Variational Problems in Feature Selection and Data Compression

Alex Kokot

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

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Marina Meila, Chair

Alex Luedtke, Chair

Zaid Harchaoui

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Variational Problems in Feature Selection and Data Compression

Alex Kokot

Co-Chairs of the Supervisory Committee:
Professor Marina Meila
Department of Statistics

Professor Alex Luedtke
Department of Statistics

Feature selection and data compression are fundamental problems in dimension reduction and representation learning. This dissertation develops a common variational framework for these tasks, posing them as quadratic optimization over spaces of probability distributions.

The first substantive chapter introduces the local expected gradient outer product (local EGOP) motivated by high-dimensional nonparametric regression. It is utilized in a recursive feature-learning algorithm which yields an intrinsic learning rate for signals parameterized by low-dimensional manifolds.

The second substantive chapter develops CO2 for selecting convexly weighted coresets to approximate measures with respect to smooth divergences. Second-order Hadamard differentiability turns a divergence's Hessian into a loss-adapted kernel, reducing local compression to maximum mean discrepancy minimization and kernel quadrature. For the Sinkhorn divergence, this geometry is equivalent to a scaled Gaussian reproducing kernel Hilbert space and supports coresets with poly-logarithmically many observations.

The final substantive chapter analyzes entropic self-transport as the regularization parameter ε tends to zero. A second-order expansion of the symmetric Schrödinger potential makes explicit its relationship to Gaussian-kernel estimators of density, score, and diffusion.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my family, in particular my mother who always dreamed of having a PhD, and instilled that dream in me.

The first time I met Marina, it was a departmental hike, and she arrived some time after me at the mountain. My group arrived at the summit, not without some ordeal, sliding on gravel and winding through narrow paths. Just as we prepared to take a well earned rest, with little hope of seeing the others soon, up comes Marina, at marching pace, skipping over boulders, unphased by the trek. I think this is exemplary of who she is as both a researcher and a person. Sure footed, knowing where she is going, always seeing the peak ahead while others are wading through the bramble. With her having led me to this summit, I feel my footing will be more clear on my next climb.

Alex has many students in the statistics department, and it is no surprise why. Somehow he treats each one of them as if they were his only student. Ok, maybe as if they were one of 3 of his students. The amount of time he spends every week going into nitty-gritty details on a smorgasbord of different research problems is something I find difficult to fathom. I am so grateful that he chose to take me in, to embrace the challenges I happily threw at him, and for the investment he gave of himself in every one of our research projects.

David Galvin inspired me, and I am sure countless other students, to be a mathematics major, to stay the course beyond honors calculus 1. I am indebted to many professors at Notre Dame, particularly to Misha Gekhtman and Sonja Mapes, and of course Liviu Nicolaescu who provided mentorship even into my early years as a PhD student.

Many great researchers at the University of Washington have shaped and guided my interests, particularly Soumik Pal, Stefan Steinerberger, and Zaid Harchaoui.

Finally, I would like to thank Anand Hemmady, Pawel Morzywolek, James Buenfil, Shreya Prakash, Erin Lipman, Elizabeth Maher, Anupreet Porwal, Seth Temple, Austin Sibu, Lars

van der Laan, Apara Venkat, Yikun Zhang, Jess Kunke, Antonio Olivas, Nina Galanter, Alana McGovern, and many more friends and colleagues who were a continual source of joy throughout the years.

Most of all, I would like to thank Vydhourie Thiyageswaran, who has been here with me every step of the way. There are not many people you could imagine surviving spending nearly every day with for five years, and even fewer you could not imagine having survived without. I am so lucky to have shared this time with you, every day I am better for it.

Chapter 1

INTRODUCTION

The problems of feature selection and data compression are intimately related, although they are not always presented as such. Feature selection is most commonly thought of in supervised contexts, including lasso penalization and other methods of model selection. Data compression refers to unsupervised data sparsification, such as through clustering or coreset selection. In statistical learning, however, both tasks ask a common question: which directions or observations are essential for a specified objective, and which can be discarded only to positive benefit?

We illustrate this with an example. Suppose one wished to select ten cafes from the many thousands in Seattle to best represent the city's coffee culture. Before selecting the cafes, one would likely identify descriptors such as ambience or bean selection to compare the different shops. The quality of the exemplars produced by this procedure would correspond to the capacity of these features to detect meaningful differences between the coffee shops.

We frame this problem variationally, realizing these tasks as optimization problems over probability distributions. For feature-learning we focus on Nadaraya-Watson style estimators which locally pool data. In the compression setting, sparse probability distributions supported on the observed samples. In both cases, the relevant loss is generally nonlinear in the measure. The main analytical strategy is to expand that nonlinear loss locally and identify the quadratic form that governs its leading behavior.

This quadratic perspective produces the main technical objects studied below: the expected gradient outer product (EGOP), which measures local variation of a regression target; the Hadamard operator, which is the Hessian of a smooth divergence on a space of probability measures; and the Schrödinger potential, whose small-regularization expansion exposes the local geometry of entropic optimal transport.

1.1 Selection as an optimization over measures

Let P denote a population distribution and

$$P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

its empirical distribution. A convexly weighted coreset of size m is a probability measure

$$P_m := \sum_{j=1}^{m} w_j \delta_{X_{i_j}}, \qquad w_j \geq 0, \qquad \sum_{j=1}^{m} w_j = 1,$$

supported on at most m of the observed points. Given a discrepancy $D(\mu, \nu)$, data compression seeks a small P_m that is nearly indistinguishable from P_n in the geometry relevant to D. For losses whose empirical fluctuation is of order n^{-1}, the desired relation is

$$D(P_m, P) = D(P_n, P) + o_p(n^{-1}).$$

Throughout the dissertation we denote by P a population law, P_n its empirical law, and P_m a coreset supported on at most m observations; μ and ν denote generic measures or localizations.

Now consider supervised data (X_i, Y_i) with

$$Y_i = f(X_i) + \epsilon_i, \qquad f(x) := \mathbb{E}[Y \mid X = x].$$

At a query point x, an anisotropic kernel estimator averages observations according to a positive-semidefinite metric M:

$$\widehat{f}_{M,h}(x) := \frac{\sum_{i=1}^{n} k\Big(\|M^{1/2}(x - X_i)\|/\sqrt{2h}\Big)\, Y_i}{\sum_{i=1}^{n} k\Big(\|M^{1/2}(x - X_i)\|/\sqrt{2h}\Big)}. \tag{1.1}$$

For a Gaussian kernel, this estimator has a continuum analogue that averages f under the localization

$$\mu_{x,M,h} := N(x, \Sigma), \qquad \Sigma = hM^{-1}.$$

Thus learning the metric M is equivalent to selecting a local probability measure. The covariance Σ determines which directions and how large a region the estimator may pool. Directions assigned large covariance are treated as locally redundant, while directions assigned small covariance are retained at higher resolution.

1.2 Quadratic Approximation

The recurring analytical step in this dissertation is to establish quadratic surrogates for the desired learning tasks. Let $\mathcal{J}$ be a functional defined on probability measures and let γ be a signed, zero-mass perturbation of P. Under an appropriate notion of differentiability,

$$\mathcal{J}(P + t\gamma) = \mathcal{J}(P) + t\,\dot{\mathcal{J}}_P(\gamma) + t^2\,\ddot{\mathcal{J}}_P(\gamma) + o(t^2). \tag{1.2}$$

For a divergence minimized at P, the first derivative vanishes. The leading local behavior is therefore the quadratic form $\ddot{\mathcal{J}}_P$. When this form is positive semidefinite, it may be represented by an operator G_P and a kernel g_P:

$$\ddot{\mathcal{J}}_P(\gamma) = \int (G_P\gamma)\,d\gamma = \iint g_P(x,y)\,d\gamma(x)d\gamma(y) =: \|\gamma\|^2_{G_P}. \tag{1.3}$$

The operator G_P is the Hadamard operator of the functional. It defines the tangent geometry in which two nearby distributions should be compared. If g_P is a reproducing kernel, then the final expression in Equation (1.3) is a squared maximum mean discrepancy (MMD). A nonlinear compression problem can consequently be reduced, locally, to kernel quadrature.

An analogous expansion governs local regression. If Z is centered with covariance Σ and concentrated near x, then

$$f(x+Z) - f(x) \approx \nabla f(x)^\top Z,$$

and hence

$$\mathbb{E}\big[(f(x+Z) - f(x))^2\big] \approx \nabla f(x)^\top \Sigma \nabla f(x) = \Big\langle \nabla f(x)\nabla f(x)^\top, \Sigma \Big\rangle_F.$$

After averaging over a localization μ, the rank-one matrix $\nabla f(x)\nabla f(x)^\top$ becomes the expected gradient outer product

$$\mathcal{L}(\mu) := \int \nabla f(z)\nabla f(z)^\top\,d\mu(z), \tag{1.4}$$

and the corresponding quadratic variation is

$$W(\mu) := \langle \mathcal{L}(\mu), \Sigma(\mu)\rangle_F = \int \nabla f(z)^\top \Sigma(\mu)\nabla f(z)\,d\mu(z). \tag{1.5}$$

We call W the EGOP form. Equations (1.3) and (1.5) are the two principal quadratic forms of the dissertation. The first measures a perturbation of a distribution in a loss-adapted

tangent geometry, the second measures a perturbation of the input in a response-adapted spatial geometry. Both identify directions in which variation is expensive and directions in which averaging is safe.

In regression, a kernel determines the local measure over which observations are pooled. In compression, a kernel embeds probability measures into a Hilbert space and determines which distributional perturbations must be preserved. In the final chapter, a Gaussian kernel is also the reference interaction in entropic transport, and its normalization produces a Markov operator. Kernel learning, kernel quadrature, and kernel diffusion are all central to this research, establishing geometry tailored to the estimation task at hand.

1.3 Dissertation outline

Below we detail the primary contributions of each of the main dissertation chapters.

1.3.1 Local Feature Learning as Distributional Optimization: Rates and Theoretical Foundations

The first substantive chapter studies feature learning for nonparametric regression. In high-dimensional problems, the ambient dimension D may substantially overstate the dimension along which the regression target varies. The classical multi-index model assumes

$$f(x) = g(Vx)$$

for a low-rank matrix V. The active directions are then globally fixed. We instead study a nonlinear extension called *continuous index learning* (CIL). Let $\mathcal{M} \subset \mathbb{R}^D$ be a smooth d-dimensional manifold and let

$$\pi(x) := \operatorname*{argmin}_{q \in \mathcal{M}} \|q - x\|$$

be the nearest-point projection within the reach of $\mathcal{M}$. The CIL model assumes

$$f(x) = g(\pi(x)). \tag{1.6}$$

Thus f may vary along a curved collection of tangent spaces while remaining invariant to normal displacement. A flat manifold recovers the multi-index model. This formulation

captures the local, spatially varying feature structure that motivates manifold learning and kernel steering.

The EGOP in (1.4) records that structure. Its leading eigenvectors identify directions in which f varies most strongly, while its null or near-null directions identify directions along which a local estimator may smooth aggressively. This suggests metrizing the kernel in Equation (1.1) by a local estimate of $\mathcal{L}(\mu)$. The empirical analogue is an average gradient outer product (AGOP),

$$\widehat{\mathcal{L}}(Q) := \int \nabla \widehat{f}(z) \nabla \widehat{f}(z)^\top \, dQ(z),$$

evaluated not over the entire empirical distribution, but over the same local kernel measure used for prediction.

This local construction is closely related to recursive feature machines and other gradient-based kernel-learning methods (Radhakrishnan, Beaglehole, et al., 2022; Beaglehole, Radhakrishnan, et al., 2023), but the distributional formulation makes the associated risk tradeoff explicit.

We formalize this geometric idea as follows. For $\mu = N(x, \Sigma)$, the Gaussian Poincaré inequality connects the integrated squared bias of a local average to $W(\mu)$. The finite-sample variance, in turn, is controlled by the volume of the smoothing region. At the level of rates, the resulting local mean integrated squared error (MISE) obeys

$$\mathbb{E}\left[\int (\widehat{f}_{M,h}(x) - f(z))^2 \, d\mu(z)\right] = O\left(\frac{1}{n\sqrt{\det \Sigma}} + W(\mu)\right). \tag{1.7}$$

The two terms expose the bias–variance tradeoff in distributional form. Making Σ small controls the EGOP form but discards observations; making Σ large increases the effective sample size but risks averaging across directions in which the target changes.

This tradeoff leads to a constrained maximum-entropy problem. For a prescribed bias scale $a > 0$, one would like to solve

$$\max_{\Sigma \succeq 0} \ \log\det \Sigma \qquad \text{subject to} \qquad \operatorname{tr}(\mathcal{L}(\mu)\Sigma) \le a. \tag{1.8}$$

Freezing the EGOP at the current localization leads to the explicit solution $\Sigma \propto \mathcal{L}(\mu)^{-1}$. Consequently, the Mahalanobis metric $M \propto \Sigma^{-1}$ is proportional to this matrix. Local EGOP

Learning iterates this relation,

$$\mu_t = N(x, \Sigma_t), \qquad \Sigma_{t+1} = h_{t+1}\big[\beta\mathcal{L}(\mu_t) + (1-\beta)\mathcal{L}(\mu_{t-1})\big]^{-1}, \tag{1.9}$$

1.3.2 Coreset Selection for the Sinkhorn Divergence and Generic Smooth Divergences

Let

$$D_P(Q) := D(Q, P)$$

be a divergence from the population P. The aim of this section is to approximate the empirical distribution with a sparse coreset. Many non-MMD losses may be costly to optimize directly over all sparse measures $P_m \ll P_n$. The proposed method, CO2 (Coresets of Order 2), uses the functional expansion (1.2) to construct a quadratic approximation tailored to D.

More precisely, suppose D_P is second-order Hadamard differentiable at P relative to a suitable tangent topology. Because $D_P(P) = 0$ and the first derivative vanishes at the null,

$$D_P(P + t\gamma) = t^2 \ddot{D}_P(\gamma) + o(t^2). \tag{1.10}$$

The Hadamard operator G_P represents $\ddot{D}_P$ as in Equation (1.3). Compression under its kernel therefore preserves the locally relevant part of the original divergence. In particular, if

$$\mathrm{MMD}_{g_P}(P_n, P_m) = o_p(n^{-1/2}),$$

then the compression error is negligible at the quadratic statistical scale, and

$$D_P(P_m) = D_P(P_n) + o_p(n^{-1}).$$

Algorithmically, a low-rank approximation to the Hadamard kernel can be combined with kernel recombination to construct a convexly weighted coreset. The required coreset size is controlled by the residual spectrum of the normalized Gram matrix. When the loss-induced kernel is well approximated at low rank, a small number of observations preserve the leading statistical behavior of the full empirical distribution.

The main application is the Sinkhorn divergence. For $\varepsilon > 0$, define entropically regularized quadratic transport by

$$\mathrm{OT}_\varepsilon(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int \|x - y\|^2 \, d\pi(x, y) + \varepsilon \, \mathrm{KL}(\pi \| \mu \otimes \nu) \right\}, \tag{1.11}$$

and

$$S_\varepsilon(\mu,\nu) := \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\nu,\nu). \tag{1.12}$$

Let

$$k_\varepsilon(x,y) := \exp\!\left(-\frac{\|x-y\|^2}{\varepsilon}\right)$$

and let a_P denote the positive self-transport scaling at P, characterized by

$$a_P = K_\varepsilon\!\left(\frac{dP}{a_P}\right), \qquad K_\varepsilon\gamma := \int k_\varepsilon(\,\cdot\,,y)\,d\gamma(y). \tag{1.13}$$

The base-point-scaled Gaussian kernel is

$$\widetilde{k}_{a_P}(x,y) := \frac{k_\varepsilon(x,y)}{a_P(x)a_P(y)}. \tag{1.14}$$

If $J_{a_P}\gamma := K_\varepsilon(d\gamma/a_P)$ and T_P is the self-transport integral operator, the Sinkhorn Hessian has the form

$$\ddot{S}_{\varepsilon,P}(\gamma) = \frac{\varepsilon}{2}\left\langle J_{a_P}\gamma, (I-\mathsf{T}_P^2)^{-1}J_{a_P}\gamma\right\rangle_{\mathcal{H}} \tag{1.15}$$

on the zero-mass tangent space. For distributions in $\mathbb{R}^d$, the eigenvalues of this analytic kernel decay sufficiently rapidly that taking

$$m = \omega(\log^d n)$$

is enough to construct a lossless Sinkhorn coreset. This result is representative of a general recipe for D-coreset construction. Differentiate the nonlinear loss, identify its local kernel geometry, approximate that kernel at low rank, and preserve the resulting quadratic form with a sparse measure.

1.3.3 Entropic Self-Transport in the Small-ε Limit: Scores, Densities, Diffusion Maps, and Fast Coresets

The final substantive chapter studies the same entropic self-transport object in a different asymptotic regime. The coreset chapter perturbs the marginal distribution while holding $\varepsilon > 0$ fixed. The final chapter instead holds the distribution fixed and lets $\varepsilon \downarrow 0$. We recover asymptotic results that asymptotically link entropic optimal transport to natural diffusion processes.

For self-transport of a measure μ, the optimal coupling has density

$$\xi_\varepsilon(x,y) := \frac{d\pi_\varepsilon}{d(\mu\otimes\mu)}(x,y) = \exp\left(\frac{f_\varepsilon(x)+f_\varepsilon(y)}{\varepsilon}\right) k_\varepsilon(x,y), \tag{1.16}$$

where the symmetric Schrödinger potential f_ε solves

$$f_\varepsilon(x) = -\varepsilon \log \int \exp(f_\varepsilon(y)/\varepsilon)\, k_\varepsilon(x,y)\, d\mu(y). \tag{1.17}$$

Suppose μ has density $p = h^2$ on $\mathbb{R}^d$, with a sufficiently smooth log-density. The main expansion is

$$f_\varepsilon = -\varepsilon\frac{d}{4}\log(\pi\varepsilon) - \varepsilon\log h - \varepsilon^2\frac{\Delta h}{8h} + o(\varepsilon^2). \tag{1.18}$$

To derive this, we reparameterize the Schrödinger equation using the heat semigroup

$$P_\varepsilon u(x) := (\pi\varepsilon)^{-d/2}\int e^{-\|x-y\|^2/\varepsilon} u(y)\, dy = e^{(\varepsilon/4)\Delta}u(x)$$

and apply an implicit-function argument to the h-transformed equation.

A key implication of these results is that self-EOT and Gaussian kernel estimators of density, score, and diffusion agree at leading order, and their first discrepancies can be written explicitly.

Chapter 2

LOCAL FEATURE LEARNING AS DISTRIBUTIONAL OPTIMIZATION: RATES AND THEORETICAL FOUNDATIONS

Abstract. We consider the setting of *Continuous Index Learning*, in which a function f of many variables varies only along a small number of directions at each point x. Any sample-efficient predictor of $f(x)$ must learn these directions, and a natural way to achieve this is via kernel regression, where the kernel shape adapts to the local variation of f. In our approach, we introduce the *local Expected Gradient Outer Product (local EGOP)*, a novel quantity that functions dually as an anisotropic metric and the inverse-covariance of a Poincaré-constrained maximum entropy distribution centered at x. This leads to a new framework connecting the Local Feature Learning, Recursive Feature Machine, and nonlinear dimension reduction literatures. The resulting *Local EGOP Learning* algorithm is a recursive kernel learning method that achieves a learning rate *adaptive* to the intrinsic regularity of the target function. We demonstrate some of its properties through controlled experiments and on real data.

2.1 Problem and approach

Learning representations adapted to structured data is a fundamental objective of modern machine learning, underpinning the success of methods from neural network architectures to manifold learning (Bengio et al., 2013; LeCun et al., 2015; Coifman and Lafon, 2006a; Meilă and Zhang, 2024). In many high-dimensional settings, including speech, images, and scientific measurements, data typically concentrates near a non-linear d-dimensional submanifold $\mathcal{M}$ (Tenenbaum et al., 2000; Das, Moll, Stamati, L. E. Kavraki, et al., 2006; Pope et al., 2021). Meaningful features are functions that capture variation along $\mathcal{M}$ while remaining invariant to orthogonal displacements. Classical techniques in image processing leverage differential information including gradients and Hessians to learn such local structure, capturing fine

details like edges and contours (Schmid and Mohr, 1997; Lowe, 1999; Wallraven et al., 2003; Takeda et al., 2007). In the kernel literature, this is known as *Local Feature Learning (LFL)*, a problem setting where kernel regressors are augmented with derivatives of the target function for the purpose of point estimation. The Recursive Feature Machine (RFM) algorithm is a recent method that falls under this umbrella, being proposed for nonparametric regression and as a theoretical model for neural network training dynamics (Radhakrishnan, Beaglehole, et al., 2022; Beaglehole, Radhakrishnan, et al., 2023; Zhu et al., 2025).

The goal of this chapter is to develop theoretical foundations for such algorithms, linking them directly to the optimization of certain regression related risk functionals such as the Mean Integrated Squared Error (MISE). This leads us to (1) the development of a new problem setting that generalizes multi-index learning, (2) a blueprint kernel learning method that achieves adaptive rates for such data, and (3) a new framework which relates these procedures to the optimization of Poincaré inequalities of central interest in nonlinear dimension reduction. We briefly explain our approach here in light of these three interrelated contributions.

Kernel learning problem Given n i.i.d. observations $(X_i, Y_i) \in \mathbb{R}^D \times \mathbb{R}$, we focus on the Nadaraya-Watson estimator of $f(x) = \mathbb{E}[Y|X = x]$ given by $\hat{f}(x) = \sum_i w_i Y_i$, with $w_i \propto k_h(x, X_i)$ (see (2.2)), with kernel k and bandwidth h. We focus on k Gaussian, with modifications for other kernels presented in Appendix 2.C.1.4.

We aim to learn *anisotropic* k_h, emulating the local structure of f through the Mahalanobis metrization $k_h(x, X_i) := k_{h,M}(x, X_i) = k(\|M^{1/2}(x - X_i)\|/\sqrt{2h})$. The goal is to learn an M that penalizes displacement in directions where f most strongly varies, to reduce bias, but remains wide in the directions where f is constant, to control variance.

(1) The data setting – Continuous Index Learning (CIL) This new setting is a generalization of multi-index learning (MIL). MIL is the task of estimating a function that only depends on x through composition with a low-rank matrix V, $f(x) = g(Vx)$. In CIL, we generalize this to the compositional problem $f(x) = g(\pi(x))$ where $\pi(x) := \text{argmin}_{q \in \mathcal{M}} \|q - x\|$ denotes the nearest point projection onto a latent manifold $\mathcal{M}$. MIL corresponds to the specific subsetting where $\mathcal{M}$ is a flat vector space. CIL is the natural setting for local kernel adaptation problems, see Figure 2.2 for an example of continuous single-index data.

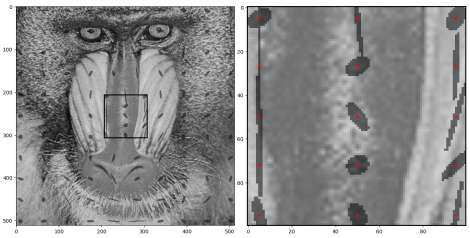


Figure 2.1: Local EGOP Learning localizations (Algorithm 1) on a grayscale mandrill image Bush (2021). Inputs x are pixel locations; $f(x)$ are grayscale values. We highlight the 125 highest-weight pixels (red) around specific center points (early stopping applied). Right: Magnified view of the boxed region.

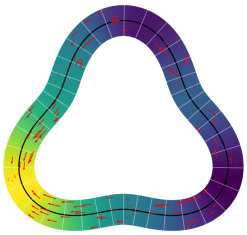


Figure 2.2: Continuous single-index data. Heatmap: a normally-invariant function about a closed curve (black). White lines indicate normal spaces. Red vectors denote sampled gradients (length $\propto$ magnitude).

(2) Recursive kernel learning – our blueprint To this end, we develop a recursive procedure we call *Local EGOP Learning*, which iteratively refines the metric M by pooling coarse differential information estimated from the observed data around a target point x. This procedure is intentionally similar to the Recursive Feature Machine (RFM) algorithm proposed for kernel ridge regression. Central in this line of research is the expected gradient outer product (EGOP), $\mathcal{L}(\mu) := \int \nabla f \nabla f^\top \, d\mu$. Its empirical counterpart is the average gradient outer product (AGOP) $\widehat{\mathcal{L}}(P_n) := \mathcal{L}_{\hat{f}}(P_n) := \frac{1}{n}\sum_{i=1}^n \nabla \hat{f}(x_i) \nabla \hat{f}^\top(x_i)$, with P_n the empirical distribution.

Our key insight is to (iteratively) *localize* the EGOP (or AGOP) around the point of interest x. At iteration t, let $\widehat{M}_t$ denote the normalized precision used by the kernel and set $w_i^{(t)} \propto \exp(-(x_i - x)^\top \widehat{M}_t (x_i - x))$. With $\hat{f}$ given by (2.2), we consider

$$\widehat{\mathcal{L}}_t := \widehat{\mathcal{L}}\left(\sum_{i=1}^n w_i^{(t)} \delta_{x_i}\right) = \sum_{i=1}^n w_i^{(t)} \nabla \hat{f}(x_i) \nabla \hat{f}(x_i)^\top. \tag{2.1}$$

That is, rather than averaging the gradient outer products over the full dataset, as has been recently emphasized, we *weigh* them with respect to the same kernel we utilize for regression.

After two initialization updates without momentum, we set $\widehat{M}_{t+1} = \beta \widehat{\mathcal{L}}_t + (1-\beta)\widehat{\mathcal{L}}_{t-1}$,

$\beta \in (0,1)$. Algorithm 1 normalizes $\widehat{M}_{t+1}$ before rescaling by the bandwidth. Corollary 2.4 shows that the corresponding population trace converges to a positive constant, so this scalar normalization does not affect the learning rate. Momentum improves stability, as discussed in Appendix 2.A.4. In Figure 2.1 we visualize the resulting regression neighborhoods. This method, described in Section 2.4, has close ties also to the heuristically motivated "kernel steering" algorithm (Takeda et al., 2007; Takeda et al., 2008); therefore, by providing formal guarantees for it, we illuminate the workings of these very successful precursors.

(3) Theoretical analysis framework We introduce a distributional interpretation of this recursive procedure through the Gaussian localizations $\mu_t = N(x, \Sigma_t)$, where $\Sigma_t \propto M_t^{-1}$ up to a covariance cap. $\int f \, d\mu_t$ is the population counterpart of $\hat{f}(x)$. Central to our analysis is a new Dirichlet functional that we call the *EGOP form*, $W(\mu_t) := \mathrm{tr}(\mathcal{L}(\mu_t)\Sigma(\mu_t))$, which naturally bounds the bias of Mahalanobis metrized kernel regression. This leads to an interpretation of Local EGOP Learning as a Poincaré constrained maximum entropy estimator, which we show achieves an intrinsic learning rate for the CIL setting.

In Section 2.3, we develop a framework for the study of anisotropic kernel regression, and an interpretation of EGOP metrization as greedy MISE optimization. In Section 2.4, Algorithm 1, we introduce the Local EGOP Learning algorithm in detail. In Section 2.5, we analyze this procedure as a matrix recurrence, demonstrating learning rates adaptive to the regularity of the target function. In Section 2.6 we highlight connections to recent methods in the compositional learning and nonlinear dimension reduction literature. In Section 2.7, we illustrate several significant aspects of the theory via numerical experiments, while Section 2.8 indicates avenues for future work.

2.2 Background and Related Work

For extended background material, refer to Appendix 2.A.

Kernel Learning: Kernel selection is a central task in the analysis of specialized data (Odone et al., 2005; Kondor and Jebara, 2003; Joachims, 1998; Chapelle et al., 1999; Barla et al., 2002; Vishwanathan et al., 2010), statistical optimization (Genton, 2001; Schölkopf, Simard, et al., 1997; Osborne, 2010), and distillation (Gong et al., 2024; Kokot and Luedtke,

2025). This setting is of broad interest to the development of machine learning theory (Belkin et al., 2018), and models like neural networks and random forests asymptotically correspond to tailored kernels (Jacot et al., 2018; Scornet, 2016).

Learning with Gradients: Classically, Local Feature Learning (Schmid and Mohr, 1997; Weickert, 1999; Lowe, 1999; Wallraven et al., 2003; Dalal and Triggs, 2005; Bay et al., 2006) and recursive partitioning (Heise, 1971; Breiman and Meisel, 1976; Friedman, 1979) have placed emphasis on leveraging information on the target function and its derivatives for representation learning. Our approach provides a rigorous basis for related "kernel steering" techniques in image processing (Takeda et al., 2007; Takeda et al., 2008). The EGOP has wide statistical applications (Samarov, 1993; Hristache et al., 2001; Xia et al., 2002; Trivedi et al., 2014; Yuan et al., 2023) and gradient metrizations appear in the setting of manifold learning (Q. Wu et al., 2010; Mukherjee et al., 2010). For supervised *noiseless* manifolds, Aswani et al. (2011) achieves intrinsic rates using isotropic kernels and sparse regression, penalizing normal and tangent spaces differently. However, their method requires estimating the local dimension d a priori and extends to infinite regressors, lacking our adaptive, local refinement. Recursive Feature Machines (Radhakrishnan, Beaglehole, et al., 2022; Beaglehole, Radhakrishnan, et al., 2023; Radhakrishnan, Belkin, et al., 2025; Zhu et al., 2025) is a recently proposed method that highlights the LFL framework. Our analysis reveals a fundamental connection to recent research in the nonlinear dimension reduction literature which approaches the problem of compositional learning through the optimization of Poincaré inequalities (Bigoni et al., 2022; Verdière et al., 2025; Nouy and Pasco, 2025).

Multi-Index Learning: Multi-index learning has frequently been studied through neural network adaptation (Mousavi-Hosseini et al., 2022; Boix-Adsera et al., 2023; Damian et al., 2023). In the single-index setting—where $f(x)$ depends solely on $v^\top x$—two-layer networks learn this dependence efficiently, rapidly improving prediction (Bietti et al., 2022; Abbe et al., 2024; J. D. Lee et al., 2024), with recent results extending to multi-index frameworks (Damian et al., 2023; Arnaboldi et al., 2024). The continuous index feature structure we adopt is standard in manifold learning (Aamari and Levrard, 2018; Genovese et al., 2012;

Kokot, Murad, et al., 2025). Continuous index models previously appeared as the "nonlinear single-variable model" of Y. Wu and Maggioni (2024) for 1D manifolds. In that paper, broad connections are made to compositional learning (Juditsky et al., 2009; Shen et al., 2021) and nonlinear dimension reduction (Yeh et al., 2008; K.-Y. Lee et al., 2013).

2.3 A Framework for Recursive Kernel Learning

In this section we develop a framework that relates kernel Mahalanobis metrization M_t, the covariance of the local sampling density μ_t, $\Sigma(\mu_t) := \mathrm{Cov}_{\mu_t}(X)$, and the EGOP $\mathcal{L}(\mu_t)$ to local Mean Integrated Squared Error (MISE) risk functionals. Central to this framework is the *EGOP form* $W(\mu) = \langle \mathcal{L}(\mu), \Sigma(\mu) \rangle_F$.

The idea is as follows. Since $\hat{f}(x)$ can be viewed as estimating the local average $\mathbb{E}_{X\sim\mu}[f(X)]$, its error can be bounded in terms of the variance of $f(X)$ under μ. This naturally ties the bias of this estimator to a Dirichlet form, and thus the local EGOP. We will leverage this to connect metrization by the EGOP to a greedy MISE reduction strategy.

2.3.1 Assumptions and Notation

We refer to the population distribution of our feature vectors $X \in \mathbb{R}^D$ as P, and its empirical distribution as P_n. We assume P has a bounded C^1 density p with respect to the Lebesgue measure on $\mathbb{R}^D$, bounded away from 0 at x the query point. In CIL, we denote by $\mathcal{M}$ the smooth base manifold of dimension d, by π the nearest point projection onto the manifold, and by $\mathcal{T}_x$ and $\mathcal{N}_x$ the tangent and normal spaces at point $\pi(x) \in \mathcal{M}$. For π to be well-defined, we further assume that X lies within the reach of $\mathcal{M}$ (Federer, 1959) almost surely.

The labels Y are assumed to be of the form $Y_i = f(X_i) + \varepsilon_i$ for ε_i i.i.d., mean zero, and finite 4th moment. We assume that $f \in C^4$ and is bounded with uniformly bounded derivatives. We reserve μ to refer to a localization or target distribution of data appropriately concentrated about the point of interest x. We denote by Σ the covariance of μ, and we enforce $\|\Sigma\| \leq \zeta < \infty$. We denote by $\langle \cdot, \cdot \rangle_F$ the Frobenius inner product, and by k a second-order kernel that satisfies typical assumptions in the nonparametric statistics literature, as in Tsybakov, 2009[Chapter 1]. For convenience, we use a Gaussian kernel for k, although modifications for the general setting are presented in Appendix 2.C.1.4. We assume

$g = \nabla f(x) \neq 0$. Finally, we define the empirical kernel estimator for predicting f at x and its continuous counterpart by:

$$\hat{f}(x) \equiv \hat{P}_{M_t}(f) := \frac{1}{\hat{C}_{M_t} n} \sum_{i=1}^{n} k_{h_t, M_t}(x, X_i) Y_i, \qquad \hat{C}_{M_t} := \frac{1}{n} \sum_{i=1}^{n} k_{h_t, M_t}(x, X_i), \tag{2.2}$$

$$P_{M_t}(f) := \frac{1}{C_{M_t}} \int k_{h_t, M_t}(x, y) f(y)\, dP(y), \quad C_{M_t} := \int k_{h_t, M_t}(x, y)\, dP(y). \tag{2.3}$$

The bandwidth h is suppressed in our notation as it will be pre-set at each iteration of our algorithm.

2.3.2 The EGOP and MISE

To assess the quality of predictor (2.3) on μ concentrated about a query point x, we introduce a Dirichlet form we call the EGOP form $W(\mu)$.

$$W(\mu) := \int \nabla f^\top \Sigma(\mu) \nabla f\, d\mu = \operatorname{tr}(\mathcal{L}(\mu)\Sigma(\mu)) = \langle \mathcal{L}(\mu), \Sigma(\mu) \rangle_F. \tag{2.4}$$

Identifying the EGOP form is a central contribution of this chapter, as we relate $W(\mu)$ to the bias of Mahalanobis metrized kernel regressors relative to a local MISE functional.

Lemma 2.1 (Gaussian-smoother bias bound)**.** *Let $\|h_t M_t^{-1}\|$ be uniformly bounded, and set $\mu_t = N(x, h_t M_t^{-1})$. Then, there exists a universal constant C depending only on x, P, such that*

$$\int (P_{M_t}(f) - f)^2\, d\mu_t \leq C W(\mu_t). \tag{2.5}$$

This bound follows from the Gaussian Poincaré inequality, and has the following natural interpretation. The matrix $\mathcal{L}(\mu)$ encodes the variation of f, and $\Sigma(\mu)$ the directions in which data is pooled for our predictor. Thus, when $\mathcal{L}(\mu)$ and $\Sigma(\mu)$ are misaligned, such that $\langle \mathcal{L}(\mu), \Sigma(\mu) \rangle_F = W(\mu)$ is small, then f is smooth with respect to μ. If data is pooled in directions of minimal variation of f, then the squared residual from the average is reduced. We note that this general structure, of metrizing by the inverse covariance, appears in disparate literatures, from classical Local Feature estimators (Schmid and Mohr, 1997), to VAEs (Chadebec and Allassonnière, 2022), and Langevin preconditioning (Cui et al., 2024).

We next seek to control the variance of the finite sample estimator $\hat{P}_M$ of f at x.

Lemma 2.2 (Kernel-regression variance bound)**.** *Let $M_t \succ 0$, $\Sigma_t = h_t M_t^{-1}$, and $\|\Sigma_t\|$ be uniformly bounded. Then, there exists a universal constant C depending only on $x, P, \|f\|_\infty, \mathbb{E}[\varepsilon^4]$ such that*

$$\mathbb{E}\left[(\hat{P}_{M_t}(f) - P_{M_t}(f))^2\right] \leq \frac{C}{n\sqrt{\det \Sigma_t}}. \tag{2.6}$$

Thus the discretization error is monotone in the entropy of the Gaussian smoother, $H(\mu_t) = \log\det(\Sigma_t)$, encouraging averaging over larger pools of data. We build on this perspective in Section 2.3.3. Combining these bounds yields our fundamental error rate.

Theorem 2.1 (Local MISE bound and dimension-adaptive rate)**.** *Let $M_t \succ 0$, and set $\Sigma_t = h_t M_t^{-1}$, $\mu_t = N(x, \Sigma_t)$, and $\|\Sigma_t\|$ be uniformly bounded. Then,*

$$\mathbb{E}\left[\int (\hat{P}_{M_t}(f) - f)^2 \, d\mu_t\right] = O\left(\frac{1}{n\sqrt{\det \Sigma_t}} + W(\mu_t)\right). \tag{2.7}$$

In particular, if $W(\mu_t) = O(h_t)$ and $\det \Sigma_t = \Omega(h_t^q)$ then the asymptotic learning rate is

$$\mathbb{E}\left[\int (\hat{P}_{M_t}(f) - f)^2 \, d\mu_t\right] = O\left(n^{-\frac{1}{1+q/2}}\right). \tag{2.8}$$

2.3.3 Recursive Kernel Learning as Constrained Entropy Maximization

We now show how EGOP based recursive kernel learning, combined with our basic kernel bounds, constitutes a greedy variance reduction strategy. Intuitively, the goal is to induce anisotropy in the kernel, stretching or steering it to include additional low-bias points that would otherwise be discarded with an isotropic metric. The inclusion of these additional high-quality points is exactly a reduction of variance, as including more data results in tighter concentration to the mean for a local averaging estimator.

To optimize the bound of Theorem 2.1, we would like to select μ_t such that $W(\mu_t) \to 0$ with entropy $\log\det \Sigma(\mu_t)$ as large as possible. In other words, if we specify the rate $W(\mu_t) = \mathrm{tr}(\mathcal{L}(\mu_t)\Sigma(\mu_t)) = h_t \to 0$, it would be optimal to smooth over the largest possible region that achieves this bias, meaning we would like to select

$$\mu_t = \mathrm{argmax}_{\mathrm{tr}(\mathcal{L}(\mu_t)\Sigma(\mu_t)) \leq h_t} \log\det \Sigma(\mu_t). \tag{2.9}$$

This, however, cannot be solved, as we lack precise knowledge of the target function f. Moreover, $\mathrm{tr}(\mathcal{L}(\mu_t)\Sigma(\mu_t)) \leq h_t$ presents a challenging feasible set, as it is posed over a space

of probability distributions, thus making it non-Euclidean. Our Gaussian ansatz reduces the problem to optimizing a finite parameterization, though the optimization is non-convex.

For $A \succ 0$, the finite-dimensional maximum-entropy problem has solution

$$\Sigma_h^\circ(A) := \operatorname{argmax}_{\Sigma \succ 0,\ \operatorname{tr}(A\Sigma) \le h} \log\det \Sigma = \frac{h}{D} A^{-1}. \tag{2.10}$$

We impose the covariance cap only after this update. For $B \succeq 0$ with eigendecomposition $B = U \operatorname{diag}(b_1, \dots, b_D) U^\top$, define the spectral truncation

$$\operatorname{Cap}_\zeta(B) := U \operatorname{diag}(\min\{b_1, \zeta\}, \dots, \min\{b_D, \zeta\}) U^\top. \tag{2.11}$$

For a possibly singular $A \succeq 0$, define the truncated inverse map

$$\mathcal{T}_{h,\zeta}(A) := \lim_{\varepsilon \downarrow 0} \operatorname{Cap}_\zeta\left(\frac{h}{D}(A + \varepsilon I)^{-1}\right). \tag{2.12}$$

This operation is empirically similar to the numerically stabilized inversion $(A + \zeta I)^{-1}$ which is common in the literature, although we elect for exact capping for theoretical convenience.

Proposition 2.1 (Truncated inverse-EGOP metric update)**.** *During initialization $\widehat{M}_{t+1} = \widehat{\mathcal{L}}_t$, while after momentum is introduced $\widehat{M}_{t+1} = \beta \widehat{\mathcal{L}}_t + (1-\beta)\widehat{\mathcal{L}}_{t-1}$. The associated covariance is*

$$\Sigma_{t+1} = \mathcal{T}_{h_{t+1},\zeta}(\widehat{M}_{t+1}). \tag{2.13}$$

If $\widehat{M}_{t+1} = U \operatorname{diag}(\lambda_1, \dots, \lambda_D) U^\top$, then

$$\Sigma_{t+1} = U \operatorname{diag}(\tau_1, \dots, \tau_D) U^\top, \qquad \tau_j = \begin{cases} \min\{\zeta, h_{t+1}/(D\lambda_j)\}, & \lambda_j > 0, \\ \zeta, & \lambda_j = 0. \end{cases} \tag{2.14}$$

Hence the update agrees with $h_{t+1}\widehat{M}_{t+1}^{-1}/D$ on every untruncated spectral block, while null and sufficiently small-EGOP directions are set to the cap.

When implemented in Algorithm 1, we use local linear regression to better fit into the EGOP estimation loop, and we develop corresponding theory in Appendix 2.D.

2.4 The Local EGOP Algorithm

In this section, we introduce Local EGOP Learning (Algorithm 1), which implements the recursive inverse-EGOP metrization motivated by Section 2.3.3. We emphasize this method for its simplicity, although it can be optimized, as discussed in Appendix 2.B.

The algorithm computes base weights with respect to the metric $\widehat{M}_t$ (Step 3) to sample points where the gradient will be estimated (Step 4). For each selected point, new local weights are computed to fit a Leave-One-Out (LOO) local linear regression, yielding a local gradient and function estimate (Steps 5–6). The resulting gradient outer products are averaged to form $\widehat{\mathcal{L}}_t$ (Step 7). After two initialization updates we add momentum $\widehat{M}_{t+1} = \beta\widehat{\mathcal{L}}_t + (1-\beta)\widehat{\mathcal{L}}_{t-1}$, $\beta \in (0,1)$ and rescale by the bandwidth. In practice, hyper-parameters α, m, β, T, and the bandwidth schedule $\{h_t\}$ are chosen via cross-validation. While the theoretically optimal iteration count is dimension-dependent (Theorems 2.2 and 2.3), Algorithm 1 incorporates the LOO validation loop directly into its execution (Steps 7–9) to provide a purely data-driven early stopping criterion. Further theoretical motivation is detailed in Section 2.3.3.

The time complexity of Algorithm 1 is $O(TmnD^2 + TmD^3)$, with memory overhead $O(nD + D^2)$.

2.5 Theoretical Analysis of Local EGOP Learning

For the two initialization steps, set $\mu_0 = N(x^*, \alpha I)$, $M_1 = \mathcal{L}(\mu_0)$, and $M_2 = \mathcal{L}(\mu_1)$. For $t \geq 2$, define the population momentum metric

$$M_{t+1} := \beta\mathcal{L}(\mu_t) + (1-\beta)\mathcal{L}(\mu_{t-1}). \tag{2.15}$$

The oracle covariance recurrence is

$$\Sigma_{t+1} = \mathcal{T}_{h_{t+1},\zeta}(M_{t+1}). \tag{2.16}$$

On every spectral block where truncation is inactive,

$$\Sigma_{t+1} = \frac{h_{t+1}}{D} M_{t+1}^{-1}. \tag{2.17}$$

We absorb the fixed factor D^{-1} into the bandwidth notation in the rate analysis. The following Taylor expansion identifies the gradient and Hessian components that determine the covariance scales.

Lemma 2.3 (Local EGOP Expansion)**.** *Let $\mu = N(x^*, \Sigma)$, $g = \nabla f(x^*)$, and $H = \nabla^2 f(x^*)$. There exist a linear map $T : \mathbb{R}^{D\times D} \to \mathbb{R}^D$ and a map $R : \mathbb{R}^{D\times D} \to \mathbb{R}^{D\times D}$ such that, for every $A \succeq 0$,*

$$\|T(A)\| \le C_T\|A\|, \qquad \|R(A)\| \le C_R\|A\|^2, \tag{2.18}$$

$$\mathcal{L}(\mu) = gg^\top + H\Sigma H + gT(\Sigma)^\top + T(\Sigma)g^\top + R(\Sigma). \tag{2.19}$$

We next compile the following geometric facts key to realizing an intrinsic learning rate.

