := \{p \in \mathbb{R}_+^k : p^\top \mathbf{1} = 1\}$ is the probability k-simplex. Moreover, the set of couplings $\Pi(P, Q)$ can be represented by the set of matrices

$$U(a, b) := \left\{M \in \mathbb{R}_+^{s\times t} : M\,\mathbf{1} = a, M^\top \mathbf{1} = b\right\}.$$

Hence, the matrix formulation of the EOT problem reads

$$\min_{M\in U(a,b)} \left[\langle M, C\rangle + \varepsilon \operatorname{Ent}(M)\right], \tag{5.9}$$

where $C \in \mathbb{R}^{s\times t}$ is the pairwise cost matrix such that $C_{ij} := c(z_i, z'_j)$ and $\operatorname{Ent}(M) := \sum_{i=1}^s \sum_{j=1}^t M_{ij} \log M_{ij}$. Due to Peyré and Cuturi (2019, Proposition 4.3), the solution to (5.9) is unique and has the form

$$M_\varepsilon = \operatorname{Diag}(u) K \operatorname{Diag}(v) \tag{5.10}$$

for two (unknown) scaling variables $u \in \mathbb{R}_+^s$ and $v \in \mathbb{R}_+^t$, where $K \in \mathbb{R}_+^{s\times t}$ is the *Gram matrix* defined as $K_{ij} := e^{-C_{ij}/\varepsilon}$. Since $M_\varepsilon \in U(a, b)$, the variables u and v must satisfy

$$u \odot (Kv) = a \quad \text{and} \quad v \odot (K^\top u) = b, \tag{5.11}$$

where $\odot$ represents the element-wise product. It can be solved by the Sinkhorn algorithm summarized in Algorithm 2, where $\oslash$ represents the element-wise division. Intuitively, this algorithm iteratively solves the two constraints in (5.11).

Remark 5.2. *The problem* (5.10) *is known as the matrix scaling problem in the literature of numerical analysis; see Peyré and Cuturi (2019, Chapter 4.2) for a historical perspective on this topic.*

Algorithm 2 Sinkhorn Algorithm

```
Input: $a$, $b$, and $K$.
Initialize $u \leftarrow \mathbf{1}_s$ and $v \leftarrow \mathbf{1}_t$.
while not converge do
    $u \leftarrow a \oslash (Kv)$ and $v \leftarrow b \oslash (K^\top u)$.
end while
Output: $u$ and $v$.
```

5.3.3 Random feature approximation

There has been a line of work on accelerating the Sinkhorn algorithm. We review here the random feature technique introduced by Scetbon and Cuturi (2020). On a high level, we approximate the Gram matrix K by its low-rank approximation $\xi\zeta^\top$, where $\xi \in \mathbb{R}_+^{s\times p}$ and $\zeta \in \mathbb{R}_+^{t\times p}$ are the matrices of random features. Concretely, let ρ be a probability measure on a measurable space $\mathcal{W}$. Consider a cost function c such that its induced *Gibbs kernel* $k := e^{-c/\varepsilon}$ admits the following form:

$$k(z, z') = \int \phi(z, w)\phi(z', w)\mathrm{d}\rho(w),$$

where $\phi : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}_+$. Note that the Gibbs kernel induced by the quadratic cost admits this expression (Scetbon and Cuturi, 2020, Lemma 1). For $p \in \mathbb{N}_+$, we first obtain an i.i.d. sample $\{w_i\}_{i=1}^p$ from ρ. We then denote $\boldsymbol{w} := (w_1, \dots, w_p)$ and approximate $k(z, z')$ by

$$k_{\boldsymbol{w}}(z, z') := \frac{1}{p}\sum_{l=1}^{p} \phi(z, w_l)\phi(z', w_l).$$

In other words, we approximate the Gibbs matrix by $\xi\zeta^\top$ where $\xi_{il} := \frac{1}{\sqrt{p}}\phi(z_i, w_l)$ and $\zeta_{jl} := \frac{1}{\sqrt{p}}\phi(z'_j, w_l)$ for $i \in [s]$, $j \in [t]$, and $l \in [p]$. Now, if we replace K by $\xi\zeta^\top$ in Algorithm 2 and compute, e.g., Kv in two steps (i.e., $\zeta^\top v$ and $\xi(\zeta^\top v)$), then we reduce the time complexity of each Sinkhorn iteration from $O(st)$ to $O((s+t)p)$. According to Scetbon and Cuturi (2020, Theorem 3.1), it suffices to choose $p = O(\log{(s+t)})$ to achieve a good accuracy.

5.3.4 Sinkhorn divergence

Since the EOT objective S_ε in (5.6) is an approximation to the OT cost when ε is small, it is tempting to use it to quantify the distance between probability measures. However, it is not centered in the sense that $S_\varepsilon(P, P)$ is not necessarily zero. One remedy is to center it by subtracting two diagonal terms:

$$\bar{S}_\varepsilon(P, Q) := S_\varepsilon(P, Q) - \frac{1}{2} S_\varepsilon(P, P) - \frac{1}{2} S_\varepsilon(Q, Q). \tag{5.12}$$

This centered version is known as the *Sinkhorn divergence* which first appears in Ramdas et al. (2017). In a follow-up work, Genevay et al. (2018) applied it to generative modeling. It is shown by Feydy et al. (2019, Theorem 1) that $\bar{S}_\varepsilon$ defines a semi-metric (metric without the triangle inequality) on the space of probability measures with bounded support if the Gibbs kernel induced by the cost is positive universal.

5.4 Entropy-Regularized Optimal Transport for Independence Testing

We introduce in this section a new independence criterion based on the entropy-regularized optimal transport. We develop an independence test whose test statistic is the plug-in estimator of the independence criterion. Finally, we design an efficient algorithm to compute the test statistic which scales quadratically both in time and space. All the proofs are deferred to Appendix D.1.

5.4.1 Entropy-regularized optimal transport independence criterion

Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a continuous cost function satisfying $c((x, y), (x', y')) = 0$ iff $(x, y) = (x', y')$. We introduce the *entropy regularized optimal transport independence criterion (ETIC)*:

$$T(X, Y) := T_\varepsilon(X, Y) := \bar{S}_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y), \tag{5.13}$$

where $\bar{S}_\varepsilon$ is the Sinkhorn divergence defined in (5.12). That is, we use the Sinkhorn divergence to quantify the distance between μ_{XY} and $\mu_X \otimes \mu_Y$ which measures the independence between

X and Y.

As we will show later, it is computationally convenient to work with additive cost functions, i.e., $c((x,y),(x',y')) = c_1(x,x') + c_2(y,y')$. For this type of cost functions, we prove that the resulting ETIC is a valid independence criterion as long as the induced *Gibbs kernels*

$$k_1(x,x') = e^{-c_1(x,x')/\varepsilon} \quad \text{and} \quad k_2(y,y') = e^{-c_2(y,y')/\varepsilon} \tag{5.14}$$

are positive universal.

Proposition 5.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces equipped with Lipschitz costs c_1 and c_2, respectively. Assume that the Gibbs kernels defined in (5.14) are positive universal. Then ETIC is a valid independence criterion on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Moreover, the claim holds true for measures with a bounded support on $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d$ with the costs $c_1(x,x') = \|x - x'\|^p / \lambda_1$ and $c_2(y,y') = \|y - y'\|^p / \lambda_2$ for $p \in \{1,2\}$ and for all $\lambda_1, \lambda_2 > 0$.*

A running example we consider in this chapter is the weighted quadratic cost.

Example 5.3 (Weighted quadratic cost)**.** *Let $\lambda_1, \lambda_2 \in (0,\infty)$. Consider the cost function*

$$c((x,y),(x',y')) = \frac{1}{\lambda_1} \|x - x'\|^2 + \frac{1}{\lambda_2} \|y - y'\|^2 . \tag{5.15}$$

This cost induces two universal kernels

$$k_1(x,x') = e^{-\|x-x'\|^2/(\varepsilon\lambda_1)} \quad \textit{and} \quad k_2(y,y') = e^{-\|y-y'\|^2/(\varepsilon\lambda_2)}.$$

They play a similar role as the two kernels used in HSIC, and $\varepsilon\lambda_1$ and $\varepsilon\lambda_2$ serve as two kernel parameters.

5.4.2 ETIC-based independence test

In order to use ETIC for independence testing, we use the plug-in estimator of $T(X,Y)$ as the test statistic, that is,

$$T_n(X,Y) := T_{\varepsilon,n}(X,Y) := \bar{S}_\varepsilon(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y), \tag{5.16}$$

Algorithm 3 Tensor Sinkhorn Algorithm

```
1: **Input:** $A$, $B$, $K_1$, and $K_2$.
2: Initialize $U \leftarrow \mathbf{1_{n\times n}}$ and $V \leftarrow \mathbf{1_{n\times n}}$.
3: **while** not converge **do**
4:     $U \leftarrow A \oslash (K_1 V K_2^\top)$ and $V = B \oslash (K_1^\top U K_2)$.
5: **end while**
6: **Output:** $U$ and $V$.
```

where $\hat{\mu}_{XY} := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ is the empirical measure of the pairs, and $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\mu}_Y := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ are the empirical measures of the two samples, respectively. Note that this is different from the plug-in estimator in the two-sample case since the product measure $\mu_X \otimes \mu_Y$ is estimated by n^2 *dependent* (rather than independent) pairs $\{(X_i, Y_j)\}_{i,j=1}^n$. It raises challenges in the analysis of its statistical behavior as elaborated in Section 5.5. The statistical test (or decision rule) is then defined as

$$\psi(\alpha) := \mathbb{1}\{T_n(X,Y) > H_n(\alpha)\}, \tag{5.17}$$

where α is a prescribed significance level, e.g., $\alpha = 0.05$, and $H_n(\alpha)$ is a threshold chosen such that the *type I error rate* $\mathbb{P}(\psi(\alpha) = 1 \mid \mathbf{H}_0)$ is bounded by α.

To avoid tuning the regularization parameter ε, we also consider an adaptive version of the test:

$$\psi_a(\alpha) := \mathbb{1}\left\{\max_{\varepsilon\in\mathcal{E}} \bar{T}_{n,\varepsilon}(X,Y) > H_{n,\mathcal{E}}(\alpha)\right\}, \tag{5.18}$$

where $\mathcal{E}$ is a finite set of positive numbers selected by the user and $\bar{T}_{n,\varepsilon}(X,Y) := [T_{n,\varepsilon}(X,Y) - \mathbb{E}[T_{n,\varepsilon}(X,Y)]]/\mathrm{Sd}(T_{n,\varepsilon}(X,Y))$ is the studentized version of $T_{n,\varepsilon}(X,Y)$. In practice, the two quantities $\mathbb{E}[T_{n,\varepsilon}(X,Y)]$ and $\mathrm{Sd}(T_{n,\varepsilon}(X,Y))$ can be estimated via resampling.

5.4.3 Efficient computation of the ETIC statistic

We then derive an efficient algorithm to compute the test statistic. When μ_{XY} admits a density, $\hat{\mu}_X \otimes \hat{\mu}_Y$ is supported on n^2 items $\{X_i\}_{i=1}^n \times \{Y_i\}_{i=1}^n$ almost surely. If we compute the ETIC statistic naively using the Sinkhorn algorithm (i.e., Algorithm 2), each iteration costs $O(n^4)$ time and space due to the matrix-vector product of sizes $n^2 \times n^2$ and $n^2 \times 1$. To speed up its computation, we adopt here a variant of the Sinkhorn algorithm to solve the EOT problem between two measures supported on the Cartesian product $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$.

Let A and B be two probability measures on $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$, where $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$. For convenience, both A and B are represented as matrices, i.e., $A_{ij} = A(x_i, y_j)$. For instance, if we choose $A = \hat{\mu}_{XY}$ and $B = \hat{\mu}_X \otimes \hat{\mu}_Y$, then, in its matrix form, $A = I_n/n$ and $B = \mathbf{1}_{n\times n}/n^2$. Consider an additive cost function c, e.g., the weighted quadratic cost, such that $c((x,y),(x',y')) = c_1(x,x') + c_2(y,y')$ for $x, x' \in \{x_i\}_{i=1}^n$ and $y, y' \in \{y_j\}_{j=1}^n$. Let C_1 and C_2 be the cost matrices of $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$, respectively. Define Gibbs matrices $K_1 := e^{-C_1/\varepsilon}$ and $K_2 := e^{-C_2/\varepsilon}$, where the exponential function is element-wise. We show in the following proposition that Algorithm 3 can be used to compute $S_\varepsilon(A, B)$, where $\oslash$ represents element-wise division. We refer to it as the *Tensor Sinkhorn* algorithm. The proof can be found in Appendix D.1. Each iteration in the Tensor Sinkhorn algorithm takes $O(n^3)$ time and $O(n^2)$ space, thanks to the additive cost function being used. This algorithm can be generalized to measures supported on the Cartesian product of $p > 2$ sets, which is also noted in Peyré and Cuturi (2019, Remark 4.17).

Proposition 5.4. *Define some constants $\kappa_1 := \max_{i,i'} k_1^{-1}(x_i, x_{i'})$, $\kappa_2 := \max_{j,j'} k_2^{-1}(y_j, y_{j'})$, and $\kappa_3 := \max_{i,j}\{a_{ij}^{-1}, b_{ij}^{-1}\}$. The Tensor Sinkhorn algorithm outputs an τ-accurate estimate of the entropic cost $S_\varepsilon(A, B)$ in $O\left(n^3 \log(\kappa_1\kappa_2\kappa_3)/\tau\right)$ arithmetic operations.*

To further speed up the computation, we apply the random feature technique introduced in Section 5.3.3. To be more concrete, let ρ_1 and ρ_2 be two probability measures on measurable spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Consider cost functions c_1 and c_2 such that their induced Gibbs kernels $k_1 := e^{-c_1/\varepsilon}$ and $k_2 := e^{-c_2/\varepsilon}$ are of the form

$$k_1(x, x') = \int \varphi(x, u)\varphi(x', u)d\rho_1(u) \quad \text{and} \quad k_2(y, y') = \int \psi(y, v)\psi(y', v)d\rho_2(v),$$

where $\varphi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}_+$ and $\psi : \mathcal{Y} \times \mathcal{V} \to \mathbb{R}_+$. Note that the Gibbs kernels induced by the weighted quadratic cost admit this expression. For $p \in \mathbb{N}_+$, we obtain two i.i.d. samples $\{u_i\}_{i=1}^p$ and $\{v_i\}_{i=1}^p$ from ρ_1 and ρ_2, respectively. We denote $\boldsymbol{u} := (u_1, \ldots, u_p)$ and approximate $k_1(x, x')$ by

$$k_{1,\boldsymbol{u}}(x, x') := \frac{1}{p} \sum_{k=1}^{p} \varphi(x, u_k)^\top \varphi(x', u_k).$$

We denote by $K_{1,\boldsymbol{u}}$ the Gram matrix of $k_{1,\boldsymbol{u}}$. Similarly, we define $\boldsymbol{v}$, $k_{2,\boldsymbol{v}}$, and $K_{2,\boldsymbol{v}}$. Replacing K_1 and K_2 by their random feature approximations $K_{1,\boldsymbol{u}}$ and $K_{2,\boldsymbol{v}}$ in Algorithm 3 leads to an algorithm with $O(pn^2)$ time complexity and $O(n^2)$ space complexity in each iteration. It is clear that this new algorithm is exactly the Sinkhorn algorithm that solves the EOT problem between A and B with cost $c_{\boldsymbol{u},\boldsymbol{v}}((x, y), (x', y')) := c_{1,\boldsymbol{u}}(x, x') + c_{2,\boldsymbol{v}}(y, y')$, where $c_{1,\boldsymbol{u}} := -\varepsilon \log k_{1,\boldsymbol{u}}$ and $c_{2,\boldsymbol{v}} := -\varepsilon \log k_{2,\boldsymbol{v}}$. Let $S_{\varepsilon, c_{\boldsymbol{u},\boldsymbol{v}}}(A, B)$ be the minimum of this EOT problem and $S_{\varepsilon,c}(A, B)$ be the minimum of the same EOT problem with cost c. The next proposition provides a high-probability guarantee for this random feature approximation.

Assumption 5.1. *There exists a constant $C > 0$ such that, for all $x, x' \in \{x_i\}_{i=1}^n$, $y, y' \in \{y_j\}_{j=1}^n$, $u \in \mathcal{U}$, and $v \in \mathcal{V}$, it holds that*

$$\varphi(x, u)\varphi(x', u)/k_1(x, x') \leq C \quad \textit{and} \quad \psi(y, v)\psi(y', v)/k_2(y, y') \leq C.$$

Proposition 5.5. *Let $\delta > 0$, $\tau > 0$, and $p = \Omega\left(\frac{C^2}{\tau^2} \log \frac{n}{\delta}\right)$. Under Assumption 5.1, with probability at least $1 - \delta$, it holds that*

$$\left| S_{\varepsilon, c_{\boldsymbol{u},\boldsymbol{v}}}(A, B) - S_{\varepsilon,c}(A, B) \right| \leq \tau.$$

Remark 5.4. *If one applies the random feature technique directly to the original Sinkhorn algorithm, then the resulting algorithm would have the same $O(pn^2)$ time complexity but $O(pn^2)$ space complexity.*

For large-scale applications, the $O(n^2)$ space complexity can sometimes be infeasible. Therefore, we also provide a memory-efficient implementation of the ETIC computation using symbolic matrices (Feydy et al., 2020). The key idea is that: when constructing the

Table 5.1: Comparison of complexities, in time and in space, of Sinkhorn, Tensor Sinkhorn (TS), and large-scale Tensor Sinkhorn (LS) algorithms, exact or with random features approximation.

	Sinkhorn		**TS**		**LS**	
	Exact	**RF**	**Exact**	**RF**	**Exact**	**RF**
Time	$O(n^4)$	$\boldsymbol{O(pn^2)}$	$O(n^3)$	$\boldsymbol{O(pn^2)}$	$O(n^3)$	$\boldsymbol{O(pn^2)}$
Space	$O(n^4)$	$O(pn^2)$	$O(n^2)$	$O(n^2)$	$O(dn)$	$\boldsymbol{O(pn)}$

Gibbs matrix $K_1 := (k_1(X_i, X_j))_{i,j=1}^n$, instead of storing K_1 as a full matrix, we store it as a symbolic matrix linked by the kernel k_1 and data $\{X_i\}_{i=1}^n$. Moreover, we only evaluate the symbolic link if the matrix K_1 is involved in a reduction operation such as $K_1 v$. This implementation improves the space complexity of the Tensor Sinkhorn algorithm from $O(n^2)$ to $O(np)$ or $O(nd)$ depending on whether the random feature approximation is used or not. We call it the *large-scale Tensor Sinkhorn* algorithm. We summarize the time and space complexities of different implementations of ETIC in Table 5.1. We also compare the runtime and memory of variants of the Tensor Sinkhorn algorithm in Figure 5.1. As expected, the large-scale Tensor Sinkhorn algorithm outperforms others both in time and memory.

5.4.4 Gradient backpropagation through ETIC

We describe here how ETIC can fit into a differentiable programming framework, i.e., how one can run the reverse mode automatic differentiation through statistical quantities based on ETIC. Recently, Li et al. (2021) proposed a self-supervised learning approach using HSIC which we summarize below. Let (W, Y) be a pair of image and its identity. Given an i.i.d. sample $\{(W_i, Y_i)\}_{i=1}^n$, the goal is to learn a feature embedding model ϕ_θ such that the

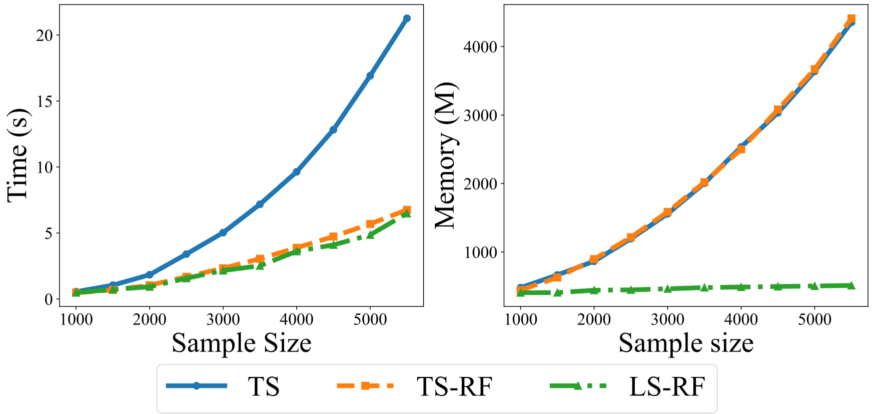

Figure 5.1: Runtime and memory comparison of Tensor Sinkhorn (TS), Tensor Sinkhorn with random feature (TS-RF), and large-scale Tensor Sinkhorn with random feature (LS-RF). The number of random features is set to $O(\log n)$.

dependence between the image feature $X := \phi_\theta(W)$ and its identity Y is maximized, i.e., $\max_{\theta\in\Theta} \mathrm{HSIC}_n(\phi_\theta(W), Y)$. Similarly, one could also maximize the dependence measured by ETIC instead. This boils down to gradient backpropagation through $T_n(\phi_\theta(W), Y)$. We use the strategy in Peyré and Cuturi (2019, Section 9.1.3) and illustrate it on the entropy regularized OT $S_\varepsilon(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y)$ defined in (5.6). For the forward pass, we construct the computational graph via the following steps. Firstly, we run Algorithm 3 (or its random feature variant) with $A = I_n/n$, $B = \mathbf{1}_{n\times n}/n^2$, $K_1 = \big(k_1(\phi_\theta(W_i), \phi_\theta(W_j))\big)_{n\times n}$, and $K_2 = \big(k_2(Y_i, Y_j)\big)_{n\times n}$ for L iterations to get $U^{(L)}$ and $V^{(L)}$. Secondly, we obtain the associated Schrödinger potentials $F^{(L)} := \varepsilon \log U^{(L)}$ and $G^{(L)} := \varepsilon \log V^{(L)}$. Thirdly, we approximate $S_\varepsilon(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y)$ by $S_\varepsilon^{(L)}(\theta) := \langle F^{(L)}, A\rangle_{\mathbf{F}} + \langle G^{(L)}, B\rangle_{\mathbf{F}}$ where $\langle\cdot,\cdot\rangle_{\mathbf{F}}$ is the Frobenius inner product. For the backward pass, we call the reverse mode automatic differentiation to evaluate $\nabla_\theta S_\varepsilon^{(L)}(\theta)$. Since computing $S_\varepsilon^{(L)}(\theta)$ only requires simple operations between matrices, the time complexity of the above procedure is of the same order as the one of Algorithm 3 for the computation of $S_\varepsilon^{(L)}(\theta)$.

5.5 Finite-Sample Analysis

We characterize the statistical behavior of the ETIC test by providing non-asymptotic bounds. We present the main results and their proof sketches here. We use C to denote a constant whose value may change from line to line, where subscripts are used to emphasize the dependency on other quantities. For instance, C_d represents a constant depending only on the dimension d. The detailed proofs are deferred to Appendix D.

5.5.1 Consistency

We first show that the ETIC statistic is a consistent estimator of its population counterpart under both the null and alternative.

Assumption 5.2. *We make the following assumptions:*

(i) $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ *and* $d := d_1 + d_2$.

(ii) c *is chosen as the quadratic cost.*

(iii) μ_X *and* μ_Y *are* $subG(\sigma^2)$.

The quadratic cost is chosen for the sake of concision. We extend the results to weighted quadratic cost in Appendix D.

Theorem 5.6. *Under Assumption 5.2, we have*

$$\mathbb{E}\,|T_n(X,Y) - T(X,Y)| \leq C_d \left(1 + \frac{\sigma^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right) \frac{\varepsilon}{\sqrt{n}}.$$

The bound in Theorem 5.6 coincides with the one obtained by Mena and Weed (2019) in the two-sample case. In terms of the sample size, it scales as $O(n^{-1/2})$ which is the standard parametric rate of convergence. As for the dimension, it contains two dimension-dependent terms. The first term C_d is a constant that only depends on the dimension. The second one involving σ^2 and ε has exponential dependency on the dimension. If we choose $\varepsilon = \sigma^2$, the bound simplifies to $C_d\sigma^2/\sqrt{n}$ which only depends on the dimension via the constant C_d.

We then outline the proof ideas of this theorem. Given a probability measure μ on $\mathbb{R}^d$ and a set of real-valued functions $\mathcal{F}$ on $\mathbb{R}^d$, we denote

$$\|\mu\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}\mu \right| .$$

We start by upper bounding the above $\mathbf{L}^1$ loss $\mathbb{E}\left|T_n(X,Y) - T(X,Y)\right|$ by the supremum of an empirical process and a U-process

$$\|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s}^2 \quad \text{and} \quad \|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}^s}^2 ,$$

respectively, where $\mathcal{F}^s$ is the set of real-valued functions satisfying

$$|f(x,y)| \le C_{s,d}(1 + \|(x,y)\|^2) \quad \text{and} \quad |D^\alpha f(x,y)| \le C_{s,d}(1 + \|(x,y)\|^{|\alpha|}),$$

for any multi-index α with $1 < |\alpha| \le s$. Mena and Weed (2019) used a similar strategy in their proofs. Empirical process theory has a long history in statistics and there are well-established tools to control them (see, e.g., van der Vaart and Wellner, 1996). However, the theory of U-processes is much less well-developed. Moreover, many of the previous works focus on one-sample U-processes (see, e.g., Peña and Giné, 1999). The second U-process here is a two-sample U-process on a paired sample, bringing about additional challenges in its analysis compared to the first empirical process. In order to control it, we develop the following results.

The first result is a metric entropy bound for *degenerate two-sample U-processes*. The main challenge comes from the dependence among the summands in $\sum_{i,j=1}^n f(X_i, Y_j)$. We get around that using the decoupling technique presented in Peña and Giné (1999). Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we say it is *degenerate* under $\mu_X \otimes \mu_Y$ if

$$\mathbb{E}_{\mu_X \otimes \mu_Y}[f(X,Y) \mid X] \overset{\mu_X\text{-a.s.}}{=} 0 \quad \text{and} \quad \mathbb{E}_{\mu_X \otimes \mu_Y}[f(X,Y) \mid Y] \overset{\mu_Y\text{-a.s.}}{=} 0.$$

Proposition 5.7. *Let $\mathcal{F}$ be a class of real-valued functions that are degenerate under $\mu_X \otimes \mu_Y$. Under Assumption 5.2, we have*

$$\mathbb{E}\,\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}}^2 \le \frac{C}{n}\,\mathbb{E}\left(\int_0^B \sqrt{\log N(\tau, \mathcal{F}, \mathbf{L}^2(\hat{\mu}_X \otimes \hat{\mu}_Y))}\mathrm{d}\tau\right)^2,$$

where B is any measurable upper bound of $2\max_{f \in \mathcal{F}} \|f\|_{\mathbf{L}^2(\hat{\mu}_X \otimes \hat{\mu}_Y)}$.

Remark 5.5. *In classical two-sample U-statistics literature, it is usually assumed that the two samples are independent, i.e., X is independent of Y. However, Proposition 5.7 allows the sample to be paired since $(X, Y) \sim \mu_{XY}$.*

With Proposition 5.7 at hand, we can control the U-process $\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}^s}^2$ by upper bounding its covering number $N(\tau, \mathcal{F}^s, \mathbf{L}^2(\hat{\mu}_X \otimes \hat{\mu}_Y))$. The proof is inspired by Mena and Weed (2019) and relies on a result in van der Vaart and Wellner (1996, Chapter 2.7) to control the covering number of a class of smooth functions.

Proposition 5.8. *Under Assumption 5.2, there exists a random variable $L \geq 1$ depending on the samples $\{(X_i, Y_i)\}_{i=1}^n$ with $\mathbb{E}[L] \leq 2$ such that, for any $s \geq 2$,*

$$\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(\hat{\mu}_X \otimes \hat{\mu}_Y)) \leq C_{s,d} \tau^{-d/s} L^{d/2s} (1 + \sigma^{2d})$$

and

$$\max_{f \in \mathcal{F}^s} \|f\|_{\mathbf{L}^2(\hat{\mu}_X \otimes \hat{\mu}_Y)}^2 \leq C_{s,d}(1 + L\sigma^4).$$

In particular, when $s > d/2$, we have

$$\mathbb{E} \|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}^s}^2 \leq C_{s,d}(1 + \sigma^{2d+4}) \frac{1}{n}.$$

5.5.2 Exponential tail bound

We also prove an exponential tail bound for the ETIC statistic. It follows from Theorem 5.6 and the McDiarmid inequality.

Theorem 5.9. *Let c be the quadratic cost. Assume that μ_X and μ_Y are supported on a bounded domain of radius R. Then we have, with probability at least $1 - \delta$,*

$$|T_n(X, Y) - T(X, Y)| \leq C_d \left(1 + \frac{R^{5d+16}}{\varepsilon^{5d/2+8}} \sqrt{\log \frac{6}{\delta}}\right) \frac{\varepsilon}{\sqrt{n}}.$$

Under $\mathbf{H}_0$, we have $T(X, Y) = 0$, so Theorem 5.9 implies that

$$|T_n(X, Y)| > C_d \left(1 + \frac{R^{5d+16}}{\varepsilon^{5d/2+8}} \sqrt{\log \frac{6}{\delta}}\right) \frac{\varepsilon}{\sqrt{n}}$$

with probability at most δ. It gives an estimate of the tail behavior of $T_n(X,Y)$ which suggests that the critical value $H_n(\alpha)$ in (5.17) should be of order $O(n^{-1/2})$. Under $\mathbf{H}_1$, Theorem 5.9 implies that

$$T_n(X,Y) > T(X,Y) - C_d\left(1 + \frac{R^{5d+16}}{\varepsilon^{5d/2+8}}\sqrt{\log\frac{6}{\delta}}\right)\frac{\varepsilon}{\sqrt{n}}$$

with probability at least $1-\delta$. When $T(X,Y) > 0$, it is clear that the right hand side in the above inequality exceeds the threshold $H_n(\alpha)$ for large n. Hence, the ETIC test has power converging to 1 as $n \to \infty$.

5.6 Experiments

We examine the empirical behavior of the proposed ETIC test for independence testing on both synthetic and real data. The code to reproduce the experiments is available online (etic, 2022).

We focus on the weighted quadratic cost

$$c((x,y),(x',y')) = \frac{1}{\lambda_1}\|x-x'\|^2 + \frac{1}{\lambda_2}\|y-y'\|^2 .$$

For convenience, we absorb the regularization parameter ε in ETIC into the weights $\{\lambda_i\}_{i=1}^2$ and set $\varepsilon = 1$. It then induces two Gibbs kernels

$$k_1(x,x') = e^{-\frac{\|x-x'\|^2}{\lambda_1}} \quad \text{and} \quad k_2(y,y') = e^{-\frac{\|y-y'\|^2}{\lambda_2}}$$

with λ_i being the parameter of kernel k_i for $i \in \{1,2\}$. To select the weights, we apply the median heuristic (Gretton et al., 2007b) widely used for HSIC, i.e.,

$$\lambda_1 = r_1 M_x \quad \text{and} \quad \lambda_2 = r_2 M_y$$

with r_1 and r_2 ranging from 0.25 to 4, where M_x and M_y are the medians of the quadratic costs $\{\|X_i - X_j\|^2\}_{i,j=1}^n$ and $\{\|Y_i - Y_j\|^2\}_{i,j=1}^n$, respectively. We also examine its random feature variant, which we call ETIC-RF, discussed in Section 5.4.3, where the number of random features is set to be 100 unless otherwise stated. We compare them with the HSIC

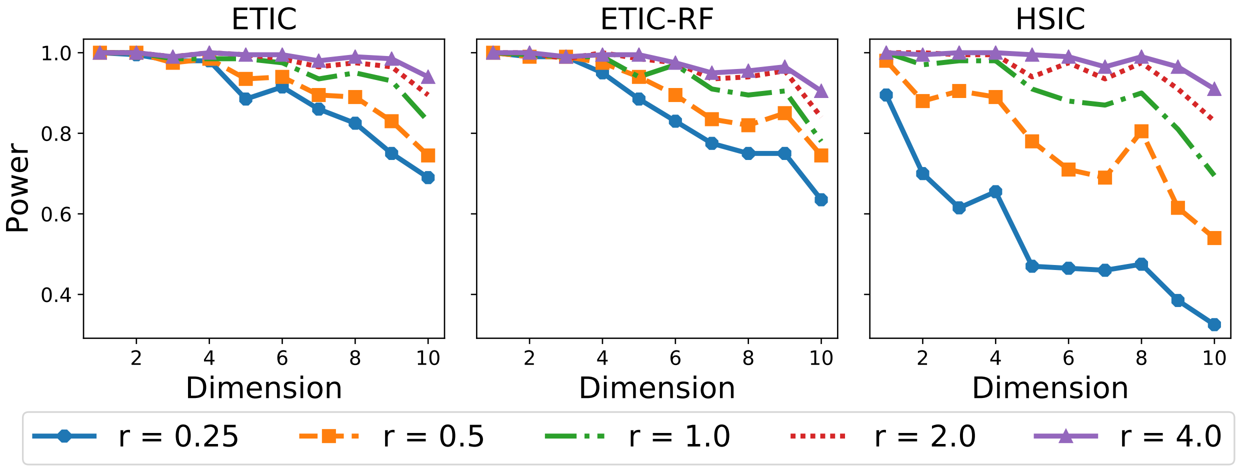

Figure 5.2: Power versus dimension in the linear dependency model (5.19).

statistic with kernels k_1 and k_2. For a fair comparison, we calibrate these tests by a Monte Carlo resampling technique (Feuerverger, 1993) with 200 permutations. For each of the experiment, we repeat the whole procedure 200 times and report the rejection frequency as either the type I error rate (when the null is true) or power (when the null is not true). Note that, even though we are using the same λ_1 and λ_2 in the cost and kernels, that does *not* mean we should compare ETIC and HSIC under the same hyper-parameters. Our goal is to explore their performance over a range of values of the hyper-parameters controlling the regularization penalties.

Our main findings are: 1) Both ETIC and ETIC-RF are consistent in power as the sample size approaches infinity. 2) In some scenarios, ETIC and ETIC-RF outperforms HSIC significantly; in the linear dependency model in particular, their power is much more robust than HSIC to the value of the hyper-parameters. 3) ETIC-RF performs reasonably good compared to ETIC with a moderate number (i.e., 100) of random features. 4) All three tests benefit from large hyper-parameters in detecting simple linear dependency, but smaller values lead to higher power when the dependency is more complicated.

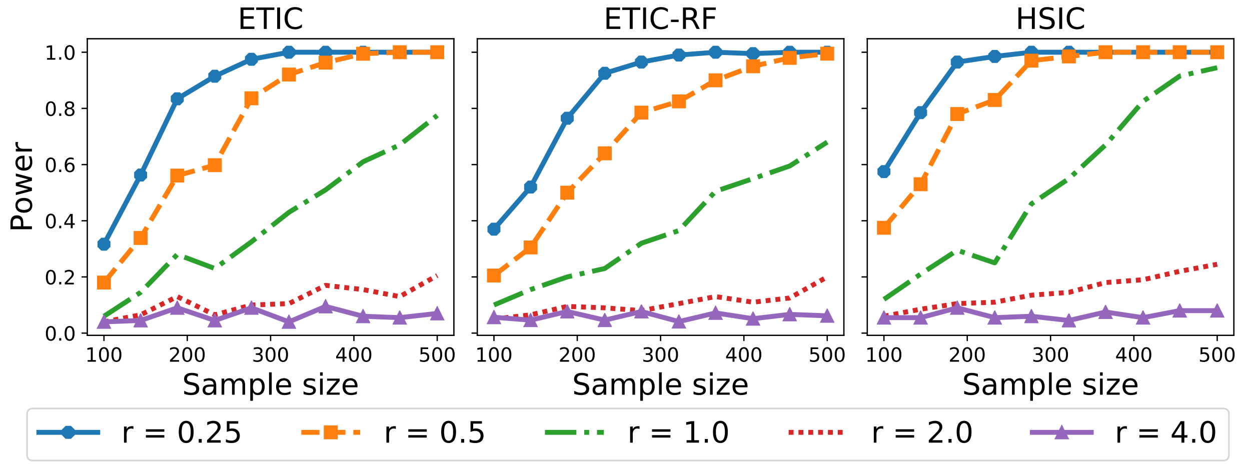

Figure 5.3: Power versus sample size in the Gaussian sign model (5.20).

5.6.1 *Synthetic data*

We first compare the performance of ETIC and ETIC-RF with HSIC on synthetic data. We consider synthetic benchmarks from Gretton et al. (2007b), Jitkrittum et al. (2017), and Zhang et al. (2018). To facilitate the comparison, we set $r_1 = r_2 = r \in \{0.25, 0.5, 1, 2, 4\}$ in this section.

Linear dependency. We begin with a simple linear dependency model. Concretely,

$$X \sim \mathcal{N}_d(0, I_d) \quad \text{and} \quad Y = X_1 + Z, \tag{5.19}$$

where X_1 is the first coordinate of X, and $Z \sim \mathcal{N}(0,1)$ is independent with X. We fix $n = 50$ and plot the power versus $d \in [1, 10]$ in Figure 5.2. All the tests have decaying power as the dimension increases. This is as expected since larger dimension results in weaker dependency between X and Y. It is clear that the power of both ETIC and HSIC increases as r *increases*, with the former more robust than the latter. While the performance of HSIC is similar to ETIC when r is large, it is much worse than ETIC when r is small. As for ETIC-RF, it has similar power curves as ETIC.

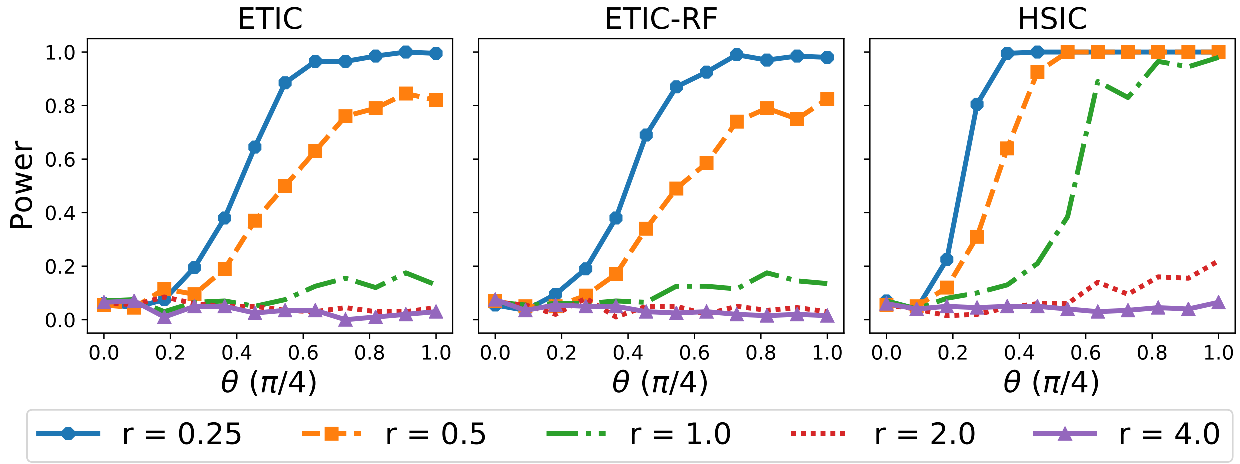

Figure 5.4: Power versus parameter in the subspace dependency model.

Gaussian sign. We then consider a Gaussian sign model, i.e.,

$$X \sim \mathcal{N}_d(0, I_d) \quad \text{and} \quad Y = |Z| \prod_{i=1}^{d} \operatorname{sgn}(X_i), \tag{5.20}$$

where $\operatorname{sgn}(\cdot)$ is the sign function and $Z \sim \mathcal{N}(0,1)$ is independent with X. This problem is challenging since Y is independent with any strict subset of $\{X_1, \ldots, X_d\}$. We fix $d = 3$ and plot the power versus $n \in [100, 500]$ in Figure 5.3. All the tests have improved power as the sample size increases. Additionally, they all benefit from a *small* regularization parameter, with HSIC performs the best and the other two perform similarly in this particular example.

Subspace dependency. One important application of independence testing is independent component analysis (Gretton et al., 2005), which involves separating random variables from their linear mixtures. The next example mimics this application. We construct our data by i) generating n i.i.d. copies of two random variables following independently $0.5\mathcal{N}(0.98, 0.04) + 0.5\mathcal{N}(-0.98, 0.04)$, ii) mixing the two random variables by a rotation matrix parameterized by $\theta \in [0, \pi/4]$ (larger θ leads to stronger dependency), iii) appending

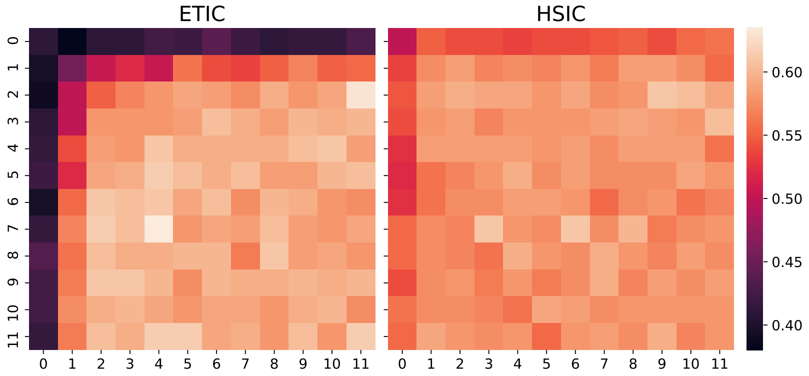

Figure 5.5: Heatmap of power on the bilingual data. The x-axis is for r_1 and y-axis is for r_2. The indices from 0 to 11 correspond to equally spaced values from 0.25 to 4. Lighter color indicates larger power.

$\mathcal{N}_{d-1}(0, I_{d-1})$ to each of the two mixtures, and iv) multiplying each vector by an independent random d-dimensional orthogonal matrix. We refer to it as the *subspace dependency model*. We fix $n = 64$, $d = 2$, and plot the power versus $\theta \in [0, \pi/4]$ in Figure 5.4. As expected, the power of all three tests improves as θ becomes closer to $\pi/4$. Moreover, they all have improved power as r *decreases*. ETIC and ETIC-RF performs similarly, and they are outperformed by HSIC in this example.

5.6.2 Dependency between bilingual text

Inspired by Gretton et al. (2007b), we now investigate the performance of the proposed tests on bilingual data using recent developments in natural language processing. Our dataset is taken from the parallel European Parliament corpus (Koehn, 2005) which consists of a large number of documents of the same content in different languages. For the hyper-parameters, we consider different values of r_1 and r_2 ranging from 0.25 to 4.

Settings. To be more specific, we randomly select $n = 64$ English documents and a paragraph in each document from the corpus. We then pair each paragraph with a random

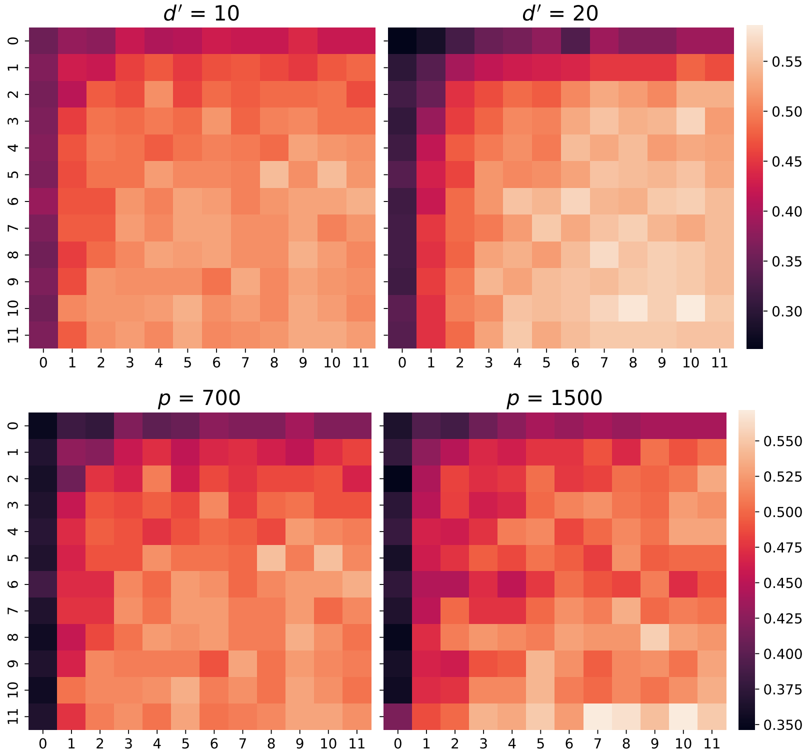

Figure 5.6: Heatmap of power for ETIC-RF with p random features and d' PCs on the bilingual data (**top:** $p = 700$; **bottom:** $d' = 10$). The x-axis is for r_1 and y-axis is for r_2. The indices from 0 to 11 correspond to equally spaced values from 0.25 to 4. Lighter color indicates larger power.

paragraph in the same document in French to form the sample. This sample is *partially dependent* in the sense that the two paragraphs, even though not correspond to the same paragraph, are in the same document. Finally, we use LaBSE (Feng et al., 2020) to embed all the paragraphs into a common feature embedding space of dimension 768 and perform independence testing on these feature vectors. LaBSE is a state-of-the-art, language agnostic, sentence embedding model based on Bidirectional Encoder Representations from Transformers. This allows us to revisit the idea of Gretton et al. (2007b) yet with modern feature embeddings.

Results. The results for ETIC and HSIC are shown in Figure 5.5. ETIC performs better than HSIC when one of r_1 and r_2 is large; while HSIC has larger power when r_1 or r_2 is small. Overall ETIC appears to perform better than HSIC for large amounts of regularization parameters.

As for ETIC-RF, the high-dimensional nature ofthe feature embeddings imposes challenges on the random feature approximation. For its performance to be comparable, we first use dimension reduction (principal component analysis) on the English embeddings and French embeddings separately to reduce the dimension to $d' \ll 768$, and then perform ETIC-RF on the low-dimensional embeddings. Since the dimension reduction step does not utilize information about the joint distribution μ_{XY}, it will not violate the level consistency of the test.

As shown in the first row of Figure 5.6, The number of PCs d' has an interesting effect on the power. Intuitively, the larger d' is the less information we lose, and thus the larger power the test has. This can be seen at the lower right corner where both r_1 and r_2 are large. However, larger d' also means the random feature approximation is harder, especially when r_1 and r_2 are small. This is reflected at the upper left corner where the power decreases as d' increases. We then investigate the effect of p—the number of random features. As shown in the second row of Figure 5.6, the power increases with the number of random features. Overall, the random feature approximation demonstrates similar performance as the exact ETIC with enough random features.

Chapter 6

CONCLUSION

In this dissertation, we addressed some challenges of statistical divergences arising from complex and high-dimensional data in modern applications including parameter estimation, generative models comparison, and the Schrödinger bridge problem. There are several promising venues for future work.

In Chapter 2 we studied the minimum KL divergence estimator in a non-asymptotic fashion using the notion of self-concordance. In a related work (Liu et al., 2022a), the techniques were utilized to obtain excess risk bound on double machine learning/orthogonal statistical learning (DML/OSL) estimators for semi-parametric models. Relying on the Neyman orthogonality, we can achieve a parametric rate of convergence even if the non-parametric part is estimated less accurately, i.e., at rate $O(n^{-1/4})$. However, in some cases such as the case of a partially linear model with non-Gaussian residual, the DML/OSL estimator can still suffer from a large bias. To address this challenge, Mackey et al. (2018) considered the notion of k-orthogonality, recovering the Neyman orthogonality at $k = 1$, and allowed the non-parametric part to be estimated at rate $O(n^{-1/(2k+2)})$. However, their analysis is asymptotic and it would be interesting to explore this direction with our techniques from a non-asymptotic viewpoint, e.g., a risk bound for the DML/OSL estimator under k-orthogonality.

In Chapter 3 we investigated the divergence frontiers for comparing generative models and established sample complexities for its two-step estimation procedure taken by practitioners. In the first quantization step, we showed the existence of an oracle quantization whose quantization error scales linearly with the inverse of the quantization level. However, it is common in practice to use data-dependent quantization schemes with deep neural networks (see, e.g., Sablayrolles et al., 2019; Hämäläinen et al., 2020) whose statistical behavior is to date unclear. Provided new theoretical results on this line of research, it would be interesting to specialize our bounds to such quantization schemes. Furthermore, while our results hold for a large class of f-divergences, it is also interesting to go beyond f-divergences and extend them to, e.g., β-divergences—a class of statistical divergences that is known to be robust against outliers (Samek et al., 2013).

In Chapter 4 we characterized the asymptotic behavior of the discrete Schrödinger bridge by developing novel theoretical tools such as the chaos decomposition and variance analysis of infinite-order U-statistics. Note that the standard two-sample first-order U-statistic with kernel η is $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\eta(X_i, Y_j)$ which can be rewritten as a conditional expectation $\mathbb{E}_{P\otimes Q}[\eta(X_1, Y_1) \mid \mathcal{G}_n]$, and the statistic T_n also admits a conditional expectation form $\mathbb{E}_{\mu}[\eta(X_1, Y_1) \mid \mathcal{G}_n]$. Hence, we can view the statistic T_n as a first-order U-statistic for paired samples. It would be interesting to extend our results to high-order U-statistics for paired samples such as $\mathbb{E}_{\mu}[\eta(X_1, X_2, Y_1, Y_2) \mid \mathcal{G}_n]$. Another interesting direction for future work is the regime when $\varepsilon = \varepsilon_n = o(1)$ as $n \to \infty$. We proved preliminary results in Section 4.3.4 showing the convergence of the discrete Schrödinger bridge towards the OT plan. It is interesting to conduct a more refined analysis by imposing smoothness conditions on the OT plan in the same spirit as Hütter and Rigollet (2021).

In Chapter 5 we proposed an independence test based on the entropy-regularized optimal transport and established finite-sample bounds for its empirical estimator. One promising venue to explore is the extension to joint independence testing, that is, testing joint independence of d random elements. Existing works have generalized distance covariance (dCov) and Hilbert-Schmidt independence criterion (HSIC) to this setting, e.g., the dCov-based measure of mutual dependence (Jin and Matteson, 2018), the distance multivariance (Böttcher et al., 2019), the joint dCov (Chakraborty and Zhang, 2019), and the d-variable HSIC (Pfister et al., 2018). It would be interesting to generalize ETIC to this setting and compare it with these methods. Another direction that is worth exploring is to compare our test with other independence tests and identify the family of alternatives under which our test performs the best. One framework for this purpose is the asymptotic efficiency (see, e.g., Nikitin, 1995).

BIBLIOGRAPHY

https://github.com/langliu95/divergence-frontier-bounds, 2021.

https://github.com/langliu95/confset, 2022.

https://github.com/langliu95/dsbridge, 2022.

https://github.com/langliu95/etic, 2022.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *AISTATS*, 2021.

Jacob Duncan Abernethy, Elad Hazan, and Alexander Rakhlin. An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18, 2017.

Morteza Alamgir, Gábor Lugosi, and Ulrike von Luxburg. Density-preserving quantization with application to graph downsampling. In *COLT*, 2014.

Theodore W. Anderson and Donald A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 1954.

Richard Arratia, Andrew D Barbour, and Simon Tavaré. *Logarithmic Combinatorial Structures: A Probabilistic Approach*, volume 1. European Mathematical Society, 2003.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with Fisher embeddings. In *NeurIPS*, 2021.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 2010.

Francis Bach. *Learning Theory from First Principles.* Online version, 2021.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 2002.

Yannick Baraud. Confidence balls in Gaussian regression. *The Annals of Statistics*, 32(2), 2004.

Isabel Beichl and Francis Sullivan. Approximating the permanent via importance sampling with application to the dimer covering problem. *Journal of computational Physics*, 149 (1), 1999.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* SIAM, 2001.

Rudolf Beran. Confidence sets centered at C_p-estimators. *Annals of the Institute of Statistical Mathematics*, 48(1), 1996.

Rudolf Beran and Lutz Dumbgen. Modulation of estimators and confidence sets. *Annals of Statistics*, 1998.

Daniel Berend and Aryeh Kontorovich. The missing mass problem. *Statistics & Probability Letters*, 82(6), 2012.

Yu M Berezansky and Yuri G Kondratiev. *Spectral Methods in Infinite-Dimensional Analysis.* Springer, 1 edition, 2013.

Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E. Jacob. Schrödinger bridge samplers. *arXiv preprint*, 2019.

Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4), 1946.

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1 edition, 1998.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2), 2019.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, third edition, 1995.

Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán Revista Series A*, 5, 1946.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NIPS*, 2011.

Björn Böttcher, Martin Keller-Ressel, and René L Schilling. Distance multivariance: New dependence measures for random vectors. *The Annals of Statistics*, 47(5), 2019.

Wacha Bounliphone, Arthur Gretton, Arthur Tenenhaus, and Matthew Blaschko. A low variance consistent test of relative dependency. In *ICML*, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Dietrich Braess and Tomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2), 2004.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of KL divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4), 2018.

Sébastien Bubeck and Ronen Eldan. The entropic barrier: Exponential families, log-concave geometry, and self-concordance. *Mathematics of Operations Research*, 44(1), 2019.

Sébastien Bubeck and Yin Tat Lee. Black-box optimization with a politician. In *ICML*. PMLR, 2016.

Haixiao Cai, Sanjeev R. Kulkarni, and Sergio Verdú. Universal divergence estimation for finite-alphabet sources. *IEEE Transactions on Information Theory*, 52(8), 2006.

T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 2011.

George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2 edition, 2001.

Shubhadeep Chakraborty and Xianyang Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528), 2019.

Raymond L Chambers, David G Steel, Suojin Wang, and Alan Welsh. *Maximum Likelihood Estimation for Sample Surveys*. CRC Press, 2012.

Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin. Variational inference and model selection with generalized evidence bounds. In *ICML*, 2018.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4), 1999.

Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *The Annals of Statistics*, 48(3), 2020.

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Review*, 63(2), 2021.

Kenneth W. Church and William A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech & Language*, 5, 1991.

Donato Michele Cifarelli and Eugenio Regazzini. On the centennial anniversary of Ginis theory of statistical relations. *Metron*, 2017.

Stéphan Clémençon and Nicolas Vayatis. Nonparametric estimation of the precision-recall curve. In *ICML*, 2009.

Stéphan Clémençon and Nicolas Vayatis. Nonparametric estimation of the precision-recall curve. In *ICML*, 2009.

G Constantine and T Savits. A multivariate Faà di Bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2), 1996.

Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the ROC curve. In *NIPS*, 2004.

Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the ROC curve. In *NIPS*, 2005.

Harald Cramér. On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, 11, 1928.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2, 1967.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1), 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

A. Philip Dawid, Monica Musio, and Laura Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1), 2016.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *NeurIPS*, 34, 2021.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 2021.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Science & Business Media, 2009.

William E. Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 1940.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv Preprint*, 2020.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John W. Paisley, and David M. Blei. Variational inference via χ upper bound minimization. In *NeurIPS*, 2017.

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Precision-recall curves using information divergence frontiers. In *AISTATS*, 2020.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1), 1969.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorns algorithm. In *ICML*, 2018.

Eugene B Dynkin and Avi Mandelbaum. Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *The Annals of Statistics*, 11(3), 1983.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *ICML*. PMLR, 2020.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *arXiv Preprint*, 2020.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3), 2014.

Andrey Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3), 1993.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, 2019.

Jean Feydy, Joan Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *NeurIPS*, 2020.

Jillian Fisher, Lang Liu, Krishna Pillutla, Yejin Choi, and Zaid Harchaoui. Statistical and computational guarantees for influence diagnostics. *arXiv preprint*, 2022.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society A*, 222(594-604), 1922.

Peter Flach. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, 2012.

Dominique Foata. Some Hermite polynomial identities and their combinatorics. *Advances in Applied Mathematics*, 2, 1981.

Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*. Springer, 1988.

Stefano Fortunati, Fulvio Gini, and Maria S Greco. The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions. *IEEE Transactions on Signal Processing*, 64(9), 2016.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162, 2015.

Edward W. Frees. Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16(1), 1989.

Sara A Geer, Sara van de Geer, and D Williams. *Empirical Processes in M-Estimation*, volume 6. Cambridge university press, 2000.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis R. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Charles J Geyer. Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, 10, 2013.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015.

I. Gohberg, S. Goldberg, and M.A. Kaashoek. *Classes of Linear Operators Vol. 1*. Springer, 1990.

Alexander Goldenshluger and Oleg Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probability Theory and Related Fields*, 143, 2009.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *arXiv preprint*, 2022.

Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4), 1953.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 27, 2014.

Arthur Gretton, Ralf Herbrich, Alexander J. Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6, 2005.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample problem. In *NIPS*, 2007a.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *NIPS*, 2007b.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13, 2012.

Chong Gu. *Smoothing Spline ANOVA Models*. Springer, 2013.

L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 30, 1978.

Gábor Halász and Gábor J. Székely. On the elementary symmetric polynomials of independent random variables. *Acta Mathematica Academiae Scientiarum Hungaricae*, 28, 1976.

Paul R. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1), 1946.

Perttu Hämäläinen, Tuure Saloheimo, and Arno Solin. Deep residual mixture models. *arXiv Preprint*, 2020.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory*, 1 (3), 2020.

Zaid Harchaoui, Lang Liu, and Soumik Pal. Asymptotics of discrete schrödinger bridges via chaos decomposition. *arXiv preprint*, 2022.