Lemma 2.4 (Invariant directions and projected Hessian structure in CIL)**.** *Let $x^* = p + \nu$, where $p \in \mathcal{M}$ and $\nu \in \mathcal{N}_p$, and set $g = \nabla f(x^*)$ and $Q = \pi_g^\perp$. Then*

$$\nabla f(p + \nu) = (I - S_p(\nu))^{-1}\nabla f(p). \tag{2.20}$$

There exists a subspace $\mathrm{Null} \subseteq g^\perp$ *of dimension at least $D - 2d$ along which the gradient is constant and which the Hessian annihilates. If* $\mathrm{rank}(QHQ) = 2d - 2$*, define*

$$\mathcal{E}_x = \mathrm{range}(QHQ), \qquad \mathcal{K}_x = \ker\big((QHQ)|_{g^\perp}\big). \tag{2.21}$$

Then $\dim(\mathcal{E}_x) = 2d - 2$, $\dim(\mathcal{K}_x) = D - 2d + 1$, *$QHQ$ is invertible on $\mathcal{E}_x$, and* $\mathrm{Null} \subseteq \mathcal{K}_x$*. The space $\mathcal{K}_x$ contains at most one additional direction; along this full space the gradient may change in magnitude but not direction. More precisely, for sufficiently local $v \in \mathcal{K}_x$,*

$$\nabla f(x^* + v) \in \mathrm{span}(g), \qquad Q\nabla^2 f(x^* + v)\pi_{\mathcal{K}_x} = 0. \tag{2.22}$$

Theorem 2.2 (Dimension-adaptive Local EGOP rate)**.** *Let $k := D - 2d + 1$. Assume that $D \ge 2d$,* $\mathrm{rank}(\pi_g^\perp H \pi_g^\perp) = 2d - 2$*, and the local CIL condition of Section 2.C.3.1. For $n \ge 3$, set*

$$\zeta_n := \frac{c_\zeta}{\log n}, \qquad \alpha_n := c_\alpha\zeta_n, \qquad h_1 := \alpha_n^2, \qquad \sqrt{h_t} := \alpha_n r^{t-1}, \quad t \ge 1,$$

where $0 < r < 1$, $0 < \beta < 1$, and $c_\zeta, c_\alpha > 0$ are sufficiently small constants. Define

$$b_n := \left(n\zeta_n^{k/2}\right)^{-2/(d+2)}, \qquad t_n := \min\{t \ge 2 : h_t \le b_n\}.$$

Then $h_{t_n} \asymp b_n$, and, uniformly for $2 \le t \le t_n$,

$$g^\top\Sigma_t g = \Theta(h_t), \qquad v^\top\Sigma_t v = \Theta(\zeta_n) \quad \textit{for unit } v \in \mathcal{K}_x,$$

and

$$e^\top \Sigma_t e = \Theta(\sqrt{h_t}) \quad \textit{for unit } e \in \mathcal{E}_x.$$

Moreover,

$$\sqrt{\det \Sigma_t} = \Theta\left(h_t^{d/2} \zeta_n^{k/2}\right), \qquad W(\mu_t) = O(h_t),$$

and

$$\mathbb{E}\left[\int (\hat{P}_{M_{t_n}}(f) - f)^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)} \zeta_n^{-k/(d+2)}\right).$$

Consequently,

$$\mathbb{E}\left[\int (\hat{P}_{M_{t_n}}(f) - f)^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)} (\log n)^{(D-2d+1)/(d+2)}\right).$$

The covariance therefore has one direction of order h_t, $2d-2$ directions of order $\sqrt{h_t}$, and $D-2d+1$ directions at the shrinking cap. The logarithmic factor is the variance cost of forcing the Gaussian localization into the local CIL tube.

When data is not CIL, and the Hessian is full rank, our guarantee remains sharp. Compared to an isotropic Nadaraya-Watson estimator, our method nearly squares the asymptotic error of the local MISE, effectively cutting the effective dimension in half for D large.

Theorem 2.3 (Full-rank projected-Hessian anisotropy and rate)**.** *Let* $\operatorname{rank}(\pi_g^\perp H \pi_g^\perp) = D-1$ *and* Σ_t *satisfy the recurrence in equation 2.16,* $\sqrt{h_t} = \alpha r^t$, $0 < r < 1$, *and* $\beta \in (0,1)$. *Then, for* α *sufficiently small,* $g^\top \Sigma_t g = \Theta(h_t)$, *and for* $v \perp g$, $v^\top \Sigma_t v = \Theta(\sqrt{h_t})$. *It follows that, selecting* $t_n = \frac{2 \log n}{(D+5)\log(1/r)}$,

$$\mathbb{E}\left[\int (\hat{P}_{M_{t_n}}(f) - f)^2 \, d\mu_{t_n}\right] = O\left(n^{-\frac{4}{D+5}}\right). \tag{2.23}$$

2.6 Connection to the Nouy and Pasco (2025) Nonlinear Dimension Reduction

Our results connect LFL to methods in compositional learning and nonlinear dimension reduction.

RFM procedures are similar to Algorithm 1, but typically use kernel ridge regression for both the predictor $\hat{f} = \hat{f}_{\text{ridge}}$ and the (global) AGOP; the latter is averaged uniformly over the full dataset, yielding $M = \widehat{\mathcal{L}}_{\hat{f}_{\text{ridge}}}(P_n)$. Other minor implementation differences include a

preference for Laplace kernels and, recently, adjustments to the metrization scheme (Zhu et al., 2025).

Nouy and Pasco (2025) compose a class of predictors $\mathcal{H}$ with a non-linear mapping $g : \mathbb{R}^D \to \mathbb{R}^d$. They bound the L^2 bias of this estimator as a function of g, $\mathcal{E}(g) = \min_{h\in\mathcal{H}} \mathbb{E}[|f(X) - h(g(X))|^2]$, via a Poincaré inequality. Specifically, $\mathcal{E}(g) \leq C\mathcal{J}(g)$, where C depends on the data density and $\mathcal{H}$, $\Pi^{\perp}_{Dg(X)}$ denotes the projection onto the orthogonal complement of the Jacobian of g, and $\mathcal{J}(g)$ is the projected energy functional $\mathcal{J}(g) = \mathbb{E}\left[\|\Pi^{\perp}_{Dg(X)}\nabla f(X)\|_2^2\right]$.

In Lemma 2.1 and Proposition 2.1 we formally linked EGOP metrization to MISE type risk functionals via the optimization of Poincaré inequalities. We now make explicit the connection to the regret optimization of Nouy and Pasco, 2025.

First, note that evaluating a Mahalanobis kernel $k(\|M^{1/2}(x - y)\|)$ is equivalent to precomposing an isotropic kernel k with the linear map $g(x) = M^{1/2}x$. Because the Jacobian of this map is exactly $M^{1/2}$, minimizing $\mathcal{J}(g)$ is achieved precisely when M spans the principal directions of the EGOP $\mathcal{L}(P)$. Thus, by performing kernel ridge regression on these linearly transformed features, RFM implicitly seeks to construct an optimal Poincaré dimension reduction.

Second, local EGOP naturally extends this linear optimization to a pointwise objective. While our primary focus is local, we briefly explore global bounds. We recast the deterministic map $g(X)$ as a stochastic feature encoder, defined via the conditional Gaussian $Z(X) \mid X \sim \mathcal{N}(X, \Sigma(X))$. Under this formulation, the minimal L^2 risk is the expected conditional variance of the target function:

$$\inf_h \mathbb{E}_{X,Z}\left[(f(X) - h(Z(X)))^2\right] = \mathbb{E}_Z\left[\mathrm{Var}(f(X) \mid Z(X))\right]. \tag{2.24}$$

This can be further upperbounded via Gaussian type Poincaré inequalities in settings where $\|\Sigma\| \to 0$, or X is log-concave.

2.7 Experiments

The following experiments illustrate various aspects of our theoretical results. Details for data generation, evaluation, and neural network training are provided in Appendix 2.B.

The importance of intrinsic rates: We compare Local EGOP Learning with two-layer neural networks on helical data; this data has a CIL structure with $d = 1$, but a global MIL structure with $d = 3$ for $D \geq 3$ odd, $d = 2$ for D even. This data closely approximates a manifold structure, which is a standard setting for accelerated learning rates (Kiani et al., 2024; Liu et al., 2021). Figure 2.4 (left) shows that many neural network architectures efficiently recover the MIL structure (in dimension $D = 5$) but, unlike Local EGOP Learning, they do not recover the CIL structure, being vastly outperformed in moderate sample settings. This demonstrates the advantage of intrinsic rates in even the simplest toy problems.

Ambient Dimension (In)dependence: We further test Local EGOP Learning across varying ambient dimensions D to isolate the impact of the ambient space on the algorithm's performance. As demonstrated in Figure 2.4 (right), the asymptotic learning rate is largely independent of D. Some shifts in error scaling between even and odd D can be seen; they are due purely to structural changes in the helix parameterization.

Feature Learning as a necessity: We compare the feature learning capabilities of Local EGOP to a deep transformer (Gorishniy, Rubachev, and Babenko, 2023) on 1-sphere CIL data. As visualized previously in Figure 2.3, both algorithms achieve similar and nearly complete data denoising. This experiment illustrates a further aspect that we have omitted for lack of space: irrespective of the representation and learning modality, some form of "kernel adaptation" is necessary for statistically efficiency in the CIL setting. We discuss the role of depth in Appendix 2.A.3, as well as unsupervised embeddings in Appendix 2.B.

Molecular Dynamics (MD), a high noise approximately CIL data set We apply our method to molecular geometries, where interatomic interactions constrain atomic coordinates near a low-dimensional manifold (Das, Moll, Stamati, L. Kavraki, et al., 2006). For malonaldehyde, the manifold has a $d = 2$ torus topology, with *backbone angles* $\tau_{1,2}$ varying along it (see Figure 2.11b). The data distribution around the torus is anisotropic, multiscale, and with significant variance. We take the target function to be τ_1, which can be computed for each configuration and is known to approximate the CIL assumption; we utilize

$n = 10^4$ MD configurations from Chmiela et al. (2017), pre-processed to $D = 50$ dimensions (Koelle et al., 2022). Evaluated on 500 hold-out points, Local EGOP Learning achieves an MSE of 0.00043, significantly outperforming a Gaussian Nadaraya-Watson estimator with cross-validated bandwidth (MSE 0.012).

2.8 Discussion, Limitations, and Future Work

High-dimensional nonparametric regression classically relies on strict global assumptions including low-rank manifold and multi-index models. We relax these constraints by formalizing Continuous Index Learning (CIL) and constructing an estimator that adapts to local function regularity. Our approach mirrors existing empirical methods, grounding them in rigorous theory. We establish a functional framework that links the EGOP directly to the MISE of kernel regressors, and connects Local Feature Learning (LFL), Recursive Feature Machines (RFM), and compositional learning. Our primary insight is that recursive EGOP metrization acts as a greedy constrained entropy maximization on the Bures-Wasserstein manifold. This provides a novel theoretical foundation for gradient and EGOP estimation techniques used across image processing (Takeda et al., 2007; Takeda et al., 2008), machine learning (Radhakrishnan, Beaglehole, et al., 2022; Beaglehole, Radhakrishnan, et al., 2023), and dimension reduction (Mukherjee et al., 2010; Q. Wu et al., 2010; Trivedi et al., 2014; Yuan et al., 2023; Radhakrishnan, Belkin, et al., 2025), yielding new efficiency guarantees for point estimation and tractable access to local EGOP metrizations.

Local EGOP Learning achieves an intrinsic rate for CIL data without requiring a priori knowledge of problem dimensionality, noise levels, or manifold regularity. However, for generic manifolds, it does not completely denoise the covariate distribution: a small orthogonal subspace contracts at the tangent rate, despite the invariance of f. This artifact arises because our optimization scheme minimizes a bias upper bound rather than the exact bias, leaving its refinement as an open challenge.

Algorithm 1 prioritizes clarity over efficiency; computational optimizations, such as partitioning schemes (Beaglehole, Holzmüller, et al., 2025), are left for future work. Furthermore, its derivation from an MSE formulation restricts it to ℓ_2 loss, requiring structural modifications for tasks like classification or image generation.

Finally, our methodology relies on a second-order Gaussian relaxation. Improved guarantees may be possible by treating $W(\mu)$ and $\log\det(\Sigma(\mu))$ as functionals and optimizing the MISE via alternative procedures such as Wasserstein gradient flows (Ambrosio et al., 2006). This would enable non-ellipsoidal localizations to better accommodate curved level-sets, provided the fundamental hurdle of the EGOP form acting only as a loose bias upper bound can be overcome.

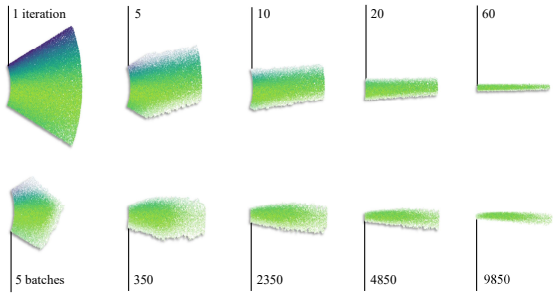


Figure 2.3: Local features fit to annulus data: Local EGOP (top) vs. deep transformer embedding (bottom). Heatmap: normally-invariant labels. About $x = (1, 0)$, opacity $w \propto \exp(-d(x, x_i)^2)$ maps AGOP Mahalanobis and transformer embedding distances across progressive iterations.

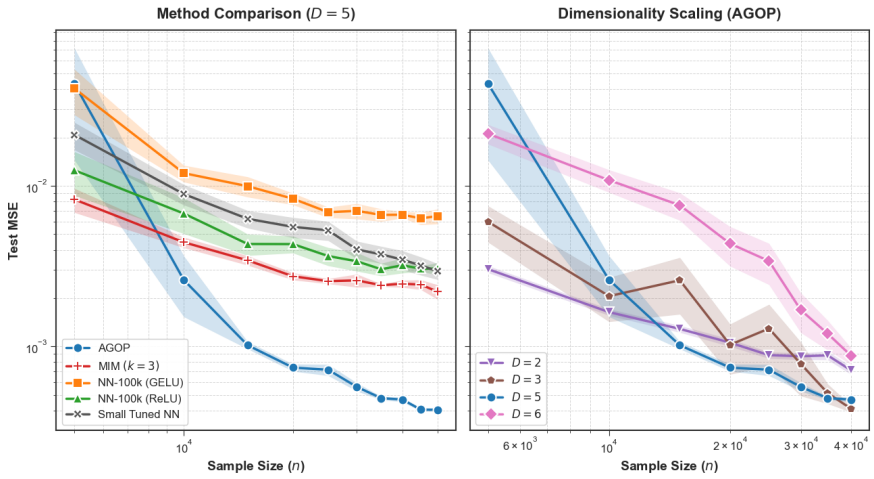


Figure 2.4: Test MSE on helical data. Left: Local EGOP vs. two-layer neural networks in ambient dimension $D = 5$. Right: Local EGOP performance across varying ambient dimensions D.

Algorithm 1 Local EGOP Learning

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, target x, subsample size m, iterations T, bandwidths $\{h_t\}$, $\beta, \alpha > 0$.

Output: Best estimate $\hat{f}_{\text{best}}(x)$.

Initialize $\widehat{M}_0 \leftarrow I/\alpha, \quad \{h_t\} \leftarrow \{\alpha h_t\}, \quad E_{\text{best}} \leftarrow \infty, \quad \hat{f}_{\text{best}}(x) \leftarrow \varnothing.$

for $t = 0$ **to** $T - 1$ **do**

  Compute base weights $w_i \propto \exp\big(-(x_i - x)^\top \widehat{M}_t (x_i - x)\big)$ and normalize.

  Sample $I_1, \ldots, I_m$ independently with $\mathbb{P}(I_\ell = i) = w_i$.

  for $\ell = 1, \ldots, m$, set $i = I_\ell$ and compute $W_{ij} \propto \exp\big(-(x_j - x_i)^\top \widehat{M}_t (x_j - x_i)\big)$ for all $j \neq i$, and normalize.

    Fit local linear regression to obtain gradient $\nabla \hat{f}[i]$ and LOO estimate $\hat{f}[i] = \sum_{j \neq i} W_{ij} y_j$.

  $\widehat{\mathcal{L}}_t \leftarrow \frac{1}{m} \sum_{\ell=1}^m \nabla \hat{f}[I_\ell] \nabla \hat{f}[I_\ell]^\top, \quad \widehat{\text{MSE}}_t \leftarrow \frac{1}{m} \sum_{\ell=1}^m (y_{I_\ell} - \hat{f}[I_\ell])^2.$

  if $\widehat{\text{MSE}}_t < E_{\text{best}}$ **then**

    $E_{\text{best}} \leftarrow \widehat{\text{MSE}}_t, \quad \hat{f}_{\text{best}}(x) \leftarrow \sum_{i=1}^n w_i y_i$

  end if

  if $t < 2$ **then**

    $\widehat{M}_{t+1} \leftarrow \widehat{\mathcal{L}}_t$ ▷ *Initialize without momentum*

  else

    $\widehat{M}_{t+1} \leftarrow \beta \widehat{\mathcal{L}}_t + (1 - \beta) \widehat{\mathcal{L}}_{t-1}$

  end if

  $\widehat{M}_{t+1} \leftarrow \widehat{M}_{t+1} / [h_{t+1} \operatorname{tr}(\widehat{M}_{t+1})]$ ▷ *Normalize kernel precision*

end for

return $\hat{f}_{\text{best}}(x)$

Local Feature Learning as Distributional Optimization: Rates and Theoretical Foundations

Supplementary Material

Appendix Table of Contents

2.A Extended Background and Problem Setting

In this section, we provide extended geometric intuition of the Local EGOP Learning algorithm and other metrization schemes. In section 2.A.5, we display representative figures of the learning dynamics of each algorithm.

2.A.1 Recursive Feature Machines (Radhakrishnan, Belkin, et al., 2025)

As a preliminary, we discuss the setting of multi-index learning and its relationship to the Recursive Feature Machine algorithm.

The EGOP matrix $\mathcal{L}(P) = \int \nabla f \nabla f^\top \, dP$ has natural motivation. As a Mahalanobis metric, of central interest is the quadratic form $v^\top \mathcal{L}(P) v = \mathbb{E}[(\partial_v f)^2]$, where $\partial_v f = \nabla f^\top v$ denotes the directional derivative. This matrix directly encodes the magnitude by which f varies on the data distribution. When f is multi-index, meaning that there exists a vector space U such that, for all $x \in \mathbb{R}^D$, all $u \in U$, $f(x+u) = f(x)$, it follows that $\partial_u f \equiv 0$ holds identically. Thus the nullspace of $\mathcal{L}(P)$ are exactly these uninformative indices. When we look at a kernel such as $\exp(-[x-y]^\top \mathcal{L}(P)[x-y])$, the metrization can be viewed as the standard Gaussian kernel composed with the linear map $\mathcal{L}(P)^{1/2}$, hence this acts as an implicit dimension reduction, stripping away those directions in which f does not vary.

In a real data setting we do not have access to the matrix $\mathcal{L}(P)$, hence this metric must be learned concurrently with our predictor $\hat{f}$. This motivates a *recursive* algorithm, where an initial predictor is fit generically, and this baseline information is used to form an estimate of the EGOP $\widehat{\mathcal{L}}(P) = \int \nabla \hat{f} \nabla \hat{f}^\top \, dP$, or typically the empirical AGOP matrix $\widehat{\mathcal{L}}(P_n) = \frac{1}{n} \sum_{i=1}^n \nabla \hat{f}(X_i) \nabla \hat{f}(X_i)^\top$. Then, this initial metric can be used to adapt the regressor via Mahalanobis metrization, improving the prediction quality, and the accuracy of the estimated gradients. This procedure is repeated iteratively, terminating either by a convergence criterion or a cross-validation loop as suggested by our approach. We summarize these steps in Figure 2.5.

Empirically, we can observe that this procedure is quite effective for this learning setting. As illustrated in Figure 2.9a (provided in Section 2.A.5), we demonstrate a simple simulation where data is generated in $\mathbb{R}^2$, depicting a heatmap of a single-index function that only

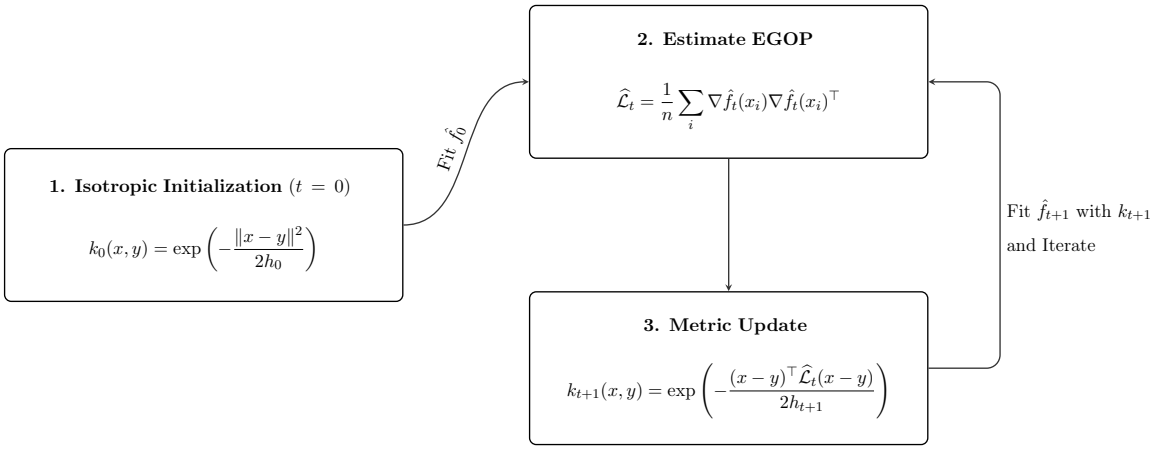


Figure 2.5: Flowchart demonstrating the primary steps of RFM adapted to Nadaraya-Watson with a Gaussian kernel.

depends on the x variable. Very rapidly, within the first several iterations, a metric can be learned that effectively removes the y variable, demonstrated by the elongation of the yellow overlay which depicts a ball in the Mahalanobis metrization.

2.A.2 Local EGOP

A fundamental motivation for this research is to leverage this framework for nonlinear dimension reduction tasks. RFM, being related to kernel ridge regression with a fixed Mahalanobis metric, is naturally suited to detect global linear structure within data. When the data is parameterized by a latent manifold with curvature, this procedure fails to capture the desired adaptivity. To make this concrete, we modify the previous experimental setting, wrapping the prescribed heatmap about the annulus in $\mathbb{R}^2$. The resulting data is multi-index in polar coordinates, being parameterized by $f(r,\theta) = \cos(\theta)$.

There are several reasons why a simple RFM-like implementation is lacking for such data. First, the direction of the gradient rotates with the angle θ, hence when the gradients are averaged uniformly across the dataset, the resulting metric is isotropic and exhibits a full-dimensional learning rate. Secondly, the direction of the manifold normal changes depending on the query point, hence no single Mahalanobis metric is capable of smoothing

the normal direction for each point simultaneously. These failure dynamics are visualized in Figure 2.9b within Section 2.A.5.

The Nadaraya-Watson estimator, on the other hand, being a local averaging estimator, is well-equipped to adapt to pointwise structures in the dataset. This weighting scheme naturally motivates our local EGOP estimation procedure in Figure 2.6, with the primary difference from the previous being the weighting of the averaged gradients. The intuition for this procedure is that, while the EGOP cannot explicitly dimension reduce the dataset, it still detects directions of principal variation of the target function f. This motivates a recursive procedure where datapoints achieving the largest values of the quadratic form $(x_i - x)^\top \widehat{\mathcal{L}}_t (x_i - x)$ are dropped from the dataset, the EGOP is estimated on the remaining data, and these values are computed again to reassess data-importance for the local regression problem. The Local EGOP Learning algorithm constitutes a continuously weighted formalization of this procedure, which we establish a rigorous basis for as an optimization of key functionals related to regression risk.

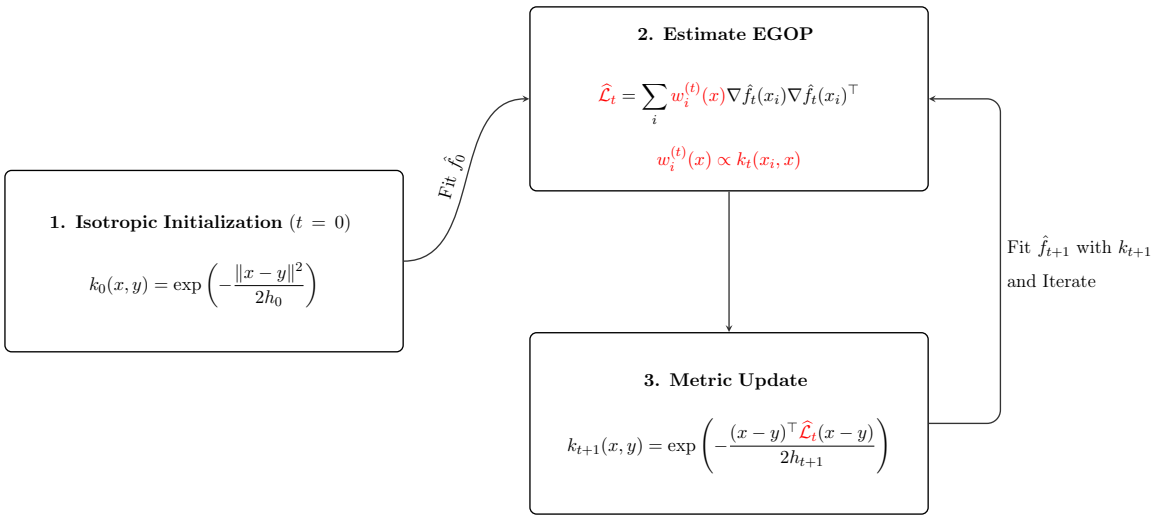


Figure 2.6: Flowchart illustrating the Local EGOP Learning algorithm, with the local weighting being highlighted in red.

2.A.3 Connections to Deep Learning

The recent revitalization of the Local Feature Learning framework in the RFM algorithm is primarily due to its strong empirical performance, and tentative connections to the training dynamics of neural network architectures. Preliminary connections between these methods have been a focus for much of the foundational research (Radhakrishnan, Beaglehole, et al., 2022; Beaglehole, Radhakrishnan, et al., 2023). Our research highlights several new connections, particularly regarding the role of depth in local feature learning. We consider architectures with a linear head, of the form $\hat{f}_{\text{NN}} = v^\top \phi(x)$. Of primary interest to us is the pullback Riemannian metric $G_\phi := D\phi D\phi^\top$, $D\phi$ the Jacobian of ϕ. This metric encodes the local displacement of ϕ, indicating the extent to which changes in x change the resulting feature embedding. For a two-layer neural network with m hidden nodes, $\phi(x) = \sum_{i=1}^m \sigma(w_i^\top x + c_i)$, and for σ the ReLU activation, the pullback metric has the simple form $W_{\text{active}} W_{\text{active}}^\top$. where W_{active} is the matrix with columns the active weight vectors w_i such that $w_i^\top x + c_i > 0$.

When developing theory regarding the learning capabilities of neural network architectures, two-layer neural networks are a typical tool, particularly for MIL feature learning as shown in Figure 2.10a. However, the CIL setting is incompatible with these predictors, as we demonstrate in Figure 2.10b, and has been previously observed in Nichani et al. (2023). As shown in Figure 2.10b, G_ϕ cannot adapt to nonlinear structures, being rendered effectively isotropic for annular data. In Figure 2.10c we see that this is overcome for three-layer neural networks. This is suggestive that global EGOP algorithms demonstrate similar adaptivity to two-layer neural networks, whereas localization has a similar effect to increasing the depth of the architecture. We leave it to future research to establish a formal connection between these algorithms.

2.A.4 Special Cases and Further Theoretical Details

Example 2.1 (Multi-Index Learning)**.** When the underlying manifold $\mathcal{M}$ is flat, CIL reduces to a multi-index setting. The key distinction is that displacement away from the manifold not only preserves the function value, but also all higher order function derivatives. In particular,

the Hessian has rank no more than d. When performing expansions such as Corollary 2.2, the orthogonal has no contribution to the EGOP, yielding a local MISE learning rate of $(d+1)/2$.

Example 2.2 (Continuous Single-Index)**.** When the underlying manifold is 1-dimensional, the tangent space coincides with the span of the gradient, hence $\pi_{g^\perp} H \pi_{g^\perp} = 0$, removing the second order anisotropy. Thus no normal directions contract asymptotically, leading to an effective dimension of 1 for the learning rate. This corroborates the result of Y. Wu and Maggioni (2024). Note that despite this common objective, the method we propose is very different than that of Y. Wu and Maggioni (2024). Their implementation requires approximating the underlying manifold through local PCA. Our procedure, by iteratively adapting the local metric, automatically detects implicit structure without this needing to be specified algorithmically.

Example 2.3 (Scalar recurrence and its fixed point)**.** Fix a schedule for $h_t \to 0$ such that h_{t+1}/h_t is larger than a fixed threshold. Consider the one dimensional recurrences

$$a_{t+1} = \frac{h_{t+1}}{\beta a_t + (1-\beta) a_{t-1}}, \qquad b_{t+1} = \frac{h_{t+1}}{c + \beta b_t + (1-\beta) b_{t-1}}. \tag{2.25}$$

It is easy to see that $b_t = O(h_t)$, as if b_t decays at all, then we have $b_t = h_{t+1}/c + o(h_t)$. Assume that $a_t = \Theta(h_t^\zeta)$. The recurrence then gives $a_{t+1} = \Theta(h_t^{1-\zeta})$, so a fixed rate requires $\zeta = 1/2$. Lemma 2.3 shows that the leading matrix recurrence consists of the rank-one constant term $gg^\top$ and the homogeneous term $H\Sigma H$. Appendix 2.C makes this correspondence precise: b_t models the gradient block of Σ_t, while a_t models the active block in $g^\perp$.

Example 2.4 (Momentum)**.** We now take a moment to highlight the necessity of the momentum term $\beta \in (0,1)$. Consider, again, the sequence a_t from Example 2.3, and suppose we set $\beta = 1$. For convenience, let us assume $\sqrt{h_{t+1}} = \sqrt{h_t} r$, an exact geometric schedule. We reparameterize, writing

$$a_{t+1} = \frac{h_{t+1}}{a_t} \iff a'_{t+1} = \frac{1}{a'_t}$$

for $a'_i := \frac{a_i}{\sqrt{h_i}}$. Clearly, a'_t has no limit, as it oscillates about 1. In the corresponding matrix recurrence, the remainder terms, as well as the estimation error, make the $\sqrt{h_t}$ rate become

unstable. Proposition 2.4 shows that every transverse linearized mode is strictly damped when $0 < \beta < 1$, while the endpoints retain undamped roots. We demonstrate the resulting localizations with and without momentum in Figure 2.7.

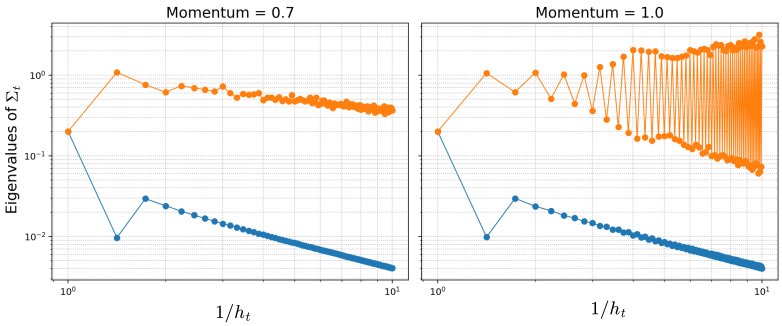


Figure 2.7: Comparison of the eigenvalue decay of Σ_t for Local EGOP Learning with and without momentum. On the left, we set $\beta = 0.7$, on the right $\beta = 1$ (no momentum). Without momentum, the second order eigenvalues are prone to wild oscillations.

Example 2.5 (Spheres)**.** Built into Theorem 2.2 are assumptions on the ambient dimension and derivatives of f. The sphere S^{D-1} is a typical toy example, however $D \geq 2(D-1)$ fails for $D \geq 3$. In these settings, the Hessian has the form

$$H = \begin{bmatrix} T & b \\ b^\top & 0 \end{bmatrix}$$

in the $\mathcal{T} \times \mathcal{N}$ basis. Thus the CIL hypothesis only guarantees that a single entry of the Hessian is 0. Hence, $\pi_{g^\perp} H \pi_{g^\perp}$ is generically full rank, driving second order anisotropy along the tangent and orthogonal, as demonstrated in Figure 2.8.

Example 2.6 (Pointwise MSE)**.** In our analysis, we emphasize the local MISE due to its direct correspondence to the EGOP and gradient based local feature algorithms. However, our analysis is still applicable to other standard problems such as pointwise MSE. A standard Taylor expansion adapted to our setting reveals the asymptotic rate

$$\mathbb{E}[(\hat{f}_{t_n}(x^*) - f(x^*))^2] = O\left(\|\pi_{\mathcal{T}_{x^*}} \Sigma_{t_n} \pi_{\mathcal{T}_{x^*}}\|^2 + \frac{1}{n\sqrt{\det \Sigma_{t_n}}} \right) = O\left(n^{-\frac{2}{d+2}} \zeta_n^{-\frac{D-2d+1}{d+2}} \right)$$

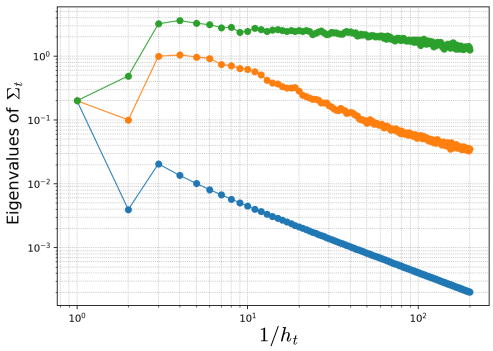


Figure 2.8: Eigenvalue decay of Σ_t with Local EGOP Learning applied to CIL data about S^2. The orthogonal exhibits light decay, approaching second order anisotropy asymptotically.

for CIL data. This is intrinsic dimensional, and in future work it is of interest to develop procedures toward the direct optimization of this regret.

Example 2.7 (Uniform-in-query extension)**.** Suppose that the local CIL radius ρ, the derivative bounds, and the smallest singular value of $(\pi_{g_x}^{\perp} H_x \pi_{g_x}^{\perp})|_{\mathcal{E}_x}$ are uniform over a compact set of query points, and that $\|\nabla f(x)\|$ is uniformly bounded away from zero there. Then the Gaussian-tail, small-cap, and spectral-separation constants in the proof of Theorem 2.2 can be chosen uniformly in x. Consequently, the shrinking-cap conclusion extends uniformly over such regions. Points at which the gradient vanishes or the active rank changes require a separate local analysis.

2.A.5 Empirical Visualizations of Metric Adaptation

The following figures display the primary algorithms studied in this chapter and the disparate learning dynamics they encourage. The data setting is kept consistent across algorithms, either displaying the single-index function $f(x, y) = \cos(x)$, or the continuous single-index function in polar coordinates $f(r, \theta) = \cos(\theta)$. In each figure, a heatmap of these function values is overlaid on the data, and a yellow ellipsoid is attached to points of interest indicating the nearest points in the prescribed Mahalanobis metrization about each query point. In EGOP related figures, gradients are indicated with red arrows. For neural networks, the

displayed metric is the pullback metric as described in Section 2.A.3.

2.B Experiment Details

The simulations were run on a machine of Intel® Xeon® processors with 48 CPU cores, and 50GB of RAM.

2.B.1 General Set-up

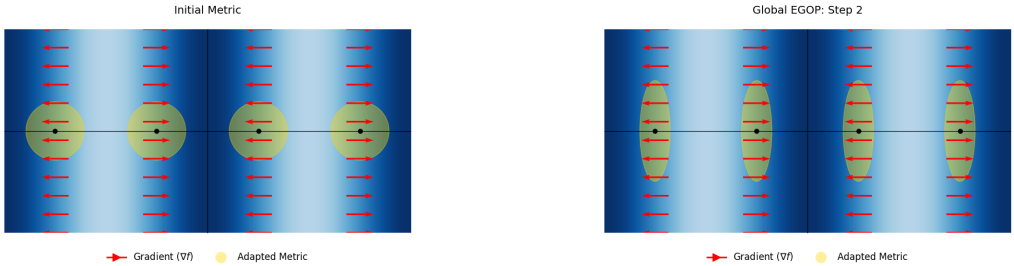
The following simulations illustrate our theoretical results. First, expanding on the noisy manifold setting, we show how the regression error rates of Local EGOP Learning remain invariant to the injection of high-dimensional noise. We demonstrate that two-layer neural networks are not able to efficiently learn in the CIL setting, with a sharp decrease in performance compared to Local EGOP Learning. We next compare the performance of our method to that of a deep neural network on toy data. In particular, we show how the feature embeddings produced by transformers trained on a simple example are qualitatively similar to the localizations generated by our procedure. Finally, we apply this procedure to estimate backbone angles in Molecular Dynamics (MD) data, where we leverage the noisy manifold structured data to improve prediction quality.

In our simulations, we consider helical data, parameterized by a curve $\theta(t) = (\sin(t + w_1), \cos(t + w_1), \sin(t + w_2), \cos(t + w_2), \ldots, g(t))$, where $g(t) = t$ is a linear term included if D is odd dimensional, and the w_i are constant offsets taken as a mesh from 0 to 2π. We rescale this data by a constant τ, then contaminate it with uniform orthogonal noise of radius r. In our simulations, we set $\tau = 0.8$, $r = 0.5$, and sample t from 0 to 2π. See Figure 2.11a for a visualization. For the outcomes y, we generate a third-degree polynomial with coefficients uniformly sampled from $(-3, 3)$, then evaluate it at the projection point onto $\theta(t)$.

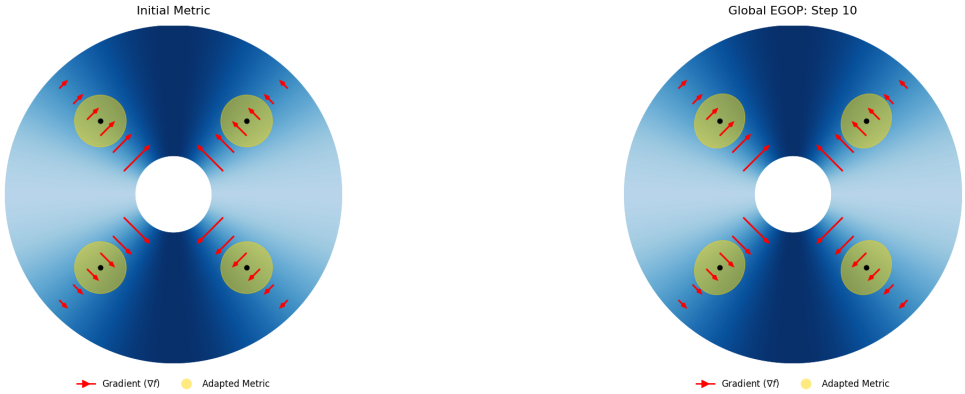
In our synthetic experiments, to generate labeled data we randomly generated a homogeneous cubic polynomial

$$f(x) = \sum_{|\alpha| \leq 3} c_\alpha x^\alpha,$$

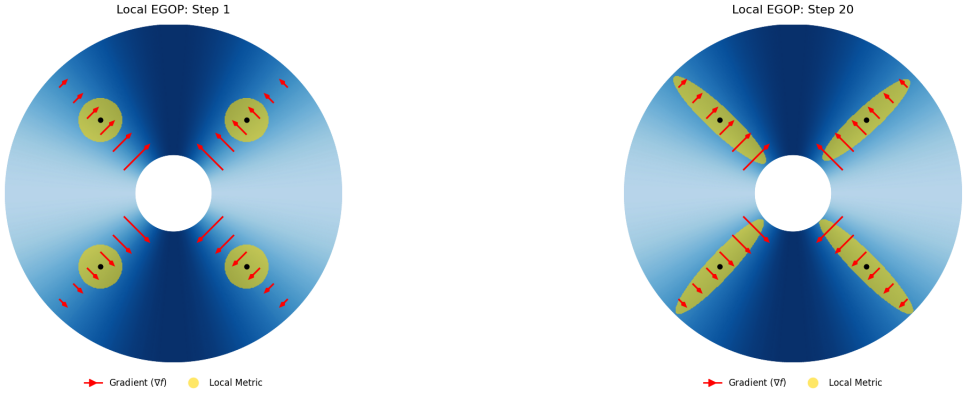
where $c_\alpha \sim \text{Uniform}[-3, 3]$. The resulting function f is then evaluated at the unperturbed



(a) Nadaraya-Watson RFM applied to multi-index data.

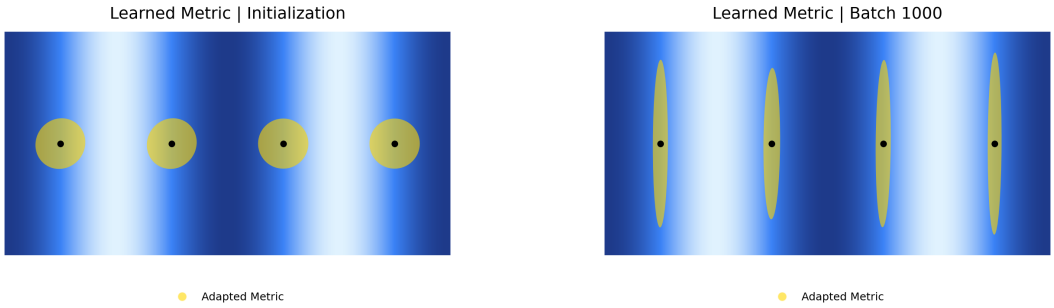


(b) Nadaraya-Watson RFM applied to CIL data.

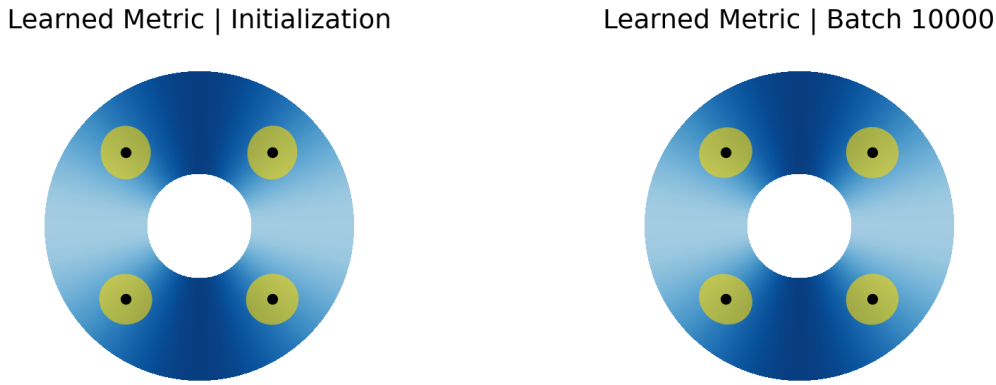


(c) Local EGOP Learning applied to CIL data.

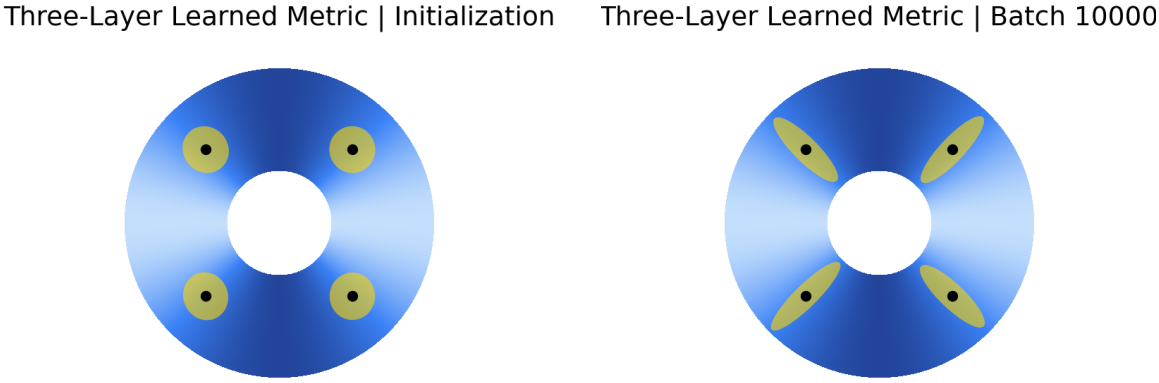
Figure 2.9: **EGOP Learning Algorithms.** Comparison of algorithmic adaptations. The left column displays the isotropic initialization, and the right column displays the final estimated metric.



(a) Pullback metric of a two-layer neural network for MIL data.