Anant Hegde, Deniz Erdogmus, Tue Lehn-Schioler, Yadunandana N Rao, and Jose C Principe. Vector-quantization by density matching in the minimum Kullback-Leibler divergence sense. In *IJCNN*, 2004.

David F Hendry and Bent Nielsen. *Econometric Modeling: A Likelihood Approach*. Princeton University Press, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3), 1948a.

Wassily Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4), 1948b.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *COLT*, 2012.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

Marc Van Hulle. Faithful representations with topographic maps. *Neural Networks*, 12(6), 1999.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5), 2007.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1 edition, 1981.

Daniel Jiwoong Im, He Ma, Graham W. Taylor, and Kristin Branson. Quantitatively evaluating GANs with divergences proposed for training. In *ICLR*, 2018.

Yuri Ingster and Irina A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer, 2003.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186, 1946.

Ze Jin and David S Matteson. Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete v-statistics. *Journal of Multivariate Analysis*, 168:304–322, 2018.

Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *ICML*, 2017.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *ICML*, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35(3), 1987.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematical Analysis*, 2(2), 2020.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit*, 2005.

Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4, 1933.

Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2), 1981.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

William H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284), 1958.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1), 1951.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *NeurIPS*, 2019.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5), 2000.

Lucien M. Le Cam. *Locally Asymptotically Normal Families of Distributions: Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Press, 1960.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

Erich L. Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5), 1966.

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1), 2020.

Christian Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4), 2012.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems*, 34(4), 2014.

Yazhe Li, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *arXiv Preprint*, 2021.

Lang Liu and Zaid Harchaoui. Likelihood score under generalized self-concordance. In *NeurIPS Workshop on Score-Based Methods*, 2022.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Divergence frontiers for generative models: sample complexity, quantization effects, and frontier integrals. In *NeurIPS*, 2021a.

Lang Liu, Joseph Salmon, and Zaid Harchaoui. Score-based change detection for gradient-based learning machines. In *ICASSP*, 2021b.

Lang Liu, Carlos Cinelli, and Zaid Harchaoui. Orthogonal statistical learning with self-concordant loss. In *COLT*, 2022a.

Lang Liu, Mahdi Milani Fard, and Sen Zhao. Distribution embedding networks for generalization from a diverse set of classification tasks. *Transactions on Machine Learning Research*, 2022b.

Lang Liu, Soumik Pal, and Zaid Harchaoui. Entropy regularized optimal transport independence criterion. In *AISTATS*, 2022c.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *ICLR*, 2017.

Russell Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41(5), 2013.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *ICML*, 2018.

Péter Major. The limit behavior of elementary symmetric polynomials of i.i.d. random variables when their order tends to infinity. *The Annals of Probability*, 27(4), 1999.

Avi Mandelbaum and Murad S. Taqqu. Invariance principle for symmetric statistics. *The Annals of Statistics*, 12(2), 1984.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *COLT*, 2019.

Blake Mason, Kwang-Sung Jun, and Lalit Jain. An experimental design approach for regret minimization in logistic bandits. In *AAAI*, 2022.

David A. McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4, 2003.

David A. McAllester and Robert E. Schapire. On the convergence rate of Good-Turing estimators. In *COLT*, 2000.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. Stochastic optimization for spectral risk measures. *arXiv preprint*, 2022.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A), 2018.

Peter Meinicke and Helge Ritter. Quantizing density estimators. In *NIPS*, 2002.

Gonzalo Mena and Jonathan Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017.

Gilles Mordant and Johan Segers. Measuring dependence between random vectors via optimal transport. *arXiv Preprint*, 2021.

Tamás F. Móri and Gábor J. Székely. Asymptotic behaviour of symmetric polynomial statistics. *The Annals of Probability*, 10(1), 1982.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, 2020.

John Ashworth Nelder and Robert W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 1972.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11), 2010.

Frank Nielsen and Rajendra Bhatia. *Matrix Information Geometry.* Springer, 2013.

Thomas G. Nies, Thomas Staudt, and Axel Munk. Transport dependency: Optimal transport based dependency measures. *arXiv Preprint*, 2021.

Yakov Nikitin. *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press, 1995.

Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. In *NeurIPS*, 2015.

Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *FOCS*, 2003.

Dmitrii M. Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1), 2021.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aäron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *NeurIPS*, 2019.

Soumik Pal and Ting-Kam Leonard Wong. Multiplicative Schrödinger problem and the Dirichlet transport. *Probability Theory and Related Fields*, 178(1), 2020.

Victor de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, 1 edition, 1999.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 1900.

Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In *Advances in neural information processing systems*, 2018.

Margaret S. Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449), 2000.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

Niklas Pfister, Peter Bühlmann, Bernhard Schlkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 2018.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019.

Grzegorz A. Rempała and Jacek Wesołowski. Limiting behavior of random permanents. *Statistics & Probability Letters*, 45(2), 1999.

Grzegorz A Rempała and Jacek Wesołowski. Approximation theorems for random permanents and associated stochastic processes. *Probability Theory and Related Fields*, 131(3), 2005.

Grzegorz A. Rempała and Jacek Wesołowski. *Symmetric functionals on random matrices and random matchings problems*, volume 147 of *The IMA Volumes in Mathematics and its Applications*. Springer Science & Business Media, 2007.

Philippe Rigollet. Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2), 2012.

Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11), 2018.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton university press, 1970.

Paul K. Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O. Tolstikhin. Practical and consistent estimation of f-divergences. In *NeurIPS*, 2019.

Hans Henrik Rugh. Cones and gauges in complex spaces: Spectral gaps and complex Perron-Frobenius theory. *Annals of Mathematics*, 171(3), 2010.

Ludger Rüschendorf and Wolfgang Thomsen. Note on the Schrödinger equation and I-projections. *Statistics & Probability Letters*, 17, 1993.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *ICLR*, 2019.

Mehdi S.M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris N. Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Wojciech Samek, Duncan Blythe, Klaus-Robert Müller, and Motoaki Kawanabe. Robust spatial filtering with beta divergence. In *NIPS*, 2013.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 1 edition, 2015.

Meyer Scetbon and Marco Cuturi. Linear time Sinkhorn divergences using positive features. In *NeurIPS*, 2020.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *AAAI*, 2022.

Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'Institut Henri Poincaré*, 2, 1932.

B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4), 1981.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics*, 2013.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, 1980a. Wiley Series in Probability and Mathematical Statistics.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, 1980b.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 1948.

Hongjian Shi, Mathias Drton, and Fang Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 2020.

Galen R Shorack. *Probability for Statisticians.* Springer, 2000.

Jorge Silva and Shrikanth Narayanan. Universal consistency of data-driven partitions for divergence estimation. In *ISIT*, 2007.

Jorge Silva and Shrikanth S Narayanan. Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11), 2010.

Loïc Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. *ICML*, 2019.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, 74(4), 1967.

Nikolai V. Smirnov. On the deviation of the empirical distribution function. *Rec. Math.[Mathematicheskii Sbornik] NS*, 6, 1939.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *ALT*. Springer, 2007.

Alexander Soen and Ke Sun. On the variance of the Fisher information for deep learning. In *NeurIPS*, 2021.

Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6), 2012.

Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincar, Probabilits et Statistiques*, 53(1), 2017.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2, 2001.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. *Mathematical Programming*, 178(1), 2019.

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2007.

Chenyang Tao, Liqun Chen, Ricardo Henao, Jianfeng Feng, and Lawrence Carin. Chi-square generative adversarial network. In *ICML*, 2018.

Levent Tunçel and Arkadi Nemirovski. Self-concordant barriers for convex approximations of structured convex sets. *Foundations of Computational Mathematics*, 10(5), 2010.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Aad W. van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1 edition, 1996.

Bert van Es. On the weak limits of elementary symmetric polynomials. *The Annals of Probability*, 14(2), 1986.

Bert van Es and Roelof Helmers. Elementary symmetric polynomials of increasing order. *Probability theory and related fields*, 80(1), 1988.

Veeravalli S Varadarajan. Weak convergence of measures on separable metric spaces. *Sankhyā: The Indian Journal of Statistics*, 19(1/2), 1958.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2021.

Richard von Mises. *Wahrscheinlichkeitsrechnun*. Deuticke, 1931.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, 2013.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9), 2005.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017.

Michael D Ward and John S Ahlquist. *Maximum Likelihood for Social Science: Strategies for Analysis*. Cambridge University Press, 2018.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A), 2019.

Johannes Wiesel. Measuring association with Wasserstein distances. *arXiv Preprint*, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020.

Scott Yang and Mehryar Mohri. Optimistic bandit convex optimization. In *NIPS*, 2016.

Ming Yu, Mladen Kolar, and Varun Gupta. Statistical inference for pairwise graphical models using score matching. In *NIPS*, 2016.

Ming Yu, Varun Gupta, and Mladen Kolar. Simultaneous inference for pairwise graphical models with generalized score matching. *Journal of Machine Learning Research*, 21(91), 2020.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20(76), 2019.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *NeurIPS*, 2019.

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28, 2018.

Yuchen Zhang and Xiao Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In *ICML*. PMLR, 2015.

Zhiyi Zhang and Michael Grabchak. Nonparametric estimation of Küllback-Leibler divergence. *Neural Computation*, 26(11), 2014.

Wenxin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of statistics*, 46(5), 2018.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Proof of Main Results

Our proofs are inspired by Ostrovskii and Bach (2021). However, there are two key differences. First, since they focus on loss functions of the form $\ell(Y, \theta^\top X)$, the Hessian is $\ell''(Y, \theta^\top X)XX^\top$ where $\ell''(y, \bar{y}) := \mathrm{d}^2\ell(y, \bar{y})/\mathrm{d}\bar{y}^2$. As a result, they can control the deviation of the empirical Hessian using inequalities for sample second-moment matrices of sub-Gaussian random vectors (Ostrovskii and Bach, 2021, Thm. A.2). In contrast, we use matrix Bernstein inequality which allows us to work with a larger class of loss functions. Second, we extend their localization result from pseudo self-concordant losses to generalized self-concordant losses (Proposition 2.10). This is enabled by a new property on the existence of a unique minimizer for generalized self-concordant functions (Proposition 2.3). We also establish the concentration of the effective dimension.

In the remainder of this section, we first prove the localization result Proposition 2.10 and the score bound Proposition 2.11 in Appendix A.1.1. It not only guarantees the existence and uniqueness of θ_n but also localizes it. We then, in Appendix A.1.2, control the empirical Hessian at θ_n as in Proposition 2.12 using a covering number argument. Finally, we prove Theorem 2.6, Theorem 2.7, and Proposition 2.8.

We write $G_\star := G(\theta_\star)$ and $H_\star := H(\theta_\star)$ for short. We use the notation C to denote a constant which may change from line to line, where subscripts are used to emphasize the dependency on other quantities. For instance, C_d represents a quantity depending only on d.

A.1.1 Localization

We start by showing that the empirical risk L_n is generalized self-concordant.

Lemma A.1. *Under Assumption 2.2, the empirical risk L_n is $(n^{\nu/2-1}R, \nu)$-generalized self-concordant.*

Proof. By Assumption 2.2, the loss $\ell(\cdot; Z_i)$ is (R, ν)-generalized self-concordant for every $i \in [n] := \{1, \dots, n\}$. Note that L_n is the empirical average of $\{\ell(\cdot; Z_i)\}_{i=1}^n$. Hence, it follows from Proposition 2.2 that L_n is $(n^{\nu/2-1}R, \nu)$-generalized self-concordant □

Applying Proposition 2.3 to L_n leads to the localization result. Let $\lambda_{n,\star} := \lambda_{\min}(H_n(\theta_\star))$ and $\lambda_n^\star := \lambda_{\min}(H_n(\theta_\star))$. Recall K_ν from Proposition 2.3. Define

$$R^\star_{n,\nu} := \begin{cases} \lambda_{n,\star}^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2 - 1)\lambda_{n,\star}^{(\nu-3)/2} n^{\nu/2-1} R & \text{if } \nu \in (2, 3] \\ (\nu/2 - 1)(\lambda_n^\star)^{(\nu-3)/2} n^{\nu/2-1} R & \text{if } \nu > 3. \end{cases} \tag{A.1}$$

We can then prove Proposition 2.10.

Proposition 2.10. *Under Assumption 2.2, whenever $R^\star_{n,\nu} \|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} \le K_\nu$, the estimator θ_n uniquely exists and satisfies*

$$\|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \le 4 \|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)} .$$

Proof. The claim follows directly from Lemma A.1 and Proposition 2.3. □

Proposition 2.10 implies that the ERM θ_n uniquely exists if $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$ is small. Hence, it remains to bound $\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}$, which can be achieved by controlling $\|S_n(\theta_\star)\|_{H_\star^{-1}}$ and $H_n(\theta_\star)$. Let $\Omega(\theta) := G(\theta)^{1/2} H(\theta)^{-1} G(\theta)^{1/2}$ and $\Omega_\star := \Omega(\theta_\star)$ Recall from Definition 2.6 that $d_\star = \operatorname{Tr}(\Omega_\star)$.

Lemma A.2. *Under Assumption 2.3, it holds that, with probability at least $1 - \delta$,*

$$\|S_n(\theta_\star)\|^2_{H_\star^{-1}} \le \frac{d_\star}{n} + C K_1^2 \log(e/\delta) \frac{\|\Omega_\star\|_2}{n}.$$

Proof. By the first order optimality condition, we have $S(\theta_\star) = 0$. As a result,

$$X := \sqrt{n} G^{-1/2}(\theta_\star) S_n(\theta_\star; Z)$$

is an isotropic random vector. Moreover, it follows from Lemma A.11 that $\|X\|_{\psi_2} \lesssim K_1$. Define $J := G_\star^{1/2} H_\star^{-1} G_\star^{1/2} / n$. Then we have

$$\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 = \|X\|_J^2 .$$

Invoking Theorem 2.4 yields the claim. □

The next result characterizes the concentration of $H_n(\theta_\star)$. Let

$$t_n := t_n(\delta) := \frac{2\sigma_H^2}{-K_2 + \sqrt{K_2^2 + 2\sigma_H^2 n / \log(4d/\delta)}}. \tag{A.2}$$

Note that it decays to 0 at rate $O(n^{-1/2})$ as $n \to \infty$.

Lemma A.3. *Under Assumption 2.4 with $r = 0$, it holds that, with probability at least $1 - \delta$,*

$$(1 - t_n) H_\star \preceq H_n(\theta_\star) \preceq (1 + t_n) H_\star.$$

Furthermore, if $n \ge 4(K_2 + 2\sigma_H^2) \log(2d/\delta)$, we have $t_n \le 1/2$ and thus

$$\frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star.$$

Proof. Due to Assumption 2.4, the standardized Hessian at $\theta_\star$

$$H_\star^{-1/2} H(\theta_\star; Z) H_\star^{-1/2} - I_d$$

satisfies a Bernstein condition with parameter K_2. It then follows from Theorem 2.5 that

$$\mathbb{P}\left(\left\| H_\star^{-1/2} H_n(\theta_\star) H_\star^{-1/2} - I_d \right\|_2 \ge t \right) \le 2d \exp\left\{ -\frac{nt^2}{2(\sigma_H^2 + K_2 t)} \right\}.$$

As a result, it holds that, with probability at least $1 - \delta$, $(1 - t_n) I_d \preceq H_\star^{-1/2} H_n(\theta_\star) H_\star^{-1/2} \le (1 + t_n) I_d$, or equivalently,

$$(1 - t_n) H_\star \preceq H_n(\theta_\star) \preceq (1 + t_n) H_\star.$$

Hence, whenever $n \geq 4(K_2 + 2\sigma_H^2) \log(2d/\delta)$, we have

$$\frac{1}{2} H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2} H_\star.$$

□

We then prove Proposition 2.11. Recall t_n from (A.2).

Proposition 2.11. *Under Assumptions 2.3 and 2.4 with $r = 0$, if $n \geq 4(K_2 + 2\sigma_H^2) \log(4d/\delta)$, then we have $t_n \leq 1/2$ and, with probability at least $1 - \delta$,*

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \leq \frac{d_\star}{n(1-t_n)} + CK_1^2 \log(e/\delta) \frac{\|\Omega_\star\|_2}{n(1-t_n)}.$$

Proof. Define two events

$$\mathcal{A} := \left\{ \|S_n(\theta_\star)\|^2_{H_\star^{-1}} \leq \frac{d_\star}{n} + CK_1^2 \log(2e/\delta) \frac{\|\Omega_\star\|_2}{n} \right\}$$
$$\mathcal{B} := \left\{ (1-t_n) H_\star \preceq H_n(\theta_\star) \preceq (1+t_n) H_\star \right\}.$$

According to Lemmas A.2 and A.3, we have $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/2$ and $\mathbb{P}(\mathcal{B}) \geq 1 - \delta/2$. On the event $\mathcal{AB}$, we have

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \leq \frac{1}{1-t_n} \|S_n(\theta_\star)\|^2_{H_\star^{-1}} \leq \frac{d_\star}{n(1-t_n)} + CK_1^2 \log(2e/\delta) \frac{\|\Omega_\star\|_2}{n(1-t_n)}.$$

Since $\mathbb{P}(\mathcal{AB}) \geq 1 - \mathbb{P}(\mathcal{A}^c) - \mathbb{P}(\mathcal{B}^c) \geq 1 - \delta$, we have, with probability at least $1 - \delta$,

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \leq \frac{d_\star}{n(1-t_n)} + CK_1^2 \log(e/\delta) \frac{\|\Omega_\star\|_2}{n(1-t_n)}.$$

If $n \geq 4(K_2 + 2\sigma_H^2) \log(4d/\delta)$, then $t_n \leq 1/2$ and thus

$$\|S_n(\theta_\star)\|^2_{H_n^{-1}(\theta_\star)} \leq \frac{2d_\star}{n} + CK_1^2 \log(e/\delta) \frac{\|\Omega_\star\|_2}{n}.$$

□

A.1.2 Proof of the main theorems

Before we prove the main theorem, we control the empirical Hessian as in Proposition 2.12. A naïve approach is to invoke Lemma A.3 to bound $H_n(\theta)$ by $H_n(\theta_\star)$. However, this would not work since the generalized self-concordance parameter of L_n, i.e., $n^{\nu/2-1}R$, is diverging as $n \to \infty$. Hence, we use a covering number argument: 1) we take a covering with radius $O(n^{1-\nu/2})$; 2) we bound $H_n(\theta)$ by $H_n(\pi(\theta))$ where $\pi(\theta)$ is the projection of θ onto the covering. The factor $n^{1-\nu/2}$ in the radius will cancel out with the factor $n^{\nu/2-1}$ in the generalized self-concordance parameter; 3) we bound $H_n(\pi(\theta))$ by $H(\pi(\theta))$ using matrix concentration; 4) we bound $H(\pi(\theta))$ by $H(\theta_\star)$ where the generalized self-concordance parameter of L is R. Recall t_n from (A.2), $\lambda_\star := \lambda_{\min}(H_\star)$ and $\lambda^\star := \lambda_{\max}(H_\star)$. Let $\omega_\nu(\tau) := e^\tau$ if $\nu = 2$ and $(1-\tau)^{-2/(\nu-2)}$ if $\nu > 2$.

$$R_\nu^\star := \begin{cases} \lambda_\star^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2 - 1)\lambda_\star^{(\nu-3)/2} R & \text{if } \nu \in (2, 3] \\ (\nu/2-1)(\lambda^\star)^{(\nu-3)/2} R & \text{if } \nu > 3. \end{cases} \tag{A.3}$$

Proposition 2.12. *Fix $\varepsilon \in (0, K_\nu]$ and let $s_n := t_n\big(3^{-d}[1.5\omega_\nu(\varepsilon)n]^{d(1-\nu/2)}\delta/2\big)$. Under Assumptions 2.2 and 2.4 with $r = K_\nu/R_\nu^\star$, it holds that, with probability at least $1-\delta$,*

$$\frac{1}{2\omega_\nu^2(\varepsilon)} H_\star \preceq \frac{1-s_n}{\omega_\nu^2(\varepsilon)} H_\star \preceq H_n(\theta) \preceq (1+s_n)\omega_\nu^2(\varepsilon) H_\star \preceq \frac{3}{2}\omega_\nu^2(\varepsilon) H_\star, \text{ for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star),$$

whenever $n \geq 4(K_2 + 2\sigma_H^2)\left\{\log(4d/\delta) + d\log\left[3(1.5\omega_\nu(\varepsilon)n)^{\nu/2-1}\right]\right\}$.

Proof. We prove the result in the following steps.

Step 1. Take a τ-covering and relate $H_n(\theta)$ to $H_n(\bar\theta)$ for some $\bar\theta$ in the covering. Let $\tau := \varepsilon/R_\nu^\star[1.5\omega_\nu(\varepsilon)n]^{\nu/2-1}$. Take an τ-covering $\mathcal{N}_\tau$ of $\Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$ w.r.t. $\|\cdot\|_{H_\star}$, and let $\pi(\theta)$ be the projection of θ onto $\mathcal{N}_\tau$. Let

$$d_{n,\nu}(\theta_1, \theta_2) := \begin{cases} n^{\nu/2-1} R \left\|\theta_2 - \theta_1\right\|_2 & \text{if } \nu = 2 \\ (\nu/2-1) n^{(\nu/2-1)} R \left\|\theta_2 - \theta_1\right\|_2^{3-\nu} \left\|\theta_2 - \theta_1\right\|_{H_n(\theta_1)}^{\nu-2} & \text{otherwise.} \end{cases}$$

By Lemma A.1 and Proposition 2.1, we have, for all $\theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$,

$$\frac{1}{\omega_\nu(d_{n,\nu}(\pi(\theta),\theta))} H_n(\pi(\theta)) \preceq H_n(\theta) \preceq \omega_\nu(d_{n,\nu}(\pi(\theta),\theta)) H_n(\pi(\theta)), \tag{A.4}$$

where it holds if $d_{n,\nu}(\pi(\theta),\theta) < 1$ for the case $\nu > 2$.

Step 2. Relate $H_n(\theta)$ to $H_\star$ for all θ in the covering. Fix an arbitrary $\theta \in \mathcal{N}_\tau$. Following the same argument as Lemma A.3, we have, with probability at least $1-\delta$,

$$(1-t_n)H(\theta) \preceq H_n(\theta) \preceq (1+t_n)H(\theta). \tag{A.5}$$

It follows from Assumption 2.2 and Lemma A.7 that

$$\frac{1}{\omega_\nu(R_\nu^\star \left\|\theta - \theta_\star\right\|_{H_\star})} H_\star \preceq H(\theta) \preceq \omega_\nu(R_\nu^\star \left\|\theta - \theta_\star\right\|_{H_\star}) H_\star, \tag{A.6}$$

since $R_\nu^\star \left\|\theta - \theta_\star\right\|_{H_\star} \le \varepsilon \le K_\nu < 1$. By the monotonicity of ω_ν, we get

$$\frac{1}{\omega_\nu(\varepsilon)} H_\star \preceq H(\theta) \preceq \omega_\nu(\varepsilon) H_\star,$$

and thus, with probability at least $1-\delta$,

$$\frac{1-t_n(\delta/2)}{\omega_\nu(\varepsilon)} H_\star \preceq H_n(\theta) \preceq [1+t_n(\delta/2)]\omega_\nu(\varepsilon) H_\star.$$

Let $s_n := t_n\big((\tau R_\nu^\star/3\varepsilon)^d \delta/2\big)$ and

$$\mathcal{A} := \left\{ \frac{1-s_n}{\omega_\nu(\varepsilon)} H_\star \preceq H_n(\pi(\theta)) \preceq (1+s_n)\omega_\nu(\varepsilon) H_\star, \text{ for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star) \right\}.$$

Since $|\mathcal{N}_\tau| \le (3\varepsilon/\tau R_\nu^\star)^d$ (Ostrovskii and Bach, 2021), by a union bound, we have $\mathbb{P}(\mathcal{A}) \ge 1-\delta$.

Step 3. Combine the previous two steps. On the event $\mathcal{A}$, we have $H_n(\pi(\theta)) \preceq (1 + s_n)\omega_\nu(\varepsilon) H_\star$ for all $\theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$. A similar argument as Lemma A.7 shows that

$$d_{n,\nu}(\pi(\theta),\theta) \le \begin{cases} \lambda_\star^{-1/2} R\tau & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_\star^{(\nu-3)/2}[(1+s_n)\omega_\nu(\varepsilon)]^{(\nu-2)/2} n^{\nu/2-1} R\tau & \text{if } \nu \in (2,3] \\ (\nu/2-1)(\lambda^\star)^{(\nu-3)/2}[(1+s_n)\omega_\nu(\varepsilon)]^{(\nu-2)/2} n^{\nu/2-1} R\tau & \text{otherwise,} \end{cases}$$

which is equal to $[(1+s_n)\omega_\nu(\varepsilon)]^{\nu/2-1} n^{\nu/2-1} R_\nu^\star \tau$. When

$$n \geq 4(K_2 + 2\sigma_H^2)\left\{\log(4d/\delta) + d\log\left[3(1.5\omega_\nu(\varepsilon)n)^{\nu/2-1}\right]\right\},$$

we have $s_n \leq 1/2$, and thus substituting τ gives $d_{n,\nu}(\pi(\theta),\theta) \leq \varepsilon \leq K_\nu < 1$. Hence, by (A.4), we obtain

$$\frac{1-s_n}{\omega_\nu^2(\varepsilon)} H_\star \preceq H_n(\theta) \preceq (1+s_n)\omega_\nu^2(\varepsilon) H_\star, \quad \text{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star).$$

on the event $\mathcal{A}$. □

We give below the precise version of Theorem 2.6. Recall K_ν and $R_\nu^\star$ from Corollary A.9 and (A.3).

Theorem 2.6. *Let $\nu \in [2,3)$. Under Assumptions 2.2 to 2.4 with $r = 0$, we have, whenever*

$$n \geq \max\left\{4(K_2 + 2\sigma_H^2)\log(4d/\delta), C\left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log(e/\delta)}{K_\nu^2}\right]^{1/(3-\nu)}\right\},$$

the empirical risk minimizer θ_n uniquely exists and satisfies, with probability at least $1-\delta$,

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq \frac{16 d_\star}{n} + CK_1^2 \log(e/\delta)\frac{\|\Omega_\star\|_2}{n}.$$

Proof. Similar to the proof of Proposition 2.11, we define two events

$$\mathcal{A} := \left\{\|S_n(\theta_\star)\|_{H_\star^{-1}}^2 \leq \frac{d_\star}{n} + CK_1^2\log(2e/\delta)\frac{\|\Omega_\star\|_2}{n}\right\} \quad \text{and} \quad \mathcal{B} := \left\{\frac{1}{2}H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2}H_\star\right\}.$$

In the following, we let

$$n \gtrsim \max\left\{4(K_2 + 2\sigma_H^2)\log(4d/\delta), \left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log(e/\delta)}{K_\nu^2}\right]^{1/(3-\nu)}\right\}.$$

Following the same argument as Proposition 2.11, we have $\mathbb{P}(\mathcal{AB}) \geq 1-\delta$ and

$$\|S_n(\theta_\star)\|_{H_n^{-1}(\theta_\star)}^2 \leq \frac{2d_\star}{n} + CK_1^2\log(e/\delta)\frac{\|\Omega_\star\|_2}{n} \leq CK_1^2\log(e/\delta)\frac{d_\star}{n}.$$

Now, it suffices to prove, on the event $\mathcal{AB}$,

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq \frac{16 d_\star}{n} + CK_1^2\log(e/\delta)\frac{\|\Omega_\star\|_2}{n}.$$

Recall $R^\star_{n,\nu}$ and $R^\star_\nu$ from (A.1) and (A.3). It is straightforward to check that $R^\star_{n,\nu} \le \sqrt{2}n^{\nu/2-1}R^\star_\nu$ for all $\nu \in [2,3]$. Consequently, it holds that

$$R^\star_{n,\nu} \left\|S_n(\theta_\star)\right\|_{H_n^{-1}(\theta_\star)} \lesssim R^\star_\nu n^{(\nu-3)/2}\sqrt{K_1^2 \log{(e/\delta)} d_\star} \le K_\nu$$

since $n^{3-\nu} \gtrsim (R^\star_\nu)^2 K_1^2 \log{(e/\delta)} d_\star / K_\nu^2$. As a result, by Proposition 2.10, we have that θ_n uniquely exists and satisfies

$$\left\|\theta_n - \theta_\star\right\|_{H_n(\theta_\star)} \le 4\left\|S_n(\theta_\star)\right\|_{H_n^{-1}(\theta_\star)},$$

and thus, using the event $\mathcal{B}$,

$$\left\|\theta_n - \theta_\star\right\|^2_{H_\star} \le 2\left\|\theta_n - \theta_\star\right\|^2_{H_n(\theta_\star)} \le \frac{16 d_\star}{n} + CK_1^2 \log{(e/\delta)} \frac{\left\|\Omega_\star\right\|_2}{n}.$$

□

We give below the precise version of Theorem 2.7.

Theorem 2.7. *Let $\nu \in [2,3)$ and $r_n := \sqrt{CK_1^2 \log{(e/\delta)} d_\star / n}$. Suppose the same assumptions in Theorem 2.6 hold true. Furthermore, suppose that Assumption 2.4 holds with $r = K_\nu / R^\star_\nu$. Let*

$$\mathcal{C}_n(\delta) := \left\{\theta \in \Theta : \left\|\theta_n - \theta\right\|^2_{H_n(\theta_n)} \le 24\omega_\nu^2(r_n R^\star_\nu)\frac{d_\star}{n} + CK_1^2 \omega_\nu^2(r_n R^\star_\nu) \log{(e/\delta)} \frac{\left\|\Omega_\star\right\|_2}{n}\right\}.$$

Then we have $\mathbb{P}(\theta_\star \in \mathcal{C}_n(\delta)) \ge 1 - \delta$ whenever n satisfies

$$n \ge C \max\left\{(K_2 + \sigma_H^2)\left[\log(2d/\delta) + d\log{(\omega_\nu(K_\nu)n)}\right], \left[\frac{(R^\star_\nu)^2 K_1^2 d_\star \log{(e/\delta)}}{K_\nu^2}\right]^{1/(3-\nu)}\right\}.$$

Here C is an absolute constant which may change from line to line.

Proof. We start by defining some events:

$$\begin{aligned}
\mathcal{A} &:= \left\{\left\|S_n(\theta_\star)\right\|^2_{H_\star^{-1}} \le \frac{d_\star}{n} + CK_1^2 \log{(3e/\delta)} \frac{\left\|\Omega_\star\right\|_2}{n}\right\} \\
\mathcal{B} &:= \left\{\frac{1}{2}H_\star \preceq H_n(\theta_\star) \preceq \frac{3}{2}H_\star\right\} \qquad (A.7) \\
\mathcal{C} &:= \left\{\frac{1}{2\omega_\nu^2(r_n R^\star_\nu)}H_\star \preceq H_n(\theta) \preceq \frac{3}{2}\omega_\nu^2(r_n R^\star_\nu) H_\star, \quad \text{for all } \theta \in \Theta_{r_n}(\theta_\star)\right\}.
\end{aligned}$$

In the following, we let

$$n \geq C \max\left\{ (K_2 + \sigma_H^2)\left[\log(2d/\delta) + d \log\left(\omega_\nu(K_\nu)n\right)\right], \left[\frac{(R_\nu^\star)^2 K_1^2 d_\star \log\left(e/\delta\right)}{K_\nu^2}\right]^{1/(3-\nu)} \right\}.$$

It then follows that $r_n R_\nu^\star \leq K_\nu$. According to Lemma A.2, Lemma A.3, and Proposition 2.12 (with $\varepsilon = r_n R_\nu^\star$), it holds that $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/3$, $\mathbb{P}(\mathcal{B}) \geq 1 - \delta/3$, and $\mathbb{P}(\mathcal{C}) \geq 1 - \delta/3$. This implies that $\mathbb{P}(\mathcal{ABC}) \geq 1 - \delta$. Now, it suffices to prove, on the event $\mathcal{ABC}$,

$$\|\theta_n - \theta_\star\|_{H_n(\theta_n)}^2 \leq 24\omega_\nu^2(r_n R_\nu^\star)\frac{d_\star}{n} + CK_1^2\omega_\nu^2(r_n R_\nu^\star)\log\left(e/\delta\right)\frac{\|\Omega_\star\|_2}{n}.$$

Following the same argument as Theorem 2.6, we obtain

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \leq 2\left\|\theta_n - \theta_\star\right\|_{H_n(\theta_\star)}^2 \leq \frac{16 d_\star}{n} + CK_1^2 \log\left(e/\delta\right)\frac{\|\Omega_\star\|_2}{n} \leq r_n^2.$$

Therefore, using the event $\mathcal{C}$, we have

$$\|\theta_n - \theta_\star\|_{H_n(\theta_n)}^2 \leq \frac{3}{2}\omega_\nu^2(r_n R_\nu^\star)\left\|\theta_n - \theta_\star\right\|_{H_\star}^2 \leq 24\omega_\nu^2(r_n R_\nu^\star)\frac{d_\star}{n} + CK_1^2\omega_\nu^2(r_n R_\nu^\star)\log\left(e/\delta\right)\frac{\|\Omega_\star\|_2}{n},$$

which completes the proof. □

A.1.3 Consistency of d_n

Now we are ready to prove Proposition 2.8. Recall t_n from (A.2) and r_n from Theorem 2.7.

Proposition 2.8. *Let $\nu \in [2, 3)$ and $s_n := CMr_n + CK_1^2(1 + Mr_n)(d/n)\log\left(Mnr_n/\delta\right)$. Under Assumptions 2.2, 2.3', 2.4, and 2.5 with $r = K_\nu/R_\nu^\star$, it holds that, with probability at least $1 - \delta$,*

$$\frac{1 - t_n}{\omega_\nu^2(r_n R_\nu^\star)(1 + s_n)} d_n \leq d_\star \leq \frac{(1 + t_n)\omega_\nu^2(r_n R_\nu^\star)}{1 - s_n} d_n$$

whenever n satisfies

$$n \geq C \max\Big\{ (K_2 + \sigma_H^2 + K_1^2)\left[\log\left(2d/\delta\right) + d\log\left(\omega_\nu(K_\nu)n/\delta\right)\right],$$
$$\left[(M + R_\nu^\star/K_\nu)^2 K_1^2 d_\star \log\left(e/\delta\right)\right]^{1/(3-\nu)} \Big\}.$$

Proof. Let $\tau := \delta/(Mn)$. Take a τ-covering of $\mathcal{N}_\tau$ of $\Theta_{r_n}(\theta_\star)$ w.r.t. $\|\cdot\|_{H_\star}$, and let $\pi(\theta)$ be the projection of θ onto $\mathcal{N}_\tau$. For simplicity of the notation, we define

$$\|A\|_B := \left\|B^{1/2}AB^{1/2}\right\|$$

for a symmetric matrix A and a psd matrix B. We start by defining some events. Let

$$\begin{aligned}
\mathcal{A} &:= \left\{\|S_n(\theta_\star)\|^2_{H_\star^{-1}} \lesssim \frac{1}{n}K_1^2 \log{(5e/\delta)} d_\star\right\}\\
\mathcal{B} &:= \{(1-t_n)H_\star \preceq H_n(\theta_\star) \preceq (1+t_n)H_\star\}\\
\mathcal{C} &:= \left\{\frac{1-t_n}{\omega_\nu^2(r_n R_\nu^\star)}H_\star \preceq H_n(\theta) \preceq (1+t_n)\omega_\nu^2(r_n R_\nu^\star)H_\star, \quad \text{for all } \theta \in \Theta_{r_n}(\theta_\star)\right\}\\
\mathcal{D} &:= \left\{\sup_{\theta\in\Theta_{r_n}(\theta_\star)} \|G_n(\theta) - G_n(\pi(\theta))\|_{G_\star^{-1}} \le 5M\tau/\delta\right\}\\
\mathcal{E} &:= \left\{\sup_{\theta\in\Theta_{r_n}(\theta_\star)} \|G_n(\pi(\theta)) - G(\pi(\theta))\|_{G_\star^{-1}}\right.\\
&\qquad \left.\lesssim K_1^2(1+Mr_n)h\left(\frac{d\log{(36r_n/\tau)} + \log{(10/\delta)}}{n}\right)\right\},
\end{aligned}$$

where $h(t) := \max\{t^2, t\}$. In the following, we let

$$\begin{aligned}
n \ge C\max\Big\{&(K_2 + \sigma_H^2 + K_1^2)\left[\log{(2d/\delta)} + d\log{(\omega_\nu(K_\nu)n/\delta)}\right],\\
&\left[(M + R_\nu^\star/K_\nu)^2 K_1^2 d_\star \log{(e/\delta)}\right]^{1/(3-\nu)}\Big\}.
\end{aligned} \tag{A.8}$$

It then follows that $t_n \le 1/2$, $r_n \le K_\nu/R_\nu^\star = r$ and $s_n < 1$. According to Lemma A.2, Lemma A.3, and Proposition 2.12, it holds that $\mathbb{P}(\mathcal{A}) \ge 1-\delta/5$, $\mathbb{P}(\mathcal{B}) \ge 1-\delta/5$, and $\mathbb{P}(\mathcal{C}) \ge 1-\delta/5$. In the following, we prove the claim in three steps.

Step 1. Control the probability of $\mathcal{D}$. By Markov's inequality, it holds that

$$\begin{aligned}
\mathbb{P}(\mathcal{D}^c) &\le \frac{\delta}{5M\tau}\mathbb{E}\left[\sup_{\theta\in\Theta_{r_n}(\theta_\star)} \|G_n(\theta) - G_n(\pi(\theta))\|_{G_\star^{-1}}\right]\\
&\overset{\text{Jensen's}}{\le} \frac{\delta}{5M\tau} \sup_{\theta\in\Theta_{r_n}(\theta_\star)} \|G(\theta) - G(\pi(\theta))\|_{G_\star^{-1}}.
\end{aligned}$$

According to Assumption 2.5, we have

$$M \left\|\theta_1 - \theta_2\right\|_{H_\star} \geq \mathbb{E}[\left\|G(\theta_1; Z) - G(\theta_2; Z)\right\|_{G_\star^{-1}}], \quad \text{for all } \theta_1, \theta_2 \in \Theta_r(\theta_\star). \tag{A.9}$$

It follows from Jensen's inequality that

$$M \left\|\theta_1 - \theta_2\right\|_{H_\star} \geq \left\|G(\theta_1) - G(\theta_2)\right\|_{G_\star^{-1}}, \quad \text{for all } \theta_1, \theta_2 \in \Theta_r(\theta_\star). \tag{A.10}$$

As a result,

$$\mathbb{P}(\mathcal{D}^c) \leq \frac{\delta}{5\tau} \left\|\theta - \pi(\theta)\right\|_{H_\star} \leq \frac{\delta}{5}.$$

Step 2. Control the probability of $\mathcal{E}$. According to Vershynin (2018, Exercise 4.4.3), we have

$$\left\|G_n(\pi(\theta)) - G(\pi(\theta))\right\|_{G_\star^{-1}} \leq \frac{1}{2} \sup_{v \in \mathcal{V}_{1/4}} \left|v^\top G_\star^{-1/2} [G_n(\pi(\theta)) - G(\pi(\theta))] G_\star^{-1/2} v\right|, \tag{A.11}$$

where $\mathcal{V}_{1/4}$ is a 1/4-covering of the unit ball in $\mathbb{R}^d$. Note that

$$v^\top G_\star^{-1/2} (G_n(\pi(\theta)) - G(\pi(\theta))) G_\star^{-1/2} v = \frac{1}{n} \sum_{i=1}^n [W_i - \mathbb{E}[W_i]],$$

where $W_i := [v^\top G_\star^{-1/2} S(\pi(\theta); Z_i)]^2$. Let $\bar{v} := G(\pi(\theta))^{1/2} G_\star^{-1/2} v$. By Assumption 2.3',

$$\begin{aligned} \left\|v^\top G_\star^{-1/2} S(\pi(\theta); Z_i)\right\|_{\psi_2} &= \left\|\bar{v}^\top G(\pi(\theta))^{-1/2} S(\pi(\theta); Z_i)\right\|_{\psi_2} \\ &\leq \left\|\bar{v}\right\|_2 K_1 \leq \left\|G(\pi(\theta))^{1/2} G_\star^{-1/2}\right\| K_1. \end{aligned}$$

Since $\pi(\theta) \in \Theta_{r_n}(\theta_\star) \subset \Theta_r(\theta_\star)$, it follows from (A.10) that

$$\left\|G_\star^{-1/2} G(\pi(\theta)) G_\star^{-1/2} - I_d\right\| \leq M \left\|\pi(\theta) - \theta_\star\right\|_{H_\star} \leq M r_n.$$

and thus

$$\left\|v^\top G_\star^{-1/2} S(\pi(\theta); Z_i)\right\|_{\psi_2} \leq \sqrt{1 + M r_n} K_1.$$

This implies, by Vershynin (2018, Lemma 2.7.6), W_i is sub-Exponential with $\|W_i\|_{\psi_1} \le K_1^2(1+Mr_n)$. It then follows from the Bernstein inequality that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n [W_i - \mathbb{E}[W_i]]\right| > t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K_1^4(1+Mr_n)^2}, \frac{t}{K_1^2(1+Mr_n)}\right\}\right).$$

Since $|\mathcal{N}_\tau| \le (3r_n/\tau)^d$ and $\mathcal{V}_{1/4} \le 12^d$, by a union bound, we get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{2}\sup_{\theta\in\Theta_{r_n}(\theta_\star)}\sup_{v\in\mathcal{V}_{1/4}}\left|v^\top G_\star^{-1/2}[G_n(\pi(\theta)) - G(\pi(\theta))]G_\star^{-1/2}v\right| > t\right)\\
&\le 2\left|\mathcal{N}_\tau\right|\left|\mathcal{V}_{1/4}\right|\exp\left(-c\min\left\{\frac{4t^2}{K_1^4(1+Mr_n)^2}, \frac{2t}{K_1^2(1+Mr_n)}\right\}\right)\\
&\le 2(36r_n/\tau)^d\exp\left(-c\min\left\{\frac{4t^2}{K_1^4(1+Mr_n)^2}, \frac{2t}{K_1^2(1+Mr_n)}\right\}\right).
\end{aligned}$$

Hence, it follows from (A.11) that $\mathbb{P}(\mathcal{E}^c) \le \delta/5$.

Step 3. Prove the bound on the event $\mathcal{ABCDE}$. Following the same argument as Theorem 2.6, we obtain

$$\|\theta_n - \theta_\star\|_{H_\star} \lesssim \|\theta_n - \theta_\star\|_{H_n(\theta_\star)} \lesssim n^{-1/2}\sqrt{K_1^2\log(e/\delta)d_\star} = r_n. \tag{A.12}$$

Using the event $\mathcal{C}$, we have

$$\frac{1}{(1+t_n)\omega_\nu^2(r_nR_\nu^\star)}H_n(\theta_n) \preceq H_\star \preceq \frac{\omega_\nu^2(r_nR_\nu^\star)}{1-t_n}H_n(\theta_n),$$

and thus

$$\begin{aligned}
d_\star &\le (1+t_n)\omega_\nu^2(r_nR_\nu^\star)\operatorname{Tr}\left(H_n(\theta_n)^{-1/2}G_\star H_n(\theta_n)^{-1/2}\right)\\
d_\star &\ge \frac{1-t_n}{\omega_\nu^2(r_nR_\nu^\star)}\operatorname{Tr}\left(H_n(\theta_n)^{-1/2}G_\star H_n(\theta_n)^{-1/2}\right).
\end{aligned} \tag{A.13}$$

Now it remains to control

$$\|G_n(\theta_n) - G_\star\|_{G_\star^{-1}} \le \|G(\theta_n) - G_\star\|_{G_\star^{-1}} + \|G_n(\theta_n) - G(\theta_n)\|_{G_\star^{-1}}.$$

We first control $\|G(\theta_n) - G_\star\|_{G_\star^{-1}}$. It follows from (A.10) and (A.12) that

$$\|G(\theta_n) - G_\star\|_{G_\star^{-1}} \le M\|\theta_n - \theta_\star\|_{H_\star} \lesssim Mr_n.$$

We then control $\|G_n(\theta_n) - G(\theta_n)\|_{G_\star^{-1}}$. By (A.12), we have

$$\|G_n(\theta_n) - G(\theta_n)\|_{G_\star^{-1}} \leq \sup_{\theta \in \Theta_{r_n}(\theta_\star)} \|G_n(\theta) - G(\theta)\|_{G_\star^{-1}} .$$

It then follows from the triangle inequality that

$$\sup_{\theta \in \Theta_{r_n}(\theta_\star)} \|G_n(\theta) - G(\theta)\|_{G_\star^{-1}} \leq A_1 + A_2 + A_3,$$

where

$$\begin{aligned} A_1 &:= \sup_{\theta \in \Theta_{r_n}(\theta_\star)} \|G(\pi(\theta)) - G(\theta)\|_{G_\star^{-1}} \\ A_2 &:= \sup_{\theta \in \Theta_{r_n}(\theta_\star)} \|G_n(\pi(\theta)) - G(\pi(\theta))\|_{G_\star^{-1}} \\ A_3 &:= \sup_{\theta \in \Theta_{r_n}(\theta_\star)} \|G_n(\theta) - G_n(\pi(\theta))\|_{G_\star^{-1}} . \end{aligned}$$

To control A_1, note that, for all $\theta \in \Theta_{r_n}(\theta_\star)$,

$$\|G(\pi(\theta)) - G(\theta)\|_{G_\star^{-1}} \overset{\text{(A.10)}}{\leq} M \|\pi(\theta) - \theta\|_{H_\star} \leq M\tau.$$

Consequently, we obtain $A_1 \leq M\tau$. To control A_2, we use the event $\mathcal{E}$ to obtain

$$A_2 \lesssim K_1^2(1 + Mr_n) h\left(\frac{d \log(36 r_n/\tau) + \log(10/\delta)}{n}\right).$$

To control A_3, we use the event $\mathcal{D}$ to obtain $A_3 \leq 5M\tau/\delta$. Therefore,

$$\begin{aligned} &\|G_n(\theta_n) - G_\star\|_{G_\star^{-1}} \\ &\leq CMr_n + M\tau + 5M\tau/\delta + CK_1^2(1 + Mr_n) h\left(\frac{d \log(36 r_n/\tau) + \log(10/\delta)}{n}\right) \\ &= CMr_n + \frac{5+\delta}{n} + CK_1^2(1 + Mr_n) h\left(\frac{d \log(36 M n r_n/\delta) + \log(10/\delta)}{n}\right). \end{aligned}$$

This yields that

$$(1 - s_n) G_\star \preceq G_n(\theta_n) \preceq (1 + s_n) G_\star,$$

and thus

$$\frac{1 - t_n}{\omega_\nu^2(r_n R_\nu^\star)(1 + s_n)} d_n \leq d_\star \leq \frac{(1 + t_n)\omega_\nu^2(r_n R_\nu^\star)}{1 - s_n} d_n.$$

□

A.2 Examples and Applications

A.2.1 Examples

Example A.1 (Generalized linear models). *Let $Z := (X, Y)$ be a pair of input and output, where $X \in \mathcal{X} \subset \mathbb{R}^\tau$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. Let $t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ and μ be a measure on $\mathcal{Y}$. Consider the statistical model*

$$p(y \mid x) \sim \frac{\exp(\langle \theta, t(x, y)\rangle)}{\int \exp(\langle \theta, t(x, \bar{y})\rangle) \mathrm{d}\mu(\bar{y})} \mathrm{d}\mu(y)$$

with $\|t(x, Y)\|_2 \le M$ a.s. under $p(y \mid x)$ for all x. It induces the loss function

$$\ell(\theta; z) := -\langle \theta, t(x, y)\rangle + \log \int \exp(\langle \theta, t(x, \bar{y})\rangle) \mathrm{d}\mu(\bar{y}).$$

We first verify Assumption 2.2, i.e., show that it is generalized self-concordant for $\nu = 2$ and $R = 2M$. We denote by $\mathbb{E}_{Y|x}$ the expectation w.r.t. $p(y \mid x)$. Note that $\log \int \langle \theta, t(x, \bar{y})\rangle \mathrm{d}\mu(\bar{y})$ is the cumulant generating function. It follows from some computation that

$$\begin{aligned}
D_\theta \ell(\theta; z)[u] &= -\langle u, t(x, y)\rangle + \mathbb{E}_{Y|x} \langle u, t(x, Y)\rangle \\
D_\theta^2 \ell(\theta; z)[u, u] &= \mathbb{E}_{Y|x}[\langle u, t(x, Y)\rangle^2] - [\mathbb{E}_{Y|x} \langle u, t(x, Y)\rangle]^2 \\
D_\theta^3 \ell(\theta; z)[u, u, v] &= \mathbb{E}_{Y|x}[\langle u, t(x, Y)\rangle^2 \langle v, t(x, Y\rangle] - \mathbb{E}_{Y|x}[\langle u, t(x, Y)\rangle^2]\, \mathbb{E}_{Y|x} \langle v, t(x, Y)\rangle \\
&\quad - 2\, \mathbb{E}[\langle u, t(x, Y)\rangle \langle v, t(x, Y)\rangle]\, \mathbb{E}[\langle u, t(x, Y)\rangle] - 2[\mathbb{E} \langle u, t(x, Y)\rangle]^2\, \mathbb{E} \langle v, t(x, Y)\rangle.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\left| D_\theta^3 \ell(\theta; z)[u, u, v] \right| \\
&= \left| \mathbb{E}_{Y|x} \left\{ \left[\langle u, t(x, Y)\rangle - \mathbb{E}_{Y|x} \langle u, t(x, Y)\rangle \right]^2 \left[\langle v, t(x, Y)\rangle - \mathbb{E}_{Y|x} \langle v, t(x, Y)\rangle \right] \right\} \right| \\
&\le 2M \|v\|_2\, \mathbb{E}_{Y|x} \left\{ \left[\langle u, t(x, Y)\rangle - \mathbb{E}_{Y|x} \langle u, t(x, Y)\rangle \right]^2 \right\}, \quad \text{by } \|t(x, Y)\|_2 \overset{a.s.}{\le} M \\
&= 2M \|v\|_2\, D_\theta^2 \ell(\theta; z)[u, u],
\end{aligned}$$

which completes the proof.

We then verify Assumption 2.3 and Assumption 2.3'. By Lemma A.10, it suffices to show that $\|S(\theta_\star; Z)\|_2$ *is a.s. bounded. In fact,*

$$S(\theta_\star; z) = -t(x, y) + \mathbb{E}_{p_{\theta_\star}(Y|x)}[t(x, Y)].$$

Since $|t(X, Y)|_2 \overset{a.s.}{\leq} M$*, we get* $\|S(\theta_\star; Z)\|_2 \overset{a.s.}{\leq} 2M$ *and thus the claim follows. Assumption 2.3' can be verified similarly.*

Next, we verify Assumption 2.4. According to Lemma A.12, it is enough to prove that $\|H(\theta; Z)\|_2$ *is a.s. bounded. In fact,*

$$H(\theta; z) = \mathbb{E}_{Y|x}[t(x, Y)t(x, Y)^\top] - \mathbb{E}_{Y|x}[t(x, Y)]\, \mathbb{E}_{Y|x}[t(x, Y)]^\top.$$

Since $\left\|t(X, Y)t(X, Y)^\top\right\|_2 \leq \|t(X, Y)\|_2^2 \overset{a.s.}{\leq} M^2$*, it follows that* $\|H(\theta, Z)\|_2 \overset{a.s.}{\leq} M^2$*.*

Finally, we verify Assumption 2.5. It suffices to show that $\|G(\theta_1; Z) - G(\theta_2; Z)\|_2 / \|\theta_1 - \theta_2\|_2$ *is a.s. bounded. Note that*

$$\begin{aligned} &G(\theta_1; z) - G(\theta_2; z) \\ &= \mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)]\, \mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)]^\top - \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)]\, \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)]^\top \\ &\quad - 2t(x, y) \left\{\mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)] - \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)]\right\}^\top. \end{aligned}$$

For the second term, we have

$$\begin{aligned} &\left\|-2t(x, y) \left\{\mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)] - \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)]\right\}^\top\right\|_2 \\ &\leq 2\, \|t(x, y)\|_2 \left\|\mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)] - \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)]\right\|_2. \end{aligned}$$

Note that

$$\begin{aligned} &\mathbb{E}_{p_{\theta_1}(Y|x)}[t(x, Y)] - \mathbb{E}_{p_{\theta_2}(Y|x)}[t(x, Y)] \\ &= \frac{\int t(x, y) \exp(\langle \theta_1, t(x, y)\rangle) \mathrm{d}\mu(y)}{\int \exp(\langle \theta_1, t(x, y)\rangle) \mathrm{d}\mu(y)} - \frac{\int t(x, y) \exp(\langle \theta_2, t(x, y)\rangle) \mathrm{d}\mu(y)}{\int \exp(\langle \theta_2, t(x, y)\rangle) \mathrm{d}\mu(y)}. \end{aligned}$$

Since $|\langle \theta, t(X,Y)\rangle| \overset{a.s.}{\le} [\|\theta - \theta_\star\|_2 + \|\theta_\star\|_2]M \le [\lambda_\star^{-1/2} r + \|\theta_\star\|_2]M$ *for all* $\theta \in \Theta_r(\theta_\star)$, *it holds that* $\int \exp(\langle \theta, t(X,y)\rangle)\mathrm{d}\mu(y) \overset{a.s.}{\ge} c$ *for some* $c > 0$ *and* $\theta \in \{\theta_1, \theta_2\}$. *Now it remains to control*

$$\begin{aligned} A_1 := \Bigg\| \int t(x,y)\exp(\langle \theta_1, t(x,y)\rangle)\mathrm{d}\mu(y) \int \exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y) \\ - \int t(x,y)\exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y) \int \exp(\langle \theta_1, t(x,y)\rangle)\mathrm{d}\mu(y) \Bigg\|_2. \end{aligned}$$

By the triangle inequality, we get $A_1 \le B_1 + B_2$ *where*

$$\begin{aligned} B_1 &:= \Bigg\| \left[\int t(x,y)\exp(\langle \theta_1, t(x,y)\rangle)\mathrm{d}\mu(y) - \int t(x,y)\exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y)\right] \\ &\qquad \int \exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y) \Bigg\|_2 \\ B_2 &:= \Bigg\| \left[\int \exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y) - \int \exp(\langle \theta_1, t(x,y)\rangle)\mathrm{d}\mu(y)\right] \\ &\qquad \int t(x,y)\exp(\langle \theta_2, t(x,y)\rangle)\mathrm{d}\mu(y) \Bigg\|_2. \end{aligned}$$

Since $|\langle \theta_2, t(X,Y)\rangle|$ *and* d *is a.s. bounded,*

Remark. As a special case, the negative log-likelihood of the softmax regression with $\mathcal{X} \subset \{x \in \mathbb{R}^\tau : \|x\| \le M\}$ and $\mathcal{Y} = \{1, \dots, K\}$ is generalized self-concordant with $\nu = 2$ and $R = 2M$. In fact, the statistical model of the softmax regression is

$$p(y = k \mid x) \sim \frac{\exp\langle w_k, x\rangle}{\sum_{j=1}^K \exp\langle w_j, x\rangle}.$$

Define $\theta^\top := (w_1^\top, \dots, w_K^\top)$ and $t(x,y)^\top := (0_\tau^\top, \dots, x^\top, \dots, 0_\tau^\top)$ whose elements from $(y-1)\tau + 1$ to $y\tau$ are given by $x^\top$ and 0 elsewhere. Then we have

$$p(y = k \mid x) \sim \frac{\exp\langle \theta, t(x,k)\rangle}{\sum_{y=1}^K \exp\langle \theta, t(x,y)\rangle}.$$

The claim then follows from the example above and $\|t(x,Y)\|_2 = \|x\|_2 \le M$.

Remark. The conditional random fields (Lafferty et al., 2001) also fall into the category of generalized linear models. For simplicity, we consider a conditional random field on a chain, i.e., for $x = (x_t)_{t=1}^T$ and $y = (y_t)_{t=1}^T$,

$$p(y \mid x) \propto \exp\left\{\sum_{t=1}^{T-1} \lambda_t f_t(x, y_t, y_{t+1}) + \sum_{t=1}^{T} \mu_t g_t(x, y_t)\right\}.$$

Define $\theta^\top := (\lambda_1, \ldots, \lambda_{T-1}, \mu_1, \ldots, \mu_T)$ and

$$t(x,y)^\top := (f_1(x, y_1, y_2), \ldots, f_{T-1}(x, y_{T-1}, y_T), g_1(x, y_1), \ldots, g_T(x, y_T)).$$

Then we have

$$p(y \mid x) \sim \frac{\exp\langle\theta, t(x,y)\rangle}{\int \exp\langle\theta, t(x,\bar{y})\rangle \mathrm{d}\bar{y}}.$$

Example A.2 (Score matching with exponential families)**.** *Assume that $\mathbb{Z} = \mathbb{R}^p$. Consider an exponential family on $\mathbb{R}^d$ with densities*

$$\log p_\theta(z) = \theta^\top t(z) + h(z) - \Lambda(\theta).$$

The non-normalized density q_θ then reads $\log q_\theta(z) = \theta^\top t(z) + h(z)$. As a result, the score matching loss becomes

$$\begin{aligned}\ell(\theta; z) &= \sum_{k=1}^{p}\left[\theta^\top \frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial^2 h(z)}{\partial z_k^2} + \frac{1}{2}\left(\theta^\top \frac{\partial t(z)}{\partial z_k} + \frac{\partial h(z)}{\partial z_k}\right)^2\right] + const \\ &= \frac{1}{2}\theta^\top A(z)\theta - b(z)^\top\theta + c(z) + const,\end{aligned}$$

where $A(z) := \sum_{k=1}^p \frac{\partial t(z)}{\partial z_k}\left(\frac{\partial t(z)}{\partial z_k}\right)^\top$ is p.s.d, $b(z) := \sum_{k=1}^p \left[\frac{\partial^2 t(z)}{\partial z_k^2} + \frac{\partial h(z)}{\partial z_k}\frac{\partial t(z)}{\partial z_k}\right]$, and $c(z) := \sum_{k=1}^p \left[\frac{\partial^2 h(z)}{\partial z_k^2} + \left(\frac{\partial h(z)}{\partial z_k}\right)^2\right]$. Therefore, the score matching loss $\ell(\theta; z)$ is convex. Moreover, since the third derivatives of $\ell(\cdot; z)$ is zero, the score matching loss is generalized self-concordant for all $\nu \geq 2$ and $R \geq 0$. When the true distribution $\mathbb{P}$ is supported on the non-negative orthant $\mathbb{R}_+^p$, the score matching loss does not apply. Fortunately, a generalized score matching (Hyvärinen, 2007; Yu et al., 2019) loss can be used to address this issue. Let $w_1, \ldots, w_m : \mathbb{R}_+ \to \mathbb{R}_+$ be functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$. Then the generalized score matching loss reads

$$\ell(\theta; z) = \sum_{j=1}^{d}\left[w_j'(z_j)\partial_j \log q(z) + w_j(z_j)\partial_{jj}\log q(z) + \frac{1}{2}w_j(z_j)(\partial_j \log q(z))^2\right] + const, \quad \text{(A.14)}$$

which consists of a weighted version of the original score matching loss with weights $\{w_j(x_j)\}_{j=1}^d$ *(the last two terms in* (A.14)*) and an additional term (the first term in* (A.14)*). According to (Yu et al., 2019, Theorem 5), the loss* (A.14) *admits a quadratic form:*

$$\ell(z, Q_\theta) = \frac{1}{2}\theta^\top \bar{A}(z)\theta - \bar{b}(z)^\top \theta + \bar{c}(z) + const,$$

where $\bar{A}(z)$ *is p.s.d. Hence, it is generalized self-concordant. Note that a particular example is the pairwise graphical models studies in (Yu et al., 2016, 2020).*

Example A.3 (Generalized score matching with exponential families)**.** *When the true distribution* $\mathbb{P}$ *is supported on the non-negative orthant,* $\mathbb{R}_+^d$*, the Hyvärinen score does not apply. Hyvärinen (Hyvärinen, 2007) proposed the non-negative score matching to address this issue, which is later generalized in (Yu et al., 2019, Section 2.2). Let* $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ *be positive functions that are absolutely continuous in every bounded sub-interval of* $\mathbb{R}_+$*. Then the generalized Hyvärinen score reads*

$$\ell(z, Q) = \sum_{j=1}^d \left[h_j'(z_j)\partial_j \log q(z) + h_j(z_j)\partial_{jj} \log q(z) + \frac{1}{2} h_j(z_j)(\partial_j \log q(z))^2 \right], \tag{A.15}$$

which is a weighted version of the original Hyvärinen score with weights $\{h_j(x_j)\}_{j=1}^d$ *(the last two terms in* (A.15)*) with an additional term (the first term in* (A.15)*).*

We then consider an exponential family on $\mathbb{R}_+^d$ *with densities*

$$\log q_\theta(z) = \theta^\top t(z) - S(\theta) + b(z).$$

According to (Yu et al., 2019, Theorem 5), the score (A.15) *admits the quadratic form:*

$$\ell(z, Q_\theta) = \frac{1}{2}\theta^\top \Gamma(z)\theta - g(z)^\top \theta + C,$$

where $\Gamma(z)$ *is p.s.d. Hence, this score is self-concordant. Note that a particular example is the pairwise graphical models studies in (Yu et al., 2016, 2020).*

A.2.2 Applications to goodness-of-fit testing

Before we start, we note that a simple modification to the confidence bound in Theorem 2.7 leads to the following risk bound that can be utilized to analyze the likelihood ratio test.

Corollary A.4. *Under the same assumptions in Theorem 2.7, we have, with probability at least $1-\delta$,*

$$L(\theta_n) - L(\theta_\star) \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log{(e/\delta)} \frac{d_\star}{n}$$

whenever n satisfies (2.8).

Proof. By Taylor's expansion, we have

$$L(\theta_n) - L(\theta_\star) = S(\theta_\star)^\top (\theta_n - \theta_\star) + \frac{1}{2} \left\| \theta_n - \theta_\star \right\|_{H_n(\bar{\theta}_n)}^2$$

for some $\bar{\theta}_n \in \operatorname{Conv}\{\theta_n, \theta_\star\} \subset \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)$. By $S(\theta_\star) = 0$ and Theorem 2.7, we get

$$L(\theta_n) - L(\theta_\star) \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log{(e/\delta)} \frac{d_\star}{n}.$$

□

We begin with the type I error rates of Rao's score test, the likelihood ratio test, and the Wald test. Note that $d_\star = d$ under $\mathbf{H}_0$.

Proposition 2.13. *Suppose that Assumptions 2.3 and 2.4 with $r = 0$ hold true. Under $\mathbf{H}_0$, we have, with probability at least $1-\delta$,*

$$T_{Rao} \lesssim K_1^2 \log{(e/\delta)} \frac{d}{n}$$

whenever $n \geq 4(K_2 + 2\sigma_H^2) \log{(4d/\delta)}$. Furthermore, if Assumptions 2.2 and 2.4 with $r = K_\nu / R_\nu^\star$ hold true, we have, with probability at least $1-\delta$,

$$T_{LR}, T_{Wald} \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log{(e/\delta)} \frac{d}{n}$$

whenever n satisfies (2.8).

Proof. Under $\mathbf{H}_0$, we have $\theta_\star = \theta_0$. It then follows from Proposition 2.11 that, with probability at least $1-\delta$,

$$T_{\text{Rao}} := \|S_n(\theta_0)\|^2_{H_n^{-1}(\theta_0)} \lesssim \frac{1}{n} K_1^2 \log(e/\delta) d$$

whenever $n \geq 4(K_2 + 2\sigma_H^2)\log(4d/\delta)$.

By Taylor's theorem, there exists $\bar{\theta}_n \in \operatorname{Conv}\{\theta_n, \theta_\star\}$ such that

$$T_{\text{LR}} = 2S_n^\top(\theta_n)(\theta_0 - \theta_n) + \|\theta_0 - \theta_n\|^2_{H_n(\bar{\theta}_n)} = \|\theta_0 - \theta_n\|^2_{H_n(\bar{\theta}_n)}.$$

Following a similar argument as Theorem 2.7, we obtain, with probability at least $1-\delta$,

$$T_{\text{LR}} \lesssim K_1^2 \omega_\nu^2(\varepsilon) \log(e/\delta) \frac{d}{n}$$

whenever n satisfies (2.8). The statement for T_{Wald} follows directly from Theorem 2.7. □

We then prove the result for statistical power given in Proposition 2.14.

Proposition 2.14 (Statistical power)**.** *Let $\theta_\star \neq \theta_0$ that may depend on n. The following statements are true for sufficiently large n.*

(a) Suppose that $S(\theta_0) \neq 0$, $H(\theta_0) \succ 0$, and Assumptions 2.2 to 2.4 hold true with $r = 0$. When $\theta_\star - \theta_0 = O(n^{-1/2})$ and $\tau_n := t_n(\alpha)/4 - \|S(\theta_0)\|^2_{H(\theta_0)^{-1}} - \operatorname{Tr}(\Omega(\theta_0))/n > 0$, we have

$$\begin{aligned}&\mathbb{P}(T_{Rao} > t_n(\alpha)) \\ &\leq 2d \exp\left(-\frac{n}{4(K_2 + 2\sigma_H^2)}\right) + \exp\left(-c \min\left\{\frac{n^2\tau_n^2}{K_1^2 \|\Omega(\theta_0)\|_2^2}, \frac{n\tau_n}{K_1 \|\Omega(\theta_0)\|_\infty}\right\}\right).\end{aligned}$$

When $\theta_* - \theta_n = \omega(n^{-1/2})$, we have

$$\begin{aligned}&\mathbb{P}(T_{Rao} > t_n(\alpha)) \\ &\geq 1 - 2d \exp\left(-\frac{n}{4(K_2 + 2\sigma_H^2)}\right) - \exp\left(-c \min\left\{\frac{n^2\bar{\tau}_n^2}{K_1^2 \|\Omega(\theta_0)\|_2^2}, \frac{n\bar{\tau}_n}{K_1 \|\Omega(\theta_0)\|_\infty}\right\}\right),\end{aligned}$$

where $\bar{\tau}_n := \left[\|S(\theta_0)\|_{H(\theta_0)^{-1}} - \sqrt{3t_n(\alpha)/4}\right]^2 - \operatorname{Tr}(\Omega(\theta_0))/n$.

(b) *Suppose that the assumptions in Theorem 2.7 hold true. When* $\theta_\star - \theta_0 = O(n^{-1/2})$ *and* $\tau_n' := t_n(\alpha)/384 - \|\theta_\star - \theta_0\|_{H(\theta_\star)}^2/64 - d/n > 0$, *we have*

$$\mathbb{P}(T_{LR} > t_n(\alpha)) \leq \exp\left(-c\min\left\{\frac{n^2(\tau_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\tau_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) + \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right).$$

When $\theta_* - \theta_n = \omega(n^{-1/2})$, *we have*

$$\begin{aligned}&\mathbb{P}(T_{LR} > t_n(\alpha))\\ &\geq 1 - \exp\left(-c\min\left\{\frac{n^2(\bar{\tau}_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\bar{\tau}_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) - \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right),\end{aligned}$$

where

$$\bar{\tau}_n' := \left[\left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)}/8 - \sqrt{t_n(\alpha)}/4\right]^2 - d/n.$$

(c) *Suppose that the assumptions in Theorem 2.7 hold true. When* $\theta_\star - \theta_0 = O(n^{-1/2})$ *and* $\tau_n' := t_n(\alpha)/384 - \|\theta_\star - \theta_0\|_{H(\theta_\star)}^2/64 - d/n > 0$, *we have*

$$\mathbb{P}(T_{Wald} > t_n(\alpha)) \leq \exp\left(-c\min\left\{\frac{n^2(\tau_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\tau_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) + \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right).$$

When $\theta_* - \theta_n = \omega(n^{-1/2})$, *we have*

$$\begin{aligned}&\mathbb{P}(T_{Wald} > t_n(\alpha))\\ &\geq 1 - \exp\left(-c\min\left\{\frac{n^2(\bar{\tau}_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\bar{\tau}_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) - \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right),\end{aligned}$$

where

$$\bar{\tau}_n' := \left[\left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)}/8 - \sqrt{t_n(\alpha)}/4\right]^2 - d/n.$$

Proof of Proposition 2.14. We are mostly interested in local alternatives, i.e., $\theta_\star \to \theta_0$ as $n \to \infty$.

Rao's score test. Define four events

$$
\begin{aligned}
&\mathcal{A} := \{T_{\text{Rao}} > t_n(\alpha)\} \\
&\mathcal{B} := \left\{\frac{1}{2}H(\theta_0) \preceq H_n(\theta_0) \preceq \frac{3}{2}H(\theta_0)\right\} \\
&\mathcal{C} := \left\{4\left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}}^2 > t_n(\alpha) - 4\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}}^2\right\} \\
&\mathcal{D} := \left\{\left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}} < \left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}} - \sqrt{3t_n(\alpha)/4}\right\}.
\end{aligned}
$$

Note that

$$S(\theta_0) = S(\theta_0) - S(\theta_\star) = H(\bar{\theta})(\theta_0 - \theta_\star),$$

where $\bar{\theta} \in \operatorname{Conv}\{\theta_0, \theta_\star\}$. Due to Assumption 2.2, we have

$$e^{-R\left\|\bar{\theta}-\theta_0\right\|_2}H(\theta_0) \preceq H(\bar{\theta}) \preceq e^{R\left\|\bar{\theta}-\theta_0\right\|_2}H(\theta_0). \tag{A.16}$$

Therefore, we conclude that, as $n \to \infty$,

$$S(\theta_0) = H(\bar{\theta})(\theta_0 - \theta_\star) = \Theta(\theta_\star - \theta_0). \tag{A.17}$$

We first consider the case when $\theta_\star - \theta_0 = O(n^{-1/2})$. On the event $\mathcal{B}$, it holds that

$$T_{\text{Rao}} \leq 2\left\|S_n(\theta_0)\right\|_{H(\theta_0)^{-1}}^2 \leq 4\left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}}^2 + 4\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}}^2.$$

This implies $\mathcal{AB} \subset \mathcal{AC}$ and thus

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{AB}) + \mathbb{P}(\mathcal{AB}^c) \leq \mathbb{P}(\mathcal{AC}) + \mathbb{P}(\mathcal{B}^c) \leq \mathbb{P}(\mathcal{C}) + \mathbb{P}(\mathcal{B}^c)$$

It follows from Theorem 2.5 that, when n is large enough,

$$\mathbb{P}(\mathcal{B}^c) \leq 2d\exp\left(-\frac{n}{4(K_2 + 2\sigma_H^2)}\right).$$

Moreover, note that $\mathcal{C} = \{\left\|S_n(\theta_0) - S(\theta_0)\right\|_{H^{-1}(\theta_0)}^2 - \operatorname{Tr}(\Omega(\theta_0))/n \geq \tau_n\}$, where

$$\tau_n = t_n(\alpha)/4 - \left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}}^2 - \operatorname{Tr}(\Omega(\theta_0))/n.$$

By Theorem 2.4, we have, whenever $\tau_n > 0$,

$$\mathbb{P}(\mathcal{C}) \le \exp\left(-c\min\left\{\frac{n^2\tau_n^2}{K_1^2\left\|\Omega(\theta_0)\right\|_2^2}, \frac{n\tau_n}{K_1\left\|\Omega(\theta_0)\right\|_\infty}\right\}\right).$$

Consequently, it holds that, whenever $\tau_n > 0$ and n is large enough,

$$\mathbb{P}(\mathcal{A}) \le 2d\exp\left(-\frac{n}{4(K_2+2\sigma_H^2)}\right) + \exp\left(-c\min\left\{\frac{n^2\tau_n^2}{K_1^2\left\|\Omega(\theta_0)\right\|_2^2}, \frac{n\tau_n}{K_1\left\|\Omega(\theta_0)\right\|_\infty}\right\}\right).$$

Note that, for large enough n, it holds that $\mathrm{Tr}(\Omega(\theta_0)) \to d$ and thus $t_n(\alpha) > \mathrm{Tr}(\Omega(\theta_0))/n$. Hence, it follows from (A.17) that, as long as $\theta_\star - \theta_0 = o(n^{-1/2})$, $\tau_n > 0$ for sufficiently large n.

We then consider the case when $\theta_\star - \theta_0 = \omega(n^{-1/2})$. On the event $\mathcal{B}$, it holds that

$$T_{\text{Rao}} \ge 2\left\|S_n(\theta_0)\right\|_{H(\theta_0)^{-1}}^2/3 \ge 4[\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}} - \left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}}]^2/3.$$

By Theorem 2.4, it holds that $\left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}} = O(n^{-1/2})$. By (A.17), we know that $\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}} = \omega(n^{-1/2})$ and thus, for sufficiently large n,

$$\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}} > \left\|S_n(\theta_0) - S(\theta_0)\right\|_{H(\theta_0)^{-1}} + \sqrt{t_n(\alpha)}.$$

This implies that $\mathcal{B}\mathcal{D} \subset \mathcal{A}\mathcal{B}$ and hence

$$\mathbb{P}(\mathcal{A}) \ge \mathbb{P}(\mathcal{A}\mathcal{B}) \ge \mathbb{P}(\mathcal{B}\mathcal{D}) \ge 1 - \mathbb{P}(\mathcal{B}^c) - \mathbb{P}(\mathcal{D}^c).$$

Following a similar argument as above, we have, whenever n is large enough,

$$\mathbb{P}(\mathcal{A}) \ge 1 - 2d\exp\left(-\frac{n}{4(K_2+2\sigma_H^2)}\right) - \exp\left(-c\min\left\{\frac{n^2\bar{\tau}_n^2}{K_1^2\left\|\Omega(\theta_0)\right\|_2^2}, \frac{n\bar{\tau}_n}{K_1\left\|\Omega(\theta_0)\right\|_\infty}\right\}\right),$$

where

$$\bar{\tau}_n := \left[\left\|S(\theta_0)\right\|_{H(\theta_0)^{-1}} - \sqrt{3t_n(\alpha)/4}\right]^2 - \mathrm{Tr}(\Omega(\theta_0))/n.$$

The Wald test. Notice that $d_\star = d$ since the model is well-specified. Fix $\varepsilon = \varepsilon_\nu$ so that $\omega_\nu(\varepsilon) \le 2$. Let $\delta := \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right)$. Define the following events

$$\begin{aligned}
\mathcal{A} &:= \left\{\|S_n(\theta_\star)\|^2_{H^{-1}(\theta_\star)} \le CK_1^2 \log{(e/\delta)}\frac{d}{n}\right\} \\
\mathcal{B} &:= \left\{\frac{1}{2}H(\theta_\star) \preceq H_n(\theta_\star) \preceq \frac{3}{2}H(\theta_\star)\right\} \\
\mathcal{C} &:= \left\{\frac{1}{2\omega_\nu^2(\varepsilon)}H(\theta_\star) \preceq H_n(\theta) \preceq \frac{3}{2}\omega_\nu^2(\varepsilon)H(\theta_\star), \quad \text{for all } \theta \in \Theta_{\varepsilon/R_\nu^\star}(\theta_\star)\right\} \qquad \text{(A.18)} \\
\mathcal{D} &:= \{T_{\text{Wald}} > t_n(\alpha)\} \\
\mathcal{E} &:= \left\{\|S_n(\theta_\star)\|^2_{H(\theta_\star)^{-1}} > t_n(\alpha)/384 - \|\theta_\star - \theta_0\|^2_{H(\theta_\star)}/64\right\} \\
\mathcal{F} &:= \left\{\|S_n(\theta_\star)\|^2_{H(\theta_\star)^{-1}} < \|\theta_\star - \theta_0\|_{H(\theta_\star)}/8 - \sqrt{t_n(\alpha)}/4\right\}.
\end{aligned}$$

Following the proof of Theorem 2.7, we get $\mathbb{P}(\mathcal{ABC}) \ge 1 - \delta$ and, on the event $\mathcal{ABC}$, we have, for sufficiently large n,

$$\frac{1}{4}H(\theta_\star) \preceq \frac{1}{2\omega_\nu^2(\varepsilon)}H(\theta_\star) \preceq H_n(\theta_n) \preceq \frac{3}{2}\omega_\nu^2(\varepsilon)H(\theta_\star) \preceq 3H(\theta_\star)$$

and

$$\|\theta_n - \theta_\star\|_{H(\theta_\star)} \le 4\sqrt{2}\,\|S_n(\theta_\star)\|_{H_n(\theta_\star)^{-1}} \le 8\,\|S_n(\theta_\star)\|_{H(\theta_\star)^{-1}}. \qquad \text{(A.19)}$$

We first consider the case $\theta_\star - \theta_0 = O(n^{-1/2})$. On the event $\mathcal{ABC}$, it holds that

$$\begin{aligned}
\|\theta_n - \theta_0\|^2_{H_n(\theta_n)} &\le 3\,\|\theta_n - \theta_0\|^2_{H(\theta_\star)} \le 6\,\|\theta_n - \theta_\star\|^2_{H(\theta_\star)} + 6\,\|\theta_\star - \theta_0\|^2_{H(\theta_\star)} \\
&\le 384\,\|S_n(\theta_\star)\|^2_{H_n(\theta_\star)^{-1}} + 6\,\|\theta_\star - \theta_0\|^2_{H(\theta_\star)}
\end{aligned}$$

This implies that $\mathcal{ABCD} \subset \mathcal{ABCE}$ and thus

$$\mathbb{P}(\mathcal{D}) = \mathbb{P}(\mathcal{ABCD}) + \mathbb{P}((\mathcal{ABC})^c\mathcal{D}) \le \mathbb{P}(\mathcal{E}) + \mathbb{P}((\mathcal{ABC})^c).$$

Moreover, note that $\mathcal{E} = \{\|S_n(\theta_0) - S(\theta_0)\|^2_{H^{-1}(\theta_0)} - d/n \ge \tau_n'\}$, where

$$\tau_n' = t_n(\alpha)/384 - \|\theta_\star - \theta_0\|^2_{H(\theta_\star)}/64 - d/n.$$

By Theorem 2.4, we have, whenever $\tau_n' > 0$,

$$\mathbb{P}(\mathcal{E}) \le \exp\left(-c\min\left\{\frac{n^2(\tau_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\tau_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right).$$

Since $\mathbb{P}((\mathcal{ABC})^c) \le \delta = \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right)$, it holds that

$$\mathbb{P}(\mathcal{D}) \le \exp\left(-c\min\left\{\frac{n^2(\tau_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\tau_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) + \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right).$$

We then consider the case $\theta_\star - \theta_0 = \omega(n^{-1/2})$. On the event $\mathcal{ABC}$, we get

$$\left\|\theta_n - \theta_0\right\|_{H_n(\theta_n)}^2 \ge \left\|\theta_n - \theta_0\right\|_{H(\theta_\star)}^2/4 \ge [\left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)} - \left\|\theta_n - \theta_\star\right\|_{H(\theta_\star)}]^2/4.$$

According to (A.19) and the event $\mathcal{A}$, we have $\left\|\theta_n - \theta_\star\right\|_{H(\theta_\star)} = O(n^{-1})$ and thus $\left\|\theta_n - \theta_\star\right\|_{H(\theta_\star)} < \left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)}$ for sufficiently large n. As a result, it holds that

$$\left\|\theta_n - \theta_0\right\|_{H_n(\theta_n)}^2 \ge \left[\left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)} - 8\left\|S_n(\theta_\star)\right\|_{H(\theta_\star)^{-1}}\right]^2/4.$$

This implies that $\mathcal{ABCF} \subset \mathcal{ABCD}$ and thus

$$\mathbb{P}(\mathcal{D}) \ge \mathbb{P}(\mathcal{ABCD}) \ge \mathbb{P}(\mathcal{ABCF}) \ge 1 - \mathbb{P}((\mathcal{ABC})^c) - \mathbb{P}(\mathcal{F}^c).$$

Let

$$\bar{\tau}_n' := \left[\left\|\theta_\star - \theta_0\right\|_{H(\theta_\star)}/8 - \sqrt{t_n(\alpha)}/4\right]^2 - d/n.$$

It is positive for sufficiently large n since $\theta_\star - \theta_0 = \omega(n^{-1/2})$. By Theorem 2.4 and $\mathbb{P}((\mathcal{ABC})^c) \le \delta$, it holds that

$$\mathbb{P}(\mathcal{D}) \ge 1 - \exp\left(-c\min\left\{\frac{n^2(\bar{\tau}_n')^2}{K_1^2\left\|\Omega(\theta_\star)\right\|_2^2}, \frac{n\bar{\tau}_n'}{K_1\left\|\Omega(\theta_\star)\right\|_\infty}\right\}\right) - \exp\left(-c\frac{\varepsilon^2 n^{3-\nu}}{(R_\nu^\star)^2 K_1^2 d}\right).$$

The likelihood ratio test. Note that

$$\ell_n(\theta_0) - \ell_n(\theta_n) = \left\|\theta_n - \theta_0\right\|_{H_n(\bar{\theta})}^2$$

for some $\bar{\theta} \in \operatorname{Conv}\{\theta_n, \theta_0\}$. The claim can be proved with the same argument as the one for the Wald test. □

A.3 Technical Tools

In this section, we first recall and prove some key properties of generalized self-concordant functions. We then review some key results regarding the concentration of random vectors and matrices.

A.3.1 Properties of generalized self-concordant functions

Throughout this section, we let $f : \mathbb{R}^d \to \mathbb{R}$ be (R, ν)-generalized self-concordant as in Definition 2.1, where $R > 0$ and $\nu \geq 2$. For simplicity of the notation, we denote $\|\cdot\|_x := \|\cdot\|_{\nabla^2 f(x)}$. Let

$$d_\nu(x, y) := \begin{cases} R\|y - x\|_2 & \text{if } \nu = 2 \\ (\nu/2 - 1)R\|y - x\|_2^{3-\nu}\|y - x\|_x^{\nu-2} & \text{if } \nu > 2 \end{cases} \tag{A.20}$$

and

$$\omega_\nu(\tau) := \begin{cases} (1 - \tau)^{-2/(\nu-2)} & \text{if } \nu > 2 \\ e^\tau & \text{if } \nu = 2 \end{cases} \tag{A.21}$$

with $\mathrm{dom}(\omega_\nu) = \mathbb{R}$ if $\nu = 2$ and $\mathrm{dom}(\omega_\nu) = (-\infty, 1)$ if $\nu > 2$.

The next proposition gives bounds for the Hessian of f.

Proposition A.5 (Sun and Tran-Dinh (2019), Prop. 8)**.** *For any $x, y \in \mathrm{dom}(f)$, we have*

$$\frac{1}{\omega_\nu(d_\nu(x, y))}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \omega_\nu(d_\nu(x, y))\nabla^2 f(x),$$

where it holds if $d_\nu(x, y) < 1$ for the case $\nu > 2$.

We then give the bounds for function values. Define two functions

$$\bar{\omega}_\nu(\tau) := \int_0^1 \omega_\nu(t\tau)\mathrm{d}t = \begin{cases} \tau^{-1}(e^\tau - 1) & \text{if } \nu = 2 \\ -\tau^{-1}\log(1 - \tau) & \text{if } \nu = 4 \\ \frac{\nu-2}{\nu-4}\frac{1-(1-\tau)^{(\nu-4)/(\nu-2)}}{\tau} & \text{otherwise} \end{cases} \tag{A.22}$$

and

$$\bar{\bar{\omega}}_\nu(\tau) := \int_0^1 t\bar{\omega}_\nu(t\tau)\mathrm{d}t = \begin{cases} \tau^{-2}(e^\tau - \tau - 1) & \text{if } \nu = 2 \\ -\tau^{-2}[\tau + \log(1-\tau)] & \text{if } \nu = 3 \\ \tau^{-2}[(1-\tau)\log(1-\tau) + \tau] & \text{if } \nu = 4 \\ \frac{\nu-2}{\nu-4}\frac{1}{\tau}\left[\frac{\nu-2}{2(3-\nu)\tau}\left((1-\tau)^{2(3-\nu)/(2-\nu)} - 1\right) - 1\right] & \text{otherwise.} \end{cases} \tag{A.23}$$

Proposition A.6 (Sun and Tran-Dinh (2019), Prop. 10)**.** *For any $x, y \in \operatorname{dom}(f)$, we have*

$$\bar{\bar{\omega}}_\nu(-d_\nu(x,y))\,\|y-x\|_x^2 \le f(y) - f(x) - \langle \nabla f(x), y-x\rangle \le \bar{\bar{\omega}}_\nu(d_\nu(x,y))\,\|y-x\|_x^2\,,$$

where it holds if $d_\nu(x,y) < 1$ for the case $\nu > 2$.

In the following, we fix $x \in \operatorname{dom}(f)$ and assume $\nabla^2 f(x) \succ 0$. We denote $\lambda_{\min} := \lambda_{\min}(\nabla^2 f(x))$ and $\lambda_{\max} := \lambda_{\max}(\nabla^2 f(x))$. The next lemma bounds $d_\nu(x,y)$ with the local norm $\|y-x\|_x$. Let

$$R_\nu := \begin{cases} \lambda_{\min}^{-1/2} R & \text{if } \nu = 2 \\ (\nu/2-1)\lambda_{\min}^{(\nu-3)/2} R & \text{if } \nu \in (2,3] \\ (\nu/2-1)\lambda_{\max}^{(\nu-3)/2} R & \text{if } \nu > 3. \end{cases} \tag{A.24}$$

Lemma A.7. *For any $\nu \ge 2$ and $y \in \operatorname{dom}(f)$, we have*

$$d_\nu(x,y) \le R_\nu \|y-x\|_x\,. \tag{A.25}$$

Moreover, it holds that

$$\frac{1}{\omega_\nu(R_\nu \|y-x\|_x)}\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \omega_\nu(R_\nu\|y-x\|_x)\nabla^2 f(x),$$

where it holds if $R_\nu \|y-x\|_x < 1$ for the case $\nu > 2$.

Proof. Recall the definition of d_ν in (A.20). If $\nu = 2$, then, by the Cauchy-Schwarz inequality,

$$d_\nu(x,y) = R\,\|y-x\|_2 \le \left\|[\nabla^2 f(x)]^{-1/2}\right\|_2 R\,\|y-x\|_x \le \lambda_{\min}^{-1/2} R\,\|y-x\|_x\,.$$

The case $\nu > 2$ can be proved similarly. □

We then prove some useful properties for the function $\bar{\bar{\omega}}$.

Lemma A.8. *For any $\nu \geq 2$, the following statements hold true:*

(a) The function $\varphi(\tau) := \bar{\bar{\omega}}_\nu(-\tau)$ is strictly decreasing on $[0, \infty)$ with $\varphi(0) = 1/2$ and $\varphi(\tau) \geq 0$ for all $\tau \geq 0$.

(b) The function $\psi(\tau) := \bar{\bar{\omega}}_\nu(-\tau)\tau$ is strictly increasing on $[0, \infty)$ with $\psi(0) = 0$.

Proof. **(a)**. By definition, ω_ν is strictly increasing on $(-\infty, 1)$. As a result, for any $\tau \in (-\infty, 1)$,

$$\bar{\omega}'_\nu(\tau) = \int_0^1 t\omega'_\nu(t\tau)\mathrm{d}t > 0.$$

It then follows that, for any $\tau \geq 0$,

$$\varphi'(\tau) = -\bar{\bar{\omega}}'_\nu(-\tau) = -\int_0^1 t^2\bar{\omega}'_\nu(-t\tau)\mathrm{d}t < 0,$$

and thus φ is strictly decreasing on $[0, \infty)$. Note that $\omega_\nu(0) = 1$ and $\omega_\nu(\tau) > 0$ for all $\tau \in (-\infty, 1)$. It is straightforward to check that $\varphi(0) = 1/2$ and $\varphi(\tau) > 0$ for all $\tau \geq 0$.

(b) Due to (A.22), it is clear that $\tau \mapsto \tau\bar{\omega}_\nu(-\tau)$ is strictly increasing on $[0, \infty)$ and equals 0 at $\tau = 0$. Note that, for any $\tau \geq 0$,

$$\psi(\tau) = \int_0^1 t\tau\bar{\omega}_\nu(-t\tau)\mathrm{d}t = \frac{1}{\tau}\int_0^\tau t\bar{\omega}_\nu(-t)\mathrm{d}t.$$

We get

$$\psi'(\tau) = \frac{1}{\tau^2}\left[\tau^2\bar{\omega}_\nu(-\tau) - \int_0^\tau t\bar{\omega}_\nu(-t)\mathrm{d}t\right].$$

By the monotonicity of $\tau \mapsto \tau\bar{\omega}_\nu(-\tau)$, it follows that $\psi'(\tau) > 0$. □

Corollary A.9. *Let $\tau \geq 0$. For any $\nu \geq 2$, there exists $K_\nu \in (0, 1/2]$ such that*

$$\bar{\bar{\omega}}_\nu(-\tau)\tau \leq K_\nu \Rightarrow \tau < 1 + \mathbb{1}\{\nu = 2\} \text{ and } \bar{\bar{\omega}}_\nu(-\tau) \geq 1/4.$$

In particular, $K_\nu = 1/2$ if $\nu = 2$ and $K_\nu = 1/4$ if $\nu = 3$.

Proof. The existence of K_ν follows directly from the strict monotonicity of φ and ψ shown in Lemma A.8. For $\nu = 2$,

$$\bar{\bar{\omega}}_\nu(-\tau)\tau = \frac{e^{-\tau} + \tau - 1}{\tau} \le 1/2 \Rightarrow \tau < 2.$$

As a result, we have $\bar{\bar{\omega}}_\nu(-\tau) \ge 1/4$. The case for $\nu = 3$ can be proved similarly. □

Now we are ready to prove Proposition 2.3.

Proof of Proposition 2.3. Consider the level set

$$\mathcal{L}_f(f(x)) := \{y \in \mathcal{X} : f(y) \le f(x)\} \ne \emptyset.$$

Take an arbitrary $y \in \mathcal{L}_f(f(x))$. According to Proposition A.6, we have

$$0 \ge f(y) - f(x) \ge \langle \nabla f(x), y - x\rangle + \bar{\bar{\omega}}_\nu(-d_\nu(x,y)) \left\|y - x\right\|_x^2 .$$

By the Cauchy-Schwarz inequality and Lemmas A.7 and A.8, we get

$$\bar{\bar{\omega}}_\nu(-R_\nu \left\|y - x\right\|_x) \left\|y - x\right\|_x^2 \le \left\|\nabla f(x)\right\|_{H^{-1}(x)} \left\|y - x\right\|_x$$

This implies

$$\bar{\bar{\omega}}_\nu(-R_\nu \left\|y - x\right\|_x) R_\nu \left\|y - x\right\|_x \le R_\nu \left\|\nabla f(x)\right\|_{H^{-1}(x)} \le K_\nu.$$

Due to Corollary A.9, it holds that $R_\nu \left\|y - x\right\|_x < 1 + \mathbb{1}\{\nu = 2\}$ and $\bar{\bar{\omega}}_\nu(-R_\nu \left\|y - x\right\|_x) \ge 1/4$. It follows that $d_\nu(x,y) < 1 + \mathbb{1}\{\nu = 2\}$ and

$$\left\|y - x\right\|_x \le 4 \left\|\nabla f(x)\right\|_{\nabla^2 f(x)^{-1}} .$$

Hence, the level set $\mathcal{L}_f(f(x))$ is compact so that f has a minimizer $\bar{x}$. Moreover, by Proposition A.5 and $\nabla^2 f(x) \succ 0$, we obtain $\nabla^2 f(y) \succ 0$ for all $y \in \mathcal{L}_f(f(x))$. This yields that $\bar{x}$ is the unique minimizer of f and it satisfies

$$\left\|\bar{x} - x\right\|_x \le 4 \left\|\nabla f(x)\right\|_{\nabla^2 f(x)^{-1}} .$$

□

Remark A.4. *A similar result also appears in (Ostrovskii and Bach, 2021, Prop. B.4). We extend their result from $\nu \in \{2, 3\}$ to $\nu \ge 2$.*

A.3.2 Concentration of random vectors and matrices

Recall the definition of sub-Gaussian random vectors from Definition 2.3.

Remark A.5. *Let S be a sub-Gaussian random vector. When S is not mean-zero, we have*

$$\|S - \mathbb{E}[S]\|_{\psi_2} = \sup_{\|s\|_2=1} \|\langle S - \mathbb{E}[S], s\rangle\|_{\psi_2} = \sup_{\|s\|_2=1} \left\|s^\top S - \mathbb{E}[s^\top S]\right\|_{\psi_2}.$$

According to Vershynin (2018, Lemma 2.6.8), we obtain

$$\|S - \mathbb{E}[S]\|_{\psi_2} \le C \sup_{\|s\|_2=1} \left\|s^\top S\right\|_{\psi_2} = C\,\|S\|_{\psi_2},$$

where C is an absolute constant.

It follows from Vershynin (2018, Eq. (2.17)) that a bounded random vector is sub-Gaussian.

Lemma A.10. *Let S be a random vector such that $\|S\|_2 \le M$ for some constant $M > 0$. Then X is sub-Gaussian with $\|X\|_{\psi_2} \le M/\sqrt{\log 2}$.*

As a direct consequence of Vershynin (2018, Prop. 2.6.1), the sum of i.i.d. sub-Gaussian random vectors is also sub-Gaussian.

Lemma A.11. *Let $S_1, \dots, S_n$ be i.i.d. random vectors, then we have $\|\sum_{i=1}^n S_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|S_i\|_{\psi_2}^2$.*

Recall the definition of the matrix Bernstein condition from Definition 2.4. The next lemma, which follows from Wainwright (2019, Eq. (6.30)), shows that a matrix with bounded spectral norm satisfies the matrix Bernstein condition.

Lemma A.12. *Let H be a zero-mean random matrix such that $\|H\|_2 \le M$ for some constant $M > 0$. Then H satisfies the matrix Bernstein condition with $b = M$ and $\sigma_H^2 = \|\mathbb{V}\mathrm{ar}(H)\|_2$. Moreover, $\sigma_H^2 \le 2M^2$.*

Appendix B

APPENDIX TO CHAPTER 3

B.1 f-Divergence: Review and Examples

We review the definition of f-divergences and give a few examples.

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let $P, Q \in \mathcal{M}_1(\mathcal{X})$ be dominated by some measure $\mu \in \mathcal{M}_1(\mathcal{X})$ with densities p and q, respectively. The f-divergence generated by f is

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x)\,,$$

with the convention that $f(0) := \lim_{t\to 0^+} f(t)$ and $0f(p/0) = pf^*(0)$, where $f^*(0) = \lim_{x\to 0^+} xf(1/x) \in [0, \infty]$. Hence, $D_f(P\|Q)$ can be rewritten as

$$D_f(P\|Q) = \int_{q>0} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x) + f^*(0)P[q=0]\,,$$

with the agreement that the last term is zero if $P[q = 0] = 0$ no matter what value $f^*(0)$ takes (which could be infinity). For any $c \in \mathbb{R}$, it holds that $D_{f_c}(P\|Q) = D_f(P\|Q)$ where $f_c(t) = f(t) + c(t-1)$. Hence, we also assume, w.l.o.g., that $f(t) \geq 0$ for all $t \in (0, \infty)$. To summarize, f is convex and nonnegative with $f(1) = 0$. As a result, f is non-increasing on $(0, 1]$ and non-decreasing on $[1, \infty)$.

The conjugate generator to f is the function $f^* : (0, \infty) \to [0, \infty)$ defined by[1]

$$f^*(t) = tf(1/t)\,,$$

where again we define $f^*(0) = \lim_{t\to 0^+} f^*(t)$. Since f^* can be constructed by the perspective transform of f, it is also convex. We can verify that $f^*(1) = 0$ and $f^*(t) \geq 0$ for all $t \in (0, \infty)$,

[1] The conjugacy between f and f^* is unrelated to the usual Fenchel or Lagrange duality in convex analysis, but is related to the perspective transform.

so it defines another divergence D_{f^*}. We call this the *conjugate divergence* to D_f since

$$D_{f^*}(P\|Q) = D_f(Q\|P)\,.$$

The divergence D_f is symmetric if and only if $f = f^*$, and we write it as $D_f(P, Q)$ to emphasize the symmetry.

Example B.1. *We illustrate a number of examples.*

(a) KL divergence: It is an f-divergence generated by $f_{\mathrm{KL}}(t) = t\log t - t + 1$.

(b) Interpolated KL divergence: For $\lambda \in (0, 1)$, the interpolated KL divergence is defined as

$$\mathrm{KL}_\lambda(P\|Q) = \mathrm{KL}((\|P)\|\lambda P + (1-\lambda)Q)\,,$$

which is a f-divergence generated by

$$f_{\mathrm{KL}(,\|\lambda)}(t) = t\log\left(\frac{t}{\lambda t + 1 - \lambda}\right) - (1-\lambda)(t-1)\,.$$

(c) Jensen-Shannon divergence: The Jensen-Shannon Divergence is defined as

$$D_{\mathrm{JS}}(P, Q) = \frac{1}{2}\mathrm{KL}_{1/2}(P\|Q) + \frac{1}{2}\mathrm{KL}_{1/2}(Q\|P).$$

More generally, we have the λ-skew Jensen-Shannon Divergence Nielsen and Bhatia (2013), which is defined for $\lambda \in (0, 1)$ as $D_{\mathrm{JS},\lambda} = \lambda\mathrm{KL}_\lambda(P\|Q) + (1-\lambda)\mathrm{KL}_{1-\lambda}(Q\|P)$. This is an f-divergence generated by

$$f_{\mathrm{JS},\lambda}(t) = \lambda t\log\left(\frac{t}{\lambda t + 1 - \lambda}\right) + (1-\lambda)\log\left(\frac{1}{\lambda t + 1 - \lambda}\right)\,.$$

Note that this is the linearized cost defined in (3.4)

(d) Frontier Integral: From Property 3.1, FI *is an f-divergence generated by*

$$f_{\mathrm{FI}}(t) = \frac{t+1}{2} - \frac{t}{t-1}\log t\,.$$

(e) Interpolated χ^2 divergence: Similar to the interpolated KL divergence, we can define the interpolated χ^2 divergence $D_{\chi^2,\lambda}$ and the corresponding convex generator $f_{\chi^2,\lambda}$ for $\lambda \in (0, 1)$ as

$$D_{\chi^2,\lambda}(P\|Q) = D_{\chi^2}(P\|\lambda P + (1-\lambda)Q)\,, \quad \textit{and,} \quad f_{\chi^2,\lambda}(t) = \frac{(t-1)^2}{\lambda t + 1 - \lambda}\,.$$

The usual Neyman and Pearson χ^2 divergences are respectively obtained in the limits $\lambda \to 1$ and $\lambda \to 0$.

(f) *Squared Le Cam distance: The squared Le Cam distance is, up to scaling, a special case of the interpolated χ^2 divergence with $\lambda = 1/2$:*

$$D_{\mathrm{LC}}(P, Q) = \frac{1}{4} D_{\chi^2, 1/2}(P \| Q)\,.$$

(g) *Squared Hellinger Distance: It is an f-divergence generated by $f_H(t) = (1 - \sqrt{t})^2$.*

B.2 Regularity Assumptions

In this section, we discuss the regularity assumptions in Assumption 3.1 required for the statistical error bounds. Throughout, we assume that $\mathcal{X}$ is a finite set (for instance, on the quantized space). We upper bound the expected error of the empirical f-divergences estimated from data.

We use the convention that all higher order derivatives of f and f^* at 0 are defined as the corresponding limits as $x \to 0^+$ (if they exist). Further, we use the notation

$$\psi(p, q) = q f(p/q) = p f^*(q/p), \tag{B.1}$$

so that $D_f(P \| Q) = \sum_{a \in \mathcal{X}} \psi(P(a), Q(a))$.

B.2.1 Examples satisfying the assumptions

We now consider the examples in Example B.1. The constants are summarized in Table B.1.

KL divergence. We have

$$f_{\mathrm{KL}}(t) = t \log t - t + 1 \quad \text{and} \quad f^*_{\mathrm{KL}}(t) = -\log t + t - 1\,.$$

We have $f(0) = 1$ but $f^*(0) = \infty$. Therefore, the KL divergence does not satisfy our assumptions. Indeed, this is because the KL divergence can be unbounded.

Table B.1: Examples of f-divergences and whether they satisfy Assumptions **(A1)**-**(A3)**. Here, $\lambda \in (0,1)$ is a parameter of the interpolated or skew divergences, and we define $\bar\lambda := 1 - \lambda$.

f-divergence	Satisfies Assumptions?	C_0	C_0^*	C_1	C_1^*	C_2	C_2^*
KL	No	1	∞				
Interpolated KL	Yes	$\bar\lambda$	$\log\frac{1}{\lambda} - \bar\lambda$	1	$\frac{\bar\lambda^2}{\lambda}$	$\frac{1}{2}$	$\frac{\bar\lambda}{8\lambda}$
JS	Yes	$\frac{1}{2}\log 2$	$\frac{1}{2}\log 2$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Skew JS	Yes	$\bar\lambda\log\frac{1}{\bar\lambda}$	$\lambda\log\frac{1}{\lambda}$	λ	$\bar\lambda$	$\frac{\lambda}{2}$	$\frac{\bar\lambda}{2}$
Frontier integral	Yes	$\frac{1}{2}$	$\frac{1}{2}$	4	4	$\frac{1}{2}$	$\frac{1}{2}$
LeCam	Yes	$\frac{1}{2}$	$\frac{1}{2}$	2	2	$\frac{8}{27}$	$\frac{8}{27}$
Interpolated χ^2	Yes	$\frac{1}{\bar\lambda}$	$\frac{1}{\lambda}$	$\frac{2}{\bar\lambda^2}$	$\frac{2}{\lambda^2}$	$\frac{4}{27\lambda\bar\lambda^2}$	$\frac{4}{27\lambda^2\bar\lambda}$
Hellinger	No	1	1	∞	∞		

Interpolated KL Divergence. Let $\lambda \in (0,1)$ be a parameter and denote $\bar\lambda = 1 - \lambda$. We have

$$f_{\mathrm{KL}(,\|\lambda)}(t) = t\log\left(\frac{t}{\lambda t + \bar\lambda}\right) - \bar\lambda(t-1) \quad \text{and} \quad f^*_{\mathrm{KL}(,\|\lambda)}(t) = -\log(\bar\lambda t + \lambda) + \bar\lambda(t-1)\,.$$

The corresponding derivatives are

$$f'_{\mathrm{KL}(,\|\lambda)}(t) = \frac{\bar\lambda}{\lambda t + \bar\lambda} + \log\left(\frac{t}{\lambda t + \bar\lambda}\right) - \bar\lambda, \qquad (f^*_{\mathrm{KL}(,\|\lambda)})'(t) = \bar\lambda - \frac{\bar\lambda}{\bar\lambda t + \lambda},$$

$$f''_{\mathrm{KL}(,\|\lambda)}(t) = \frac{\bar\lambda^2}{t(\lambda t + \bar\lambda)^2}, \qquad (f^*_{\mathrm{KL}(,\|\lambda)})''(t) = \frac{\bar\lambda^2}{(\bar\lambda t + \lambda)^2}\,.$$

Proposition B.1. *The interpolated KL divergence generated by $f_{\mathrm{KL}(,\|\lambda)}$ satisfies Assump-*

tion 3.1 with

$$C_0 = 1 - \lambda, \quad C_0^* = \log\frac{1}{\lambda} - 1 + \lambda, \quad C_1 = 1, \quad C_1^* = \frac{(1-\lambda)^2}{\lambda}, \quad C_2 = \frac{1}{2}, \quad C_2^* = \frac{1-\lambda}{8\lambda}\,.$$

Proof. First, C_0, C_0^* can be computed directly. Second, it is clear that

$$-f'_{\mathrm{KL}(,\|\lambda)}(t) = \log\frac{1}{t} + \log(\lambda t + \bar\lambda) - \frac{\bar\lambda}{\lambda t + \bar\lambda} + \bar\lambda \le \log\frac{1}{t} + \log 1 - \bar\lambda + \bar\lambda = \log\frac{1}{t}$$

for all $x \in (0,1)$. Moreover, since f is convex and $f'_{\mathrm{KL}(,\|\lambda)}(1) = 0$, it holds that $f'_{\mathrm{KL}(,\|\lambda)}(x) \le 0$ for all $x \in (0,1)$, and thus $C_1 = 1$. Next, we note that $|(f^*_{\mathrm{KL}(,\|\lambda)})'(x)| \le \bar\lambda^2/\lambda$ holds uniformly on $(0,1)$ (or equivalently that $f^*_{\mathrm{KL}(,\|\lambda)}$ is Lipschitz); this gives C_1^*. Next, we have

$$C_2 = \sup_{t>0}\left\{\frac{1}{2} t f''_{\mathrm{KL}(,\|\lambda)}(t)\right\} \le \frac{1}{2}\,,$$

since the function inside the sup is monotonic decreasing on $(0,\infty)$. Finally, we have

$$C_2^* = \sup_{t>0}\left\{\frac{1}{2} t (f^*_{\mathrm{KL}(,\|\lambda)})''(t)\right\} = \frac{\bar\lambda}{8\lambda}\,,$$

since the term inside the sup is maximized at $t = \lambda/\bar\lambda$. □

Skew Jensen-Shannon Divergence. Let $\lambda \in (0,1)$ be a parameter and $\bar\lambda = 1 - \lambda$. We have,

$$f_{\mathrm{JS},\lambda}(t) = \lambda t \log\left(\frac{t}{\lambda t + \bar\lambda}\right) + \bar\lambda \log\left(\frac{1}{\lambda t + \bar\lambda}\right) = f^*_{\mathrm{JS},1-\lambda}(t)\,.$$

Its derivatives are

$$f'_{\mathrm{JS},\lambda}(t) = \lambda \log\left(\frac{t}{\lambda t + \bar\lambda}\right) \quad \text{and} \quad f''_{\mathrm{JS},\lambda}(t) = \frac{\lambda\bar\lambda}{t(\lambda t + \bar\lambda)}\,.$$

Proposition B.2. *The λ-skew JS divergence generated by $f_{\mathrm{JS},\lambda}$ above satisfies Assumption 3.1 with*

$$C_0 = (1-\lambda)\log\frac{1}{1-\lambda}, \quad C_0^* = \lambda \log\frac{1}{\lambda}, \quad C_1 = \lambda, \quad C_1^* = 1 - \lambda, \quad C_2 = \frac{\lambda}{2}, \quad C_2^* = \frac{1-\lambda}{2}\,.$$

Proof. For C_1, we have

$$-f'_{\mathrm{JS},\lambda}(t) = \lambda \log \frac{1}{t} + \lambda \log(\lambda t + \bar{\lambda}) \le \lambda \log \frac{1}{t}$$

for $x \in (0, 1)$. Next, we have

$$C_2 = \frac{\lambda \bar{\lambda}}{2} \sup_{t>0} \frac{1}{\lambda t + \bar{\lambda}} = \frac{\lambda}{2} \,.$$

□

Frontier integral. We have

$$f_{\mathrm{FI}}(t) = \frac{t+1}{2} - \frac{t}{t-1} \log t = f^*_{\mathrm{FI}}(t) \,.$$

Its derivatives are

$$f'_{\mathrm{FI}}(t) = \frac{(1-t)(3-t) + 2\log t}{2(1-t)^2} \quad \text{and} \quad f''_{\mathrm{FI}}(t) = \frac{2t\log t - t^2 + 1}{t(1-t)^3} \,.$$

Proposition B.3. *The frontier integral satisfies Assumption 3.1 with*

$$C_0 = \frac{1}{2} = C_0^*, \quad C_1 = 1 = C_1^*, \quad C_2 = \frac{1}{2} = C_2^* \,.$$

Proof. We get C_0 by calculating the limit as $t \to 0$ using L'Hôpital's rule. For C_2, we note that the term inside the sup below is decreasing in t to get

$$C_2 = \sup_{t>0} \frac{2t\log t - t^2 + 1}{(1-t)^3} = \frac{1}{2} \,.$$

By definition,

$$f_{\mathrm{FI}}(t) = 2 \int_0^1 f_{\mathrm{JS},\lambda}(t) \mathrm{d}\lambda \,,$$

so that, by Proposition B.2,

$$-f'_{\mathrm{FI}}(t) = -2 \int_0^1 f'_{\mathrm{JS},\lambda}(t) \mathrm{d}\lambda \le 2 \int_0^1 \lambda \log \frac{1}{t} \mathrm{d}\lambda = \log \frac{1}{t} \,.$$

□

Interpolated χ^2 divergence. Let $\lambda \in (0,1)$ be a parameter and denote $\bar\lambda = 1-\lambda$. We have,

$$f_{\chi^2,\lambda}(t) = \frac{(t-1)^2}{\lambda t + 1 - \lambda} = f^*_{\chi^2, 1-\lambda}(t)\,.$$

Its derivatives are

$$f'_{\chi^2,\lambda}(t) = \frac{(t-1)(\lambda t + \bar\lambda + 1)}{(\lambda t + \bar\lambda)^2} \quad \text{and} \quad f''_{\chi^2,\lambda}(t) = \frac{2}{(\lambda t + \bar\lambda)^2}\,.$$

Proposition B.4. *For $\lambda \in (0,1)$, the interpolated χ^2-divergence satisfies Assumption 3.1 with*

$$\begin{aligned} C_0 &= \frac{1}{1-\lambda}, \quad C_0^* = \frac{1}{\lambda}, \quad C_1 = \frac{2}{(1-\lambda)^2}, \quad C_1^* = \frac{2}{\lambda^2} \\ C_2 &= \frac{4}{27\lambda(1-\lambda)^2}, \quad C_2^* = \frac{4}{27\lambda^2(1-\lambda)}\,. \end{aligned}$$

Proof. Note that $0 \ge f'_{\chi^2,\lambda}(0) = -(1+\bar\lambda)/\bar\lambda^2 \ge -2/\bar\lambda^2$ is bounded. Since $f'_{\chi^2,\lambda}$ is monotonic increasing with $f'_{\chi^2,\lambda}(1) = 0$, this gives the bound on C_1. Next, we bound

$$C_2 = \sup_{t>0} \frac{t}{(\lambda t + \bar\lambda)^3} = \frac{4}{27\lambda\bar\lambda^2}\,,$$

since the supremeum is attained at $t = \bar\lambda/(2\lambda)$. □

Squared Hellinger distance. We have,

$$f_H(t) = (1-\sqrt{t})^2 = f_H^*(t), \quad f'_H(t) = 1 - \frac{1}{\sqrt{t}}, \quad f''_H(t) = \frac{1}{2}t^{-3/2}\,.$$

The squared Hellinger divergence does not satisfy our assumptions since for $t<1$, $|f'_H(x)| \approx 1/\sqrt{t}$ diverges faster than the $\log 1/t$ rate required by Assumption **(A2)**.

B.2.2 Properties and useful lemmas

We state here some useful properties and lemmas that we use throughout the paper.

First, we express the derivatives of $\psi(p,q) = qf(p/q)$ in terms of the derivatives of f:

$$\frac{\partial \psi}{\partial p}(p,q) = f'\left(\frac{p}{q}\right) = f^*\left(\frac{q}{p}\right) - \frac{q}{p}(f^*)'\left(\frac{q}{p}\right) \tag{B.2a}$$

$$\frac{\partial \psi}{\partial q}(p,q) = f\left(\frac{p}{q}\right) - \frac{p}{q}f'\left(\frac{p}{q}\right) = (f^*)'\left(\frac{q}{p}\right) \tag{B.2b}$$

$$\frac{\partial^2 \psi}{\partial p^2}(p,q) = \frac{1}{q}f''\left(\frac{p}{q}\right) = \frac{q^2}{p^3}(f^*)''\left(\frac{q}{p}\right) \ge 0 \tag{B.2c}$$

$$\frac{\partial^2 \psi}{\partial q^2}(p,q) = \frac{p^2}{q^3}f''\left(\frac{p}{q}\right) = \frac{1}{p}(f^*)''\left(\frac{q}{p}\right) \ge 0 \tag{B.2d}$$

$$\frac{\partial^2 \psi}{\partial p \partial q}(p,q) = -\frac{p}{q^2}f''\left(\frac{p}{q}\right) = -\frac{q}{p^2}(f^*)''\left(\frac{q}{p}\right) \le 0\,, \tag{B.2e}$$

where the inequalities $f'', (f^*)'' \ge 0$ followed from convexity of f and f^* respectively.

The next lemma shows that the function ψ is nearly Lipschitz, up to a log factor. This lemma can be leveraged to directly obtain a bound on statistical error of the f-divergence in terms of the expected total variation distance, provided the probabilities are not too small.

Lemma B.5. *Suppose that f satisfies Assumption 3.1. Consider $\psi : [0,1] \times [0,1] \to [0,\infty)$ given by $\psi(p,q) = qf(p/q)$. We have, for all $p, p', q, q' \in [0,1]$ with $p \vee p' > 0$, $q \vee q' > 0$, that*

$$|\psi(p',q) - \psi(p,q)| \le \left(C_1 \max\left\{1, \log\frac{1}{p \vee p'}\right\} + C_0^* \vee C_2\right)|p - p'|$$

$$|\psi(p,q') - \psi(p,q)| \le \left(C_1^* \max\left\{1, \log\frac{1}{q \vee q'}\right\} + C_0 \vee C_2^*\right)|q - q'|\,.$$

Proof. We only prove the first inequality. The second one is identical with the use of f^* rather than f. Suppose $p' \ge p$. From the fact that ψ is convex in p together with a Taylor expansion of $\psi(\cdot, q)$ around p', we get,

$$\begin{aligned} 0 \le \psi(p,q) - \psi(p',q) - (p - p')\frac{\partial \psi}{\partial p}(p',q) &= \frac{1}{2}\int_{p'}^{p} \frac{\partial^2 \psi}{\partial p^2}(s,q)(p-s)\mathrm{d}s \\ &= -\frac{p}{2}\int_{p}^{p'} \frac{\partial^2 \psi}{\partial p^2}(s,q)\mathrm{d}s + \frac{1}{2}\int_{p}^{p'} s\frac{\partial^2 \psi}{\partial p^2}(s,q)\mathrm{d}s \\ &\le 0 + C_2(p' - p)\,, \end{aligned}$$

where we used $\partial^2\psi/\partial p^2$ is non-negative due to convexity and, by (B.2c) and Assumption **(A3)**,

$$s\frac{\partial^2\psi}{\partial p^2}(s,q) = \frac{s}{q}f''\left(s/q\right) \le 2C_2\,.$$

This yields

$$-(p'-p)\frac{\partial\psi}{\partial p}(p',q) \le \psi(p,q) - \psi(p',q) \le -(p'-p)\frac{\partial\psi}{\partial p}(p',q) + C_2(p'-p)\,.$$

We consider two cases based on the sign of $\frac{\partial\psi}{\partial p}(p',q) = f'(p/q)$ (cf. Eq. (B.2a)).