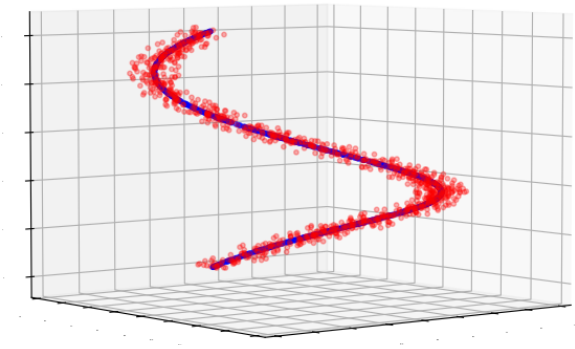


(b) Pullback metric of a two-layer neural network for CIL data.

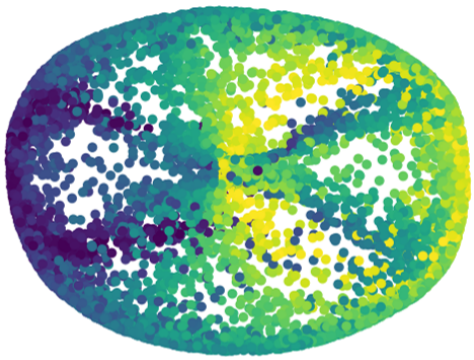


(c) Pullback metric of a three-layer neural network for CIL data.

Figure 2.10: **Neural Network Training Dynamics.** Comparison of representation learning over time. The left column displays the initialization, and the right column displays the final iteration (1,000 or 10,000).



(a) 3-dimensional Helical data.



(b) Coordinates of Malonaldehyde embedded into 2 dimensions via Diffusion Maps (Coifman and Lafon, 2006a) and colored by the backbone angles τ_1.

Figure 2.11: Visualizations of the experimental datasets.

datapoints before the features X were contaminated with noise orthogonal to the underlying manifold $\mathcal{M}$. The coefficients c_α are randomly sampled across each replicate of the experiments. In our plots, confidence bands are displayed with a width of ± 1 standard error. We generically apply the below hyperparameters to the Local EGOP Learning algorithm in each example.

- Number of EGOP iterations $T_{\text{test}} = 150$,
- Monte Carlo sample size $m = 300$,
- Bandwidth decay $h_t = (1+t)^{-1.2}$,
- Initialization scale $\alpha = 0.2$,
- Exclusion radius $\varepsilon = 1.6$.

2.B.2 Two-layer Neural Network

We train fully connected NNs with ReLU and GELU activations. All first layers use Kaiming initialization, and output layers are initialized at variance $1/\sqrt{m}$ where m is hidden width. All networks are trained with Adam.

For moderate-scale neural models, we conduct a grid search over hyperparameters. Widths are selected from $\{128, 256, 512, 1024, 2048, 4096, 8192\}$, learning rates are selected from $\{10^{-3}, 3 \times 10^{-3}\}$, weight decay from $\{0, 10^{-5}, 10^{-4}\}$, and batch size from $\{16, 32, 64\}$. The default Adam momentum parameters are used. Early stopping is applied with patience 200 and improvement tolerance 10^{-4} monitored on a validation split. In each experiment, only the optimal test error of these methods is aggregated into the "Small Tuned NN" plot.

We train overparameterized two-layer models with width 100,000, Adam with learning rate 0.003, batch size 64, and fixed weight decay 10^{-4}, applying early stopping with patience 300 and improvement tolerance 10^{-4}.

We also consider multi-index models $f(x) = g(Ux)$, g a two-layer NN, with width 256 and GELU activations, U a rank three projection matrix updated at each iteration.

2.B.3 Feature Learning

We generated $n = 50{,}000$ samples on a spherical cap in $\mathbb{R}^2$, restricted to an angular width of 25°. Each point x_i, we evaluate $f(x_i)$ for f a cubic polynomial with Uniform$[-1, 1]$ coefficients. The features x_i are then scaled radially by a random factor $r_i \sim \text{Uniform}[r, R]$, with $r = 0.6$ and $R = 1.8$. Additive Gaussian noise with standard deviation 0.05 was applied to the labels. We set $x^* = (1, 0)^\top$.

We trained an `FTTransformer` regression model (Gorishniy, Rubachev, Khrulkov, et al., 2021) on this dataset using the `rtdl` and `zero` frameworks. The network was trained for 40 epochs with a batch size of 1024 and learning rate 2×10^{-3}. After every 50 training steps, we computed weights:

$$w_i = \frac{\exp(-\|z_i - z^*\|^2/8)}{\sum_j \exp(-\|z_j - z^*\|^2/8)},$$

where z_i denotes the learned feature representation of x_i.

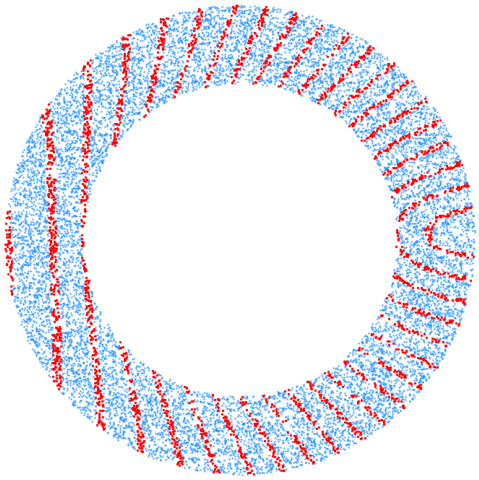


Figure 2.12: Data sampled from an annulus, with red rings displaying uniform distances in diffusion map embedding distance from the point $(1, 0)^\top$, for a 2-dimensional embedding.

2.B.4 Unsupervised Embedding

As noted in works such as (Coifman and Lafon, 2006a; El Karoui, 2010; Landa and Cheng, 2023b; Kokot, Murad, et al., 2025), unsupervised methods such as diffusion maps are robust to noise. The resulting low-frequency eigenfunctions closely match those of the underlying manifold. However, as demonstrated in Figure 2.12, they fail to capture this structure exactly, and thus cannot match the metrization produced by Local EGOP Learning and deep learning architectures in Figure 2.3.

2.B.5 Algorithmic Modifications

We chose to highlight the Local EGOP Learning Algorithm 1 presented in the text for its simplicity. We now discuss several modifications that can be implemented for improved numerical efficiency. To explain these, we walk through the algorithm's basic steps.

Step 1: Compute weights $w_i \propto \exp(-(x_i - x)^\top \widehat{M}_t (x_i - x))$. Evaluating $(x_i - x)^\top \widehat{M}_t (x_i - x)$ is D^2, and thus the full weight vector results in $O(nD^2)$.

Step 2: Sample m points with respect to these weights, and recompute these weights at

each of the new centers. This yields the dependence $O(mnD^2)$, and over the course of T iterations $O(TmnD^2)$.

The evaluation of the quadratic form can be improved when D is very large. A random projection (Johnson-Lindenstrauss) of dimension $r \propto \log(n)$ reduces the overall complexity to $O(T(rD^2 + nmrD))$ with good approximation guarantees. Targeted dimension reductions can be applied to similar effect, including local PCA as studied in Y. Wu and Maggioni (2024), or Nyström approximations leveraging the spectral structure of the Mahalanobis metric.

Step 3: Perform local linear regression at each of the m subsampled points. If the linear fit is performed on the full dataset this is $O(Tm(nD^2 + D^3))$, however if the linear regression is fit on m subsampled datapoints, this reduces to $O(Tm(mD^2 + D^3))$. Further, if we approximate matrix inversion with τ conjugate gradient steps, this can be reduced to $O(Tm(mD^2 + \tau D^2))$.

The resulting time complexity is $O(T(nmrD + (m^2 + m\tau + r)D^2))$.

2.C Proofs

In our theory, we focus on the kernel $k = \exp(-\|x\|^2)$. We discuss general kernels in Appendix 2.C.1.4. For notational clarity, we refer to the query point of our kernel predictor as x^*, and when it is unambiguous we denote $p = \pi(x^*)$ as the projection point onto the manifold.

2.C.1 Bias Control via EGOP

2.C.1.1 Bias Bound

To derive the upper bound from Lemma 2.1, we argue in stages. We first decompose the kernel smoother into two parts. The first is given by a pure Gaussian convolution and a density dependent covariance relating f and p. Both terms can readily be controlled via the EGOP form, leading to our bound. In the following, we absorb the bandwidth into the metric and covariance, suppressing h in the notation.

Define $\mu_M := N(x^*, M^{-1})$, $\mathbb{E}_M[\cdot] := \mathbb{E}_{\mu_M}[\cdot]$, Cov_M, Var_M similarly.

Lemma 2.5 (Density-weighted Gaussian smoother decomposition)**.**

$$P_M f = \mathbb{E}_M f + \frac{\mathrm{Cov}_M[f,p]}{\mathbb{E}_M p} =: G_M f + A_M f$$

Proof. We can simply compute

$$\begin{aligned} P_M f &= \frac{\mathbb{E}_M[fp]}{\mathbb{E}_M[p]} = \frac{\mathbb{E}_M[f]\mathbb{E}_M[p]}{\mathbb{E}_M[p]} + \frac{\mathrm{Cov}_M[f,p]}{\mathbb{E}_M[p]} \\ &= \mathbb{E}_M[f] + \frac{\mathrm{Cov}_M[f,p]}{\mathbb{E}_M[p]} =: G_M[f] + A_M f. \end{aligned}$$

□

Lemma 2.6 (Gaussian mean-squared bias bound)**.**

$$\int (\mathbb{E}_M[f] - f)^2 \, d\mu_M \leq W(\mu_M)$$

Proof. This is simply the Gaussian Poincaré inequality, see for instance Bakry et al., 2013[Proposition 4.1.1]. □

Lemma 2.7 (Density-correction bias bound)**.**

$$(A_M f)^2 \leq W(\mu_M) \frac{\mathrm{Var}_M(p)}{\mathbb{E}_M[p]^2}$$

Proof.

$$A_M f^2 = \frac{\mathrm{Cov}_M[f,p]^2}{\mathbb{E}_M[p]^2} \leq \frac{\mathrm{Var}_M(f)\,\mathrm{Var}_M(p)}{\mathbb{E}_M[p]^2} \leq W(\mu_M) \frac{\mathrm{Var}_M(p)}{\mathbb{E}_M[p]^2},$$

applying the Gaussian Poincaré inequality Bakry et al., 2013[Proposition 4.1.1] for the last line. □

Proof of Lemma 2.1. We can apply Lemmas 2.6 and 2.7 immediately to yield

$$\begin{aligned} \int (P_M - f)^2 \, d\mu_M &= \int (\mathbb{E}_M[f] - f + A_M f)^2 \, d\mu_M = \int (\mathbb{E}_M[f] - f)^2 \, d\mu_M + A_M f^2 \\ &\leq \left(1 + \frac{\mathrm{Var}_M(p)}{\mathbb{E}_M[p]^2}\right) W(\mu_M). \end{aligned}$$

Note that the density dependent factor $\frac{\mathrm{Var}_M(p)}{\mathbb{E}_M[p]^2}$ is bounded by the continuity of p and the capped covariance assumptions. □

2.C.1.2 Variance Bound

We operate by a typical Nadaraya-Watson style argument. Our first steps are to show that the denominator of our estimator concentrates, allowing us to treat it as a constant up to a negligible residual. Let $\gamma_M \propto \det(M)^{1/2}$ by the normalization factor of μ_M. Define

$$\hat{C}'_M := \gamma_M \hat{C}_M := \gamma_M \frac{1}{n} \sum_{i=1}^{n} \exp(-(X_i - x^*)^\top M (X_i - x^*)).$$

Lemma 2.8 (Kernel-denominator concentration)**.**

$$\mathbb{P}(|\hat{C}'_M - \mathbb{E}_M(p)| \geq \mathbb{E}_M(p)/2) \leq 2 \exp\left(-\frac{n\mathbb{E}_M(p)}{10\gamma_M}\right)$$

Proof. We prove this by applying Bernstein's inequality. First, observe $\mathbb{E}[\hat{C}_M] = G_M(p)$. Next, $\exp(-v^\top M v) \in (0, 1]$ for any v, hence $\hat{C}'_M = \frac{\gamma_M}{n} \sum_{i=1}^{n} \exp(-(X_i - x)^\top M (X_i - x))$ is the sample mean of iid bounded random variables in $(0, \gamma_M]$. Further, we can compute

$$\begin{aligned}
\operatorname{Var}[\gamma_M \exp(-(X_i - x)^\top M (X_i - x))] &\leq \gamma_M \mathbb{E}[\gamma_M \exp(-(X_i - x)^\top M (X_i - x))^2] \\
&\leq \gamma_M \mathbb{E}[\exp(-(X_i - x)^\top M (X_i - x))] = \gamma_M G_M(p).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{P}(|\hat{C}'_M - G_M(p)| \geq G_M(p)/2) \\
&\qquad \leq 2 \exp\left(-\frac{n G_M(p)^2}{8 \operatorname{Var}(\gamma_M \exp(-(X_i - x)^\top M (X_i - x))) + \frac{4}{3}\gamma_M G_M(p)^2}\right) \\
&\qquad \leq 2 \exp\left(-\frac{n G_M(p)^2}{8\gamma_M G_M(p) + \frac{4}{3}\gamma_M G_M(p)}\right) \\
&\qquad \leq 2 \exp\left(-\frac{n G_M(p)}{10\gamma_M}\right).
\end{aligned}$$

□

Lemma 2.9 (Finite-sample kernel-regression variance decomposition)**.** *Assume that the noise ε_i have finite fourth moments. Then, there exists a constant $C > 0$ such that*

$$\begin{aligned}
\mathbb{E}[(\hat{P}_M(f) - P_M(f))^2] \leq\ & 2\frac{\mathbb{E}[\gamma_M^2 \exp(-(X_i - x)^\top M (X_i - x))^2 \{Y_i - P_M(f)\}^2]}{n G_M(p)^2} \\
& + 2\frac{\mathbb{E}[\{\gamma_M \exp(-(X_i - x)^\top M (X_i - x)) Y_i\}^2]}{n G_M(p)^2} + C \exp\left(-\frac{n G_M(p)}{10\gamma_M}\right)
\end{aligned}$$

Proof. Let E be the event $|\hat{C}'_M - G_M(p)| \le G_M(p)/2$. We compute

$$\begin{aligned}
\hat{P}_M(f) - P_M(f) &= \frac{\frac{1}{n}\sum_{i=1}^n \gamma_M \exp(-(X_i - x^*)^\top M(X_i - x))Y_i}{\hat{C}'_M} - \frac{G_M(fp)}{G_M(p)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n [\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))\{Y_i - P_M(f)\}]}{G_M(p)} \mathbf{1}_E \\
&\quad + \frac{G_M(p) - \hat{C}'_M}{G_M(p)\hat{C}'_M} \frac{1}{n}\sum_{i=1}^n [\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i] \mathbf{1}_E \\
&\quad + \left\{ \frac{\frac{1}{n}\sum_{i=1}^n \gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i}{\hat{C}'_M} - \frac{G_M(fp)}{G_M(p)} \right\} \mathbf{1}_{E^c}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}[(\hat{P}_M(f) - P_M(f))^2] &\le 2\frac{\mathbb{E}[\gamma_M^2 \exp(-(X_i - x^*)^\top M(X_i - x^*))^2 \{Y_i - P_M(f)\}^2]}{nP_M(p)^2} \\
&\quad + 2\mathbb{E}\left[\left\{ \frac{G_M(p) - \hat{C}'_M}{G_M(p)\hat{C}'_M} \frac{1}{n}\sum_{i=1}^n [\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i] \right\}^2 \mathbf{1}_E \right] \\
&\quad + 2\mathbb{E}\left[\left\{ \hat{P}_M(f) - P_M(f) \right\}^2 \mathbf{1}_{E^c} \right].
\end{aligned}$$

The first term is as desired. Next, by definition of the event, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\{ \frac{G_M(p) - \hat{C}'_M}{G_M(p)\hat{C}'_M} \frac{1}{n}\sum_{i=1}^n [\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i] \right\}^2 \mathbf{1}_E \right] \\
&\qquad \le \frac{\mathbb{E}[\{\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i\}^2]}{nG_M(p)^2}.
\end{aligned}$$

For the remaining term, observe that

$$\hat{P}_M(f) = \sum_{i=1}^n w_i Y_i$$

where w_i are a convex combination. Hence,

$$\begin{aligned}
&\mathbb{E}\left[\left\{ \hat{P}_M(f) - P_M(f) \right\}^2 \mathbf{1}_{E^c} \right] \le \\
&\qquad \sqrt{\mathbb{E}\left[\left\{ 2\hat{P}_M(f)^2 + 2P_M(f)^2 \right\}^2 \right]} \mathbb{P}[E^c] \le C \exp\left(-\frac{nG_M(p)}{10\gamma_M} \right),
\end{aligned}$$

by Lemma 2.8 and the hypothesis on the moments of ε_i and the boundedness of f. □

With this out of the way, we bound the two remaining expectations.

Lemma 2.10 (Gaussian-kernel second-moment bounds)**.** *There exists a constant $C > 0$ such that*

$$\mathbb{E}[\gamma_M^2 \exp(-(X_i - x^*)^\top M(X_i - x^*))^2 \{Y_i - P_M(f)\}^2] \leq C\gamma_M G_{2M}(p),$$
$$\mathbb{E}[\{\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i\}^2] \leq C\gamma_M G_{2M}(p).$$

Proof. Notice that $\gamma_M/\gamma_{4M} = C$ for some constant. Let $\varepsilon \sim \rho$. Focusing on the first expression,

$$\begin{aligned}
&\mathbb{E}[\gamma_M^2 \exp(-(X_i - x^*)^\top M(X_i - x^*))^2 \{Y_i - P_M(f)\}^2] \\
&\quad = C\gamma_M \int \gamma_{4M} \exp(-(z - x)^\top [2M](z - x))p(y)(f(y) + \varepsilon - P_M(f))^2 \, dz \, d\rho(\varepsilon) \\
&\quad = C\gamma_M \mathbb{E}_{2M}[p(x + Z)\mathbb{E}_\rho[(f(x + Z) + \mathcal{E} - P_M(f)(x))^2]] \leq C'\gamma_M G_{2M}(p)(x),
\end{aligned}$$

using bounded moments and bounded f assumptions. Similarly,

$$\mathbb{E}[\{\gamma_M \exp(-(X_i - x^*)^\top M(X_i - x^*))Y_i\}^2] = C\gamma_M \mathbb{E}_{2M,\rho}[p(Z)[f(Z) + \mathcal{E}]^2] = C\gamma_M G_{2M}(p).$$

□

Combining these yields the desired variance bound multiplied by density dependent factors.

Proof of Lemma 2.2. By Lemmas 2.9 and 2.10, there exists a constant $C > 0$ such that

$$\begin{aligned}
\mathbb{E}[(\hat{P}_M(f) - P_M(f))^2] &\leq C\frac{\gamma_M}{n}\left(\frac{G_{2M}(p)}{G_M(p)^2}\right) + C\exp\left(-\frac{n}{\gamma_M}\left(\frac{G_M(p)}{10}\right)\right) \\
&= O\left(\frac{\sqrt{\det M}}{n}\right).
\end{aligned}$$

□

2.C.1.3 Local MISE

We compile the results from the previous section to provide guarantees for the MISE against Gaussians.

Proof of Theorem 2.1. By adding and subtracting $P_M(f)$, then applying Lemmas 2.1 and 2.2,

$$\mathbb{E}\left[\int (\hat{P}_M(f) - f)^2 \, d\mu_M\right] \leq 2\mathbb{E}[(\hat{P}_M(f) - P_M(f))^2] + 2\int (P_M(f) - f)^2 \, d\mu_M$$
$$= O\left(\frac{\sqrt{\det M}}{n} + W(\mu_M)\right).$$

Since $\Sigma = M^{-1}$ after absorbing the bandwidth into M, this is the claimed bound. □

2.C.1.4 General Kernels

In this section, we discuss which of results above are applicable to kernels of the form $k(M^{1/2}x)$. We start with the following Poincaré inequality. Let $\mu_{k,M}$ denote the probability distribution with density $\propto k(M^{1/2}[x - x^*])$.

Lemma 2.11 (Poincaré inequality under Mahalanobis rescaling)**.** *The measure $\mu_{k,M}$ has covariance proportionate to M^{-1}. If $\mu_{k,I}$ satisfies a Poincaré inequality with constant $1/C_I$, then*

$$\mathrm{Var}_{\mu_{k,M}}(f) \leq \frac{1}{C_I}\int \nabla f^\top M^{-1} \nabla f \, d\mu_{k,M}.$$

Proof. For convenience, we translate $x^* = 0$ without loss of generality. For the first claim, notice that when $M = I$, the density is symmetric about the origin, hence the covariance is proportionate to the identity. Thus the claim follows as $\mu_{k,M}$ is the pushforward of $\mu_{k,I}$ by $M^{-1/2}$.

Further,

$$\mathrm{Var}_{\mu_{k,M}}(f) = \frac{1}{\sqrt{\det M}}\mathrm{Var}_{\mu_{k,I}}(f \circ M^{-1/2}) \leq \frac{1}{C_I}\int \frac{1}{\sqrt{\det M}} \|\nabla f(M^{-1/2}y)\|^2 d\mu_{k,I}(y).$$

An application of the chain rule gives that

$$\nabla_y f(M^{-1/2}y) = M^{-1/2}\nabla f(M^{-1/2}y).$$

Hence,

$$\mathrm{Var}_{\mu_{k,M}}(f) \leq \frac{1}{C_I}\int \frac{1}{\sqrt{\det M}} \|M^{-1/2}\nabla f(M^{-1/2}y)\|^2 \, \mu_{k,I}(y).$$

Changing variable back to $\mu_{k,M}$ yields the desired result. □

With this in hand, the following is immediate for $\hat{P}_M$ with k generic. The claim for the covariance similarly follows by change of variables, after observing that $\mu_{k,I}$ is rotation invariant, and therefore has isotropic covariance.

Corollary 2.1 (MSE bound for general radial kernels)**.** *Take k to be nonnegative, and satisfying a Poincaré inequality for isotropic covariance. Assume that f is bounded and ε_i has finite fourth moments. Then there exists an absolute constant C such that*

$$\mathbb{E}\left[\int (\hat{P}_M(f) - f)^2 \, d\mu_{k,M}\right] \leq C\frac{\sqrt{\det M}}{n}\left(\frac{\mathbb{E}_{\mu_{k^2,2M}}[p]}{\mathbb{E}_{\mu_{k,2M}}[p]^2}\right) + C\exp\left(-\frac{n}{\sqrt{\det M}}\left(\frac{\mathbb{E}_{\mu_{k,2M}}[p]}{10}\right)\right) + C\left(1 + \frac{\mathrm{Var}_{\mu_{k,M}}(p)}{\mathbb{E}_{\mu_{k,M}}[p]^2}\right) W(\mu_{k,M}).$$

2.C.2 Recursive Kernel Learning as Variance Minimization

The proof of Proposition 2.1 is clear from the main text, but we fill in the missing details here.

Proof of Proposition 2.1. For $A \succ 0$, the Lagrangian of the uncapped entropy problem is

$$\log\det X - \nu(\mathrm{tr}(AX) - h).$$

The first-order condition gives $X^{-1} = \nu A$, and the active trace constraint gives $\nu = D/h$. Hence $\Sigma_h^\circ(A) = hA^{-1}/D$. For $A \succeq 0$, diagonalize $A = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_D)U^\top$. Then

$$\mathrm{Cap}_\zeta\left(\frac{h}{D}(A + \varepsilon I)^{-1}\right) = U\,\mathrm{diag}\left(\min\left\{\zeta, \frac{h}{D(\lambda_j + \varepsilon)}\right\}\right)_{j=1}^{D} U^\top.$$

Taking $\varepsilon \downarrow 0$ gives the stated eigenvalue formula. In particular, if $\lambda_j = 0$, the regularized inverse diverges and the truncation maps that direction to ζ. □

2.C.3 Guarantees

To derive the learning rate for the Local EGOP Learning algorithm, we mostly operate at the level of matrices, studying the recurrence from Equation 2.16. To supplement this in the CIL setting, we rely on critical geometric relations verified in Appendix 2.C.3.2.

2.C.3.1 Local CIL Geometry and Shrinking-Cap Localization

We impose the CIL structure only locally. Fix $\rho > 0$ such that $B(x^*, 2\rho)$ lies inside a tubular neighborhood on which the nearest-point projection is unique and

$$f(x) = q(\pi(x)).$$

All geometric identities in Lemma 2.4 are assumed only on this neighborhood. Away from it, we use only the global boundedness of f and its derivatives stated in Section 2.3.

The shrinking covariance cap makes this local formulation sufficient. The following bound will be used repeatedly.

Lemma 2.12 (Gaussian localization under a shrinking cap)**.** *Let $Z \sim N(0, \Sigma)$ with $0 \preceq \Sigma \preceq \zeta I_D$ and $0 < \zeta \leq 1$. For every fixed integer $q \geq 0$, there are constants $c, C_q > 0$, depending only on D, q, such that*

$$\mathbb{P}(\|Z\| > \rho) + \mathbb{E}\left[(1 + \|Z\|^q)\mathbf{1}\{\|Z\| > \rho\}\right] \leq C_q \exp\left(-\frac{c\rho^2}{\zeta}\right).$$

Consequently, if $\zeta_n = c_\zeta / \log n$, then for any prescribed $A > 0$, choosing c_ζ sufficiently small gives

$$\varepsilon_n := C \exp\left(-\frac{c\rho^2}{\zeta_n}\right) \leq C n^{-A}.$$

Proof. Write $Z = \Sigma^{1/2} G$ with $G \sim N(0, I_D)$. Since $\|\Sigma^{1/2}\|_{\mathrm{op}} \leq \sqrt{\zeta}$,

$$\{\|Z\| > \rho\} \subseteq \{\|G\| > \rho/\sqrt{\zeta}\}.$$

The standard finite-dimensional Gaussian tail bound gives $\mathbb{P}(\|G\| > x) \leq C \exp(-cx^2)$ for all sufficiently large x. For the weighted bound, Cauchy–Schwarz yields

$$\mathbb{E}\left[(1 + \|Z\|^q)\mathbf{1}\{\|Z\| > \rho\}\right] \leq \left\{\mathbb{E}(1 + \|Z\|^q)^2\right\}^{1/2} \mathbb{P}(\|Z\| > \rho)^{1/2}.$$

The first factor is uniformly bounded for $\zeta \leq 1$, and the second has the asserted exponential decay. Substituting $\zeta_n = c_\zeta / \log n$ proves the polynomial bound. Since c_ζ is a free sufficiently small constant, its exponent can be made arbitrarily large. □

Because every covariance produced by the truncated update satisfies $\|\Sigma_t\|_{\mathrm{op}} \le \zeta_n$, Lemma 2.12 applies at every iteration. The bounded derivatives of f imply that the contribution to any EGOP or Taylor block from $B(x^*, \rho)^c$ is $O(\varepsilon_n)$. In the proof of Theorem 2.2, c_ζ is chosen so that, uniformly for $t \le t_n$,

$$\varepsilon_n = o(h_{t_n}), \qquad \frac{t_n \varepsilon_n}{\sqrt{h_{t_n}}} = o(1).$$

Thus all nonlocal contributions are smaller than the retained Taylor remainders and remain negligible after normalization and iteration.

2.C.3.2 CIL Geometry

We collect the geometric facts used to identify the exact gradient-invariant and projected-null subspaces in the CIL recurrence. We start with basic properties of our function of interest f.

Lemma 2.13 (Tangent-gradient and normal-Hessian identities)**.** *Suppose that $f = f \circ \pi$. Then $\nabla f(x) \in \mathcal{T}_x$ and for any $u, v \in \mathcal{N}_x$, $u^\top \nabla^2 f(x) v = 0$.*

Proof. Let $u, v \in \mathcal{N}_x$ be arbitrary. Then,

$$\begin{aligned}\nabla f(x)^\top v &= \lim_{t\to 0}[f(x+tv) - f(x)]/t = \lim_{t\to 0}[f(\pi[x+tv]) - f(\pi[x])]/t \\ &= \lim_{t\to 0}[f(\pi[x]) - f(\pi[x])]/t = 0,\end{aligned}$$

as for small t orthogonal perturbations do not change the nearest point on $\mathcal{M}$. As this holds for generic v the first claim follows. To simplify notation, define $\partial_\eta h(x) := \lim_{t\to 0}[h(x + t\eta) - h(x)]/t$. We can thus compute

$$u^\top \nabla^2 f(x) v = \partial_u(\nabla f^\top v)(x) = 0$$

as $\nabla f^\top v$ is identically 0 along the path $x + tv$, $t \in [0, \delta]$ for δ small enough. That is, as ∇f remains tangent along this path, as verified above, it is always orthogonal to v. □

The remaining CIL proof uses the explicit gradient formula in the tubular neighborhood, from which the exact gradient-invariant and projected-null subspaces follow.

2.C.3.3 Gradient Geometry and Projected Hessian Structure

Lemma 2.14 (Differential of the nearest-point projection)**.** *Let $x \in \mathcal{M}$ and $\eta \in \mathcal{N}_x$. Then $D\pi_{x+\eta} = (I - S_x(\eta))^{-1}\pi_{\mathcal{T}_x}$.*

Proof. Clearly, for any normal vector $v \in \mathcal{N}_x$, $D\pi_{x+\eta}v = 0$, as projection is invariant to normal displacement. Thus it suffices to verify the formula for any tangent vector. Our proof has two primary parts. First, we pass to normal coordinates, equating the tubular neighborhood about $\mathcal{M}$ with the normal bundle, and compute the differential of the exponential map. From this parameterization, we can then analyze the projection via the chain rule.

Starting from the normal bundle $N\mathcal{M}$, for $(x, v) \in N_x\mathcal{M}$, the exponential map $E(x, v) = x + v$ is a diffeomorphism to the tube. Let $\gamma(t) = (u(t), \xi(t)) \in N\mathcal{M}$, $\gamma'(t) \in TN\mathcal{M}$. We compute $\xi'(0) = -S_x(v)u'(0) + \nabla^{\perp}_{u'(0)}\xi(0)$, hence

$$DE_{(x,v)}\gamma'(0) = (I - S_x(v))u'(0) + \nabla^{\perp}_{u'(0)}\xi(0).$$

Composing with the projection map, $\pi \circ E(x, v) = x$, hence

$$D\pi_{x+v}DE_{(x,v)}\gamma'(0) = u'(0) \quad \iff \quad D\pi_{x+v}(I - S_x(v))u'(0) = u'(0),$$

and inverting yields the claim, as $D\pi_{x+v}$ is a linear operator. □

Proof of the gradient-geometry assertions in Lemma 2.4. The first claim follows immediately from Lemma 2.14 and the chain rule, indeed,

$$\nabla f(x + v) = \nabla g(\pi(x + v)) = (I - S_x(v))^{-1}\nabla g(x).$$

Let $u = (I - S_x(v))^{-1}\nabla g(x)$. Applying linearity of the shape operator, observe that $(I - S_x(v + \eta))u = (I - S_x(v))u - S_x(\eta)(u) = \nabla g(x) - S_x(\eta)u$. Hence, if $S_x(\eta)u = 0$ then $u = \nabla f(x + v + \eta)$. Thus, the second claim follows if we can show that there is a $D - 2d$ dimensional subspace Null that annihilates the solution vector u. For $w \in \mathcal{T}_x$,

$$\langle S_x(\eta)u, w\rangle = \langle \mathrm{I\!I}(u, w), \eta\rangle$$

for $\mathrm{I\!I}$ the second fundamental form. This is a bilinear map, and the map defined by fixing the u argument, $\phi_u(w) := \mathrm{I\!I}(u, w)$ is a linear map from $\mathcal{T}_x \to \mathcal{N}_x$, and thus has image no larger than d. Its complement $\phi_u(\mathcal{T}_x)^{\perp}$ has dimension at least $\dim(\mathcal{N}_x) - d = D - 2d$, as desired.

This proves the invariance of the gradient along a space of dimension at least $D - 2d$, which immediately implies $\nabla^2 f(x + v + \eta)z = \partial_z \nabla f(x + v) = 0$. □

Proof of the projected-Hessian assertions in Lemma 2.4. Write $g^\perp = (\mathcal{T}_p \cap g^\perp) \oplus \mathcal{N}_p$. In this basis,

$$(QHQ)|_{g^\perp} = \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix},$$

where $B : \mathcal{T}_p \cap g^\perp \to \mathcal{N}_p$. Since the tangent block has dimension $d - 1$, rank $2d - 2$ forces B to have full column rank. Hence the kernel on $g^\perp$ consists of normal vectors v satisfying $B^\top v = 0$. For such v, $QHv = 0$, so $Hv \in \mathrm{span}(g)$. Symmetry implies that QHQ is invertible on its range. Finally, $H \,\mathrm{Null} = 0$ gives $\mathrm{Null} \subseteq \mathcal{K}_x$, and the dimension bound shows that $\mathcal{K}_x$ contains at most one additional direction.

To obtain the local assertion along the full projected-null space, write $A_\nu = I - S_p(\nu)$ and $g_0 = \nabla f(p)$, so $g = A_\nu^{-1} g_0$. For $v \in \mathcal{K}_x$, differentiation in the normal direction gives

$$Hv = A_\nu^{-1} S_p(v) g = c(v) g$$

for a linear functional c on $\mathcal{K}_x$. Thus $S_p(v)g = c(v)g_0$ and

$$(A_\nu - S_p(v))g = (1 - c(v))g_0.$$

For sufficiently local v, the gradient formula therefore yields

$$\nabla f(x^* + v) = (1 - c(v))^{-1} g \in \mathrm{span}(g).$$

Applying the same calculation at $x^* + v$ in a direction $w \in \mathcal{K}_x$ gives $\nabla^2 f(x^* + v)w \in \mathrm{span}(g)$, and hence $Q\nabla^2 f(x^* + v)\pi_{\mathcal{K}_x} = 0$. □

2.C.3.4 Generic and Conditional Local EGOP Expansions

We first prove the Local EGOP Expansion stated in Lemma 2.3. We then condition on the exact gradient-invariant block to obtain the conditional expansion used in the CIL argument.

Proof of Lemma 2.3. We start by Taylor expanding the gradient. For a D tuple ξ, as $\partial_\xi \partial_i f = \partial_i \partial_\xi f$, for f sufficiently smooth we can express

$$\partial_i f(x) = \sum_{|\xi| \le q} \frac{1}{q!} \partial_i [\partial_\xi f](x) \prod_{q_i \in \xi} x_i^{q_i} + O(\|x\|^{q+1}).$$

In particular, we have

$$\partial_i f(x) = \partial_i f(0) + \sum_j \partial_i \partial_j f(0) x_j + \frac{1}{2} \sum_{j,k} \partial_i \partial_j \partial_k f(0) x_j x_k + O(\|x\|^3).$$

Working at the level of the gradient, this yields

$$\nabla f(x) = \nabla f(0) + \nabla^2 f(0) x + T(x,x) + O(\|X\|^3)$$

where T is the bilinear for

$$[T[x,y]]_i = \frac{1}{2} \sum_{j,k} \partial_i \partial_j \partial_k f(0) x_j y_k.$$

We use the first three terms of this Taylor expansion as a starting point. Set $g := \nabla f(0), H := \nabla^2 f(0)$.

The remainder of the Taylor expansion can be exactly expressed as

$$\tilde{R}(x) = \frac{1}{6} \int_0^1 (1-s)^2 \nabla^4 f(sx)[x,x,x]\, ds, \quad [\nabla^4 f(sx)[x,x,x]]_i := \sum_{j,k,\ell} \partial_i \partial_j \partial_k \partial_\ell f(sx) x_i x_j x_k.$$

Hence, under our uniform boundedness assumption, we have

$$\|T[x,x]\| \le C\|x\|^2, \quad \|\tilde{R}(x)\| \le C\|x\|^3$$

for some constant C. Throughout the argument, for convenience we will absorb constants into C if necessary. Hence,

$$\mathcal{L}(\mu) = \int \nabla f \nabla f^\top \, d\mu = g g^\top + H\Sigma H + g T(\Sigma)^\top + T(\Sigma) g^\top + R(\Sigma),$$

for

$$T(\Sigma)_i := \mathbb{E}_\mu[T(x,x)_i] = \frac{1}{2} \sum_{j,k} \partial_i \partial_j \partial_k f(0) \Sigma_{jk}$$

$$R(\Sigma) := \mathbb{E}_\mu[T[Z,Z]T[Z,Z]^\top + HZ\tilde{R}(Z)^\top + \tilde{R}(Z)(HZ)^\top + \tilde{R}(Z)\tilde{R}(Z)^\top].$$

That T is continuous and linear in Σ is immediate. For the remainder, we have

$$\begin{aligned}\|R(\Sigma)\| &\leq \mathbb{E}_\mu[\|T[Z,Z]T[Z,Z]^\top\| + \|HZ\tilde{R}(Z)^\top\| + \|\tilde{R}(Z)(HZ)^\top\| + \|\tilde{R}(Z)\tilde{R}(Z)^\top\|] \\ &\leq C\mathbb{E}[\|Z\|^4 + \|Z\|^6] \leq C\mathbb{E}_\mu[\mathrm{tr}(\Sigma^2) + \mathrm{tr}(\Sigma)^2]\end{aligned}$$

where we apply Isserlis's Theorem Isserlis, 1918 which verifies the vanishing of third order terms, and relates the higher moments to the norm of the covariance matrix. Our capped covariance assumption is used to compress all higher order terms to quadratic order. Our proof is completed upon observing

$$\mathrm{tr}(\Sigma^2) = \|\Sigma\|_F^2, \quad \mathrm{tr}(\Sigma)^2 = \left(\sum_i \lambda_i\right)^2 \leq D\sum_i \lambda_i^2 = D\|\Sigma\|_F^2,$$

and in finite dimensions the frobenius and operator norms are equivalent. □

For the CIL and empirical arguments, we also record the expansion obtained by conditioning on the exact gradient-invariant subspace from Lemma 2.4.

Corollary 2.2 (Conditional Local EGOP Expansion)**.** *Let $x^* = p + \eta$, $v \in \mathrm{Null}_{p+\eta}$. Define μ_v as the measure $\mu = N(x^*, \Sigma)$ conditioned on $\pi_{\mathrm{Null}}[X - x^*] = v$, i.e. the draws where the* Null *component is fixed at v. Let Σ_v be the covariance of this distribution. Then,*

$$\mathcal{L}(\mu_v) = gg^\top + H_v\Sigma_v H_v + gT_v(\Sigma_v)^\top + T_v(\Sigma_v)g^\top + R_v(\Sigma_v)$$

where $\mathrm{Null} \subseteq \mathrm{nullspace}(H_v)$, $\sup_{v\in\mathrm{Null}} \|R_v(A)\| \leq C_{\mathrm{Null}}\|A\|^2$, $T_v(A) \leq C_{\mathrm{Null}}\|A\|$ *for a uniform constant C_{Null}, and H_v, T_v, R_v are continuous in v and Lipschitz in $\|A\|$.*

Proof. The basic claim is immediate from Lemma 2.3, and that the terms stemming from higher order derivatives are uniformly bounded follows from the smoothness assumptions on f. Further, the nullspace of the Hessian and the zeroth-order expansion being the gradient $g = \nabla f(x^*)$ follow immediately from Lemma 2.4. □

2.C.3.5 Rescaled Inverse Recurrences

The full-rank projected-Hessian argument reduces to the following fixed-congruence recurrence. We rescale its active block by $\sqrt{h_t}$, obtaining a fixed time-homogeneous inverse map plus a summable perturbation.

Homogeneous active block. We first study the signature-normalized recurrence

$$C_{t+1} = h_{t+1} J\{\beta C_t + (1-\beta) C_{t-1} + F_t\}^{-1} J, \tag{2.26}$$

where $C_t \succ 0$, $J = J^\top = J^{-1}$, and $\sqrt{h_t} = \alpha r^t$.

Proposition 2.2 (Rescaled Homogeneous Recurrence)**.** *Define*

$$c(r,\beta) := \frac{1}{\beta/r + (1-\beta)/r^2}, \qquad \eta := c(r,\beta)\frac{\beta}{r}, \tag{2.27}$$

$$X_t := \frac{C_t}{\sqrt{c(r,\beta) h_t}}, \qquad R_t := \frac{\sqrt{c(r,\beta)} F_t}{\sqrt{h_{t+1}}}. \tag{2.28}$$

Then $\eta \in (0,1)$, $1-\eta = c(r,\beta)(1-\beta)/r^2$, and

$$X_{t+1} = J\{\eta X_t + (1-\eta) X_{t-1} + R_t\}^{-1} J. \tag{2.29}$$

If $\|F_t\| = O(h_t)$, then $\|R_t\| = O(\sqrt{h_t}) = O(\alpha r^t)$, so the normalized recurrence is a summably perturbed time-homogeneous system.

Proof. Substitute $C_s = \sqrt{c(r,\beta) h_s}\, X_s$ into (2.26). Since

$$\sqrt{h_t} = \frac{\sqrt{h_{t+1}}}{r}, \qquad \sqrt{h_{t-1}} = \frac{\sqrt{h_{t+1}}}{r^2}, \tag{2.30}$$

the matrix inside the inverse becomes

$$\sqrt{c(r,\beta) h_{t+1}} \left\{ \frac{\beta}{r} X_t + \frac{1-\beta}{r^2} X_{t-1} + \frac{F_t}{\sqrt{c(r,\beta) h_{t+1}}} \right\}. \tag{2.31}$$

After dividing the resulting identity by $\sqrt{c(r,\beta) h_{t+1}}$, the choice of $c(r,\beta)$ makes the first two coefficients sum to one and gives (2.29). The bound on R_t follows from $h_{t+1} = r^2 h_t$. □

Lemma 2.15 (Spectral Stability of the Rescaled Recurrence)**.** *Suppose that $C_0/\sqrt{h_0}$ and $C_1/\sqrt{h_1}$ have spectra bounded above and away from zero, and that $\sum_t \|R_t\|$ is sufficiently small. Then there are constants $0 < c < C < \infty$ such that*

$$cI \preceq X_t \preceq CI \tag{2.32}$$

for every t. Consequently, $C_t = \Theta(\sqrt{h_t})$.

Proof. For positive-definite matrices, write the Thompson metric as

$$d_T(A, B) := \left\| \log \left(A^{-1/2} B A^{-1/2} \right) \right\|_{\text{op}}. \tag{2.33}$$

It is invariant under inversion and congruence, and satisfies

$$d_T(\eta A + (1-\eta)B, \eta A' + (1-\eta)B') \leq \max\{d_T(A, A'), d_T(B, B')\}. \tag{2.34}$$

Let

$$L_t := \max\{d_T(X_t, I), d_T(X_{t-1}, I)\}. \tag{2.35}$$

Then

$$e^{-L_t} I \preceq X_t, X_{t-1} \preceq e^{L_t} I. \tag{2.36}$$

Set $Q_t = \eta X_t + (1-\eta) X_{t-1}$ and

$$\xi_t := \|Q_t^{-1/2} R_t Q_t^{-1/2}\| \leq e^{L_t} \|R_t\|. \tag{2.37}$$

Whenever $\xi_t < 1$,

$$d_T(Q_t + R_t, Q_t) \leq \log \frac{1+\xi_t}{1-\xi_t}. \tag{2.38}$$

Applying these two properties to (2.29) and then using the preceding perturbation bound gives

$$L_{t+1} \leq L_t + \log \frac{1 + e^{L_t}\|R_t\|}{1 - e^{L_t}\|R_t\|}. \tag{2.39}$$

Writing $\rho_t = e^{-L_t}$ gives

$$\rho_{t+1} \geq \rho_t \frac{\rho_t - \|R_t\|}{\rho_t + \|R_t\|} \geq \rho_t - 2\|R_t\|. \tag{2.40}$$

Choose the total perturbation so that

$$2 \sum_{s=1}^{\infty} \|R_s\| < \frac{\rho_1}{2}. \tag{2.41}$$

We now close the estimate by induction. If $\rho_j \geq \rho_1/2$ through time t, then $\|R_t\| < \rho_t$, so $\xi_t < 1$ and the preceding inequality is valid. It gives

$$\rho_{t+1} \geq \rho_1 - 2\sum_{s=1}^{t} \|R_s\| \geq \frac{\rho_1}{2}. \tag{2.42}$$

Thus ρ_t remains uniformly positive for every t, and hence L_t is uniformly bounded. This is equivalent to a common spectral band $cI \preceq X_t \preceq CI$. Rescaling X_t yields $C_t = \Theta(\sqrt{h_t})$. □

Proposition 2.3 (Active-space covariance recurrence and scaling)**.** *Let H_0 be a fixed invertible symmetric operator on a finite-dimensional active space. Suppose that $C_t \succ 0$ satisfies*

$$C_{t+1} = h_{t+1}\left[H_0\{\beta C_t + (1-\beta)C_{t-1}\}H_0 + F_t\right]^{-1}, \qquad \|F_t\| = O(h_t), \tag{2.43}$$

where $\sqrt{h_t} = \alpha r^t$. Assume that two consecutive normalized initial matrices have spectra bounded above and away from zero, and take α small enough that the total normalized perturbation is sufficiently small. After a fixed congruence by H_0 and normalization by $\sqrt{h_t}$, the sequence satisfies

$$X_{t+1} = J\{\eta X_t + (1-\eta)X_{t-1} + R_t\}^{-1}J, \qquad \|R_t\| = O(\sqrt{h_t}), \tag{2.44}$$

where J is a signature matrix and $\eta \in (0,1)$ is constant. Since $\sum_t \|R_t\| < \infty$, the normalized sequence remains bounded above and away from zero. Consequently, $C_t = \Theta(\sqrt{h_t})$.