Case 1. $\frac{\partial\psi}{\partial p}(p',q) \ge 0$. Since $q \mapsto f'(p/q)$ is decreasing in q, we have

$$0 \le (p'-p)\frac{\partial\psi}{\partial p}(p',q) = (p'-p)f'(p/q) \le \lim_{q\to 0}(p'-p)f'(p/q) = (p'-p)f^*(0)\,,$$

where we used $f'(\infty) = f^*(0)$ from Lemma B.6. From Assumption **(A1)**, we get the bound

$$|\psi(p,q) - \psi(p',q)| \le (C_0^* \vee C_2)(p'-p)\,.$$

Case 2. $\frac{\partial\psi}{\partial p}(p',q) < 0$. By Assumption **(A2)**, it holds that

$$\left|\frac{\partial\psi}{\partial p}(p',q)\right| \le C_1 \max\{1, \log(q/p')\} \le C_1 \max\{1, \log(1/p')\}\,,$$

and thus

$$|\psi(p,q) - \psi(p',q)| \le \left(C_1 \max\left\{1, \log\frac{1}{p'}\right\} + C_2\right)(p'-p)\,.$$

□

With the above lemma, the estimation error of the empirical f-divergence can be upper bounded by the total variation distance between the empirical measure and its population counterpart up to a logarithmic factor, where:

$$\|P_n - P\|_{\mathrm{TV}} = \sum_{a\in\mathcal{X}} |P_n(a) - P(a)|\,. \tag{B.3}$$

Next, we state and prove a technical lemma.

Lemma B.6. *Suppose the generator f satisfies Assumptions **(A1)** and **(A2)**. Then,*

$$\lim_{t\to\infty} f'(t) = f^*(0)\,, \quad \text{and} \lim_{t\to\infty}(f^*)'(t) = f(0)\,.$$

Proof. We start by observing that

$$\lim_{t\to 0} t|f'(t)| \le C_1 \lim_{t\to 0} t \vee t \log \frac{1}{t} = 0 \,.$$

Next, a direct calculation gives

$$(f^*)'(1/t) = f(t) - tf'(t) \,,$$

so that taking the limit $t \to 0$ gives

$$\lim_{t\to\infty} (f^*)'(t) = f(0) - \lim_{t\to 0} tf'(t) = f(0) \,.$$

The proof of the other part is identical. □

B.3 Plug-in Estimator: Statistical Error

In this section, we prove the high probability concentration bound for the plug-in estimator. There are two keys steps: bounding the statistical error and giving a deviation bound.

Throughout this section, we assume that P and Q are discrete. Let $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ be two independent i.i.d. samples from P and Q, respectively. We consider the plug-in estimator of the f-divergences, i.e., $D_f(P_n \| Q_m)$. The main results are (a) an upper bound for its statistical error, and (b) a high probability concentration bound. They all hold for the linearized cost $\mathcal{L}_\lambda(P_n, Q_m)$ and the frontier integral $\mathrm{FI}(P_n, Q_m)$ due to Proposition B.2 and Proposition B.3.

B.3.1 Statistical error

Proposition B.7. *Suppose that f satisfies Assumption 3.1 and $k := |Supp(P)| \vee |Supp(Q)| \in \mathbb{N} \cup \{\infty\}$. Let $n, m \ge 3$. Let $c_1 = C_1 + C_1^*$ and $c_2 = C_2 \vee C_0^* + C_2^* \vee C_0$. We have,*

$$\begin{aligned}\mathbb{E}\,|D_f(P\|Q) - D_f(P_n\|Q_m)| \le \big(C_1 \log n + C_0^* \vee C_2\big)\alpha_n(P) + \big(C_1^* \log m + C_0 \vee C_2^*\big)\alpha_m(Q) \\ + \big(C_1 + C_0^* \vee C_2\big)\beta_n(P) + \big(C_1^* + C_0 \vee C_2^*\big)\beta_m(Q) \,, \qquad \text{(B.4)}\end{aligned}$$

where $\alpha_n(P) = \sum_{a\in\mathcal{X}} \sqrt{n^{-1}P(a)}$ *and* $\beta_n(P) = \mathbb{E}\left[\sum_{a:P_n(a)=0} P(a)\max\{1, \log(1/P(a))\}\right]$. *Furthermore, if* $k < \infty$*, then*

$$\mathbb{E}\,|D_f(P\|Q) - D_f(P_n\|Q_m)| \le \big(c_1 \log(n \wedge m) + c_2\big)\left(\sqrt{\frac{k}{n\wedge m}} + \frac{k}{n\wedge m}\right). \tag{B.5}$$

The proof relies on two key lemmas—the approximate Lipschitz lemma (Lemma B.5) and the missing mass lemma (Lemma B.9). The argument breaks into two cases in P (and analogously for Q) for each atom $a \in \mathcal{X}$:

(a) $P_n(a) > 0$: Since P_n is an empirical measure, we have that $P_n(a) \ge 1/n$. In this case the approximate Lipschitz lemma gives us the Lipschitzness in $\|P - P_n\|_{\mathrm{TV}}$ up to a factor of $\log n$.

(b) $P_n(a) = 0$: In this case, the mass corresponding to $P(a)$ is missing in the empirical measure and we directly bound its expectation following similar arguments as in the missing mass literature; see, e.g., Berend and Kontorovich (2012); McAllester and Ortiz (2003).

For the first part, we further upper bound the expected total variation distance of the plug-in estimator, which is

$$\|P_n - P\|_{\mathrm{TV}} = \sum_{a\in\mathcal{X}} |P_n(a) - P(a)|\,.$$

Lemma B.8. *Assume that* P *is discrete. For any* $n \ge 1$*, it holds that*

$$\mathbb{E}\,\|P_n - P\|_{\mathrm{TV}} \le \alpha_n(P).$$

Furthermore, if $k = |Supp(P)| < \infty$*, then*

$$\mathbb{E}\,\|P_n - P\|_{\mathrm{TV}} \le \alpha_n(P) \le \sqrt{\frac{k}{n}}\,.$$

Proof. Using Jensen's inequality, we have,

$$\begin{aligned}\mathbb{E} \sum_{a \in \mathrm{Supp}(P)} |P_n(a) - P(a)| &\le \sum_{a \in \mathrm{Supp}(P)} \sqrt{\mathbb{E}(P_n(a) - P(a))^2} \\ &= \sum_{a \in \mathrm{Supp}(P)} \sqrt{\frac{P(a)(1-P(a))}{n}} \le \alpha_n(P) \,,\end{aligned}$$

If $k < \infty$, then it follows from Jensen's inequality applied to the concave function $t \mapsto \sqrt{t}$ that

$$\frac{1}{k} \sum_{i=1}^{k} \sqrt{a_k} \le \sqrt{\frac{1}{k} \sum_{i=1}^{k} a_k} \,.$$

Hence, $\alpha_n(P) \le \sqrt{k/n}$ and it completes the proof. □

For the second part, we treat the missing mass directly.

Lemma B.9 (Missing Mass)**.** *Assume that $k = |Supp(P)| < \infty$. Then, for any $n \ge 3$,*

$$\mathbb{E}\left[\sum_{a \in \mathcal{X}} \mathbb{1}\{P_n(a) = 0\} P(a)\right] \le \frac{k}{n} \tag{B.6}$$

$$\beta_n(P) := \mathbb{E}\left[\sum_{a \in \mathcal{X}} \mathbb{1}\{P_n(a) = 0\} P(a) \left(1 \vee \log \frac{1}{P(a)}\right)\right] \le \frac{k \log n}{n} \,, \tag{B.7}$$

where $a \vee b := \max\{a, b\}$.

Proof. We prove the second inequality. The first one is identical. Note that $\mathbb{E}[\mathbb{1}\{P_n(a) = 0\}] = \mathbb{P}(P_n(a) = 0) = (1 - P(a))^n$. Therefore, the left hand side (LHS) of the second inequality is

$$\begin{aligned}\mathrm{LHS} &= \sum_{a \in \mathcal{X}} (1 - P(a))^n P(a) \max\{1, -\log P(a)\} \\ &\le \sum_{a \in \mathcal{X}} \frac{1}{n} \vee \frac{\log n}{n} = \frac{k \log n}{n} \,,\end{aligned}$$

where we used Lemma B.19 and Lemma B.20. □

Remark B.2. *According to (Berend and Kontorovich, 2012, Prop. 3), the bound k/n in* (B.6) *is tight up to a constant factor.*

Now, we are ready to prove Proposition B.7.

Proof of Proposition B.7. Define $\Delta_{n,m}(a) := \big|\psi\big(P(a), Q(a)\big) - \psi\big(P_n(a), Q_m(a)\big)\big|$. We have from the triangle inequality that

$$\Delta_{n,m}(a) \le \underbrace{\big|\psi\big(P(a), Q(a)\big) - \psi\big(P_n(a), Q(a)\big)\big|}_{=:\mathcal{T}_1(a)} + \underbrace{\big|\psi\big(P_n(a), Q(a)\big) - \psi\big(P_n(a), Q_m(a)\big)\big|}_{=:\mathcal{T}_2(a)} .$$

Since $P_n(a) = 0$ or $P_n(a) \ge 1/n$, the approximate Lipschitz lemma (Lemma B.5) gives

$$\mathcal{T}_1(a) \le \begin{cases} P(a)\left(C_1 \max\{1, \log(1/P(a))\} + C_0^* \vee C_2\right) , & \text{if } P_n(a) = 0, \\ |P(a) - P_n(a)|\,\big(C_1 \log n + C_0^* \vee C_2\big) , & \text{else.} \end{cases}$$

Consequently, Lemma B.8 yields

$$\begin{aligned} \sum_{a\in\mathcal{X}} \mathbb{E}[\mathcal{T}_1] &\le \sum_{a\in\mathcal{X}} \mathbb{E}\left[\mathbb{1}\{P_n(a) = 0\} P(a)\left(C_1 \max\{1, \log(1/P(a))\} + C_0^* \vee C_2\right)\right] \\ &\quad + \sum_{a\in\mathcal{X}} \mathbb{E}\left[|P_n(a) - P(a)|\right]\big(C_1 \log n + C_0^* \vee C_2\big) \\ &\le \left(C_1 + C_0^* \vee C_2\right)\beta_n(P) + \big(C_1 \log n + C_0^* \vee C_2\big)\alpha_n(P) . \end{aligned}$$

Since $\psi(p, q) = qf(p/q) = pf^*(q/p)$, an analogous bound holds for $\mathcal{T}_2$ with the appropriate adjustment of constants. Hence, the inequality (B.4) holds. Moreover, when $k < \infty$, the inequality (B.5) follows by invoking again Lemma B.9 and Lemma B.8. □

Invoking Proposition B.1 and Proposition B.7 for the interpolated KL divergence leads to the following result.

Proposition B.10. *Assume that $k = |Supp(P)| \vee |Supp(Q)| < \infty$. For any $\lambda \in (0, 1)$, it holds that*

$$\begin{aligned} &\mathbb{E}\left|\mathrm{KL}_\lambda(P_n \| Q_m) - \mathrm{KL}_\lambda(P \| Q)\right| \\ &\le \left[\left(1 + \frac{(1-\lambda)^2}{\lambda}\right)\log(n \wedge m) + \left(\log\frac{1}{\lambda} - 1 + \lambda\right) \vee \frac{1}{2} + (1-\lambda) \vee \frac{1-\lambda}{8\lambda}\right] \\ &\times \left(\sqrt{\frac{k}{n \wedge m}} + \frac{k}{n \wedge m}\right) . \end{aligned}$$

Moreover, for any $\lambda_{n,m} \in (0, 1/2)$,

$$\mathbb{E}\left[\sup_{\lambda\in[\lambda_{n,m},1-\lambda_{n,m}]} \{|\mathrm{KL}_\lambda(P_n\|Q_m) - \mathrm{KL}_\lambda(P\|Q)| + |\mathrm{KL}_{1-\lambda}(Q_m\|P_n) - \mathrm{KL}_{1-\lambda}(Q\|P)|\}\right]$$
$$\leq 2\left((1+1/\lambda_{n,m})\log n + \log\frac{1}{\lambda_{n,m}} \vee \frac{1}{2} + 1 \vee \frac{1}{8\lambda_{n,m}}\right)\left(\sqrt{\frac{k}{n\wedge m}} + \frac{k}{n\wedge m}\right).$$

Proof. We only prove the second inequality. The first one is a direct consequence of Proposition B.1 and Proposition B.7. From the proof of Proposition B.7 we have

$$\begin{aligned}
&|\mathrm{KL}_\lambda(P_n\|Q_m) - \mathrm{KL}_\lambda(P\|Q)| \\
\leq &\sum_{a\in\mathcal{X}} \mathbb{1}\{P_n(a)=0\}\, P(a)\,(C_1\max\{1, \log(1/P(a))\} + C_0^*\vee C_2) \\
&+ \sum_{a\in\mathcal{X}} \mathbb{1}\{Q_m(a)=0\}\, Q(a)\,(C_1^*\max\{1, \log(1/Q(a))\} + C_0\vee C_2^*) \\
&+ \sum_{a\in\mathcal{X}} \big|P(a) - P_n(a)\big|\big(C_1\log n + C_0^*\vee C_2\big) + \sum_{a\in\mathcal{X}} \big|Q(a) - Q_m(a)\big|\big(C_1^*\log m + C_0\vee C_2^*\big).
\end{aligned}$$

Note that, for the interpolated KL divergence, we have

$$\begin{aligned}
C_0 = 1-\lambda \leq 1, &\quad C_0^* = \log\frac{1}{\lambda} - 1 + \lambda \leq \log\frac{1}{\lambda_{n,m}} \\
C_1 = 1, &\quad C_1^* = \frac{(1-\lambda)^2}{\lambda} \leq \frac{1}{\lambda_{n,m}} \\
C_2 = 1/2, &\quad C_2^* = \frac{1-\lambda}{8\lambda} \leq \frac{1}{8\lambda_{n,m}}
\end{aligned}$$

for all $\lambda \in [\lambda_{n,m}, 1-\lambda_{n,m}]$. The claim then follows from the same steps of Proposition B.7. □

B.3.2 Concentration bound

We now state and prove the concentration bound for general f-divergences which satisfy our regularity assumptions. We start by considering concentration around the expectation.

Proposition B.11. *Consider the f-divergence D_f where f satisfies Assumptions **(A1)**-**(A3)**. For any $t > 0$ and any discrete distributions P, Q, we have,*

$$\mathbb{P}\left(|D_f(P_n\|Q_m) - \mathbb{E}[D_f(P_n\|Q_m)]| > \varepsilon\right) \leq 2\exp\left(-\frac{(n\wedge m)\varepsilon^2}{2(c_1\log(n\wedge m) + c_2)^2}\right),$$

where $c_1 = C_1 + C_1^*$ *and* $c_2 = C_2 \vee C_0^* + C_2^* \vee C_0$.

Proof. We first establish that D_f satisfies the bounded deviation property and then invoke McDiarmid's inequality.

We start with some notation. As before, define $\psi(p, q) = qf(p/q)$. Without loss of generality, let $\mathcal{X} = \mathrm{Supp}(P) \cup \mathrm{Supp}(Q)$. Define the function $\Phi : \mathcal{X}^{n+m} \to \mathbb{R}$ so that

$$\Phi(X_1, \cdots, X_n, Y_1, \cdots, Y_m) = D_f(P_n \| Q_m) .$$

We show the bounded deviation property of Φ. Fix some $T = (x_1, \cdots, x_n, y_1, \cdots, y_m) \in \mathcal{X}^{n+m}$ and let $T' = (x_1', \cdots, x_n', y_1', \cdots, y_m') \in \mathcal{X}^{n+m}$ be such that T and T' differ only on $x_i = a \neq a' = x_i'$. Suppose the number of occurrences of a in the x-component of T is l and of a' is l', while their corresponding y-components are mq and mq' respectively. We now have

$$\begin{aligned} |\Phi(T') - \Phi(T)| &= \left| \psi\left(\frac{s-1}{n}, q\right) - \psi\left(\frac{s}{n}, q\right) + \psi\left(\frac{s'+1}{n}, q'\right) - \psi\left(\frac{s'}{n}, q'\right) \right| \\ &\le \left| \psi\left(\frac{s-1}{n}, q\right) - \psi\left(\frac{s}{n}, q\right) \right| + \left| \psi\left(\frac{s'+1}{n}, q'\right) - \psi\left(\frac{s'}{n}, q'\right) \right| \\ &\le \frac{2}{n}(C_1 \log n + C_0^* \vee C_2) =: B_i , \end{aligned}$$

where we used the triangle inequality first and then invoked Lemma B.5. Likewise, if A and A' differ only in y_i and y_i', an analogous argument gives

$$|\Phi(T') - \Phi(T)| \le \frac{2}{m}(C_1^* \log m + C_0 \vee C_2^*) =: B_i^* .$$

With this we can use McDiarmid's inequality (cf. Theorem B.18) to bound

$$\mathbb{P}\left(|D_f(P_n \| Q_m) - \mathbb{E}[D_f(P_n \| Q_m)]| > \varepsilon\right) \le h(\varepsilon) ,$$

where

$$h(\varepsilon) = 2 \exp\left(- \frac{2\varepsilon^2}{\sum_{i=1}^n B_i^2 + \sum_{i=n+1}^{n+m} (B_i^*)^2} \right) \le 2 \exp\left(- \frac{(n \wedge m)\varepsilon^2}{2(c_1 \log (n \wedge m) + c_2)^2} \right) .$$

$\square$

Hence, the concentration bound around the population f-divergence follows directly from Proposition B.7 and Proposition B.11.

Theorem B.12. *Assume that P and Q are discrete and let $k = |Supp(P)| \vee |Supp(Q)| \in \mathbb{N} \cup \{\infty\}$. For any $\delta \in (0,1)$, it holds that, with probability at least $1 - \delta$,*

$$\begin{aligned}|D_f(P_n\|Q_m) - D_f(P\|Q)| \le \big(c_1 \log(n \wedge m) + c_2\big)\sqrt{\frac{2}{n \wedge m}} \log\frac{2}{\delta} \\ + \big(C_1 \log n + C_0^* \vee C_2\big)\alpha_n(P) + \big(C_1^* \log m + C_0 \vee C_2^*\big)\alpha_m(Q) \\ + \big(C_1 + C_0^* \vee C_2\big)\beta_n(P) + \big(C_1^* + C_0 \vee C_2^*\big)\beta_m(Q)\,.\end{aligned}$$

Furthermore, if $k < \infty$, then, with probability at least $1 - \delta$,

$$|D_f(P_n\|Q_m) - D_f(P\|Q)| \le \big(c_1 \log(n \wedge m) + c_2\big)\left(\sqrt{\frac{2}{n \wedge m}} \log\frac{2}{\delta} + \sqrt{\frac{k}{n \wedge m}} + \frac{k}{n \wedge m}\right).$$

Proof of Theorem B.12. We only prove the second inequality. The first one follows from a similar argument. According to Proposition B.7, we have

$$\begin{aligned}&|D_f(P_n\|Q_m) - \mathbb{E}[D_f(P_n\|Q_m)]| \\ &\ge |D_f(P_n\|Q_m) - D_f(P\|Q)| - |\mathbb{E}[D_f(P_n\|Q_m)] - D_f(P\|Q)| \\ &\ge |D_f(P_n\|Q_m) - D_f(P\|Q)| - \big(c_1 \log(n \wedge m) + c_2\big)\left(\sqrt{\frac{k}{n \wedge m}} + \frac{k}{n \wedge m}\right).\end{aligned}$$

By Proposition B.11, it holds that

$$\mathbb{P}\left(|D_f(P_n\|Q_m) - D_f(P\|Q)| > \varepsilon + \big(c_1 \log(n \wedge m) + c_2\big)\left(\sqrt{\frac{k}{n \wedge m}} + \frac{k}{n \wedge m}\right)\right) \le h(\epsilon)\,,$$

where

$$h(\epsilon) = 2\exp\left(-\frac{(n \wedge m)\varepsilon^2}{2(c_1 \log(n \wedge m) + c_2)^2}\right).$$

The claim then follows from setting $h(\epsilon) = \delta$ and solving for ϵ. □

B.4 Add-Constant Smoothing: Statistical Error

In this section, we apply add-constant smoothing to estimate the f-divergences and study its statistical error. All the results hold for the linearized cost $\mathcal{L}_\lambda(P_n, Q_m)$ and the frontier integral $\mathrm{FI}(P_n, Q_m)$ due to Proposition B.2 and Proposition B.3.

For notational simplicity, we assume that P and Q are supported on a common finite alphabet with size $k < \infty$. Without loss of generality, let $\mathcal{X}$ be the support. Consider $P \in \mathcal{M}_1(\mathcal{X})$ and an i.i.d. sample $\{X_i\}_{i=1}^n \sim P$. The add-constant estimator of P is defined by

$$P_{n,b}(a) = \frac{N_a + b}{n + kb}, \quad \text{for all } a \in \mathcal{X}\,,$$

where $b > 0$ is a constant and $N_a = |\{i \in [n] : X_i = a\}|$ is the number of times the symbol a appears in the sample. In practice, $b = b_a$ could be different depending on the value of N_a, but we use the same constant b for simplicity. Similarly, We define $Q_{m,b}$ with $M_a = |\{i \in [m] : Y_i = a\}|$. The goal is to upper bound the statistical error

$$\mathbb{E}\,|D_f(P\|Q) - D_f(P_{n,b}\|Q_{m,b})| \tag{B.8}$$

under Assumption 3.1.

Compared to the statistical error of the plug-in estimator, a key difference is that each entry in the add-constant estimator is at least $(n + kb)^{-1} \wedge (m + kb)^{-1}$. Hence, we can directly apply the approximate Lipschitz lemma without the need to control the missing mass part. Another difference is that the total variation distance is now between the add-constant estimator and its population counterpart, which can be bounded as follows.

Lemma B.13. *Assume that $k = Supp(P) < \infty$. Then, for any $b > 0$,*

$$\sum_{a\in\mathcal{X}} \mathbb{E}\,|P_{n,b}(a) - P(a)| \le \sum_{a\in\mathcal{X}} \frac{\sqrt{nP(a)(1 - P(a))} + bk\,|P(a) - 1/k|}{n + kb} \le \frac{\sqrt{kn} + 2b(k-1)}{n + kb}\,.$$

Proof. Note that

$$|P_{n,b}(a) - P(a)| = \left|\frac{N_a - nP(a)}{n + kb} + \frac{b(1 - kP(a))}{n + kb}\right| \le \left|\frac{N_a - nP(a)}{n + kb}\right| + \left|\frac{b(1 - kP(a))}{n + kb}\right|.$$

Using Jensen's inequality, we have

$$\sum_{a\in\mathcal{X}} \mathbb{E}\left|P_{n,b}(a) - P(a)\right| \leq \sum_{a\in\mathcal{X}} \left[\sqrt{\mathbb{E}\left|\frac{N_a - nP(a)}{n+kb}\right|^2} + \frac{c\left|1-kP(a)\right|}{n+kb}\right]$$
$$= \sum_{a\in\mathcal{X}} \left[\frac{\sqrt{nP(a)(1-P(a))}}{n+kb} + \frac{bk\left|1/k - P(a)\right|}{n+kb}\right].$$

We claim that

$$\sum_{a\in\mathcal{X}} \left|P(a) - \frac{1}{k}\right| \leq \frac{2(k-1)}{k}.$$

If this is true, we have

$$\sum_{a\in\mathcal{X}} \mathbb{E}\left|P_{n,b}(a) - P(a)\right| \leq \frac{\sqrt{kn} + 2b(k-1)}{n+kb},$$

since $\sum_{a\in\mathcal{X}} \sqrt{P(a)(1-P(a))} \leq \sqrt{k}$ It then remains to prove the claim. Take $a_1, a_2 \in \mathcal{X}$ such that $P(a_1) \geq k^{-1} \geq P(a_2)$. It is clear that

$$\left|P(a_1) - \frac{1}{k}\right| + \left|P(a_2) - \frac{1}{k}\right| \leq \left|P(a_1) + P(a_2) - \frac{1}{k}\right| + \left|P(a_2) - P(a_2) - \frac{1}{k}\right|$$
$$= P(a_1) + P(a_2).$$

Repeating this argument gives

$$\sum_{a\in\mathcal{X}} \left|P(a) - \frac{1}{k}\right| \leq 1 - \frac{1}{k} + \frac{k-1}{k} = \frac{2(k-1)}{k}.$$

□

The next proposition gives the upper bound for the statistical error of the add-constant estimator.

Proposition B.14. *Suppose that f satisfies Assumption 3.1 and $k = |\mathcal{X}| < \infty$. We have,* *for any* $n, m \geq 3$,

$$\begin{aligned}\mathbb{E}\left|D_f(P\|Q) - D_f(P_{n,b}\|Q_{m,b})\right| &\leq \left[\frac{n\alpha_n(P)}{n+kb} + \gamma_{n,k}(P)\right] \big(C_1 \log(n/b+k) + C_0^* \vee C_2\big) \\ &\quad + \left[\frac{m\alpha_m(Q)}{m+kb} + \gamma_{m,k}(Q)\right] \big(C_1^* \log(m/b+k) + C_0 \vee C_2^*\big) \\ &\leq \big(C_1 \log(n/b+k) + C_0^* \vee C_2\big) \frac{\sqrt{kn} + 2b(k-1)}{n+kb} \\ &\quad + \big(C_1^* \log(m/b+k) + C_0 \vee C_2^*\big) \frac{\sqrt{km} + 2b(k-1)}{m+kb} \,,\end{aligned}$$

where $\gamma_{n,k}(P) = (n+bk)^{-1} bk \sum_{a\in\mathcal{X}} |P(a) - 1/k|$.

Proof. Following the proof of Proposition B.7, we define

$$\Delta_{n,m}(a) := |\psi(P(a), Q(a)) - \psi(P_{n,b}(a), Q_{m,b}(a))| \,.$$

We have from the triangle inequality that

$$\Delta_{n,m}(a) \leq \underbrace{\big|\psi\big(P(a), Q(a)\big) - \psi\big(P_{n,b}(a), Q(a)\big)\big|}_{=:\mathcal{T}_1(a)} + \underbrace{\big|\psi\big(P_{n,b}(a), Q(a)\big) - \psi\big(P_{n,b}(a), Q_{m,b}(a)\big)\big|}_{=:\mathcal{T}_2(a)} \,.$$

Since $P_{n,b}(a) \geq b/(n+kb)$, the approximate Lipschitz lemma (Lemma B.5) gives

$$\mathcal{T}_1(a) \leq |P(a) - P_{n,b}(a)| \big(C_1 \log(n/b+k) + C_0^* \vee C_2\big) \,,$$

By Lemma B.13, it holds that

$$\begin{aligned}\frac{\sum_{a\in\mathcal{X}} \mathbb{E}[\mathcal{T}_1(a)]}{C_1 \log(n/b+k) + C_0^* \vee C_2} &\leq \sum_{a\in\mathcal{X}} \left[\frac{\sqrt{nP(a)}}{n+kb} + \frac{bk\,|1/k - P(a)|}{n+kb}\right] = \frac{n\alpha_n(P)}{n+kb} + \gamma_{n,k}(P) \\ &\leq \frac{\sqrt{kn} + 2b(k-1)}{n+kb} \,.\end{aligned}$$

Since $\psi(p,q) = qf(p/q) = pf^*(q/p)$, an analogous bound holds for $\mathcal{T}_2(a)$ with the appropriate adjustment of constants and the sample size. Putting these together, we get,

$$
\begin{aligned}
&\mathbb{E}\left|D_f(P\|Q) - D_f(P_{n,b}\|Q_{m,b})\right| \le \mathbb{E}\left[\sum_{a\in\mathcal{X}} |\Delta_n(a)|\right] \\
&\le \left[\frac{n\alpha_n(P)}{n+kb} + \gamma_{n,k}(P)\right] \big(C_1\log(n/b+k) + C_0^* \vee C_2\big) \\
&\quad + \left[\frac{m\alpha_m(Q)}{m+kb} + \gamma_{m,k}(Q)\right] \big(C_1^*\log(m/b+k) + C_0 \vee C_2^*\big) \\
&\le \big(C_1\log(n/b+k) + C_0^* \vee C_2\big)\frac{\sqrt{kn} + 2b(k-1)}{n+kb} \\
&\quad + \big(C_1^*\log(m/b+k) + C_0 \vee C_2^*\big)\frac{\sqrt{km} + 2b(k-1)}{m+kb}\,.
\end{aligned}
$$

□

The concentration bound for the add-constant estimator can be proved similarly.

B.5 Quantization Error

In this section, we study the quantization error of f-divergences, i.e.,

$$
\inf_{|\mathcal{S}|\le k} \left|D_f(P\|Q) - D_f(P_{\mathcal{S}}\|Q_{\mathcal{S}})\right|, \tag{B.9}
$$

where the infimum is over all partitions of $\mathcal{X}$ of size no larger than k, and $P_{\mathcal{S}}$ and $Q_{\mathcal{S}}$ are the quantized versions of P and Q according to $\mathcal{S}$, respectively. Note that we do not assume $\mathcal{X}$ to be discrete in this section. All the results hold for the linearized cost $\mathcal{L}_\lambda(P_n, Q_m)$ and the frontier integral $\mathrm{FI}(P_n, Q_m)$ due to Proposition B.2 and Proposition B.3.

Our analysis is inspired by the following result, which shows that the f-divergence can be approximated by its quantized counterpart; see, e.g., (Györfi and Nemetz, 1978, Theorem 6).

Theorem B.15. *For any $P, Q \in \mathcal{M}_1(\mathcal{X})$, it holds that*

$$
D_f(P\|Q) = \sup_{\mathcal{S}} D_f(P_{\mathcal{S}}\|Q_{\mathcal{S}}), \tag{B.10}
$$

where the supremum is over all finite partitions of $\mathcal{X}$.

The next theorem holds for general f-divergences without the requirement of Assumption 3.1.

Theorem B.16. *For any $k \ge 1$, we have*

$$\sup_{P,Q} \inf_{|\mathcal{S}| \le 2k} |D_f(P\|Q) - D_f(P_{\mathcal{S}}\|Q_{\mathcal{S}})| \le \frac{f(0) + f^*(0)}{k} \,.$$

Proof. Assume $f(0)+f^*(0) < \infty$. Otherwise, there is nothing to prove. Fix two distributions P, Q over $\mathcal{X}$. Partition the measurable space $\mathcal{X}$ into

$$\mathcal{X}_1 = \left\{ x \in \mathcal{X} \,:\, \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \le 1 \right\}, \quad \text{and,} \quad \mathcal{X}_2 = \left\{ x \in \mathcal{X} \,:\, \frac{\mathrm{d}P}{\mathrm{d}Q}(x) > 1 \right\},$$

so that

$$D_f(P\|Q) = \int_{\mathcal{X}_1} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x) + \int_{\mathcal{X}_2} f^*\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x) =: D_f^+(P\|Q) + D_{f^*}^+(Q\|P) \,.$$

We quantize $\mathcal{X}_1$ and $\mathcal{X}_2$ separately, starting with $\mathcal{X}_1$. Define sets $S_1, \cdots, S_k$ as

$$S_m = \left\{ x \in \mathcal{X}_1 \,:\, \frac{f(0)(m-1)}{k} \le f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) < \frac{f(0)m}{k} \right\},$$

where the last set S_k is also extended to include $\{x \in \mathcal{X}_1 \,:\, f((\mathrm{d}P/\mathrm{d}Q)(x)) = f(0)\}$. Since f is non-increasing on $(0, 1]$, it follows that $\sup_{x\in\mathcal{X}_1} f((\mathrm{d}P/\mathrm{d}Q)(x)) \le f(0)$. As a result, the collection $\mathcal{S} = \{S_1, \cdots, S_k\}$ is a partition of $\mathcal{X}_1$. This gives

$$\frac{f(0)}{k} \sum_{m=1}^{k} (m-1)\, Q[S_m] \le D_f^+(P\|Q) \le \frac{f(0)}{k} \sum_{m=1}^{k} m\, Q[S_m] \,. \tag{B.11}$$

Further, since f is non-increasing on $(0, 1]$, we also have

$$\frac{f(0)(m-1)}{k} \le f\left(\sup_{x\in F_m} \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \le f\left(\frac{P[F_m]}{Q[F_m]}\right) \le f\left(\inf_{x\in F_m} \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \le \frac{f(0)m}{k} \,.$$

Hence, it follows that

$$\frac{f(0)}{k} \sum_{m=1}^{k} (m-1)\, Q[S_m] \le D_f^+(P_{\mathcal{S}_1}\|Q_{\mathcal{S}_1}) \le \frac{f(0)}{k} \sum_{m=1}^{k} m\, Q[S_m] \,. \tag{B.12}$$

Putting (B.11) and (B.12) together gives

$$\inf_{|\mathcal{S}_1|\le k} \left| D_f^+(P\|Q) - D_f^+(P_{\mathcal{S}_1}\|Q_{\mathcal{S}_1})\right| \le \frac{f(0)}{k}\sum_{m=1}^{k} Q[S_m] \le \frac{f(0)}{k}\,, \tag{B.13}$$

since $\sum_{m=1}^{k} Q[S_m] = Q[\mathcal{X}_1] \le 1$. Repeating the same argument with P and Q interchanged and replacing f by f^* gives

$$\inf_{|\mathcal{S}_2|\le k} \left| D_{f^*}^+(Q\|P) - D_{f^*}^+(Q_{\mathcal{S}_2}\|P_{\mathcal{S}_2})\right| \le \frac{f^*(0)}{k}\,. \tag{B.14}$$

To complete the proof, we upper bound the infimum of $\mathcal{S}$ over all partitions of $\mathcal{X}$ with $|\mathcal{S}| = k$ by the infimum over $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ with partitions $\mathcal{S}_1$ of $\mathcal{X}_1$ and $\mathcal{S}_2$ of $\mathcal{X}_2$, and $|\mathcal{S}_1| = |\mathcal{S}_2| = k$. Now, under this partitioning, we have, $D_f^+(P_{\mathcal{S}}\|Q_{\mathcal{S}}) = D_f^+(P_{\mathcal{S}_1}\|Q_{\mathcal{S}_1})$ and $D_{f^*}^+(Q_{\mathcal{S}}\|P_{\mathcal{S}}) = D_{f^*}^+(Q_{\mathcal{S}_2}\|P_{\mathcal{S}_2})$. Putting this together with the triangle inequality, we get,

$$\begin{aligned}
&\inf_{|\mathcal{S}|\le 2k} \left| D_f(P\|Q) - D_f(P_{\mathcal{S}}\|Q_{\mathcal{S}})\right| \\
&\le \inf_{\mathcal{S}=\mathcal{S}_1\cup\mathcal{S}_2} \left\{ \left| D_f^+(P\|Q) - D_f^+(P_{\mathcal{S}}\|Q_{\mathcal{S}})\right| + \left| D_{f^*}^+(Q\|P) - D_{f^*}^+(Q_{\mathcal{S}}\|P_{\mathcal{S}})\right| \right\} \\
&= \inf_{|\mathcal{S}_1|\le k} \left| D_f^+(P\|Q) - D_f^+(P_{\mathcal{S}_1}\|Q_{\mathcal{S}_1})\right| + \inf_{|\mathcal{S}_2|\le k} \left| D_{f^*}^+(Q\|P) - D_{f^*}^+(Q_{\mathcal{S}_2}\|P_{\mathcal{S}_2})\right| \\
&\le \frac{f(0)+f^*(0)}{k}\,.
\end{aligned}$$

□

Now, combining Proposition B.7 and Theorem B.16 leads to an upper bound for the overall estimation error.

Theorem B.17. *Let $\mathcal{S}_k$ be a partition of $\mathcal{X}$ such that $|\mathcal{S}| = k \ge 2$ and its quantization error satisfies the bound in Theorem B.16, i.e.,*

$$|D_f(P\|Q) - D_f(P_{\mathcal{S}_k}\|Q_{\mathcal{S}_k})| \le \frac{f(0)+f^*(0)}{k}.$$

Then, for any $n, m \geq 3$,

$$\begin{aligned}
&\mathbb{E}\left| D_f(\hat{P}_{\mathcal{S}_k,n} \| \hat{Q}_{\mathcal{S}_k,m}) - D_f(P\|Q) \right| \\
&\leq \big(C_1 \log n + C_0^* \vee C_2\big)\alpha_n(P) + \big(C_1^* \log m + C_0 \vee C_2^*\big)\alpha_m(Q) \\
&\quad + \big(C_1 + C_0^* \vee C_2\big)\beta_n(P) + \big(C_1^* + C_0 \vee C_2^*\big)\beta_m(Q) + \frac{f(0) + f^*(0)}{k} \\
&\leq \big(c_1 \log(n \wedge m) + c_2\big)\left(\sqrt{\frac{k}{n \wedge m}} + \frac{k}{n \wedge m}\right) + \frac{f(0) + f^*(0)}{k}\,,
\end{aligned}$$

where $c_1 = C_1 + C_1^*$ and $c_2 = C_2 \vee C_0^* + C_2^* \vee C_0$.

According to Theorem B.17, a good choice of quantization level k is of order $\Theta(n^{1/3})$ which balances between the two types of errors.

B.6 Technical Lemmas

We state here some technical results used in the paper.

Theorem B.18 (McDiarmid's Inequality)**.** *Let $X_1, \cdots, X_m$ be independent random variables such that X_i has range $\mathcal{X}_i$. Let $\Phi : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function which satisfies the bounded difference property. That is, there exist constants $B_1, \cdots, B_n > 0$ such that for every $i = 1, \cdots, n$ and $(x_1, \cdots, x_n), (x_1', \cdots, x_n') \in \mathcal{X}_1 \times \cdots \mathcal{X}_n$ which differ only on the i^{th} coordinate (i.e., $x_j = x_j'$ for $j \neq i$), we have,*

$$|\Phi(x_1, \cdots, x_n) - \Phi(x_1', \cdots, x_n')| \leq B_i\,.$$

Then, for any $t > 0$, we have,

$$\mathbb{P}\left(|\Phi(X_1, \cdots, X_n) - \mathbb{E}[\Phi(X_1, \cdots, X_n)]| > t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n B_i^2}\right).$$

Property B.3. *Suppose $f : (0, \infty) \to [0, \infty)$ is convex and continuously differentiable with $f(1) = 0 = f'(1)$. Then, $f'(x) \leq 0$ for all $x \in (0, 1)$ and $f'(x) \geq 0$ for all $x \in (1, \infty)$.*

Proof. Monotonicity of f' means that we have for any $x \in (0, 1)$ and $y \in (1, \infty)$ that $f'(x) \leq f'(1) = 0 \leq f'(y)$. □

Lemma B.19. *For all $x \in (0,1)$ and $n \geq 3$, we have*

$$0 \leq (1-x)^n x \log\frac{1}{x} \leq \frac{\log n}{n}\,.$$

Proof. Let $h(x) = (1-x)^n x\log(1/x)$ be defined on $(0,1)$. Since $\lim_{x\to 0} h(x) = 0 < h(1/n)$, the global supremum does not occur as $x \to 0$. We first argue that h obtains its global maximum in $(0, 1/n]$. We calculate

$$h'(x) = (1-x)^{n-1}\left(-nx\log\frac{1}{x} + (1-x)\left(\log\frac{1}{x} - 1\right)\right) \leq (1-x)^{n-1}(1-nx)\log\frac{1}{x}\,.$$

Note that $h'(x) < 0$ for $x > 1/n$, so h is strictly decreasing on $(1/n, 1)$. Therefore, it must obtain its global maximum on $(0, 1/n]$. On this interval, we have,

$$(1-x)^n x\log\frac{1}{x} \leq x\log\frac{1}{x} \leq \frac{\log n}{n}\,,$$

since $x\log(1/x)$ is increasing on $(0, \exp(-1))$. □

The next lemma comes from (Berend and Kontorovich, 2012, Theorem 1).

Lemma B.20. *For all $x \in (0,1)$ and $n \geq 1$, we have*

$$0 \leq (1-x)^n x \leq \exp(-1)/(n+1) < 1/n\,.$$

Appendix C

APPENDIX TO CHAPTER 4

C.1 From the Schrödinger Bridge to the Optimal Transport Plan

In this section, we show that the discrete Schrödinger bridge converges to the optimal transport plan with a decaying $\varepsilon := \varepsilon_n$ as $n \to \infty$. We denote by c and C absolute constants which may change from line to line. We start by proving two useful lemmas.

Lemma C.1. *Let Z be a random vector in $\mathbb{R}^d$ whose coordinates are sub-Gaussian with parameter K. Let $\{Z_i\}_{i=1}^n$ be i.i.d. copies of Z. Then we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \|Z_i\|^2 - \mathbb{E}[\|Z\|^2] > CdK^2 h\left(\frac{\log(1/\delta)}{n}\right)\right) \le \delta,$$

where $h(t) := \max\{\sqrt{t}, t\}$.

Proof. According to Vershynin (2018, Lemma 2.7.6), we know that the k-th coordinate $Z^{(k)}$ is sub-exponential with parameter K^2 for each $k \in [d]$. Since the sub-exponential norm $\|\cdot\|_{\psi_1}$ is a norm, by the triangle inequality, it holds that

$$\left\|\|Z\|^2\right\|_{\psi_1} = \left\|\sum_{k=1}^d (Z^{(k)})^2\right\|_{\psi_1} \le \sum_{k=1}^d \left\|(Z^{(k)})^2\right\|_{\psi_1} \le dK^2.$$

By Vershynin (2018, Exercise 2.7.10), we get $\|Z\|^2 - \mathbb{E}[\|Z\|^2]$ is sub-exponential with parameter CdK^2. Now the claim follows from the Bernstein inequality. □

Lemma C.2. *Let $Z \sim \nu$ on $\mathbb{R}^d$ with $d > 4$ such that $\mathbb{E}[\exp(\gamma\|Z\|^\alpha)] < \infty$ for some $\alpha > 2$ and $\gamma > 0$. Let $\{Z_i\}_{i=1}^n$ be i.i.d. copies of Z and ν_n be the empirical measure. Fix $\delta \in (0, 1)$. For sufficiently large n, we have*

$$\mathbb{P}\left(\mathsf{W}_2^2(\nu_n, \nu) \ge \bar{C} n^{-2/d} \log^{2/d}(1/\delta)\right) \le \delta,$$

where $\bar{C}$ and $\bar{c}$ only depend on d, α, γ, ν.

Proof. Due to Fournier and Guillin (2015, Theorem 2), for all $n \geq 1$ and $t > 0$, we have

$$\mathbb{P}\left(\mathsf{W}_2^2(\nu_n, \nu) \geq t\right) \leq \bar{C}\exp(-\bar{c}nt^{d/2})\mathbb{1}\{t \leq 1\} + \bar{C}\exp(-\bar{c}nt^{\alpha/2})\mathbb{1}\{t > 1\},$$

where $\bar{C}$ and $\bar{c}$ only depend on d, α, γ, ν. Let $t = [\log(\bar{C}/\delta)/(\bar{c}n)]^{2/d}$. For sufficiently large n, we have $t \leq 1$ and thus

$$\mathbb{P}\left(\mathsf{W}_2^2(\nu_n, \nu) \geq \bar{C}n^{-2/d}\log^{2/d}(1/\delta)\right) \leq \delta.$$

□

Now we are ready to prove Proposition 4.2. The argument is inspired by Pal and Wong (2020, Theorem 10).

Proof of Proposition 4.2. Recall that $T_\star$ is the optimal transport map from P to Q and μ_n is the discrete Schrödinger bridge defined in (4.9). For each random sample X_i, let $X_i^\star = T(X_i)$ be the image of X_i under T. For each n, let μ_n' denote the empirical distribution

$$\mu_n' = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, X_i^\star)}.$$

Since W_2 is a metric, so, by the triangle inequality,

$$\mathsf{W}_2^2(\mu_n, \mu_\star) \leq 2\mathsf{W}_2^2(\mu_n, \mu_n') + 2\mathsf{W}_2^2(\mu_n', \mu_\star). \tag{C.1}$$

Note that $\{(X_i, X_i^\star\})$ is an i.i.d. sample of $\mu_\star$. Define the following events

$$\begin{aligned}
\mathcal{E}_1 &:= \left\{\frac{1}{n}\sum_{i=1}^n \|Y_i\|^2 \leq \mathbb{E}[\|Y\|^2] + CdK^2h\left(\frac{\log(1/\delta)}{n}\right)\right\} \\
\mathcal{E}_2 &:= \left\{\frac{1}{n}\sum_{i=1}^n \|X_i^\star\|^2 \leq \mathbb{E}[\|Y\|^2] + CdK^2h\left(\frac{\log(1/\delta)}{n}\right)\right\} \\
\mathcal{E}_3 &:= \left\{\mathsf{W}_2^2(Q_n, Q) \leq \bar{C}n^{-2/d}\log^{2/d}(1/\delta)\right\} \\
\mathcal{E}_4 &:= \left\{\mathsf{W}_2^2(Q_n', Q) \leq \bar{C}n^{-2/d}\log^{2/d}(1/\delta)\right\} \\
\mathcal{E}_5 &:= \left\{\mathsf{W}_2^2(\mu_n', \mu_\star) \leq \bar{C}n^{-2/d}\log^{2/d}(1/\delta)\right\},
\end{aligned}$$

where $Q_n := n^{-1} \sum_{i=1}^n \delta_{Y_i}$ and $Q'_n := n^{-1} \sum_{i=1}^n \delta_{X_i^\star}$. By Lemmas C.1 and C.2, each of these events holds with probability at least $1-\delta/5$. Therefore, it suffices to prove the upper bound on the event $\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3\mathcal{E}_4\mathcal{E}_5$ which boils down to an upper bound for $\mathsf{W}_2^2(\mu_n, \mu'_n)$ since we already have an upper bound for $\mathsf{W}_2^2(\mu'_n, \mu_\star)$ on the event $\mathcal{E}_5$.

Step 1. Express the weights $\gamma_\varepsilon(\sigma)$ in terms of the divergence D in (4.12). Fix $\varepsilon > 0$. Recall that

$$\begin{aligned}\gamma_\varepsilon(\sigma) &:= \frac{\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n c\left(X_i, Y_{\sigma_i}\right)\right]}{\sum_{\tau\in\mathcal{S}_n}\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n c\left(X_i, Y_{\tau_i}\right)\right]} \\ &= \frac{\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n [c\left(X_i, Y_{\sigma_i}\right) - \phi(X_i) - \psi(Y_{\sigma_i})]\right]}{\sum_{\tau\in\mathcal{S}_n}\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n [c\left(X_i, Y_{\tau_i}\right) - \phi(X_i) - \psi(Y_{\tau_i})]\right]} \\ &= \frac{\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n D[Y_{\sigma_i} \mid X_i]\right]}{\sum_{\tau\in\mathcal{S}_n}\exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n D[Y_{\tau_i} \mid X_i]\right]}.\end{aligned} \tag{C.2}$$

We denote by $\omega(\sigma) = \exp\{-\varepsilon^{-1}\sum_{i=1}^n D[Y_{\sigma_i} \mid X_i]\}$ for each $\sigma \in \mathcal{S}_n$.

Step 2. Bound $\omega(\sigma)$. Since empirical measures do not depend on the labeling of indices, we will relabel $\{Y_i, i \in [n]\}$ such that

$$\frac{1}{n}\sum_{i=1}^n \|Y_i - X_i^\star\|^2 = \min_{\sigma\in\mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n \|Y_{\sigma_i} - X_i^\star\|^2 =: W_n^2. \tag{C.3}$$

That is, after the relabeling, the identity permutation id minimizes the $\mathbf{L}^2$-matching distance in (C.3).

We now use the quadratic approxmation of D from Assumption 4.1. Note that, for any $\sigma \neq id$ we have

$$\sum_{i=1}^n D[Y_{\sigma_i} \mid X_i] \geq L \sum_{i=1}^n \|Y_{\sigma_i} - X_i^\star\|^2 ,$$

giving us

$$\omega(\sigma) = \exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^n D[Y_{\sigma_i} \mid X_i]\right] \leq \exp\left(-\frac{1}{\varepsilon} L \sum_{i=1}^n \|Y_{\sigma_i} - X_i^\star\|^2\right).$$

On the other hand, by a similar argument, we can get a lower bound for the identity permutation:

$$\omega(id) = \exp\left[-\frac{1}{\varepsilon}\sum_{i=1}^{n} D[Y_i \mid X_i]\right] \geq \exp\left[-\frac{1}{\varepsilon}L'\sum_{i=1}^{n}\|Y_i - X_i^\star\|^2\right], \quad \text{by Assumption 4.1}$$
$$= \exp\left(-\frac{1}{\varepsilon}L'nW_n^2\right), \quad \text{by (C.3).}$$

Therefore, for any $\sigma \in \mathcal{S}_n$, we have

$$\frac{\omega(\sigma)}{\omega(id)} \leq \exp\left(-\frac{1}{\varepsilon}L\sum_{i=1}^{n}\|Y_{\sigma_i} - X_i^\star\|^2 + \frac{1}{\varepsilon}L'nW_n^2\right). \tag{C.4}$$

Step 3. Bound $\mathsf{W}_2^2(\mu_n, \mu_n')$. Let $\{\delta_n\}_{n\geq 1}$ be a positive decreasing sequence, to be chosen later, such that $\lim_{n\to\infty}\delta_n = 0$. Partition $\mathcal{S}_n$ into two disjoint subsets based on the $\mathbf{L}^2$-matching distance:

$$\mathcal{G}_n := \left\{\sigma \in \mathcal{S}_n : \frac{1}{n}\sum_{i=1}^{n}\|Y_{\sigma_i} - X_i^\star\|^2 \leq \delta_n\right\}, \quad \mathcal{G}_n^c = \mathcal{S}_n \setminus \mathcal{G}_n.$$

Consider $\sigma \in \mathcal{G}_n$ and the probability measures M_σ and μ_n' on $\mathbb{R}^d \times \mathbb{R}^d$. There is a coupling between them that couples the atom (X_i, Y_{σ_i}) of M_σ with the atom $(X_i, X_i^\star)$ of μ_n' with mass $1/n$. The squared Euclidean distance (in $\mathbb{R}^{2d}$) between these two atoms is exactly $\|Y_{\sigma_i} - X_i^\star\|^2$, which implies that

$$\mathsf{W}_2^2(M_\sigma, \mu_n') \leq n^{-1}\sum_{i=1}^{n}\|Y_{\sigma_i} - X_i^\star\|^2 \leq \delta_n \to 0.$$

For $\sigma \notin \mathcal{G}_n$ we have the bound

$$\begin{aligned}\mathsf{W}_2^2(M_\sigma, \mu_n') &\leq \frac{2}{n}\sum_{i=1}^{n}\|X_i^\star\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|Y_i\|^2, \quad \text{by the triangle inequality}\\ &\leq C\left[\mathbb{E}[\|Y\|^2] + dK^2h\left(\frac{\log(1/\delta)}{n}\right)\right], \quad \text{by the events } \mathcal{E}_1 \text{ and } \mathcal{E}_2\\ &=: \beta_n\end{aligned}$$

Since μ_n is the mixture of $\{M_\sigma\}$ with weights $\{\gamma_\varepsilon(\sigma)\}$, the natural mixture coupling, *i.e.*, couples the atom (X_i, Y_{σ_i}) of μ_n with the atom $(X_i, X_i^\star)$ of μ_n' with mass $\gamma_\varepsilon(\sigma)/n$ gives

$$\begin{aligned}\mathsf{W}_2^2(\mu_n, \mu_n') &\leq \sum_{\sigma\in\mathcal{S}_n}\frac{\gamma_\varepsilon(\sigma)}{n}\sum_{i=1}^{n}\|Y_{\sigma_i} - X_i^\star\|^2 \leq \delta_n\sum_{\sigma\in\mathcal{G}_n}\gamma_\varepsilon(\sigma) + \beta_n\sum_{\sigma\in\mathcal{G}_n^c}\gamma_\varepsilon(\sigma)\\ &\leq \delta_n + \beta_n\sum_{\sigma\in\mathcal{G}_n^c}\gamma_\varepsilon(\sigma).\end{aligned} \tag{C.5}$$

We then control $\sum_{\sigma \in \mathcal{G}_n^c} \gamma_\varepsilon(\sigma)$.

To this end, note that from (C.2), we have

$$\begin{aligned}
\sum_{\sigma \in \mathcal{G}_n^c} \gamma_\varepsilon(\sigma) &= \frac{\sum_{\sigma \in \mathcal{G}_n^c} \omega(\sigma)}{\omega(id) + \sum_{\sigma \neq id} \omega(\sigma)} \leq \sum_{\sigma \in \mathcal{G}_n^c} \frac{\omega(\sigma)}{\omega(id)} \\
&\leq \sum_{\sigma \in \mathcal{G}_n^c} \exp\left(-\frac{1}{\varepsilon} L \sum_{i=1}^n \|Y_{\sigma_i} - X_i^\star\|^2 + \frac{1}{\varepsilon} L' n W_n^2 \right), \quad \text{by (C.4)} \qquad \text{(C.6)} \\
&\leq n! \exp\left[\frac{1}{\varepsilon} L' n W_n^2 - \frac{1}{\varepsilon} L n \delta_n \right],
\end{aligned}$$

where the last inequality uses the crude estimate $|\mathcal{G}_n^c| \leq |\mathcal{S}_n| = n!$ as well as the fact that, for $\sigma \in \mathcal{G}_n^c$,

$$\sum_{i=1}^n \|Y_{\sigma_i} - X_i^\star\|^2 > n\delta_n.$$

Step 4. Bound W_n^2 in (C.3). We now let $\varepsilon = \varepsilon_n$ depend on n. By the trivial bound $n! \leq n^n$, we can bound (C.6) above by

$$\exp\left[\frac{1}{\varepsilon_n} L' n W_n^2 - \frac{1}{\varepsilon_n} L n \delta_n + n \log n \right]. \qquad \text{(C.7)}$$

We will choose δ_n and ε_n suitably such that (C.7) tends to zero exponentially fast as $n \to \infty$. Before that, let us obtain a bound for W_n^2. By the triangle inequality, we have

$$W_n^2 \leq 2\left[\mathsf{W}_2^2(Q_n, Q) + \mathsf{W}_2^2(Q_n', Q) \right] \leq \bar{C} n^{-2/d} \log^{2/d}(1/\delta). \qquad \text{(C.8)}$$

where the last inequality follows from the events $\mathcal{E}_3$ and $\mathcal{E}_4$. Combining everything, it follows from (C.5) that

$$\mathsf{W}_2^2(\mu_n, \mu_n') \leq \delta_n + \beta_n \exp\left[\frac{\bar{C} L' n}{\varepsilon_n} \Big(\frac{\log(1/\delta)}{n} \Big)^{2/d} - \frac{L n \delta_n}{\varepsilon_n} + n \log n \right]. \qquad \text{(C.9)}$$

Step 5. Choose δ_n and ε_n. For $\varepsilon_n = n^{-2/d}$, we now choose δ_n so that the upper bound in (C.9) is minimized. For this choice of ε_n, the exponent in the upper bound reads

$$\bar{C} L' n (\log(1/\delta))^{2/d} - L n^{1+2/d} \delta_n + n \log n.$$

For this term to be negative, we choose $\delta_n = 3n^{-2/d}\log n/L$. Therefore, for all large enough n,

$$\bar{C}L'n(\log(1/\delta))^{2/d} - Ln^{1+2/d}\delta_n + n\log n \geq -n\log n,$$

and thus

$$\mathsf{W}_2^2(\mu_n, \mu_n') \leq 3n^{-2/d}\log n/L + \beta_n e^{-n\log n} \leq C_{L,d,K,Q}n^{-2/d}\log n. \tag{C.10}$$

□

C.2 Properties of the Schrödinger Bridge Statistic

We prove in this section two propositions regarding the Schrödinger bridge statistic—the continuity in Proposition 4.3 and the conditional expectation expression in Proposition 4.8.

To prove Proposition 4.3, we first establish the continuity for the Schrödinger bridge cost. In fact, Proposition 4.3 is a direct consequence of the following Lemma C.3.

Lemma C.3. *The Schrödinger bridge cost $T_\varepsilon(P,Q)$ is increasing and continuous in ε on $(0,\infty)$. Moreover, if c is bounded and continuous, then*

$$T_\infty(P,Q) := \lim_{\varepsilon\uparrow\infty} T_\varepsilon(P,Q) = \int c\, \mathrm{d}(P\otimes Q). \tag{C.11}$$

Proof. Take any $0 < \varepsilon < \varepsilon' < \infty$. By the optimality, we have

$$\int c(x,y)\mathrm{d}\mu_\varepsilon(x,y) + \varepsilon\,\mathrm{KL}(\mu_\varepsilon\|P\otimes Q) \leq \int c(x,y)\mathrm{d}\mu_{\varepsilon'}(x,y) + \varepsilon\,\mathrm{KL}(\mu_{\varepsilon'}\|P\otimes Q),$$

and thus

$$T_{\varepsilon'}(P,Q) - T_\varepsilon(P,Q) \geq \varepsilon\left[\mathrm{KL}(\mu_\varepsilon\|P\otimes Q) - \mathrm{KL}(\mu_{\varepsilon'}\|P\otimes Q)\right].$$

Similarly, we obtain

$$T_{\varepsilon'}(P,Q) - T_\varepsilon(P,Q) \leq \varepsilon'\left[\mathrm{KL}(\mu_\varepsilon\|P\otimes Q) - \mathrm{KL}(\mu_{\varepsilon'}\|P\otimes Q)\right].$$

Combining these two inequalities implies

$$\varepsilon D_{\varepsilon,\varepsilon'} \leq T_{\varepsilon'}(P,Q) - T_\varepsilon(P,Q) \leq \varepsilon' D_{\varepsilon,\varepsilon'}, \tag{C.12}$$

where $D_{\varepsilon,\varepsilon'} := [\mathrm{KL}(\mu_\varepsilon \| P \otimes Q) - \mathrm{KL}(\mu_{\varepsilon'} \| P \otimes Q)]$. It then follows from $\varepsilon < \varepsilon'$ that $D_{\varepsilon,\varepsilon'} \geq 0$, or equivalently,

$$\mathrm{KL}(\mu_{\varepsilon'} \| P \otimes Q) \leq \mathrm{KL}(\mu_\varepsilon \| P \otimes Q). \tag{C.13}$$

This yields $T_\varepsilon(P, Q) \leq T_{\varepsilon'}(P, Q)$ and thus $T_\varepsilon(P, Q)$ is increasing in ε.