Proof. Diagonalize $H_0 = U\Lambda U^\top$ and set

$$W = U|\Lambda|^{1/2}, \qquad J = \operatorname{sign}(\Lambda), \qquad Y_t = W^\top C_t W. \tag{2.45}$$

Since $H_0 = WJW^\top$, recurrence (2.43) becomes

$$Y_{t+1} = h_{t+1}J\{\beta Y_t + (1-\beta)Y_{t-1} + \widetilde{F}_t\}^{-1}J, \tag{2.46}$$

where

$$\widetilde{F}_t = JW^{-1}F_tW^{-T}J = O(h_t). \tag{2.47}$$

Proposition 2.2 reduces this to (2.29), whose unperturbed update is independent of t and whose perturbation is summable of order $O(\sqrt{h_t})$. Lemma 2.15 gives $Y_t = \Theta(\sqrt{h_t})$. Because W is fixed and invertible, the same order holds for C_t. □

2.C.3.6 Full-Rank Projected Hessian

Lemma 2.16 (Gradient-aligned Schur-complement decomposition)**.** *Let $u = g/\|g\|$, $Q = I - uu^\top$, and write the momentum metric in the basis* $\mathrm{span}(u) \oplus g^\perp$ *as*

$$M_{t+1} = \begin{bmatrix} a_t & d_t^\top \\ d_t & D_t \end{bmatrix}, \qquad S_t := D_t - \frac{d_t d_t^\top}{a_t}. \tag{2.48}$$

Whenever truncation is inactive on this block,

$$\Sigma_{t+1} = h_{t+1} \begin{bmatrix} a_t^{-1} + a_t^{-2} d_t^\top S_t^{-1} d_t & -a_t^{-1} d_t^\top S_t^{-1} \\ -S_t^{-1} d_t a_t^{-1} & S_t^{-1} \end{bmatrix}. \tag{2.49}$$

Suppose that, for $s = t, t-1$, $u^\top \Sigma_s u = \Theta(h_s)$, $Q\Sigma_s Q = \Theta(\sqrt{h_s})$, and $\|Q\Sigma_s u\| = O(\sqrt{h_s})$. Then

$$a_t = \|g\|^2 + O(\sqrt{h_t}), \qquad \|d_t\| = O(\sqrt{h_t}), \tag{2.50}$$

$$S_t = H_Q\{\beta Q\Sigma_t Q + (1-\beta) Q\Sigma_{t-1} Q\} H_Q + F_t, \qquad \|F_t\| = O(h_t), \tag{2.51}$$

where $H_Q = (QHQ)|_{g^\perp}$.

Proof. The inverse formula is the ordinary Schur complement with pivot a_t. For the block estimates, apply Lemma 2.3 to μ_t and μ_{t-1} and average with the momentum weights. The rank-one term $gg^\top$ contributes $\|g\|^2$ only to a_t. Every remaining term containing the gradient covariance or a Taylor remainder is $O(\sqrt{h_t})$ in the gradient cross block and $O(h_t)$ after taking the Schur complement. The only order-$\sqrt{h_t}$ contribution to S_t is the displayed homogeneous Hessian term. □

Corollary 2.3 (Full-Rank Initialization)**.** *Suppose that $\Sigma_0 = \alpha I$ and $h_1 = \alpha^2$. If H_Q is invertible, then*

$$u^\top \Sigma_1 u = \frac{\alpha^2}{\|g\|^2} + O(\alpha^3), \quad Q\Sigma_1 Q = \alpha (QH^2Q)^{-1} + O(\alpha^2), \quad \|Q\Sigma_1 u\| = O(\alpha^2), \tag{2.52}$$

and the same three orders hold at the second initialization step.

Proof. At the isotropic initialization, $a_0 = \|g\|^2 + O(\alpha)$, $d_0 = O(\alpha)$, and $S_0 = \alpha QH^2Q + O(\alpha^2)$. Substituting these expressions into (2.49) gives the result. At the next step the same calculation applies, with $H_Q \Sigma_1 H_Q$ as the leading projected term. □

Proof of Theorem 2.3. Corollary 2.3 supplies the initial orders. Assume that they hold through iteration t. By Lemma 2.16, a_t remains bounded above and away from zero and the gradient Schur complement satisfies

$$S_t = H_Q\{\beta C_t + (1-\beta)C_{t-1}\}H_Q + F_t, \qquad C_t := Q\Sigma_t Q, \qquad \|F_t\| = O(h_t). \tag{2.53}$$

Because H_Q is invertible, Proposition 2.3 gives $C_{t+1} = \Theta(\sqrt{h_{t+1}})$. These active covariance eigenvalues and the gradient covariance eigenvalue tend to zero, so for α sufficiently small they remain strictly below the fixed cap ζ. Thus truncation is inactive on the full-rank spectral block throughout the induction. The upper-left entry of (2.49) is $\Theta(h_{t+1})$, since $d_t^\top S_t^{-1} d_t = O(\sqrt{h_t})$. The off-diagonal entry is

$$O\left(h_{t+1}\|S_t^{-1}\|\|d_t\|\right) = O(h_{t+1}), \tag{2.54}$$

which is smaller than the asserted $O(\sqrt{h_{t+1}})$ induction bound. This closes the induction.

There is one covariance eigenvalue of order h_t and $D-1$ eigenvalues of order $\sqrt{h_t}$. Hence

$$\sqrt{\det \Sigma_t} = \Theta\left(h_t^{(D+1)/4}\right), \qquad W(\mu_t) = O(h_t), \tag{2.55}$$

and Theorem 2.1 gives the stated learning rate. □

2.C.3.7 CIL Momentum Analysis

Let $u := g/\|g\|$ and

$$\mathcal{A} := \operatorname{span}(u) \oplus \mathcal{E}_x = \mathcal{K}_x^\perp.$$

For a symmetric matrix A, write its block decomposition on $\operatorname{span}(u) \oplus u^\perp$ as

$$A = \begin{bmatrix} a_A & b_A^\top \\ b_A & D_A \end{bmatrix}, \qquad \bar{A} := D_A - \frac{b_A b_A^\top}{a_A}$$

whenever $a_A > 0$. Thus $\bar{A}$ is the Schur complement of the gradient block. We also set

$$C_t := \pi_{\mathcal{E}_x} \Sigma_t \pi_{\mathcal{E}_x}.$$

For $v \in \mathcal{K}_x$ with $\|v\| < \rho$, define

$$H_{\mathcal{E}}(v) := \pi_{\mathcal{E}_x} Q \nabla^2 f(x^* + v) Q \pi_{\mathcal{E}_x}\big|_{\mathcal{E}_x}, \qquad H_0 := H_{\mathcal{E}}(0).$$

The maximal-rank hypothesis makes H_0 invertible on $\mathcal{E}_x$. Choose an even smooth cutoff χ on $\mathcal{K}_x$ such that $\chi(v) = 1$ for $\|v\| \le \rho/4$ and $\chi(v) = 0$ for $\|v\| \ge \rho/2$. For $\nu_\zeta = N(0, \zeta I_{\mathcal{K}_x})$, define

$$\Phi_\zeta(B) := \int \chi(v) H_{\mathcal{E}}(v) B H_{\mathcal{E}}(v)\, d\nu_\zeta(v). \tag{2.56}$$

The main idea of our argument is to sufficiently localize so that the fixed-congruence dynamics dominate up to a negligible error.

Lemma 2.17 (Small-cap integrated-Hessian expansion)**.** *There are $\zeta_0, C > 0$ such that, for $0 < \zeta \le \zeta_0$ and every symmetric B on $\mathcal{E}_x$,*

$$\Phi_\zeta(B) = H_0 B H_0 + \Delta_\zeta(B), \qquad \|\Delta_\zeta(B)\| \le C\zeta\|B\|.$$

Proof. On the support of χ, Taylor expansion gives

$$H_{\mathcal{E}}(v) = H_0 + L(v) + R(v), \qquad \|R(v)\| \le C\|v\|^2,$$

where L is linear. Since both χ and ν_ζ are even, the two terms that are linear in $L(v)$ integrate to zero. Every remaining term contains either two factors of $L(v)$ or at least one factor of $R(v)$, and hence is bounded by $C\mathbb{E}\|V\|^2\|B\| = O(\zeta\|B\|)$ for $V \sim \nu_\zeta$. Finally, $\int \chi\, d\nu_\zeta = 1 + O(\exp(-c\rho^2/\zeta))$; this exponential term is also $O(\zeta)$ for sufficiently small ζ. □

For

$$B_t := \beta C_t + (1-\beta) C_{t-1}, \qquad C_t := \pi_{\mathcal{E}_x} \Sigma_t \pi_{\mathcal{E}_x},$$

we first prove

$$\pi_{\mathcal{E}_x} \bar{M}_{t+1} \pi_{\mathcal{E}_x} = H_0 B_t H_0 + F_t, \qquad \|F_t\| \le C\{\zeta_n \sqrt{h_t} + h_t + \varepsilon_n\}. \tag{2.57}$$

Here $H_0 = (QHQ)|_{\mathcal{E}_x}$ is invertible by the maximal-rank assumption, and ε_n is the nonlocal Gaussian-tail bound from Lemma 2.12.

Write

$$q_{r,\beta} := \frac{\beta}{r} + \frac{1-\beta}{r^2}, \qquad c_D(r,\beta) := \frac{1}{D q_{r,\beta}}, \qquad \eta := \frac{\beta/r}{q_{r,\beta}} = \frac{\beta r}{1 - \beta + \beta r}.$$

If $H_0 = U \Lambda U^\top$, set

$$W := U|\Lambda|^{1/2}, \qquad J := \operatorname{sign}(\Lambda), \qquad X_t := \frac{W^\top C_t W}{\sqrt{c_D(r,\beta) h_t}}.$$

We then derive

$$X_{t+1} = J\{\eta X_t + (1-\eta)X_{t-1} + R_t\}^{-1}J \tag{2.58}$$

with

$$\|R_t\| \le C\left\{\zeta_n + \sqrt{h_t} + \frac{\varepsilon_n}{\sqrt{h_t}}\right\}. \tag{2.59}$$

Once (2.58)– (2.59) are established, the bound

$$\sum_{t=2}^{t_n} \|R_t\| < \delta_0$$

for the constant δ_0 in Lemma 2.15 gives $cI \preceq X_t \preceq CI$ for $2 \le t \le t_n$.

For $s = t, t-1$, write the covariance in the decomposition $\mathcal{A} \oplus \mathcal{K}_x$ as

$$\Sigma_s = \begin{bmatrix} \Sigma_{\mathcal{A}\mathcal{A},s} & \Sigma_{\mathcal{A}\mathcal{K},s} \\ \Sigma_{\mathcal{K}\mathcal{A},s} & K_s \end{bmatrix},$$

where

$$K_s := \pi_{\mathcal{K}_x}\Sigma_s\pi_{\mathcal{K}_x}, \qquad \Sigma_{\mathcal{A}\mathcal{K},s} := \pi_{\mathcal{A}}\Sigma_s\pi_{\mathcal{K}_x}.$$

Lemma 2.18 (Localized CIL Schur-complement expansion)**.** *Fix n and an integer $T \ge 2$. Write $\zeta = \zeta_n$ and $\varepsilon = \varepsilon_n$ as in Lemma 2.12. Suppose $\varepsilon \le h_T$ and, for $s = t, t-1 \le T$,*

$$\begin{aligned} u^\top\Sigma_s u &= \Theta(h_s), & c\sqrt{h_s}I_{\mathcal{E}_x} \preceq C_s \preceq C\sqrt{h_s}I_{\mathcal{E}_x}, \\ K_s &= \zeta I_{\mathcal{K}_x} + O(h_s+\varepsilon), & \|\Sigma_{\mathcal{A}\mathcal{K},s}\| \le C(\zeta\sqrt{h_s} + h_s + \varepsilon), \\ \|\pi_{\mathcal{E}_x}\Sigma_s u\| &= O(h_s+\varepsilon). \end{aligned} \tag{2.60}$$

Assume also that $h_s \le c_0\zeta^2$. Then, uniformly for $t < T$,

$$u^\top M_{t+1}u = \Theta(1), \qquad \|\pi_{u^\perp}M_{t+1}u\| = O(\sqrt{h_t}), \tag{2.61}$$

$$\pi_{\mathcal{E}_x}\bar{M}_{t+1}\pi_{\mathcal{E}_x} = H_0B_tH_0 + F_t, \qquad \|F_t\| \le C(\zeta\sqrt{h_t} + h_t + \varepsilon), \tag{2.62}$$

$$\|\pi_{\mathcal{E}_x}\bar{M}_{t+1}\pi_{\mathcal{K}_x}\| \le C(h_t+\varepsilon), \qquad 0 \preceq \pi_{\mathcal{K}_x}\bar{M}_{t+1}\pi_{\mathcal{K}_x} \preceq C(h_t+\varepsilon)I, \tag{2.63}$$

where

$$B_t := \beta C_t + (1-\beta)C_{t-1}.$$

Proof. For a single iteration $s \in \{t, t-1\}$, set

$$V_s := \pi_{\mathcal{K}_x}(X - x^*) \sim N(0, K_s).$$

The conditional distribution of the $\mathcal{A}$ component is Gaussian with mean and covariance

$$m_s(v) := \mathbb{E}[\pi_{\mathcal{A}}(X - x^*) \mid V_s = v] = \Sigma_{\mathcal{A}\mathcal{K},s} K_s^{-1} v, \tag{2.64}$$

$$\begin{aligned}\Sigma_{\mathcal{A}|\mathcal{K},s} &:= \mathrm{Cov}(\pi_{\mathcal{A}}(X - x^*) \mid V_s = v) \\ &= \Sigma_{\mathcal{A}\mathcal{A},s} - \Sigma_{\mathcal{A}\mathcal{K},s} K_s^{-1} \Sigma_{\mathcal{K}\mathcal{A},s}.\end{aligned} \tag{2.65}$$

Because $h_s \le c_0\zeta^2$, $\varepsilon \le h_s$, and $K_s = \zeta I + O(h_s + \varepsilon)$, the eigenvalues of K_s lie in $[\zeta/2, 2\zeta]$ for sufficiently small c_0. Hence $\|K_s^{-1}\| \le 2/\zeta$. Moreover,

$$\|\Sigma_{\mathcal{A}\mathcal{K},s}\| \le C(\zeta\sqrt{h_s} + h_s + \varepsilon) \le C'\zeta\sqrt{h_s},$$

where the final inequality again uses $h_s \le c_0\zeta^2$ and $\varepsilon \le h_s$. It follows that

$$\left\|\Sigma_{\mathcal{A}\mathcal{K},s} K_s^{-1} \Sigma_{\mathcal{K}\mathcal{A},s}\right\| \le \|\Sigma_{\mathcal{A}\mathcal{K},s}\|^2 \|K_s^{-1}\| = O(\zeta h_s) = O(h_s), \tag{2.66}$$

$$\mathbb{E}\|m_s(V_s)\|^2 = \mathrm{tr}\left(\Sigma_{\mathcal{A}\mathcal{K},s} K_s^{-1} \Sigma_{\mathcal{K}\mathcal{A},s}\right) = O(h_s). \tag{2.67}$$

Thus

$$\Sigma_{\mathcal{A}|\mathcal{K},s} = \Sigma_{\mathcal{A}\mathcal{A},s} + O(h_s).$$

Choose a smooth radial cutoff ψ on $\mathbb{R}^D$ such that $\psi(x) = 1$ on $B(x^*, \rho)$ and $\psi(x) = 0$ outside $B(x^*, 3\rho/2)$. Split each conditional EGOP integral into its ψ-weighted part and its complement. By Lemma 2.12, the probability and all polynomially weighted moments of the complement are $O(\varepsilon)$. Since f and its first four derivatives are globally bounded, every EGOP or Taylor block contributed by this complement has operator norm $O(\varepsilon)$.

On the support of ψ, the local CIL identities give

$$\nabla f(x^* + v) = a(v) g, \qquad Q \nabla^2 f(x^* + v) \pi_{\mathcal{K}_x} = 0,$$

where a is bounded and $a(0) = 1$. Let $H_v := \nabla^2 f(x^* + v)$ and let $\mathcal{L}_{s|v}$ denote the EGOP of the conditional Gaussian, recentered at $x^* + v$. In the decomposition $\mathrm{span}(u) \oplus u^\perp$, write

$$\mathcal{L}_{s|v} = \begin{bmatrix} \ell_s(v) & d_s(v)^\top \\ d_s(v) & D_s(v) \end{bmatrix}.$$

Applying Lemma 2.3, together with (2.66)– (2.67), gives uniformly on the localized region

$$\ell_s(v) = a(v)^2\|g\|^2 + O(\sqrt{h_s}), \qquad \|d_s(v)\| = O(\sqrt{h_s}), \tag{2.68}$$

$$\pi_{\mathcal{E}_x} D_s(v)\pi_{\mathcal{E}_x} = H_{\mathcal{E}}(v)C_s H_{\mathcal{E}}(v) + O(h_s), \qquad \|\pi_{\mathcal{E}_x} D_s(v)\pi_{\mathcal{K}_x}\| = O(h_s), \tag{2.69}$$

$$0 \preceq \pi_{\mathcal{K}_x} D_s(v)\pi_{\mathcal{K}_x} \preceq C h_s I. \tag{2.70}$$

The $O(h_s)$ terms in these displays contain the conditional-mean correction, the linear map T, the gradient-row Schur correction, and the quadratic Taylor remainder. The identity $QH_v\pi_{\mathcal{K}_x} = 0$ removes every leading Hessian contribution involving $\mathcal{K}_x$.

Let

$$\mathcal{L}(\mu_s) = \begin{bmatrix} a_s & b_s^\top \\ b_s & D_s \end{bmatrix}, \qquad \bar{\mathcal{L}}_s := D_s - \frac{b_s b_s^\top}{a_s}.$$

After integrating (2.68)– (2.70) over $V_s \sim N(0, K_s)$ and adding the $O(\varepsilon)$ nonlocal term, we obtain

$$a_s = \Theta(1), \qquad \|b_s\| = O(\sqrt{h_s}), \tag{2.71}$$

$$\pi_{\mathcal{E}_x}\bar{\mathcal{L}}_s\pi_{\mathcal{E}_x} = \int H_{\mathcal{E}}(v)C_s H_{\mathcal{E}}(v)\, d\nu_s(v) + O(h_s + \varepsilon), \tag{2.72}$$

$$\|\pi_{\mathcal{E}_x}\bar{\mathcal{L}}_s\pi_{\mathcal{K}_x}\| \le C(h_s + \varepsilon), \qquad 0 \preceq \pi_{\mathcal{K}_x}\bar{\mathcal{L}}_s\pi_{\mathcal{K}_x} \preceq C(h_s + \varepsilon)I, \tag{2.73}$$

where $\nu_s = N(0, K_s)$. Notice that $\|b_s b_s^\top / a_s\| = O(h_s)$, so taking the gradient Schur complement does not alter the leading active term.

We next replace ν_s by $\nu_\zeta = N(0, \zeta I)$. Couple

$$V_s = K_s^{1/2} Z, \qquad V_\zeta = \sqrt{\zeta} Z, \qquad Z \sim N(0, I_{\mathcal{K}_x}).$$

The square-root map is Lipschitz on matrices with spectra in $[\zeta/2, 2\zeta]$, and therefore

$$\|K_s^{1/2} - \sqrt{\zeta} I\| \le C\frac{h_s + \varepsilon}{\sqrt{\zeta}}, \qquad \mathbb{E}\|V_s - V_\zeta\| \le C\frac{h_s + \varepsilon}{\sqrt{\zeta}}.$$

The localized map $v \mapsto \chi(v)H_{\mathcal{E}}(v)C_s H_{\mathcal{E}}(v)$ is Lipschitz with constant $O(\|C_s\|) = O(\sqrt{h_s})$. Hence

$$\left\| \int H_{\mathcal{E}}(v)C_s H_{\mathcal{E}}(v)\, d\nu_s(v) - \Phi_\zeta(C_s) \right\| \le C\sqrt{h_s}\frac{h_s + \varepsilon}{\sqrt{\zeta}} = O(h_s), \tag{2.74}$$

where the last equality uses $h_s \le c_0\zeta^2$ and $\varepsilon \le h_s$. Lemma 2.17 now yields

$$\pi_{\mathcal{E}_x}\bar{\mathcal{L}}_s\pi_{\mathcal{E}_x} = H_0 C_s H_0 + F_s, \qquad \|F_s\| \le C(\zeta\sqrt{h_s} + h_s + \varepsilon). \tag{2.75}$$

Finally, write the momentum metric as

$$M_{t+1} = \beta\mathcal{L}(\mu_t) + (1-\beta)\mathcal{L}(\mu_{t-1}) = \begin{bmatrix} a_t^M & (b_t^M)^\top \\ b_t^M & D_t^M \end{bmatrix}.$$

Its gradient Schur complement satisfies

$$\begin{aligned} \bar{M}_{t+1} &= D_t^M - \frac{b_t^M (b_t^M)^\top}{a_t^M} \\ &= \beta\bar{\mathcal{L}}_t + (1-\beta)\bar{\mathcal{L}}_{t-1} + \mathcal{R}_t^{\mathrm{sc}}, \end{aligned} \tag{2.76}$$

where

$$\mathcal{R}_t^{\mathrm{sc}} = \beta\frac{b_t b_t^\top}{a_t} + (1-\beta)\frac{b_{t-1}b_{t-1}^\top}{a_{t-1}} - \frac{b_t^M (b_t^M)^\top}{a_t^M}. \tag{2.77}$$

Because the scalar blocks are bounded above and away from zero and all three gradient cross blocks are $O(\sqrt{h_t})$, $\|\mathcal{R}_t^{\mathrm{sc}}\| = O(h_t)$. Combining this identity with (2.75) and $h_{t-1} = h_t/r^2$ proves (2.61)– (2.63). □

Corollary 2.4 (Scalar trace normalization)**.** *Under the assumptions of Lemma 2.18, define*

$$\tau_\zeta := \int \chi(v)\|\nabla f(x^* + v)\|^2 \, d\nu_\zeta(v).$$

Then

$$\tau_\zeta = \|g\|^2 + O(\zeta), \qquad \operatorname{tr}(M_{t+1}) = \tau_\zeta + O(\sqrt{h_t} + \varepsilon_n).$$

For each fixed n, the positive scalar τ_{ζ_n} may therefore be absorbed into the bandwidth scale. The remaining relative normalization error is $O(\sqrt{h_t} + \varepsilon_n)$ and is included in the perturbation of the autonomous recurrence.

Proof. For a single localization, the trace of the conditional rank-one term is

$$\int \chi(v)\|\nabla f(x^* + v)\|^2 \, d\nu_s(v).$$

The same coupling used in (2.74) changes this integral by $O(h_s + \varepsilon_n)$. The Hessian and Taylor terms have trace $O(\sqrt{h_s})$ under the bounds in (2.60), while the nonlocal contribution is $O(\varepsilon_n)$. Taking the momentum average gives the asserted trace expansion. The identity $\tau_\zeta = \|g\|^2 + O(\zeta)$ follows by the even-cutoff Taylor argument used in Lemma 2.17. □

Corollary 2.5 (Initialization at iterations 2 and 3)**.** *Fix n and write $\zeta := \zeta_n$. Suppose $\Sigma_0 = \alpha I$, $h_1 = \alpha^2$, the first two updates use $\beta = 1$, and $\alpha \le c_\alpha \zeta$. If $\varepsilon_n \le c h_2$, then, for ζ and c_α sufficiently small, the pair (Σ_2, Σ_3) satisfies* (2.60)*, with constants independent of n.*

Proof. At the isotropic initialization, Lemma 2.3 and the localization bound give

$$\mathcal{L}(\mu_0) = gg^\top + \alpha H^2 + gT(\alpha I)^\top + T(\alpha I)g^\top + O(\alpha^2 + \varepsilon_n).$$

The gradient eigenvalue is $\Theta(1)$. After taking its Schur complement, the active $\mathcal{E}_x$ block is uniformly positive definite of order α: for every unit $e \in \mathcal{E}_x$,

$$e^\top \pi_{\mathcal{E}_x} H^2 \pi_{\mathcal{E}_x} e = \|He\|^2 \ge \|QHe\|^2,$$

and QHQ is invertible on $\mathcal{E}_x$. Thus the first update gives

$$u^\top \Sigma_1 u = \Theta(h_1), \qquad \pi_{\mathcal{E}_x} \Sigma_1 \pi_{\mathcal{E}_x} = \Theta\left(\frac{h_1}{\alpha}\right) = \Theta(\alpha) = \Theta(\sqrt{h_1}).$$

The true Null directions have metric eigenvalues $O(\alpha^2 + \varepsilon_n)$ and are therefore capped immediately.

Let

$$\mathcal{K}_x^\circ := \mathcal{K}_x \cap \mathrm{Null}^\perp,$$

which has dimension at most one. A direction $v \in \mathcal{K}_x^\circ$ may remain uncapped at the first update, because Hv can be a nonzero multiple of g. At the next iteration, however,

$$(Hv)^\top \Sigma_1 (Hv) = O(g^\top \Sigma_1 g) = O(h_1).$$

Consequently, after the gradient Schur complement, the full $\mathcal{K}_x \times \mathcal{K}_x$ metric block of $\mathcal{L}(\mu_1)$ is $O(h_1 + \varepsilon_n)$, and its $\mathcal{E}_x \times \mathcal{K}_x$ block has the same order. The active block remains $\Theta(\alpha) = \Theta(\sqrt{h_1})$.

The threshold used to form Σ_2 is

$$\theta_1 := \frac{h_2}{D\zeta} = \frac{r^2\alpha^2}{D\zeta}.$$

Choose ζ sufficiently small and then c_α sufficiently small. Since $\varepsilon_n \le ch_2$, these choices ensure

$$C(h_1 + \varepsilon_n) < \theta_1 < c\alpha.$$

Thus every eigenvalue associated with $\mathcal{K}_x$ lies below θ_1, while every eigenvalue associated with $\mathcal{E}_x$ lies above θ_1. The gradient and active spectral gaps are fixed multiples of 1 and α, respectively, so the two graph transforms have angles $O(\alpha) = O(\sqrt{h_1})$. Rotating back to the fixed decomposition gives

$$\begin{aligned} u^\top \Sigma_2 u &= \Theta(h_2), & C_2 &= \Theta(\sqrt{h_2}), \\ K_2 &= \zeta I + O(h_2 + \varepsilon_n), & \|\pi_{\mathcal{E}_x}\Sigma_2 u\| &= O(h_2 + \varepsilon_n), \\ \|\Sigma_{\mathcal{AK},2}\| &\le C(\zeta\sqrt{h_2} + h_2 + \varepsilon_n). \end{aligned} \tag{2.78}$$

It remains to check the first momentum update. In the decomposition $\mathrm{span}(u) \oplus u^\perp$, write

$$\mathcal{L}(\mu_s) = \begin{bmatrix} a_s & b_s^\top \\ b_s & D_s \end{bmatrix}, \qquad s = 1, 2.$$

Here $a_s = \Theta(1)$ and $\|b_s\| = O(\sqrt{h_1})$. Therefore

$$\bar{M}_3 = \beta\bar{\mathcal{L}}_2 + (1-\beta)\bar{\mathcal{L}}_1 + O(h_1 + \varepsilon_n). \tag{2.79}$$

The active Schur blocks of the two summands are respectively $\Theta(\sqrt{h_2})$ and $\Theta(\sqrt{h_1})$, which have the same order because $h_2 = r^2 h_1$. Their positive weighted sum is therefore $\Theta(\sqrt{h_2})$. The projected-null and active–projected-null blocks are $O(h_1 + \varepsilon_n) = O(h_2 + \varepsilon_n)$. The same threshold comparison and graph-transform argument used above gives the bounds in (2.60) for Σ_3. Hence (Σ_2, Σ_3) initializes the induction. □

Lemma 2.19 (Spectral separation and truncation)**.** *Fix n and an integer $T \ge 2$, and write $\zeta := \zeta_n$. Suppose* (2.60) *holds at iterations $t-1, t \le T$, and suppose $\varepsilon_n \le h_T$. Let*

$$m := \dim(\mathcal{E}_x) = 2d - 2, \qquad k := \dim(\mathcal{K}_x) = D - 2d + 1, \qquad \theta_t := \frac{h_{t+1}}{D\zeta}.$$

There exist constants

$$c_g, C_g, c_E, C_E, C_K > 0,$$

independent of n, t, T*, such that, for* ζ *and* α/ζ *sufficiently small,* M_{t+1} *has spectral projectors* $P_{g,t+1}$*,* $P_{E,t+1}$*, and* $P_{\mathcal{K},t+1}$ *satisfying*

$$P_{g,t+1} + P_{E,t+1} + P_{\mathcal{K},t+1} = I,$$

$$\operatorname{rank}(P_{g,t+1}) = 1, \qquad \operatorname{rank}(P_{E,t+1}) = m, \qquad \operatorname{rank}(P_{\mathcal{K},t+1}) = k,$$

and

$$\sigma\left(M_{t+1}\big|_{\operatorname{range}(P_{g,t+1})}\right) \subseteq [c_g, C_g], \tag{2.80}$$

$$\sigma\left(M_{t+1}\big|_{\operatorname{range}(P_{E,t+1})}\right) \subseteq [c_E\sqrt{h_t}, C_E\sqrt{h_t}], \tag{2.81}$$

$$\sigma\left(M_{t+1}\big|_{\operatorname{range}(P_{\mathcal{K},t+1})}\right) \subseteq [0, C_K(h_t + \varepsilon_n)]. \tag{2.82}$$

Moreover,

$$\|P_{g,t+1} - uu^\top\| = O(\sqrt{h_t}), \tag{2.83}$$

$$\|P_{E,t+1} - \pi_{\mathcal{E}_x}\| = O(\sqrt{h_t}), \tag{2.84}$$

$$\|P_{\mathcal{K},t+1} - \pi_{\mathcal{K}_x}\| = O(\sqrt{h_t}). \tag{2.85}$$

The threshold satisfies $C_K(h_t + \varepsilon_n) < \theta_t < c_E\sqrt{h_t}$*, hence*

$$P_{\mathcal{K},t+1} = \mathbf{1}_{[0,\theta_t)}(M_{t+1}).$$

Therefore the update $\Sigma_{t+1} = \mathcal{T}_{h_{t+1},\zeta}(M_{t+1})$ *satisfies*

$$u^\top \Sigma_{t+1} u = \Theta(h_{t+1}), \qquad c\sqrt{h_{t+1}} I_{\mathcal{E}_x} \preceq C_{t+1} \preceq C\sqrt{h_{t+1}} I_{\mathcal{E}_x}, \tag{2.86}$$

$$K_{t+1} = \zeta I_{\mathcal{K}_x} + O(h_{t+1} + \varepsilon_n), \qquad \|\pi_{\mathcal{E}_x} \Sigma_{t+1} u\| = O(h_{t+1} + \varepsilon_n),$$

$$\|\Sigma_{\mathcal{A}\mathcal{K},t+1}\| \le C\left\{\zeta\sqrt{h_{t+1}} + h_{t+1} + \varepsilon_n\right\}. \tag{2.87}$$

In particular, (2.60) *holds at iteration* $t+1$*.*

Proof. Write the momentum metric in the decomposition $\operatorname{span}(u) \oplus u^\perp$ as

$$M_{t+1} = \begin{bmatrix} a_t & b_t^\top \\ b_t & D_t \end{bmatrix}, \qquad \bar{M}_{t+1} := D_t - \frac{b_t b_t^\top}{a_t}.$$

Lemma 2.18 gives

$$c_a \le a_t \le C_a, \qquad \|b_t\| = O(\sqrt{h_t}), \qquad \|D_t\| = O(\sqrt{h_t}).$$

We first separate the eigenvalue associated with the gradient direction. The invariant subspace complementary to this eigenvalue is the graph of a map

$$Y_t : u^\perp \longrightarrow \mathrm{span}(u)$$

satisfying

$$a_t Y_t - Y_t D_t + b_t^\top - Y_t b_t Y_t = 0.$$

The linearized Sylvester operator

$$Y \longmapsto a_t Y - Y D_t$$

has a uniformly bounded inverse because a_t is bounded away from zero and $\|D_t\| = o(1)$. Its fixed-point equation is therefore a contraction on a ball of radius $C\sqrt{h_t}$, yielding

$$\|Y_t\| = O(\sqrt{h_t}), \qquad Y_t = -a_t^{-1} b_t^\top + O(h_t).$$

The associated orthogonal change of coordinates separates the gradient eigenvalue and reduces the operator on $u^\perp$ to

$$\widetilde{M}_{t+1} = \bar{M}_{t+1} + E_{g,t}, \qquad \|E_{g,t}\| = O(h_t^{3/2}). \tag{2.88}$$

Decompose $u^\perp = \mathcal{E}_x \oplus \mathcal{K}_x$ and write

$$\widetilde{M}_{t+1} = \begin{bmatrix} \mathsf{A}_t & \mathsf{C}_t \\ \mathsf{C}_t^\top & \mathsf{K}_t \end{bmatrix}.$$

Equations (2.62)– (2.63) and (2.88) imply

$$\mathsf{A}_t = H_0 B_t H_0 + \widehat{F}_t, \qquad \|\widehat{F}_t\| \le C\left\{\zeta\sqrt{h_t} + h_t + \varepsilon_n\right\}, \tag{2.89}$$

$$\|\mathsf{C}_t\| \le C(h_t + \varepsilon_n), \qquad 0 \preceq \mathsf{K}_t \preceq C(h_t + \varepsilon_n) I. \tag{2.90}$$

Here

$$B_t := \beta C_t + (1-\beta) C_{t-1}.$$

The bounds in (2.60) and $h_{t-1} = h_t/r^2$ give

$$c_B\sqrt{h_t}I_{\mathcal{E}_x} \preceq B_t \preceq C_B\sqrt{h_t}I_{\mathcal{E}_x}.$$

Since H_0 is invertible, reducing ζ, α/ζ, and the tail level if necessary gives

$$c_E\sqrt{h_t}I_{\mathcal{E}_x} \preceq \mathsf{A}_t \preceq C_E\sqrt{h_t}I_{\mathcal{E}_x}.$$

Because $\varepsilon_n \le h_T \le h_t$,

$$\|\mathsf{C}_t\| = O(h_t), \qquad 0 \preceq \mathsf{K}_t \preceq C_K h_t I.$$

The distance between the spectra of A_t and K_t is therefore at least $c\sqrt{h_t}$. The invariant subspace associated with the k smallest eigenvalues is the graph of a map

$$Z_t : \mathcal{K}_x \longrightarrow \mathcal{E}_x$$

satisfying

$$\mathsf{A}_t Z_t - Z_t\mathsf{K}_t + \mathsf{C}_t - Z_t\mathsf{C}_t^\top Z_t = 0.$$

The inverse of the linearized Sylvester operator has norm $O(h_t^{-1/2})$, and hence

$$\|Z_t\| = O(h_t^{-1/2})\,O(h_t) = O(\sqrt{h_t}).$$

The corresponding orthogonal change of coordinates makes $\widetilde{M}_{t+1}$ block diagonal, changing each diagonal block by

$$O(\|\mathsf{C}_t\|\,\|Z_t\|) = O(h_t^{3/2}).$$

It follows that the non-gradient eigenvalues consist of m eigenvalues in $[c_E\sqrt{h_t}, C_E\sqrt{h_t}]$ and k eigenvalues in $[0, C_K(h_t + \varepsilon_n)]$. Combining the two changes of coordinates gives (2.85).

The truncation threshold is $\theta_t = \frac{h_{t+1}}{D\zeta}$. Since $\varepsilon_n \le h_t$,

$$C_K(h_t + \varepsilon_n) \le 2C_K h_t.$$

Choosing ζ so that $2C_K < \frac{r^2}{D\zeta}$ gives

$$C_K(h_t + \varepsilon_n) < \theta_t.$$

For the other inequality, $\theta_t < c_E\sqrt{h_t}$ is equivalent to

$$\sqrt{h_t} < \frac{Dc_E}{r^2}\zeta.$$

Since

$$\sqrt{h_t} \le \sqrt{h_1} = \alpha,$$

this holds whenever α/ζ is sufficiently small. Thus

$$P_{\mathcal{K},t+1} = \mathbf{1}_{[0,\theta_t)}(M_{t+1}),$$

and the truncated inverse assigns covariance ζ exactly on range$(P_{\mathcal{K},t+1})$. On the gradient and active spectral subspaces, inversion gives covariance orders h_{t+1} and

$$\frac{h_{t+1}}{\sqrt{h_t}} = r\sqrt{h_{t+1}},$$

respectively. Rotating these covariances back to

$$\mathrm{span}(u) \oplus \mathcal{E}_x \oplus \mathcal{K}_x$$

using (2.85) yields (2.86) and (2.87). The term $\zeta\sqrt{h_{t+1}}$ in (2.87) is the first-order off-diagonal contribution of the capped projector. Its contribution to either fixed diagonal block is

$$O(\zeta h_{t+1}) = O(h_{t+1}),$$

because $\zeta \le 1$. □

Lemma 2.20 (Reduction to the normalized active recurrence)**.** *Under the hypotheses of Lemma 2.19, there exist symmetric matrices $\widehat{F}_t$ and G_t such that*

$$C_{t+1} = \frac{h_{t+1}}{D}\left[H_0 B_t H_0 + \widehat{F}_t\right]^{-1} + G_t, \tag{2.91}$$

where

$$\|\widehat{F}_t\| \le C\left\{\zeta\sqrt{h_t} + h_t + \varepsilon_n\right\}, \tag{2.92}$$

$$\|G_t\| \le C\left\{\zeta h_t + h_t^{3/2} + \varepsilon_n\right\}. \tag{2.93}$$

Let

$$H_0 = U\Lambda U^\top, \qquad W := U|\Lambda|^{1/2}, \qquad J := \operatorname{sign}(\Lambda),$$

and define

$$X_t := \frac{W^\top C_t W}{\sqrt{c_D(r,\beta)h_t}}.$$

Then there exists a symmetric matrix R_t such that

$$X_{t+1} = J\left\{\eta X_t + (1-\eta)X_{t-1} + R_t\right\}^{-1} J, \tag{2.94}$$

and

$$\|R_t\| \le C\left\{\zeta + \sqrt{h_t} + \frac{\varepsilon_n}{\sqrt{h_t}}\right\}. \tag{2.95}$$

Proof. The spectral decomposition in Lemma 2.19 shows that the fixed-coordinate active covariance differs from the inverse of its active metric block for two reasons.

First, the active spectral subspace differs from $\mathcal{E}_x$ by an angle $O(\sqrt{h_t})$. Rotating a covariance of order $\sqrt{h_{t+1}}$ back to $\mathcal{E}_x$ produces an $O(h_t^{3/2})$ correction to the active diagonal block.

Second, the capped projector contributes

$$\zeta\, \pi_{\mathcal{E}_x} P_{\mathcal{K},t+1} \pi_{\mathcal{E}_x} = O(\zeta h_t),$$

because

$$\|\pi_{\mathcal{E}_x} P_{\mathcal{K},t+1}\| = O(\sqrt{h_t}).$$

The nonlocal metric contribution ε_n induces an $O(\varepsilon_n)$ covariance correction. Combining these terms with (2.89) gives (2.91)– (2.93).

Set

$$Y_t := W^\top C_t W.$$

Since

$$H_0 = WJW^\top,$$

equation (2.91) becomes

$$Y_{t+1} = \frac{h_{t+1}}{D} J\left[\beta Y_t + (1-\beta)Y_{t-1} + \widetilde{F}_t\right]^{-1} J + \widetilde{G}_t, \tag{2.96}$$

where

$$\widetilde{F}_t := JW^{-1}\widehat{F}_t W^{-T} J, \qquad \widetilde{G}_t := W^\top G_t W.$$

Because W is fixed and invertible, $\widetilde{F}_t$ and $\widetilde{G}_t$ satisfy the same respective orders as $\widehat{F}_t$ and G_t.

Substitute

$$Y_s = \sqrt{c_D(r,\beta)h_s}\, X_s.$$

Using

$$\frac{\sqrt{h_t}}{\sqrt{h_{t+1}}} = \frac{1}{r}, \qquad \frac{\sqrt{h_{t-1}}}{\sqrt{h_{t+1}}} = \frac{1}{r^2},$$

and

$$D\, c_D(r,\beta)\, q_{r,\beta} = 1,$$

equation (2.96) gives

$$X_{t+1} = J\left[\eta X_t + (1-\eta)X_{t-1} + R_t^{\text{met}}\right]^{-1} J + E_t, \tag{2.97}$$

where

$$R_t^{\text{met}} := \frac{\widetilde{F}_t}{q_{r,\beta}\sqrt{c_D(r,\beta)h_{t+1}}}, \qquad E_t := \frac{\widetilde{G}_t}{\sqrt{c_D(r,\beta)h_{t+1}}}.$$

Equations (2.92) and (2.93) imply

$$\|R_t^{\text{met}}\| \le C\left\{\zeta + \sqrt{h_t} + \frac{\varepsilon_n}{\sqrt{h_t}}\right\}, \tag{2.98}$$

$$\|E_t\| \le C\left\{\zeta\sqrt{h_t} + h_t + \frac{\varepsilon_n}{\sqrt{h_t}}\right\}. \tag{2.99}$$

Define

$$Q_t := \eta X_t + (1-\eta)X_{t-1}, \qquad R_t := (JX_{t+1}J)^{-1} - Q_t.$$

Then (2.94) holds by definition. To prove (2.95), let

$$\widehat{X}_{t+1} := J(Q_t + R_t^{\text{met}})^{-1}J.$$

By (2.60) and Lemma 2.19, there are constants $0 < c < C < \infty$ such that

$$cI \preceq X_{t+1}, \widehat{X}_{t+1} \preceq CI.$$

Equation (2.97) gives $X_{t+1} - \widehat{X}_{t+1} = E_t$. The inverse identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ therefore yields

$$\|R_t - R_t^{\mathrm{met}}\| = \left\|(JX_{t+1}J)^{-1} - (J\widehat{X}_{t+1}J)^{-1}\right\| \tag{2.100}$$

$$\leq C\|E_t\|. \tag{2.101}$$

Combining this with (2.98) and (2.99) proves (2.95). The residual trace-normalization error has the same or smaller order and is absorbed into R_t. □

Corollary 2.6 (Cumulative normalized recurrence error)**.** *Under the hypotheses of Lemma 2.20, for every integer S satisfying $2 \leq S \leq T$,*

$$\sum_{t=2}^{S} \|R_t\| \leq C\left\{(S-1)\zeta + \frac{\sqrt{h_1}}{1-r} + \frac{(S-1)\varepsilon_n}{\sqrt{h_S}}\right\}. \tag{2.102}$$

Proof. Equation (2.95) gives

$$\sum_{t=2}^{S} \|R_t\| \leq C\left\{(S-1)\zeta + \sum_{t=2}^{S}\sqrt{h_t} + \varepsilon_n \sum_{t=2}^{S}\frac{1}{\sqrt{h_t}}\right\}.$$

Since $\sqrt{h_t} = \sqrt{h_1}\, r^{t-1}$, we have

$$\sum_{t=2}^{S}\sqrt{h_t} \leq \frac{\sqrt{h_1}}{1-r}.$$

Moreover, $\sqrt{h_t}$ decreases in t, so

$$\sum_{t=2}^{S}\frac{1}{\sqrt{h_t}} \leq \frac{S-1}{\sqrt{h_S}}.$$

Substitution gives (2.102). □

Proposition 2.4 (Linearization of the normalized recurrence)**.** *For the unperturbed autonomous map in* (2.94)*, linearization at I gives*

$$E_{t+1} = -\mathcal{J}\{\eta E_t + (1-\eta)E_{t-1}\}, \qquad \mathcal{J}(E) := JEJ.$$

On a mode satisfying $\mathcal{J}(E) = sE$, $s \in \{1, -1\}$, the characteristic polynomials are

$$z^2 + \eta z + (1-\eta), \qquad s = 1,$$

and

$$z^2 - \eta z - (1-\eta) = (z-1)(z+1-\eta), \qquad s = -1.$$

Every root transverse to the fixed-point manifold has modulus strictly below one for $0 < \eta < 1$. *At either endpoint, undamped roots remain.*

Proof. For $s = 1$, the Jury conditions for a monic quadratic z^2+az+b are $|b| < 1$, $1+a+b > 0$, and $1-a+b > 0$. With $a = \eta$ and $b = 1-\eta$, all three inequalities hold strictly for $0 < \eta < 1$. For $s = -1$, the roots are 1 and $-(1-\eta)$. The root 1 is tangent to $\{X \succ 0 : (XJ)^2 = I\}$, while the transverse root has modulus $1 - \eta < 1$. □

Lemma 2.21 (Consecutive covariance comparison and EGOP-form bound)**.** *Fix* n *and write* $\zeta := \zeta_n$. *Under the conclusions of Lemma 2.19, there is a constant* $C < \infty$, *independent of* $t \le T$, *such that*

$$\Sigma_t \preceq C\Sigma_{t+1}.$$

Consequently,

$$W(\mu_t) = \operatorname{tr}\{\Sigma_t \mathcal{L}(\mu_t)\} = O(h_t).$$

Proof. Define

$$\mathcal{D}_s := h_s uu^\top + \sqrt{h_s}\,\pi_{\mathcal{E}_x} + \zeta \pi_{\mathcal{K}_x}$$

and

$$\widehat{\Sigma}_s := \mathcal{D}_s^{-1/2} \Sigma_s \mathcal{D}_s^{-1/2}.$$

In the decomposition $\operatorname{span}(u) \oplus \mathcal{E}_x \oplus \mathcal{K}_x$, the diagonal blocks satisfy

$$u^\top \widehat{\Sigma}_s u = \Theta(1), \tag{2.103}$$

$$cI_{\mathcal{E}_x} \preceq \pi_{\mathcal{E}_x} \widehat{\Sigma}_s \pi_{\mathcal{E}_x} \preceq CI_{\mathcal{E}_x}, \tag{2.104}$$

$$\pi_{\mathcal{K}_x} \widehat{\Sigma}_s \pi_{\mathcal{K}_x} = I_{\mathcal{K}_x} + O\left(\frac{h_s + \varepsilon_n}{\zeta}\right) = I_{\mathcal{K}_x} + O(\zeta), \tag{2.105}$$

where the last equality uses $h_s \le c_0 \zeta^2$ and $\varepsilon_n \le h_s$.

The normalized off-diagonal blocks satisfy

$$\left\| \frac{\pi_{\mathcal{E}_x} \Sigma_s u}{h_s^{3/4}} \right\| \le C \frac{h_s + \varepsilon_n}{h_s^{3/4}} = O(h_s^{1/4}), \tag{2.106}$$

$$\left\| \frac{\pi_{\mathcal{K}_x} \Sigma_s u}{\sqrt{h_s \zeta}} \right\| \le C \frac{\zeta \sqrt{h_s} + h_s + \varepsilon_n}{\sqrt{h_s \zeta}} = O(\sqrt{\zeta}), \tag{2.107}$$

$$\left\| \frac{\pi_{\mathcal{E}_x} \Sigma_s \pi_{\mathcal{K}_x}}{h_s^{1/4} \sqrt{\zeta}} \right\| \le C \frac{\zeta \sqrt{h_s} + h_s + \varepsilon_n}{h_s^{1/4} \sqrt{\zeta}} = O(\zeta). \tag{2.108}$$

Thus, after choosing c_0 and ζ sufficiently small, there are constants $0 < c < C < \infty$ such that

$$cI \preceq \widehat{\Sigma}_s \preceq CI.$$

Equivalently,

$$c\mathcal{D}_s \preceq \Sigma_s \preceq C\mathcal{D}_s. \tag{2.109}$$

Since $h_t = r^{-2} h_{t+1}$,

$$\mathcal{D}_t = r^{-2} h_{t+1} u u^\top + r^{-1} \sqrt{h_{t+1}}\, \pi_{\mathcal{E}_x} + \zeta \pi_{\mathcal{K}_x} \preceq r^{-2} \mathcal{D}_{t+1}.$$

Applying (2.109) at times t and $t+1$ gives

$$\Sigma_t \preceq C\mathcal{D}_t \preceq C r^{-2} \mathcal{D}_{t+1} \preceq C' \Sigma_{t+1}.$$

Moreover, $M_{t+1} = \beta \mathcal{L}(\mu_t) + (1-\beta)\mathcal{L}(\mu_{t-1}) \succeq \beta \mathcal{L}(\mu_t)$, yielding

$$\beta W(\mu_t) \le \operatorname{tr}(\Sigma_t M_{t+1}) \le C' \operatorname{tr}(\Sigma_{t+1} M_{t+1}).$$

Let $\lambda_1, \dots, \lambda_D$ be the eigenvalues of M_{t+1}. The corresponding eigenvalues of $\Sigma_{t+1} = \mathcal{T}_{h_{t+1},\zeta}(M_{t+1})$ are

$$\tau_j = \begin{cases} \min\left\{ \zeta, \dfrac{h_{t+1}}{D\lambda_j} \right\}, & \lambda_j > 0, \\ \zeta, & \lambda_j = 0. \end{cases}$$

Hence

$$0 \le \lambda_j \tau_j \le \frac{h_{t+1}}{D}$$

for every j, and

$$\operatorname{tr}(\Sigma_{t+1} M_{t+1}) = \sum_{j=1}^{D} \lambda_j \tau_j \le h_{t+1} = r^2 h_t.$$

Thus $W(\mu_t) = O(h_t)$. □

Proof of Theorem 2.2. Let

$$k := D - 2d + 1, \qquad b_n := \left(n\zeta_n^{k/2}\right)^{-2/(d+2)}.$$

The stopping time t_n is the smallest integer satisfying $h_{t_n} \le b_n$. Since $h_{t+1} = r^2 h_t$,

$$r^2 b_n < h_{t_n} \le b_n,$$

and hence

$$h_{t_n} \asymp b_n. \tag{2.110}$$

Moreover,

$$h_t = \alpha_n^2 r^{2(t-1)}, \qquad \zeta_n = \frac{c_\zeta}{\log n}, \qquad \alpha_n = c_\alpha \zeta_n,$$

so there is a constant C_T such that

$$t_n \le C_T \log n \tag{2.111}$$

for all sufficiently large n.

Let $\delta_0 > 0$ be a value for which Lemma 2.15 applies to the initial normalized matrices supplied by Corollary 2.5. By Lemma 2.12, for every fixed $A > 0$, c_ζ can be chosen so that

$$\varepsilon_n \le C n^{-A}.$$

Choose A large enough that

$$\frac{\varepsilon_n}{h_{t_n}} \longrightarrow 0, \qquad \frac{t_n \varepsilon_n}{\sqrt{h_{t_n}}} \longrightarrow 0. \tag{2.112}$$

This is possible because (2.110) is a negative power of n multiplied by a power of $\log n$.

Applying (2.102) with $S = t_n$ gives

$$\sum_{t=2}^{t_n} \|R_t\| \le C\left\{(t_n - 1)\zeta_n + \frac{\alpha_n}{1-r} + \frac{(t_n-1)\varepsilon_n}{\sqrt{h_{t_n}}}\right\} \tag{2.113}$$

$$\le C\left\{C_T c_\zeta + \frac{c_\alpha c_\zeta}{(1-r)\log n} + \frac{t_n \varepsilon_n}{\sqrt{h_{t_n}}}\right\}. \tag{2.114}$$

Choose c_ζ so that

$$C C_T c_\zeta < \frac{\delta_0}{3},$$

then choose c_α so that

$$\frac{Cc_\alpha c_\zeta}{1-r} < \frac{\delta_0}{3}.$$

By (2.112), the last term in (2.114) is smaller than $\delta_0/3$ for all sufficiently large n. Therefore

$$\sum_{t=2}^{t_n} \|R_t\| < \delta_0. \tag{2.115}$$

Corollary 2.5 gives (2.60) at $s = 2, 3$. Let T^* be the largest integer in $\{3, \ldots, t_n\}$ such that (2.60) holds for every $2 \le s \le T^*$. Suppose $T^* < t_n$.

For $2 \le t \le T^*$, Lemma 2.19 gives (2.94). Its partial error sum is at most the left side of (2.115). Lemma 2.15 therefore gives constants $0 < c_X < C_X < \infty$, independent of n and T^*, such that

$$c_X I \preceq X_s \preceq C_X I, \qquad 2 \le s \le T^* + 1.$$

Applying Lemma 2.19 at $t = T^*$ also gives all remaining inequalities in (2.60) at $s = T^* + 1$. This contradicts the definition of T^*. Hence

$$T^* = t_n.$$

It follows that, for every $2 \le t \le t_n$,

$$c_X I \preceq X_t \preceq C_X I$$

and (2.60) holds. In particular, there are constants $0 < c < C < \infty$ such that

$$c\left(h_t uu^\top + \sqrt{h_t}\,\pi_{\mathcal{E}_x} + \zeta_n \pi_{\mathcal{K}_x}\right) \preceq \Sigma_t \preceq C\left(h_t uu^\top + \sqrt{h_t}\,\pi_{\mathcal{E}_x} + \zeta_n \pi_{\mathcal{K}_x}\right).$$

Therefore

$$\sqrt{\det \Sigma_t} = \Theta\left(h_t^{d/2} \zeta_n^{k/2}\right).$$

Lemma 2.21 gives $W(\mu_t) = O(h_t)$, hence

$$\frac{1}{n\sqrt{\det \Sigma_t}} + W(\mu_t) = O\left(\frac{1}{n h_t^{d/2} \zeta_n^{k/2}} + h_t\right).$$

By (2.110), the two terms have the same order at $t = t_n$, and

$$\mathbb{E}\left[\int \{\widehat{P}_{M_{t_n}}(f) - f\}^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)} \zeta_n^{-k/(d+2)}\right).$$

Since $\zeta_n = c_\zeta / \log n$,

$$\mathbb{E}\left[\int \{\widehat{P}_{M_{t_n}}(f) - f\}^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)} (\log n)^{(D-2d+1)/(d+2)}\right).$$

Finally, solving $\alpha_n^2 r^{2(t_n-1)} \asymp b_n$ yields

$$t_n = \frac{\log n + (k/2)\log \zeta_n + (d+2)\log \alpha_n}{(d+2)\log(1/r)} + O(1).$$

□

2.D EGOP Approximation

In this section, we show that metric anisotropy not only improves point regression, but accelerates the estimation of the EGOP matrix in our two model settings. We argue that the smoothed gradients, while potentially biased estimators at individual points, in the aggregate form a matrix that differs from the population quantity by a benign perturbation, leading to an equivalent convergence analysis at adequate sample sizes.

Define the smoothed gradient and EGOP as

$$\nabla_\Sigma f(x) = \operatorname*{argmin}_v \min_a \int \|f(z) - a - v(z-x)\|^2 \, dN(x, \Sigma), \quad \mathcal{L}_\Sigma(\mu) = \int \nabla_\Sigma f \nabla_\Sigma f^\top \, d\mu.$$

Define the empirical momentum metric and covariance recurrence by

$$\widehat{M}_{t+1} := \beta \widehat{\mathcal{L}}_{\Sigma_t}(\mu_t) + (1-\beta)\widehat{\mathcal{L}}_{\Sigma_{t-1}}(\mu_{t-1}), \tag{2.116}$$

$$\Sigma_{t+1} := \mathcal{T}_{h_{t+1},\zeta}(\widehat{M}_{t+1}). \tag{2.117}$$

Lemma 2.22 (Smoothed-gradient and EGOP expansion)**.** *There exists T_x, C, $T_x(\Sigma) \le C\|\Sigma\|$ such that, for $\mu = N(x^*, \Sigma)$,*

$$\nabla_\Sigma f = \nabla f + T_x(\Sigma), \quad \mathcal{L}_\Sigma(\mu) = \mathcal{L}(\mu) + \mathbb{E}_\mu[g T_X(\Sigma)^\top + T_X(\Sigma) g^\top] + O(\|\Sigma\|^2).$$

Proof. The standard OLS solution yields

$$\nabla_\Sigma f(x) = \Sigma^{-1} \operatorname{Cov}_{N(x,\Sigma)}(X, f(X)) = \mathbb{E}_{N(x,\Sigma)}[\nabla f(X)],$$

with the last equality following from Stein's Identity (Stein, 1981). Applying a Taylor expansion, we see that

$$\mathbb{E}_{N(x,\Sigma)}[\nabla f(X)] = \int \nabla f(x) + H(y-x) + T_x(y-x, y-x) \, dN(0,\Sigma) =: \nabla f(x) + T_x(\Sigma),$$

where T_x is a second order remainder. For the second claim, the proof follows by a simple expansion,

$$\begin{aligned}\mathcal{L}_\Sigma(\mu) &= \int \nabla_\Sigma f \nabla_\Sigma f^\top \, d\mu \\ &= \int \nabla f \nabla f^\top \, d\mu + \int T_x(\Sigma) \nabla f^\top + \nabla f T_x(\Sigma) \, d\mu \\ &= \mathcal{L}(\mu) + \int T_x(\Sigma) g^\top + g T_x(\Sigma) \, d\mu + \int T_x(\Sigma)(\nabla f - g)^\top + (\nabla f - g) T_x(\Sigma) \, d\mu,\end{aligned}$$

satisfying the claim, with full details being similar to those in the proof of Lemma 2.3. □

The preceding expansions show that the empirical recurrence is controlled by the error in the momentum metric itself. We therefore state the finite-sample guarantees under the direct perturbation condition

$$\|\widehat{M}_{t+1} - M_{t+1}\| \le C h_t \tag{2.118}$$

uniformly through the stopping time. Establishing (2.118) for a particular gradient estimator is a separate nonparametric estimation problem.

Proposition 2.5 (Empirical full-rank Local EGOP rate)**.** *Let* $\operatorname{rank}(\pi_g^\perp H \pi_g^\perp) = D - 1$ *and adopt the assumptions of Theorem 2.3. Suppose that the empirical recurrence* (2.116) *satisfies* (2.118) *uniformly for* $t \le t_n$*. Then, for*

$$t_n = \frac{2 \log n}{(D+5) \log(1/r)}, \tag{2.119}$$

$$\mathbb{E}\left[\int (\hat{P}_{\widehat{M}_{t_n}}(f) - f)^2 \, d\mu_{t_n}\right] = O\left(n^{-\frac{4}{D+5}}\right). \tag{2.120}$$

Proof. The perturbation in (2.118) enters the gradient Schur complement as an additional $O(h_t)$ remainder. It is therefore absorbed into F_t in Lemma 2.16. Proposition 2.3 preserves the h_t gradient scale and the $\sqrt{h_t}$ scale on $g^\perp$. The determinant and risk calculation is then identical to the proof of Theorem 2.3. □

Proposition 2.6 (Empirical Local EGOP rate for continuous-index learning)**.** *Let* (X, Y) *be CIL, suppose that* $\operatorname{rank}(\pi_g^\perp H \pi_g^\perp) = 2d - 2$*, and adopt the assumptions and the choices*

$\zeta_n = c_\zeta/\log n$, $\alpha_n = c_\alpha\zeta_n$ *of Theorem 2.2. Suppose that the empirical recurrence* (2.116) *satisfies* (2.118) *uniformly for* $t \le t_n$. *Then*

$$\mathbb{E}\left[\int (\hat{P}_{\widehat{M}_{t_n}}(f) - f)^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)}(\log n)^{(D-2d+1)/(d+2)}\right).$$

Proof. Set

$$\Delta_{t+1} := \widehat{M}_{t+1} - M_{t+1}, \qquad \|\Delta_{t+1}\| \le Ch_t.$$

Write M_{t+1} and $\widehat{M}_{t+1}$ in the decomposition $\mathrm{span}(u) \oplus u^\perp$. Since the population gradient block is $\Theta(1)$ and its cross block is $O(\sqrt{h_t})$, the ordinary Schur-complement formula gives

$$\left\|\overline{\widehat{M}}_{t+1} - \bar{M}_{t+1}\right\| \le C\left\{\|\Delta_{t+1}\| + \sqrt{h_t}\|\Delta_{t+1}\| + \|\Delta_{t+1}\|^2\right\} = O(h_t).$$

Hence the empirical active, active–projected-null, and projected-null Schur blocks satisfy the same bounds as their population counterparts, with different constants:

$$\pi_{\mathcal{E}_x}\overline{\widehat{M}}_{t+1}\pi_{\mathcal{E}_x} = H_0 B_t H_0 + O\left(\zeta_n\sqrt{h_t} + h_t + \varepsilon_n\right), \tag{2.121}$$

$$\left\|\pi_{\mathcal{E}_x}\overline{\widehat{M}}_{t+1}\pi_{\mathcal{K}_x}\right\| = O(h_t + \varepsilon_n), \tag{2.122}$$

$$0 \preceq \pi_{\mathcal{K}_x}\overline{\widehat{M}}_{t+1}\pi_{\mathcal{K}_x} \preceq C(h_t + \varepsilon_n)I. \tag{2.123}$$

The spectral-separation proof of Lemma 2.19 therefore applies without change. In particular, the empirical active eigenvalues remain of order $\sqrt{h_t}$, the projected-null eigenvalues remain $O(h_t + \varepsilon_n)$, and

$$C(h_t + \varepsilon_n) < \frac{h_{t+1}}{D\zeta_n} < c\sqrt{h_t}.$$

Thus the empirical truncated update caps exactly the projected-null spectral subspace.

In the normalized active recurrence, the additional $O(h_t)$ metric error is divided by the active scale $\sqrt{h_t}$. Consequently, the empirical recurrence can be written as

$$X_{t+1} = J\left\{\eta X_t + (1-\eta)X_{t-1} + R_t + R_t^{\mathrm{emp}}\right\}^{-1} J, \qquad \|R_t^{\mathrm{emp}}\| \le C\sqrt{h_t}.$$

Since $r \in (0,1)$ is fixed independently of n,

$$\sum_{t=2}^{t_n}\|R_t^{\mathrm{emp}}\| \le C\sum_{t=2}^{t_n}\sqrt{h_t} \le \frac{C\alpha_n}{1-r} = O\left(\frac{1}{\log n}\right) = o(1).$$

The population constants c_ζ and c_α are chosen so that $\sum_{t=2}^{t_n} \|R_t\|$ is below the perturbation threshold in Lemma 2.15. The preceding display shows that the empirical contribution is eventually smaller than the remaining perturbation margin. Therefore the same common spectral band holds for the empirical active covariance through t_n.

It follows that, uniformly for $2 \le t \le t_n$, there is one covariance eigenvalue of order h_t, $2d-2$ eigenvalues of order $\sqrt{h_t}$, and $k := D - 2d + 1$ eigenvalues of order ζ_n. Hence

$$\sqrt{\det \Sigma_t} = \Theta\left(h_t^{d/2} \zeta_n^{k/2}\right).$$

It remains to bound the population EGOP form. The covariance block bounds imply

$$\Sigma_t \preceq C \Sigma_{t+1}.$$

Moreover,

$$M_{t+1} \succeq \beta \mathcal{L}(\mu_t),$$

so

$$\beta W(\mu_t) \le C \operatorname{tr}(\Sigma_{t+1} M_{t+1}).$$

Using $M_{t+1} = \widehat{M}_{t+1} - \Delta_{t+1}$ gives

$$\operatorname{tr}(\Sigma_{t+1} M_{t+1}) \le \operatorname{tr}(\Sigma_{t+1} \widehat{M}_{t+1}) + \|\Delta_{t+1}\| \operatorname{tr}(\Sigma_{t+1}) \tag{2.124}$$

$$\le h_{t+1} + C h_t \zeta_n = O(h_t). \tag{2.125}$$

Here the first inequality uses $\Sigma_{t+1} = \mathcal{T}_{h_{t+1},\zeta_n}(\widehat{M}_{t+1})$, so every paired metric–covariance eigenvalue product is at most h_{t+1}/D. Therefore

$$W(\mu_t) = O(h_t).$$

The generic risk bound now gives

$$\mathbb{E}\left[\int \{\widehat{P}_{\widehat{M}_t}(f) - f\}^2 \, d\mu_t\right] = O\left(\frac{1}{n h_t^{d/2} \zeta_n^{k/2}} + h_t\right).$$

At $t = t_n$,

$$h_{t_n} \asymp \left(n \zeta_n^{k/2}\right)^{-2/(d+2)},$$

so

$$\mathbb{E}\left[\int \{\widehat{P}_{\widehat{M}_{t_n}}(f) - f\}^2 \, d\mu_{t_n}\right] = O\left(n^{-2/(d+2)}\zeta_n^{-k/(d+2)}\right).$$

Substituting $\zeta_n = c_\zeta / \log n$ and $k = D - 2d + 1$ proves the result. □

Chapter 3

CORESET SELECTION FOR THE SINKHORN DIVERGENCE AND GENERIC SMOOTH DIVERGENCES

Abstract. We introduce CO2, an efficient algorithm to produce convexly-weighted coresets with respect to generic smooth divergences. By employing a functional Taylor expansion, we show a local equivalence between sufficiently regular losses and their second order approximations, reducing the coreset selection problem to maximum mean discrepancy minimization. We apply CO2 to the Sinkhorn divergence, providing a novel sampling procedure that requires poly-logarithmically many data points to match the approximation guarantees of random sampling. Central to this result, we identify the Hessian of the Sinkhorn divergence and establish second-order Hadamard differentiability in its induced topology, which is equivalent to a scaled Gaussian RKHS topology. Our approach leads to a new perspective linking coreset selection and kernel quadrature to classical statistical methods such as moment and score matching. We showcase this method with a practical application of subsampling image data, and highlight key directions to explore for improved algorithmic efficiency and theoretical guarantees.

Keywords. coresets, kernel quadrature, Hadamard differentiability, entropic optimal transport, Sinkhorn divergence

3.1 Introduction

We consider the problem of compressing a dataset by selecting a convexly weighted subset of the observations, optimized for an error metric D. More concretely, given n independent draws from a compactly supported P on $\mathcal{X} \subseteq \mathbb{R}^d$ with empirical distribution P_n, we seek to construct P_m supported on a 'coreset' of m of these observations such that $D(P_n, P_m)$ is on the order of $D(P_n, P)$ as $m, n \to \infty$. We call such a weighted coreset (asymptotically) lossless. A typical example of D is an integral probability metric (Sriperumbudur et al.,

2012), $D(P_n, P_m) = \sup_{f \in \mathcal{F}} |\int f d(P_n - P_m)|$ — when $\mathcal{F}$ is Donsker (A. W. v. d. Vaart, 2000), we shall seek a P_m such that this quantity is $O_p(n^{-1/2})$. Another example of interest is the Sinkhorn divergence which we introduce in detail later (Ramdas et al., 2015). As we are concerned with empirical approximation of the population distribution, we frequently adopt the notation $D(Q) := D(Q, P)$.

Compression techniques are commonly employed in settings where it is costly to work with a large sample. In computational cardiology, for each configuration x of a simulated heart, thousands of machine hours must be used to evaluate a function $f(x)$ that approximates the mechanism for calcium signaling (Strocchi et al., 2020; Dwivedi and Mackey, 2023). This precludes evaluating common statistics of interest, such as $P_n f$, if P_n is defined using a large sample of heart configurations. However, if m configurations could be identified and weighted so that $P_m f \approx P_n f$, the computational task could be made tractable. If we had knowledge that f belonged to a function class $\mathcal{F}$, then this would be achieved by selecting P_m to make the corresponding integral probability metric small.

The construction of such P_m has been well studied in the case where $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space (Schölkopf and Smola, 2002). An IPM in this setting is referred to as maximum mean discrepancy (MMD) (Gretton et al., 2008). As these spaces are Donsker, the optimal error rate is $O(n^{-1/2})$.

Kernel herding was shown to achieve this error rate with $m = O(n^{1/2})$ in the case where the kernel is finite dimensional (Chen, Welling, et al., 2012). Kernel thinning subsequently met this bound (up to logarithmic factors) in more general cases, depending on the kernel's tail behavior (Dwivedi and Mackey, 2023). An alternative approach based on quadrature was developed in (Hayakawa et al., 2022) and refined in (Li et al., 2024a), leading to guarantees controlled by low-rank approximation of the empirical kernel matrix. Key to our analysis is a related kernel quadrature algorithm tailored to data-dependent kernels.

Often in unsupervised learning settings, data compression algorithms optimize for non-MMD error metrics. For example, K-means forms clusters to minimize the average squared deviation from the centroids (Steinley, 2006). This algorithm can be equivalently expressed in terms of the 2-Wasserstein distance, $W_2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\int d(x,y)^2 \pi(x,y)\right)^{1/2}$ (Pollard, 1982; Canas and Rosasco, 2012). In this framework, a distribution P_K supported on K

centroids is selected to minimize $W_2(P_n, P_K)$. Works such as Claici et al., 2020 fit more complex parametric distributions for the surrogate measure P_K, with Gaussian mixtures being a primary example. By considering compressed data, we make the more restrictive assumption that $P_K \ll P_n$. This forces the support of P_K to retain any structure enjoyed by that of P_n—for example, if P_n is supported on images of faces, then so too is P_K.

In practice, the Wasserstein distance is often too computationally expensive to be computed on large datasets, but it can be approximated efficiently using its entropically regularized counterpart (Cuturi, 2013). The Sinkhorn divergence similarly approximates the Wasserstein loss, but with slight modifications to satisfy the properties of a statistical divergence (Ramdas et al., 2015). Beyond their computational benefits, these quantities are also easier to estimate than the Wasserstein distance, especially in high dimensions (Vacher et al., 2021), and are more robust to noise (Rigollet and Weed, 2018; Bigot et al., 2019).

In this chapter, we explore whether divergences such as the Sinkhorn divergence can also yield benefits over the Wasserstein distance for coreset selection. This problem is important since, despite the Wasserstein distance's appealing relationship to familiar clustering algorithms, no compression algorithm can asymptotically improve on a random sample with respect to this loss (Weed and Bach, 2019). Our approach culminates in a procedure applicable to arbitrary divergences, where MMD compression algorithms can be applied to analytically derived Taylor expansions of the functional of interest, a two-step procedure we call CO2 (Coresets of Order 2).

Our main contributions are as follows:

1. We develop a framework for lossless data compression with respect to generic smooth divergences.

2. We identify the Hadamard operator, or Hessian, of the Sinkhorn divergence and establish second-order Hadamard differentiability in its induced topology. This requires first-order Hadamard differentiability of the entropic optimal transport potentials in scaled RKHS tangent spaces.

3. Using these results, we show that poly-logarithmically many data points m are sufficient

to ensure $D(P_m) = O_p(n^{-1})$ when D is the Sinkhorn divergence.

Precisely, we show $m = \omega(\log^d n)$ samples suffice for asymptotically lossless Sinkhorn reconstruction. To our knowledge, ours is the first work in the literature to even consider whether $m = o(n)$ samples suffices. The recent work of Yin et al., 2025, which also developed a method for constructing Sinkhorn coresets, instead focused on providing stability and convergence guarantees for their proposed optimization scheme.

We highlight in our experiments how Sinkhorn-CO2 applied to MNIST data not only improves performance with respect to the Sinkhorn divergence, but also better approximates concrete quantities of interest, such as the proportions of each label present in the resulting coreset. Other recent works have related coreset selection to improved efficiency in learning downstream tasks (Xiong et al., 2024; Chen, Wang, et al., n.d.; Gong et al., 2024). This places these methods under the umbrella of data distillation (Sachdeva and McAuley, 2023), the study of compactly representing large datasets with the goal of improving algorithmic efficiency. Our experiments demonstrate that Sinkhorn coresets provide a similar enhancement.

3.2 The General Case

3.2.1 Notation and definitions

Our results apply to sufficiently smooth divergences D, made precise by a second-order Hadamard differentiability condition (Fernholz, 1983). Denote by $\mathcal{P}(\mathcal{X})$ the space of probability distributions supported on $\mathcal{X}$. For an RKHS $\mathcal{F}$ on $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{F}}$, let $\mathcal{F}_1 := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1\}$ denote the unit ball and $\ell^\infty(\mathcal{F}_1)$ be the space of distributions on $\mathcal{X}$ with pseudometric $\|\mu\|_{\ell^\infty(\mathcal{F}_1)} := \sup_{\|f\|_{\mathcal{F}} \leq 1} |\int f d\mu|$. We say that a functional D with $\text{Domain}(D) \subseteq \ell^\infty(\mathcal{F}_1)$ is Hadamard differentiable at P (relative to $\ell^\infty(\mathcal{F}_1)$) if there exists a continuous linear functional $\dot{D} : \ell^\infty(\mathcal{F}_1) \to \mathbb{R}$ such that, for all $\{P + th_t : t \in [-1, 1], h_t \in \ell^\infty(\mathcal{F}_1)\} \subseteq \text{Domain}(D)$ with $\|h_t - h\|_{\ell^\infty(\mathcal{F}_1)} = o(1)$ for some h,

$$\lim_{t \to 0} \frac{D(P + th_t) - D(P)}{t} = \dot{D}(h).$$

The functional D is second order Hadamard differentiable at P if there exists a continuous

quadratic form $\ddot{D} : \ell^\infty(\mathcal{F}_1) \to \mathbb{R}$ such that

$$\lim_{t\to 0} \frac{D(P + th_t) - D(P) - \dot{D}(th_t)}{t^2} = \ddot{D}(h).$$

As the dual RKHS topology $\ell^\infty(\mathcal{F}_1)$ will be central to our analysis, we introduce the corresponding kernel function k and kernel embedding

$$Kh := \int k(\cdot, y)\, dh(y).$$

Depending on which is clearer in context, we denote the RKHS norm by either $\|\cdot\|_{\mathcal{F}}$ or $\|\cdot\|_K$. The dual norm has the quadratic representation

$$\|\mu\|^2_{\ell^\infty(\mathcal{F}_1)} = \iint k(x, y)\, d\mu(x) d\mu(y) =: Q_K(\mu).$$

In particular, $\|\mu - \nu\|_{\ell^\infty(\mathcal{F}_1)} = \mathrm{MMD}_K(\mu, \nu)$.

The continuous quadratic form $\ddot{D}$ can be identified with a bounded positive self-adjoint operator $\widetilde{G}$:

$$\ddot{D}(h) = \langle Kh, \widetilde{G}Kh\rangle_{\mathcal{F}}.$$

Equivalently, with

$$g(x, y) := \langle k(\cdot, x), \widetilde{G}k(\cdot, y)\rangle_{\mathcal{F}},$$

we may write $Gh = \int g(\cdot, y)\, dh(y)$ and $\ddot{D}(h) = \int Gh\, dh$. We refer to G as the Hadamard operator of D. See Appendix 3.A for further discussion of this quadratic expansion.

On the tangent space, the Hadamard operator induces the quadratic seminorm

$$\|h\|^2_G := \int Gh\, dh = \ddot{D}(h).$$

This object is fundamental to our analysis, as we will show that, under appropriate regularity conditions, data compression with respect to this quadratic form results in sharp compression of D.

Much of our analysis centers on P_n and on coresets P_m supported on the observed sample. For a possibly data-dependent positive-semidefinite kernel $\widehat{k}_n$, let

$$\widehat{\mathbf{K}}_n := \frac{1}{n}\left[\widehat{k}_n(X_i, X_j)\right]_{i,j=1}^n$$

be its normalized Gram matrix. If $\lambda_{1,n} \geq \cdots \geq \lambda_{n,n} \geq 0$ are its eigenvalues, define the rank-r residual trace

$$\mathrm{Tail}_r(\widehat{\mathbf{K}}_n) := \sum_{j>r} \lambda_{j,n}.$$

Equivalently, this is the trace error of the best positive-semidefinite rank-r approximation to $\widehat{\mathbf{K}}_n$.

When studying operators, $A : F \to G$, we will denote $\|A\|_{F\to G} = \sup_{\|f\|_F \leq 1} \|Af\|_G$, and when $F = G$ we shorten this to $\|A\|_F$. In technical arguments, we may often express a scalar constant on each bound by C that may represent different values in different arguments. For self-adjoint positive-semidefinite operators K_1 and K_2 on the same tangent space, we write $K_1 \succeq K_2$ if

$$\int (K_1 - K_2)\mu \, d\mu \geq 0$$

for every admissible signed tangent μ.

3.2.2 Main results

The main idea of our method is that compression with respect to a carefully chosen MMD leads to compression with respect to D. This leads to the CO2 Algorithm in 2. The KernelSelection step associates the divergence and the observed empirical distribution with a kernel for which MMD compression controls the loss of interest. We achieve this by establishing second-order Hadamard differentiability at the population distribution in an RKHS tangent topology. For computational efficiency, we reduce the rank of this objective by utilizing a Nyström approximation (Tropp et al., 2017; Chen, Epperly, et al., 2025) up to an appropriately negligible Schatten-1 error $o_p(n^{-1})$. We then apply an MMD compression method (Hayakawa et al., 2023; Li et al., 2024a) on the resulting quadratic form. Our analysis centers on the recombination algorithm (Litterer and Lyons, 2012b), but this can be easily modified to the user's preferred choice.

We make this concrete in the following results, which establish a criteria for the selection of D coresets, and a tractable algorithm for their computation.

Algorithm 2 CO2-recombination

Input: D, P_n, target size m, oversampling parameter $\theta \geq 2$

1: $K \leftarrow \text{KernelSelection}(D, P_n)$

2: $(\cdot, U, \Lambda) \leftarrow \text{Nyström}\big(K, \theta(m-1)\big)$

3: $P_m \leftarrow \text{Recombination}\big(U[:, 1{:}m-1], [k(X_i, X_i)]_{i=1}^n, \Lambda[1{:}m-1]\big)$ } KernelCompress

4: **return** P_m

Lemma 3.1 (Quadratic loss control from MMD compression)**.** *Let D be second order Hadamard differentiable at P relative to $\ell^\infty(\mathcal{F}_1)$, where $\mathcal{F}_1$ is the unit ball in an RKHS with characteristic kernel k. Then,* $\text{MMD}_K(P_n, P_m) = o_p(n^{-1/2})$ *implies* $D(P_m) = O_p(n^{-1})$ *and*

$$D(P_m) = \ddot{D}(P_m - P) + o_p(n^{-1}) = D(P_n) + o_p(n^{-1}).$$

Our proof of the above uses that $D(P) = 0$ and $\dot{D} \equiv 0$ for a divergence about the null. Combining this with the fact that $P_m = P + tH_t$ when $H_t := [P_m - P]/t$ yields

$$D(P_m) = t^2 \left[\frac{D(P + tH_t) - D(P) - t\dot{D}(H_t)}{t^2} \right].$$

Taking $t = n^{-1/2}$, using that $\text{MMD}_K(P_n, P_m) := \|P_n - P_m\|_{\ell^\infty(\mathcal{F}_1)}$, and applying the second order functional delta method yields Lemma 3.1.

Lemma 3.2 (MMD compression from a Gram-matrix tail bound)**.** *Let $\widehat{k}_n$ be any positive-semidefinite kernel selected after observing $X_1, \ldots, X_n$, and let $\widehat{\mathbf{K}}_n$ be its normalized Gram matrix. Fix an integer $r \geq 1$ and an oversampling parameter $\theta \geq 2$ such that $r(\theta - 1) > 1$. Conditionally on the observed sample and the selected kernel, a probability measure P_m supported on no more than $m \leq r + 1$ observations can be randomly constructed such that*

$$\mathbb{E}\left[\text{MMD}^2_{\widehat{k}_n}(P_n, P_m) \,\middle|\, X_1, \ldots, X_n, \widehat{k}_n \right] \leq 4 \left(1 + \frac{r}{r(\theta - 1) - 1} \right) \text{Tail}_r(\widehat{\mathbf{K}}_n).$$

The construction has time complexity $O(n^2 r + \theta^3 r^3 + r^3 \log(n/r))$.

The oversampling parameter θ increases the quality of the approximation at the expense of additional computational overhead.

A similar approach appears in a recent paper (Li et al., 2024a, Algorithm 4). The primary difference in our algorithms is the choice of Nyström approximation, where they used the randomly pivoted Cholesky (Chen, Epperly, et al., 2025) while we used one based on random projections (Tropp et al., 2017). Comparing these two approximations, the method of Tropp et al., 2017 yields a sharper low-rank approximation at the cost of increased memory and time complexity, respectively at $O(n^2)$ and $O(n^2m)$. This discrepancy leads to improved algorithmic efficiency in Li et al., 2024a, with $O(nm)$ memory requirements and a Nyström implementation with $O(nm^2)$ time complexity, in addition to the cost of evaluating mn Gram-matrix entries. This makes their procedure particularly effective for standard MMD coreset problems, where kernel entries can be evaluated on demand. In our applications, the selected kernel may itself depend on the empirical distribution. For the Sinkhorn divergence, computing the empirical scaling q_n requires solving the empirical self-transport problem; once q_n is available, however, the scaled-kernel entries are explicit and no inversion of the empirical Hessian is needed. In the dense implementation analyzed here, construction of the empirical transport kernel already incurs quadratic cost, so the random-projection Nyström approximation adds comparatively little overhead while providing the low-rank guarantee used in our analysis. For additional state-of-the-art MMD compression algorithms, see Dwivedi and Mackey, 2025 and Carrell et al., 2025.

These results combine to yield our main compression guarantee. In situations where the Hadamard operator is data-dependent, we include additional details regarding the use of a suitable estimator.

Theorem 3.1 (Lossless compression for smooth divergences)**.** *Fix an oversampling parameter $\theta \geq 2$. Let k be a fixed characteristic population kernel with RKHS $\mathcal{F}$, and suppose that D is second-order Hadamard differentiable at P relative to $\ell^\infty(\mathcal{F}_1)$. Let $\widehat{k}_n$ be a possibly data-dependent positive-semidefinite kernel selected from the observed sample, with normalized Gram matrix $\widehat{\mathbf{K}}_n$. Let $m = m_n$ and put $r_n := m_n - 1$. Suppose $r_n(\theta - 1) > 1$ eventually and*

$$\mathrm{Tail}_{r_n}(\widehat{\mathbf{K}}_n) = o_p(n^{-1}),$$

and suppose that, for every possibly data-dependent coreset $Q_n \ll P_n$,

$$\mathrm{MMD}_{\widehat{k}_n}(Q_n, P_n) = o_p(n^{-1/2}) \quad \Longrightarrow \quad \mathrm{MMD}_k(Q_n, P_n) = o_p(n^{-1/2}).$$

Then Algorithm 2 constructs a coreset P_m such that

$$D(P_m) = D(P_n) + o_p(n^{-1}).$$

Its time complexity is $O(n^2 m + m^3 \log(n/m))$.

Proof. Apply Lemma 3.2 conditionally on the observed sample and the selected kernel, with $r = r_n$. Writing

$$C_{r,\theta} := 4\left(1 + \frac{r}{r(\theta-1)-1}\right),$$

the lemma gives

$$\mathbb{E}\left[\mathrm{MMD}^2_{\widehat{k}_n}(P_n, P_m) \,\middle|\, X_1, \ldots, X_n, \widehat{k}_n\right] \le C_{r_n,\theta}\,\mathrm{Tail}_{r_n}(\widehat{\mathbf{K}}_n).$$

Since θ is fixed and $r_n(\theta-1) > 1$ eventually, $C_{r_n,\theta}$ is uniformly bounded.

Fix $\eta > 0$. Conditional Markov inequality therefore yields

$$\begin{aligned} &P\left(\sqrt{n}\ \mathrm{MMD}_{\widehat{k}_n}(P_n, P_m) > \eta \,\middle|\, X_1, \ldots, X_n, \widehat{k}_n\right) \\ &\quad = P\left(\mathrm{MMD}^2_{\widehat{k}_n}(P_n, P_m) > \frac{\eta^2}{n} \,\middle|\, X_1, \ldots, X_n, \widehat{k}_n\right) \\ &\quad \le \min\left\{1,\ \frac{C_{r_n,\theta} n\,\mathrm{Tail}_{r_n}(\widehat{\mathbf{K}}_n)}{\eta^2}\right\}. \end{aligned}$$

By assumption,

$$n\,\mathrm{Tail}_{r_n}(\widehat{\mathbf{K}}_n) = o_p(1),$$

so the random upper bound on the right converges to zero in probability. It is also bounded by one. Hence its expectation converges to zero. Taking expectations of the conditional probability and using the tower property gives

$$P\left(\sqrt{n}\ \mathrm{MMD}_{\widehat{k}_n}(P_n, P_m) > \eta\right) \longrightarrow 0.$$

Since this holds for every $\eta > 0$,

$$\mathrm{MMD}_{\widehat{k}_n}(P_m, P_n) = o_p(n^{-1/2}).$$

Thus, by assumption

$$\mathrm{MMD}_k(P_m, P_n) = o_p(n^{-1/2}).$$

The result follows by Lemmas 3.1 and 3.2. □

Algorithm 3 m selection: recombination sweep

$U \leftarrow m_{\max} - 1$ leading Nyström eigenvectors

$K \leftarrow$ kernel matrix

$V \leftarrow$ orthogonal complement of $U \oplus \mathbf{1}$

$\tau \leftarrow$ specified tolerance

err $\leftarrow 0$, % initialized to 0

count $\leftarrow 0$, % initialized to 0

$\mu \leftarrow$ initial empirical distribution

while err $\leq \tau$ **and** count $\leq n - m_{\max} + 1$ **do**

$\mu_{\text{prev}} \leftarrow \mu$

count$\leftarrow$ count + 1

$\alpha \leftarrow \min d\mu/V[:, \text{count}]$ % choose α so that next iterate has one less non-zero entry

$\mu \leftarrow \mu - \alpha V[:, \text{count}]$

err $\leftarrow (\mu - P_n)^\top K(\mu - P_n)$

end while

return μ_{prev}

3.2.3 Selecting m

Our theoretical results match earlier work in MMD compression and coreset selection, which aims to make the coreset error negligible relative to the empirical sampling error. In practice, rather than selecting m solely from an asymptotic rate, we favor direct evaluation of $D(P_n, P_m)$ or, more efficiently, the compression error under the kernel returned by KernelSelection. Thus we recommend choosing the smallest m for which

$$\text{MMD}^2_{\hat{k}_n}(P_n, P_m) \leq \tau,$$

or, when computationally feasible, $D(P_n, P_m) \leq \tau$.

There are several ways to determine this minimal m. Binary search is a simple strategy. Algorithm 3 is more efficient when there is an upper bound $m_{\max}$ on the number of samples the user is willing to retain. It performs one recombination sweep and returns the smallest

available coreset for which

$$\mathrm{MMD}^2_{\widehat{k}_n}(P_n, P_{m'}) \le \tau,$$

taking $m = m_{\max}$ if no smaller coreset meets the tolerance.

Theorem 3.1 suggests taking $\tau = o(n^{-1})$. For a fixed population kernel, the proof of Lemma 3.1 gives the more informative comparison

$$|D(P_m) - D(P_n)| \le (R^2_{m,n} + 2R_{m,n})\,\mathrm{MMD}^2_k(P_n, P) + o_p(n^{-1}),$$

where

$$R_{m,n} := \frac{\mathrm{MMD}_k(P_m, P_n)}{\mathrm{MMD}_k(P_n, P)}.$$

Thus $R_{m,n} = o_p(1)$ is sufficient. In practice, one may estimate the scale of the empirical MMD from the eigenvalues of the selected normalized Gram matrix and choose τ as a small multiple of n^{-1}; a formal finite-sample calibration is beyond the asymptotic guarantees developed here.

3.2.4 Hadamard operator and kernel selection

The Hadamard operator G is the canonical population target of KernelSelection because it represents the local quadratic form and hence the tangent geometry relevant for compression. A primary technical challenge is not just to compute G, but to verify that D is second-order Hadamard differentiable in the topology it induces. The following lemma motivates this selection.

Lemma 3.3 (Domination of the Hadamard form by the RKHS kernel)**.** *If k is a characteristic kernel and D is second-order Hadamard differentiable at P relative to $\ell^\infty(\mathcal{F}_1)$, then there exists a constant $C > 0$ such that $CK \succeq G$.*

Concretely, $CK \succeq G$ implies

$$\|h\|_G \le C^{1/2}\|h\|_K.$$

Thus every RKHS topology where the second order expansion is valid is at least as strong as the Hadamard topology, making Lemma 3.1 a strictly more stringent condition for a coreset to satisfy. An equivalent kernel may still be preferable computationally when it is easier to estimate or approximate at low rank.

3.3 Illustration: Sinkhorn Coresets

3.3.1 Overview

As an illustration of our general framework, we consider compression with respect to the Sinkhorn divergence (Ramdas et al., 2015)

$$S(\mu,\nu) := \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\mu,\mu) + \mathrm{OT}_\varepsilon(\nu,\nu)),$$

with the regularization parameter $\varepsilon > 0$ suppressed in the notation. Here, OT_ε denotes the entropically regularized Wasserstein-2 distance

$$\mathrm{OT}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 d\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi \parallel \mu\otimes\nu), \tag{3.1}$$

for $\Pi(\mu,\nu)$ the set of joint distributions with marginals μ,ν.