To prove the continuity, it suffices to show that $\varepsilon' D_{\varepsilon,\varepsilon'} - \varepsilon D_{\varepsilon,\varepsilon'} \to 0$ as $\varepsilon' \to \varepsilon$ for any $\varepsilon \in (0, \infty)$. Fix $\varepsilon \in (0, \infty)$ and consider a neighborhood of ε so that $|\varepsilon' - \varepsilon| \leq \varepsilon/2$. Note that

$$\begin{aligned} |\varepsilon' D_{\varepsilon,\varepsilon'} - \varepsilon D_{\varepsilon,\varepsilon'}| = |(\varepsilon' - \varepsilon) D_{\varepsilon,\varepsilon'}| &\leq |\varepsilon' - \varepsilon| \max\{\mathrm{KL}(\mu_\varepsilon \| P \otimes Q), \mathrm{KL}(\mu_{\varepsilon'} \| P \otimes Q)\} \\ &\leq |\varepsilon' - \varepsilon| \, \mathrm{KL}(\mu_{\varepsilon/2} \| P \otimes Q), \quad \text{by (C.13)}. \end{aligned}$$

Now the claim follows from the fact that $\mathrm{KL}(\mu_{\varepsilon/2} \| P \otimes Q) < \infty$.

We then study the limit of T_ε as $\varepsilon \to \infty$. Let $C := \sup_{\nu \in \Pi(P,Q)} \int c \mathrm{d}\nu < \infty$ since c is bounded. Note that

$$\sup_{\nu \in \Pi(P,Q)} \left| \frac{1}{\varepsilon} \int c \mathrm{d}\nu + \mathrm{KL}(\nu \| P \otimes Q) - \mathrm{KL}(\nu \| P \otimes Q) \right| \leq \frac{C}{\varepsilon} < \infty.$$

It follows that

$$\inf_{\nu \in \Pi(P,Q)} \left[\frac{1}{\varepsilon} \int c \mathrm{d}\nu + \mathrm{KL}(\nu \| P \otimes Q) \right] \to \inf_{\nu \in \Pi(P,Q)} \mathrm{KL}(\nu \| P \otimes Q) = 0, \quad \text{as } \varepsilon \to \infty.$$

Furthermore, the problem on the LHS has a unique minimizer μ_ε and the one one the RHS has a unique minimizer $\mu_\infty := P \otimes Q$. Due to the tightness of $\Pi(P, Q)$ (Santambrogio, 2015, Theorem 1.7) and Prokhorov's theorem, every sequence of measures in $\{\mu_\varepsilon\}$ has a weakly converging subsequence whose limit must be μ_∞. Hence, the equality (C.11) follows from the definition of weak convergence. □

We then prove Proposition 4.8.

Proof of Proposition 4.8. For simplicity of the notation, let $\bar{\eta}(X, Y_\sigma) := \frac{1}{n} \sum_{i=1}^n \eta(X_i, Y_{\sigma_i})$ for each $\sigma \in \mathcal{S}_n$. By exchangeability of $\{(X_i, Y_i)\}_{i=1}^n$, it holds that $\mathbb{E}_\mu[\eta(X_i, Y_i) \mid \mathcal{F}_n] =$

$\mathbb{E}_\mu[\eta(X_j, Y_j) \mid \mathcal{F}_n]$ for all $1 \leq i, j \leq n$ which implies that $\mathbb{E}_\mu[\eta(X_1, Y_1) \mid \mathcal{F}_n] = \mathbb{E}_\mu[\bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{F}_n]$ where I is the identity permutation. Since $\bar{\eta}(X, Y_{\mathrm{I}})$ is $\mathcal{F}_n$-measurable, it follows that $\mathbb{E}_\mu\left[\eta(X_1, Y_1) \mid \mathcal{F}_n\right] = \bar{\eta}(X, Y_{\mathrm{I}})$. By the tower property of conditional expectations,

$$h_n := \mathbb{E}_\mu\left[\eta(X_1, Y_1) \mid \mathcal{G}_n\right] = \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\eta(X_1, Y_1) \mid \mathcal{F}_n\right] \mid \mathcal{G}_n\right] = \mathbb{E}_\mu\left[\bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{G}_n\right].$$

By definition, the last expression is the a.s. unique $\mathcal{G}_n$-measurable function such that for any bounded $\mathcal{G}_n$-measurable ϕ, it holds that $\mathbb{E}_\mu[\bar{\eta}(X, Y_{\mathrm{I}})\phi] = \mathbb{E}_\mu[h_n \phi]$. By (4.26), we have

$$\begin{aligned}\mathbb{E}_\mu\left[\bar{\eta}(X, Y_{\mathrm{I}})\phi\right] &= \mathbb{E}\left[f_n \bar{\eta}(X, Y_{\mathrm{I}})\phi\right] = \mathbb{E}\left[\mathbb{E}\left[f_n \bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{G}_n\right]\phi\right] \\ &= \mathbb{E}_\mu\left[\frac{dR^n}{dS^n}\,\mathbb{E}\left[f_n \bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{G}_n\right]\phi\right],\end{aligned}$$

which implies that $h_n = \frac{dR^n}{dS^n}\,\mathbb{E}\left[f_n \bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{G}_n\right]$. Similar to (4.27), we have

$$\mathbb{E}\left[f_n \bar{\eta}(X, Y_{\mathrm{I}}) \mid \mathcal{G}_n\right] = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \bar{\eta}(X, Y_\sigma) \xi^{\otimes}(X, Y_\sigma)$$

Now, according to Fact 4.5,

$$h_n = \frac{1}{D_n} \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \bar{\eta}(X, Y_\sigma) \xi^{\otimes}(X, Y_\sigma) = T_n.$$

Hence, the unbiasedness of T_n under μ follows by the tower property of conditional expectations. Now consider the reverse σ-algebra $\overline{\mathcal{G}}_n = \sigma\left(\mathcal{G}_n, (X_i, Y_i),\ i \geq N+1\right)$. Since $\{(X_i, Y_i)\}_{i \geq n+1}$ are independent of $\{(X_i, Y_i)\}_{i=1}^n$, we have $T_n = \mathbb{E}_\mu\left[\eta(X_1, Y_1) \mid \overline{\mathcal{G}}_n\right]$. Thus, $(T_n, \overline{\mathcal{G}}_n)_{n \geq 1}$ is a reverse martingale and T_n converges almost surely to $\mathbb{E}_\mu[\eta(X_1, Y_1)] = \theta$. □

C.3 First Order Chaos

We verify in this section the first order chaos given in Proposition 4.10. Before that, we give some useful properties for the operators $\mathcal{A}$ and $\mathcal{A}^*$. We denote by $\mathbb{E}_\mu$ the expectation under the model μ.

Lemma C.4. *Let $(X, Y) \sim \mu$. Under Assumption 4.2, the following statements hold true:*

(a) For any $f \in \mathbf{L}^2(P)$ and $g \in \mathbf{L}^2(Q)$, it holds $\mathbb{E}_\mu[f(X) \mid Y](y) = \mathcal{A}f(y)$ and $\mathbb{E}_\mu[g(Y) \mid X](x) = \mathcal{A}^* g(x)$. In particular, $\mathcal{A}f \in \mathbf{L}^2(Q)$ and $\mathcal{A}^* g \in \mathbf{L}^2(P)$.

(b) The largest eigenvalue of $\mathcal{A}$ and $\mathcal{A}^*$ is 1, and $\mathcal{A}\,\mathbf{1} = \mathcal{A}^*\,\mathbf{1} = \mathbf{1}$.

(c) The operator $\mathcal{A}$ maps $\mathbf{L}_0^2(P)$ to $\mathbf{L}_0^2(Q)$, and $\mathcal{A}^*$ maps $\mathbf{L}_0^2(Q)$ to $\mathbf{L}_0^2(P)$.

(d) The operators $(I - \mathcal{A}^*\mathcal{A})^{-1} : \mathbf{L}_0^2(P) \to \mathbf{L}_0^2(P)$ and $(I - \mathcal{A}\mathcal{A}^*)^{-1} : \mathbf{L}_0^2(Q) \to \mathbf{L}_0^2(Q)$ are well-defined.

(e) It holds that $\mathcal{A}(I - \mathcal{A}^*\mathcal{A})^{-1} = (I - \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}$ and $\mathcal{A}^*(I - \mathcal{A}\mathcal{A}^*)^{-1} = (I - \mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*$ on their domains defined above. Moreover, for any $f \in \mathbf{L}_0^2(P)$ and $g \in \mathbf{L}_0^2(Q)$, we have

$$\begin{aligned} &\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)(X) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)(Y) \mid X\right] = f(X) \\ &\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)(X) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)(Y) \mid X\right] = g(Y). \end{aligned} \tag{C.14}$$

Proof. (a) According to (4.29), it holds that $\mathcal{A}f(y) = \mathbb{E}_\mu[f(X) \mid Y](y)$ and thus, by Jensen's inequality,

$$\|\mathcal{A}f\|^2_{\mathbf{L}^2(Q)} = \mathbb{E}_\mu[(\mathcal{A}f)^2(Y)] = \mathbb{E}_\mu[\mathbb{E}_\mu[f(X) \mid Y]^2] \le \mathbb{E}_\mu[f^2(X)] = \|f\|^2_{\mathbf{L}^2(P)} < \infty, \tag{C.15}$$

which implies $\mathcal{A}f \in \mathbf{L}^2(Q)$. A similar argument holds for $\mathcal{A}^* g$.

(b) Since $\mu \in \Pi(P, Q)$, we get, for any $y \in \mathbb{R}^d$,

$$\mathcal{A}\,\mathbf{1}(y) = \int \mathbf{1}(x)\xi(x, y)\mathrm{d}P(x) \overset{\text{a.s.}}{=} 1.$$

This implies $(1, \mathbf{1})$ is a (eigenvalue, eigenvector) pair of $\mathcal{A}$. It then follows from (C.15) that 1 is the largest eigenvalue of $\mathcal{A}$.

(c) For any $f \in \mathbf{L}_0^2(P)$, it holds

$$\int \mathcal{A}f(y)\mathrm{d}Q(y) = \int \mathrm{d}Q(y) \int f(x)\xi(x, y)\mathrm{d}P(x) = \int f(x)\mathrm{d}P(x) = 0.$$

It then follows that $\mathcal{A}f \in \mathbf{L}_0^2(Q)$.

(d) From (b) and (c) we know $\mathcal{A}^*\mathcal{A}$ maps from $\mathbf{L}_0^2(P)$ to $\mathbf{L}_0^2(P)$ with the largest eigenvalue being 1. Recall that we assume $\mathcal{A}^*\mathcal{A}$ has positive eigenvalue gap, in other words, $\mathbf{1}$ is the only eigenfunction corresponds to the eigenvalue 1. Given $f, g \in \mathbf{L}_0^2(P)$, if $(I - \mathcal{A}^*\mathcal{A})f =$

$(I - \mathcal{A}^*\mathcal{A})g$, then $f - g = c\,\mathbf{1}$ for some constant c. Since $f - g \in \mathbf{L}_0^2(P)$ is orthogonal to $\mathbf{1}$, it holds that $f = g$ and thus $I - \mathcal{A}^*\mathcal{A}$ is injective on $\mathbf{L}_0^2(P)$. Moreover, for every $f \in \mathbf{L}_0^2(P)$,

$$\tilde{f} := \left[I + \sum_{k \geq 1} (\mathcal{A}^*\mathcal{A})^k \right] f$$

converges in $\mathbf{L}^2(P)$ and $(I - \mathcal{A}^*\mathcal{A})\tilde{f} = f$. It follows that $I - \mathcal{A}^*\mathcal{A}$ is also surjective. Therefore, $(I - \mathcal{A}^*\mathcal{A})^{-1} f$ is well-defined and is equal to $\tilde{f}$.

(e) From (d) we get, for any $f \in \mathbf{L}_0^2(P)$,

$$\mathcal{A}(I - \mathcal{A}^*\mathcal{A})^{-1} f = \mathcal{A}\left[I + \sum_{k \geq 1} (\mathcal{A}^*\mathcal{A})^k \right] f = \left[I + \sum_{k \geq 1} (\mathcal{A}\mathcal{A}^*)^k \right] \mathcal{A} f = (I - \mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} f.$$

This implies $\mathcal{A}(I - \mathcal{A}^*\mathcal{A})^{-1} = (I - \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}$. The other identity can be proved analogously. Finally, we prove the first equation in (C.14). In fact,

$$\begin{aligned}
&\mathbb{E}_\mu \left[(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)(X) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f)(Y) \mid X \right] \\
&= (I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)(X) + \mathcal{A}^*(I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f)(X) \\
&= (I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^* g)(X) + (I - \mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(g - \mathcal{A} f)(X) = f(X),
\end{aligned}$$

where the last equality follows from a simple algebra. □

Now we are ready to give the first order chaos of T_n, *i.e.*, $\mathrm{Proj}_{H_1}(T_n)$.

Proof of Proposition 4.10. By the definition of orthogonal projection, it suffices to show that, for any $i \in [n]$,

$$\mathbb{E}_\mu[T_n - \theta - \mathcal{L}_n \mid X_i] = 0 \quad \text{and} \quad \mathbb{E}_\mu[T_n - \theta - \mathcal{L}_n \mid Y_i] = 0$$

almost surely. We will prove it for X_1, and the rest of them can be proved similarly. Note that $\kappa_{1,0} \in \mathbf{L}_0^2(P)$ and $\kappa_{0,1} \in \mathbf{L}_0^2(Q)$. By (c) and (d) in Lemma C.4, we know, for every $i \in [n]$,

$$\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_i) \right] = 0.$$

It then follows from (C.14) that $\mathbb{E}_\mu[\mathcal{L}_n \mid X_1]$ is equal to

$$\frac{1}{n}\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_1) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_1) \mid X_1\right] = \frac{\kappa_{1,0}(X_1)}{n}.$$

We only need to show $\mathbb{E}_\mu[T_n - \theta \mid X_1] = \frac{1}{n}\kappa_{1,0}(X_1)$. Let $h(x) := \mathbb{E}_\mu[T_n - \theta \mid X_1](x)$. Since $T_n - \theta$ is invariant to a permutation on $\{X_i\}_{i=1}^n$, we get $\mathbb{E}_\mu[T_n - \theta \mid X_i](x) \equiv h(x)$ for all $i \in [n]$. As a result, for any $\phi \in \mathbf{L}^2(P)$, it holds that

$$\mathbb{E}_\mu\left[(T_n - \theta)\sum_{i=1}^n \phi(X_i)\right] = \sum_{i=1}^n \mathbb{E}_\mu[(T_n - \theta)\phi(X_i)] = n\mathbb{E}_\mu[h(X_1)\phi(X_1)].$$

Recall from Proposition 4.8 that $T_n = \mathbb{E}_\mu[\eta(X_1, Y_1) \mid \mathcal{G}_n]$. Since $\sum_{i=1}^n \phi(X_i)$ is $\mathcal{G}_n$ measurable, by the tower property of conditional expectation, we get

$$\mathbb{E}_\mu\left[(T_n - \theta)\sum_{i=1}^n \phi(X_i)\right] = \mathbb{E}_\mu\left[(\eta(X_1, Y_1) - \theta)\sum_{i=1}^n \phi(X_i)\right] = \mathbb{E}_\mu[\kappa_{1,0}(X_1)\phi(X_1)].$$

It follows that $\mathbb{E}_\mu[\kappa_{1,0}(X_1)\phi(X_1)] = N\mathbb{E}_\mu[h(X_1)\phi(X_1)]$. Hence, we have $h(X_1) = \frac{1}{n}\kappa_{1,0}(X_1)$. □

We then prove some properties for the operator $\mathcal{B}$, including the identity in Lemma 4.11.

Lemma C.5. *Under Assumption 4.2, the following statements hold true:*

(a) Let $(X_1, Y_1), (X_2, Y_2) \overset{i.i.d}{\sim} \mu$. *It holds that* $\mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1](x, y) = \mathcal{B}f(x, y)$ *for any* $f \in \mathbf{L}^2(P \otimes Q)$. *In particular,* $\mathcal{B}f \in \mathbf{L}^2(P \otimes Q)$.

(b) The operator $\mathcal{B}$ *maps* $\mathbf{L}_0^2(P \otimes Q)$ *to* $\mathbf{L}_0^2(P \otimes Q)$.

(c) For any $f \oplus g \in \mathbf{L}^2(P \otimes Q)$, *we have* $\mathcal{B}(f \oplus g) = \mathcal{A}^*g \oplus \mathcal{A}f$.

(d) The operator $(I + \mathcal{B})^{-1}$ *is well-defined on* $\mathbf{L}_0^2(P \otimes Q)$.

(e) For any $f \in \mathbf{L}_0^2(P)$ *and* $g \in \mathbf{L}_0^2(Q)$, *it holds that*

$$(I + \mathcal{B})^{-1}(f \oplus g) = [(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^*g)] \oplus [(I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)]. \tag{C.16}$$

Proof. (a) Let $f \in \mathbf{L}^2(P \otimes Q)$. By the definition of B, we have

$$\mathcal{B}f(x, y) = \iint f(x', y')\xi(x', y)\xi(x, y')\mathrm{d}P(x')\mathrm{d}Q(y') = \mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1](x, y).$$

By Jensen's inequality, $\|\mathcal{B}f\|^2_{\mathbf{L}^2(P\otimes Q)} = \mathbb{E}_\mu[\mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1]^2] \le \mathbb{E}_\mu[f^2(X_1, Y_2)] < \infty$, and thus $\mathcal{B}f \in \mathbf{L}^2(P \otimes Q)$.

(b) Take any $f \in \mathbf{L}^2_0(P \otimes Q)$, we have, by (a),

$$\mathbb{E}_\mu[\mathcal{B}f(X, Y)] = \mathbb{E}_\mu[\mathcal{B}f(X_2, Y_1)] = \mathbb{E}_\mu[\mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1]] = \mathbb{E}_\mu[f(X_1, Y_2)] = 0,$$

and thus $\mathcal{B}f \in \mathbf{L}^2_0(P \otimes Q)$.

(c) Recall $\mathcal{B} = \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)$. Take any $f \oplus g \in \mathbf{L}^2(P \otimes Q)$, we have

$$\mathcal{B}(f \oplus g)(x, y) = (\mathcal{A} \otimes \mathcal{A}^*)(f \oplus g)(y, x) = \mathcal{A}f(y) + \mathcal{A}^*g(x) = (\mathcal{A}^*g \oplus \mathcal{A}f)(x, y).$$

(d) Recall from Assumption 4.3 that $\mathcal{A}$ admits a singular value decomposition: $\mathcal{A}\alpha_k = s_k\beta_k$ and $\mathcal{A}^*\beta_k = s_k\alpha_k$ for all $k \ge 0$ with $s_0 = 1$ and $\alpha_0 = \beta_0 = \mathbf{1}$, where $\{\alpha_k\}$ and $\{\beta_k\}$ are orthonormal bases of $\mathbf{L}^2(P)$ and $\mathbf{L}^2(Q)$, respectively. Take any $f \in \mathbf{L}^2_0(P \otimes Q)$. According to (Berezansky and Kondratiev, 2013, Page 90), $\{\alpha_i \otimes \beta_j\}_{i,j\ge 0}$ forms an orthonormal basis of $\mathbf{L}^2(P \otimes Q)$. As a result, we get that f has an expansion

$$f = \sum_{i,j\ge 0, i+j>0} \gamma_{ij}(\alpha_i \otimes \beta_j),$$

where $\sum_{i,j\ge 0, i+j>0} \gamma_{ij}^2 < \infty$. Define a function

$$\tilde{f} := \sum_{i,j\ge 0, i+j>0} \frac{\gamma_{ij}}{1 + s_i s_j}(\alpha_i \otimes \beta_j).$$

Since $s_k \ge 0$ for all $k \ge 0$, it holds that $\tilde{f} \in \mathbf{L}^2(P \otimes Q)$. Furthermore, we have $(P \otimes Q)[\tilde{f}] = 0$ as $\alpha_i \in \mathbf{L}^2_0(P)$ and $\beta_i \in \mathbf{L}^2_0(Q)$ for all $i > 0$. This implies $\tilde{f} \in \mathbf{L}^2_0(P \otimes Q)$. Moreover, we have

$$(I + \mathcal{B})\tilde{f} = \sum_{i,j\ge 0, i+j>0} \frac{\gamma_{ij}}{1 + s_i s_j}(\alpha_i \otimes \beta_j) + \sum_{i,j\ge 0, i+j>0} \frac{\gamma_{ij}}{1 + s_i s_j} s_i s_j(\alpha_i \otimes \beta_j) = f, \tag{C.17}$$

and thus $I + \mathcal{B} : \mathbf{L}^2_0(P \otimes Q) \to \mathbf{L}^2_0(P \otimes Q)$ is a surjective. On the other hand, if $(I + \mathcal{B})f = 0$ for some $f \in \mathbf{L}^2_0(P \otimes Q)$, then we must have $\langle \mathcal{B}f, f\rangle_{\mathbf{L}^2_0(P\otimes Q)} = -\|f\|_{\mathbf{L}^2_0(P\otimes Q)}$. However, we also know $\langle \mathcal{B}f, f\rangle_{\mathbf{L}^2_0(P\otimes Q)} = \sum_{i,j\ge 0, i+j>0} s_i s_j \gamma_{ij}^2 \ge 0$. Consequently, it holds $f \equiv 0$ and thus $I + \mathcal{B}$ is also an injective. Hence, the inverse operator $(I + \mathcal{B})^{-1}$ is well-defined on $\mathbf{L}^2_0(P \otimes Q)$.

(e) Take any $f \oplus g \in \mathbf{L}_0^2(P \otimes Q)$, it follows from (d) that $(I + \mathcal{B})^{-1}(f \oplus g)$ exists. It then suffices to verify

$$(I + \mathcal{B}) \left[(I - \mathcal{A}^* \mathcal{A})^{-1}(f - \mathcal{A}^* g) \oplus (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f) \right] = f \oplus g.$$

By (c), we know

$$\begin{aligned} &\mathcal{B} \left[(I - \mathcal{A}^* \mathcal{A})^{-1}(f - \mathcal{A}^* g) \oplus (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f) \right] \\ &= \mathcal{A}^*(I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f) \oplus \mathcal{A}(I - \mathcal{A}^* \mathcal{A})^{-1}(f - \mathcal{A}^* g) \\ &= (I - \mathcal{A}^* \mathcal{A})^{-1} \mathcal{A}^*(g - \mathcal{A} f) \oplus (I - \mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A}(f - \mathcal{A}^* g), \end{aligned}$$

where the last equality follows from (e) in Lemma C.4. Consequently,

$$(I + \mathcal{B}) \left[(I - \mathcal{A}^* \mathcal{A})^{-1}(f - \mathcal{A}^* g) \oplus (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A} f) \right] = f \oplus g.$$

□

C.4 The Denominator and the Remainder

C.4.1 Hoeffding decomposition under the product measure

Definition C.1. *Given $A, B \subset [n]$, we denote by H_{AB} the subspace of $\mathbf{L}^2((P \otimes Q)^n)$ spanned by functions of the form $f(X_A, Y_B)$ such that*

$$\mathbb{E}[f(X_A, Y_B) \mid X_C, Y_D] \overset{a.s.}{=} 0, \quad \text{for all } C \subset A, D \subset B \text{ and } |C| + |D| < |A| + |B| . \tag{C.18}$$

We say such an $f(X_A, Y_B)$ is completely degenerate. In particular, when $|A| = |B| = 1$, we write $f \in \mathbf{L}_{0,0}^2(P \otimes Q)$. By definition, for distinct choices of the pair (A, B), the subspaces H_{AB} are orthogonal. Take an arbitrary mean-zero statistic $T \in \mathbf{L}_0^2((P \otimes Q)^n)$. If T can be decomposed as

$$T = \sum_{A,B \subset [n]} T_{AB}, \quad \text{with} \quad T_{AB} \in H_{AB}, \tag{C.19}$$

then we call it the Hoeffding decomposition *of T (van der Vaart, 2000, Chapter 11). Its variance can then be computed as $\mathbb{E}[T^2] = \sum_{A,B \subset [n]} \mathbb{E}[T_{AB}^2]$.*

For example, both $\tilde{\xi}(X_1, Y_1) := \xi(X_1, Y_1) - 1$ and $h(X_1, Y_1) := \tilde{\eta}(X_1, Y_1)\xi(X_1, Y_1)$ are completely degenerate according to the following lemma.

Lemma C.6. *Assume that $\xi, \eta\xi \in \mathbf{L}^2(P \otimes Q)$, then $\tilde{\xi}, \tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$.*

Proof. The claim $\tilde{\xi} \in \mathbf{L}^2_{0,0}(P \otimes Q)$ follows from $\mathbb{E}[\xi(X_i, Y_j) \mid X_i] \overset{\text{a.s.}}{=} \mathbb{E}[\xi(X_i, Y_j) \mid Y_j] \overset{\text{a.s.}}{=} 1$ for all $i, j \in [n]$ since $\mu := \xi \cdot (P \otimes Q) \in \Pi(P, Q)$. To prove the other claim, note that, by (C.14),

$$\kappa_{1,0}(x) = \int \big[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y)\big]\xi(x, y)\mathrm{d}Q(y).$$

By definition, $\kappa_{1,0}(x) = \int[\eta(x, y) - \theta]\xi(x, y)\mathrm{d}Q(y)$. This yields $\int \tilde{\eta}(x, y)\xi(x, y)\mathrm{d}Q(y) = 0$. Similarly, $\int \tilde{\eta}(x, y)\xi(x, y)\mathrm{d}P(x) = 0$ and thus $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$. □

We then derive the Hoeffding decompositions of D_n and U_n as in Proposition 4.13. We start with two useful lemmas.

Lemma C.7. *Let $A_1, A_2, B_1, B_2 \subset [n]$ be such that $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$. Assume $T_1 := f_1(X_{A_1}, Y_{B_1}) \in \mathbf{L}^2((P \otimes Q)^n)$ and $T_2 := f_2(X_{A_2}, Y_{B_2}) \in \mathbf{L}^2((P \otimes Q)^n)$ are completely degenerate. Then $T_1T_2 \in \mathbf{L}^2((P \otimes Q)^n)$ is also completely degenerate.*

Proof. Take any $A' \subset A_1 \cup A_2$ and $B' \subset B_1 \cup B_2$ such that $|A'| + |B'| < |A_1| + |A_2| + |B_1| + |B_2|$. Let $A'_1 := A' \cap A_1$, $A'_2 := A' \cap A_2$, $B'_1 := B' \cap B_1$ and $B'_2 := B' \cap B_2$. Then $A' = A'_1 \cup A'_2$ and $B' = B'_1 \cup B'_2$. Furthermore, without loss of generality, we may assume $|A'_1| + |B'_1| < |A_1| + |B_1|$. By independence, we have

$$\mathbb{E}[T_1T_2 \mid X_{A'}, Y_{B'}] = \mathbb{E}[T_1 \mid X_{A'_1}, Y_{B'_1}]\,\mathbb{E}[T_2 \mid X_{A'_2}, Y_{B'_2}] = 0,$$

since $\mathbb{E}[T_1 \mid X_{A'_1}, Y_{B'_1}] = 0$. □

Lemma C.8. *Let $A \subset [n]$ be a subset. For any $\sigma \in \mathcal{S}_n$, the following identity holds:*

$$\prod_{i \in A} \xi(X_i, Y_{\sigma_i}) = \sum_{C \subset A} \prod_{i \in C} \tilde{\xi}(X_i, Y_{\sigma_i}), \tag{C.20}$$

where $\prod_{i \in \emptyset} \tilde{\xi}(X_i, Y_{\sigma_i}) := 1$. Moreover, (C.20) gives the Hoeffding decomposition of $\prod_{i \in A} \xi(X_i, Y_{\sigma_i})$.

Proof. By Lemma C.7, $\prod_{i\in C}\tilde{\xi}(X_i, Y_{\sigma_i})$ is completely degenerate for each $C \subset A$. It then suffices to prove the identity (C.20). Without loss of generality, we prove it for $A = [n]$ by induction. For $n = 1$, the identity reduces to $\xi(X_1, Y_1) = 1 + \tilde{\xi}(X_1, Y_1)$. Assume the identity holds for $n - 1$. Consequently,

$$\begin{aligned}\prod_{i=1}^{n}\xi(X_i, Y_{\sigma_i}) &= \sum_{C\subset[n-1]}\prod_{i\in C}\tilde{\xi}(X_i, Y_{\sigma_i}) \times \xi(X_n, Y_{\sigma_n})\\ &= \sum_{C\subset[n],n\in C}\prod_{i\in C}\tilde{\xi}(X_i, Y_{\sigma_i}) + \sum_{C\subset[n-1]}\prod_{i\in C}\tilde{\xi}(X_i, Y_{\sigma_i}) = \sum_{C\subset[n]}\prod_{i\in C}\tilde{\xi}(X_i, Y_{\sigma_i}).\end{aligned}$$

Thus, the identity holds for n. □

Proof of Proposition 4.13. We only prove the results for U_n. The proof for D_n is similar. By definition,

$$\begin{aligned}U_n &:= \frac{1}{n\cdot n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i=1}^{n} h(X_i, Y_{\sigma_i}) \prod_{j\in[n]\setminus\{i\}}\xi(X_j, Y_{\sigma_j})\\ &= \frac{1}{n\cdot n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i=1}^{n} h(X_i, Y_{\sigma_i}) \sum_{C\subset[n]\setminus\{i\}}\prod_{j\in C}\tilde{\xi}(X_j, Y_{\sigma_j}), \quad \text{by Lemma C.8.}\end{aligned}$$

Take $A, B \subset [n]$ such that $|A| = |B| > 0$. We will write U_n as a sum of terms that only contain $X_A := (X_i)_{i\in A}$ and $Y_B := (Y_i)_{i\in B}$. The terms corresponding to X_A in the above decomposition are

$$\frac{1}{n\cdot n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i\in A} h(X_i, Y_{\sigma_i}) \prod_{j\in A\setminus\{i\}}\tilde{\xi}(X_j, Y_{\sigma_j}).$$

Consequently, the terms corresponding to (X_A, Y_B) are

$$\frac{1}{n\cdot n!}U_{AB} := \frac{1}{n\cdot n!}\sum_{\sigma\in\mathcal{S}_n:\sigma_A=B}\sum_{i\in A} h(X_i, Y_{\sigma_i}) \prod_{j\in A\setminus\{i\}}\tilde{\xi}(X_j, Y_{\sigma_j}).$$

Hence, the identity (4.36) follows. Moreover, since $h \in \mathbf{L}^2_{0,0}(P\otimes Q)$, we get, by Lemma C.7, that

$$h(X_i, Y_{\sigma_i}) \prod_{j\in A\setminus\{i\}}\tilde{\xi}(X_j, Y_{\sigma_j}) \in H_{AB}, \quad \text{for any } i \in A \text{ and } \sigma \in \mathcal{S}_n \text{ such that } \sigma_A = B.$$

This implies $U_{AB} \in H_{AB}$, and thus (4.36) is the Hoeffding decomposition of U_n.

Let us compute $\mathbb{E}[U_n^2]$. For any $A, B \subset [n]$ such that $|A| = |B| = r > 0$, we get, by the exchangeability $X_{[n]}$ and $Y_{[n]}$ under the measure $(P \otimes Q)^n$, $\mathbb{E}[U_{AB}^2] = \mathbb{E}[U_{[r][r]}^2]$. Furthermore, since there are $(n-r)!$ permutations that map $[r]$ to $[r]$, we get

$$\mathbb{E}[U_{[r][r]}^2] = (n-r)!^2\, \mathbb{E}\left[\sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} h(X_i, Y_{\sigma_i}) \prod_{j \in [r] \setminus \{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right]^2 .$$

As a result, $\mathbb{E}[U_{[r][r]}^2]$ is equal to

$$(n-r)!^2 \sum_{\tau \in \mathcal{S}_n} \sum_{i=1}^{r} \mathbb{E}\left[h(X_i, Y_{\tau_i}) \prod_{k \in [r] \setminus \{l\}} \tilde{\xi}(X_k, Y_{\tau_k}) \times \sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} h(X_i, Y_{\sigma_i}) \prod_{j \in [r] \setminus \{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right].$$

By symmetry, the contribution from every τ is the same, so $\mathbb{E}[U_{[r][r]}^2]$ is equal to

$$(n-r)!^2 r!\, \mathbb{E}\left[\sum_{l=1}^{r} h(X_l, Y_l) \prod_{k \in [r] \setminus \{l\}} \tilde{\xi}(X_k, Y_k) \sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} h(X_i, Y_{\sigma_i}) \prod_{j \in [r] \setminus \{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right].$$

It then follows from the exchangeability of $\{(X_i, Y_i)\}_{i \in [n]}$ that

$$\mathbb{E}[U_{[r][r]}^2] = (n-r)!^2 r! r\, \mathbb{E}\left[h(X_1, Y_1) \prod_{k=2}^{r} \tilde{\xi}(X_k, Y_k) \sum_{\sigma \in \mathcal{S}_r} \sum_{i \in A} h(X_i, Y_{\sigma_i}) \prod_{j \in A \setminus \{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right].$$

As a result,

$$\begin{aligned}
\mathbb{E}[U_n^2] &= \frac{1}{n^2 (n!)^2} \sum_{r=1}^{n} \sum_{|A|=|B|=r} \mathbb{E}[U_{AB}^2] = \frac{1}{n^2 (n!)^2} \sum_{r=1}^{n} \binom{n}{r}^2 \mathbb{E}[U_{[r][r]}^2] \\
&= \frac{1}{n^2} \sum_{r=1}^{n} \frac{r}{r!} \sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} \mathbb{E}\left[h(X_1, Y_1) \prod_{j=2}^{r} \tilde{\xi}(X_k, Y_k) h(X_i, Y_{\sigma_i}) \prod_{j \in [r] \setminus \{i\}} \tilde{\xi}(X_j, Y_{\sigma_j})\right].
\end{aligned}$$

□

C.4.2 Variance bound

We then bound the variance of D_n and U_n using the spectral gap of operators $\mathcal{A}$ and $\mathcal{A}^*$. Assumption 4.2 guarantees that such spectral gap does exist. We first prove the contraction property in Lemma 4.14.

Proof of Lemma 4.14. Take $f \in \mathbf{L}^2_{0,0}(P \otimes P)$. By definition, we have $(I_P \otimes \mathcal{A})f(x,y) = \int f(x,x')\xi(x',y)\mathrm{d}P(x')$, and thus

$$\begin{aligned}\mathbb{E}[(I_P \otimes \mathcal{A})f(X_1,Y_1) \mid X_1] &= \int f(X_1,x')\left[\int \xi(x',y)\mathrm{d}Q(y)\right]\mathrm{d}P(x') \\ &= \int f(X_1,x')\mathrm{d}P(x') \overset{\text{a.s.}}{=} 0.\end{aligned}$$

Similarly, $\mathbb{E}[(I_P \otimes \mathcal{A})f(X_1,Y_1) \mid Y_1] \overset{\text{a.s.}}{=} 0$. Consequently, $(I_P \otimes \mathcal{A})f \in \mathbf{L}^2_{0,0}(P \otimes Q)$. Now, by Berezansky and Kondratiev (2013, Page 90), $\{\alpha_i \otimes \beta_j\}_{i,j\geq 0}$ forms an orthonormal basis of $\mathbf{L}^2(P \otimes Q)$, and thus f admits the following expansion $f = \sum_{i,j\geq 1} \gamma_{ij}\alpha_i \otimes \alpha_j$ where $\sum_{i,j\geq 1}\gamma_{ij}^2 < \infty$. It then follows that

$$\|(I_P \otimes \mathcal{A})f\|^2_{\mathbf{L}^2(P\otimes Q)} = \left\|\sum_{i,j\geq 1}\gamma_{ij}s_j\alpha_i \otimes \beta_j\right\|^2_{\mathbf{L}^2(P\otimes Q)} = \sum_{i,j\geq 1}\gamma_{ij}^2 s_j^2 \leq s_1^2\|f\|^2_{\mathbf{L}^2(P\otimes P)}.$$

□

In order to control the expectation in Lemma 4.15, we decompose a permutation into disjoint cycles. By independence, the expectation then equals the product of expectations with respect to each cycle. We first give a simple example to illustrate the idea.

Example C.1. *Consider the case when $r = 3$, $i = 3$, and σ is given by $\sigma_1 = 2$, $\sigma_2 = 1$ and $\sigma_3 = 3$. We are interested in bounding the following expectation:*

$$\mathbb{E}[h(X_1,Y_1)\tilde{\xi}(X_2,Y_2)\tilde{\xi}(X_3,Y_3)h(X_3,Y_3)\tilde{\xi}(X_1,Y_2)\tilde{\xi}(X_2,Y_1)]. \tag{C.21}$$

By construction, σ contains two cycles, $1 \to 2 \to 1$ and $3 \to 3$, and the above expectation reads

$$\mathbb{E}[h(X_1,Y_1)\tilde{\xi}(X_2,Y_2)\tilde{\xi}(X_1,Y_2)\tilde{\xi}(X_2,Y_1)] \cdot \mathbb{E}[h(X_3,Y_3)\tilde{\xi}(X_3,Y_3)].$$

The second expectation is upper bounded by $\|h\|_{\mathbf{L}^2(P\otimes Q)}\left\|\tilde{\xi}\right\|_{\mathbf{L}^2(P\otimes Q)}$ by the Cauchy-Schwarz inequality. It then suffices to bound the first expectation. We simplify this expectation by *iteratively integrating with respect to a single variable, while keeping the rest being fixed. We first integrate with respect to X_1 given X_2, Y_1, Y_2. This gives us*

$$\begin{aligned}&\mathbb{E}[h(X_1, Y_1)\tilde{\xi}(X_1, Y_2) \mid X_2, Y_1, Y_2] \cdot \tilde{\xi}(X_2, Y_2)\tilde{\xi}(X_2, Y_1)\\ &= (\mathcal{A} \otimes I_Q)h(Y_2, Y_1) \cdot \tilde{\xi}(X_2, Y_2)\tilde{\xi}(X_2, Y_1),\end{aligned}$$

where we have used $\mathbb{E}[h(X_1, Y_1)\tilde{\xi}(X_1, Y_2) \mid X_2, Y_1, Y_2] = \mathbb{E}[h(X_1, Y_1)\xi(X_1, Y_2) \mid Y_1, Y_2] = (\mathcal{A} \otimes I_Q)h(Y_2, Y_1)$. We then integrate with respect to Y_2 given X_2 and Y_1. This yields

$$\mathbb{E}[(\mathcal{A} \otimes I_Q)h(Y_2, Y_1)\tilde{\xi}(X_2, Y_2) \mid X_2, Y_1] \cdot \tilde{\xi}(X_2, Y_1) = (\mathcal{A}^* \otimes I_Q)(\mathcal{A} \otimes I_Q)h(X_2, Y_1) \cdot \tilde{\xi}(X_2, Y_1).$$

By the Cauchy-Schwarz inequality and Lemma 4.14, its expectation is upper bounded by

$$\|(\mathcal{A}^* \otimes I_Q)(\mathcal{A} \otimes I_Q)h\|_{\mathbf{L}^2(P\otimes Q)} \left\|\tilde{\xi}\right\|_{\mathbf{L}^2(P\otimes Q)} \le s_1^2 \|h\|_{\mathbf{L}^2(P\otimes Q)} \left\|\tilde{\xi}\right\|_{P\otimes Q}.$$

Hence, the expectation in (C.21) is upper bounded by $s_1^2 \|h\|^2_{\mathbf{L}^2(P\otimes Q)} \|\tilde{\xi}\|^2_{\mathbf{L}^2(P\otimes Q)}$.

The following lemma generalizes this example to an arbitrary cycle $k_1 \to k_2 \to \cdots \to k_l \to k_1$.

Lemma C.9. *Suppose Assumption 4.2 holds and $f, g \in \mathbf{L}^2_{0,0}(P\otimes Q)$. Define $\varsigma_f := \|f\|_{\mathbf{L}^2(P\otimes Q)}$ and $\varsigma_g := \|g\|_{\mathbf{L}^2(P\otimes Q)}$. For any $l > 0$ and l distinct indices $\{k_1, \ldots, k_l\} \subset [n]$, we have, for all $t, t' \in [l]$,*

$$\mathbb{E}\left[f(X_{k_t}, Y_{k_t})g(X_{k_{t'}}, Y_{k_{t'+1}}) \prod_{i\neq t} \tilde{\xi}(X_{k_i}, Y_{k_i}) \prod_{j\neq t'} \tilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)} \varsigma_f \varsigma_g. \tag{C.22}$$

Proof. There are two cases to consider: $t = t'$ and $t \neq t'$. The proofs are similar so we only prove it for $t = t'$. By exchangeability, it suffices to consider $t = t' = 1$. The strategy is again to iteratively take expectation with respective to one variable, while keeping the rest being fixed. Note that

$$\begin{aligned}&\mathbb{E}[f(X_{k_1}, Y_{k_1})\tilde{\xi}(X_{k_l}, Y_{k_1}) \mid X_{k_1}, X_{k_l}]\\ &= \mathbb{E}[f(X_{k_1}, Y_{k_1})\xi(X_{k_l}, Y_{k_1}) \mid X_{k_1}, X_{k_l}] = (I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l}).\end{aligned}$$

Taking expectation with respect to Y_{k_1} in (C.22), while keeping others being fixed, we get

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[f(X_{k_1}, Y_{k_1})\tilde{\xi}(X_{k_l}, Y_{k_1}) \mid X_{k_1}, X_{k_l}]g(X_{k_1}, Y_{k_2})\prod_{i=2}^{l}\tilde{\xi}(X_{k_i}, Y_{k_i})\prod_{i=2}^{l-1}\tilde{\xi}(X_{k_i}, Y_{k_{i+1}})\right] \\
&= \mathbb{E}\left[(I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l})g(X_{k_1}, Y_{k_2})\prod_{i=2}^{l}\tilde{\xi}(X_{k_i}, Y_{k_i})\prod_{i=2}^{l-1}\tilde{\xi}(X_{k_i}, Y_{k_{i+1}})\right].
\end{aligned}$$

Now taking expectation with respect to X_{k_l}, while keeping others being fixed, we get

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[(I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l})\tilde{\xi}(X_{k_l}, Y_{k_l}) \mid X_{k_1}, Y_{k_l}]g(X_{k_1}, Y_{k_2})\prod_{i=2}^{l-1}\tilde{\xi}(X_{k_i}, Y_{k_i})\tilde{\xi}(X_{k_i}, Y_{k_{i+1}})\right] \\
&= \mathbb{E}\left[(I_P \otimes \mathcal{A}\mathcal{A}^*)f(X_{k_1}, Y_{k_l})g(X_{k_1}, Y_{k_2})\prod_{i=2}^{l-1}\tilde{\xi}(X_{k_i}, Y_{k_i})\tilde{\xi}(X_{k_i}, Y_{k_{i+1}})\right],
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E}[(I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l})\tilde{\xi}(X_{k_l}, Y_{k_l}) \mid X_{k_1}, Y_{k_l}] \\
&= \mathbb{E}[(I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l})\xi(X_{k_l}, Y_{k_l}) \mid X_{k_1}, Y_{k_l}] - \mathbb{E}[(I_P \otimes \mathcal{A}^*)f(X_{k_1}, X_{k_l}) \mid X_{k_1}] \\
&= (I_P \otimes \mathcal{A}\mathcal{A}^*)f(X_{k_1}, Y_{k_l}).
\end{aligned}$$

Keep repeating this argument, we ultimately get

$$\begin{aligned}
&\mathbb{E}\left[f(X_{k_1}, Y_{k_1})g(X_{k_1}, Y_{k_2})\prod_{i=2}^{l}\tilde{\xi}(X_{k_i}, Y_{k_i})\tilde{\xi}(X_{k_i}, Y_{k_{i+1}})\right] \\
&\leq \left\|(I_P \otimes \mathcal{A}\mathcal{A}^*)^{l-1}f\right\|_{\mathbf{L}^2(P\otimes Q)}\|g\|_{\mathbf{L}^2(P\otimes Q)} \leq s_1^{2(l-1)}\varsigma_f\varsigma_g.
\end{aligned}$$

□

Now we are ready to prove Lemma 4.15.

Proof of Lemma 4.15. We first consider the case when $i \neq 1$. It is well-known that every permutation can be decomposed as disjoint cycles. Take a cycle $k_1 \to k_2 \to \cdots \to k_l \to k_1$ of σ. If it contains both 1 and i, then we assume, w.l.o.g., $k_1 = 1$ and $k_2 = i$. Consequently, all the terms that involve $X_{k_{[l]}}$ and $Y_{k_{[l]}}$ are

$$h(X_1, Y_1)h(X_i, Y_{\sigma_i})\prod_{j=2}^{l}\tilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j\in[l]\setminus\{2\}}\tilde{\xi}(X_{k_j}, Y_{k_{j+1}}).$$

Using Lemma C.9 with $f = h$ and $g = h$, it holds that

$$\mathbb{E}\left[h(X_1, Y_1)h(X_i, Y_{\sigma_i})\prod_{j=2}^{l}\tilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j\in[l]\setminus\{2\}}\tilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h^2.$$

If this cycle only contains 1, then a similar argument gives

$$\mathbb{E}\left[h(X_1, Y_1)\prod_{j=2}^{l}\tilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j=1}^{l}\tilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h\varsigma_0.$$

If this cycle only contains i, with $k_1 = i$, then we have

$$\mathbb{E}\left[h(X_i, Y_{\sigma_i})\prod_{j=1}^{l}\tilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j=2}^{l}\tilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h\varsigma_0.$$

Finally, if this cycle does not contain either 1 or i, then it holds

$$\mathbb{E}\left[\prod_{j=1}^{l}\tilde{\xi}(X_{k_j}, Y_{k_j})\tilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_0^2.$$

Here we are invoking Lemma C.9 with $f = g = \xi - 1$. Putting all together, we obtain

$$\mathbb{E}\left[h(X_1, Y_1)\prod_{j=2}^{n}\tilde{\xi}(X_j, Y_j)h(X_i, Y_{\sigma_i})\prod_{j\in[n]\setminus\{i\}}\tilde{\xi}(X_j, Y_{\sigma_j})\right] \le s_1^{2(n-\#\sigma)}\varsigma_h^2\varsigma_0^{2(\#\sigma-1)}.$$

When $i \neq 1$, we can invoke Lemma C.9 to get the same bound, since we allow $t = t'$ in this lemma. □

C.4.3 Limit Law of the Denominator

Finally, we prove Theorem 4.5 regarding the limiting distribution of D_n. According to the singular value decomposition in Assumption 4.2, it holds that

$$\xi(x, y) = 1 + \sum_{k=1}^{\infty} s_k\alpha_k(x)\beta_k(y), \quad \text{in } \mathbf{L}^2(P \otimes Q),$$

where $0 \le s_k < 1$ is decreasing in k. Hence, we start by considering a truncated version of ξ, i.e., $\xi^K(x, y) := 1 + \sum_{k=1}^{K} s_k\alpha_k(x)\beta_k(y)$ for some integer K and derive the limit law of

$$D_n^K := \frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\prod_{i=1}^{n}\xi^K(X_i, Y_{\sigma_i}).$$

Note that all the results in Sections C.4.1 and C.4.2 hold for U_n^K with ξ being replaced by ξ^K.

Proposition C.10. *Under Assumption 4.2, it holds that*

$$D_n^K \to_d D^K := \frac{1}{\sqrt{\prod_{k=1}^K (1-s_k^2)}} \exp\left\{ \frac{1}{2} \sum_{k=1}^K \left[-\frac{s_k^2}{1-s_k^2}(U_k^2+V_k^2) + \frac{2s_k}{1-s_k^2} U_k V_k \right] \right\}, \quad \text{(C.23)}$$

where $\{U_k\}_{k=1}^K$ and $\{V_k\}_{k=1}^K$ are independent standard normal random variables.

Proof. We will prove the convergence using characteristic functions, i.e., $\mathbb{E}[e^{itD_n^K}] \to \mathbb{E}[e^{itD^K}]$.

Step 1. Truncation. Recall from (4.39) that $D_n = 1 + \sum_{r=1}^n D_{n,r}$. Applying it to D_n^K yields $D_n^K = 1 + \sum_{r=1}^n D_{n,r}^K$ where $D_{n,r}^K$ is $D_{n,r}$. We further truncate D_n^K so that it becomes a two-sample U-statistic of fixed order $R > 0$, that is, we consider $D_n^{K,R} := 1 + \sum_{r=1}^R D_{n,r}^K$. We then truncate the limit D^K. By the multi-linear Mehler formula (see, e.g., Foata, 1981), we have

$$D^K = \sum_{p_1,\ldots,p_K \geq 0} \prod_{k=1}^K \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k), \quad \text{(C.24)}$$

where $\{H_p\}_{p\geq 0}$ are the Hermite polynomials satisfying

$$\int H_p(x) H_p(x) e^{-x^2/2} dx = \sqrt{2\pi} p! \mathbb{1}\{p=q\}. \quad \text{(C.25)}$$

Therefore, it is natural to define

$$D^{K,R} := 1 + \sum_{r=1}^R \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^K \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k).$$

By the triangle inequality, $\left|\mathbb{E}[e^{itD_n^K}] - \mathbb{E}[e^{itD^K}]\right| \leq C_1 + C_2 + C_3$ where

$$C_1 := \left|\mathbb{E}[e^{itD_n^K} - e^{itD_n^{K,R}}]\right|, \; C_2 := \left|\mathbb{E}[e^{itD_n^{K,R}} - e^{itD^{K,R}}]\right|, \; C_3 := \left|\mathbb{E}[e^{itD^{K,R}} - e^{itD^K}]\right|.$$

We fix some arbitrary $\delta > 0$ and show that $C_1, C_2, C_3 \leq \delta$ for sufficiently large N and R.

Step 2. Control C_1 and C_3. Using the inequality $|e^{iz} - 1| \leq |z|$, we get

$$C_1 \leq \mathbb{E}\left|e^{itD_n^K} - e^{itD_n^{K,R}}\right| \leq |t| \, \mathbb{E}\left|D_n^K - D_n^{K,R}\right| \leq |t| \sqrt{\mathbb{E}\left|D_n^K - D_n^{K,R}\right|^2}.$$

Invoking Proposition 4.16 for D_n^K implies that, for sufficiently large R, we have $C_1 \leq \delta$. Similarly, it holds that $C_3 \leq |t| \sqrt{\mathbb{E}\left|D^{K,R} - D^K\right|^2}$ where

$$\begin{aligned}\mathbb{E}\left|D^{K,R} - D^K\right|^2 &= \mathbb{E}\left|\sum_{r=R+1}^{\infty} \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^{K} \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k)\right|^2 \\ &= \sum_{r=R+1}^{\infty} \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^{K} s_k^{2p_k} \leq \sum_{r=R+1}^{\infty} s_1^{2r}, \quad \text{since } s_k \leq s_1.\end{aligned}$$

Here the two equations follow from (C.24) and (C.25), respectively. Since $s_1 < 1$, we have $C_3 \leq \delta$ for sufficiently large R.

Step 3. Control C_2. It suffices to show that $D_n^{K,R} \to_d D^{K,R}$ as $n \to \infty$ for any $R > 0$. Note that

$$\begin{aligned}D_{n,r}^K &:= \frac{1}{n!} \sum_{|A|=|B|=r} \sum_{\sigma_A=B} \prod_{i \in A} \tilde{\xi}(X_i, Y_{\sigma_i}) = \frac{(n-r)!}{n!} \sum_{\substack{1 \leq i_1 < \cdots < i_r \leq n \\ 1 \leq j_1 < \cdots < j_r \leq n}} \sum_{\sigma \in \mathcal{S}_r} \prod_{t=1}^{r} \tilde{\xi}(X_{i_t}, Y_{j_{\sigma_t}}) \\ &= \frac{(n-r)!}{r!n!} \sum_{\substack{i_1 \neq \cdots \neq i_r \\ j_1 \neq \cdots \neq j_r}} \prod_{t=1}^{r} \tilde{\xi}(X_{i_t}, Y_{j_t}) = \frac{(n-r)!}{r!n!} \sum_{\substack{i_1 \neq \cdots \neq i_r \\ j_1 \neq \cdots \neq j_r}} \sum_{k_1,\ldots,k_r=1}^{K} \prod_{t=1}^{r} s_{k_t} \alpha_{k_t}(X_{i_t}) \beta_{k_t}(Y_{j_t}) \\ &= \frac{1}{r!} \sum_{k_1,\ldots,k_r=1}^{K} \left(\prod_{t=1}^{r} s_{k_t}\right) \frac{(n-r)!}{n!} \left[\sum_{i_1 \neq \cdots \neq i_r} \prod_{t=1}^{r} \alpha_{k_t}(X_{i_t})\right] \left[\sum_{j_1 \neq \cdots \neq j_r} \prod_{t=1}^{r} \beta_{k_t}(X_{j_t})\right].\end{aligned}$$

The last term above can be rewritten as follows. Take an arbitrary sequence $(k_t)_{t=1}^r \subset [K]^r$. For each $k \in [K]$, let p_k be the number of times k appears among $(k_t)_{t=1}^r$, then, for any permutation symmetric $f : [K]^r \to \mathbb{R}$, we have $\frac{1}{r!} \sum_{k_1,\ldots,k_r=1}^{K} f(k_1, \ldots, k_r) = \sum_{p_1+\cdots+p_K=r} \frac{1}{p_1! \ldots p_K!} f(l_1, \ldots, l_r)$, where $l_1, \ldots, l_r$ is an arbitrary sequence such that k appears exactly p_k times for all $k \in [K]$. Moreover, it follows from (van der Vaart, 2000, Theorem 12.10) that

$$\begin{aligned}\sqrt{\frac{(n-r)!}{n!}} \sum_{i_1 \neq \cdots \neq i_r} \prod_{t=1}^{r} \alpha_{k_t}(X_{i_t}) &= \prod_{k=1}^{K} H_{p_k}(\mathbb{G}_n^{(X)} \alpha_k) + o_p(1) \\ \sqrt{\frac{(n-r)!}{n!}} \sum_{j_1 \neq \cdots \neq j_r} \prod_{t=1}^{r} \beta_{k_t}(Y_{j_t}) &= \prod_{k=1}^{K} H_{p_k}(\mathbb{G}_n^{(Y)} \beta_k) + o_p(1),\end{aligned}$$

where $\mathbb{G}_n^{(X)}\alpha := \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha(X_i)$ and $\mathbb{G}_n^{(Y)}\beta$ similarly. As a result,

$$D_{n,r}^K = \sum_{p_1+\cdots+p_K=r}\prod_{k=1}^K \frac{s_k^{p_k}}{p_k!}H_{p_k}(\mathbb{G}_n^{(X)}\alpha_k)H_{p_k}(\mathbb{G}_n^{(Y)}\beta_k) + o_p(1),$$

and thus $D_n^{K,R} = 1 + \sum_{r=1}^R \sum_{p_1+\cdots+p_K=r}\prod_{k=1}^K \frac{s_k^{p_k}}{p_k!}H_{p_k}(\mathbb{G}_n^{(X)}\alpha_k)H_{p_k}(\mathbb{G}_n^{(Y)}\beta_k) + o_p(1)$. According to the multivariate CLT (Billingsley, 1995, Section 29), the random vector $(\mathbb{G}_n^{(X)}\alpha_k, \mathbb{G}_n^{(Y)}\beta_k)_{k=1}^K$ converges in distribution to $\backslash_{2K}(0, I_{2K})$ by the orthonormality of $\{\alpha_k\}_{k=1}^K$ and $\{\beta_k\}_{k=1}^K$. It then follows from the continuous mapping theorem that

$$D_n^{K,R} \to_d 1 + \sum_{r=1}^R \sum_{p_1+\cdots+p_K=r}\prod_{k=1}^K \frac{s_k^{p_k}}{p_k!}H_{p_k}(U_k)H_{p_k}(V_k) = D^{K,R},$$

which completes the proof. □

Proof of Theorem 4.5. We again prove the convergence using the characteristic functions. *Step 0. Verify the validity of the limit.* We first show $1/\prod_{k=1}^\infty(1-s_k^2) < \infty$. In fact,

$$\frac{1}{\prod_{k=1}^\infty(1-s_k^2)} = \exp\left\{\sum_{k=1}^\infty \log\frac{1}{1-s_k^2}\right\} \le \exp\left\{\sum_{k=1}^\infty \frac{s_k^2}{1-s_k^2}\right\} \le \exp\left\{\frac{\sum_{k=1}^\infty s_k^2}{1-s_1^2}\right\} < \infty, \tag{C.26}$$

where the first inequality follows from $\log(1+x) \ge \frac{x}{1+x}$ for all $x > -1$ and the last inequality follows from the square summability of $\{s_k\}_{k\ge1}$. It suffices to show that $D \in \mathbf{L}^2(P\otimes Q)$. For any $k \ge 1$, let

$$Z_k := \frac{1}{\sqrt{1-s_k^2}}\exp\left\{-\frac{s_k^2}{2(1-s_k^2)}(U_k^2+V_k^2) + \frac{s_k}{1-s_k^2}U_kV_k\right\}. \tag{C.27}$$

Then $\{Z_k\}_{k\ge1}$ are mutually independent and $D = \prod_{k=1}^\infty Z_k$. By a standard computation, we get $\mathbb{E}[Z_k^2] = 1/(1-s_k^2)$. Therefore, by (C.26), $\mathbb{E}[D^2] = \prod_{k=1}^\infty \mathbb{E}[Z_k^2] = 1/\prod_{k=1}^\infty(1-s_k^2) < \infty$.