In this section, we verify that lossless Sinkhorn compression requires m to grow only slightly faster than $\log^d n$.

Theorem 3.2 (Polylogarithmic Sinkhorn compression)**.** *For $m = \omega(\log^d n)$, P_m can be constructed in $O(n^2 m + m^3 \log(n/m))$ time by applying the compression routine to the empirical scaled Gaussian kernel, such that*

$$S(P, P_m) = S(P, P_n) + o_p(n^{-1}).$$

The main theoretical step is to verify second-order Hadamard differentiability in the topology induced by the Sinkhorn Hessian. Section 3.3.2 establishes first-order differentiability of the entropic transport potentials in a scaled Gaussian RKHS. Section 3.3.3 then identifies the Sinkhorn Hadamard operator and proves that it induces this topology. Finally, we verify that the empirical self-transport plan supplies a computable equivalent kernel, yielding the desired sample complexity.

3.3.2 Entropic optimal transport potentials

The OT_ε functional also admits a representation in terms of the entropic optimal transport potentials $\phi_{\mu,\nu}, \psi_{\mu,\nu} : \mathcal{X} \to \mathbb{R}$:

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \int \phi_{\mu,\nu}\, d\mu + \int \psi_{\mu,\nu}\, d\nu. \tag{3.2}$$

Thus, to study the differentiability of S, we first provide corresponding results for $\phi_{\mu,\nu}, \psi_{\mu,\nu}$. Equivalently, the optimal entropic transport plan $\pi^{\varepsilon}_{\mu,\nu}$, which attains the minimum in (3.1), has density

$$\xi^{\varepsilon}_{\mu,\nu}(x,y) := \frac{d\pi^{\varepsilon}_{\mu,\nu}}{d(\mu \otimes \nu)}(x,y) = \exp\left(\frac{\phi_{\mu,\nu}(x) + \psi_{\mu,\nu}(y) - \|x-y\|^2}{\varepsilon}\right). \tag{3.3}$$

This decomposition originates from Csiszár, 1975 and Rüschendorf and Thomsen, 1993, and the functions $\phi_{\mu,\nu}$ and $\psi_{\mu,\nu}$ are called the entropic optimal transport potentials. Recently, asymptotic distributions of plug-in estimates of the potentials and related quantities have been verified via the implicit function theorem (Goldfeld et al., 2022; Gonzalez-Sanz et al., 2022). Specifically, evaluating the marginal constraints of $\pi^{\varepsilon}_{\mu,\nu}$ yields the Schrödinger system:

$$\begin{aligned}
\phi_{\mu,\nu}(x) &= -\varepsilon \log \int \exp\left(\frac{\psi_{\mu,\nu}(y) - \|x-y\|^2}{\varepsilon}\right) d\nu(y), \\
\psi_{\mu,\nu}(x) &= -\varepsilon \log \int \exp\left(\frac{\phi_{\mu,\nu}(y) - \|x-y\|^2}{\varepsilon}\right) d\mu(y).
\end{aligned}$$

Goldfeld et al. (2022) study this problem through Hadamard differentiability, while Gonzalez-Sanz et al. (2022) provide a direct analytic expression for the limiting distribution. These analyses establish regularity of the potentials in Sobolev-type topologies. We will see that the second-order Sinkhorn expansion is governed by a finer P dependent tangent topology, thus necessitating a tightened analysis. We reparameterize the equations in terms of the scalings

$$(u_{\mu,\nu}, v_{\mu,\nu}) := \big(\exp(-\phi_{\mu,\nu}/\varepsilon), \exp(-\psi_{\mu,\nu}/\varepsilon)\big),$$

which satisfy

$$u_{\mu,\nu} = K\left(\frac{d\nu}{v_{\mu,\nu}}\right) := J_{v_{\mu,\nu}}(\nu), \qquad v_{\mu,\nu} = K\left(\frac{d\mu}{u_{\mu,\nu}}\right) := J_{u_{\mu,\nu}}(\mu), \tag{3.4}$$

where $K\gamma := \int k(\cdot, y) d\gamma(y)$ and $k(x,y) = \exp(-\|x-y\|^2/\varepsilon)$. Let $u = u_{\mu,\nu}$ and $v = v_{\mu,\nu}$. From this perspective, the operators J_u and J_v define the system, and thus induce a natural topology for the verification of differentiability.

For a fixed strictly positive $a \in \mathcal{H}$, define

$$\widetilde{k}_a(x,y) := \frac{k(x,y)}{a(x)a(y)}, \qquad J_a\gamma := K\left(\frac{d\gamma}{a}\right), \qquad \|\gamma\|_{\mathcal{E}_a} := \|J_a\gamma\|_{\mathcal{H}}.$$

Let $\mathcal{E}_a$ be the completion of finite signed measures in this norm and let $\mathcal{E}_a^0 := \{\gamma \in \mathcal{E}_a : \gamma(\mathbf{1}) = 0\}$. To enforce uniqueness of the solution to this system of equations we follow the convention of Goldfeld et al. (2022), fixing $x_0 \in \mathcal{X}$ and selecting the anchored representatives satisfying $u_{\mu,\nu}(x_0) = v_{\mu,\nu}(x_0)$. Define

$$\mathsf{T}_\nu h := K\left(h\frac{d\nu}{v^2}\right), \qquad \mathsf{T}_\mu h := K\left(h\frac{d\mu}{u^2}\right).$$

Theorem 3.3 (Hadamard derivative of the EOT potentials)**.** *The map*

$$\eta : (\mu,\nu) \longmapsto (\phi_{\mu,\nu}, \psi_{\mu,\nu}) = \left(-\varepsilon \log u_{\mu,\nu}, -\varepsilon \log v_{\mu,\nu}\right)$$

is Hadamard differentiable at (μ,ν), tangentially to $\mathcal{E}_u^0 \times \mathcal{E}_v^0$, as a map into $C(\mathcal{X}) \times C(\mathcal{X})$. For $(\gamma^1,\gamma^2) \in \mathcal{E}_u^0 \times \mathcal{E}_v^0$, let $(\dot{u},\dot{v})$ be the unique solution of

$$\dot{u} + \mathsf{T}_\nu \dot{v} = J_v \gamma^2, \tag{3.5}$$

$$\mathsf{T}_\mu \dot{u} + \dot{v} = J_u \gamma^1, \tag{3.6}$$

$$\dot{u}(x_0) - \dot{v}(x_0) = 0. \tag{3.7}$$

Then

$$\dot{\eta}_{\mu,\nu}[\gamma^1,\gamma^2] = \left(-\varepsilon\frac{\dot{u}}{u}, -\varepsilon\frac{\dot{v}}{v}\right).$$

At a diagonal base point $\mu = \nu = P$, the scalings also agree, setting $u = v = q$. This leads to a common topology $\mathcal{E}_q^0$ on the product, and we will show it agrees with the topology induced by the Sinkhorn Hessian.

3.3.3 Sinkhorn Hadamard operator and topology

At the diagonal, write

$$q := u_{P,P} = v_{P,P}, \qquad \mathsf{T}_P h := K\left(h\frac{dP}{q^2}\right), \qquad \mathsf{R}_P := (I - \mathsf{T}_P^2)^{-1}\big|_{q^\perp}.$$

The operator T_P is compact, positive, and self-adjoint; $\mathsf{T}_P q = q$, the eigenvalue 1 is simple, and therefore $\rho := \|\mathsf{T}_P|_{q^\perp}\| < 1$. Let $\Pi_{q^\perp} : \mathcal{H} \to q^\perp$ denote the orthogonal projection.

Theorem 3.4 (Sinkhorn Hessian and equivalent RKHS topology)**.** *Let $D(\nu) := S(P, \nu)$. Then D is second-order Hadamard differentiable at P, tangentially to $\mathcal{E}_q^0$, and*

$$\ddot{D}_P(\gamma) = \frac{\varepsilon}{2} \langle J_q\gamma, \mathsf{R}_P J_q\gamma\rangle_{\mathcal{H}}\,. \tag{3.8}$$

Equivalently, if

$$\Phi_q^0(x) := \Pi_{q^\perp}\left[\frac{k(\cdot, x)}{q(x)}\right],$$

the corresponding positive-definite Hadamard kernel is

$$g_{S,P}(x, y) := \frac{\varepsilon}{2}\left\langle \mathsf{R}_P^{1/2}\Phi_q^0(x), \mathsf{R}_P^{1/2}\Phi_q^0(y)\right\rangle_{\mathcal{H}}.$$

Moreover, for every $\gamma \in \mathcal{E}_q^0$,

$$\frac{\varepsilon}{2}\|\gamma\|_{\mathcal{E}_q}^2 \le \ddot{D}_P(\gamma) \le \frac{\varepsilon}{2(1-\rho^2)}\|\gamma\|_{\mathcal{E}_q}^2.$$

Consequently, on zero-mass tangents, the topology induced by the Sinkhorn Hadamard kernel $g_{S,P}$ coincides with the base-point-scaled Gaussian topology $\mathcal{E}_q^0$ generated by $\widetilde{k}_q$.

For computation, we use the the empirical analogue of the Gaussian scaled kernel. Let $q_n := u_{P_n,P_n}$ and define

$$\widetilde{k}_{q_n}(x, y) := \frac{k(x, y)}{q_n(x)q_n(y)}, \qquad J_{q_n}\gamma := K\left(\frac{d\gamma}{q_n}\right).$$

Lemma 3.4 (Empirical-to-population Sinkhorn embedding)**.** *Let $P_m = P_{m,n}$ be any possibly data-dependent probability measure. If*

$$\|J_{q_n}(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}),$$

then

$$\|J_q(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}).$$

Theorem 3.5 (Compression with the empirical Sinkhorn kernel)**.** *If $m = \omega(\log^d n)$ and the MMD-compression routine in Algorithm 2 is applied conditionally to the empirical scaled kernel $\widetilde{k}_{q_n}$, then*

$$\mathrm{MMD}_{\widetilde{k}_{q_n}}(P_m, P_n) = o_p(n^{-1/2}) \quad \textit{and} \quad \mathrm{MMD}_{\widetilde{k}_q}(P_m, P_n) = o_p(n^{-1/2}).$$

Consequently, the population quadratic form in (3.8) *is preserved up to $o_p(n^{-1})$.*

The preceding results lead to the following implementation of Sinkhorn CO2. Let

$$\Pi_n^\varepsilon = [\pi_{ij}]$$

denote the entropic self-transport plan $\pi^\varepsilon_{P_n,P_n}$, and define

$$A_n := n\Pi_n^\varepsilon = \frac{1}{n}\widetilde{K}_{q_n}, \qquad \widetilde{K}_{q_n} := \left[\frac{k_\varepsilon(X_i,X_j)}{q_n(X_i)q_n(X_j)}\right]_{i,j=1}^n.$$

For a coreset $P_m = \sum_{i=1}^n w_{m,i}\delta_{X_i}$, write $w_m \in \mathbb{R}^n$ for its weight vector and set $d_m := w_m - \frac{1}{n}\mathbf{1} \in \mathbf{1}^\perp$. The empirical Hessian quadratic form at P_n is

$$\ddot{S}_{P_n}(d_m) = \frac{\varepsilon n}{2} d_m^\top A_n (I - A_n^2)^{-1} d_m. \tag{3.9}$$

Here the inverse is taken on the zero-sum subspace $\mathbf{1}^\perp$, where $I - A_n^2$ is invertible. Equivalently, this is

$$\frac{\varepsilon}{2}\left\langle J_{q_n}(P_m - P_n), \mathsf{R}_{P_n} J_{q_n}(P_m - P_n)\right\rangle_{\mathcal{H}}, \qquad \mathsf{R}_{P_n} = (I - \mathsf{T}_{P_n}^2)^{-1}\big|_{q_n^\perp}.$$

Computing the resolvent $(I - A_n^2)^{-1}$ is expensive, hence we propose using the substantially simpler quadratic form

$$\begin{aligned}\widetilde{S}_n(P_n,P_m) &:= \frac{\varepsilon n}{2} d_m^\top A_n d_m \\ &= \frac{\varepsilon}{2} d_m^\top \widetilde{K}_{q_n} d_m = \frac{\varepsilon}{2}\,\mathrm{MMD}^2_{k_{q_n}}(P_m,P_n).\end{aligned} \tag{3.10}$$

The recombination algorithm applied to the eigenvectors of this kernel leads to the following guarantee.

Lemma 3.5 (Sinkhorn fast computation)**.** *Let T_{P_n} be the empirical analogue of T_P and define*

$$\mathsf{R}_{P_n} := (I - \mathsf{T}_{P_n}^2)^{-1}\big|_{q_n^\perp}.$$

For a zero-mass discrepancy Δ, set $z = J_{q_n}\Delta \in q_n^\perp$. If z is orthogonal to the first r eigenvectors retained by the recombination step and $\lambda_{r+1,n} < 1$ is the next eigenvalue of $\mathsf{T}_{P_n}|_{q_n^\perp}$, then

$$\left|\frac{\varepsilon}{2}\langle z, \mathsf{R}_{P_n} z\rangle_{\mathcal{H}} - \frac{\varepsilon}{2}\|z\|^2_{\mathcal{H}}\right| \le \frac{\varepsilon}{2}\frac{\lambda^2_{r+1,n}}{1-\lambda^2_{r+1,n}}\|z\|^2_{\mathcal{H}}.$$

3.4 Numerical experiments

For efficient computation of the Sinkhorn divergence and entropic optimal transport potentials we utilize the geomloss Python package (Feydy et al., 2019a). CO2 is performed with respect to the empirical scaled kernel $\widetilde{k}_{q_n}$. In our Nyström approximation, we select a $\theta = 3$ oversampling parameter except in our MNIST experiment in Section 3.4.5, where we set $\theta = 5$. Code for all simulations can be found at https://github.com/amkokot/sinkhorn_coresets.

3.4.1 Visualization

We first qualitatively demonstrate that our compression algorithm rapidly recovers key characteristics of input datasets. We consider a Gaussian mixture generative model $\frac{1}{K}\sum_{i=1}^{K} N(\mu_i, I)$ in two dimensions, which we approximate with P_n, $n = 10^4$. We consider a mixture of $K = 8$ Gaussians with means placed at $[\pm 3, \pm 3], [0, \pm 6], [\pm 6, 0]$, which we compress to $m = 16$ points based on the Sinkhorn divergence with $\varepsilon = 0.75$. As seen in Figure 3.1, these coresets typically recover the primary clusters of the dataset, with samples being reasonably spread across each of the 8 clusters.

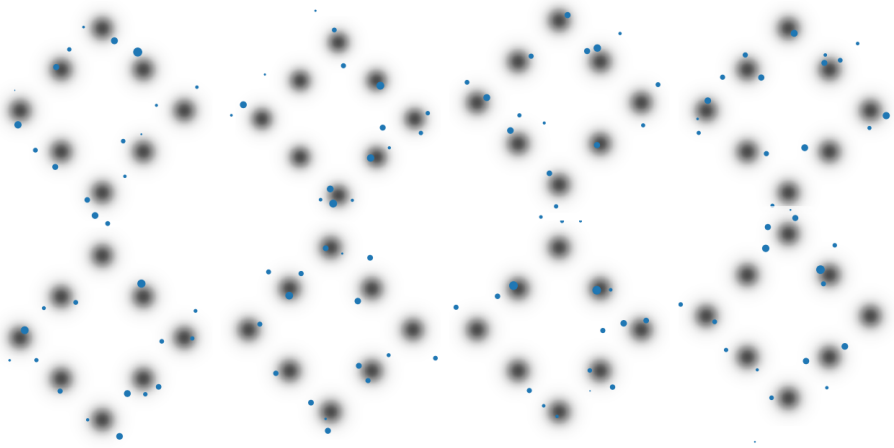


Figure 3.1: Eight repetitions of the simulation described in Section 3.4.1.

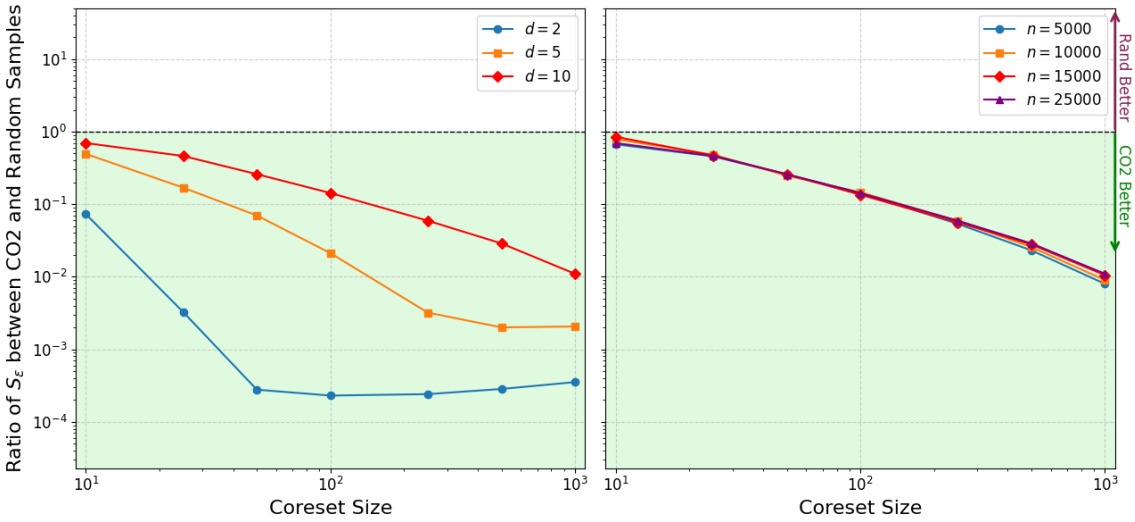


Figure 3.2: The reconstruction error $S_\varepsilon(P_n, P_m)$ in various dimensions (left) and dataset sizes n (right). In the first plot the sample size is fixed at $n = 2.5 \times 10^4$, for the latter the dimension is fixed at $d = 10$.

3.4.2 Distribution recovery

We now demonstrate how our method quickly reduces the reconstruction error $S_\varepsilon(P_n, P_m)$. Following Section 7.3 of (Dwivedi and Mackey, 2023), we take $N(0, I)$ as our generative distribution in dimensions $d = 2, 5, 10$, and we set the regularization parameter to be $\varepsilon = 2d$ in each setting. We then compare $S_\varepsilon(P_n, P_m)$ for P_m generated by Sinkhorn compression to P_m generated by random sampling. We perform 40 trials for each configuration of m, d, n. In Figure 3.2 we see a strong dependence on the dimension, and minimal dependence on n.

3.4.3 Quadratic approximation

A key element of our theory is that compression with respect to the second order expansion of a divergence is sufficient for compression with respect to the actual divergence of interest. We compute the relative error

$$\frac{|S(P_n, P_m) - \widetilde{S}_n(P_n, P_m)|}{S(P_n, P_m)}.$$

For P_m generated via our recombination method, and samples P_n generated with respect to the generative process in Section 3.4.2, with $d = 10$, $n = 2.5 \times 10^4$, we perform 80 trials for each m. Figure 3.3 reports this relative error. Lemma 3.5 quantifies the difference between this inexpensive form and the exact empirical Hessian on the recombination residual.

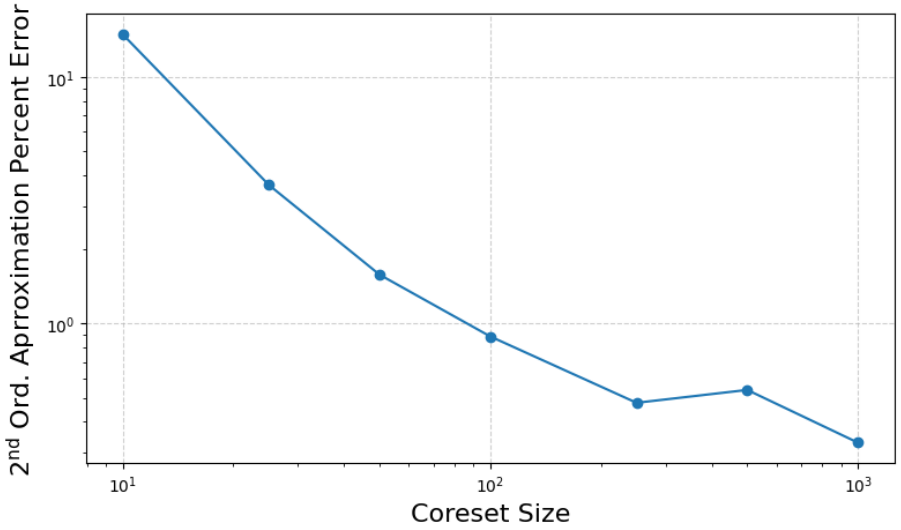


Figure 3.3: Percent relative error of the empirical scaled-kernel quadratic approximation $\widetilde{S}_n(P_n, P_m)$ compared to $S(P_n, P_m)$, with P_m generated via recombination compression.

3.4.4 MMD Compression Algorithms

We demonstrate how a variety of MMD compression algorithms can be paired with CO2 to provide efficient coresets with respect to the Sinkhorn divergence. Our second order approximation results allow one to reduce the coreset selection problem for smooth divergences such as the Sinkhorn to a corresponding MMD, allowing practitioners to select an algorithm of their preference at this step of our method. This is particularly useful as this is a growing area of research, and we expect the development of algorithms that greatly improve our basic recombination approach, particularly in terms of practical finite sample performance.

We draw $n = 2.5 \times 10^4$ samples uniformly from the unit cube of dimension 10, and compute the Sinkhorn divergence with regularization $\varepsilon = 20$. We compare 3 different coreset

selection procedures. Two of these algorithms are aimed at directly optimizing the MMD compression error, kernel herding and recombination, and we optimize the weights of the selected samples post selection. Halton sampling is a more generic discrepancy minimization procedure used in quasi-Monte Carlo (Jank, 2005). We compare these procedures to the baseline random sampling, and perform 80 trials for each m and sampling algorithm.

As seen in Figure 3.4, procedures that directly targeted MMD minimization of the second order term displayed the best performance. The recombination algorithm, our recommended MMD compression procedure, performed near optimally in each setting, with particularly impressive results for smaller targeted compression sizes.

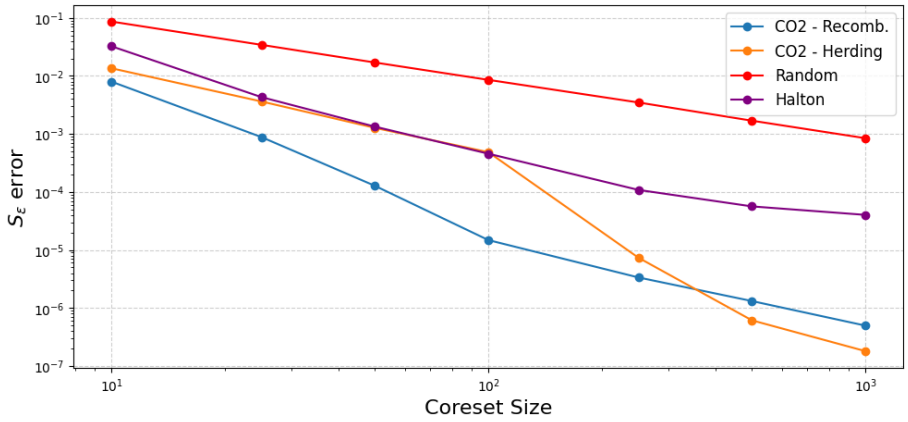


Figure 3.4: MMD compression algorithms as well as other sampling methods employed at the secondary stage of our Sinkhorn compression method and the resulting compression error $S(P_n, P_m)$ induced by these algorithms.

3.4.5 Image processing

We employ our algorithm on the entire 70,000 image MNIST dataset, where after standardizing the raw features to have mean 0 and unit variance, we select a coreset of size $m = 1000$ for the Sinkhorn divergence with regularization $\varepsilon = 1.5 \times 10^4$. This regularization is large, but not atypical in this high-dimensional setting. As a point of comparison, a typical heuristic for kernel bandwidth selection is to use the median squared distance between observations

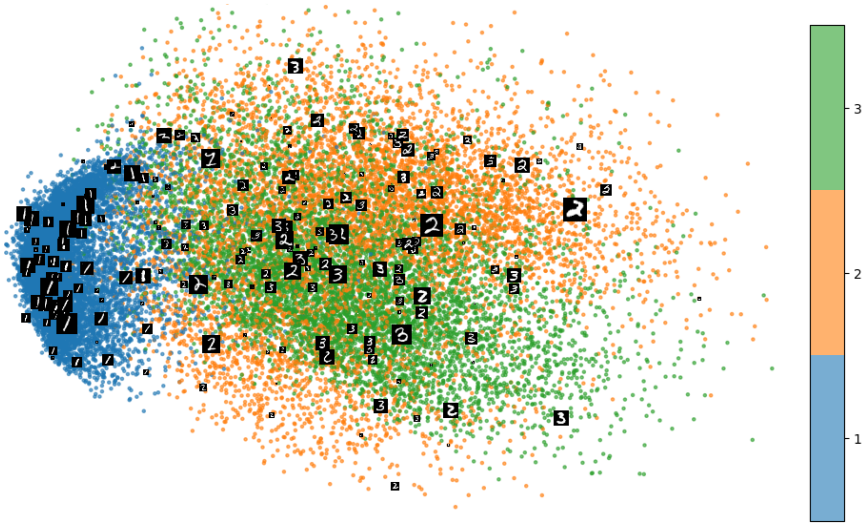


Figure 3.5: PCA coordinates of MNIST data with a Sinkhorn CO2 coreset overlaid on the sample, with figure sizes proportional to the sample weights. For interpretability, only digits 1, 2, and 3 are displayed here.

(Garreau et al., 2018), and for our data this is ≈ 1000, and thus of a comparable magnitude. We perform 500 trials of coreset generation. High dimensional image data is representative of a particularly relevant setting for data compression. By enforcing that our coreset belongs to the observed dataset, there is no ambiguity regarding which of the labels these datapoints correspond to, as compared to a procedure that selects centroids from the ambient space. Our selection thus respects the underlying structure of the data, which we expect to practically be of a much lower intrinsic dimension.

As a visualization of the resulting weighted coreset, in Figure 3.5 we plot the first 2 PCA coordinates of the MNIST dataset corresponding to a subset of the labels considered, $\{1, 2, 3\}$, in a scatter plot, and overlay the images corresponding to the selected datapoints with size proportionate to their weights. This figure gives a sense of the geometry of the sample, as we

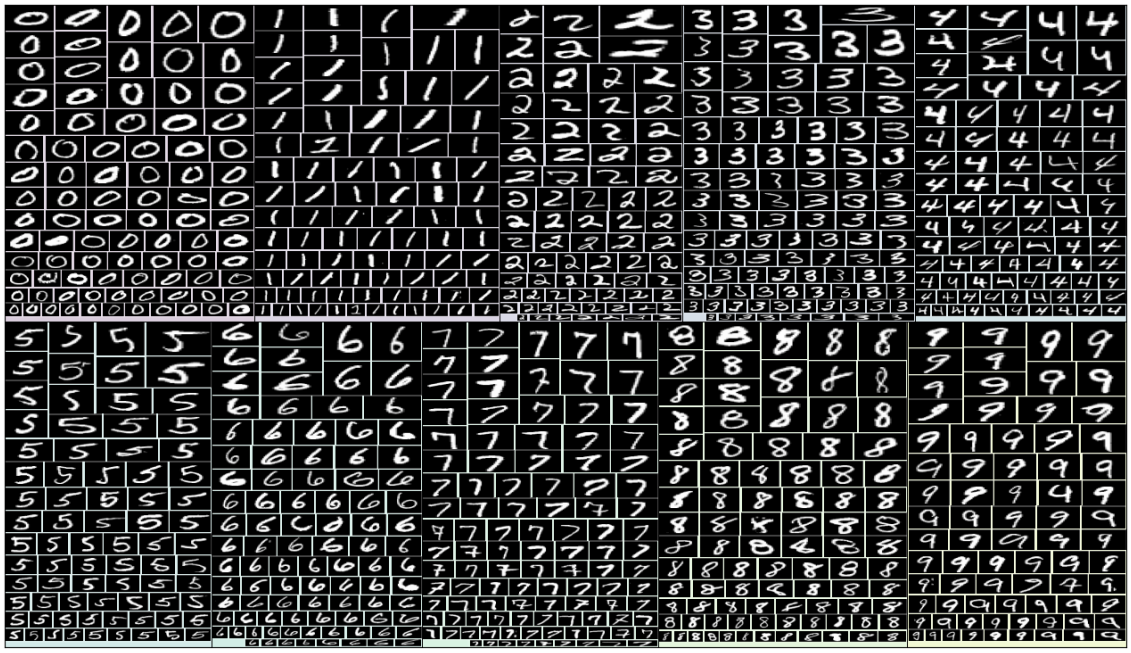


Figure 3.6: 1000 MNIST digits selected via Sinkhorn CO2, sorted by digit and with areas proportional to their weights.
Note: the treemap software used to produce this visualization could not maintain the 1:1 aspect ratio of the original MNIST images.

see a tight cluster of 1s, compared to the 2 and 3 digits which have more significant overlap. In Figure 3.6 we display a treemap of the 1000 selected digits realized by one trial of the CO2 algorithm. This visualization reveals that each digit receives about 1/10 of the total weight in the convexly weighted coreset.

The results of this experiment are shown in Figure 3.7, which displays Q-Q (quantile-quantile) plots comparing the percentiles of compression error as compared to random sampling, as well as the ℓ_1 error of sampled label proportions relative to the population values, $\sum_{i=0}^{9} |P(\text{label} = i) - \sum_{j=1}^{n} w_j \mathbf{1}\{Y_j = i\}|$. The Q-Q plots capture the variation of both the random samples as well as the random projection employed in the Nyström approximation in Algorithm 2. Due to the significant regularization level, it is unsurprising that we see our compression scheme dramatically reduces the Sinkhorn error as compared

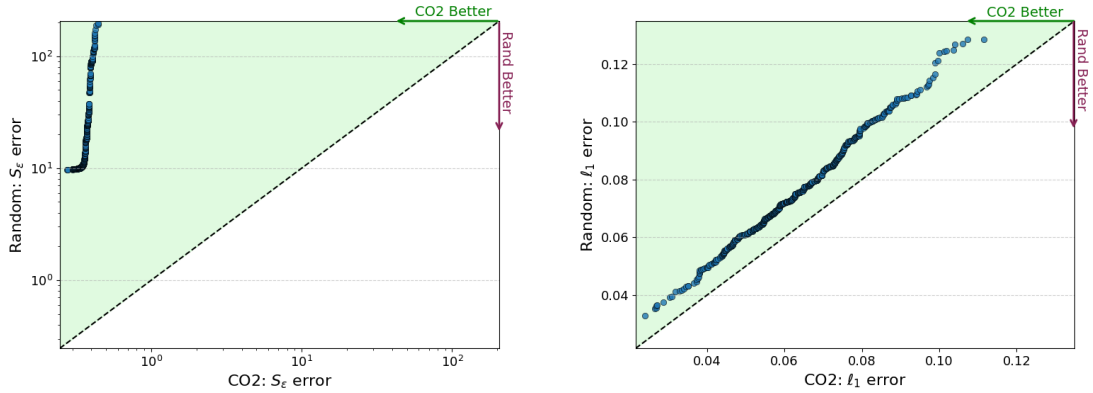


Figure 3.7: Q-Q plots of the Sinkhorn reconstruction error (left) and ℓ_1 error between the label proportions (right) of the compressed data as compared to random samples.

to random subsampling. However, we additionally see that this compression adequately encapsulates the dataset relative to the more concrete ℓ_1 error on label proportions defined above. This indicates that the quality of the approximation is meaningful beyond this particular divergence.

3.5 Discussion

The CO2 algorithm attains rapid, asymptotically lossless compression by targeting the second order expansion of a divergence of interest. This framework invites many natural follow-up questions, and paths towards improved algorithmic efficiency.

One problem of interest is the development of finite sample guarantees for this procedure. The second stage of the algorithm, MMD compression, depends on how accurately the selected empirical kernel can be approximated at low rank, which is amenable to finite-sample analysis; however, the relationship of these bounds to the performance of the coreset for the desired divergence is not immediate. Key to our analysis were asymptotic properties of the divergence of interest via a functional delta method. To achieve finite sample guarantees, explicit bounds on the remainder of the divergence relative to its quadratic approximation need to be developed.

Our procedure can benefit from future developments in convex MMD compression. Our current preferred instantiation of CO2 uses a recombination procedure. However, we anticipate that this algorithm can be improved as the state-of-the-art in this domain improves. In particular, the resulting rates may be suboptimal when the selected empirical kernel is only slowly approximable at low rank. As methodology improves in this domain to achieve minimax MMD compression guarantees, we believe pairing CO2 with such a procedure would lead to asymptotically minimax compression for all sufficiently regular divergences. Likewise, we speculate that in finite sample settings minimax MMD compression would lead to minimax CO2 compression whenever the remainder term satisfies appropriate regularity conditions.

A drawback of CO2 is that it requires identifying a kernel that controls the second-order expansion of the divergence. This may involve deriving the Hadamard operator itself or, as in the Sinkhorn application, proving that a simpler computable kernel is equivalent to it on the relevant tangent space. The guarantees also require second-order Hadamard differentiability in the topology natural to the functional and, for data-dependent kernels, an empirical-to-population transfer argument. It would therefore be useful to develop general conditions that verify these steps for broad classes of divergences. On the computational side, automatic differentiation may help approximate the required local quadratic form; related methods have recently been used to estimate first-order derivatives of statistical functionals (Luedtke, 2024).

3.A Functional Taylor expansion

Here we provide further details on the intuition of our method, linking compression with respect to a smooth divergence D to a carefully chosen MMD.

As we are in the setting of coreset selection, we desire the size of our compressed dataset $m \ll n$. The condition $D(P_m) = O_p(n^{-1})$ places a natural constraint on the class of functionals of interest. Typically, such D are required to be sufficiently smooth, made precise by the notion of Hadamard differentiability (Fernholz, 1983).

To understand how to minimize a Hadamard differentiable divergence, we first consider minimizing a smooth function f. If we Taylor expand about a point x, then we get a local expansion

$$f(y) - f(x) = \nabla f(x)^\top (y - x) + \tfrac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + o(\|y - x\|^2).$$

If x was a minimizer of f, then the first order gradient would vanish, leaving a dominant quadratic term from the Hessian. Now, suppose we generated a sequence $\{y_m\}$ such that

$\|(y_m - x)^\top \nabla^2 f(x)(y_m - x)\| = O(n^{-1})$. If the Hessian were non-degenerate, this would necessitate that the remainder term be $o(n^{-1})$, hence the overall error would be $|f(y_m) - f(x)| = O(n^{-1})$.

Second-order Hadamard differentiability gives the corresponding von Mises expansion along admissible moving directions (A. W. v. d. Vaart, 2000, Chapter 9). In particular,

$$D(P + th_t) = D(P) + t\dot{D}_P(h_t) + t^2\ddot{D}_P(h) + o(t^2)$$

whenever $h_t \to h$ in the chosen tangent topology. At the minimum of a differentiable divergence, the first derivative vanishes, so the leading term is the continuous quadratic form $\ddot{D}_P$. Lemma 3.1 uses precisely this expansion: if the coreset and empirical measure differ by $o_p(n^{-1/2})$ in a fixed RKHS tangent norm, then they have the same first-order tangent limit and their divergence values differ by $o_p(n^{-1})$.

The topology is part of the regularity statement. The kernel representing the quadratic form need not itself control the remainder unless second-order Hadamard differentiability holds in that kernel topology, or in an equivalent stronger topology. This is why KernelSelection may return a dominating or equivalent kernel rather than the exact Hadamard kernel. If the selected kernel is estimated from the data, one must additionally transfer the empirical compression error to the population topology before applying the expansion.

3.B Algorithms

In this section, we briefly present the internal algorithms used in our implementation of Algorithm 2. Algorithm 4 presents a Nyström approximation algorithm. The recombination step is the standard routine described in (Litterer and Lyons, 2012a). For a more detailed and efficient implementation, see https://github.com/satoshi-hayakawa/kernel-quadrature.

3.C General facts

Where convenient, we denote $k_\varepsilon(x, y) := \exp\left(\frac{-\|x-y\|^2}{\varepsilon}\right)$, the kernel function of $\mathcal{H}$.

Algorithm 4 Nyström approximation (Tropp et al., 2017, Algorithm 3)

Input: matrix A, must be PSD; $\theta \in \mathbb{N},\ 2 \le \theta \le n$; rank parameter $1 \le r \le \theta$

$\Omega \leftarrow \text{randn}(n, \theta)$

$\mu \leftarrow$ small value to improve numerical stability

$\Omega, \sim \leftarrow \text{QR}(\Omega)$

$Y \leftarrow A\Omega$

$\nu \leftarrow \mu \text{norm}(Y)$

$Y \leftarrow Y + \nu\Omega$

$B \leftarrow \Omega^* Y$

$C \leftarrow \text{chol}\left((B + B^*)/2\right)$ % Cholesky decomposition

$(U, \Sigma, \sim) \leftarrow \text{svd}(Y/C, \text{'econ'})$ % thin SVD

$U \leftarrow U(:, 1:r)$, $\Sigma \leftarrow \Sigma(1:r, 1:r)$ % truncate to rank r

$\Lambda \leftarrow \max\left\{0, \Sigma^2 - \nu I\right\}$

return $U\Lambda U^\top, U, \Lambda$

3.C.1 Entropic optimal transport

We briefly summarize key properties of entropic optimal transport used in our argument. See Section 3.3.2 for basic definitions. The foundational results were established in (Csiszár, 1975; Rüschendorf and Thomsen, 1993).

Lemma 3.6 (EOT coupling and marginal identities)**.**

$$\int f(x,y)\, d\pi^{\varepsilon}_{\mu,\nu}(x,y) = \int \exp\left(\frac{\phi_{\mu,\nu}(x) + \psi_{\mu,\nu}(y) - \|x-y\|^2}{\varepsilon}\right) f(x,y)\, d\mu(x) d\nu(y).$$

In particular,

$$\int f(x)g(y)\, d\pi^{\varepsilon}_{\mu,\nu}(x,y) = \int f\mathcal{A}^* g\, d\mu = \int g\mathcal{A} f\, d\nu$$
$$\int f(x) + g(y)\, d\pi^{\varepsilon}_{\mu,\nu}(x,y) = \int f\, d\mu + \int g\, d\nu.$$

Proof. The first identity is definitional, and the second and third follow by integrating along marginals. □

Corollary 3.1 (Schrödinger normalization identities)**.**

$$\int \exp\left(\frac{\phi_{\mu,\nu}(x)+\psi_{\mu,\nu}(y)-\|x-y\|^2}{\varepsilon}\right)d\mu(x)d\nu(y)=1$$

In fact, we have,

$$\int \exp\left(\frac{\phi_{\mu,\nu}(x)+\psi_{\mu,\nu}(\cdot)-\|x-\cdot\|^2}{\varepsilon}\right)d\mu(x)\equiv 1\equiv \int \exp\left(\frac{\phi_{\mu,\nu}(\cdot)+\psi_{\mu,\nu}(y)-\|\cdot-y\|^2}{\varepsilon}\right)d\nu(y)$$

Proof. The first identity is immediate, since

$$\int \exp\left(\frac{\phi_{\mu,\nu}(x)+\psi_{\mu,\nu}(y)-\|x-y\|^2}{\varepsilon}\right)d\mu(x)d\nu(y)=\int d\pi^{\varepsilon}_{\mu,\nu}=1,$$

as $\pi^{\varepsilon}_{\mu,\nu}$ is a probability distribution. The second claim is a consequence of the Schrödinger equations. □

Proposition 3.1 (Dual-potential representation of entropic transport)**.**

$$\mathrm{OT}_{\varepsilon}(\mu,\nu)=\int \phi_{\mu,\nu}d\mu+\int \psi_{\mu,\nu}d\nu.$$

Proof.

$$\begin{aligned}
\mathrm{OT}_{\varepsilon}(\mu,\nu) &= \int \|x-y\|^2 d\pi^{\varepsilon}_{\mu,\nu}+\varepsilon\int \log\left(\frac{d\pi^{\varepsilon}_{\mu,\nu}}{d\mu\otimes d\nu}\right)d\pi^{\varepsilon}_{\mu,\nu}\\
&= \int \|x-y\|^2 d\pi^{\varepsilon}_{\mu,\nu}+\varepsilon\int \log\circ\exp\left(\frac{\phi_{\mu,\nu}(x)+\psi_{\mu,\nu}(y)-\|x-y\|^2}{\varepsilon}\right)d\pi^{\varepsilon}_{\mu,\nu}\\
&= \int \phi_{\mu,\nu}(x)+\psi_{\mu,\nu}(y)d\pi^{\varepsilon}_{\mu,\nu}\\
&= \int \phi_{\mu,\nu}d\mu+\int \psi_{\mu,\nu}d\nu.
\end{aligned}$$

□

Lemma 3.7 (Symmetry and normalization of self-transport potentials)**.** $\phi_{\mu,\mu}=\psi_{\mu,\mu}$ *under the constraint* $\phi_{\mu,\mu}(x_0)=\psi_{\mu,\mu}(x_0)$.

Additionally, for μ,ν *we have*

$$\int \exp\left(\frac{\phi_{\mu,\mu}(x)+\psi_{\mu,\mu}(y)-\|x-y\|^2}{\varepsilon}\right)d\mu(x)d\nu(y)=1,$$
$$\int \exp\left(\frac{\phi_{\nu,\nu}(x)+\psi_{\nu,\nu}(y)-\|x-y\|^2}{\varepsilon}\right)d\mu(x)d\nu(y)=1.$$

Proof. The first identity was observed in (Feydy et al., 2019a) and (Goldfeld et al., 2022). The second comes from the Schrödinger equations, as

$$\int \exp\left(\frac{\phi_{\mu,\mu}(x) + \psi_{\mu,\mu}(y) - \|x-y\|^2}{\varepsilon}\right) d\mu(x) = 1$$

holds identically for all y. □

3.C.2 Reproducing kernel Hilbert space

Here we recall properties of the RKHS which will be essential for our analysis.

Lemma 3.8 (RKHS sup-norm bound)**.** *For $h \in \mathcal{H}$,*

$$\|h\|_\infty \le \|h\|_{\mathcal{H}}.$$

Proof.

$$\begin{aligned}\|h\|_\infty &= \sup_{y\in\mathcal{X}} |h(y)| = \sup_{y\in\mathcal{X}} \langle h, k_\varepsilon(y,\cdot)\rangle_{\mathcal{H}} \le \|h\|_{\mathcal{H}} \sup_{y\in\mathcal{X}} \|k_\varepsilon(y,\cdot)\|_{\mathcal{H}} \\ &= \|h\|_{\mathcal{H}} \sup_{y\in\mathcal{X}} \sqrt{k_\varepsilon(y,y)} = \|h\|_{\mathcal{H}}.\end{aligned}$$

□

Lemma 3.9 (Kernel mean embedding bound)**.** *For $\mu \in \mathcal{P}(\mathcal{X})$ and $f \in C(\mathcal{X})$,*

$$\|K(f d\mu)\|_{\mathcal{H}} \le \|f\|_\infty.$$

Proof. Using $|k_\varepsilon(x,y)| \le 1$,

$$\begin{aligned}\|K(f d\mu)\|_{\mathcal{H}}^2 &= \iint k_\varepsilon(x,y) f(x) f(y)\, d\mu(x) d\mu(y) \\ &\le \iint |f(x) f(y)|\, d\mu(x) d\mu(y) \\ &\le \|f\|_\infty^2.\end{aligned}$$

□

Lemma 3.10 (Scaled tangent space and exact isometry)**.** *Let $a \in \mathcal{H}$ satisfy $\inf_{x\in\mathcal{X}} a(x) > 0$, and define*

$$\widetilde{k}_a(x,y) := \frac{k_\varepsilon(x,y)}{a(x)a(y)}.$$

For a finite signed measure γ, put

$$J_a\gamma := K\left(\frac{d\gamma}{a}\right), \qquad \|\gamma\|_{\mathcal{E}_a} := \|J_a\gamma\|_{\mathcal{H}},$$

and let $\mathcal{E}_a$ be the completion of finite signed measures in this norm. Then $\widetilde{k}_a$ is positive definite, its RKHS is

$$\widetilde{\mathcal{H}}_a = \left\{\frac{h}{a} : h \in \mathcal{H}\right\}, \qquad \left\|\frac{h}{a}\right\|_{\widetilde{\mathcal{H}}_a} = \|h\|_{\mathcal{H}},$$

and J_a extends uniquely to an isometric isomorphism $J_a : \mathcal{E}_a \to \mathcal{H}$. Moreover,

$$\gamma(\mathbf{1}) = \langle J_a\gamma, a\rangle_{\mathcal{H}},$$

so $\mathcal{E}_a^0 := \{\gamma : \gamma(\mathbf{1}) = 0\}$ is a closed subspace.

Proof. A convenient feature map for $\widetilde{k}_a$, taking values in the original RKHS $\mathcal{H}$, is

$$\Phi_a(x) := \frac{k_\varepsilon(\cdot, x)}{a(x)}.$$

Indeed,

$$\langle \Phi_a(x), \Phi_a(y)\rangle_{\mathcal{H}} = \frac{\langle k_\varepsilon(\cdot, x), k_\varepsilon(\cdot, y)\rangle_{\mathcal{H}}}{a(x)a(y)} = \frac{k_\varepsilon(x, y)}{a(x)a(y)} = \widetilde{k}_a(x, y).$$

To make the corresponding RKHS identification explicit, consider first a finite linear combination of kernel sections,

$$f = \sum_{i=1}^{m} c_i \widetilde{k}_a(\cdot, x_i).$$

Using the definition of $\widetilde{k}_a$,

$$f(\cdot) = \frac{1}{a(\cdot)} \sum_{i=1}^{m} \frac{c_i}{a(x_i)} k_\varepsilon(\cdot, x_i).$$

Thus, if

$$h := \sum_{i=1}^{m} \left(\frac{c_i}{a(x_i)}\right) k_\varepsilon(\cdot, x_i) \in \mathcal{H},$$

then $f = h/a$. Moreover,

$$\begin{aligned}
\|f\|^2_{\widetilde{\mathcal{H}}_a} &= \sum_{i,j=1}^{m} c_i c_j \widetilde{k}_a(x_i, x_j) \\
&= \sum_{i,j=1}^{m} \frac{c_i c_j}{a(x_i)a(x_j)} k_\varepsilon(x_i, x_j) \\
&= \left\| \sum_{i=1}^{m} \frac{c_i}{a(x_i)} k_\varepsilon(\cdot, x_i) \right\|^2_{\mathcal{H}} \\
&= \|h\|^2_{\mathcal{H}}.
\end{aligned}$$

This proves the first claim Now consider a finite signed measure γ. By definition,

$$J_a\gamma := K\left(\frac{d\gamma}{a}\right) = \int \frac{k_\varepsilon(\cdot, x)}{a(x)}\, d\gamma(x).$$

Hence

$$\begin{aligned}
\|J_a\gamma\|^2_{\mathcal{H}} &= \left\langle \int \frac{k_\varepsilon(\cdot, x)}{a(x)}\, d\gamma(x), \int \frac{k_\varepsilon(\cdot, y)}{a(y)}\, d\gamma(y) \right\rangle_{\mathcal{H}} \\
&= \iint \frac{k_\varepsilon(x, y)}{a(x)a(y)}\, d\gamma(x)\, d\gamma(y) \\
&= \|\gamma\|^2_{\mathcal{E}_a}.
\end{aligned}$$

Thus J_a is an isometry on finite signed measures and therefore extends uniquely by continuity to an isometry

$$J_a : \mathcal{E}_a \longrightarrow \mathcal{H}.$$

For each $x \in \mathcal{X}$,

$$J_a\delta_x = \frac{k_\varepsilon(\cdot, x)}{a(x)}.$$

Consequently, the range of J_a contains

$$\operatorname{span}\left\{ \frac{k_\varepsilon(\cdot, x)}{a(x)} : x \in \mathcal{X} \right\}.$$

Since $a(x) > 0$ for every x, this span is the same as $\operatorname{span}\{k_\varepsilon(\cdot, x) : x \in \mathcal{X}\}$, which is dense in $\mathcal{H}$. The extension of J_a is an isometry from the complete space $\mathcal{E}_a$, so its range is closed. Its range is therefore both dense and closed in $\mathcal{H}$, and hence J_a is onto.

Finally, for a finite signed measure γ, the reproducing property gives

$$\langle J_a\gamma, a\rangle_{\mathcal{H}} = \int \left\langle \frac{k_\varepsilon(\cdot,x)}{a(x)}, a \right\rangle_{\mathcal{H}} d\gamma(x) = \int \frac{a(x)}{a(x)}\, d\gamma(x) = \gamma(\mathbf{1}).$$

Since both sides are continuous with respect to the $\mathcal{E}_a$ norm, this identity extends to all $\gamma \in \mathcal{E}_a$. □

Lemma 3.11 (Weak topology on probability measures)**.** *The scaled RKHS $\widetilde{\mathcal{H}}_a$ is universal on compact $\mathcal{X}$. Consequently, the MMD induced by $\widetilde{k}_a$ metrizes weak convergence on $\mathcal{P}(\mathcal{X})$.*

Proof. For any $f \in C(\mathcal{X})$, the product af is continuous. Universality of the Gaussian RKHS provides $h_j \in \mathcal{H}$ with $h_j \to af$ uniformly, and then $h_j/a \to f$ uniformly. Thus $\widetilde{\mathcal{H}}_a$ is universal. A bounded continuous universal kernel on a compact space metrizes weak convergence of probability measures through its MMD. □

Lemma 3.12 (Compact RKHS test classes)**.** *The unit ball $\mathcal{H}_1$ is uniformly bounded and equicontinuous in $C(\mathcal{X})$, hence relatively compact in the uniform norm. If a is continuous and bounded away from zero, then each of the families*

$$\left\{ \frac{h}{a} : h \in \mathcal{H}_1 \right\}, \qquad \left\{ \frac{gh}{a^2} : g, h \in \mathcal{H}_1 \right\}$$

is also relatively compact in $C(\mathcal{X})$.

Proof. The reproducing property gives $\|h\|_\infty \le \|h\|_{\mathcal{H}}$ and

$$|h(x) - h(y)| = |\langle h, k_\varepsilon(\cdot,x) - k_\varepsilon(\cdot,y)\rangle_{\mathcal{H}}|. \le \|h\|_{\mathcal{H}} \|k_\varepsilon(\cdot,x) - k_\varepsilon(\cdot,y)\|_{\mathcal{H}}.$$

Continuity of the kernel on compact $\mathcal{X}^2$ makes the right-hand side uniformly small as $y \to x$. Arzelà–Ascoli gives relative compactness. Division by a fixed positive continuous function and multiplication are continuous operations in the uniform norm, proving the remaining assertions. □

Corollary 3.2 (Uniform integration over compact classes)**.** *If $\rho_n, \rho \in \mathcal{P}(\mathcal{X})$ and $\rho_n \rightsquigarrow \rho$, while $\mathcal{F} \subset C(\mathcal{X})$ is relatively compact in the uniform norm, then*

$$\sup_{f\in\mathcal{F}} \left| \int f\, d(\rho_n - \rho) \right| \longrightarrow 0.$$

The same conclusion holds for $\Delta_n = \rho_n - \eta_n$ whenever $\Delta_n \rightsquigarrow 0$ and $\|\Delta_n\|_{\mathrm{tv}} \le 2$.

Proof. Fix $\delta > 0$ and choose a finite δ-net $f_1, \ldots, f_N$ for the uniform closure of $\mathcal{F}$. Since the relevant signed differences have total variation at most two,

$$\sup_{f\in\mathcal{F}} |(\rho_n - \rho)f| \le \sup_{f\in\mathcal{F}} \inf_j |(\rho_n - \rho)f_j| + |(\rho_n - \rho)(f - f_j)| \le \max_{1\le j\le N} |(\rho_n - \rho)f_j| + 2\delta.$$

The maximum tends to zero by weak convergence. Letting $\delta \downarrow 0$ proves the claim; the proof for $\rho_n - \eta_n$ is identical. □

3.D Functional compression

Lemma 3.13 (Vanishing derivative at a divergence minimum)**.** *If D is a divergence and $D(\cdot) := D(\,\cdot\,, P)$ is Hadamard differentiable, then the Hadamard derivative vanishes at P.*

Proof. This argument is analogous to the Euclidean case where the gradient of a differentiable function vanishes at a local minimum. Indeed, observe for any $t, h_t \to h$, $D(P+th_t) - D(P) \ge 0$, and so

$$\lim_{t\uparrow 0} \frac{D(P + th_t) - D(P)}{t} \le 0, \quad \lim_{t\downarrow 0} \frac{D(P + th_t) - D(P)}{t} \ge 0.$$

The Hadamard differentiability of D implies the two limits above agree, and so

$$\dot{D}(h) := \lim_{t\to 0} \frac{D(P + th_t) - D(P)}{t} = 0.$$

As h was arbitrary, $\dot{D} = 0$. □

Lemma 3.14 (Donsker property of a continuous-kernel RKHS ball)**.** *Suppose that $\mathcal{F}$ is a RKHS with continuous kernel. Then $\mathcal{F}_1$, the unit ball of this space, is P-Donsker.*

Proof. This is an immediate consequence of Mercer's theorem (Berlinet and Thomas-Agnan, 2011, Theorem 40) in combination with A. v. d. Vaart and Wellner, 2023, Theorem 2.13.2. □

Proof of Lemma 3.1. By Lemma 3.13, the first Hadamard derivative vanishes at P. $\sqrt{n}(P_m - P) = \sqrt{n}(P_n - P) + \sqrt{n}(P_m - P_n) = \sqrt{n}(P_n - P) + o_p(1) \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F}_1)$ by Lemma 3.14. Hence, we can apply the second order functional delta method Goldfeld et al., 2022, Lemma 13 to get

$$nD(P_m) = n[D(P_m) - D(P) - \dot{D}(P_m - P)] \rightsquigarrow Q_G(\mathbb{G}) = O_p(1).$$

To get sharper control, we can apply the second part of Römisch, 2004, Theorem 2 to get

$$n[D(P_m) - D(P) - \ddot{D}(P_m - P)] = o_p(1),$$
$$n[D(P_n) - D(P) - \ddot{D}(P_n - P)] = o_p(1).$$

Hence, by the continuous mapping theorem,

$$n[D(P_m) - D(P_n) - \ddot{D}(P_m - P) + \ddot{D}(P_n - P)] = o_p(1),$$

and so

$$\begin{aligned} D(P_m) - D(P_n) &= \ddot{D}(P_m - P) - \ddot{D}(P_n - P) + o_p(1/n) \\ &= \int G(P_m - P)d(P_m - P) - \int G(P_n - P)d(P_n - P) + o_p(1/n) \\ &= \int G(P_m - P_n)d(P_m - P_n) + 2\int G(P_m - P_n)d(P_n - P) + o_p(1/n) \\ &= \mathrm{MMD}_G^2(P_m, P_n) + 2\int G(P_m - P_n)d(P_n - P) + o_p(1/n). \end{aligned}$$

The second term above is an inner product in the RKHS with kernel g between the differences of the g-kernel mean embeddings of P_m and P_n and P_n and P. Hence, applying Cauchy-Schwarz yields

$$|D(P_m) - D(P_n)| \leq \mathrm{MMD}_G^2(P_m, P_n) + 2\,\mathrm{MMD}_G(P_m, P_n)\,\mathrm{MMD}_G(P_n, P) + o_p(1/n).$$

Since P_n is the empirical distribution of an iid sample from P, $\mathrm{MMD}_G(P_n, P) = O_p(n^{-1/2})$, thus in combination with Lemma 3.3,

$$D(P_m) = D(P_n) + o_p(1/n).$$

□

Proof of Lemma 3.3. Since $\ddot{D}$ is a continuous quadratic form in the K-tangent there exists $C < \infty$ such that

$$|\ddot{D}(h)| \leq C\|Kh\|_{\mathcal{F}}^2$$

for every admissible signed tangent h. Since

$$\ddot{D}(h) = \int Gh\,dh, \qquad \|Kh\|_{\mathcal{F}}^2 = \int Kh\,dh,$$

we obtain

$$\int (CK - G)h \, dh \geq 0.$$

Thus $CK \succeq G$. □

3.E Entropic optimal transport

3.E.1 Regularity of potentials

We now establish Hadamard differentiability of $\tau(\mu, \nu) := (u_{\mu,\nu}, v_{\mu,\nu})$ at a fixed $(\mu, \nu) \in \mathcal{P}(\mathcal{X})^2$, tangentially to the scaled spaces $\mathcal{E}_u^0 \times \mathcal{E}_v^0$ introduced in Appendix 3.C.2. The zero-mass restriction is necessary because the two perturbed marginals remain probability measures.

Recall that

$$K\gamma := \int k_\varepsilon(\cdot, y) \, d\gamma(y).$$

For any strictly positive $a \in \mathcal{H}$ and $\rho \in \mathcal{P}(\mathcal{X})$, define

$$\mathsf{T}_{a,\rho} f := K\left(\frac{f \, d\rho}{a^2}\right). \tag{3.11}$$

Fix the anchored representative satisfying $u_{\mu,\nu}(x_0) = v_{\mu,\nu}(x_0)$. For brevity, write $u = u_{\mu,\nu}$ and $v = v_{\mu,\nu}$, and abbreviate

$$\mathsf{T}_\nu := \mathsf{T}_{v,\nu}, \qquad \mathsf{T}_\mu := \mathsf{T}_{u,\mu}, \qquad \mathsf{L} := \begin{bmatrix} I & \mathsf{T}_\nu \\ \mathsf{T}_\mu & I \end{bmatrix}.$$

$$(\mathcal{A}f)(y) := \int \xi_{\mu,\nu}^\varepsilon(x, y) f(x) \, d\mu(x), \qquad (\mathcal{A}^* g)(x) := \int \xi_{\mu,\nu}^\varepsilon(x, y) g(y) \, d\nu(y).$$

Thus

$$\mathsf{L}(h_1, h_2) = (h_1 + \mathsf{T}_\nu h_2, \mathsf{T}_\mu h_1 + h_2), \qquad \mathsf{T}_\nu v = u, \qquad \mathsf{T}_\mu u = v.$$

Lemma 3.6 gives

$$\langle \mathcal{A}f, g \rangle_{L^2(\nu)} = \langle f, \mathcal{A}^* g \rangle_{L^2(\mu)}.$$

They are contractions. For example, Jensen's inequality and the first marginal constraint give

$$|\mathcal{A}^* g(x)|^2 \leq \int \xi_{\mu,\nu}^\varepsilon(x, y) g(y)^2 \, d\nu(y),$$

and integration with respect to μ yields $\|\mathcal{A}^* g\|_{L^2(\mu)} \leq \|g\|_{L^2(\nu)}$; the argument for $\mathcal{A}$ is symmetric. Finally,

$$\mathsf{T}_\nu(vg) = u\mathcal{A}^* g, \qquad \mathsf{T}_\mu(uf) = v\mathcal{A} f.$$

Lemma 3.15 (Fredholm structure)**.** *The operators T_ν and T_μ are compact, self-adjoint, and positive on $\mathcal{H}$. Moreover,*

$$\ker(\mathsf{L}) = \operatorname{span}\{(u, -v)\}, \qquad \ker(\mathsf{L}^*) = \operatorname{span}\{(v, -u)\},$$

and

$$\operatorname{ran}(\mathsf{L}) = \mathcal{R} := \left\{(r_1, r_2) \in \mathcal{H}^2 : \langle r_1, v\rangle_{\mathcal{H}} = \langle r_2, u\rangle_{\mathcal{H}}\right\}. \tag{3.12}$$

Proof. Self-adjointness and positivity follow from

$$\langle f, \mathsf{T}_\nu g\rangle_{\mathcal{H}} = \int \frac{fg}{v^2} d\nu, \qquad \langle f, \mathsf{T}_\mu g\rangle_{\mathcal{H}} = \int \frac{fg}{u^2} d\mu.$$

Compactness follows from the compact embedding of bounded subsets of $\mathcal{H}$ into $C(\mathcal{X})$ and Lemma 3.9.

Suppose $\mathsf{L}(h_1, h_2) = 0$. Write $h_2 = vg$. The first equation gives

$$h_1 = -\mathsf{T}_\nu(vg) = -u\mathcal{A}^* g.$$

Substituting this into the second equation and using the identities above gives

$$0 = \mathsf{T}_\mu h_1 + h_2 = -v\mathcal{A}\mathcal{A}^* g + vg,$$

and hence

$$\mathcal{A}\mathcal{A}^* g = g.$$

Since $\mathcal{A}$ and $\mathcal{A}^*$ are contractions,

$$\|g\|_{L^2(\nu)} = \|\mathcal{A}\mathcal{A}^* g\|_{L^2(\nu)} \leq \|\mathcal{A}^* g\|_{L^2(\mu)} \leq \|g\|_{L^2(\nu)},$$

so equality holds throughout. For fixed x, let

$$m(x) := (\mathcal{A}^* g)(x), \qquad d\nu_x(y) := \xi^{\varepsilon}_{\mu,\nu}(x, y) d\nu(y).$$

Since ν_x is a probability measure,

$$\int (g(y) - m(x))^2 \, d\nu_x(y) = \int \xi^{\varepsilon}_{\mu,\nu}(x,y) g(y)^2 \, d\nu(y) - |\mathcal{A}^* g(x)|^2 \geq 0.$$

Integrating this identity with respect to μ gives

$$\begin{aligned} &\int \left[\int (g(y) - m(x))^2 \, d\nu_x(y) \right] d\mu(x) \\ &\quad = \iint g(y)^2 \xi^{\varepsilon}_{\mu,\nu}(x,y) \, d\nu(y) d\mu(x) - \int |\mathcal{A}^* g(x)|^2 \, d\mu(x) \\ &\quad = \|g\|^2_{L^2(\nu)} - \|\mathcal{A}^* g\|^2_{L^2(\mu)} = 0, \end{aligned}$$

where the second equality uses that the Y-marginal of $\pi^{\varepsilon}_{\mu,\nu}$ is ν, and the final equality follows from equality in the contraction bound above. Since the inner integral is nonnegative, it follows that

$$\int (g(y) - m(x))^2 \, d\nu_x(y) = 0$$

for μ-almost every x. Therefore $g(y) = m(x)$ ν_x-almost everywhere. Since $\xi^{\varepsilon}_{\mu,\nu}(x,y) > 0$, ν_x is equivalent to ν, so $g(y) = m(x)$ ν-almost everywhere. Thus g is constant ν-almost everywhere. Hence $g = c$ for some $c \in \mathbb{R}$, so that

$$h_2 = cv, \qquad h_1 = -cu.$$

Therefore,

$$\ker(\mathsf{L}) = \operatorname{span}\{(u, -v)\}.$$

Since T_ν and T_μ are self-adjoint,

$$\mathsf{L}^* = \begin{bmatrix} I & \mathsf{T}_\mu \\ \mathsf{T}_\nu & I \end{bmatrix}.$$

The same argument with the roles of the two marginals interchanged yields

$$\ker(\mathsf{L}^*) = \operatorname{span}\{(v, -u)\}.$$

Finally, L is the identity plus a compact operator, so it is Fredholm of index zero. The Fredholm alternative gives

$$\operatorname{ran}(\mathsf{L}) = \ker(\mathsf{L}^*)^{\perp}.$$

Since

$$\ker(\mathsf{L}^*) = \operatorname{span}\{(v, -u)\},$$

this is precisely

$$\operatorname{ran}(\mathsf{L}) = \left\{(r_1, r_2) \in \mathcal{H}^2 : \langle r_1, v\rangle_{\mathcal{H}} = \langle r_2, u\rangle_{\mathcal{H}}\right\},$$

which is (3.12). □

Lemma 3.16 (Anchored inverse)**.** *Let*

$$\mathcal{D}_0 := \{(h_1, h_2) \in \mathcal{H}^2 : h_1(x_0) - h_2(x_0) = 0\}.$$

Then the restriction

$$\mathsf{L}^\circ := \mathsf{L}|_{\mathcal{D}_0} : \mathcal{D}_0 \longrightarrow \mathcal{R}$$

is a bounded bijection and has a bounded inverse.

Proof. By Lemma 3.15, the only null direction is $(u, -v)$, which is not in $\mathcal{D}_0$ because $u(x_0) = v(x_0) > 0$. Thus the restriction is injective. For $r \in \mathcal{R}$, choose h with $\mathsf{L}h = r$ and add the unique multiple of $(u, -v)$ that enforces the anchor. This proves surjectivity, and the open mapping theorem gives boundedness of the inverse. □

Lemma 3.17 (Uniform bounds and weak stability)**.** *Let* $\kappa := \min_{x,y\in\mathcal{X}} k_\varepsilon(x, y) > 0$*. For every* $\alpha, \beta \in \mathcal{P}(\mathcal{X})$*, the anchored scalings satisfy*

$$\kappa^2 \le u_{\alpha,\beta}(x), v_{\alpha,\beta}(x) \le \kappa^{-3/2}, \qquad x \in \mathcal{X},$$

and the family of all anchored scalings is equicontinuous. If $\alpha_n \rightsquigarrow \alpha$ *and* $\beta_n \rightsquigarrow \beta$*, then the corresponding anchored scalings converge uniformly.*

Proof. Writing

$$a := u_{\alpha,\beta}(x_0) = v_{\alpha,\beta}(x_0),$$

observe first that

$$\kappa \le \frac{k_\varepsilon(x, y)}{k_\varepsilon(x_0, y)} \le \kappa^{-1},$$

since $\kappa \le k_\varepsilon \le 1$. Hence

$$\begin{aligned} u(x) &= \int \frac{k_\varepsilon(x,y)}{k_\varepsilon(x_0,y)} \frac{k_\varepsilon(x_0,y)}{v(y)}\, d\beta(y) \\ &\in \left[\kappa \int \frac{k_\varepsilon(x_0,y)}{v(y)}\, d\beta(y),\ \kappa^{-1} \int \frac{k_\varepsilon(x_0,y)}{v(y)}\, d\beta(y)\right] \\ &= [\kappa a, a/\kappa]. \end{aligned}$$

The same argument applied to the second scaling equation gives

$$\kappa a \le v(x) \le a/\kappa.$$

Evaluating the first scaling equation at x_0 and using these bounds,

$$a = \int \frac{k_\varepsilon(x_0,y)}{v(y)}\, d\beta(y) \ge \int \frac{\kappa}{a/\kappa}\, d\beta(y) = \frac{\kappa^2}{a},$$

so $a \ge \kappa$. Likewise,

$$a = \int \frac{k_\varepsilon(x_0,y)}{v(y)}\, d\beta(y) \le \int \frac{1}{\kappa a}\, d\beta(y) = \frac{1}{\kappa a},$$

so $a \le \kappa^{-1/2}$. Therefore

$$\kappa^2 \le u_{\alpha,\beta}(x), v_{\alpha,\beta}(x) \le \kappa^{-3/2}.$$

To establish equicontinuity, let

$$\omega_k(\delta) := \sup_{\substack{x,x',y \in \mathcal{X} \\ \|x-x'\| \le \delta}} |k_\varepsilon(x,y) - k_\varepsilon(x',y)|.$$

Since k_ε is uniformly continuous on the compact set $\mathcal{X} \times \mathcal{X}$, $\omega_k(\delta) \to 0$ as $\delta \downarrow 0$. Using the uniform lower bound $v_{\alpha,\beta} \ge \kappa^2$, for any $x, x' \in \mathcal{X}$,

$$\begin{aligned} |u_{\alpha,\beta}(x) - u_{\alpha,\beta}(x')| &= \left| \int \frac{k_\varepsilon(x,y) - k_\varepsilon(x',y)}{v_{\alpha,\beta}(y)}\, d\beta(y) \right| \\ &\le \kappa^{-2} \sup_{y \in \mathcal{X}} |k_\varepsilon(x,y) - k_\varepsilon(x',y)| \\ &\le \kappa^{-2} \omega_k(\|x - x'\|). \end{aligned}$$

The same calculation, using $u_{\alpha,\beta} \ge \kappa^2$, gives

$$|v_{\alpha,\beta}(x) - v_{\alpha,\beta}(x')| \le \kappa^{-2} \omega_k(\|x - x'\|).$$

Thus the family of anchored scalings is uniformly bounded and equicontinuous.

Now suppose $\alpha_n \rightsquigarrow \alpha$ and $\beta_n \rightsquigarrow \beta$, and write

$$u_n := u_{\alpha_n,\beta_n}, \qquad v_n := v_{\alpha_n,\beta_n}.$$

By the uniform bounds and equicontinuity, Arzelà–Ascoli implies that every subsequence of (u_n, v_n) has a further subsequence, which we continue to index by n, such that

$$u_n \to \bar{u}, \qquad v_n \to \bar{v}$$

uniformly on $\mathcal{X}$. The uniform lower bounds pass to the limit, so

$$\bar{u}, \bar{v} \geq \kappa^2.$$

We next identify the limiting pair. From the first scaling equation,

$$u_n = K\left(\frac{d\beta_n}{v_n}\right).$$

Hence

$$u_n - K\left(\frac{d\beta}{\bar{v}}\right) = K\left[\left(\frac{1}{v_n} - \frac{1}{\bar{v}}\right) d\beta_n\right] + K\left(\frac{d\beta_n - d\beta}{\bar{v}}\right).$$

For the first term, Lemma 3.9 gives

$$\left\| K\left[\left(\frac{1}{v_n} - \frac{1}{\bar{v}}\right) d\beta_n\right] \right\|_{\mathcal{H}} \leq \left\| \frac{1}{v_n} - \frac{1}{\bar{v}} \right\|_{\infty} \longrightarrow 0,$$

because $v_n \to \bar{v}$ uniformly and all these functions are uniformly bounded away from zero.

For the second term, using the reproducing property,

$$\left\| K\left(\frac{d\beta_n - d\beta}{\bar{v}}\right) \right\|_{\mathcal{H}} = \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \int \frac{h}{\bar{v}}\, d(\beta_n - \beta) \right|.$$

By Lemma 3.12, the class

$$\left\{ \frac{h}{\bar{v}} : \|h\|_{\mathcal{H}} \leq 1 \right\}$$

is relatively compact in $C(\mathcal{X})$. Since $\beta_n \rightsquigarrow \beta$, Corollary 3.2 therefore implies

$$\left\| K\left(\frac{d\beta_n - d\beta}{\bar{v}}\right) \right\|_{\mathcal{H}} \longrightarrow 0.$$

Consequently,

$$u_n \longrightarrow K\left(\frac{d\beta}{\bar{v}}\right) \qquad \text{in } \mathcal{H},$$

and hence also uniformly by Lemma 3.8. Since we already know that $u_n \to \bar{u}$ uniformly, uniqueness of the uniform limit gives

$$\bar{u} = K\left(\frac{d\beta}{\bar{v}}\right).$$

The same argument applied to the second scaling equation yields

$$\bar{v} = K\left(\frac{d\alpha}{\bar{u}}\right).$$

Moreover, since $u_n(x_0) = v_n(x_0)$ for every n, uniform convergence gives

$$\bar{u}(x_0) = \bar{v}(x_0).$$

Thus $(\bar{u}, \bar{v})$ satisfies the Schrödinger scaling equations for (α, β) together with the anchoring constraint. By uniqueness of the anchored solution,

$$(\bar{u}, \bar{v}) = (u_{\alpha,\beta}, v_{\alpha,\beta}).$$

We have therefore shown that every subsequence of (u_n, v_n) has a further subsequence converging uniformly to $(u_{\alpha,\beta}, v_{\alpha,\beta})$. It follows that the full sequence converges:

$$u_{\alpha_n,\beta_n} \to u_{\alpha,\beta}, \qquad v_{\alpha_n,\beta_n} \to v_{\alpha,\beta}$$

uniformly on $\mathcal{X}$. □

Lemma 3.18 ($\mathcal{H}$-stability along scaled paths)**.** *Suppose*

$$\mu_t = \mu + t\gamma_t^1, \qquad \nu_t = \nu + t\gamma_t^2$$

are probability measures with $\gamma_t^1 \to \gamma^1$ in $\mathcal{E}_u^0$ and $\gamma_t^2 \to \gamma^2$ in $\mathcal{E}_v^0$. Then the corresponding anchored scalings satisfy $u_t \to u$ and $v_t \to v$ in $\mathcal{H}$.

Proof. Scaled-MMD convergence implies weak convergence of the probability paths by Lemma 3.11, and Lemma 3.17 gives uniform convergence of the scalings. Using the scaling

equations and inserting the intermediate term $K(d\nu_t/v)$,

$$\begin{aligned} u_t - u &= K\left(\frac{d\nu_t}{v_t}\right) - K\left(\frac{d\nu}{v}\right) \\ &= K\left[\left(\frac{1}{v_t} - \frac{1}{v}\right) d\nu_t\right] + K\left(\frac{d\nu_t - d\nu}{v}\right) \\ &= K\left[\left(\frac{1}{v_t} - \frac{1}{v}\right) d\nu_t\right] + tJ_v\gamma_t^2. \end{aligned}$$

Therefore, by the triangle inequality,

$$\|u_t - u\|_{\mathcal{H}} \leq \left\| K\left[\left(\frac{1}{v_t} - \frac{1}{v}\right) d\nu_t\right]\right\|_{\mathcal{H}} + |t|\|J_v\gamma_t^2\|_{\mathcal{H}}.$$

Since ν_t is a probability measure, Lemma 3.9 gives

$$\left\| K\left[\left(\frac{1}{v_t} - \frac{1}{v}\right) d\nu_t\right]\right\|_{\mathcal{H}} \leq \left\|\frac{1}{v_t} - \frac{1}{v}\right\|_{\infty}.$$

By the uniform lower bound from Lemma 3.17, $v_t, v \geq \kappa^2$, and hence

$$\left\|\frac{1}{v_t} - \frac{1}{v}\right\|_{\infty} = \left\|\frac{v - v_t}{v_t v}\right\|_{\infty} \leq \kappa^{-4}\|v_t - v\|_{\infty} \longrightarrow 0,$$

because $v_t \to v$ uniformly.

For the second term, convergence $\gamma_t^2 \to \gamma^2$ in $\mathcal{E}_v$ and the isometry of $J_v : \mathcal{E}_v \to \mathcal{H}$ imply

$$\|J_v\gamma_t^2 - J_v\gamma^2\|_{\mathcal{H}} = \|\gamma_t^2 - \gamma^2\|_{\mathcal{E}_v} \longrightarrow 0.$$

In particular, $\|J_v\gamma_t^2\|_{\mathcal{H}} = O(1)$, so

$$|t|\|J_v\gamma_t^2\|_{\mathcal{H}} \longrightarrow 0.$$

Combining the two bounds yields

$$\|u_t - u\|_{\mathcal{H}} \longrightarrow 0.$$

The argument for $v_t \to v$ in $\mathcal{H}$ is identical. □

Lemma 3.19 (Secant equations)**.** *Define*

$$h_{1,t} := \frac{u_t - u}{t}, \qquad h_{2,t} := \frac{v_t - v}{t},$$

and

$$\mathsf{T}_{\nu,t}h := K\left(h\frac{d\nu_t}{vv_t}\right), \qquad \mathsf{T}_{\mu,t}h := K\left(h\frac{d\mu_t}{uu_t}\right).$$

Then the following identities are exact:

$$h_{1,t} + \mathsf{T}_{\nu,t}h_{2,t} = J_v\gamma_t^2, \tag{3.13}$$

$$\mathsf{T}_{\mu,t}h_{1,t} + h_{2,t} = J_u\gamma_t^1, \tag{3.14}$$

and $(h_{1,t}, h_{2,t}) \in \mathcal{D}_0$.

Proof. Subtract the first equations in (3.4) and insert the intermediate term $K(d\nu_t/v)$:

$$\frac{u_t - u}{t} = K\left(d\nu_t\frac{v - v_t}{tvv_t}\right) + K\left(\frac{d\nu_t - d\nu}{tv}\right).$$

This is the first identity; the second is symmetric. The anchor gives membership in $\mathcal{D}_0$. □

Lemma 3.20 (Operator convergence of the secants)**.** *Under the assumptions of Lemma 3.18,*

$$\|\mathsf{T}_{\nu,t} - \mathsf{T}_\nu\|_{\mathcal{H}\to\mathcal{H}} \to 0, \qquad \|\mathsf{T}_{\mu,t} - \mathsf{T}_\mu\|_{\mathcal{H}\to\mathcal{H}} \to 0.$$

Proof. For $\|h\|_{\mathcal{H}} \le 1$,

$$(\mathsf{T}_{\nu,t} - \mathsf{T}_\nu)h = K\left[h\left(\frac{1}{vv_t} - \frac{1}{v^2}\right)d\nu_t\right] + K\left(\frac{h}{v^2}d(\nu_t - \nu)\right).$$

The first term tends to zero uniformly in h by uniform convergence of v_t. For the second,

$$\sup_{\|h\|_{\mathcal{H}}\le 1}\left\|K\left(\frac{h}{v^2}d(\nu_t - \nu)\right)\right\|_{\mathcal{H}} = \sup_{\substack{\|g\|_{\mathcal{H}}\le 1\\ \|h\|_{\mathcal{H}}\le 1}}\left|\int\frac{gh}{v^2}d(\nu_t - \nu)\right|.$$

The test family is relatively compact by Lemma 3.12, so Corollary 3.2 proves convergence. The proof for $\mathsf{T}_{\mu,t}$ is identical. □

Lemma 3.21 (Fixed-range perturbation)**.** *Let* $\Pi_{\mathcal{R}}$ *be the orthogonal projection onto* $\mathcal{R}$ *and define*

$$\mathsf{L}_t := \begin{bmatrix} I & \mathsf{T}_{\nu,t} \\ \mathsf{T}_{\mu,t} & I \end{bmatrix}, \qquad \mathsf{L}_t^\circ := \Pi_{\mathcal{R}}\mathsf{L}_t\big|_{\mathcal{D}_0} : \mathcal{D}_0 \to \mathcal{R}.$$

Then $\|\mathsf{L}_t^\circ - \mathsf{L}^\circ\| \to 0$. *Hence, for all sufficiently small* $|t|$, L_t° *is invertible and* $(\mathsf{L}_t^\circ)^{-1} \to (\mathsf{L}^\circ)^{-1}$ *in operator norm.*

Proof. The convergence follows from Lemma 3.20. Since L° is boundedly invertible by Lemma 3.16 and $\|\mathsf{L}_t^\circ - \mathsf{L}^\circ\| \to 0$, for all sufficiently small $|t|$,

$$\|(\mathsf{L}^\circ)^{-1}(\mathsf{L}_t^\circ - \mathsf{L}^\circ)\| < 1.$$

Hence L_t° is invertible by the Neumann-series perturbation argument, and

$$(\mathsf{L}_t^\circ)^{-1} = \left[I + (\mathsf{L}^\circ)^{-1}(\mathsf{L}_t^\circ - \mathsf{L}^\circ)\right]^{-1} (\mathsf{L}^\circ)^{-1}.$$

It follows that

$$(\mathsf{L}_t^\circ)^{-1} \longrightarrow (\mathsf{L}^\circ)^{-1}$$

in operator norm. □

Lemma 3.22 (Compatibility of the secant right-hand side)**.** *For every t,*

$$(J_v\gamma_t^2, J_u\gamma_t^1) \in \mathcal{R},$$

and

$$(J_v\gamma_t^2, J_u\gamma_t^1) \longrightarrow (J_v\gamma^2, J_u\gamma^1) \quad \textit{in } \mathcal{H}^2.$$

Proof. Because $\mu_t = \mu + t\gamma_t^1$ and $\nu_t = \nu + t\gamma_t^2$ are probability measures, while μ and ν are also probability measures,

$$1 = \mu_t(\mathbf{1}) = 1 + t\gamma_t^1(\mathbf{1}), \qquad 1 = \nu_t(\mathbf{1}) = 1 + t\gamma_t^2(\mathbf{1}),$$

and hence

$$\gamma_t^1(\mathbf{1}) = \gamma_t^2(\mathbf{1}) = 0.$$

By Lemma 3.10,

$$\langle J_v\gamma_t^2, v\rangle_{\mathcal{H}} = \gamma_t^2(\mathbf{1}) = 0, \qquad \langle J_u\gamma_t^1, u\rangle_{\mathcal{H}} = \gamma_t^1(\mathbf{1}) = 0.$$

Therefore

$$\langle J_v\gamma_t^2, v\rangle_{\mathcal{H}} = \langle J_u\gamma_t^1, u\rangle_{\mathcal{H}},$$

which is precisely the compatibility condition defining $\mathcal{R}$. Thus

$$(J_v\gamma_t^2, J_u\gamma_t^1) \in \mathcal{R}.$$

Finally, since $\gamma_t^2 \to \gamma^2$ in $\mathcal{E}_v$ and $\gamma_t^1 \to \gamma^1$ in $\mathcal{E}_u$, the isometry in Lemma 3.10 gives

$$\|J_v(\gamma_t^2 - \gamma^2)\|_{\mathcal{H}} = \|\gamma_t^2 - \gamma^2\|_{\mathcal{E}_v} \to 0,$$

and

$$\|J_u(\gamma_t^1 - \gamma^1)\|_{\mathcal{H}} = \|\gamma_t^1 - \gamma^1\|_{\mathcal{E}_u} \to 0.$$

Hence

$$(J_v\gamma_t^2, J_u\gamma_t^1) \longrightarrow (J_v\gamma^2, J_u\gamma^1) \qquad \text{in } \mathcal{H}^2.$$

□

Lemma 3.23 (Convergence of the difference quotients)**.** *Under the assumptions of Lemma 3.18,*

$$(h_{1,t}, h_{2,t}) \longrightarrow (\mathsf{L}^\circ)^{-1}(J_v\gamma^2, J_u\gamma^1) \quad \textit{in } \mathcal{H}^2.$$

Proof. By Lemma 3.19, the difference quotients satisfy

$$(h_{1,t}, h_{2,t}) \in \mathcal{D}_0$$

and the secant system

$$\begin{bmatrix} I & \mathsf{T}_{\nu,t} \\ \mathsf{T}_{\mu,t} & I \end{bmatrix} \begin{bmatrix} h_{1,t} \\ h_{2,t} \end{bmatrix} = \begin{bmatrix} J_v\gamma_t^2 \\ J_u\gamma_t^1 \end{bmatrix}.$$

By Lemma 3.22, the right-hand side belongs to $\mathcal{R}$. Therefore applying the orthogonal projection $\Pi_{\mathcal{R}}$ to both sides leaves the right-hand side unchanged. Since $(h_{1,t}, h_{2,t}) \in \mathcal{D}_0$, the definition of L_t° gives

$$\mathsf{L}_t^\circ(h_{1,t}, h_{2,t}) = (J_v\gamma_t^2, J_u\gamma_t^1).$$

For all sufficiently small $|t|$, Lemma 3.21 shows that L_t° is invertible. Hence

$$(h_{1,t}, h_{2,t}) = (\mathsf{L}_t^\circ)^{-1}(J_v\gamma_t^2, J_u\gamma_t^1).$$

Set

$$r_t := (J_v\gamma_t^2, J_u\gamma_t^1), \qquad r := (J_v\gamma^2, J_u\gamma^1).$$

Then

$$r_t \to r \qquad \text{in } \mathcal{H}^2$$

by Lemma 3.22, while

$$(\mathsf{L}_t^\circ)^{-1} \longrightarrow (\mathsf{L}^\circ)^{-1}$$

in operator norm by Lemma 3.21. Consequently,

$$\begin{aligned}
&\big\|(h_{1,t}, h_{2,t}) - (\mathsf{L}^\circ)^{-1} r\big\|_{\mathcal{H}^2} \\
&\quad \le \big\|(\mathsf{L}_t^\circ)^{-1}(r_t - r)\big\|_{\mathcal{H}^2} + \big\|\big[(\mathsf{L}_t^\circ)^{-1} - (\mathsf{L}^\circ)^{-1}\big] r\big\|_{\mathcal{H}^2} \\
&\quad \le \|(\mathsf{L}_t^\circ)^{-1}\|\, \|r_t - r\|_{\mathcal{H}^2} + \big\|(\mathsf{L}_t^\circ)^{-1} - (\mathsf{L}^\circ)^{-1}\big\|\, \|r\|_{\mathcal{H}^2} \longrightarrow 0.
\end{aligned}$$

Here the inverse norms are uniformly bounded for small $|t|$ because $(\mathsf{L}_t^\circ)^{-1} \to (\mathsf{L}^\circ)^{-1}$ in operator norm. Therefore

$$(h_{1,t}, h_{2,t}) \longrightarrow (\mathsf{L}^\circ)^{-1}(J_v\gamma^2, J_u\gamma^1) \qquad \text{in } \mathcal{H}^2.$$

□

Theorem 3.6 (Hadamard differentiability of the scalings)**.** *The map*

$$\tau : (\mu, \nu) \longmapsto (u_{\mu,\nu}, v_{\mu,\nu})$$

is Hadamard differentiable at (μ, ν), tangentially to $\mathcal{E}_u^0 \times \mathcal{E}_v^0$, as a map into $\mathcal{H}^2$.

More precisely, for $(\gamma^1, \gamma^2) \in \mathcal{E}_u^0 \times \mathcal{E}_v^0$, define

$$\dot\tau_{\mu,\nu}[\gamma^1, \gamma^2] := (\dot u, \dot v) := (\mathsf{L}^\circ)^{-1}\big(J_v\gamma^2, J_u\gamma^1\big).$$

Equivalently, $(\dot u, \dot v)$ is the unique solution of

$$\begin{aligned}
\dot u + \mathsf{T}_\nu \dot v &= J_v\gamma^2, \\
\mathsf{T}_\mu \dot u + \dot v &= J_u\gamma^1, \\
\dot u(x_0) - \dot v(x_0) &= 0.
\end{aligned} \tag{3.15}$$

In particular, if $t \to 0$ and

$$\mu_t = \mu + t\gamma_t^1, \qquad \nu_t = \nu + t\gamma_t^2$$

are probability measures satisfying

$$\gamma_t^1 \to \gamma^1 \quad \text{in } \mathcal{E}_u^0, \qquad \gamma_t^2 \to \gamma^2 \quad \text{in } \mathcal{E}_v^0,$$

then

$$\left\| \frac{\tau(\mu_t, \nu_t) - \tau(\mu, \nu)}{t} - \dot{\tau}_{\mu,\nu}[\gamma^1, \gamma^2] \right\|_{\mathcal{H}^2} \longrightarrow 0.$$

Proof. We first verify that the displayed derivative is well-defined. Since $\gamma^1 \in \mathcal{E}_u^0$ and $\gamma^2 \in \mathcal{E}_v^0$, Lemma 3.10 gives

$$\langle J_u \gamma^1, u \rangle_{\mathcal{H}} = \gamma^1(\mathbf{1}) = 0, \qquad \langle J_v \gamma^2, v \rangle_{\mathcal{H}} = \gamma^2(\mathbf{1}) = 0.$$

Hence

$$\langle J_v \gamma^2, v \rangle_{\mathcal{H}} = \langle J_u \gamma^1, u \rangle_{\mathcal{H}},$$

and therefore

$$(J_v \gamma^2, J_u \gamma^1) \in \mathcal{R},$$

where $\mathcal{R} = \operatorname{ran}(\mathsf{L})$ is the compatibility space in (3.12). By Lemma 3.16,

$$\mathsf{L}^\circ : \mathcal{D}_0 \longrightarrow \mathcal{R}$$

is a bounded bijection with bounded inverse. Thus

$$(\dot{u}, \dot{v}) = (\mathsf{L}^\circ)^{-1}(J_v \gamma^2, J_u \gamma^1)$$

exists uniquely and satisfies the anchoring constraint.

The derivative map is linear because J_u, J_v, and $(\mathsf{L}^\circ)^{-1}$ are linear. It is also bounded. Indeed, using the product norm on $\mathcal{H}^2$ and the isometries from Lemma 3.10,

$$\begin{aligned} \left\| \dot{\tau}_{\mu,\nu}[\gamma^1, \gamma^2] \right\|_{\mathcal{H}^2} &\le \|(\mathsf{L}^\circ)^{-1}\| \left\| (J_v \gamma^2, J_u \gamma^1) \right\|_{\mathcal{H}^2} \\ &= \|(\mathsf{L}^\circ)^{-1}\| \left(\|\gamma^2\|_{\mathcal{E}_v}^2 + \|\gamma^1\|_{\mathcal{E}_u}^2 \right)^{1/2}. \end{aligned}$$

Thus $\dot{\tau}_{\mu,\nu}$ is a continuous linear map on $\mathcal{E}_u^0 \times \mathcal{E}_v^0$.