Step 1. Control the difference between the characteristic functions. Recall D_n^K and D^K be from Proposition C.10. By the triangle inequality, we have $\left|\mathbb{E}[e^{itD_n}] - \mathbb{E}[e^{itD}]\right| \le C_1+C_2+C_3$ where

$$C_1 := \left|\mathbb{E}[e^{itD_n}] - \mathbb{E}[e^{itD_n^K}]\right|,\ C_2 := \left|\mathbb{E}[e^{itD_n^K}] - \mathbb{E}[e^{itD^K}]\right|,\ C_3 := \left|\mathbb{E}[e^{itD^K}] - \mathbb{E}[e^{itD}]\right|.$$

Fix $\delta > 0$. By Proposition C.10, $C_2 \leq \delta$ for sufficiently large n. It then remains to control C_1 and C_3.

Step 2. Control C_1. By construction, it holds that

$$D_n - D_n^K = \sum_{r=1}^{n} \frac{1}{n!} \sum_{|A|=|B|=r} \sum_{\sigma_A = B} \prod_{i \in A} \xi^{-K}(X_i, Y_{\sigma_i}),$$

where $\xi^{-K} := \xi - \xi^K \in \mathbf{L}^2_{0,0}(P \otimes Q)$ and $\varsigma_K^2 := (P \otimes Q)[(\xi^{-K})^2] = \sum_{k \geq K+1} s_k^2$. Invoking Proposition 4.16 for ξ^{-K}, we obtain $\mathbb{E}[(D_n - D_n^K)^2] \leq \sum_{r=1}^n \frac{1}{r!} \sum_{\sigma \in \mathcal{S}_r} s_1^{2(r-\#\sigma)} \varsigma_K^{2\#\sigma}$. For sufficiently large K, since ς_K^2 can be arbitrarily small, we have $C_1 \leq |t| \, \mathbb{E}[(D_n - D_n^K)^2] \leq \delta$.

Step 3. Control C_3. Again, it suffices to control $\mathbb{E}[(D^K - D)^2]$. Recall Z_k in (C.27). By independence,

$$\begin{aligned} \mathbb{E}[(D^K - D)^2] &= \mathbb{E}\left[\left(\prod_{k=1}^K Z_k - \prod_{k=1}^\infty Z_k\right)^2\right] = \mathbb{E}\left[\prod_{k=1}^K Z_k^2\right] \mathbb{E}\left[\left(1 - \prod_{k \geq K+1} Z_k\right)^2\right] \\ &= \frac{1}{\prod_{k=1}^K (1 - s_k^2)} \left[\frac{1}{\prod_{k \geq K+1}(1 - s_k^2)} - 1\right], \quad \text{since } \mathbb{E}[Z_k] = 1. \end{aligned}$$

It follows from (C.26) that $\prod_{k=1}^K (1 - s_k^2)^{-1} < \infty$ and

$$1 \leq \frac{1}{\prod_{k \geq K+1}(1 - s_k^2)} \leq \exp\left\{\frac{1}{1 - s_1^2} \sum_{k \geq K+1} s_k^2\right\} \to 1, \quad \text{as } K \to \infty.$$

Hence, we have $\mathbb{E}[(D^K - D)^2] \to 0$ as $K \to \infty$, which completes the proof. □

C.5 Second Order Chaos

We derive in this section the second order chaos of T_n. Before that, we define the operator $\mathcal{C}$ which appears in the second order chaos.

Definition C.2. *Define the operator* $\mathcal{C}$ *on* $\mathbf{L}^2(P \otimes Q)$ *by*

$$\mathcal{C} := (I - \mathcal{A}^* \mathcal{A}) \otimes (I - \mathcal{A}\mathcal{A}^*).$$

We will prove later that the operator $\mathcal{C}$ is well-defined.

Assumption C.1. *We make the additional assumptions that $\xi \in \mathbf{L}^{2\mathfrak{p}}(P \otimes Q)$ and $\mathcal{C}^{-1}(\tilde{\eta}\xi) \in \mathbf{L}^{2\mathfrak{p}/(\mathfrak{p}-2)}(P \otimes Q)$ for some*[1] $\mathfrak{p} \in [2, \infty]$.

Let $\kappa_{2,0} := -(I_P \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi)$, $\kappa_{0,2} := -(\mathcal{A} \otimes I_Q)\mathcal{C}^{-1}(\tilde{\eta}\xi)$, and $\kappa_{1,1'} := (I + \mathcal{B})\mathcal{C}^{-1}(\tilde{\eta}\xi)$.

Theorem C.11. *Assume, for some $\eta \in \mathbf{L}^2(\mu)$, $\varsigma^2 = 0$ in Theorem 4.6. Let $\theta_{1,1'} := \iint \kappa_{1,1'}(x, y)\mathrm{d}\mu(x, y)$. Then*

$$T_n - \theta + \frac{\theta_{1,1'}}{n} = \frac{1}{n(n-1)}\left[\sum_{i \neq j}(\kappa_{2,0}(X_i, X_j) + \kappa_{0,2}(Y_i, Y_j)) + \sum_{i,j=1}^{n}\kappa_{1,1'}(X_i, Y_j)\right] + o_p(n^{-1}).$$

Furthermore, suppose that the function $(\eta - \theta)\xi$ has a spectral expansion in $\mathbf{L}^2(P \otimes Q)$ with respect to the orthonormal basis $\{\alpha_k \otimes \beta_l\}_{k,l\geq 0}$ of $\mathbf{L}^2(P \otimes Q)$ with coefficients $(\gamma_{kl}, k, l \geq 0)$, i.e., $(\eta - \theta)\xi = \sum_{k,l\geq 0}\gamma_{kl}(\alpha_k \otimes \beta_l)$. Then, as $n \to \infty$, the sequence of random variables $n(T_n - \theta) + \theta_{1,1'}$ converges in law to the mean-zero random variable

$$\sum_{k,l\geq 1}\frac{\gamma_{kl}}{(1 - s_k^2)(1 - s_l^2)}\big\{U_k V_l + s_k s_l U_l V_k - s_l(U_k U_l - \mathbb{1}\{k = l\}) - s_k(V_k V_l - \mathbb{1}\{k = l\})\big\},$$

where $\{U_k\}_{k\geq 1}$ and $\{V_l\}_{l\geq 1}$ are independent i.i.d. standard normal random variables.

C.5.1 Second order chaos

We first prove that the operator $\mathcal{C}$ is well-defined. Given a measure ν on $\mathbb{R}^d \times \mathbb{R}^d$, let

$$\mathbf{L}^2_{0,0}(\nu) := \{f \in \mathbf{L}^2(\nu) : \mathbb{E}[f(X, Y) \mid Y] \overset{\text{a.s.}}{=} \mathbb{E}[f(X, Y) \mid X] \overset{\text{a.s.}}{=} 0 \text{ for all } (X, Y) \sim \nu\}. \quad \text{(C.28)}$$

For $f \in \mathbf{L}^2_{0,0}(\nu)$, we say f is degenerate with respect to ν. For example, we will show in the next lemma that the function $\tilde{\eta}$ defined in (4.32),

$$\tilde{\eta}(x, y) := \eta(x, y) - \theta - (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x) - (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y),$$

belongs to $\mathbf{L}^2_{0,0}(\mu)$, and then, by Assumption 4.3, $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$.

[1] When $\mathfrak{p} = 2$, we assume $\xi \in \mathbf{L}^4(P \otimes Q)$ and $\mathcal{C}^{-1}(\eta\xi) \in \mathbf{L}^\infty(P \otimes Q)$; when $\mathfrak{p} = \infty$, we only assume $\xi \in \mathbf{L}^\infty(P \otimes Q)$, i.e., ξ is bounded.

Lemma C.12. *Under Assumption 4.3, the inverse operator* $\mathcal{C}^{-1} : \mathbf{L}^2_{0,0}(P \otimes Q) \to \mathbf{L}^2_{0,0}(P \otimes Q)$ *is well-defined. Moreover, it is equal to* $(I - \mathcal{A}^*\mathcal{A})^{-1} \otimes (I - \mathcal{A}\mathcal{A}^*)^{-1}$. *In particular,* $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$ *so that* $\mathcal{C}^{-1}(\tilde{\eta}\xi)$ *is well-defined.*

Proof of Lemma C.12. We will prove that $\mathcal{C} : \mathbf{L}^2_{0,0}(P \otimes Q) \to \mathbf{L}^2_{0,0}(P \otimes Q)$ is bijective. On the one hand, take any $f \in \mathbf{L}^2_{0,0}(P \otimes Q)$, since $\{\alpha_i \otimes \beta_j\}_{i,j \geq 0}$ forms an orthonormal basis of $\mathbf{L}^2(P \otimes Q)$, we know f must admit the following expansion:

$$f = \sum_{i,j \geq 1} \gamma_{ij} \alpha_i \otimes \beta_j, \quad \text{where } \sum_{i,j \geq 1} \gamma_{ij}^2 < \infty.$$

Recall from Assumption 4.2 that $s_k < 1$ for all $k \geq 1$. Define

$$\tilde{f} := \sum_{i,j \geq 1} \frac{\gamma_{ij}}{(1 - s_i^2)(1 - s_j^2)} \alpha_i \otimes \beta_j,$$

then, similar to (C.17), we have $\mathcal{C}\tilde{f} = f$ and $\tilde{f} \in \mathbf{L}^2_{0,0}(P \otimes Q)$. Hence, $\mathcal{C}$ is surjective. On the other hand, if $\mathcal{C}f = 0$, then $\mathcal{C}f = \sum_{i,j \geq 1}(1 - s_i^2)(1 - s_j^2)\gamma_{ij}(\alpha_i \otimes \beta_j) = 0$. It follows that $\gamma_{ij} = 0$ for all $i, j \geq 1$, and thus $\mathcal{C}$ is injective.

By (C.14) we get

$$\mathbb{E}_\mu\big[(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(X_1) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(Y_1) \mid X_1\big] = \kappa_{1,0}(X_1).$$

By definition, $\kappa_{1,0}(X_1) = \int [\eta(X_1, y) - \theta]\xi(X_1, y)\mathrm{d}Q(y) = \mathbb{E}_\mu[\eta(X_1, Y_1) - \theta \mid X_1]$. This yields $\mathbb{E}_\mu[\tilde{\eta}(X_1, Y_1) \mid X_1] = 0$. Similarly, $\mathbb{E}_\mu[\tilde{\eta}(X_1, Y_1) \mid Y_1] = 0$. We obtain $\tilde{\eta} \in \mathbf{L}^2_{0,0}(\mu)$, and then, by Assumption 4.3, $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$ since

$$\begin{aligned} 0 &= \mathbb{E}_\mu[\tilde{\eta}(X_1, Y_1) \mid X_1](x) = \int \tilde{\eta}(x, y)\xi(x, y)\mathrm{d}Q(y) \\ 0 &= \mathbb{E}_\mu[\tilde{\eta}(X_1, Y_1) \mid Y_1](y) = \int \tilde{\eta}(x, y)\xi(x, y)\mathrm{d}P(x). \end{aligned}$$

□

From Lemma C.12 we know $\mathcal{C}$ preserves the degeneracy with respect to $P \otimes Q$. The following lemma verifies similar properties for other operators under consideration.

Lemma C.13. *Let $\mathcal{A}_k \in \{\mathcal{A}, \mathcal{A}^*, I_P, I_Q\}$ be an operator mapping from $\mathbf{L}^2(\nu_k)$ to $\mathbf{L}^2(\nu_k')$ for $k \in \{1, 2\}$. Then $\mathcal{A}_1 \otimes \mathcal{A}_2$ maps $\mathbf{L}^2_{0,0}(\nu_1 \otimes \nu_2)$ to $\mathbf{L}^2_{0,0}(\nu_1' \otimes \nu_2')$. In particular, the operator $\mathcal{B}$ maps $\mathbf{L}^2_{0,0}(P \otimes Q)$ to $\mathbf{L}^2_{0,0}(P \otimes Q)$.*

Proof. We prove the claim for $\mathcal{A}_1 = \mathcal{A} : \mathbf{L}^2(P) \to \mathbf{L}^2(Q)$ and $\mathcal{A}_2 = \mathcal{A}^* : \mathbf{L}^2(Q) \to \mathbf{L}^2(P)$. The rest follows similarly. Take any $f \in \mathbf{L}^2_{0,0}(P \otimes Q)$, we know $(\mathcal{A} \otimes \mathcal{A}^*)f(Y_1, X_2) = \mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1]$. Hence, by the tower property, it holds that

$$\mathbb{E}_\mu[(\mathcal{A} \otimes \mathcal{A}^*)f(Y_1, X_2) \mid X_2] = \mathbb{E}_\mu[f(X_1, Y_2) \mid X_2] = \mathbb{E}_\mu\big[\mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_2] \mid X_2\big] = 0.$$

Analogously, $\mathbb{E}_\mu[(\mathcal{A} \otimes \mathcal{A}^*)f(Y_1, X_2) \mid Y_1] = 0$. This implies $(\mathcal{A} \otimes \mathcal{A}^*)f(Y_1, X_2) \in \mathbf{L}^2_{0,0}(Q \otimes P)$, and the claim follows. Now, observe that $(\mathcal{A} \otimes \mathcal{A}^*)f(Y_1, X_2) \in \mathbf{L}^2_{0,0}(Q \otimes P)$ yields $\mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)f(X_2, Y_1) \in \mathbf{L}^2_{0,0}(P \otimes Q)$ and $\mathcal{B} = \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)$, we get $\mathcal{B}$ maps $\mathbf{L}^2_{0,0}(P \otimes Q)$ to $\mathbf{L}^2_{0,0}(P \otimes Q)$. □

Unlike the first order chaos, we will give an approximation to the second order chaos, i.e., the projection onto H_2, of T_n. According to Lemma C.12, we know $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$ and $\mathcal{C}^{-1}(\tilde{\eta}\xi)$ is well-defined. We define

$$\mathcal{Q}_n := \frac{1}{n(n-1)} \left\{ \sum_{i \neq j} [\kappa_{2,0}(X_i, X_j) + \kappa_{0,2}(Y_i, Y_j)] + \sum_{i,j=1}^{n} \kappa_{1,1'}(X_i, Y_j) - \sum_{i=1}^{n} \ell_{1,1'}(X_i, Y_i) \right\}, \tag{C.29}$$

where $\ell_{1,1'}(X_1, Y_1)$ is an affine function such that $\kappa_{1,1'} - \ell_{1,1'} \in \mathbf{L}^2_{0,0}(\mu)$. We will show in the next lemma that $\kappa_{1,1'} \in \mathbf{L}^2(\mu)$, so $\ell_{1,1'}$ can be derived the same way we obtain $\tilde{\eta}$. Note that $\mathcal{Q}_n$ is permutation symmetric due to the affineness of $\ell_{1,1'}$.

Lemma C.14. *The functions $\kappa_{2,0}$, $\kappa_{0,2}$ and $\kappa_{1,1'}$ are degenerate, i.e., $\kappa_{2,0} \in \mathbf{L}^2_{0,0}(P \otimes P)$, $\kappa_{0,2} \in \mathbf{L}^2_{0,0}(Q \otimes Q)$ and $\kappa_{1,1'} \in \mathbf{L}^2_{0,0}(P \otimes Q)$. Under Assumption C.1, the function $\kappa_{1,1'}$ also belongs to $\mathbf{L}^2(\mu)$, and thus $\mathcal{Q}_n \in H_2$. Moreover, the following identities hold:*

$$(I + \mathcal{T})[\kappa_{2,0} + (\mathcal{A}^* \otimes \mathcal{A}^*)\kappa_{0,2} + (I_P \otimes \mathcal{A}^*)\kappa_{1,1'}] \equiv 0$$

$$(I + \mathcal{T})[(\mathcal{A} \otimes \mathcal{A})\kappa_{2,0} + \kappa_{0,2} + (\mathcal{A} \otimes I_Q)\kappa_{1,1'}] \equiv 0$$

$$(I_P \otimes \mathcal{A})(I + \mathcal{T})\kappa_{2,0} + (\mathcal{A}^* \otimes I_Q)(I + \mathcal{T})\kappa_{0,2} + (I + \mathcal{B})\kappa_{1,1'} \equiv \tilde{\eta}\xi.$$

Proof. Since $\tilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$, we know from Lemma C.12 and Lemma C.13 that $\kappa_{2,0} \in \mathbf{L}^2_{0,0}(P \otimes P)$, $\kappa_{0,2} \in \mathbf{L}^2_{0,0}(Q \otimes Q)$ and $\kappa_{1,1'} \in \mathbf{L}^2_{0,0}(P \otimes Q)$. Let $f := \mathcal{C}^{-1}(\tilde{\eta}\xi)$. Recall from Assumption C.1 that $\xi \in \mathbf{L}^{2\mathfrak{p}}(P \otimes Q)$ and $f \in \mathbf{L}^{2\mathfrak{q}}(P \otimes Q)$. As a result,

$$\mu\left[f^2\right] \overset{\text{Hölder}}{\leq} \left[\int f^{2\mathfrak{q}}(x,y)\mathrm{d}P(x)\mathrm{d}Q(y)\right]^{\frac{1}{\mathfrak{q}}} \left[\int \xi^{\mathfrak{p}}(x,y)\mathrm{d}P(x)\mathrm{d}Q(y)\right]^{\frac{1}{\mathfrak{p}}} < \infty. \tag{C.30}$$

Furthermore,

$$\begin{aligned}
\int (\mathcal{B}f)^{2\mathfrak{q}}(x,y)\mathrm{d}P(x)\mathrm{d}Q(y) &= \int \left[\int f(x',y')\xi(x',y)\xi(x,y')P(x')Q(y')dx'dy'\right]^{2\mathfrak{q}} \mathrm{d}P(x)\mathrm{d}Q(y) \\
&\overset{\text{Jensen}}{\leq} \iint f^{2\mathfrak{q}}(x',y')\xi(x',y)\xi(x,y')\mathrm{d}P(x')\mathrm{d}Q(y')\mathrm{d}P(x)\mathrm{d}Q(y) \\
&\overset{\text{(i)}}{=} \int f^{2\mathfrak{q}}(x',y')\mathrm{d}P(x')\mathrm{d}Q(y') < \infty,
\end{aligned}$$

where (i) follows from $\int \xi(x',y)\mathrm{d}Q(y) \overset{\text{a.s.}}{=} \int \xi(x,y')\mathrm{d}P(x) \overset{\text{a.s.}}{=} 1$. Similar to (C.30), it then holds that

$$\mu[(\mathcal{B}f)^2] \leq \left[\int (\mathcal{B}f)^{2\mathfrak{q}}(x,y)\mathrm{d}P(x)\mathrm{d}Q(y)\right]^{\frac{1}{\mathfrak{q}}} \left[\int \xi^{\mathfrak{p}}(x,y)\mathrm{d}P(x)\mathrm{d}Q(y)\right]^{\frac{1}{\mathfrak{p}}} < \infty.$$

This yields that $\kappa_{1,1'} := (I + \mathcal{B})f \in \mathbf{L}^2(\mu)$. Now, by the degeneracy (C.28) of $\kappa_{2,0}$, $\kappa_{0,2}$ and $\kappa_{1,1'}$, we obtain $\mathcal{Q}_n \in H_0^{\perp} \cap H_1^{\perp}$. It then follows from the permutation symmetry of $\mathcal{Q}_n$ that $\mathcal{Q}_n \in H_2$.

Notice that $(\mathcal{A}^* \otimes \mathcal{A}^*)\kappa_{0,2} = -(\mathcal{A}^*\mathcal{A} \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi)$ and

$$\begin{aligned}
(I_P \otimes \mathcal{A}^*)\kappa_{1,1'} &= ((I_P \otimes \mathcal{A}^*) + (I_P \otimes \mathcal{A}^*)\mathcal{B})\mathcal{C}^{-1}(\tilde{\eta}\xi) \overset{\text{(i)}}{=} -\kappa_{2,0} + \mathcal{T}(\mathcal{A}^* \otimes I_P)(\mathcal{A} \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi) \\
&= -\kappa_{2,0} + \mathcal{T}(\mathcal{A}^*\mathcal{A} \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi),
\end{aligned}$$

where we have used $\mathcal{B} = \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)$ in (i). It then follows that

$$(I + \mathcal{T})[\kappa_{2,0} + (\mathcal{A}^* \otimes \mathcal{A}^*)\kappa_{0,2} + (I_P \otimes \mathcal{A}^*)\kappa_{1,1'}] = (I + \mathcal{T})(\mathcal{T} - I)(\mathcal{A}^*\mathcal{A} \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi) \equiv 0,$$

since $(I + \mathcal{T})(\mathcal{T} - I) = \mathcal{T} - I + \mathcal{T}\mathcal{T} - \mathcal{T} = 0$. Similarly, $(I + \mathcal{T})[(\mathcal{A} \otimes \mathcal{A})\kappa_{2,0} + \kappa_{0,2} + (\mathcal{A} \otimes I_Q)\kappa_{1,1'}] \equiv 0$.

Let us verify the last identity in the statement of Lemma C.14. Note that

$$(I_P \otimes \mathcal{A})(I+\mathcal{T})\kappa_{2,0} = [(I_P \otimes \mathcal{A}) + \mathcal{T}(\mathcal{A} \otimes I_P)]\kappa_{2,0} = -[(I_P \otimes \mathcal{A}\mathcal{A}^*) + \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)]\mathcal{C}^{-1}(\tilde{\eta}\xi)$$
$$= -[(I_P \otimes \mathcal{A}\mathcal{A}^*) + \mathcal{B}]\mathcal{C}^{-1}(\tilde{\eta}\xi).$$

Analogously, $(\mathcal{A}^* \otimes I_Q)(I+\mathcal{T})\kappa_{0,2} = -[(\mathcal{A}^*\mathcal{A} \otimes I_Q) + \mathcal{B}]\mathcal{C}^{-1}(\tilde{\eta}\xi)$ and

$$(I+\mathcal{B})\kappa_{1,1'} = (I+\mathcal{B})(I+\mathcal{B})\mathcal{C}^{-1}(\tilde{\eta}\xi) = [I + 2\mathcal{B} + (\mathcal{A}^* \otimes \mathcal{A})\mathcal{T}\mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*)]\mathcal{C}^{-1}(\tilde{\eta}\xi).$$

Hence,

$$(I_P \otimes \mathcal{A})(I+\mathcal{T})\kappa_{2,0} + (\mathcal{A}^* \otimes I_Q)(I+\mathcal{T})\kappa_{0,2} + (I+\mathcal{B})\kappa_{1,1'}$$
$$= [I - (I_P \otimes \mathcal{A}\mathcal{A}^*) - (\mathcal{A}^*\mathcal{A} \otimes I_Q) + (\mathcal{A}^*\mathcal{A} \otimes \mathcal{A}\mathcal{A}^*)]\mathcal{C}^{-1}(\tilde{\eta}\xi) = \tilde{\eta}\xi,$$

where the last equality follows from $\mathcal{C} := (I - \mathcal{A}^*\mathcal{A}) \otimes (I - \mathcal{A}\mathcal{A}^*) = I - I_P \otimes \mathcal{A}\mathcal{A}^* - \mathcal{A}^*\mathcal{A} \otimes I_Q + \mathcal{A}^*\mathcal{A} \otimes \mathcal{A}\mathcal{A}^*$. □

The next proposition shows that $\mathcal{Q}_n$ is equal to the second order chaos of T_n up to an $o_p(n^{-1})$ term.

Proposition C.15. *Suppose Assumption C.1 holds and* $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$. *Let the second order chaos of* T_n *be* $\operatorname{Proj}_{H_2}(T_n)$. *Then we have* $\operatorname{Proj}_{H_2}(T_n) = \mathcal{Q}_n + o_p(n^{-1})$.

Proof. Define

$$\tilde{\mathcal{Q}}_n := \frac{1}{n(n-1)} \sum_{i \neq j} [\kappa_{2,0}(X_i, X_j) + \kappa_{0,2}(Y_i, Y_j) + \kappa_{1,1'}(X_i, Y_j)]. \tag{C.31}$$

It follows from LLN that $\mathcal{Q}_n - \tilde{\mathcal{Q}}_n = o_p(n^{-1})$. It then suffices to show $\tilde{\mathcal{Q}}_n - \operatorname{Proj}_{H_2}(T_n) = o_p(n^{-1})$. According to the degeneracy in Lemma C.14, we know $\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n] = 0$ and $\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid X_i] = \mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid Y_i] = 0$ for all $i \in [n]$, which implies $\tilde{\mathcal{Q}}_n \in H_0^\perp \cap H_1^\perp$. Note that $\tilde{\mathcal{Q}}_n$ is not permutation symmetric since it lacks the diagonal terms $\kappa_{1,1'}(X_i, Y_i)$, so it is not in H_2. Moreover, we have

$$\mathbb{E}_\mu[\mathcal{Q}_n \mid X_i, Y_i] = 0, \quad \text{for all } i \in [n]. \tag{C.32}$$

Step 1. We show $\mathrm{Proj}_{H_2}(T_n) = \mathrm{Proj}_{H_2}(\tilde{T}_n)$, where

$$\tilde{T}_n := \frac{1}{n}\sum_{i=1}^{n}[\eta(X_i, Y_i) - \theta] - \mathcal{L}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{\eta}(X_i, Y_i). \tag{C.33}$$

In fact, since $\theta \perp H_2$ and $\mathcal{L}_n \perp H_2$, we have, for any $U \in H_2$,

$$\mathbb{E}_\mu[(\tilde{T}_n - T_n)U] = \mathbb{E}_\mu\left[\left(\frac{1}{n}\sum_{i=1}^{n}\eta(X_i, Y_i) - T_n\right)U\right].$$

By exchangeability of $\{(X_i, Y_i)\}_{i\in[n]}$, it holds that $\mathbb{E}_\mu\left[\frac{1}{n}\sum_{i=1}^{n}\eta(X_i, Y_i)U\right] = \mathbb{E}_\mu[\eta(X_1, Y_1)U]$, and thus

$$\begin{aligned}\mathbb{E}_\mu[(\tilde{T}_n - T_n)U] &= \mathbb{E}_\mu[\eta(X_1, Y_1)U] - \mathbb{E}_\mu[T_n U] \\ &\overset{\text{(i)}}{=} \mathbb{E}_\mu[\eta(X_1, Y_1)U] - \mathbb{E}_\mu[\mathbb{E}_\mu[\eta(X_1, Y_1)U \mid \mathcal{G}_n]] = 0,\end{aligned}$$

where (i) follows from the tower property. Hence, $\tilde{T}_n - T_n \in H_2^\perp$ and thus the claim follows. Moreover, since $\tilde{\eta} \in \mathbf{L}^2_{0,0}(\mu)$, we have $\tilde{T}_n \in H_0^\perp \cap H_1^\perp$,

$$\mathbb{E}_\mu[\tilde{T}_n \mid X_i, Y_i] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_\mu[\tilde{\eta}(X_k, Y_k) \mid X_i, Y_i] = \frac{1}{n}\tilde{\eta}(X_i, Y_i), \quad \text{for all } i \in [n], \tag{C.34}$$

and

$$\mathbb{E}_\mu[\tilde{T}_n \mid X_i, X_j] = \mathbb{E}_\mu[\tilde{T}_n \mid Y_i, Y_j] = \mathbb{E}_\mu[\tilde{T}_n \mid X_i, Y_j] = 0, \quad \text{for all } i \neq j. \tag{C.35}$$

Step 2. We show $\mathrm{Proj}_{H_2}(T_n) = \mathrm{Proj}_{H_2}(\tilde{\mathcal{Q}}_n)$. By Step 1, it suffices to prove $\tilde{T}_n - \tilde{\mathcal{Q}}_n \in H_2^\perp$. We will prove $\mathbb{E}_\mu[(\tilde{T}_n - \tilde{\mathcal{Q}}_n)U] = 0$ for every

$$U := \sum_{i<j}[f_{2,0}(X_i, X_j) + f_{0,2}(Y_i, Y_j)] + \sum_{i,j=1}^{n} f_{1,1}(X_i, Y_j) \in \mathbf{L}^2(\mu^n).$$

We first compute $\mathbb{E}_\mu[\tilde{T}_n - \tilde{\mathcal{Q}}_n \mid X_1, X_2]$. Since $\kappa_{2,0} \in \mathbf{L}^2_{0,0}(P \otimes P)$, so it holds

$$\begin{aligned}\mathbb{E}_\mu\left[\sum_{i\neq j}\kappa_{2,0}(X_i, X_j) \;\middle|\; X_1, X_2\right] &= \mathbb{E}_\mu\left[\sum_{\{i,j\}=\{1,2\}}\kappa_{2,0}(X_i, X_j) \;\middle|\; X_1, X_2\right] && \text{(C.36)} \\ &= (I + \mathcal{T})\kappa_{2,0}(X_1, X_2). && \text{(C.37)}\end{aligned}$$

Since $\kappa_{0,2} \in \mathbf{L}^2_{0,0}(Q \otimes Q)$ and $\mathbb{E}_\mu[f(Y_1, Y_2) \mid X_1, X_2] = (\mathcal{A}^* \otimes \mathcal{A}^*) f(X_1, X_2)$ for any $f \in \mathbf{L}^2(Q \otimes Q)$, we get

$$\mathbb{E}_\mu\left[\sum_{i \neq j} \kappa_{0,2}(Y_i, Y_j) \;\middle|\; X_1, X_2\right] = (I + \mathcal{T})(\mathcal{A}^* \otimes \mathcal{A}^*)\kappa_{0,2}(X_1, X_2). \tag{C.38}$$

Furthermore, since $\mathbb{E}_\mu[f(X_1, Y_2) \mid X_1, X_2] = (I_P \otimes \mathcal{A}^*) f(X_1, X_2)$, we have

$$\mathbb{E}_\mu\left[\sum_{i \neq j} \kappa_{1,1'}(X_i, Y_j) \;\middle|\; X_1, X_2\right] = (I + \mathcal{T})(I_P \otimes \mathcal{A}^*)\kappa_{1,1'}(X_1, X_2). \tag{C.39}$$

Putting (C.36), (C.38) and (C.39) together, we get $\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid X_1, X_2] = 0$ by the first identity in Lemma C.14. Consequently, by (C.35),

$$\mathbb{E}_\mu[\tilde{T}_n - \tilde{\mathcal{Q}}_n \mid X_1, X_2] = \mathbb{E}_\mu[\tilde{T}_n \mid X_1, X_2] = 0.$$

By the exchangeability of $\{(X_i, Y_i)\}_{i=1}^n$, we obtain $\mathbb{E}_\mu[\tilde{T}_n - \tilde{\mathcal{Q}}_n \mid X_i, X_j] = 0$ for all $i \neq j$. Similarly, $\mathbb{E}_\mu[\tilde{T}_n - \tilde{\mathcal{Q}}_n \mid Y_i, Y_j] = 0$ for all $i \neq j$. Hence, we only need to prove

$$\mathbb{E}_\mu\left[(\tilde{T}_n - \tilde{\mathcal{Q}}_n) \sum_{i,j} f_{1,1}(X_i, Y_j)\right] = 0.$$

For that purpose, we will compute $\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid X_i, Y_j]$. We have shown in (C.32) that $\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid X_i, Y_i] = 0$ for all $i \in [n]$. For $(i, j) = (1, 2)$, it holds that

$$\begin{aligned}
\mathbb{E}_\mu\left[\sum_{i \neq j} \kappa_{2,0}(X_i, X_j) \;\middle|\; X_1, Y_2\right] &= (I_P \otimes \mathcal{A})(I + \mathcal{T})\kappa_{2,0}(X_1, Y_2) \\
\mathbb{E}_\mu\left[\sum_{i \neq j} \kappa_{0,2}(Y_i, Y_j) \;\middle|\; X_1, Y_2\right] &= (\mathcal{A}^* \otimes I_Q)(I + \mathcal{T})\kappa_{0,2}(X_1, Y_2) \\
\mathbb{E}_\mu\left[\sum_{i \neq j} \kappa_{1,1'}(X_i, Y_j) \;\middle|\; X_1, Y_2\right] &= (I + \mathcal{B})\kappa_{1,1'}(X_1, X_2).
\end{aligned}$$

It then follows from the third identity in Lemma C.14 that

$$\mathbb{E}_\mu[\tilde{\mathcal{Q}}_n \mid X_1, Y_2] = \frac{1}{n(n-1)} \tilde{\eta}(X_1, Y_2)\xi(X_1, Y_2).$$

By the exchangeability of $\{(X_i, Y_i)\}_{i=1}^n$ again, we get

$$\mathbb{E}_\mu\left[\tilde{\mathcal{Q}}_n \sum_{i,j=1}^n f_{1,1}(X_i, Y_j)\right] = \sum_{i \neq j} \mathbb{E}_\mu[\tilde{\mathcal{Q}}_n f_{1,1}(X_i, Y_j)] = \mathbb{E}_\mu[\tilde{\eta}(X_1, Y_2)\xi(X_1, Y_2) f_{1,1}(X_1, Y_2)]$$
$$= \mathbb{E}_\mu\left[\tilde{\eta}(X_1, Y_1) f_{1,1}(X_1, Y_1)\right],$$

since ξ is the Radon-Nikodym derivative of μ with respect to $P \otimes Q$ under $\mathbb{E}_\mu$. On the other hand, we also have, by (C.34) and (C.35),

$$\mathbb{E}_\mu\left[\tilde{T}_n \sum_{i,j=1}^n f_{1,1}(X_i, Y_j)\right] = n\mathbb{E}_\mu[\tilde{T}_n f_{1,1}(X_1, Y_1)] = \mathbb{E}_\mu[\tilde{\eta}(X_1, Y_1) f_{1,1}(X_1, Y_1)].$$

Hence, $\mathbb{E}_\mu\left[(\tilde{T}_n - \tilde{\mathcal{Q}}_n)\sum_{i,j=1}^n f(X_i, Y_j)\right] = 0$ and the claim follows.

Step 3. We control the variance of $\mathrm{Proj}_{H_2}(T_n) - \tilde{\mathcal{Q}}_n$. From Step 2 we know $\mathrm{Proj}_{H_2}(\tilde{\mathcal{Q}}_n) = \mathrm{Proj}_{H_2}(T_n)$. By the definition of $\mathbf{L}^2$ projection, it holds

$$\mathbb{E}_\mu[(\mathrm{Proj}_{H_2}(T_n) - \tilde{\mathcal{Q}}_n)^2] = \mathbb{E}_\mu[(\mathrm{Proj}_{H_2}(\tilde{\mathcal{Q}}_n) - \tilde{\mathcal{Q}}_n)^2] = \min_{V \in H_2} \mathbb{E}_\mu[(\tilde{\mathcal{Q}}_n - V)^2] \leq \mathbb{E}_\mu[(\tilde{\mathcal{Q}}_n - \mathcal{Q}_n)^2],$$

since $\mathcal{Q}_n \in H_2$. Note that

$$\mathcal{Q}_n - \tilde{\mathcal{Q}}_n = \frac{1}{n(n-1)} \sum_{i=1}^n [\kappa_{1,1'}(X_i, Y_i) - \ell_{1,1'}(X_i, Y_i)].$$

By independence, we get

$$\mathbb{E}_\mu[(\tilde{\mathcal{Q}}_n - \mathcal{Q}_n)^2] = \frac{1}{n^2(n-1)^2} \sum_{i=1}^n \mathbb{E}_\mu[(\kappa_{1,1'}(X_i, Y_i) - \ell_{1,1'}(X_i, Y_i))^2] = O(n^{-3}).$$

It follows that $\tilde{\mathcal{Q}}_n = \mathrm{Proj}_{H_2}(T_n) + o_p(n^{-1})$. □

C.5.2 Variance bound for the second order remainder

Npte that the second order remainder is $R_2 := T_n - \theta - \mathcal{L}_n - \mathcal{Q}_n = (U_n - \mathcal{Q}_n D_n)/D_n$, where

$$\mathcal{Q}_n := \frac{1}{n(n-1)} \left\{ \sum_{i \neq j} [\kappa_{2,0}(X_i, X_j) + \kappa_{0,2}(Y_i, Y_j)] + \sum_{i,j=1}^n \kappa_{1,1'}(X_i, Y_j) - \sum_{i=1}^n \ell_{1,1'}(X_i, Y_i) \right\}.$$

is an approximate second order chaos of T_n by Proposition C.15. We will decompose $\mathcal{Q}_n D_n$ into manageable pieces. Let $K_{2,0}(x,x',y,y') := \kappa_{2,0}(x,x')\xi(x,y)\xi(x',y')$ and $K_{0,2}(x,x',y,y') := \kappa_{0,2}(y,y')\xi(x,y)\xi(x',y')$. Then we have

$$\sum_{i\neq j}\kappa_{2,0}(X_i,X_j)D_n = \frac{1}{n!}\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n} K_{2,0}(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in[n]\setminus\{i,j\}}\xi(X_k,Y_{\sigma_k}) \tag{C.40}$$

$$\begin{aligned}\sum_{i\neq j}\kappa_{0,2}(Y_i,Y_j)D_n &= \frac{1}{n!}\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}\kappa_{0,2}(Y_i,Y_j)\xi(X_{\sigma_i^{-1}},Y_i)\xi(X_{\sigma_j^{-1}},Y_j)\prod_{k\in[n]\setminus\{\sigma_i^{-1},\sigma_j^{-1}\}}\xi(X_k,Y_{\sigma_k})\\ &= \frac{1}{n!}\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n} K_{0,2}(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in[n]\setminus\{i,j\}}\xi(X_k,Y_{\sigma_k}). \end{aligned}\tag{C.41}$$

Furthermore, let $K_{1,1'}(x,x',y,y') := \kappa_{1,1'}(x,y')\xi(x,y)\xi(x',y')$, then

$$\begin{aligned}&\frac{1}{n!}\sum_{i,j=1}^n\sum_{\sigma_i\neq j}\kappa_{1,1'}(X_i,Y_j)\xi^{\otimes}(X,Y_\sigma)\\ &= \frac{1}{n!}\sum_{i,j=1}^n\sum_{j'\in[n]\setminus\{i\}}\sum_{\sigma_{j'}=j} K_{1,1'}(X_i,X_{j'},Y_{\sigma_i},Y_{\sigma_{j'}})\prod_{k\in[n]\setminus\{i,j'\}}\xi(X_k,Y_{\sigma_k})\\ &= \frac{1}{n!}\sum_{i\neq j'}\sum_{\sigma\in\mathcal{S}_n} K_{1,1'}(X_i,X_{j'},Y_{\sigma_i},Y_{\sigma_{j'}})\prod_{k\in[n]\setminus\{i,j'\}}\xi(X_k,Y_{\sigma_k}).\end{aligned}\tag{C.42}$$

Note that $\sum_{i=1}^n \ell_{1,1'}(X_i,Y_i) = \sum_{i=1}^n \ell_{1,1'}(X_i,Y_{\sigma_i})$ by affineness, and

$$\frac{1}{n!}\sum_{i,j=1}^n\sum_{\sigma_i=j}\kappa_{1,1'}(X_i,Y_j)\xi^{\otimes}(X,Y_\sigma) = \frac{1}{n!}\sum_{i=1}^n\sum_{\sigma\in\mathcal{S}_n}\kappa_{1,1'}(X_i,Y_{\sigma_i})\xi^{\otimes}(X,Y_\sigma).$$

It follows that

$$\begin{aligned}&\frac{1}{n!}\sum_{i,j=1}^n\sum_{\sigma_i=j}\kappa_{1,1'}(X_i,Y_j)\xi^{\otimes}(X,Y_\sigma) - \sum_{i=1}^n \ell_{1,1'}(X_i,Y_i)D_n\\ &= \frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i=1}^n[\kappa_{1,1'}-\ell_{1,1'}](X_i,Y_{\sigma_i})\xi^{\otimes}(X,Y_\sigma).\end{aligned}$$

Repeating the argument in Proposition 4.9 for $\tilde{\eta}$ replaced by $\kappa_{1,1'}-\ell_{1,1'}\in\mathbf{L}^2_{0,0}(\mu)$ gives

$$\frac{1}{n(n-1)}\frac{1}{n!}\sum_{i,j=1}^n\sum_{\sigma_i=j}\kappa_{1,1'}(X_i,Y_j)\xi^{\otimes}(X,Y_\sigma) - \sum_{i=1}^n \ell_{1,1'}(X_i,Y_i)D_n = O(n^{-2}). \tag{C.43}$$

Here we say a random variable $\phi_n = O(n^{-2})$ if $\mathbb{V}\mathrm{ar}(\phi_n) = O(n^{-4})$. Putting (C.40), (C.41), (C.42) and (C.43) together, we know that $\mathcal{Q}_n D_n$ is equal to

$$\frac{1}{n(n-1)}\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i\neq j}(K_{2,0}+K_{0,2}+K_{1,1'})(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in[n]\setminus\{i,j\}}\xi(X_k,Y_{\sigma_k})+O(n^{-2}).$$

In the following, we further decompose $K_{2,0}+K_{0,2}+K_{1,1'}$ into second, third and fourth order terms using Hoeffding decomposition, and show that the second order terms cancel out U_n and the rest of the terms are negligible.

The following lemma gives the second order terms of $K_{2,0}$, $K_{0,2}$ and $K_{1,1'}$.

Lemma C.16. *Let*

$$\begin{aligned}
k_{2,0}(x,x',y,y') &:= \kappa_{2,0}(x,x') + (\mathcal{A}\otimes\mathcal{A})\kappa_{2,0}(y,y') + (I_P\otimes\mathcal{A})\kappa_{2,0}(x,y') \\
&\quad + (I_P\otimes\mathcal{A})\mathcal{T}\kappa_{2,0}(x',y) \\
k_{0,2}(x,x',y,y') &:= (\mathcal{A}^*\otimes\mathcal{A}^*)\kappa_{0,2}(x,x') + \kappa_{0,2}(y,y') + (\mathcal{A}^*\otimes I_Q)\kappa_{0,2}(x,y') \\
&\quad + (\mathcal{A}^*\otimes I_Q)\mathcal{T}\kappa_{0,2}(x',y) \\
k_{1,1'}(x,x',y,y') &:= (I_P\otimes\mathcal{A}^*)\kappa_{1,1'}(x,x') + \mathcal{T}(\mathcal{A}\otimes I_Q)\kappa_{1,1'}(y,y') + \kappa_{1,1'}(x,y') + \mathcal{B}\kappa_{1,1'}(x',y).
\end{aligned}$$

For any $i\neq i'$ and $j\neq j'$, the function $\bar{K}_I(X_i,X_{i'},Y_j,Y_{j'}) := (K_I - k_I)(X_i,X_{i'},Y_j,Y_{j'})$ is 2-degenerate for every $I = \{2,0\},\{0,2\},\{1,1'\}$.

Proof. We only prove the claim for $I=\{2,0\}$. The rest of them can be proved similarly. Recall that $K_{2,0}(x,x',y,y') := \kappa_{2,0}(x,x')\xi(x,y)\xi(x',y')$. Conditioning on $X_i, X_{i'}$, we have

$$\begin{aligned}
&\mathbb{E}[K_{2,0}(X_i,X_{i'},Y_j,Y_{j'}) \mid X_i,X_{i'}] \\
&= \kappa_{2,0}(X_i,X_{i'})\,\mathbb{E}[\xi(X_i,Y_j)\mid X_i]\,\mathbb{E}[\xi(X_{i'},Y_{j'})\mid X_{i'}] = \kappa_{2,0}(X_i,X_{i'}).
\end{aligned}$$

It then follows from degeneracy that $\mathbb{E}[(K_{2,0}-k_{2,0})(X_i,X_{i'},Y_j,Y_{j'}) \mid X_i,X_{i'}] = 0$. Conditioning on X_i, Y_j, we have

$$\begin{aligned}
&\mathbb{E}[K_{2,0}(X_i,X_{i'},Y_j,Y_{j'}) \mid X_i,Y_j] \\
&= \xi(X_i,Y_j)\,\mathbb{E}[\kappa_{2,0}(X_i,X_{i'})\mid X_i,Y_j] = 0 = \mathbb{E}[k_{2,0}(X_i,X_{i'},Y_j,Y_{j'})\mid X_i,Y_j].
\end{aligned}$$

Conditioning on $X_i, Y_{j'}$, we have

$$\begin{aligned}\mathbb{E}[K_{2,0}(X_i, X_{i'}, Y_j, Y_{j'}) \mid X_i, Y_{j'}] &= \mathbb{E}[\kappa_{2,0}(X_i, X_{i'})\xi(X_{i'}, Y_{j'}) \mid X_i, Y_{j'}] = (I_P \otimes \mathcal{A})\kappa_{2,0}(X_i, Y_{j'}) \\ &= \mathbb{E}[k_{2,0}(X_i, X_{i'}, Y_j, Y_{j'}) \mid X_i, Y_{j'}].\end{aligned}$$

The rest follows analogously. □

Now, we get

$$\mathcal{Q}_n D_n = W_n + V_n + O(n^{-2}), \tag{C.44}$$

where

$$W_n := \frac{1}{n(n-1)}\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i\neq j}(\bar{K}_{2,0} + \bar{K}_{0,2} + \bar{K}_{1,1'})(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j})\prod_{k\in[n]\setminus\{i,j\}}\xi(X_k, Y_{\sigma_k}) \tag{C.45}$$

$$V_n := \frac{1}{n(n-1)}\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i\neq j}(k_{2,0} + k_{0,2} + k_{1,1'})(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j})\prod_{k\in[n]\setminus\{i,j\}}\xi(X_k, Y_{\sigma_k}). \tag{C.46}$$

We will show that $\mathbb{E}[(U_n - V_n)^2] = O(n^{-4})$ and $\mathbb{E}[W_n^2] = O(n^{-4})$. As a result, $\mathbb{E}[(U_n - \mathcal{Q}_n D_n)^2] = O(n^{-4})$, which implies $R_2 := R_1 - \mathcal{Q}_n = o_p(n^{-1})$.

Lemma C.17. *The following algebraic identity holds:*

$$V_n = \frac{1}{n(n-1)}\frac{1}{n!}\sum_{i,j=1}^{n}\sum_{\sigma_i\neq j}\tilde{\eta}(X_i, Y_j)\xi(X_i, Y_j)\prod_{k\in[n]\setminus\{i,\sigma_j^{-1}\}}\xi(X_k, Y_{\sigma_k}). \tag{C.47}$$

Moreover, under Assumptions 4.2 and 4.3, $\mathbb{E}[(U_n - V_n)^2] = O(n^{-4})$.

Proof. We consider the terms involving (X_i, X_j) and $(Y_{\sigma_i}, Y_{\sigma_j})$ in $\sum_{i\neq j}(k_{2,0} + k_{0,2} + k_{1,1'})(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j})$. By Lemma C.14, we get

$$\sum_{i\neq j}[\kappa_{2,0}(X_i, X_j) + (\mathcal{A}^* \otimes \mathcal{A}^*)\kappa_{0,2}(X_i, X_j) + (I_P \otimes \mathcal{A}^*)\kappa_{1,1'}(X_i, X_j)] = 0$$

$$\sum_{i\neq j}[(\mathcal{A} \otimes \mathcal{A})\kappa_{2,0}(Y_{\sigma_i}, Y_{\sigma_j}) + \kappa_{0,2}(Y_{\sigma_i}, Y_{\sigma_j}) + (\mathcal{A} \otimes I_Q)\kappa_{1,1'}(Y_{\sigma_i}, Y_{\sigma_j})] = 0.$$

We then consider the terms involving (X_i, Y_{σ_j}) and (X_j, Y_{σ_i}). Notice that

$$\begin{aligned}
&\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}(I_P\otimes\mathcal{A})\kappa_{2,0}(X_i,Y_{\sigma_j})\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k})\\
&=\sum_{i\neq j}\sum_{j'=1}^{n}\sum_{\sigma_j=j'}(I_P\otimes\mathcal{A})\kappa_{2,0}(X_i,Y_{j'})\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k})\\
&=\sum_{i,j'=1}^{n}\sum_{\sigma_i\neq j'}(I_P\otimes\mathcal{A})\kappa_{2,0}(X_i,Y_{j'})\prod_{k\in[n]\backslash\{i,\sigma_{j'}^{-1}\}}\xi(X_k,Y_{\sigma_k}).
\end{aligned}$$

A similar argument gives

$$\begin{aligned}
&\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}(I_P\otimes\mathcal{A})\mathcal{T}\kappa_{2,0}(X_j,Y_{\sigma_i})\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k})\\
&=\sum_{i',j=1}^{n}\sum_{\sigma_j\neq i'}(I_P\otimes\mathcal{A})\mathcal{T}\kappa_{2,0}(X_j,Y_{i'})\prod_{k\in[n]\backslash\{j,\sigma_{i'}^{-1}\}}\xi(X_k,Y_{\sigma_k}).
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}[(I_P\otimes\mathcal{A})\kappa_{2,0}(X_i,Y_{\sigma_j})+(I_P\otimes\mathcal{A})\mathcal{T}\kappa_{2,0}(X_j,Y_{\sigma_i})]\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k}) \qquad \text{(C.48)}\\
&=\sum_{i,j=1}^{n}\sum_{\sigma_i\neq j}(I_P\otimes\mathcal{A})(I+\mathcal{T})\kappa_{2,0}(X_i,Y_j)\prod_{k\in[n]\backslash\{i,\sigma_j^{-1}\}}\xi(X_k,Y_{\sigma_k}).
\end{aligned}$$

Analogously,

$$\begin{aligned}
&\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}[(\mathcal{A}^*\otimes I_Q)\kappa_{0,2}(X_i,Y_{\sigma_j})+(\mathcal{A}^*\otimes I_Q)\mathcal{T}\kappa_{0,2}(X_j,Y_{\sigma_i})]\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k}) \qquad \text{(C.49)}\\
&=\sum_{i,j=1}^{n}\sum_{\sigma_i\neq j}(\mathcal{A}^*\otimes I_Q)(I+\mathcal{T})\kappa_{0,2}(X_i,Y_j)\prod_{k\in[n]\backslash\{i,\sigma_j^{-1}\}}\xi(X_k,Y_{\sigma_k}),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i\neq j}\sum_{\sigma\in\mathcal{S}_n}[\kappa_{1,1'}(X_i,Y_{\sigma_j})+\mathcal{B}\kappa_{1,1'}(X_j,Y_{\sigma_i})]\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k}) \qquad \text{(C.50)}\\
&=\sum_{i,j=1}^{n}\sum_{\sigma_i\neq j}(I+\mathcal{B})\kappa_{1,1'}(X_i,Y_j)\prod_{k\in[n]\backslash\{i,\sigma_j^{-1}\}}\xi(X_k,Y_{\sigma_k}).
\end{aligned}$$

Hence, the identity (C.47) follows from the third identity in Lemma C.14.

Let us compute $\mathbb{E}[(U_n - V_n)^2]$. Denote $h := \tilde{\eta}\xi$. Recall from (4.35) that

$$U_n := \frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\frac{1}{n}\sum_{i=1}^{n}\tilde{\eta}(X_i,Y_{\sigma_i})\xi^{\otimes}(X,Y_\sigma) = \frac{1}{n\cdot n!}\sum_{i,j=1}^{n}\sum_{\sigma_i=j}h(X_i,Y_j)\prod_{k\in[n]\setminus\{i\}}\xi(X_k,Y_{\sigma_k}).$$

By Lemma C.8, we get

$$U_n = \frac{1}{n\cdot n!}\sum_{i,j=1}^{n}\sum_{\sigma_i=j}h(X_i,Y_j)\sum_{A\subset[n]\setminus\{i\}}\prod_{k\in A}[\xi(X_k,Y_{\sigma_k})-1]. \tag{C.51}$$

Similarly,

$$V_n = \frac{1}{n(n-1)}\frac{1}{n!}\sum_{i,j=1}^{n}\sum_{\sigma_i\neq j}h(X_i,Y_j)\sum_{A\subset[n]\setminus\{i,\sigma_j^{-1}\}}\prod_{k\in A}[\xi(X_k,Y_{\sigma_k})-1]. \tag{C.52}$$

Define the set of sequences of length r to be

$$\mathrm{S}_{n,r} := \{(k_i)_{i=1}^r : k_i\in[n], |\{k_1,\dots,k_r\}| = r\}, \quad \text{for } r\in[n].$$

Take $r\in[n]$ and $(k_i)_{i=1}^r, (k_i')_{i=1}^r \in \mathrm{S}_{N,r}$. Let us count the number of times the term

$$h(X_{k_1},Y_{k_1'})\prod_{s=2}^{r}[\xi(X_{k_s},Y_{k_s'})-1] \tag{C.53}$$

appears in (C.51) and (C.52), respectively. In order to get this term, we must have $i = k_1$, $j = k_1'$, $A = \{k_2,\dots,k_r\}$ and $\sigma_{k_s} = k_s'$ for all $s\in\{2,\dots,r\}$. Note that $\sigma_i = j$ in (C.51), so there are $(n-r)!$ such terms in (C.51). Similarly, there are $(n-r)(n-r)!$ such terms in (C.52). Hence, the coefficient of this term in $U_n - V_n$ is

$$C_{n,r} = \frac{(n-r)!}{n\cdot n!} - \frac{(n-r)(n-r)!}{n(n-1)\cdot n!} = \frac{r-1}{n-1}\frac{(n-r)!}{n\cdot n!}.$$

We claim that

$$U_n - V_n = \frac{1}{n(n-1)}\frac{1}{n!}\sum_{r=1}^{n}(r-1)\sum_{|A|=|B|=r}\sum_{\sigma\in\mathcal{S}_n:\sigma_A=B}\sum_{i\in A}h(X_i,Y_{\sigma_i})\prod_{j\in A\setminus\{i\}}[\xi(X_j,Y_{\sigma_j})-1]. \tag{C.54}$$

To see this, we only need to prove that the coefficient of the term (C.53) on the right hand side of (C.54) is exactly $C_{n,r}$. In other words, it appears $(n-r)!$ times in the following sum:

$$\sum_{|A|=|B|=r} \sum_{\sigma \in \mathcal{S}_n : \sigma_A = B} \sum_{i \in A} h(X_i, Y_{\sigma_i}) \prod_{j \in A \setminus \{i\}} [\xi(X_j, Y_{\sigma_j}) - 1].$$

To get this term, we must have $A = \{k_1, \dots, k_r\}$, $B = \{k'_1, \dots, k'_r\}$, $i = k_1$ and $\sigma_{k_s} = k'_s$ for all $s \in [r]$. There are $(n-r)!$ permutations satisfy this condition, and thus it appears $(n-r)!$ times.

A derivation analogous to the one for Proposition 4.13 implies that $\mathbb{E}[(U_n - V_n)^2]$ equals

$$\frac{1}{n^2(n-1)^2} \sum_{r=1}^{n} \frac{r(r-1)^2}{r!} \sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} \mathbb{E}\Bigg[h(X_1, Y_1) \prod_{j=2}^{r} [\xi(X_j, Y_j) - 1] \\ h(X_i, Y_{\sigma_i}) \prod_{j \in [n] \setminus \{i\}} [\xi(X_j, Y_{\sigma_j}) - 1] \Bigg].$$

Repeating the argument in Proposition 4.9, we know $\mathbb{E}[(U_n - V_n)^2] = O(n^{-4})$. □

Before we bound $\mathbb{E}[W_n^2]$, let us give a result similar to Lemma C.9 for functions with 3 and 4 arguments. Let $\phi \in \mathbf{L}^2(P \otimes P \otimes Q \otimes Q)$ and $\psi \in \mathbf{L}^2(P \otimes P \otimes Q)$ such that $\phi(X_1, X_2, Y_1, Y_2)$ and $\psi(X_1, X_2, Y_1)$ are completely degenerate under the measure $(P \otimes Q)^n$.

Lemma C.18. *Assume $\|\phi\|_{\mathbf{L}^2(P \otimes P \otimes Q \otimes Q)} < \infty$ and $\|\psi\|_{\mathbf{L}^2(P \otimes P \otimes Q)} < \infty$. Under Assumptions 4.2, 4.3 and C.1, there exists a constant C such that, for any $\sigma \in \mathcal{S}_n$ and $i \neq j \in [n]$,*

$$\mathbb{E}\left[\phi(X_1, X_2, Y_1, Y_2) \prod_{k=3}^{n} [\xi(X_k, Y_k) - 1] \phi(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j}) \prod_{k \in [n] \setminus \{i,j\}} [\xi(X_k, Y_{\sigma_k}) - 1] \right] \\ \leq s_1^{2(N - \#\sigma - 2)} C^{\#\sigma}$$

$$\mathbb{E}\left[\psi(X_1, X_2, Y_1) \prod_{k=3}^{n} [\xi(X_k, Y_k) - 1] \psi(X_i, X_j, Y_{\sigma_i}) \prod_{k \in [n] \setminus \{i,j\}} [\xi(X_k, Y_{\sigma_k}) - 1] \right] \\ \leq s_1^{2(N - \#\sigma - 2)} C^{\#\sigma},$$

where $\#\sigma$ is the number of cycles of $\sigma \in \mathcal{S}_n$.

The proof of Lemma C.18 is similar to Lemma C.9—we iteratively take expectation with respect to a single variable, while keeping the rest being fixed. In consideration of the space, we only give an example here.

Example C.2. *Consider $n = 4$, $i = 2$, $j = 3$ and σ given by $\sigma_i = i + 1$ for $i \in [3]$. By construction, σ only has one cycle $1 \to 2 \to 3 \to 4 \to 1$. The expectation of interest then reads*

$$\begin{aligned}\mathbb{E}\big[&\phi(X_1, X_2, Y_1, Y_2)[\xi(X_3, Y_3) - 1][\xi(X_4, Y_4) - 1]\\ &\phi(X_2, X_3, Y_3, Y_4)[\xi(X_1, Y_2) - 1][\xi(X_4, Y_1) - 1]\big].\end{aligned}$$

Let $\mathcal{A}_4$ be a shorthand notation for $I_P \otimes I_P \otimes I_Q \otimes \mathcal{A}$, and $\mathcal{A}_4^*$ similarly. Taking expectation with respect to Y_4, while keeping others being fixed, we get

$$\mathbb{E}\left[\phi(X_1, X_2, Y_1, Y_2)[\xi(X_3, Y_3) - 1](\mathcal{A}_4^*\phi)(X_2, X_3, Y_3, X_4)[\xi(X_1, Y_2) - 1][\xi(X_4, Y_1) - 1]\right],$$

since

$$\begin{aligned}&\mathbb{E}\left[\phi(X_2, X_3, Y_3, Y_4)[\xi(X_4, Y_4) - 1] \mid X_2, X_3, X_4, Y_3\right] && \text{(C.55)}\\ =\ &\mathbb{E}\left[\phi(X_2, X_3, Y_3, Y_4)\xi(X_4, Y_4) \mid X_2, X_3, X_4, Y_3\right]\\ =\ &\mathcal{A}_4^*\phi(X_2, X_3, Y_3, X_4). && \text{(C.56)}\end{aligned}$$

Now taking expectation with respect to X_4, while keeping others being fixed, we get

$$\mathbb{E}\left[\phi(X_1, X_2, Y_1, Y_2)[\xi(X_3, Y_3) - 1](\mathcal{A}_4\mathcal{A}_4^*\phi)(X_2, X_3, Y_3, Y_1)[\xi(X_1, Y_2) - 1]\right]$$

Now, both X_1 and Y_2 in $\xi(X_1, Y_2) - 1$ appears in $\phi(X_1, X_2, Y_1, Y_2)$, and both X_3 and Y_3 in $\xi(X_3, Y_3) - 1$ appears in $\phi(X_2, X_3, Y_3, Y_4)$, so we stop here and use the Cauchy-Schwarz inequality to get an upper bound

$$\begin{aligned}&\sqrt{\mathbb{E}[(\mathcal{A}_4\mathcal{A}_4^*\phi)^2(X_2, X_3, Y_3, Y_1)[\xi(X_1, Y_2) - 1]^2] \times \mathbb{E}\left[\phi^2(X_1, X_2, Y_1, Y_2)[\xi(X_3, Y_3) - 1]^2\right]}\\ =\ &\|(\mathcal{A}_4\mathcal{A}_4^*)\phi\|_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} \|\phi\|_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} \|\xi - 1\|^2_{\mathbf{L}^2(P\otimes Q)}, \quad \textit{by independence.} && \text{(C.57)}\end{aligned}$$

Let $C := \|\phi\|^2_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} \|\xi - 1\|^2_{\mathbf{L}^2(P\otimes Q)}$. *Then* (C.57) *can be further bounded above by* Cs_1^2.

For the expectation associated with ψ, *we view* ψ *as a function with four arguments such that it is constant in its fourth argument and then repeat the argument for* ϕ. *It only makes a difference at places where we apply* $\mathcal{A}_4$ *or* $\mathcal{A}_4^*$ *to* ϕ*—instead of applying this operator, the expectation is exactly zero, and thus the bound holds trivially. To be more specific, in the first step of the above example, where we take expectation with respect to* Y_4, *we should have, in* (C.56), *that*

$$\mathbb{E}\left[\psi(X_2, X_3, Y_3)[\xi(X_4, Y_4) - 1] \mid X_2, X_3, X_4, Y_3\right] = \psi(X_2, X_3, Y_3)\, \mathbb{E}[\xi(X_4, Y_4) - 1 \mid X_4] \overset{a.s.}{=} 0.$$

Recall from (C.45) that

$$W_n := \frac{1}{n(n-1)} \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \sum_{i \neq j} (\bar{K}_{2,0} + \bar{K}_{0,2} + \bar{K}_{1,1'})(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j}) \prod_{k \in [n] \setminus \{i,j\}} \xi(X_k, Y_{\sigma_k}).$$

To prove $\mathbb{E}[W_n^2] = O(n^{-4})$, we again use Hoeffding decomposition. From Lemma C.16 we know $(\bar{K}_{2,0} + \bar{K}_{0,2} + \bar{K}_{1,1'})(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j})$ is 2-degenerate, so each term in its Hoeffding decomposition should contain at least 3 variables. We assume it is given by the following form:

$$\phi(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j}) + \psi_0(X_i, X_j, Y_{\sigma_i}) + \psi_1(X_i, X_j, Y_{\sigma_j}) + \psi_2(X_i, Y_{\sigma_i}, Y_{\sigma_j}) + \psi_3(X_j, Y_{\sigma_i}, Y_{\sigma_j}).$$

Define

$$\begin{aligned} W_n^{\phi} &:= \frac{1}{n(n-1)} \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \sum_{i \neq j} \phi(X_i, X_j, Y_{\sigma_i}, Y_{\sigma_j}) \prod_{k \in [n] \setminus \{i,j\}} \xi(X_k, Y_{\sigma_k}) \\ W_n^{\psi_0} &:= \frac{1}{n(n-1)} \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \sum_{i \neq j} \psi_0(X_i, X_j, Y_{\sigma_i}) \prod_{k \in [n] \setminus \{i,j\}} \xi(X_k, Y_{\sigma_k}), \end{aligned}$$

and $W_n^{\psi_1}$, $W_n^{\psi_2}$ and $W_n^{\psi_3}$, similarly. Consequently, $W_n = W_n^{\phi} + W_n^{\psi_0} + W_n^{\psi_1} + W_n^{\psi_2} + W_n^{\psi_3}$. It then suffices to show $\mathbb{E}[(W_n^{\phi})^2] = O(n^{-4})$ and $\mathbb{E}[(W_n^{\psi_i})^2] = O(n^{-4})$ for $i \in \{0, 1, 2, 3\}$. The strategy here is the same as Proposition 4.9.

Corollary C.19. *Suppose the same assumptions in Lemma C.18 hold. Then*

$$\mathbb{E}[(W_n^{\phi})^2] \leq \frac{1}{n^2(n-1)^2}\sum_{r=2}^{n}\frac{r^2(r-1)^2}{r!}\sum_{\sigma\in\mathcal{S}_r} s_1^{2(r-\#\sigma-2)}C^{\#\sigma}$$

$$\mathbb{E}[(W_n^{\psi_i})^2] \leq \frac{1}{n^2(n-1)^2}\sum_{r=2}^{n}\frac{r^2(r-1)^2}{r!}\sum_{\sigma\in\mathcal{S}_r} s_1^{2(r-\#\sigma-2)}C^{\#\sigma}, \quad \textit{for } i\in\{0,1,2,3\}$$

In particular, $\mathbb{E}[(W_n^{\phi})^2] = O(n^{-4})$ *and* $\mathbb{E}[(W_n^{\psi_i})^2] = O(n^{-4})$ *for* $i \in \{0,1,2,3\}$.

Proof. We only prove the bound for $\mathbb{E}[(W_n^{\phi})^2]$. Notice that, using Lemma C.8 for $A = [n]\backslash\{i,j\}$, we have $\prod_{k\in[n]\backslash\{i,j\}}\xi(X_k,Y_{\sigma_k}) = \sum_{C\subset[n]\backslash\{i,j\}}\prod_{k\in C}[\xi(X_k,Y_{\sigma_k})-1]$ for every pair $i\neq j$. As a result,

$$W_n^{\phi} = \frac{1}{n(n-1)}\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\sum_{i\neq j}\phi(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\sum_{C\subset[n]\backslash\{i,j\}}\prod_{k\in C}[\xi(X_k,Y_{\sigma_k})-1]. \tag{C.58}$$

Because $\phi(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})$ is completely degenerate, an argument similar to the one in Proposition 4.13 shows that the Hoeffding decomposition of W_n^{ϕ} is given by

$$W_n^{\phi} := \frac{1}{n(n-1)}\frac{1}{n!}\sum_{|A|=|B|>1} W_{AB}^{\phi},$$

where

$$W_{AB}^{\phi} := \sum_{\sigma\in\mathcal{S}_n:\sigma_A=B}\sum_{i\neq j\in A}\phi(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in A\backslash\{i,j\}}[\xi(X_k,Y_{\sigma_k})-1].$$

Consequently,

$$\begin{aligned}\mathbb{E}[(W_n^{\phi})^2] &= \frac{1}{n^2(n-1)^2(n!)^2}\sum_{r=2}^{n}\sum_{|A|=|B|=r}\mathbb{E}[(W_{AB}^{\phi})^2]\\ &= \frac{1}{n^2(n-1)^2(n!)^2}\sum_{r=2}^{n}\binom{n}{r}^2\mathbb{E}[(W_{[r][r]}^{\phi})^2],\end{aligned} \tag{C.59}$$

where the last equality follows from exchangeability. Using a derivation similar to the one for Proposition 4.9,

$$
\begin{aligned}
\mathbb{E}[(W^{\phi}_{[r][r]})^2] &= ((n-r)!)^2\,\mathbb{E}\left[\sum_{\sigma\in\mathcal{S}_r}\sum_{1\le i\neq j\le r}\phi(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in[r]\setminus\{i,j\}}[\xi(X_k,Y_{\sigma_k})-1]\right]^2\\
&= \big((n-r)!\big)^2 r!r(r-1)\sum_{\sigma\in\mathcal{S}_r}\sum_{1\le i\neq j\le r}\mathbb{E}\Bigg[\phi(X_1,X_2,Y_1,Y_2)\prod_{k=3}^{r}[\xi(X_k,Y_k)-1]\\
&\qquad \phi(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})\prod_{k\in[r]\setminus\{i,j\}}[\xi(X_k,Y_{\sigma_k})-1]\Bigg]\\
&\le \big((n-r)!\big)^2 r!r(r-1)\sum_{\sigma\in\mathcal{S}_r}\sum_{1\le i\neq j\le r}s_1^{2(r-\#\sigma-2)}C^{\#\sigma},\quad \text{by Lemma C.18.} \qquad \text{(C.60)}
\end{aligned}
$$

Now, putting (C.59) and (C.60) together, we get

$$
\begin{aligned}
\mathbb{E}[(W^{\phi}_{n})^2] &\le \frac{1}{n^2(n-1)^2(n!)^2}\sum_{r=2}^{n}\binom{n}{r}^2\big((n-r)!\big)^2 r!r(r-1)\sum_{\sigma\in\mathcal{S}_r}\sum_{1\le i\neq j\le r}s_1^{2(r-\#\sigma-2)}C^{\#\sigma}\\
&= \frac{1}{n^2(n-1)^2}\sum_{r=2}^{n}\frac{r^2(r-1)^2}{r!}\sum_{\sigma\in\mathcal{S}_r}s_1^{2(r-\#\sigma-2)}C^{\#\sigma}.
\end{aligned}
$$

□

Proposition C.20. *Under Assumptions 4.2, 4.3 and C.1, the second order remainder* $R_2 = o_p(n^{-1})$.

Proof. Let $f := \mathcal{C}^{-1}(\tilde{\eta}\xi)$. Recall $\mathfrak{p}$ and $\mathfrak{q}$ from Assumption C.1. Note that

$$
\begin{aligned}
\mathbb{E}[\kappa^{2\mathfrak{q}}_{2,0}(X_1,X_2)] &= \int[(I_P\otimes\mathcal{A}^*)f(x,x')]^{2\mathfrak{q}}\mathrm{d}P(x)\mathrm{d}P(x')\\
&= \int\left[\int f(x,y')\xi(x',y')\mathrm{d}Q(y')\right]^{2\mathfrak{q}}\mathrm{d}P(x)\mathrm{d}P(x')\\
&\overset{\text{Jensen}}{\le} \iint f^{2\mathfrak{q}}(x,y')\xi(x',y')\mathrm{d}Q(y')\mathrm{d}P(x)\mathrm{d}P(x').
\end{aligned}
$$

Since $\int\xi(x',y')\mathrm{d}P(x') \overset{\text{a.s.}}{=} 1$, integrating with respect to x' in the above upper bound gives

$$
\int f^{2\mathfrak{q}}(x,y')\mathrm{d}Q(y')\mathrm{d}P(x) = \mathbb{E}[f^{2\mathfrak{q}}(X_1,Y_1)] < \infty.
$$

As a result,

$$\begin{aligned}\|K_{2,0}\|^2_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} &= \mathbb{E}\left[\kappa^2_{2,0}(X_1,X_2)\xi^2(X_1,Y_1)\xi^2(X_2,Y_2)\right] \\ &\overset{\text{Hölder}}{\leq} \mathbb{E}[\kappa^{2\mathfrak{q}}_{2,0}(X_1,X_2)]^{\frac{1}{\mathfrak{q}}}\,\mathbb{E}[\xi^{2\mathfrak{p}}(X_1,Y_1)\xi^{2\mathfrak{p}}(X_2,Y_2)]^{\frac{1}{\mathfrak{p}}} \\ &= \mathbb{E}[\kappa^{2\mathfrak{q}}_{2,0}(X_1,X_2)]^{\frac{1}{\mathfrak{q}}}\,\mathbb{E}[\xi^{2\mathfrak{p}}(X_1,Y_1)]^{\frac{2}{\mathfrak{p}}} < \infty.\end{aligned}$$

Analogously, we have $\|K_{0,2}\|_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} < \infty$ and $\|K_{1,1'}\|_{\mathbf{L}^2(P\otimes P\otimes Q\otimes Q)} < \infty$. As discussed before Corollary C.19, we can then decompose $(\bar{K}_{2,0}+\bar{K}_{0,2}+\bar{K}_{1,1'})(X_i,X_j,Y_{\sigma_i},Y_{\sigma_j})$ into third and fourth order terms using Hoeffding decomposition and invoke Corollary C.19 to show $\mathbb{E}[W_n^2] = O(n^{-4})$. Recall from (C.44) that $\mathbb{E}[(\mathcal{Q}_nD_n - W_n - V_n)^2] = O(n^{-4})$. Hence, by Lemma C.17,

$$\mathbb{E}[(U_n - \mathcal{Q}_nD_n)^2] \leq 3\left\{\mathbb{E}[(U_n-V_n)^2] + \mathbb{E}[W_n^2] + \mathbb{E}[(\mathcal{Q}_nD_n - V_n - W_n)^2]\right\} = O(n^{-4}).$$

It then follows from Theorem 4.5 that $R_2 = o_p(n^{-1})$. □

C.5.3 Proof of Theorem C.11

Proof of Theorem C.11. By the assumption that $\varsigma^2 = 0$, we know the first order chaos $\mathcal{L}_n = 0$ almost surely. According to Proposition C.20, it holds that $T_n - \theta - \mathcal{Q}_n = o_p(n^{-1})$. Recall from (C.29) that

$$\mathcal{Q}_n := \frac{1}{n(n-1)}\left\{\sum_{i\neq j}[\kappa_{2,0}(X_i,X_j)+\kappa_{0,2}(Y_i,Y_j)] + \sum_{i,j=1}^n \kappa_{1,1'}(X_i,Y_j) - \sum_{i=1}^n \ell_{1,1'}(X_i,Y_i)\right\},$$

where $\ell_{1,1'}$ is an affine function such that $\kappa_{1,1'} - \ell_{1,1'} \in \mathbf{L}^2_{0,0}(\mu)$. This implies $(P\otimes Q)[\ell_{1,1'}] \overset{\text{affine}}{=} \mu[\ell_{1,1'}] = \mu[\kappa_{1,1'}] = \theta_{1,1'}$. By LLN, we know $\frac{1}{n}\sum_{i=1}^n \ell_{1,1'}(X_i,Y_i) = \theta_{1,1'} + o_p(1)$. Therefore,

$$T_n - \theta + \frac{\theta_{1,1'}}{n} = \frac{1}{n(n-1)}\left\{\sum_{i\neq j}[\kappa_{2,0}(X_i,X_j)+\kappa_{0,2}(Y_i,Y_j)] + \sum_{i,j=1}^n \kappa_{1,1'}(X_i,Y_j)\right\} + o_p(n^{-1}).$$

We then prove the limit law of the second order chaos. Recall from (C.14) that $\kappa_{1,1'} \in \mathbf{L}^2_{0,0}(P\otimes Q)$, so it holds that $\frac{1}{n(n-1)}\sum_{i=1}^n \kappa_{1,1'}(X_i,Y_i) = o_p(n^{-1})$ by LLN. Hence, we will ignore this term in the following derivation.

To begin with, we show the limiting distribution is well-defined. Since $\varsigma^2 = 0$ in Theorem 4.6, we know

$$(I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x) \overset{\text{a.s.}}{=} 0 \quad \text{and} \quad (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y) \overset{\text{a.s.}}{=} 0,$$

which implies

$$\begin{aligned}\tilde{\eta}(x,y) &:= \eta(x,y) - \theta - (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})(x) - (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0})(y) \\ &\overset{\text{a.s.}}{=} \eta(x,y) - \theta.\end{aligned}$$

Consequently, $(\eta - \theta)\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$ and thus it has expansion

$$(\eta - \theta)\xi = \sum_{k,l\geq 1} \gamma_{kl}(\alpha_k \otimes \beta_l), \quad \text{in } \mathbf{L}^2(P \otimes Q), \tag{C.61}$$

where $\sum_{k,l\geq 1} \gamma_{kl}^2 < \infty$. Recall from Assumption 4.2 that $0 \leq s_k \leq s_1 < 1$ for all $k \geq 1$, we have

$$\sum_{k,l\geq 1} \frac{\gamma_{kl}^2}{(1 - s_k^2)^2(1 - s_l^2)^2} \leq \sum_{k,l\geq 1} \frac{\gamma_{kl}^2}{(1 - s_1^2)^4} < \infty. \tag{C.62}$$

Let $\{U_k\}, \{V_l\}$ be independent sequences of *i.i.d.* standard normal random variables. We define

$$\begin{aligned} Z &:= \sum_{k,l\geq 1} \frac{\gamma_{kl}}{(1 - s_k^2)(1 - s_l^2)} \left\{U_k V_l + s_k s_l U_l V_k - s_l(U_k U_l - \mathbb{1}\{k = l\}) - s_k(V_k V_l - \mathbb{1}\{k = l\})\right\} \\ &= \sum_{k,l\geq 1} \frac{1}{(1 - s_k^2)(1 - s_l^2)} \big\{(\gamma_{kl} + s_k s_l \gamma_{lk}) U_k V_l \\ &\qquad - s_l \gamma_{kl}(U_k U_l - \mathbb{1}\{k = l\}) - s_k \gamma_{kl}(V_k V_l - \mathbb{1}\{k = l\})\big\}, \end{aligned}$$

where the sum converges in $\mathbf{L}^2$. We will show $Z_n := n\mathcal{Q}_n \to_d Z$ by using characteristic functions, *i.e.*, by showing that, for each $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(itZ_n)] \to \mathbb{E}[\exp(itZ)], \quad \text{as } n \to \infty.$$

The following proof is inspired by Serfling (1980a, Chapter 5.5.2).

Step 1. We expand Z_n on $\{\alpha_k \otimes \beta_l\}_{k,l\geq 0}$. For $k \geq 1$, we denote

$$\tilde{\alpha}_k := (I - \mathcal{A}^*\mathcal{A})^{-1}\alpha_k = (1-s_k^2)^{-1}\alpha_k \quad \text{and} \quad \tilde{\beta}_k := (I - \mathcal{A}\mathcal{A}^*)^{-1}\beta_k = (1-s_k^2)^{-1}\beta_k.$$

By Lemma C.12 it holds that $\mathcal{C}^{-1}(\alpha_k \otimes \beta_l) = \tilde{\alpha}_k \otimes \tilde{\beta}_l$, and then we get

$$\mathcal{C}^{-1}[(\eta-\theta)\xi] = \sum_{k,l\geq 1} \gamma_{kl}(\tilde{\alpha}_k \otimes \tilde{\beta}_l) = \sum_{k,l\geq 1} \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)}(\alpha_k \otimes \beta_l).$$

It follows that

$$\begin{aligned}
\kappa_{1,1'}(X_i, Y_j) &:= (I+\mathcal{B})\mathcal{C}^{-1}(\tilde{\eta}\xi)(X_i, Y_j) \\
&\overset{\text{a.s.}}{=} \sum_{k,l\geq 1} \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)}[\alpha_k(X_i)\beta_l(Y_j) + s_k s_l \alpha_l(X_i)\beta_k(Y_j)] \\
\kappa_{2,0}(X_i, X_j) &:= (I_P \otimes \mathcal{A}^*)\mathcal{C}^{-1}(\tilde{\eta}\xi)(X_i, X_j) \overset{\text{a.s.}}{=} \sum_{k,l\geq 1} \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)} s_l \alpha_k(X_i)\alpha_l(X_j) \\
\kappa_{0,2}(Y_i, Y_j) &:= (\mathcal{A} \otimes I_Q)\mathcal{C}^{-1}(\tilde{\eta}\xi)(Y_i, Y_j) \overset{\text{a.s.}}{=} \sum_{k,l\geq 1} \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)} s_k \beta_k(Y_i)\beta_l(Y_j).
\end{aligned}$$

Hence, Z_n admits the following expansion:

$$\begin{aligned}
Z_n &= \frac{1}{n-1}\sum_{i\neq j}\sum_{k,l\geq 1} \frac{\gamma_{kl}[\alpha_k(X_i)\beta_l(Y_j) + s_k s_l \alpha_l(X_i)\beta_k(Y_j) - s_l \alpha_k(X_i)\alpha_l(X_j) - s_k \beta_k(Y_i)\beta_l(Y_j)]}{(1-s_k^2)(1-s_l^2)} \\
&= \frac{1}{n-1}\sum_{i\neq j}\sum_{k,l\geq 1} \frac{(\gamma_{kl} + s_k s_l \gamma_{lk})\alpha_k(X_i)\beta_l(Y_j) - s_l \gamma_{kl}\alpha_k(X_i)\alpha_l(X_j) - s_k \gamma_{kl}\beta_k(Y_i)\beta_l(Y_j)}{(1-s_k^2)(1-s_l^2)}.
\end{aligned}$$

Step 2. We truncate the inner infinite sum. Fix an arbitrary integer $K > 0$. Let

$$\begin{aligned}
Z_n^K &:= \frac{1}{n-1}\sum_{i\neq j}\sum_{k,l=1}^{K} \frac{(\gamma_{kl} + s_k s_l \gamma_{lk})\alpha_k(X_i)\beta_l(Y_j) - s_l \gamma_{kl}\alpha_k(X_i)\alpha_l(X_j) - s_k \gamma_{kl}\beta_k(Y_i)\beta_l(Y_j)}{(1-s_k^2)(1-s_l^2)} \\
Z^K &:= \sum_{k,l=1}^{K} \frac{[(\gamma_{kl} + s_k s_l \gamma_{lk})U_k V_l - s_l \gamma_{kl}(U_k U_l - \mathbb{1}\{k=l\}) - s_k \gamma_{kl}(V_k V_l - \mathbb{1}\{k=l\})]}{(1-s_k^2)(1-s_l^2)}.
\end{aligned}$$

By triangle inequality, we have

$$\begin{aligned}
\left|\mathbb{E}[e^{itZ_n}] - \mathbb{E}[e^{itZ}]\right| &\leq \left|\mathbb{E}[e^{itZ_n}] - \mathbb{E}[e^{itZ_n^K}]\right| + \left|\mathbb{E}[e^{itZ_n^K}] - \mathbb{E}[e^{itZ^K}]\right| + \left|\mathbb{E}[e^{itZ^K}] - \mathbb{E}[e^{itZ}]\right| \\
&=: A + B + C \qquad \text{(C.63)}
\end{aligned}$$

Fix arbitrary $t \in \mathbb{R}$ and $\epsilon > 0$, it now suffices to show that $A, B, C \leq \epsilon$ for all sufficiently large n with an appropriate choice of K.

Step 3. We bound A and C. Using the inequality $|e^{iz} - 1| \leq |z|$, we get

$$A \leq \mathbb{E}\left|e^{itZ_n} - e^{itZ_n^K}\right| \leq |t|\,\mathbb{E}\left|Z_n - Z_n^K\right| \leq |t|\,[\mathbb{E}(Z_n - Z_n^K)^2]^{1/2}. \tag{C.64}$$

We rewrite $Z_n - Z_n^K$ as $\frac{1}{n-1}\sum_{i\neq j}[g_K^{\alpha\beta}(X_i, Y_j) - g_K^{\alpha\alpha}(X_i, X_j) - g_K^{\beta\beta}(Y_i, Y_j)]$, where

$$\begin{aligned} g_K^{\alpha\beta}(x, y) &:= \sum_{k,l>K} \frac{\gamma_{kl} + s_k s_l \gamma_{lk}}{(1-s_k^2)(1-s_l^2)}\alpha_k(x)\beta_l(y) \\ g_K^{\alpha\alpha}(x, x') &:= \sum_{k,l>K} \frac{\gamma_{kl} s_l}{(1-s_k^2)(1-s_l^2)}\alpha_k(x)\alpha_l(x') \\ g_K^{\beta\beta}(y, y') &:= \sum_{k,l>K} \frac{\gamma_{kl} s_k}{(1-s_k^2)(1-s_l^2)}\beta_k(y)\beta_l(y'). \end{aligned}$$

By the orthogonality of $\{\alpha_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq 0}$, we know $\mathbb{E}[\alpha_k(X_i)\beta_l(Y_j)\alpha_{k'}(X_i)\beta_{l'}(Y_j)] = 0$ for all $k, l \geq 1$ and $i \neq j$. This implies $g_K^{\alpha\beta}(X_i, Y_j)$ and $g_K^{\alpha\alpha}(X_i, X_j)$ are uncorrelated. Analogously, we have $g_K^{\alpha\beta}(X_i, Y_j)$, $g_K^{\alpha\alpha}(X_i, X_j)$ and $g_K^{\beta\beta}(Y_i, Y_j)$ are mutually uncorrelated for all $i \neq j$. As a result, $\mathbb{E}[(Z_n - Z_n^K)^2]$ reads

$$\mathbb{E}[(Z_n - Z_n^K)^2] \tag{C.65}$$

$$= \frac{1}{(n-1)^2}\,\mathbb{E}\left\{\left[\sum_{i\neq j} g_K^{\alpha\beta}(X_i, Y_j)\right]^2 + \left[\sum_{i\neq j} g_K^{\alpha\alpha}(X_i, X_j)\right]^2 + \left[\sum_{i\neq j} g_K^{\beta\beta}(Y_i, Y_j)\right]^2\right\}. \tag{C.66}$$

Notice that $\mathbb{E}[\alpha_k(X_1)\beta_l(Y_2) \mid X_1] = \mathbb{E}[\alpha_k(X_1)\beta_l(Y_2) \mid Y_2] = 0$ for all $k, l \geq 1$, then

$$\mathbb{E}[g_K^{\alpha\beta}(X_1, Y_2) \mid X_1] = \mathbb{E}[g_K^{\alpha\beta}(X_1, Y_2) \mid Y_2] = 0.$$

As a result,

$$\mathbb{E}\left[\sum_{i\neq j} g_K^{\alpha\beta}(X_i, Y_j)\right]^2 = N(n-1)\,\mathbb{E}[g_K^{\alpha\beta}(X_1, Y_2)^2] = N(n-1)\sum_{k,l>K}\left[\frac{\gamma_{kl} + s_k s_l \gamma_{lk}}{(1-s_k^2)(1-s_l^2)}\right]^2.$$

Let $\delta > 0$ be such that $|t|\,\delta < \epsilon$. It then follows from (C.62) that, for all sufficiently large K, we have

$$\frac{1}{(n-1)^2}\,\mathbb{E}\left[\sum_{i\neq j} g_K^{\alpha\beta}(X_i, Y_j)\right]^2 \leq \frac{n}{n-1}\sum_{k,l>K}\left[\frac{\gamma_{kl} + s_k s_l \gamma_{lk}}{(1-s_k^2)(1-s_l^2)}\right]^2 \leq \frac{n}{6(n-1)}\delta^2.$$

The same bound for the rest of the two terms in (C.65) can be shown using similar arguments. Therefore, by (C.64),

$$A \leq |t| \, [\mathbb{E}(Z_n - Z_n^K)^2]^{1/2} \leq \sqrt{\frac{n}{2(n-1)}} \, |t| \, \delta < \epsilon, \qquad \text{for all } n \geq 2.$$

Repeating the above argument for Z^K and Z gives $C < \epsilon$ for all $n \geq 2$.

Step 4. We bound B by proving $Z_n^K \to_d Z^K$ as $n \to \infty$. Consider $W_n := (W_\alpha^\top, W_\beta^\top)$ with

$$W_\alpha := \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \alpha_k(X_i) \right)_{k=1}^K \quad \text{and} \quad W_\beta := \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \beta_k(Y_i) \right)_{k=1}^K .$$

According to the multivariate CLT (Billingsley, 1995, Section 29), it holds $W_n \to_d \mathcal{N}_{2K}(0, I_{2K})$, where the covariance matrix I_{2K} follows from the orthonormality of $\{\alpha_k\}_{k\geq 1}$ and $\{\beta_k\}_{k\geq 1}$. We then rewrite Z_n^K as a quadratic form of W_n. Notice that

$$\begin{aligned}
&\frac{1}{n} \sum_{i\neq j} \sum_{k,l=1}^K \frac{1}{(1-s_k^2)(1-s_l^2)} (\gamma_{kl} + s_k s_l \gamma_{lk}) \alpha_k(X_i) \beta_l(Y_j) \\
&= \frac{1}{n} \sum_{k,l=1}^K \frac{(\gamma_{kl} + s_k s_l \gamma_{lk})}{(1-s_k^2)(1-s_l^2)} \left\{ \left[\sum_{i=1}^n \alpha_k(X_i) \right] \left[\sum_{i=1}^n \beta_l(Y_i) \right] - \sum_{i=1}^n \alpha_k(X_i) \beta_l(Y_i) \right\} \\
&= 2 W_\alpha^\top \Sigma^{\alpha\beta} W_\beta - \sum_{k,l=1}^K \frac{(\gamma_{kl} + s_k s_l \gamma_{lk})}{(1-s_k^2)(1-s_l^2)} \frac{1}{n} \sum_{i=1}^n \alpha_k(X_i) \beta_l(Y_i),
\end{aligned}$$

where $\Sigma_{kl}^{\alpha\beta} = \frac{(\gamma_{kl} + s_k s_l \gamma_{lk})}{2(1-s_k^2)(1-s_l^2)}$ is the (k,l)-element in the matrix Σ. Similarly, it holds that

$$\begin{aligned}
&\frac{1}{n} \sum_{i\neq j} \sum_{k,l=1}^K \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)} s_l \alpha_k(X_i) \alpha_l(X_j) \\
&= W_\alpha^\top \Sigma^{\alpha\alpha} W_\alpha - \sum_{k,l=1}^K \frac{\gamma_{kl} s_l}{(1-s_k^2)(1-s_l^2)} \frac{1}{n} \sum_{i=1}^n \alpha_k(X_i) \alpha_l(X_i) \\
&\frac{1}{n} \sum_{i\neq j} \sum_{k,l=1}^K \frac{\gamma_{kl}}{(1-s_k^2)(1-s_l^2)} s_k \beta_k(Y_i) \beta_l(Y_j) \\
&= W_\beta^\top \Sigma^{\beta\beta} W_\beta - \sum_{k,l=1}^K \frac{\gamma_{kl} s_k}{(1-s_k^2)(1-s_l^2)} \frac{1}{n} \sum_{i=1}^n \beta_k(Y_i) \beta_l(Y_i),
\end{aligned}$$

where $\Sigma^{\alpha\alpha}_{kl} = \frac{\gamma_{kl} s_l}{(1-s_k^2)(1-s_l^2)}$ and $\Sigma^{\beta\beta}_{kl} = \frac{\gamma_{kl} s_k}{(1-s_k^2)(1-s_l^2)}$. Hence,

$$Z_n^K := \frac{n}{n-1} W_n^\top \begin{pmatrix} -\Sigma^{\alpha\alpha} & \Sigma^{\alpha\beta} \\ [\Sigma^{\alpha\beta}]^\top & -\Sigma^{\beta\beta} \end{pmatrix} W_n - \frac{n}{n-1} \sum_{k,l=1}^{K} \frac{1}{(1-s_k^2)(1-s_l^2)} \frac{1}{n} \sum_{i=1}^{n} \big[$$

$$(\gamma_{kl} + s_k s_l \gamma_{lk}) \alpha_k(X_i) \beta_l(Y_i) - s_l \gamma_{kl} \alpha_k(X_i) \alpha_l(X_i) - s_k \gamma_{kl} \beta_k(Y_i) \beta_l(Y_i) \big].$$

Since $\mathbb{E}[\alpha_k(X_i)\beta_l(Y_i)] = 0$ and $\mathbb{E}[\alpha_k(X_i)\alpha_l(X_i)] = \mathbb{E}[\beta_k(Y_i)\beta_l(Y_i)] = \mathbf{1}\{k = l\}$ for all $k, l \geq 1$ and $i \in [n]$, we know from LLN that

$$\frac{1}{n} \sum_{i=1}^{n} [(\gamma_{kl} + s_k s_l \gamma_{lk}) \alpha_k(X_i) \beta_l(Y_i) - s_l \alpha_k(X_i) \alpha_l(X_i) - s_k \beta_k(Y_i) \beta_l(Y_i)]$$

$$\to_p -s_l \mathbf{1}\{k = l\} - s_k \mathbf{1}\{k = l\}.$$

By Slutsky's lemma, it holds $Z_n^K \to_d Z^K$, and thus we have $B < \epsilon$ for all sufficiently large N. Now, by (C.63), we get $\left|\mathbb{E}[e^{itZ_n}] - \mathbb{E}[e^{itZ}]\right| \leq 3\epsilon$ for all sufficiently large n. Since ϵ is arbitrary, this completes the proof. □

Appendix D

APPENDIX TO CHAPTER 5

D.1 Properties of Entropy-Regularized Optimal Transport Independence Criterion

In this section, we prove the properties of ETIC discussed in Section 5.4. For the sake of generality, we state the problem for general notations P and Q while keeping in mind that $P, Q \in \{\mu_{XY}, \mu_X \otimes \mu_Y\}$ in our case. Let $P \in \mathcal{M}_1(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ and P_X and P_Y be the marginals on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Define Q, Q_X, and Q_Y similarly. We are interested in the EOT cost between P and Q under the cost function c:

$$S_\varepsilon(P, Q) := \inf_{\gamma \in \Pi(P,Q)} \left[\int c d\gamma + \varepsilon \, \mathrm{KL}(\gamma \| P \otimes Q) \right]. \tag{D.1}$$

When $\varepsilon = 0$, $S_0(P, Q)$ is the optimal transport cost between P and Q. When $\varepsilon > 0$, it admits a dual representation:

$$S_\varepsilon(P, Q) := \sup_{f,g \in \mathcal{C}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})} \left[\int f dP + \int g dQ + \varepsilon - \varepsilon \int e^{\frac{1}{\varepsilon}[f(z)+g(z')-c(z,z')]} dP(z) dQ(z') \right]. \tag{D.2}$$

The Schrödinger bridge potentials $(f_\varepsilon, g_\varepsilon)$ satisfy the optimality conditions:

$$\begin{aligned} \int e^{\frac{1}{\varepsilon}[f_\varepsilon(z)+g_\varepsilon(z)-c(z,z')]} dQ(z') &\overset{\text{a.s.}}{=} 1 \\ \int e^{\frac{1}{\varepsilon}[f_\varepsilon(z)+g_\varepsilon(z)-c(z,z')]} dP(z) &\overset{\text{a.s.}}{=} 1. \end{aligned} \tag{D.3}$$

We first prove the validity of ETIC as a dependence measure as stated in Proposition 5.3.

Proof of Proposition 5.3. Due to Blanchard et al. (2011, Lemma 5.2), the Gibbs kernel

$$k_\varepsilon(z, z') := e^{-c(z,z')/\varepsilon} = k_1(x, x') k_2(y, y')$$

is universal since both k_x and k_y are. It is also clear that k_ε is positive since both k_x and k_y are. Consequently, the Sinkhorn divergence $\bar{S}_\varepsilon$ defines a semi-metric on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ according to Feydy et al. (2019, Theorem 1). Hence, if $\mu_{XY}, \mu_X \otimes \mu_Y \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, then $T_\varepsilon(X, Y) := \bar{S}_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) = 0$ iff $\mu_{XY} = \mu_X \otimes \mu_Y$. □

Next, we analyze the computational complexity of the Tensor Sinkhorn algorithm for additive cost functions, i.e.,

$$c(z, z') := c_1(x, x') + c_2(y, y'), \tag{D.4}$$

where $z = (x, y)$ and $z' = (x', y')$.

Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$ be two sets of atoms. Note that the two sets are assumed to be of the same size for convenience. Let A and B be two probability measures on $\{x_i\}_{i=1}^n \times \{y_j\}_{j=1}^n$. For convenience, both A and B are represented as a matrix, i.e., $A_{ij} = A(x_i, y_j)$. For instance, if we choose $A = \hat{\mu}_{XY}$ and $B = \hat{\mu}_X \otimes \hat{\mu}_Y$, then, in its matrix form, $A = I_n/n$ and $B = \mathbf{1}_{n\times n}/n^2$. Denote C_1 and C_2 as the cost matrices of $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$, respectively. Define Gibbs matrices $K_1 := e^{-C_1/\varepsilon}$ and $K_2 := e^{-C_2/\varepsilon}$, where the exponential function is applied elementwisely. Let $K := K_2 \otimes K_1 \in \mathbb{R}^{n^2 \times n^2}$ be the Gibbs matrix associated with the cost matrix on the pairs $\{(x_1, y_1), (x_2, y_1), \ldots, (x_n, y_n)\}$, where $\otimes$ is the Kronecker product.

Proof of Proposition 5.4. Let $a := \mathrm{Vect}(A) \in \mathbb{R}^{n^2}$ and $b := \mathrm{Vect}(B) \in \mathbb{R}^{n^2}$ be the probability vectors corresponding to A and B, respectively. Denote $u := \mathrm{Vect}(U) \in \mathbb{R}^{n^2}$ and $v := \mathrm{Vect}(V) \in \mathbb{R}^{n^2}$. The Sinkhorn algorithm to solve $S_\varepsilon(a, b)$ has the following two update steps:

$$u = a \oslash Kv \quad \text{and} \quad v = b \oslash K^\top u.$$

By the identity $\mathrm{Vect}(MNL) = (L^\top \otimes M)\,\mathrm{Vect}(N)$ for matrices M, N, and L of compatible dimensions, we obtain

$$\mathrm{Vect}(K_1 V K_2^\top) = (K_2 \otimes K_1)\,\mathrm{Vect}(V) = Kv.$$

Thus, the update $U = A \oslash (K_1 V K_2^\top)$ is equivalent to $u = a \oslash Kv$. Similarly, the updated $V = B \oslash (K_1^\top U K_2)$ is equivalent to $v = b \oslash K^\top u$. Due to Dvurechensky et al.

(2018, Theorem 1), the Tensor Sinkhorn algorithm therefore outputs an τ-accurate estimate in $O(\log(\kappa_1\kappa_2\kappa_3)/\tau)$ iterations. Since each iteration costs $O(n^3)$ time, it has overall time complexity $O(n^3\log(\kappa_1\kappa_2\kappa_3)/\tau)$. □

Remark D.1. *A direct application of the Sinkhorn algorithm leads to $O(n^4\log(\kappa_1\kappa_2\kappa_3)/\tau)$ time complexity, which is n times slower than the Tensor Sinkhorn algorithm.*

We then characterize the convergence of the Tensor Sinkhorn algorithm with the random feature approximation as presented in Proposition 5.5.

Proof of Proposition 5.5. The proof is heavily inspired by Scetbon and Cuturi (2020, Proof of Theorem 3.1). In consideration of the space, we only present the part that is significantly different from theirs, i.e., a counterpart of Scetbon and Cuturi (2020, Proposition 3.1). This proposition gives a uniform tail bound for the ratio between the approximated kernel and the original kernel. In our case, we are approximating the kernel $K := K_2 \otimes K_1$ by $K_{\boldsymbol{u},\boldsymbol{v}} := K_{2,\boldsymbol{v}} \otimes K_{1,\boldsymbol{u}}$. Hence, it suffices to bound

$$\sup_{x,x'\in\{x_i\}_{i=1}^n, y,y'\in\{y_i\}_{i=1}^n} \left| \frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')} - 1 \right|.$$

Note that

$$\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} = \frac{1}{p}\sum_{k=1}^{p} \frac{\varphi(x,u_k)^\top \varphi(x',u_k)}{k_1(x,x')}$$

is a sum of nonnegative i.i.d. random variables with mean 1. Due to Assumption 5.1, they are also bounded. It follows from the Hoeffding inequality that

$$\mathbb{P}\left(\left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| \ge t \right) \le 2\exp\left(-\frac{pt^2}{C^2} \right).$$

The same inequality holds for the ratio $k_{2,\boldsymbol{v}}(y,y')/k_2(y,y')$. Since

$$\begin{aligned} &\left| \frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')} - 1 \right| \\ &\le \left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| \left| \frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')} - 1 \right| + \left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| + \left| \frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')} - 1 \right|, \end{aligned}$$

it follows that

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\le t^2+2t\right)\\
&\ge \mathbb{P}\left(\left\{\left|\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')}-1\right|\le t\right\}\bigcap\left\{\left|\frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')}-1\right|\le t\right\}\right)\\
&= \mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')}-1\right|\le t\right)\mathbb{P}\left(\left|\frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')}-1\right|\le t\right)\\
&\ge 1-4\exp\left(-\frac{pt^2}{C^2}\right).
\end{aligned}
$$

Equivalently,

$$
\mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\ge t\right)\le 4\exp\left(-\frac{p(\sqrt{t+1}-1)^2}{C^2}\right).
$$

A uniform bound yields

$$
\mathbb{P}\left(\sup_{x,x'\in\{x_i\}_{i=1}^n,y,y'\in\{y_i\}_{i=1}^n}\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\ge t\right)\le 4n^4\exp\left(-\frac{p(\sqrt{t+1}-1)^2}{C^2}\right).
$$

□

Remark D.2. *Let $\hat{S}_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)$ be the cost computed from Algorithm 3. Following Dvurechensky et al. (2018, Theorem 1), we can get that*

$$
\left|\hat{S}_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)-S_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)\right|\le\tau
$$

in $O\left(pn^2\log(\kappa_1\kappa_2\kappa_3)/\tau\right)$ arithmetic operations, where $\kappa_1 := \max_{i,i'} k_{1,\boldsymbol{u}}^{-1}(x_i,x_{i'})$, $\kappa_2 := \max_{j,j'} k_{2,\boldsymbol{v}}^{-1}(y_j,y_{j'})$, and $\kappa_3 := \max_{i,j}\{a_{ij}^{-1},b_{ij}^{-1}\}$.

Finally, we derive the limit of ETIC as $\varepsilon\to 0$ and $\varepsilon\to\infty$.

Proposition D.1. *Let c be a continuous cost function. If either c is bounded or P and Q have compact support, it holds that*

$$
T_\varepsilon(X,Y)\to\begin{cases}0 & \text{if } c=c_1\oplus c_2\\ -\frac{1}{2}\,\mathrm{HSIC}_{c_1,c_2}(X,Y) & \text{if } c=c_1\otimes c_2,\end{cases}\quad \text{as } \varepsilon\to\infty. \tag{D.5}
$$

Moreover, if both P and Q are densities (or discrete measures), then

$$
T_\varepsilon(X,Y)\to S_0(\mu_{XY},\mu_X\otimes\mu_Y),\quad \text{as } \varepsilon\to 0. \tag{D.6}
$$

Proof. To show (D.5), we claim that, for all $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$,

$$S_0(P,Q) \le S_\varepsilon(P,Q) \le (P \otimes Q)[c], \tag{D.7}$$

and

$$\lim_{\varepsilon\to\infty} S_\varepsilon(P,Q) = (P \otimes Q)[c]. \tag{D.8}$$

In fact, for any $\varepsilon_1 < \varepsilon_2$, we have

$$\int c d\gamma + \varepsilon_1 \operatorname{KL}(\gamma \| P \otimes Q) \le \int c d\gamma + \varepsilon_2 \operatorname{KL}(\gamma \| P \otimes Q), \quad \text{for all } \gamma \in \Pi(P,Q).$$

This yields that

$$S_{\varepsilon_1}(P,Q) \le S_{\varepsilon_2}(P,Q), \quad \text{for all } \varepsilon_1 \le \varepsilon_2,$$

and thus (D.7) follows.

We then study the limit of S_ε as $\varepsilon \to \infty$. By the assumption that c is bounded or P and Q have compact support, there exists $M > 0$ such that $\sup_{\gamma\in\Pi(P,Q)} \int c d\gamma \le M < \infty$. As a result,

$$\sup_{\gamma\in\Pi(P,Q)} \left| \frac{1}{\varepsilon} \int c d\gamma + \operatorname{KL}(\gamma \| P \otimes Q) - \operatorname{KL}(\gamma \| P \otimes Q) \right| \le \frac{M}{\varepsilon},$$

which implies that

$$\inf_{\gamma\in\Pi(P,Q)} \left[\frac{1}{\varepsilon} \int c d\gamma + \operatorname{KL}(\gamma \| P \otimes Q) \right] \to \inf_{\gamma\in\Pi(P,Q)} \operatorname{KL}(\gamma \| P \otimes Q) = 0, \quad \text{as } \varepsilon \to \infty.$$

By the strict convexity of KL, the problem on the LHS has a unique minimizer γ_ε and the problem on the RHS has a unique minimizer $\gamma_* = P \otimes Q$. Now, by the tightness of $\Pi(P,Q)$ (e.g., (Santambrogio, 2015, Theorem. 1.7)), every sequence of $\{\gamma_\varepsilon\}$ has a weakly converging subsequence whose limit must be γ_*. Therefore, the claim (D.8) holds true.

Let $c = c_1 \oplus c_2$. According to (D.8), we have

$$\lim_{\varepsilon\to\infty} S_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) = (\mu_{XY} \otimes \mu_X \otimes \mu_Y)[c] = (\mu_X \otimes \mu_X)[c_1] + (\mu_Y \otimes \mu_Y)[c_2].$$

Similarly, it holds that

$$\lim_{\varepsilon\to\infty} S_\varepsilon(\mu_{XY}, \mu_{XY}) = (\mu_X \otimes \mu_X)[c_1] + (\mu_Y \otimes \mu_Y)[c_2]$$
$$\lim_{\varepsilon\to\infty} S_\varepsilon(\mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y) = (\mu_X \otimes \mu_X)[c_1] + (\mu_Y \otimes \mu_Y)[c_2].$$

Consequently, $\lim_{\varepsilon\to\infty} T_\varepsilon(X, Y) = 0$. An analogous argument implies that, when $c = c_1 \otimes c_2$

$$\begin{aligned}\lim_{\varepsilon\to\infty} T_\varepsilon(X, Y) &= \mathbb{E}_{\mu_{XY}}\left[\mathbb{E}_{\mu_X}[c_1(X, X') \mid X]\, \mathbb{E}_{\mu_Y}[c_2(Y, Y') \mid Y]\right] \\ &- \frac{1}{2}\,\mathbb{E}_{\mu_{XY}^2}[c_1(X, X')c_2(Y, Y')] - \frac{1}{2}\,\mathbb{E}_{(\mu_X\otimes\mu_Y)^2}[c_1(X, X')c_2(Y, Y')] = -\frac{1}{2}\,\mathrm{HSIC}_{c_1,c_2}(X, Y).\end{aligned}$$

Note that

$$\lim_{\varepsilon\to 0} S_\varepsilon(P, Q) = S_0(P, Q)$$

when both P and Q are densities (Léonard, 2012) and when both of them are discrete measures (Peyré and Cuturi, 2019, Proposition 4.1). The statement (D.6) follows immediately from the fact that $S_0(P, P) = 0$ for all P. □

D.2 Consistency of the Test Statistic

In this section, we prove the main results in Section 5.5. For the sake of generality, we start by considering the formulation in (D.1). We focus on the weighted quadratic cost function

$$c(z, z') := w_1 \left\|x - x'\right\|^2 + w_2 \left\|y - y'\right\|^2,$$

where $z = (x, y)$, $z' = (x', y')$ and $w_1, w_2 \in \mathbb{R}_+$. Denote $w := \max\{w_1, w_2\}$. Due to Lemma D.17, we assume, w.l.o.g., that $\varepsilon = 1$ and write $S(P, Q) := S_1(P, Q)$.

D.2.1 Smoothness Properties of the Schrödinger Potentials

We start by deriving some smoothness properties of the Schrödinger potentials. Our proofs are deeply inspired by Mena and Weed (2019). Our results generalize theirs to weighted quadratic cost functions.

Assumption D.1. *We assume that P_X, P_Y, Q_X, and Q_Y are all subG(σ^2).*

Proposition D.2. *Under Assumption D.1. there exist smooth Schrödinger potentials (f, g) for $S(P, Q)$ such that the optimality conditions (D.3) hold for all $z, z' \in \mathbb{R}^d$. Moreover, we have*

$$\begin{aligned} f(z) &\geq -d\sigma^2 \left[2w_1 + 2w_2 + 4w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2 + 4w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2\right] - 1 \\ f(z) &\leq w_1(\|x\| + \sqrt{2d_1}\sigma)^2 + w_2(\|y\| + \sqrt{2d_2}\sigma)^2, \end{aligned}$$

and for g similarly.

Proof. Let (f_0, g_0) be a pair of Schrödinger potentials. Since $(f_0 + C, g_0 - C)$ is also a pair of Schrödinger potentials for any constant $C \in \mathbb{R}$, we assume, w.l.o.g., that $P[f_0] = Q[g_0] = \frac{1}{2}S(P, Q) \geq 0$. Define

$$f(z) := -\log \int e^{g_0(z') - c(z,z')} dQ(z') \quad \text{and} \quad g(z') := -\log \int e^{f(z) - c(z,z')} dP(z). \tag{D.9}$$

We claim that the pair (f, g) satisfies the requirements.

Since (f_0, g_0) is a pair of Schrödinger potentials, it holds that

$$g_0(z') \overset{\text{a.s.}}{=} -\log \int e^{f_0(z) - c(z,z')} dP(z) \leq -P[f_0] + w_1 \mathbb{E}_{P_X}[\|X - x'\|^2] + w_2 \mathbb{E}_{P_Y}[\|Y - y'\|^2],$$

by Jensen's inequality. Note that $P[f_0] \geq 0$ and, by Lemma D.11, $\mathbb{E}_{P_X}[\|X\|^2] \leq 2d_1\sigma^2$. It follows that

$$g_0(z') - c(z, z') \leq w_1 \left[2d_1\sigma^2 + 2\|x'\|(\sqrt{2d_1}\sigma + \|x\|)\right] + w_2 \left[2d_2\sigma^2 + 2\|y'\|(\sqrt{2d_2}\sigma + \|y\|)\right],$$

and thus

$$\begin{aligned} &\int e^{g_0(z') - c(z,z')} dQ(z') \\ &\leq e^{2(w_1d_1 + w_2d_2)\sigma^2} \left[\int e^{4w_1\|x'\|(\sqrt{2d_1}\sigma + \|x\|)} dQ_X(x') \int e^{4w_2\|y'\|(\sqrt{2d_2}\sigma + \|y\|)} dQ_Y(y')\right]^{1/2} \\ &\leq 2e^{2(w_1d_1 + w_2d_2)\sigma^2} e^{4d_1\sigma^2 w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2 + 4d_2\sigma^2 w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2} < \infty, \quad \text{by Lemma D.11.} \end{aligned}$$

Hence, $f(z)$ is well-defined for all $z \in \mathbb{R}^d$. Moreover, we have the lower bound

$$\begin{aligned} f(z) &\geq -d_1\sigma^2 \left[2w_1 + 4w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2\right] - d_2\sigma^2 \left[2w_2 + 4w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2\right] - 1 \\ &\geq -d\sigma^2 \left[4w + 4w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2 + 4w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2\right] - 1 \end{aligned}$$

For the upper bound, by Jensen's inequality, it holds that

$$\begin{aligned} f(z) &\leq -Q[g_0] + w_1\, \mathbb{E}_{Q_X} \|x - X'\|^2 + w_2\, \mathbb{E}_{Q_Y} \|y - Y'\|^2 \\ &\leq w_1(\|x\| + \sqrt{2d_1}\sigma)^2 + w_2(\|y\| + \sqrt{2d_2}\sigma)^2. \end{aligned}$$

Similar arguments prove the claim for g. Now, it remains to show that (f, g) satisfies the optimality conditions (D.3) for all $z, z \in \mathbb{R}^d$. By definition, it is clear that

$$\int e^{f(z)+g(z')-c(z,z')} dP(z) = 1 \quad \text{and} \quad \int e^{f(z)+g_0(z')-c(z,z')} dQ(z') = 1, \quad \forall z, z' \in \mathbb{R}^d.$$

Since (f_0, g_0) is a pair of Schrödinger potentials, we also have

$$\int e^{f_0(z)+g_0(z')-c(z,z')} dP(z) dQ(z') = 1.$$

Consequently, by Jensen's inequality

$$\begin{aligned} &\int (f - f_0) dP + \int (g - g_0) dQ \\ &\geq -\log \int e^{f_0 - f} dP - \log \int e^{g_0 - g} dQ \\ &= -\log \int e^{f_0(z)+g_0(z')-c(z,z')} dP(z) dQ(z') - \log \int e^{f(z)+g_0(z')-c(z,z')} dP(z) dQ(z') \\ &= 0. \end{aligned}$$

Since both (f_0, g_0) and (f, g) are Schrödinger potentials, the above equality holds true. This implies that $\int (g_0 - g) dQ = \log \int e^{g_0 - g} dQ$, and thus $g = g_0 + C$ Q-almost surely by the strict concavity of log. Therefore, we have

$$\int e^{f(z)+g(z')-c(z,z')} dQ(z') = e^C \int e^{f(z)+g_0(z')-c(z,z')} dQ(z') = e^C, \quad \forall z, z' \in \mathbb{R}^d.$$

Taking integrals with respect to P implies that $C = 0$, which completes the proof. □

The next proposition shows that there exist Schrödinger potentials satisfying Hölder-type conditions.

Definition D.1. *For any $\sigma \in \mathbb{R}_+$, $d \in \mathbb{N}_+$, and $w = (w_1, w_2) \in \mathbb{R}_+^2$, let $\mathcal{F}_\sigma := \mathcal{F}_{\sigma,d,w}$ be the set of smooth functions such that, for any $k \in \mathbb{N}_+$ and any multi-index α with $|\alpha| = k$,*

$$\left|D^\alpha \left(f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2\right)\right| \leq C_{k,d,w} \begin{cases} (1+\sigma^4) & \text{if } k = 0 \\ \sigma^k(1+\sigma)^k & \text{otherwise,} \end{cases} \tag{D.10}$$

if $\|z\| \leq \sqrt{d}\sigma$, and

$$\left|D^\alpha \left(f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2\right)\right| \leq C_{k,d,w} \begin{cases} [1+(1+\sigma^2)\|x\|^2] & \text{if } k = 0 \\ \sigma^k(\sqrt{\sigma \|x\|} + \sigma \|x\|)^k & \text{otherwise,} \end{cases} \tag{D.11}$$

if $\|z\| > \sqrt{d}\sigma$, where $C_{k,d,w}$ is a constant depending on k, d, and w.

Proposition D.3. *Under Assumption D.1, there exist Schrödinger potentials (f, g) such that the optimality conditions (D.3) hold for all $z, z' \in \mathbb{R}^d$ and $f, g \in \mathcal{F}_\sigma$.*

Proof. Let (f, g) be a pair of Schrödinger potentials satisfying the requirements in Proposition D.2. Denote $\bar{f}(x,y) := f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2$. Note that

$$\begin{aligned} \bar{f}(z) = -\log e^{-\bar{f}(x,y)} &= -\log \int e^{w_1\|x\|^2 + w_2\|y\|^2 + g(z') - c(z,z')} dQ(z') \\ &= -\log \int e^{g(z') - w_1\|x'\|^2 - w_2\|y'\|^2 + 2w_1\langle x,x'\rangle + 2w_2\langle y,y'\rangle} dQ(z'). \end{aligned}$$

The desired inequalities for $k = 0$ follow directly from Proposition D.2. We focus on $k > 0$. According to the multivariate Faá di Bruno formula (Constantine and Savits, 1996), we have

$$D^\alpha \bar{f}(z) = \sum_{\lambda_1 + \cdots + \lambda_k = \alpha} C_{\alpha,\lambda_1,\ldots,\lambda_k} \prod_{i=1}^k M_{\lambda_i},$$

where

$$M_\lambda = \frac{\int (\tilde{z}')^\lambda \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}. \tag{D.12}$$

Here $\tilde{z}' = (2w_1x'; 2w_2y')$ and $z^\lambda = \prod_{i=1}^d z_i^{\lambda_i}$. By Lemma D.4 below, it holds that

$$\left|D^\alpha \bar{f}(z)\right| \le C_{k,d,w}\begin{cases}\sigma^k(1+\sigma^k) & \text{if } \|z\| \le \sqrt{d}\sigma\\ \sigma^k(\sigma\|z\| + \sqrt{\sigma\|z\|})^k & \text{if } \|z\| > \sqrt{d}\sigma,\end{cases}$$

which proves the claim. □

Lemma D.4. *Recall* M_λ *in* (D.12). *Under Assumption* D.1, *for* $|\lambda| > 0$, *we have*

$$|M_\lambda| \le C_{|\lambda|,d,w}\begin{cases}\sigma^{|\lambda|}(\sigma+\sigma^2)^{|\lambda|} & \textit{if } \|z\| \le \sqrt{d}\sigma\\ \sigma^{|\lambda|}(\sigma\|z\| + \sqrt{\sigma\|z\|})^{|\lambda|} & \textit{if } \|z\| > \sqrt{d}\sigma\end{cases}.$$

Proof. We first bound the denominator. By the optimality conditions (D.3), it holds that

$$\left(\int \exp\left\{g(z') - w_1\|x'\|^2 - w_2\|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')\right)^{-1}$$
$$= e^{f(x,y) - w_1\|x\|^2 - w_2\|y\|^2} \le e^{w_1(2d_1\sigma^2 + 2\sqrt{2d_1}\sigma\|x\|) + w_2(2d_2\sigma^2 + 2\sqrt{2d_2}\sigma\|y\|)},$$

where the last inequality follows from Proposition D.2. To bound the numerator, we use the truncation technique. Let $A := \{(x', y') : \|2w_1x'\| \le K, \|2w_2y'\| \le K\}$ for some constant K to be determined later. On the set A, it is clear that $(\tilde{z}')^\lambda \le \|\tilde{z}'\|^{|\lambda|} \le K^{|\lambda|}$, and thus

$$\frac{\int_A (\tilde{z}')^\lambda \exp\left\{g(z') - w_1\|x'\|^2 - w_2\|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1\|x'\|^2 - w_2\|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')} \le K^{|\lambda|}.$$

On the set A^c, we proceed as follows. According to Proposition D.2, we have

$$e^{g(x',y') - w_1\|x'\|^2 - w_2\|y'\|^2} \le e^{w_1(2d_1\sigma^2 + 2\sqrt{2d_1}\sigma\|x'\|) + w_2(2d_2\sigma^2 + 2\sqrt{2d_2}\sigma\|y'\|)},$$

which yields

$$\int_{A^c} (\tilde{z}')^\lambda \exp\left\{g(z') - w_1\|x'\|^2 - w_2\|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')$$
$$\le e^{2(w_1d_1 + w_2d_2)\sigma^2}\left[\int_{A^c} (\tilde{z}')^{2\lambda} dQ(z') \int_{A^c} e^{2w_1\|x'\|(\|x\| + \sqrt{2d_1}\sigma) + 2w_2\|y'\|(\|y\| + \sqrt{2d_2}\sigma)} dQ(z')\right]^{1/2}.$$

For any $z' \in A^c$, we have either $\|2w_1x'\| > K$ or $\|2w_2y'\| > K$. If the former is true, then

$$\int_{A^c} (\tilde{z}')^{2\lambda} dQ(z') \le \int_{A^c} e^{-\frac{K^2}{16w_1^2d_1\sigma^2}} e^{\frac{\|2w_1x'\|^2}{16w_1^2d_1\sigma^2}} (\tilde{z}')^{2\lambda} dQ(z') \le C_{|\lambda|,d,w} e^{-\frac{K^2}{16w^2d\sigma^2}} \sigma^{2|\lambda|},$$

where $w = \max\{w_1, w_2\}$. The same bound holds if the latter is true. Furthermore, by the Cauchy-Schwartz inequality and Lemma D.11 in Appendix D.4, we have

$$\int_{A^c} e^{2w_1\|x'\|(\|x\|+\sqrt{2d_1}\sigma)+2w_2\|y'\|(\|y\|+\sqrt{2d_2}\sigma)} dQ(z') \le e^{4w_1^2d_1\sigma^2(\|x\|+\sqrt{2d_1}\sigma)^2+4w_2^2d_2\sigma^2(\|y\|+\sqrt{2d_2}\sigma)^2}$$

Putting all together, we get

$$\begin{aligned}
&\frac{\int_{A^c} (\tilde{z}')^{\lambda} \exp\left\{g(z') - w_1\left\|x'\right\|^2 - w_2\left\|y'\right\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1\left\|x'\right\|^2 - w_2\left\|y'\right\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')} \\
&\le C_{|\lambda|,d,w} e^{-\frac{K^2}{32w^2d\sigma^2}} e^{2w_1^2d_1\sigma^2(\|x\|+\sqrt{2d_1}\sigma)^2+2w_2^2d_2\sigma^2(\|y\|+\sqrt{2d_2}\sigma)^2} \sigma^{|\lambda|} \\
&\le C_{|\lambda|,d,w} e^{-\frac{K^2}{32w^2d\sigma^2}} e^{2w^2d\sigma^2[(\|x\|+\sqrt{2d}\sigma)^2+(\|y\|+\sqrt{2d}\sigma)^2]} \sigma^{|\lambda|}
\end{aligned}$$

When $\|z\| \le \sqrt{d}\sigma$, it holds that $\|x\| \le \sqrt{2d}\sigma$ and $\|y\| \le \sqrt{2d}\sigma$. Hence, if we choose $K^2 = C_{|\lambda|,d,w}(\sigma^4 + \sigma^6)$ for some sufficiently large constant $C_{|\lambda|,d,w}$, then we have

$$|M_\lambda| \le C_{|\lambda|,d,w}\sigma^{|\lambda|}(\sigma + \sigma^2)^{|\lambda|}.$$

When $\|z\| > \sqrt{d}\sigma$, if we choose $K^2 = C_{|\lambda|,d,w}(\sigma^4\left\|z\right\|^2 + \sigma^3\left\|z\right\|)$, then we have

$$|M_\lambda| \le C_{|\lambda|,d,w}\sigma^{|\lambda|}\left(\sigma\left\|z\right\| + \sqrt{\sigma\left\|z\right\|}\right)^{|\lambda|}.$$

□

When P and Q have bounded support, we can further show that the Schrödinger potentials can be chosen to be bounded.

Proposition D.5. *Assume that P and Q are supported on a bounded domain of radius D. Then there exist Schrödinger potentials (f, g) such that 1) the optimality conditions* (D.3) *hold for all $x, y \in \mathbb{R}^d$ and 2) $\|f\|_\infty \le 8wD^2$ and $\|g\|_\infty \le 8wD^2$.*

Proof. Let (f, g) the Schrödinger potentials defined in (D.9). By the proof of Proposition D.2, they satisfy (D.3) everywhere. Moreover, we have

$$f(z) \leq w_1 \, \mathbb{E}_{Q_X} \left\| x - X' \right\|^2 + w_2 \, \mathbb{E}_{Q_Y} \left\| y - Y' \right\|^2 \leq 8wD^2$$

and g similarly. □

D.2.2 Controlling the Empirical Process and the U-Process

We then upper bound the $\mathbf{L}^1$ loss $\mathbb{E} \left| T_n(X, Y) - T(X, Y) \right|$ by empirical processes and U-processes.

Proposition D.6 (Corollary 2 (Mena and Weed, 2019))**.** *Let $P, Q, P', Q' \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Then we have*

$$|S(P', Q') - S(P, Q)| \leq \sup_{f \in \mathcal{F}_\sigma} \left| \int f(dP' - dP) \right| + \sup_{g \in \mathcal{F}_\sigma} \left| \int g(dQ' - dQ) \right|,$$

where $\mathcal{F}_\sigma$ is defined in Definition D.1.

To simply the function class $\mathcal{F}_\sigma$, we show in Lemma D.15 in Appendix D.4 that $(1 + \sigma^{3s})^{-1} \mathcal{F}_\sigma \subset \mathcal{F}^s$ for $\mathcal{F}^s$ defined below. Consequently, we can separate the sub-Gaussian parameter σ from the function class $\mathcal{F}_\sigma$.

Definition D.2. *For any $s \geq 2$, $d \in \mathbb{N}_+$, and $w = (w_1, w_2) \in \mathbb{R}_+^2$, let $\mathcal{F}^s := \mathcal{F}^{s,d,w}$ be the set of functions satisfying*

$$\begin{aligned} |f(z)| &\leq C_{s,d,w}(1 + \|z\|^2) \\ |D^\alpha f(z)| &\leq C_{s,d,w}(1 + \|z\|^{|\alpha|}), \quad \forall 1 \leq |\alpha| \leq s, \end{aligned}$$

where $C_{s,d,w}$ is a constant depending on s, d, and w.

In order to handle the U-process, we also need a variant function class of $\mathcal{F}^s$ which we also define below.

Definition D.3. *For any $\sigma \in \mathbb{R}_+$, $s \geq 2$, $d \in \mathbb{N}_+$, and $w = (w_1, w_2) \in \mathbb{R}_+^2$, let $\mathcal{F}_\sigma^s := \mathcal{F}_\sigma^{s,d,w}$ be the set of functions satisfying*

$$
\begin{aligned}
|f(z)| &\leq C_{s,d,w}(1 + \max\{\|z\|^2, \sigma^2\}) \\
|D^\alpha f(z)| &\leq C_{s,d,w}(1 + \max\{\|z\|^{|\alpha|}, \sigma^{|\alpha|}\}), \quad \forall 1 \leq |\alpha| \leq s,
\end{aligned}
$$

where $C_{s,d,w}$ is a constant depending on s, d, and w.

Let us control the complexity of $\mathcal{F}^s$ and $\mathcal{F}_\sigma^s$, which is achieved by the following covering number bound.

Proposition D.7. *Let $P \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Let $\{Z_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and P_n be the empirical measure. There exists a random variable $L \geq 1$ depending on the sample $\{Z_i\}_{i=1}^n$ with $\mathbb{E}[L] \leq 2$ such that*

$$
\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(P_n)) \leq C_{s,d,w}\tau^{-d/s}L^{d/2s}(1 + \sigma^{2d}) \quad \text{and} \quad \max_{f \in \mathcal{F}^s} \|f\|_{\mathbf{L}^2(P_n)}^2 \leq C_{s,d,w}(1 + L\sigma^4).
$$

Moreover, the same bounds hold for $\mathcal{F}_\sigma^s$.

Proof of Proposition D.7. Define $L := \frac{1}{n}\sum_{i=1}^n e^{\|Z_i\|^2/2d\sigma^2} \geq 1$. By the sub-Gaussianity of P, we have $\mathbb{E}[L] \leq 2$. In order to apply (van der Vaart and Wellner, 1996, Corollary 2.7.4), we partition $\mathbb{R}^d$ into $\cup_{j \geq 1} B_j$ where $B_1 := [-\sigma, \sigma]^d$ and $B_j := [-j\sigma, j\sigma]^d \backslash [-(j-1)\sigma, (j-1)\sigma]^d$ for $j \geq 2$. Since B_j is *not convex* for $j \geq 2$, we further partition it into disjoint hypercubes $\{B_{j,k}\}_{k=1}^{2d}$, e.g.,

$$
B_{j,1} = [(j-1)\sigma, j\sigma] \times [-j\sigma, j\sigma]^{d-1}.
$$

Take any $j \geq 2$ and $k \in [2d]$. Firstly, it holds that

$$
\lambda\{x : d(x, B_{j,k}) \leq 1\} \leq (\sigma + 2)(2j\sigma + 2)^{d-1} \leq C_d(1 + j^d\sigma^d),
$$

where λ is the Lebesgue measure. Secondly, the mass that P_n assigns to $B_{j,k}$ can be bounded as follows:

$$
P_n(Z \in B_{j,k}) \leq P_n\left(\|Z\|^2 > d\sigma^2(j-1)^2\right) \leq P_n\left[e^{\|Z\|^2/2d\sigma^2}\right]e^{-(j-1)^2/2} = Le^{-(j-1)^2/2}. \quad \text{(D.13)}
$$

Finally, we prove that $\mathcal{F}^s \subset \mathcal{C}^s_M(B_{j,k})$ with $M = C_{s,d,w}(1 + j^s\sigma^s)$, where $\mathcal{C}^s_M(B_{j,k})$ is the set of continuous functions satisfying

$$\|f\|_s := \max_{|\alpha|\leq s} \sup_{z\in B_{j,k}} |D^\alpha f(z)| + \max_{|\alpha|=s} \sup_{z,w\in B_{j,k}} |D^\alpha f(z) - D^\alpha f(w)| \leq M.$$

In fact, for any $f \in \mathcal{F}^s$, we have

$$\max_{|\alpha|\leq s} \sup_{z\in B_{j,k}} |D^\alpha f(z)| \leq C_{s,d,w} \sup_{z\in B_{j,k}} (1 + \|z\|^s) \leq C_{s,d,w}(1 + j^s\sigma^s),$$

and

$$\max_{|\alpha|=s} \sup_{z,w\in B_{j,k}} |D^\alpha f(z) - D^\alpha f(w)| \leq 2 \max_{|\alpha|=s} \sup_{z\in B_{j,k}} |D^\alpha f(z)| \leq C_{s,d}(1 + j^s\sigma^s).$$

Note that the same argument holds for any $f \in \mathcal{F}^s_\sigma$ since we can simply replace $1 + \|z\|^s$ by $1 + \max\{\|z\|^s, \sigma^s\}$. Now, applying (van der Vaart and Wellner, 1996, Corollary 2.7.4) with $r = 2$ and $V = d/s$ leads to

$$\begin{aligned}
&\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(P_n)) \\
&\leq C_{s,d,w}\tau^{-d/s}L^{d/2s}\left(1 + \sum_{j=2}^{\infty}\sum_{k=1}^{2d}(1 + j^d\sigma^d)^{\frac{2s}{d+2s}}(1 + j^s\sigma^s)^{\frac{2d}{d+2s}}e^{-\frac{d(j-1)^2}{d+2s}}\right)^{\frac{d+2s}{2s}} \\
&\leq C_{s,d,w}\tau^{-d/s}L^{d/2s}(1 + \sigma^{2d})\left(2d\sum_{j=1}^{\infty} j^{\frac{4ds}{d+2s}}e^{-\frac{d(j-1)^2}{d+2s}}\right)^{\frac{d+2s}{2s}} \\
&\leq C_{s,d,w}\tau^{-d/s}L^{d/2s}(1 + \sigma^{2d}), \quad \text{by the summability.}
\end{aligned}$$

To verify the second inequality, we obtain

$$\max_{f\in\mathcal{F}^s} \|f\|^2_{\mathbf{L}^2(P_n)} = \max_{f\in\mathcal{F}^s} P_n[|f(Z)|^2] \leq C_{s,d,w}P_n[(1 + \|Z\|^4)]. \tag{D.14}$$

Note that $\|Z\|^4 \leq C_d e^{\|Z\|^2/2d\sigma^2}\sigma^4$. It follows that $P_n[\|Z\|^4] \leq C_d L\sigma^4$, and thus

$$\max_{f\in\mathcal{F}^s} \|f\|^2_{\mathbf{L}^2(P_n)} \leq C_{s,d,w}(1 + L\sigma^4).$$

Again, the same argument hold for $\mathcal{F}^s_\sigma$ by replacing $\|Z\|^4$ with $\max\{\|Z\|^4, \sigma^4\}$. □

With this covering number bound at hand, we can control the empirical process by the metric entropy.

Proposition D.8. *Let $P \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Let $\{Z_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and P_n be the empirical measure. Then,*

$$\mathbb{E}\, \|P_n - P\|_{\mathcal{F}^s}^2 \leq C_{s,d,w}(1+\sigma^{2d+4})\frac{1}{n}, \quad \textit{for all } s > d/2.$$

Moreover, the same bound holds for $\mathcal{F}_\sigma^s$.

Proof. Define the symmetrized version of $\|P_n - P\|_{\mathcal{F}^s}$ by

$$\left\|\hat{\mathbb{S}}_n\right\|_{\mathcal{F}^s} := \sup_{f\in\mathcal{F}^s} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i)\right|, \tag{D.15}$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables that are independent with $\{Z_i\}_{i=1}^n$. According to (Wainwright, 2019, Proposition 4.11), it holds that

$$\mathbb{E}\, \|P_n - P\|_{\mathcal{F}^s}^2 \leq 4\, \mathbb{E} \left\|\hat{\mathbb{S}}_n\right\|_{\mathcal{F}^s}^2 .$$

Conditioning on $\{Z_i\}_{i=1}^n$, the random variable $Z(f) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(Z_i)$ is a linear combination of independent Rademacher random variables. Hence, $Z(f)$ is a sub-Gaussian process (see Definition 5.3) with respect to

$$\|f-g\|_{\mathbf{L}^2(P_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n [f(Z_i) - g(Z_i)]^2}.$$

It then follows from Proposition D.9 below that

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}^s} |Z(f)|^2 &\leq C\left(\int_0^{2\max_{f\in\mathcal{F}^s}\|f\|_{\mathbf{L}^2(P_n)}} \sqrt{\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(P_n))}d\tau\right)^2 \\
&\leq C_{s,d,w}\left(\int_0^{C_{s,d}\sqrt{1+L\sigma^4}} \tau^{-d/2s} L^{d/4s}\sqrt{1+\sigma^{2d}}d\tau\right)^2, \quad \text{by Proposition D.7} \\
&= C_{s,d,w}(1+\sigma^{2d})L^{d/2s}(1+L\sigma^4)^{1-d/2s}, \quad \text{by } s > d/2 \\
&\leq C_{s,d,w}(1+\sigma^{2d+4})L, \quad \text{by } L \geq 1.
\end{aligned}$$

Note that $\mathbb{E}\left\|\hat{\mathbb{S}}_n\right\|_{\mathcal{F}^s}^2 = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}^s}|Z(f)|^2$. Consequently, we have

$$\mathbb{E}\left\|P_n - P\right\|_{\mathcal{F}^s}^2 \leq C_{s,d,w}(1+\sigma^{2d+4})\frac{1}{n}. \tag{D.16}$$

The same argument holds for $\mathcal{F}_\sigma^s$ since Proposition D.7 holds true for $\mathcal{F}_\sigma^s$. □

The following proposition controls the $\mathbf{L}^2$ norm of the supremum of a sub-Gaussian process. It can be obtained from Giné and Nickl (2015, Exercise 2.3.1). We give its proof here for self-completeness.

Proposition D.9. *Let $\{Z(\theta)\}_{\theta\in\Theta}$ be a sub-Gaussian process with respect to a metric ρ in Θ such that $\int_0^\infty \sqrt{\log N(\tau,\Theta,\rho)}d\tau < \infty$. Then it holds that, for any separable version of Z,*

$$\left\|\sup_{\theta\in\Theta}|Z(\theta)|\right\|_{\mathbf{L}^2} \leq \|Z(\theta_0)\|_{\mathbf{L}^2} + C\int_0^D \sqrt{\log N(\tau,\Theta,\rho)}d\tau, \tag{D.17}$$

where $\theta_0\in\Theta$ is arbitrary and D is the ρ-diameter of Θ.

Proof. Due to the separability, it suffices to prove

$$\left\|\sup_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} \leq \|Z(\theta_0)\|_{\mathbf{L}^2} + C\int_0^D \sqrt{\log N(\tau,\Theta,\rho)}d\tau \tag{D.18}$$

for any finite $\Theta'\subset\Theta$. When the diameter $D=0$, the claim holds trivially and thus we only need to focus on the case when $|\Theta'|\geq 2$. By considering $(Z(\theta)-Z(\theta_0))/(1+\delta)D$ and $\rho/(1+\delta)D$ instead of $Z(\theta)$ and ρ for some any small $\delta>0$, we may assume that $Z(\theta_0)=0$ and $D\in(1/2,1)$. Our proof relies on the classical chaining argument.

Step 1. Construct a chain of projections. Let $r_1\in\mathbb{N}$ be such that, for any $\theta\in\Theta$, the ball $B(\theta,2^{-r_1})$ centered at θ of radius 2^{-r_1} contains at most 1 element in Θ'. Denote $\Theta_{r_1} := \Theta'$ and $\Theta_0 := \{\theta_0\}$. For each $1\leq r<r_1$, we take a 2^{-r} covering of Θ and let Θ_r be the collection of these centers. By definition, we get $|\Theta_r|\leq N(2^{-r},\Theta,\rho)$ for all $0\leq r\leq r_1$. For each $\theta\in\Theta'$, we construct a chain $(\pi_{r_1}(\theta),\pi_{r_1-1}(\theta),\dots,\pi_0(\theta))$ such that $\pi_r(\theta)\in\Theta_r$ as follows. For $r=r_1$, we let $\pi_r(\theta)=\theta$. For any $0\leq r<r_1$, we define $\pi_r(\theta)$ to be a point in Θ_r for which the ball $B(\pi_r(\theta),2^{-r})$ contains $\pi_{r+1}(\theta)$. Note that there may be multiple points satisfying this requirement, but we select the same one for θ and θ' as long as $\pi_{r+1}(\theta)=\pi_{r+1}(\theta')$.

Step 2. Telescoping. By the triangle inequality, we have

$$\left\|\max_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} = \left\|\max_{\theta\in\Theta'}|Z(\pi_{r_1}(\theta)) - Z(\pi_0(\theta))|\right\|_{\mathbf{L}^2} \leq \sum_{r=1}^{r_1}\left\|\max_{\theta\in\Theta'}|Z(\pi_r(\theta)) - Z(\pi_{r-1}(\theta))|\right\|_{\mathbf{L}^2}.$$

Note that

$$|\{(\pi_r(\theta), \pi_{r-1}(\theta)) : \theta \in \Theta'\}| = |\{\pi_r(\theta) : \theta \in \Theta'\}| \leq |\Theta_r| \leq N(2^{-r}, \Theta, \rho).$$

According to (Giné and Nickl, 2015, Lemma 2.3.3), we obtain

$$\begin{aligned}\left\|\max_{\theta\in\Theta'}|Z(\pi_r(\theta)) - Z(\pi_{r-1}(\theta))|\right\|_{\mathbf{L}^2} &\leq C\sqrt{\log N(2^{-r}, \Theta, \rho)}\max_{\theta\in\Theta'}\|Z(\pi_r(\theta)) - Z(\pi_{r-1}(\theta))\| \\ &\leq C2^{-r+1}\sqrt{\log N(2^{-r}, \Theta, \rho)}.\end{aligned}$$

Consequently, it holds that

$$\left\|\max_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} \leq C\sum_{r=1}^{r_1}2^{-r+1}\sqrt{\log N(2^{-r}, \Theta, \rho)} \leq C\int_0^1\sqrt{\log N(\tau, \Theta, \rho)}d\tau,$$

which completes the proof. □

D.2.3 Proofs of Main Results

We now prove the main consistency results in Section 5.5. For simplicity of the notation, we focus on the quadratic cost function, i.e., $w_1 = w_2 = 1$, and drop the dependency on w (e.g., we write $C_{s,d} = C_{s,d,w}$. The proofs can be adapted to weighted quadratic costs with minor modifications. Let $\mu_X \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $\mu_Y \in \mathcal{M}_1(\mathbb{R}^{d_2})$ with $d := d_1 + d_2$. Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from some joint distribution μ_{XY} with marginals μ_X and μ_Y, where μ_{XY} may or may not equal $\mu_X \otimes \mu_Y$. Let P_n and Q_n be the empirical measures of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively.

Proof of Proposition 5.7. Step 1. Decoupling. Due to the degeneracy, it suffices to bound

$$\mathbb{E}\|\hat{\mu}_X \otimes \hat{\mu}_Y\|_{\mathcal{F}}^2 = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i, Y_j)\right|^2\right]. \tag{D.19}$$

We prove in the following that it boils down to control (D.19) under the product measure $\mu_X \otimes \mu_Y$. When $\mu_{XY} = \mu_X \otimes \mu_Y$, the claim holds trivially. When $\mu_{XY} \neq \mu_X \otimes \mu_Y$, we use the decoupling technique (Peña and Giné, 1999). Note that, by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] \leq C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i\neq j}^n f(X_i,Y_j)\right|^2 + \sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n f(X_i,Y_i)\right|^2\right].$$

Note that the second term on the RHS is a lower order term and can be taken care of by Proposition D.8. Hence, it suffices to upper bound the first term. Let $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. Rademacher random variables and $\{(X_i', Y_i')\}_{i=1}^n$ be an independent copy of $\{(X_i, Y_i)\}_{i=1}^n$. Define

$$A_i := \begin{cases} X_i & \text{if } \varepsilon_i = 1 \\ X_i' & \text{if } \varepsilon_i = -1 \end{cases} \quad \text{and} \quad B_i := \begin{cases} Y_i' & \text{if } \varepsilon_i = 1 \\ Y_i & \text{if } \varepsilon_i = -1 \end{cases}.$$

For any functional $F : \mathcal{F} \to \mathbb{R}_+$, let $\Phi(F) := \sup_{f\in\mathcal{F}} F(f)^2$. For instance, we define $U_{X,Y}(f) := \frac{1}{n^2}\left|\sum_{i\neq j} f(X_i, Y_j)\right|$. It is clear that Φ is convex and increasing, and the target reads

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] = \mathbb{E}\left[\Phi\left(\left|\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[f(X_i,Y_j) + f(X_i',Y_j) + f(X_i,Y_j') + f(X_i',Y_j') \mid \mathcal{Z}\right]\right|\right)\right],$$

where $\mathcal{Z} := \{(X_i, Y_i)\}_{i=1}^n$. Since, for any $i \neq j$,

$$f(X_i,Y_j) + f(X_i',Y_j) + f(X_i,Y_j') + f(X_i',Y_j') = 4\,\mathbb{E}\left[f(A_i,B_j) \mid \mathcal{Z},\mathcal{Z}'\right],$$

it follows from the convexity and the monotonicity of Φ that

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] \leq \mathbb{E}\left[\Phi(4U_{A,B})\right].$$

Finally, the joint distribution of $(X_1, \ldots, X_n, Y_1', \ldots, Y_n')$ is the same as the one of $(A_1, \ldots, A_n, B_1, \ldots, B_n)$, so we have

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] \leq \mathbb{E}\left[\Phi(4U_{X,Y'})\right].$$

Adding back the diagonal terms proves the claim since $(X_i, Y_i') \sim \mu_X \otimes \mu_Y$.

Step 2. Randomization. We work under the measure $\mu_{XY} = \mu_X \otimes \mu_Y$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] \\
&= \mathbb{E}_Y\,\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\left[\sum_{j=1}^n f(X_i,Y_j) - \mathbb{E}_{X'}\left[\sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right]\right|^2\right], \quad \text{by (D.23)} \\
&\leq \mathbb{E}_Y\,\mathbb{E}_{X,X'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\left[\sum_{j=1}^n f(X_i,Y_j) - \sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right|^2\right], \quad \text{by Jensen's inequality} \\
&= \mathbb{E}_Y\,\mathbb{E}_{X,X',\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\varepsilon_i\left[\sum_{j=1}^n \bar{f}(X_i,Y_j) - \sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right|^2\right] \\
&\leq C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\varepsilon_i f(X_i,Y_j)\right|^2\right], \quad \text{by the Cauchy-Schwarz inequality.}
\end{aligned}$$

Repeating above arguments gives

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] &\leq C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2\right] \\
&\leq C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2\right],
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and Jensen's inequality. Hence, it suffices to bound

$$A := \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2.$$

Step 3. Metric entropy. Define the process $Z(f) := \frac{1}{n^{3/2}}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)$ for any $f \in \mathcal{F}$. We claim that it is a sub-Gaussian process with respect to

$$\|f-g\|_{\mathbf{L}^2(P_n\otimes Q_n)} = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^n [f(X_i,Y_j) - g(X_i,Y_j)]^2}. \tag{D.20}$$

To prove it, let us control the moment generating function of the increment $Z(f) - Z(g)$. Denote $a_i := \sum_{j=1}^n \varepsilon_j'[f(X_i, Y_j) - g(X_i, Y_j)]$. Conditioning on $\{X_i, Y_i, \varepsilon_i'\}_{i=1}^n$,

$$Z(f) - Z(g) = \frac{1}{n^{3/2}} \sum_{i=1}^n a_i \varepsilon_i$$

is a linear combination of independent Rademacher random variables. Consequently,

$$\mathbb{E}_\varepsilon \exp\{\lambda[Z(f) - Z(g)]\} \le \exp\left\{\frac{\lambda^2 \sum_{i=1}^n a_i^2}{2n^3}\right\}. \tag{D.21}$$

Note that, by the Cauchy-Schwarz inequality,

$$a_i^2 \le \left[\sum_{j=1}^n (\varepsilon_j')^2\right]\left[\sum_{j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2\right] = n\left[\sum_{j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2\right].$$

This yields that

$$\begin{aligned}\mathbb{E}_\varepsilon \exp\{\lambda[Z(f) - Z(g)]\} &\le \exp\left\{\frac{\lambda^2 \sum_{i,j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2}{2n^2}\right\} \\ &= \exp\left\{\frac{\lambda^2 \|f - g\|_{\mathbf{L}^2(P_n \otimes Q_n)}^2}{2}\right\},\end{aligned}$$

and thus the claim follows. Therefore, the conclusion in Proposition 5.7 holds true due to Proposition D.9. □

Proof of Proposition 5.8. The proof of the first part is similar to Proposition D.7. Define $L_1 := \hat{\mu}_X[e^{\|X\|^2/2d\sigma^2}] \ge 1$ and $L_2 := \hat{\mu}_Y[e^{\|Y\|^2/2d\sigma^2}] \ge 1$. By the sub-Gaussian assumption, it is clear that $\mathbb{E}[L_1] \le 2$ and $\mathbb{E}[L_2] \le 2$. There are two places in the proof of Proposition D.7 where the measure is involved. The first place is (D.13), where we replace it by

$$\begin{aligned}&(\hat{\mu}_X \otimes \hat{\mu}_Y)\{(X, Y) \in B_{j,k}\} \\ &\le (\hat{\mu}_X \otimes \hat{\mu}_Y)\left\{\|X\|^2 + \|Y\|^2 > d\sigma^2 (j-1)^2\right\} \\ &\le (\hat{\mu}_X \otimes \hat{\mu}_Y)\left[\exp\left(\frac{\|X\|^2 + \|Y\|^2}{4d\sigma^2}\right)\right] e^{-(j-1)^2/4}, \quad \text{by the Chernoff bound} \\ &= L_1 L_2 e^{-(j-1)^2/4}.\end{aligned}$$

The second place is (D.14), where we replace it by

$$\max_{f\in\mathcal{F}^s}\|f\|^2_{\mathbf{L}^2(\hat\mu_X\otimes\hat\mu_Y)}=\max_{f\in\mathcal{F}^s}(\hat\mu_X\otimes\hat\mu_Y)[|f(X,Y)|^2]\le C_{s,d}(\hat\mu_X\otimes\hat\mu_Y)[1+\|X\|^4+\|Y\|^4].$$

Note that $\|Z\|^4\le C_d e^{\|Z\|^2/2d\sigma^2}\sigma^4$. It follows that $(\hat\mu_X\otimes\hat\mu_Y)[\|X\|^4+\|Y\|^4]\le C_d(L_1+L_2)\sigma^4$. Hence, the claim holds true for $L:=(L_1+L_2)/2$.

For the second part, we define $\theta_f:=\mathbb{E}_{\mu_X\otimes\mu_Y}[f(X,Y)]$,

$$f_{1,0}(X):=\mathbb{E}_{\mu_X\otimes\mu_Y}[f(X,Y)\mid X]\quad\text{and}\quad f_{0,1}(Y):=\mathbb{E}_{\mu_X\otimes\mu_Y}[f(X,Y)\mid Y]\tag{D.22}$$

for each $f\in\mathcal{F}^s$. As a result, $\bar f(x,y):=f(x,y)-f_{1,0}(x)-f_{0,1}(y)+\theta_f$ satisfies

$$\mathbb{E}_{\mu_X\otimes\mu_Y}[\bar f(X,Y)\mid X]\overset{\text{a.s.}}{=}0\overset{\text{a.s.}}{=}\mathbb{E}_{\mu_X\otimes\mu_Y}[\bar f(X,Y)\mid Y].\tag{D.23}$$

Note that

$$\begin{aligned}&\mathbb{E}\,\|\hat\mu_X\otimes\hat\mu_Y-\mu_X\otimes\mu_Y\|^2_{\mathcal{F}^s}\\&=\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n\big(f(X_i,Y_j)-\theta_f\big)\right|^2\right]\\&\le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n\bar f(X_i,Y_j)\right|^2+\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n}\sum_{i=1}^n f_{1,0}(X_i)-\theta_f\right|^2+\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n}\sum_{i=1}^n f_{0,1}(Y_i)-\theta_f\right|^2\right]\\&\le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n\bar f(X_i,Y_j)\right|^2+\|\hat\mu_X-\mu_X\|^2_{\mathcal{F}^s_\sigma}+\|\hat\mu_Y-\mu_Y\|^2_{\mathcal{F}^s_\sigma}\right],\quad\text{by Lemma D.16.}\end{aligned}\tag{D.24}$$

Since the last two terms above can be controlled by Proposition D.8, it remains to consider the first term. Analogous to the proof of Proposition D.8, we obtain, by Proposition 5.7 and the first part, that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n\bar f(X_i,Y_j)\right|^2\right]\le C_{s,d}(1+\sigma^{2d+4})\frac{1}{n}.$$

Therefore, by (D.24), we have

$$\mathbb{E}\,\|P_n\otimes Q_n-P\otimes Q\|^2_{\mathcal{F}^s}\le C_{s,d}(1+\sigma^{2d+4})\frac{1}{n}.$$

□

Now we are ready to prove Theorem 5.6.

Proof of Theorem 5.6. We prove the statement for $\varepsilon = 1$ and write $S := S_1$. The result for general $\varepsilon > 0$ follows immediately from Lemma D.17. By the triangle inequality, it holds that

$$\begin{aligned} &|T_n(X,Y) - T(X,Y)| \\ &\le |S(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y) - S(\mu_{XY}, \mu_X \otimes \mu_Y)| + \frac{1}{2} |S(\hat{\mu}_{XY}, \hat{\mu}_{XY}) - S(\mu_{XY}, \mu_{XY})| \\ &\quad + \frac{1}{2} |S(\hat{\mu}_X \otimes \hat{\mu}_Y, \hat{\mu}_X \otimes \hat{\mu}_Y) - S(\mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y)| . \end{aligned} \tag{D.25}$$

We begin with deriving the bound for the first term

$$A := |S(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y) - S(\mu_{XY}, \mu_X \otimes \mu_Y)| . \tag{D.26}$$

Step 1. Upper bound via empirical processes. According to Lemma D.12 and Lemma D.13, the joint distribution μ_{XY} is subG$(2\sigma^2)$, and thus there exist a zero-measure set $S_{\mu_{XY}} \subset \Omega$ and a random variable $\sigma^2_{\mu_{XY}}$ such that $\hat{\mu}_{XY}(\omega)$ and μ_{XY} are subG$(\sigma^2_{\mu_{XY}}(\omega))$ for every $\omega \in S^c_{\mu_{XY}}$. Similarly, by Lemma D.14, there exist a zero-measure set $S_{\mu_X,\mu_Y} \subset \Omega$ and a random variable $\sigma^2_{\mu_X,\mu_Y}$ such that $\hat{\mu}_X(\omega) \otimes \hat{\mu}_Y(\omega)$ and $\mu_X \otimes \mu_Y$ are subG$(\sigma^2_{P,Q}(\omega))$ for every $\omega \in S^c_{\mu_X,\mu_Y}$. Take $S := S^c_{\mu_{XY}} \cap S^c_{\mu_X,\mu_Y}$ and $\bar{\sigma}^2 := \max\{\sigma^2_{\mu_{XY}}, \sigma^2_{\mu_X,\mu_Y}\}$. It follows that $\hat{\mu}_{XY}(\omega)$, $\hat{\mu}_X(\omega) \otimes \hat{\mu}_Y(\omega)$, μ_{XY}, and $\mu_X \otimes \mu_Y$ are subG$(\bar{\sigma}^2(\omega))$ for every $\omega \in S$. Now, by Proposition D.6,

$$\begin{aligned} &|S(\hat{\mu}_{XY}(\omega), \hat{\mu}_X(\omega) \otimes \hat{\mu}_Y(\omega)) - S(\mu_{XY}, \mu_X \otimes \mu_Y)| \\ &\le \sup_{f \in \mathcal{F}_{\bar{\sigma}(\omega)}} \left| \int f(d\hat{\mu}_{XY}(\omega) - d\mu_{XY}) \right| + \sup_{g \in \mathcal{F}_{\bar{\sigma}(\omega)}} \left| \int g(d\hat{\mu}_X(\omega) \otimes \hat{\mu}_Y(\omega) - d\mu_X \otimes \mu_Y) \right|, \quad \forall \omega \in S. \end{aligned}$$

Note that $\mathbb{P}(S) = \mathbb{P}(S^c_{\mu_{XY}} \cap S^c_{\mu_X,\mu_Y}) = 1$. This implies, almost surely,

$$A \le \sup_{f \in \mathcal{F}_{\bar{\sigma}}} \left| \int f(d\hat{\mu}_{XY} - d\mu_{XY}) \right| + \sup_{g \in \mathcal{F}_{\bar{\sigma}}} \left| \int g(d\hat{\mu}_X \otimes \hat{\mu}_Y - d\mu_X \otimes \mu_Y) \right|. \tag{D.27}$$

According to Lemma D.15, we have

$$\begin{aligned} \mathbb{E}[A] &\le \mathbb{E}\left[(1 + \bar{\sigma}^{3s}) \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s}\right] + \mathbb{E}\left[(1 + \bar{\sigma}^{3s}) \|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}^s}\right] \\ &\le \sqrt{\mathbb{E}[(1 + \bar{\sigma}^{3s})^2]} \left[\sqrt{\mathbb{E}\|\hat{\mu}_{XY} - \mu_{XY}\|^2_{\mathcal{F}^s}} + \sqrt{\mathbb{E}\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|^2_{\mathcal{F}^s}} \right]. \end{aligned}$$

Step 2. Control empirical processes via metric entropy. Let $s = \lceil d/2 \rceil + 1$. Since the joint probability P_{XY} is $\mathrm{subG}(2\sigma^2)$, it follows from Proposition D.8 that

$$\sqrt{\mathbb{E}\,\|\hat{\mu}_{XY} - \mu_{XY}\|^2_{\mathcal{F}^s}} \le C_d(1 + \sigma^{d+2})\frac{1}{\sqrt{n}}. \tag{D.28}$$

The same bound holds for $\sqrt{\mathbb{E}\,\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|^2_{\mathcal{F}^s}}$ by Proposition 5.7. Note that

$$\mathbb{E}[(1 + \tilde{\sigma}^{3s})^2] \le C(1 + \mathbb{E}\,\tilde{\sigma}^{6s}) \le C_s(1 + \mathbb{E}\,\sigma^{6s}_{P_{XY}} + \mathbb{E}\,\sigma^{6s}_{P_X, P_Y}) \le C_s(1 + \sigma^{6s}),$$

where the last inequality follows from Lemma D.12 and Lemma D.14. Recall that we have chosen $s = \lceil d/2 \rceil + 1$. As a result, $\mathbb{E}[A] \le C_d(1 + \sigma^{\lceil 5d/2 \rceil + 6})n^{-1/2}$. A similar argument shows that the same bound hold for the second and third term in (D.25). Hence,

$$\mathbb{E}\,|T_n(X, Y)| \le C_d(1 + \sigma^{\lceil 5d/2 \rceil + 6})\frac{1}{\sqrt{n}}. \tag{D.29}$$

□

D.3 Exponential Tail Bounds

We now prove the exponential tail bound in Theorem 5.9. For simplicity of the notation, we focus on the quadratic cost function, i.e., $w_1 = w_2 = 1$, and drop the dependency on w (e.g., we write $C_{s,d} = C_{s,d,w}$. The proofs can be adapted to weighted quadratic costs with minor modifications. Let $\mu_X \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $\mu_Y \in \mathcal{M}_1(\mathbb{R}^{d_2})$ with $d := d_1 + d_2$. Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from some joint distribution μ_{XY} with marginals μ_X and μ_Y, where μ_{XY} may or may not equal $\mu_X \otimes \mu_Y$. Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be the empirical measures of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively.

Proposition D.10. *For any b-uniformly bounded class of functions $\mathcal{F}$, we have*

$$\mathbb{P}\left\{\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}} - \mathbb{E}\,\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}} > t\right\} \le \exp\left(-\frac{nt^2}{8b^2}\right),$$

for any $t \ge 0$.

Proof. For any function f defined on $\mathbb{R}^d$, we define $\bar{f}(x, y) = f(x, y) - (\mu_X \otimes \mu_Y)[f]$. As a results, we have $\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(X_i, Y_j)\right|$. Consider the function

$$F(z_1, \ldots, z_n) := \sup_{f\in\mathcal{F}} \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right|, \tag{D.30}$$

where $z_i = (x_i, y_i) \in \mathbb{R}^d$. We claim that F satisfies the bounded difference property required in the McDiarmid inequality. Since F is permutation invariant, it suffices to verify the property for the first coordinate. Let $z_1' \neq z_1$ and $z_i' = z_i$ for all $i \neq 1$. It holds that

$$\begin{aligned}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right| - F(z_1', \ldots, z_n') &\leq \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right| - \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i', y_j')\right| \\ &\leq \frac{1}{n^2}\sum_{i=1 \text{ or } j=1} \left|\bar{f}(x_i, y_j) - \bar{f}(x_i', y_j')\right| \leq \frac{4b}{n},\end{aligned}$$

where the last inequality uses the boundedness of f. Taking the supremum over $\mathcal{F}$ yields that $F(z_1, \ldots, z_n) - F(z_1', \ldots, z_n') \leq 4b/n$. By symmetry, it follows that $|F(z_1, \ldots, z_n) - F(z_1', \ldots, z_n')| \leq 4b/n$. Note that $\{Z_i := (X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample. According to the McDiarmid inequality, it holds that, for any $t \geq 0$,

$$\mathbb{P}\left\{\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}} - \mathbb{E}\,\|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}} > t\right\} \leq \exp\left(-\frac{nt^2}{8b^2}\right).$$

□

Proof of Theorem 5.9. We prove the statement for $\varepsilon = 1$ and write $S := S_1$. The result for general $\varepsilon > 0$ follows immediately from Lemma D.17. By the bounded support assumption, it holds that μ_X and μ_Y are both $\mathrm{subG}(D^2/d)$. According to the proof of Lemma D.12, we have $\{\hat{\mu}_X\}_{n\geq 1}$, $\{\hat{\mu}_Y\}_{n\geq 1}$, μ_X, and μ_Y are uniformly $\mathrm{subG}(\tau^2)$ for $\tau^2 := D^2 e^{1/2}/d \leq 2D^2/d$. Moreover, it follows from Lemma D.13 that $\{\hat{\mu}_{XY}\}_{n\geq 1}$ and μ_{XY} are uniformly $\mathrm{subG}(2\tau^2)$. As a result, we obtain, by Proposition D.6,

$$\begin{aligned}A &:= |S(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y) - S(\mu_{XY}, \mu_X \otimes \mu_Y)| \\ &\leq \sup_{f\in\mathcal{F}_{2\tau}} \left|\int f(d\hat{\mu}_{XY} - d\mu_{XY})\right| + \sup_{g\in\mathcal{F}_{2\tau}} \left|\int g(d\hat{\mu}_X \otimes \hat{\mu}_Y - d\mu_X \otimes \mu_Y)\right|.\end{aligned}$$

Fix $s = \lceil d/2 \rceil + 1$. According to Lemma D.15, we have

$$A \leq C_d(1 + D^{3d+12}) \left[\|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s} + \|\hat{\mu}_X \otimes \hat{\mu}_Y - \mu_X \otimes \mu_Y\|_{\mathcal{F}^s} \right], \tag{D.31}$$

where we have used $\tau^{3s} \leq C_d D^{3d+12}$. Proposition D.5 shows that we can further constraint the function class $\mathcal{F}^s$ to $\mathcal{F}^s_b := \{f \in \mathcal{F}^s : \|f\|_\infty \leq b\}$ for $b = 2D^2$. Hence, by (Wainwright, 2019, Theorem 4.10), it holds that

$$\mathbb{P}\left\{ \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s_b} - \mathbb{E}\, \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s_b} > t \right\} \leq \exp\left(-\frac{nt^2}{2b^2}\right), \quad \text{for any } t \geq 0.$$

It is clear from Proposition D.8 that

$$\mathbb{E}\, \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s_b} \leq \mathbb{E}\, \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s} \leq C_d(1 + D^{2d+4}) \frac{1}{\sqrt{n}}.$$

Consequently, we get

$$\mathbb{P}\left\{ \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s_b} > t + C_d(1 + D^{2d+4}) \frac{1}{\sqrt{n}} \right\} \leq \exp\left(-\frac{nt^2}{2b^2}\right), \quad \text{for any } t \geq 0.$$

Similarly, using Proposition 5.7 and Proposition D.10, we obtain

$$\mathbb{P}\left\{ \|\hat{\mu}_{XY} - \mu_{XY}\|_{\mathcal{F}^s_b} > t + C_d(1 + D^{2d+4}) \frac{1}{\sqrt{n}} \right\} \leq \exp\left(-\frac{nt^2}{8b^2}\right), \quad \text{for any } t \geq 0.$$

Now it follows from (D.31) that

$$\mathbb{P}\left\{ A \geq C_d(1 + D^{3d+12}) \left[t + (1 + D^{2d+4}) \frac{1}{\sqrt{n}} \right] \right\} \leq 2 \exp\left(-\frac{nt^2}{8b^2}\right), \quad \text{for any } t \geq 0.$$

Analogously, we have, for any $t \geq 0$

$$\begin{aligned} \mathbb{P}\left\{ B \geq C_d(1 + D^{3d+12}) \left[t + (1 + D^{2d+4}) \frac{1}{\sqrt{n}} \right] \right\} &\leq 2 \exp\left(-\frac{nt^2}{8b^2}\right) \\ \mathbb{P}\left\{ B' \geq C_d(1 + D^{3d+12}) \left[t + (1 + D^{2d+4}) \frac{1}{\sqrt{n}} \right] \right\} &\leq 2 \exp\left(-\frac{nt^2}{8b^2}\right), \end{aligned}$$

where $B := |S(\hat{\mu}_{XY}, \hat{\mu}_{XY}) - S(\mu_{XY}, \mu_{XY})|$ and

$$B' := |S(\hat{\mu}_X \otimes \hat{\mu}_Y, \hat{\mu}_X \otimes \hat{\mu}_Y) - S(\mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y)|\,.$$

Since $|T_n(X,Y) - T(X,Y)| \leq A + \frac{B}{2} + \frac{B'}{2}$, it holds that

$$\mathbb{P}\left\{|T_n(X,Y) - T(X,Y)| \geq C_d(1 + D^{3d+12})\left[t + (1 + D^{2d+4})\frac{1}{\sqrt{n}}\right]\right\} \leq 6\exp\left(-\frac{nt^2}{8b^2}\right). \tag{D.32}$$

Therefore, we have, with probability at least $1 - \delta$,

$$|T_n(X,Y) - T(X,Y)| \leq C_d\left(1 + D^{2d+2}\sqrt{\log\frac{6}{\delta}}\right)\frac{D^{3d+14}}{\sqrt{n}}.$$

□

D.4 Technical Lemmas

In this section, we give several technical lemmas used to prove the main results. We use C to denote a constant whose value may change from line to line.

Lemma D.11. *If $P \in \mathcal{M}_1(\mathbb{R}^d)$ is subG(σ^2), then, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}_P[\|Z\|^{2k}] \leq (2d\sigma^2)^k k!.$$

Moreover, for any $v \in \mathbb{R}^d$, it holds that

$$\mathbb{E}_P\, e^{\langle v, Z\rangle} \leq \mathbb{E}_P\, e^{\|v\|\|Z\|} \leq 2e^{d\sigma^2\|v\|^2/2}. \tag{D.33}$$

Proof. By Taylor's expansion, we have

$$e^{\|Z\|^2/2d\sigma^2} - 1 \geq \frac{\|Z\|^{2k}}{(2d\sigma^2)^k k!}.$$

Taking the expectation on both sides gives

$$\mathbb{E}_P[\|Z\|^{2k}] \leq (2d\sigma^2)^k k!.$$

The inequalities (D.33) follows from the Cauchy-Schwarz inequality and the sub-gaussianity of P. □

Lemma D.12. *Let $P \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2) and P_n be the empirical measure. There exist a zero-measure set $S_P \subset \Omega$ and a random variable σ_P^2 depending on the sample $\{Z_i\}_{i=1}^n$ such that $P_n(\omega)$ and P are subG$(\sigma_P^2(\omega))$ for any $\omega \in S_P^c$, and, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}\, \sigma_P^{2k} \leq 2k^k \sigma^{2k}.$$

Proof. By the strong law of large numbers, there exists a zero-measure set $S_P \subset \Omega$ such that, for all $\omega \in S_P$,

$$P_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2}\right] \to P\left[e^{\|Z\|^2/2d\sigma^2}\right] \leq 2, \quad \text{as } n \to \infty. \tag{D.34}$$

Let $\tau^2 := \sup_n P_n\left[e^{\|Z\|^2/2d\sigma^2}\right]$. It follows from (D.34) that $\tau^2(\omega)$ is finite for all $\omega \in S_P$. Since $\tau^2(\omega) \geq 1$, by Jensen's inequality, we obtain, for all $\omega \in S_P$

$$P_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2\tau^2(\omega)}\right] \leq \left(P_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2}\right]\right)^{1/\tau^2(\omega)} = \left(\tau^2(\omega)\right)^{1/\tau^2(\omega)} < 2.$$

As a result, $P_n(\omega)$ is subG$(\sigma^2\tau^2(\omega))$. Moreover, P is also subG$(\sigma^2\tau^2(\omega))$ since $\tau^2(\omega) \geq 1$. Applying the same argument to $\tau_k^2 := \sup_n P_n\left[e^{\|Z\|^2/2kd\sigma^2}\right]$ implies that $P_n(\omega)$ and P are both subG$(k\sigma^2\tau_k^2(\omega))$. Define $\sigma_P^2 := \min_{k\geq 1} k\sigma^2\tau_k^2$. Then we have, for each $k \geq 1$,

$$\mathbb{E}_P[\sigma_P^{2k}] \leq \mathbb{E}_P\left[P_n\left[k^k\sigma^{2k}e^{\|Z\|^2/2d\sigma^2}\right]\right] = k^k\sigma^{2k}\,\mathbb{E}_P[e^{\|Z\|^2/2d\sigma^2}] \leq 2k^k\sigma^{2k}.$$

□

The sub-Gaussianity of two marginals implies the sub-Gaussianity of the joint.

Lemma D.13. *If μ_X and μ_Y are subG(σ^2), then μ_{XY} is subG$(2\sigma^2)$ for any $\mu_{XY} \in \Pi(\mu_X, \mu_Y)$.*

Proof. By the Cauchy-Schwarz inequality,

$$\mathbb{E}_{\mu_{XY}}\, e^{\|Z\|^2/4d\sigma^2} = \mathbb{E}_{\mu_{XY}}[e^{\|X\|^2/4d\sigma^2}e^{\|Y\|^2/4d\sigma^2}] \leq \sqrt{\mathbb{E}_{\mu_X}[e^{\|X\|^2/2d\sigma^2}]\,\mathbb{E}_{\mu_Y}[e^{\|Y\|^2/2d\sigma^2}]}.$$

Since μ_X and μ_Y are subG(σ^2), it follows that $\mathbb{E}_{\mu_{XY}}\, e^{\|Z\|^2/4d\sigma^2} \leq 2$ and thus μ_{XY} is subG$(2\sigma^2)$.

□

The next result is for the uniform sub-Gaussianity of the product of two empirical measures.

Lemma D.14. *If μ_X and μ_Y are subG(σ^2), then there exist a zero-measure set $S_{\mu_X,\mu_Y} \subset \Omega$ and a random variable $\sigma^2_{\mu_X,\mu_Y}$ depending on the sample $\{(X_i, Y_i)\}_{i=1}^n$ such that $\hat{\mu}_X(\omega)\otimes\hat{\mu}_Y(\omega)$ and $\mu_X \otimes \mu_Y$ are subG$(\sigma^2_{\mu_X,\mu_Y}(\omega))$ for any $\omega \in S^c_{\mu_X,\mu_Y}$, and, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}\,\sigma^{2k}_{\mu_X,\mu_Y} \leq 2^{k+1}k^k\sigma^{2k}.$$

Proof. Similar to Lemma D.12. □

The sub-Gaussian processes play an central role in our analysis. We give its definition here; see, e.g., (Wainwright, 2019, Section 5.3).

Definition D.4 (Sub-Gaussian process)**.** *Let $\{Z(\theta) : \theta \in \Theta\}$ be a collection of mean-zero random variables. We call it a sub-Gaussian process with respect to a metric ρ in Θ if*

$$\mathbb{E}[e^{\lambda(Z(\theta)-Z(\theta'))}] \leq \exp\left[\lambda^2\rho^2(\theta,\theta')/2\right].$$

To facilitate the analysis of $\mathcal{F}_\sigma$ defined in Definition D.1, it is convenient to separate the sub-Gaussian parameter from the function class by the following lemma. Note that this result is used in (Mena and Weed, 2019) without proof.

Lemma D.15. *For any $\sigma > 0$ and $s \geq 2$. we have $\frac{1}{1+\sigma^{3s}}\mathcal{F}_\sigma \subset \mathcal{F}^s$, where $\mathcal{F}^s := \mathcal{F}^{s,d,w}$ is defined in Definition D.2.*

Proof. Take any $f \in \mathcal{F}_\sigma$, it suffices to show $f/(1+\sigma^{3s}) \in \mathcal{F}^s$. According to Proposition D.3, it holds that

$$\begin{aligned} |f(z)| - w_1\|x\|^2 - w_2\|y\|^2 &\leq \left|f(z) - w_1\|x\|^2 - w_2\|y\|^2\right| \\ &\leq C_{k,d,w}\begin{cases}(1+\sigma^4) & \text{if } \|z\| \leq \sqrt{d}\sigma \\ [1+(1+\sigma^2)\|z\|^2] & \text{if } \|z\| > \sqrt{d}\sigma.\end{cases}\end{aligned}$$

Consequently,

$$\left|\frac{f(z)}{1+\sigma^{3s}}\right| \leq C_{k,d,w} \begin{cases} \frac{1+\sigma^4}{1+\sigma^{3s}} & \text{if } \|z\| \leq \sqrt{d}\sigma \\ \frac{1+(1+\sigma^2)\|z\|^2}{1+\sigma^{3s}} & \text{if } \|z\| > \sqrt{d}\sigma. \end{cases}$$

Since $s \geq 2$, it is clear that $\frac{1+\sigma^4}{1+\sigma^{3s}} \leq C$ and $\frac{1+\sigma^2}{1+\sigma^{3s}} \leq C$, and thus

$$\left|\frac{f(z)}{1+\sigma^{3s}}\right| \leq C_{k,d,w}(1+\|z\|^2).$$

The other inequality can be proved analogously. □

Lemma D.16. *Let $P \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $Q \in \mathcal{M}_1(\mathbb{R}^{d_2})$ be subG(σ^2). Denote $d := d_1 + d_2$. For any $s \geq 1$ and $f \in \mathcal{F}^s$, there exist constants $C_{s,d,w}$ such that $f_{1,0} \in \mathcal{F}_\sigma^s$ and $f_{0,1} \in \mathcal{F}_\sigma^s$, where $\mathcal{F}_\sigma^s$ is defined in Definition D.3,*

$$f_{1,0}(x) := \int f(x,y)dQ(y) \quad \text{and} \quad f_{0,1}(y) := \int f(x,y)dP(x).$$

Proof. We only prove it for $f_{1,0}$. By Jensen's inequality, it holds that

$$\begin{aligned} |f_{1,0}(x)| &\leq \int |f(x,y)|\, dQ(y) \leq C_{s,d,w}\left(1+\|x\|^2+\int \|y\|^2\, dQ(y)\right) \\ &\leq C_{s,d,w}(1+\max\{\|x\|^2, \sigma^2\}), \end{aligned}$$

where the last inequality follows from Lemma D.11. The inequality for $|D^\alpha f_{1,0}(x)|$ can be verified similarly. □

The next lemma suggests that it is enough to consider the case $\varepsilon = 1$ for S_ε.

Lemma D.17. *Let $\varepsilon > 0$. For any $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$, it holds that*

$$S_\varepsilon(P,Q) = \varepsilon S(P^\varepsilon, Q^\varepsilon),$$

where P^ε and Q^ε are the pushforwards of P and Q under the map $x \mapsto \varepsilon^{-1/2}x$, respectively.

Proof. By a change of variable argument. □