It remains to establish the Hadamard limit. Consider any admissible path

$$\mu_t = \mu + t\gamma_t^1, \qquad \nu_t = \nu + t\gamma_t^2,$$

with

$$\gamma_t^1 \to \gamma^1 \quad \text{in } \mathcal{E}_u^0, \qquad \gamma_t^2 \to \gamma^2 \quad \text{in } \mathcal{E}_v^0.$$

Define

$$h_{1,t} := \frac{u_{\mu_t,\nu_t} - u}{t}, \qquad h_{2,t} := \frac{v_{\mu_t,\nu_t} - v}{t}.$$

By Lemma 3.23,

$$(h_{1,t}, h_{2,t}) \longrightarrow (\mathsf{L}^\circ)^{-1}(J_v\gamma^2, J_u\gamma^1) \qquad \text{in } \mathcal{H}^2.$$

But, by definition of τ,

$$(h_{1,t}, h_{2,t}) = \frac{\tau(\mu_t,\nu_t) - \tau(\mu,\nu)}{t}.$$

Consequently,

$$\frac{\tau(\mu_t,\nu_t) - \tau(\mu,\nu)}{t} \longrightarrow \dot{\tau}_{\mu,\nu}[\gamma^1,\gamma^2] \qquad \text{in } \mathcal{H}^2.$$

Since this holds for every admissible probability path whose tangent directions converge in $\mathcal{E}_u^0 \times \mathcal{E}_v^0$, and since the limiting map is continuous and linear, this is precisely Hadamard differentiability of τ at (μ,ν) tangentially to $\mathcal{E}_u^0 \times \mathcal{E}_v^0$. □

Corollary 3.3 (Diagonal derivative structure)**.** *At the diagonal base point (P,P), write*

$$q := u_{P,P} = v_{P,P}, \qquad \mathsf{T}_P h := K\left(h\frac{dP}{q^2}\right).$$

Let $\gamma^1, \gamma^2 \in \mathcal{E}_q^0$ and define

$$z_i := J_q\gamma^i, \qquad i = 1,2.$$

Then

$$z_i \in q^\perp, \qquad i = 1,2.$$

Let

$$a := \dot{u}_{P,P}[\gamma^1,\gamma^2], \qquad b := \dot{v}_{P,P}[\gamma^1,\gamma^2].$$

Then (a,b) is the unique anchored solution of

$$\begin{aligned} a + \mathsf{T}_P b &= z_2, \\ \mathsf{T}_P a + b &= z_1, \\ a(x_0) - b(x_0) &= 0. \end{aligned} \tag{3.16}$$

Consequently,

$$(I + \mathsf{T}_P)(a+b) = z_1 + z_2, \tag{3.17}$$

$$(I - \mathsf{T}_P)(a-b) = z_2 - z_1. \tag{3.18}$$

In particular, for a diagonal perturbation (γ, γ),

$$\dot{u}_{P,P}[\gamma, \gamma] = \dot{v}_{P,P}[\gamma, \gamma] = (I + \mathsf{T}_P)^{-1} J_q \gamma.$$

Thus, writing

$$s_i := \dot{u}_{P,P}[\gamma^i, \gamma^i] = \dot{v}_{P,P}[\gamma^i, \gamma^i],$$

we have

$$s_i = (I + \mathsf{T}_P)^{-1} z_i, \qquad i = 1, 2.$$

Proof. Since $\gamma^i \in \mathcal{E}_q^0$, Lemma 3.10 gives

$$\langle z_i, q \rangle_{\mathcal{H}} = \langle J_q \gamma^i, q \rangle_{\mathcal{H}} = \gamma^i(\mathbf{1}) = 0,$$

so $z_i \in q^\perp$.

Specializing (3.15) to $\mu = \nu = P$ and $u = v = q$ gives (3.16), since in this case

$$\mathsf{T}_\mu = \mathsf{T}_\nu = \mathsf{T}_P.$$

Uniqueness is inherited directly from Theorem 3.6.

Adding the first two equations in (3.16) gives

$$(I + \mathsf{T}_P)(a + b) = z_1 + z_2,$$

while subtracting the second from the first gives

$$(I - \mathsf{T}_P)(a - b) = z_2 - z_1.$$

Now take a diagonal perturbation $(\gamma^1, \gamma^2) = (\gamma, \gamma)$ and write

$$z := J_q \gamma.$$

Then (3.18) becomes

$$(I - \mathsf{T}_P)(a - b) = 0.$$

If $\mathsf{T}_P h = h$, then

$$\mathsf{L} \begin{bmatrix} h \\ -h \end{bmatrix} = 0.$$

By Lemma 3.15, specialized to the diagonal,

$$\ker(\mathsf{L}) = \operatorname{span}\{(q, -q)\},$$

and therefore

$$\ker(I - \mathsf{T}_P) = \operatorname{span}\{q\}.$$

Hence

$$a - b = cq$$

for some $c \in \mathbb{R}$. The anchoring condition gives

$$0 = a(x_0) - b(x_0) = cq(x_0).$$

Since $q(x_0) > 0$, $c = 0$, and therefore

$$a = b.$$

Writing their common value as s, either of the first two equations in (3.16) gives

$$(I + \mathsf{T}_P)s = z.$$

Because T_P is positive and self-adjoint,

$$\langle h, (I + \mathsf{T}_P)h\rangle_{\mathcal{H}} = \|h\|_{\mathcal{H}}^2 + \langle h, \mathsf{T}_P h\rangle_{\mathcal{H}} \geq \|h\|_{\mathcal{H}}^2.$$

Hence $I + \mathsf{T}_P$ is boundedly invertible, and

$$s = (I + \mathsf{T}_P)^{-1} z = (I + \mathsf{T}_P)^{-1} J_q \gamma.$$

Taking $\gamma = \gamma^i$ yields

$$s_i = (I + \mathsf{T}_P)^{-1} z_i, \qquad i = 1, 2.$$

□

Lemma 3.24 (Differentiability of the logarithm)**.** *On any uniform neighborhood of a function bounded away from zero, the map $f \mapsto \log f$ from $C(\mathcal{X})$ to itself is continuously Fréchet differentiable with derivative $h \mapsto h/f$.*

Proof. Taylor's theorem and a uniform lower bound give

$$\left\| \frac{\log(f + t h_t) - \log f}{t} - \frac{h}{f} \right\|_\infty \to 0$$

whenever $h_t \to h$ uniformly. □

Corollary 3.4 (Hadamard differentiability of the EOT potentials)**.** *The map*

$$\eta : (\mu, \nu) \longmapsto (\phi_{\mu,\nu}, \psi_{\mu,\nu}) = \big(-\varepsilon \log u_{\mu,\nu}, -\varepsilon \log v_{\mu,\nu}\big)$$

is Hadamard differentiable at (μ, ν), tangentially to $\mathcal{E}_u^0 \times \mathcal{E}_v^0$, as a map into $C(\mathcal{X})^2$. If $(\dot{u}, \dot{v})$ is the solution of (3.15)*, then*

$$\dot{\eta}_{\mu,\nu}[\gamma^1, \gamma^2] = \left(-\varepsilon \frac{\dot{u}}{u}, -\varepsilon \frac{\dot{v}}{v} \right).$$

Proof. Apply the Hadamard chain rule to Theorem 3.6 and Lemma 3.24. □

3.E.2 Sinkhorn divergence

We follow the local-expansion strategy of Gonzalez-Sanz et al., 2022. Write $\phi_\alpha := \phi_{\alpha,\alpha} = \psi_{\alpha,\alpha}$.

Lemma 3.25 (Local EOT expansion)**.** *Let $f, g \in C(\mathcal{X})$ satisfy*

$$\iint \exp\left(\frac{f(x) + g(y) - \|x - y\|^2}{\varepsilon} \right) d\mu(x) d\nu(y) = 1,$$

and put $H(x, y) = f(x) + g(y) - \phi_{\mu,\nu}(x) - \psi_{\mu,\nu}(y)$. Then

$$\left| \mathrm{OT}_\varepsilon(\mu, \nu) - \int f \, d\mu - \int g \, d\nu - \frac{1}{2\varepsilon} \iint H(x, y)^2 d\pi^\varepsilon_{\mu,\nu}(x, y) \right| \\ \leq \frac{1}{6\varepsilon^2} \|H\|_\infty^3 \exp\left(\frac{\|H\|_\infty}{\varepsilon} \right). \tag{3.19}$$

Proof. Set $r = H/\varepsilon$. The normalization condition and the density formula for $\pi^\varepsilon_{\mu,\nu}$ imply $\int (e^r - 1) d\pi^\varepsilon_{\mu,\nu} = 0$. The marginal constraints also give

$$\int f \, d\mu + \int g \, d\nu - \mathrm{OT}_\varepsilon(\mu, \nu) = \varepsilon \int r \, d\pi^\varepsilon_{\mu,\nu}.$$

Therefore the expression in absolute value equals $\varepsilon \int (e^r - 1 - r - r^2/2) d\pi^\varepsilon_{\mu,\nu}$. The pointwise bound $|e^s - 1 - s - s^2/2| \leq |s|^3 e^{|s|}/6$ proves (3.19). □

For probability measures α, β, define

$$S^{\mathrm{loc}}(\alpha,\beta) := \frac{1}{2}\int(\phi_\alpha - \phi_\beta)d(\beta-\alpha) + \frac{1}{4\varepsilon}\iint\left[(\phi_{\alpha,\beta}-\phi_\beta)(x)+(\psi_{\alpha,\beta}-\phi_\beta)(y)\right]^2 d\pi^\varepsilon_{\alpha,\beta}(x,y) + \frac{1}{4\varepsilon}\iint\left[(\phi_{\alpha,\beta}-\phi_\alpha)(x)+(\psi_{\alpha,\beta}-\phi_\alpha)(y)\right]^2 d\pi^\varepsilon_{\alpha,\beta}(x,y). \tag{3.20}$$

Lemma 3.26 (Second-order accuracy of the local expansion)**.** *Let $\alpha_t = P + t\gamma_t^1$ and $\beta_t = P + t\gamma_t^2$ be probability measures with $\gamma_t^i \to \gamma^i$ in $\mathcal{E}_q^0$, where $q = u_{P,P}$. Then*

$$S(\alpha_t,\beta_t) - S^{\mathrm{loc}}(\alpha_t,\beta_t) = o(t^2).$$

Proof. Write

$$\phi_{\alpha_t} := \phi_{\alpha_t,\alpha_t} = \psi_{\alpha_t,\alpha_t}, \qquad \phi_{\beta_t} := \phi_{\beta_t,\beta_t} = \psi_{\beta_t,\beta_t}.$$

By Lemma 3.7, both pairs $(\phi_{\alpha_t},\phi_{\alpha_t})$ and $(\phi_{\beta_t},\phi_{\beta_t})$ satisfy the normalization condition in Lemma 3.25 when integrated against $\alpha_t \otimes \beta_t$.

We first apply Lemma 3.25 to $\mathrm{OT}_\varepsilon(\alpha_t,\beta_t)$ using the β_t-self potentials. Define

$$H_{\beta,t}(x,y) := \phi_{\beta_t}(x) + \phi_{\beta_t}(y) - \phi_{\alpha_t,\beta_t}(x) - \psi_{\alpha_t,\beta_t}(y).$$

Then

$$\mathrm{OT}_\varepsilon(\alpha_t,\beta_t) = \int \phi_{\beta_t}\,d\alpha_t + \int \phi_{\beta_t}\,d\beta_t + \frac{1}{2\varepsilon}\iint H_{\beta,t}(x,y)^2\,d\pi^\varepsilon_{\alpha_t,\beta_t}(x,y) + R_{\beta,t}, \tag{3.21}$$

where

$$|R_{\beta,t}| \le \frac{1}{6\varepsilon^2}\|H_{\beta,t}\|_\infty^3 \exp\left(\frac{\|H_{\beta,t}\|_\infty}{\varepsilon}\right).$$

Likewise, applying Lemma 3.25 using the α_t-self potentials and defining

$$H_{\alpha,t}(x,y) := \phi_{\alpha_t}(x) + \phi_{\alpha_t}(y) - \phi_{\alpha_t,\beta_t}(x) - \psi_{\alpha_t,\beta_t}(y),$$

we obtain

$$\mathrm{OT}_\varepsilon(\alpha_t,\beta_t) = \int \phi_{\alpha_t}\,d\alpha_t + \int \phi_{\alpha_t}\,d\beta_t + \frac{1}{2\varepsilon}\iint H_{\alpha,t}(x,y)^2\,d\pi^\varepsilon_{\alpha_t,\beta_t}(x,y) + R_{\alpha,t}. \tag{3.22}$$

Averaging (3.21) and (3.22) gives

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha_t,\beta_t) = \frac{1}{2}\left[\int \phi_{\beta_t}\,d\alpha_t + \int \phi_{\beta_t}\,d\beta_t + \int \phi_{\alpha_t}\,d\alpha_t + \int \phi_{\alpha_t}\,d\beta_t\right] \\ + \frac{1}{4\varepsilon}\iint H_{\beta,t}^2\,d\pi^\varepsilon_{\alpha_t,\beta_t} + \frac{1}{4\varepsilon}\iint H_{\alpha,t}^2\,d\pi^\varepsilon_{\alpha_t,\beta_t} \\ + \frac{1}{2}(R_{\alpha,t}+R_{\beta,t}).\end{aligned}$$

On the diagonal, Proposition 3.1 gives

$$\mathrm{OT}_\varepsilon(\alpha_t,\alpha_t) = 2\int \phi_{\alpha_t}\,d\alpha_t, \qquad \mathrm{OT}_\varepsilon(\beta_t,\beta_t) = 2\int \phi_{\beta_t}\,d\beta_t.$$

Therefore, after subtracting the two self-costs in the definition of the Sinkhorn divergence,

$$\begin{aligned}S(\alpha_t,\beta_t) &= \frac{1}{2}\left[\int \phi_{\beta_t}\,d\alpha_t - \int \phi_{\beta_t}\,d\beta_t + \int \phi_{\alpha_t}\,d\beta_t - \int \phi_{\alpha_t}\,d\alpha_t\right] \\ &\quad + \frac{1}{4\varepsilon}\iint H_{\beta,t}^2\,d\pi^\varepsilon_{\alpha_t,\beta_t} + \frac{1}{4\varepsilon}\iint H_{\alpha,t}^2\,d\pi^\varepsilon_{\alpha_t,\beta_t} \\ &\quad + \frac{1}{2}(R_{\alpha,t}+R_{\beta,t}) \\ &= \frac{1}{2}\int (\phi_{\alpha_t}-\phi_{\beta_t})\,d(\beta_t-\alpha_t) \\ &\quad + \frac{1}{4\varepsilon}\iint \left[(\phi_{\alpha_t,\beta_t}-\phi_{\beta_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\beta_t})(y)\right]^2 d\pi^\varepsilon_{\alpha_t,\beta_t} \\ &\quad + \frac{1}{4\varepsilon}\iint \left[(\phi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(y)\right]^2 d\pi^\varepsilon_{\alpha_t,\beta_t} \\ &\quad + \frac{1}{2}(R_{\alpha,t}+R_{\beta,t}).\end{aligned}$$

The first three terms are exactly $S^{\mathrm{loc}}(\alpha_t,\beta_t)$ as defined in (3.20). Hence

$$S(\alpha_t,\beta_t) - S^{\mathrm{loc}}(\alpha_t,\beta_t) = \frac{1}{2}(R_{\alpha,t}+R_{\beta,t}).$$

It remains only to control these two remainders. By Corollary 3.4, Hadamard differentiability of the potentials along the paths (α_t,β_t), (α_t,α_t), and (β_t,β_t) implies

$$\begin{aligned}\|\phi_{\alpha_t,\beta_t}-\phi_{P,P}\|_\infty = O(|t|), &\qquad \|\psi_{\alpha_t,\beta_t}-\psi_{P,P}\|_\infty = O(|t|), \\ \|\phi_{\alpha_t}-\phi_{P,P}\|_\infty = O(|t|), &\qquad \|\phi_{\beta_t}-\phi_{P,P}\|_\infty = O(|t|).\end{aligned}$$

Consequently,

$$\begin{aligned}\|H_{\beta,t}\|_\infty \le \|\phi_{\beta_t}-\phi_{\alpha_t,\beta_t}\|_\infty + \|\phi_{\beta_t}-\psi_{\alpha_t,\beta_t}\|_\infty = O(|t|), \\ \|H_{\alpha,t}\|_\infty \le \|\phi_{\alpha_t}-\phi_{\alpha_t,\beta_t}\|_\infty + \|\phi_{\alpha_t}-\psi_{\alpha_t,\beta_t}\|_\infty = O(|t|).\end{aligned}$$

The remainder bound in Lemma 3.25 therefore gives

$$|R_{\alpha,t}| + |R_{\beta,t}| = O(|t|^3) = o(t^2).$$

Thus

$$S(\alpha_t, \beta_t) - S^{\mathrm{loc}}(\alpha_t, \beta_t) = o(t^2),$$

as claimed. □

Lemma 3.27 (Diagonal spectral gap)**.** *At the diagonal base point (P, P), write*

$$q := u_{P,P} = v_{P,P}, \qquad \mathsf{T}_P h := K\left(h\frac{dP}{q^2}\right).$$

Then T_P is compact, positive, and self-adjoint on $\mathcal{H}$, and

$$\mathsf{T}_P q = q, \qquad \ker(I - \mathsf{T}_P) = \operatorname{span}\{q\}.$$

Moreover, every eigenvalue of T_P belongs to $[0, 1]$. Consequently, $q^\perp$ is invariant under T_P and

$$\rho := \left\| \mathsf{T}_P\big|_{q^\perp} \right\| < 1.$$

In particular,

$$\mathsf{R}_P := (I - \mathsf{T}_P^2)^{-1}\big|_{q^\perp} : q^\perp \to q^\perp$$

is a bounded positive self-adjoint operator, with

$$\|\mathsf{R}_P\| \le \frac{1}{1 - \rho^2}.$$

Proof. Compactness, positivity, and self-adjointness follow from Lemma 3.15 specialized to $\mu = \nu = P$ and $u = v = q$. In particular,

$$\langle f, \mathsf{T}_P g\rangle_{\mathcal{H}} = \int \frac{fg}{q^2}\, dP,$$

so

$$\langle f, \mathsf{T}_P f\rangle_{\mathcal{H}} = \int \frac{f^2}{q^2}\, dP \ge 0.$$

The diagonal scaling equation is

$$q = K\left(\frac{dP}{q}\right),$$

and therefore

$$\mathsf{T}_P q = K\left(q\frac{dP}{q^2}\right) = K\left(\frac{dP}{q}\right) = q.$$

Thus q is an eigenfunction with eigenvalue 1.

We next verify that this eigenvalue is simple. If $\mathsf{T}_P h = h$, then

$$\mathsf{L}(h,-h) = \begin{bmatrix} I & \mathsf{T}_P \\ \mathsf{T}_P & I \end{bmatrix}\begin{bmatrix} h \\ -h \end{bmatrix} = \begin{bmatrix} h - \mathsf{T}_P h \\ \mathsf{T}_P h - h \end{bmatrix} = 0.$$

By Lemma 3.15,

$$\ker(\mathsf{L}) = \operatorname{span}\{(q,-q)\}.$$

Hence $h = cq$ for some $c \in \mathbb{R}$, and therefore

$$\ker(I - \mathsf{T}_P) = \operatorname{span}\{q\}.$$

It remains to locate the rest of the spectrum. Recall the normalized transport operator $\mathcal{A}$ defined by

$$(\mathcal{A}f)(y) = \int \frac{k_\varepsilon(x,y)}{q(x)q(y)} f(x)\, dP(x).$$

At the diagonal, by definition

$$\mathcal{A}f = \frac{1}{q}\mathsf{T}_P(qf).$$

Suppose

$$\mathsf{T}_P h = \lambda h, \qquad h \neq 0, \qquad f := \frac{h}{q}.$$

Then

$$\mathcal{A}f = \frac{1}{q}\mathsf{T}_P(qf) = \frac{1}{q}\mathsf{T}_P h = \lambda f.$$

Since $\mathcal{A}$ is an $L^2(P)$ contraction,

$$|\lambda|\,\|f\|_{L^2(P)} = \|\mathcal{A}f\|_{L^2(P)} \leq \|f\|_{L^2(P)}.$$

As $f \neq 0$, this yields

$$|\lambda| \leq 1.$$

On the other hand, positivity of T_P implies

$$\lambda\|h\|_{\mathcal{H}}^2 = \langle h, \mathsf{T}_P h\rangle_{\mathcal{H}} \geq 0,$$

and hence $\lambda \geq 0$. Thus every eigenvalue lies in $[0, 1]$.

Since T_P is self-adjoint and $\mathsf{T}_P q = q$, the subspace $q^\perp$ is invariant: for $h \in q^\perp$,

$$\langle \mathsf{T}_P h, q\rangle_{\mathcal{H}} = \langle h, \mathsf{T}_P q\rangle_{\mathcal{H}} = \langle h, q\rangle_{\mathcal{H}} = 0.$$

The restriction $\mathsf{T}_P|_{q^\perp}$ is again compact, positive, and self-adjoint. If its norm were equal to 1, compact self-adjointness would imply the existence of an eigenvector in $q^\perp$ with eigenvalue 1, contradicting

$$\ker(I - \mathsf{T}_P) = \operatorname{span}\{q\}.$$

Therefore

$$\rho = \left\| \mathsf{T}_P\big|_{q^\perp} \right\| < 1.$$

Hence the Neumann series converges in operator norm on $q^\perp$:

$$\mathsf{R}_P = (I - \mathsf{T}_P^2)^{-1} = \sum_{j=0}^{\infty} \mathsf{T}_P^{2j}.$$

Each term is positive and self-adjoint, so R_P is positive and self-adjoint, and

$$\|\mathsf{R}_P\| \leq \sum_{j=0}^{\infty} \rho^{2j} = \frac{1}{1-\rho^2}.$$

□

Lemma 3.28 (Limit for logarithmic difference quotients)**.** *Let a_t, b_t be strictly positive functions such that*

$$\frac{a_t - q}{t} \to a, \qquad \frac{b_t - q}{t} \to b \quad \text{in } \mathcal{H}.$$

Let

$$\alpha_t = P + t\gamma_t^1, \qquad \beta_t = P + t\gamma_t^2$$

be probability measures such that

$$\gamma_t^1 \to \gamma^1, \qquad \gamma_t^2 \to \gamma^2 \quad \text{in } \mathcal{E}_q^0.$$

Then

$$\int \frac{\log a_t - \log b_t}{t}\, d(\gamma_t^2 - \gamma_t^1) \longrightarrow \left\langle a - b, J_q(\gamma^2 - \gamma^1)\right\rangle_{\mathcal{H}}.$$

Proof. Set

$$d_t := \frac{a_t - b_t}{t}, \qquad \delta_t := \gamma_t^2 - \gamma_t^1, \qquad \delta := \gamma^2 - \gamma^1.$$

By the assumed first-order limits,

$$d_t \to a - b \quad \text{in } \mathcal{H}, \qquad J_q \delta_t \to J_q \delta \quad \text{in } \mathcal{H}.$$

By the fundamental theorem of calculus,

$$\frac{\log a_t - \log b_t}{t} = c_t d_t, \qquad c_t(x) = \int_0^1 \frac{ds}{b_t(x) + s(a_t(x) - b_t(x))}.$$

Since $a_t = q + O(t)$ and $b_t = q + O(t)$ uniformly, with q bounded away from zero,

$$c_t = \frac{1}{q} + t r_t,$$

where r_t converges uniformly. Hence

$$\begin{aligned} \int \frac{\log a_t - \log b_t}{t}\, d\delta_t &= \int \frac{d_t}{q}\, d\delta_t + t \int r_t d_t\, d\delta_t \\ &= \langle d_t, J_q \delta_t \rangle_{\mathcal{H}} + \int r_t d_t\, d(\beta_t - \alpha_t). \end{aligned}$$

The first term converges to

$$\langle a - b, J_q \delta \rangle_{\mathcal{H}}$$

by the two $\mathcal{H}$-limits above. For the second term, $r_t d_t$ converges uniformly, while $\alpha_t \rightsquigarrow P$ and $\beta_t \rightsquigarrow P$; hence

$$\int r_t d_t\, d(\beta_t - \alpha_t) \to 0.$$

Therefore

$$\int \frac{\log a_t - \log b_t}{t}\, d(\gamma_t^2 - \gamma_t^1) \longrightarrow \left\langle a - b, J_q(\gamma^2 - \gamma^1) \right\rangle_{\mathcal{H}}.$$

□

Theorem 3.7 (Second-order Hadamard differentiability)**.** *Let*

$$\alpha_t = P + t\gamma_t^1, \qquad \beta_t = P + t\gamma_t^2$$

be probability measures with

$$\gamma_t^i \to \gamma^i \quad \textit{in } \mathcal{E}_q^0, \qquad i = 1, 2.$$

Set

$$\delta := \gamma^2 - \gamma^1, \qquad z := J_q\delta \in q^\perp.$$

Then

$$\frac{S(\alpha_t,\beta_t)}{t^2} \longrightarrow \frac{\varepsilon}{2}\,\langle z, \mathsf{R}_P z\rangle_{\mathcal{H}}\,, \qquad \mathsf{R}_P := (I-\mathsf{T}_P^2)^{-1}\big|_{q^\perp}\,. \tag{3.23}$$

Consequently, S is second-order Hadamard differentiable at (P,P), tangentially to $\mathcal{E}_q^0\times\mathcal{E}_q^0$, with zero first derivative and continuous quadratic form

$$\ddot{S}_{(P,P)}(\gamma^1,\gamma^2) = \frac{\varepsilon}{2}\,\big\langle J_q(\gamma^2-\gamma^1), \mathsf{R}_P J_q(\gamma^2-\gamma^1)\big\rangle_{\mathcal{H}}\,.$$

Proof. By Lemma 3.26,

$$S(\alpha_t,\beta_t) = S^{\mathrm{loc}}(\alpha_t,\beta_t) + o(t^2),$$

so it suffices to evaluate $S^{\mathrm{loc}}(\alpha_t,\beta_t)/t^2$.

Write

$$z_i := J_q\gamma^i, \qquad z = z_2 - z_1.$$

By Corollary 3.3, define

$$(a,b) := \dot{\tau}_{P,P}[\gamma^1,\gamma^2]$$

and, for $i=1,2$,

$$(s_i,s_i) := \dot{\tau}_{P,P}[\gamma^i,\gamma^i].$$

Then

$$\begin{aligned} a+\mathsf{T}_P b = z_2, \quad \mathsf{T}_P a + b = z_1,\\ (I+\mathsf{T}_P)s_i = z_i, \qquad\qquad i=1,2. \end{aligned} \tag{3.24}$$

In particular,

$$s_2 - s_1 = (I+\mathsf{T}_P)^{-1}z.$$

We evaluate the three terms in (3.20) separately.

Term 1. By Hadamard differentiability of the scalings,

$$\frac{u_{\alpha_t,\alpha_t}-q}{t} \to s_1, \qquad \frac{u_{\beta_t,\beta_t}-q}{t} \to s_2 \quad \text{in } \mathcal{H}.$$

Since

$$\phi_{\alpha_t} = -\varepsilon \log u_{\alpha_t,\alpha_t}, \qquad \phi_{\beta_t} = -\varepsilon \log u_{\beta_t,\beta_t},$$

and

$$\beta_t - \alpha_t = t(\gamma_t^2 - \gamma_t^1),$$

Lemma 3.28 gives

$$\begin{aligned}
&\frac{1}{2t^2}\int (\phi_{\alpha_t} - \phi_{\beta_t})\, d(\beta_t - \alpha_t) \\
&\quad = -\frac{\varepsilon}{2}\int \frac{\log u_{\alpha_t,\alpha_t} - \log u_{\beta_t,\beta_t}}{t}\, d(\gamma_t^2 - \gamma_t^1) \\
&\quad \longrightarrow -\frac{\varepsilon}{2}\langle s_1 - s_2, z\rangle_{\mathcal{H}} \\
&\quad = \frac{\varepsilon}{2}\left\langle z, (I + \mathsf{T}_P)^{-1} z\right\rangle_{\mathcal{H}}.
\end{aligned}$$

Term 2. Set

$$r := \mathsf{R}_P z, \qquad f := \frac{r}{q}.$$

By Lemma 3.27, $r \in q^\perp$ is the unique element of $q^\perp$ satisfying

$$(I - \mathsf{T}_P^2) r = z.$$

Subtracting the derivative system corresponding to (γ^2, γ^2) from that corresponding to (γ^1, γ^2) in (3.24) gives

$$(a - s_2) + \mathsf{T}_P(b - s_2) = 0, \qquad \mathsf{T}_P(a - s_2) + (b - s_2) = -z.$$

Hence

$$(I - \mathsf{T}_P^2)(b - s_2) = -z.$$

Since

$$\ker(I - \mathsf{T}_P^2) = \operatorname{span}\{q\},$$

there exists $c_2 \in \mathbb{R}$ such that

$$a - s_2 = \mathsf{T}_P r - c_2 q, \qquad b - s_2 = -r + c_2 q. \tag{3.25}$$

Corollary 3.4 and (3.25) therefore give, uniformly on $\mathcal{X}^2$,

$$\begin{aligned}\frac{1}{t}\big[(\phi_{\alpha_t,\beta_t}-\phi_{\beta_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\beta_t})(y)\big]\\ \longrightarrow -\varepsilon\left[\frac{a-s_2}{q}(x)+\frac{b-s_2}{q}(y)\right]\\ =\varepsilon\left[f(y)-\frac{\mathsf{T}_P r}{q}(x)\right].\end{aligned}$$

The constants c_2 cancel. By the identity above,

$$\frac{\mathsf{T}_P r}{q}=\mathcal{A}f,$$

where

$$(\mathcal{A}f)(x)=\int \frac{k_\varepsilon(x,y)}{q(x)q(y)}f(y)\,dP(y).$$

Thus the preceding limit is

$$\varepsilon\big[f(y)-\mathcal{A}f(x)\big].$$

Term 3. Similarly, subtracting the derivative system corresponding to (γ^1,γ^1) from that corresponding to (γ^1,γ^2) gives

$$(a-s_1)+\mathsf{T}_P(b-s_1)=z,\qquad \mathsf{T}_P(a-s_1)+(b-s_1)=0.$$

Consequently, for some $c_1\in\mathbb{R}$,

$$a-s_1=r-c_1q,\qquad b-s_1=-\mathsf{T}_P r+c_1q.$$

Hence, again uniformly on $\mathcal{X}^2$,

$$\frac{1}{t}\big[(\phi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(y)\big]\longrightarrow\varepsilon\big[\mathcal{A}f(y)-f(x)\big].$$

It remains to integrate these two uniform limits. Let$\pi_t:=\pi^{\varepsilon}_{\alpha_t,\beta_t}$. Since $\alpha_t\rightsquigarrow P$ and $\beta_t\rightsquigarrow P$, and since $u_{\alpha_t,\beta_t}\to q$ and $v_{\alpha_t,\beta_t}\to q$ uniformly, the densities

$$\frac{k_\varepsilon(x,y)}{u_{\alpha_t,\beta_t}(x)v_{\alpha_t,\beta_t}(y)}$$

converge uniformly to

$$\frac{k_\varepsilon(x,y)}{q(x)q(y)}.$$

Therefore $\pi_t \rightsquigarrow \pi^{\varepsilon}_{P,P}$. Together with the uniform limits above, this yields

$$\frac{1}{t^2}\iint \left[(\phi_{\alpha_t,\beta_t}-\phi_{\beta_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\beta_t})(y)\right]^2 d\pi_t \\ \longrightarrow \varepsilon^2 \iint \left[f(y)-\mathcal{A}f(x)\right]^2 d\pi^{\varepsilon}_{P,P}(x,y), \tag{3.26}$$

and

$$\frac{1}{t^2}\iint \left[(\phi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(x)+(\psi_{\alpha_t,\beta_t}-\phi_{\alpha_t})(y)\right]^2 d\pi_t \\ \longrightarrow \varepsilon^2 \iint \left[\mathcal{A}f(y)-f(x)\right]^2 d\pi^{\varepsilon}_{P,P}(x,y). \tag{3.27}$$

The limiting plan is symmetric, so the two integrals on the right-hand sides are equal. To evaluate the first one, write

$$d\pi^{\varepsilon}_{P,P}(x,y) = \frac{k_\varepsilon(x,y)}{q(x)q(y)}\,dP(x)dP(y).$$

Using the marginal constraint and the definition of $\mathcal{A}$,

$$\iint \left[f(y)-\mathcal{A}f(x)\right]^2 d\pi^{\varepsilon}_{P,P}(x,y) \\ = \|f\|^2_{L^2(P)} - \|\mathcal{A}f\|^2_{L^2(P)}.$$

Since $r = qf$ and $\mathsf{T}_P r = q\mathcal{A}f$,

$$\|f\|^2_{L^2(P)} = \langle r, \mathsf{T}_P r\rangle_{\mathcal{H}}, \qquad \|\mathcal{A}f\|^2_{L^2(P)} = \langle r, \mathsf{T}_P^3 r\rangle_{\mathcal{H}}.$$

Hence

$$\begin{aligned}\|f\|^2_{L^2(P)} - \|\mathcal{A}f\|^2_{L^2(P)} &= \left\langle r, \mathsf{T}_P(I-\mathsf{T}_P^2)r\right\rangle_{\mathcal{H}} \\ &= \langle r, \mathsf{T}_P z\rangle_{\mathcal{H}} \\ &= \langle z, \mathsf{T}_P \mathsf{R}_P z\rangle_{\mathcal{H}}\,.\end{aligned}$$

Each quadratic term in (3.20) has coefficient $1/(4\varepsilon)$. Therefore, by (3.26)–(3.27), their combined contribution is

$$\frac{\varepsilon}{2}\,\langle z, \mathsf{T}_P \mathsf{R}_P z\rangle_{\mathcal{H}}\,.$$

Together with the first term this gives

$$\frac{S^{\mathrm{loc}}(\alpha_t, \beta_t)}{t^2} \longrightarrow \frac{\varepsilon}{2} \left\langle z, (I + \mathsf{T}_P)^{-1} z \right\rangle_{\mathcal{H}} + \frac{\varepsilon}{2} \left\langle z, \mathsf{T}_P \mathsf{R}_P z \right\rangle_{\mathcal{H}}.$$

On $q^{\perp}$,

$$(I + \mathsf{T}_P)^{-1} = (I - \mathsf{T}_P)(I - \mathsf{T}_P^2)^{-1} = (I - \mathsf{T}_P)\mathsf{R}_P.$$

Thus

$$(I + \mathsf{T}_P)^{-1} + \mathsf{T}_P \mathsf{R}_P = \mathsf{R}_P,$$

and consequently

$$\frac{S^{\mathrm{loc}}(\alpha_t, \beta_t)}{t^2} \longrightarrow \frac{\varepsilon}{2} \langle z, \mathsf{R}_P z \rangle_{\mathcal{H}}.$$

Lemma 3.26 now yields (3.23).

Finally,

$$(\gamma^1, \gamma^2) \longmapsto \frac{\varepsilon}{2} \left\langle J_q(\gamma^2 - \gamma^1), \mathsf{R}_P J_q(\gamma^2 - \gamma^1) \right\rangle_{\mathcal{H}}$$

is a continuous quadratic form because J_q is an isometry and R_P is bounded. The expansion above is $O(t^2)$, so the first derivative of S at (P, P) is zero. This proves second-order Hadamard differentiability. □

3.E.3 Sinkhorn Kernel Estimation

Lemma 3.29 (Hilbert CLT and empirical scaling)**.** *The random element*

$$\sqrt{n}\, J_q(P_n - P) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{k_\varepsilon(\cdot, X_i)}{q(X_i)} - \mathbb{E} \frac{k_\varepsilon(\cdot, X)}{q(X)} \right)$$

converges weakly in the separable Hilbert space $\mathcal{H}$ to a centered Gaussian element. Consequently, $\sqrt{n}(q_n - q)$ is tight in $\mathcal{H}$.

Proof. The summands are centered, iid, and uniformly bounded in $\mathcal{H}$ because q is bounded away from zero. The Hilbert-space central limit theorem applies. The second assertion follows from the functional delta method and Theorem 3.6; at the diagonal the derivative is $(I + \mathsf{T}_P)^{-1}$ applied to the Gaussian limit. □

Lemma 3.30 (Uniform low-rank approximation of the scaled Gaussian kernel)**.** *Let $Q \in \mathcal{P}(\mathcal{X})$ and let $q_Q := u_{Q,Q} = v_{Q,Q}$. Suppose*

$$\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\| \le R\},$$

and put

$$\kappa := \min_{x,y \in \mathcal{X}} k_\varepsilon(x,y) > 0.$$

For

$$\widetilde{k}_Q(x,y) := \frac{k_\varepsilon(x,y)}{q_Q(x) q_Q(y)},$$

and every integer $s \ge 0$, there exists a positive-semidefinite kernel $\widetilde{k}_Q^{(s)}$ of rank at most

$$r_s := \binom{s+d}{d}$$

such that $\widetilde{k}_Q - \widetilde{k}_Q^{(s)}$ is positive semidefinite and

$$\sup_{x \in \mathcal{X}} \left[\widetilde{k}_Q(x,x) - \widetilde{k}_Q^{(s)}(x,x) \right] \le \kappa^{-4} \sum_{\ell=s+1}^{\infty} \frac{(2R^2/\varepsilon)^\ell}{\ell!}.$$

Consequently, for any sample $x_1, \dots, x_n \in \mathcal{X}$, let

$$\widetilde{\mathbf{K}}_{Q,n} := \frac{1}{n} \big[\widetilde{k}_Q(x_i, x_j) \big]_{i,j=1}^n.$$

Then

$$\operatorname{Tail}_{r_s}(\widetilde{\mathbf{K}}_{Q,n}) \le \kappa^{-4} \sum_{\ell=s+1}^{\infty} \frac{(2R^2/\varepsilon)^\ell}{\ell!}, \tag{3.28}$$

uniformly in Q, n, and the sample locations.

Proof. By Lemma 3.17,

$$q_Q(x) \ge \kappa^2 \qquad \text{uniformly in } Q \text{ and } x,$$

and therefore

$$|b_Q(x)| := \frac{e^{-\|x\|^2/\varepsilon}}{q_Q(x)} \le \kappa^{-2}.$$

Since

$$k_\varepsilon(x,y) = e^{-\|x\|^2/\varepsilon} e^{-\|y\|^2/\varepsilon} e^{2x^\top y/\varepsilon},$$

we have

$$\widetilde{k}_Q(x,y) = b_Q(x)b_Q(y)\sum_{\alpha\in\mathbb{N}^d}\frac{(2/\varepsilon)^{|\alpha|}}{\alpha!}x^\alpha y^\alpha.$$

Define

$$\widetilde{k}_Q^{(s)}(x,y) := b_Q(x)b_Q(y)\sum_{|\alpha|\le s}\frac{(2/\varepsilon)^{|\alpha|}}{\alpha!}x^\alpha y^\alpha.$$

This is a positive-semidefinite feature kernel of rank at most

$$\#\{\alpha\in\mathbb{N}^d : |\alpha|\le s\} = \binom{s+d}{d} = r_s,$$

and the omitted feature sum $\widetilde{k}_Q - \widetilde{k}_Q^{(s)}$ is also positive semidefinite.

On the diagonal,

$$\begin{aligned}\widetilde{k}_Q(x,x) - \widetilde{k}_Q^{(s)}(x,x) &= b_Q(x)^2\sum_{\ell=s+1}^{\infty}\left(\frac{2}{\varepsilon}\right)^\ell\sum_{|\alpha|=\ell}\frac{x^{2\alpha}}{\alpha!}\\ &= b_Q(x)^2\sum_{\ell=s+1}^{\infty}\frac{(2\|x\|^2/\varepsilon)^\ell}{\ell!}\\ &\le \kappa^{-4}\sum_{\ell=s+1}^{\infty}\frac{(2R^2/\varepsilon)^\ell}{\ell!},\end{aligned}$$

where the second equality follows from the multinomial theorem.

Now fix sample locations $x_1,\dots,x_n$, and let $\widetilde{\mathbf{K}}_{Q,n}$ and $\widetilde{\mathbf{K}}_{Q,n}^{(s)}$ be the normalized Gram matrices of $\widetilde{k}_Q$ and $\widetilde{k}_Q^{(s)}$, respectively. Put

$$E_{Q,n}^{(s)} := \widetilde{\mathbf{K}}_{Q,n} - \widetilde{\mathbf{K}}_{Q,n}^{(s)} \succeq 0.$$

Since $\operatorname{rank}(\widetilde{\mathbf{K}}_{Q,n}^{(s)}) \le r_s$, the variational characterization of the eigenvalue tail gives

$$\operatorname{Tail}_{r_s}(\widetilde{\mathbf{K}}_{Q,n}) \le \operatorname{tr}\big(E_{Q,n}^{(s)}\big).$$

Indeed, one may take the projection onto the range of $\widetilde{\mathbf{K}}_{Q,n}^{(s)}$ in the corresponding Ky Fan variational formula. Finally,

$$\begin{aligned}\operatorname{tr}\big(E_{Q,n}^{(s)}\big) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\widetilde{k}_Q(x_i,x_i) - \widetilde{k}_Q^{(s)}(x_i,x_i)\Big]\\ &\le \kappa^{-4}\sum_{\ell=s+1}^{\infty}\frac{(2R^2/\varepsilon)^\ell}{\ell!}.\end{aligned}$$

This proves (3.28). The bound is uniform because κ and R do not depend on Q, n, or the sample locations. □

Lemma 3.31 (Polylogarithmic empirical scaled compression)**.** *If $m = \omega(\log^d n)$ and Algorithm 2 is applied conditionally to the empirical scaled kernel $\widetilde{k}_{q_n}$, then*

$$\|J_{q_n}(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}).$$

Proof. Recall that

$$r_s = \binom{s+d}{d} \asymp s^d.$$

Since $m/\log^d n \to \infty$, there exists a sequence s_n such that

$$\frac{s_n}{\log n} \longrightarrow \infty, \qquad r_{s_n} \le m - 1$$

for all sufficiently large n. Put $\Lambda := 2R^2/\varepsilon$. For s sufficiently large,

$$\sum_{\ell > s} \frac{\Lambda^\ell}{\ell!} \le \frac{\Lambda^{s+1}}{(s+1)!} \frac{1}{1 - \Lambda/(s+2)} \le 2 \left(\frac{e\Lambda}{s+1} \right)^{s+1}.$$

Since $s_n/\log n \to \infty$, the right-hand side is $o(n^{-1})$. Lemma 3.30, applied with $Q = P_n$, gives

$$\mathrm{Tail}_{r_{s_n}}(\widetilde{\mathbf{K}}_{q_n,n}) = o(n^{-1})$$

uniformly over the realized sample. Because $r_{s_n} \le m - 1$ and $r \mapsto \mathrm{Tail}_r$ is nonincreasing,

$$\mathrm{Tail}_{m-1}(\widetilde{\mathbf{K}}_{q_n,n}) = o(n^{-1}).$$

Conditionally on $X_1, \ldots, X_n$, Lemma 3.2, with $r = m - 1$, now gives

$$\mathbb{E}\left[\mathrm{MMD}^2_{\widetilde{k}_{q_n}}(P_m, P_n) \,\middle|\, X_1, \ldots, X_n \right] = o(n^{-1}),$$

where the expectation is over the internal randomness of the compression routine. Hence, for every $\eta > 0$, conditional Markov inequality gives

$$\begin{aligned} &P\left(\sqrt{n}\, \mathrm{MMD}_{\widetilde{k}_{q_n}}(P_m, P_n) > \eta \,\middle|\, X_1, \ldots, X_n \right) \\ &\quad \le \frac{n}{\eta^2} \mathbb{E}\left[\mathrm{MMD}^2_{\widetilde{k}_{q_n}}(P_m, P_n) \,\middle|\, X_1, \ldots, X_n \right] \longrightarrow 0. \end{aligned}$$

The conditional probability is bounded uniformly over the realized sample by a deterministic $o(1)$ sequence; taking expectations gives the same unconditional conclusion. Finally, Lemma 3.10 gives

$$\mathrm{MMD}_{\widetilde{k}_{q_n}}(P_m, P_n) = \|J_{q_n}(P_m - P_n)\|_{\mathcal{H}},$$

which proves the result. □

Proof of Lemma 3.4. Put

$$\Delta_n := P_m - P_n, \qquad t_n := n^{-1/2}.$$

By assumption, $\|J_{q_n}\Delta_n\|_{\mathcal{H}} = o_p(t_n)$. Since $\|\Delta_n\|_{\mathrm{tv}} \le 2$,

$$\begin{aligned}\|J_q\Delta_n\|_{\mathcal{H}} &\le \|J_{q_n}\Delta_n\|_{\mathcal{H}} + \left\| K\left[\left(\frac{1}{q} - \frac{1}{q_n}\right) d\Delta_n\right]\right\|_{\mathcal{H}} \\ &\le o_p(t_n) + 2\|q^{-1} - q_n^{-1}\|_\infty = O_p(t_n),\end{aligned}$$

where the last step follows from Lemma 3.29, the embedding $\mathcal{H} \hookrightarrow C(\mathcal{X})$, and the uniform lower bounds on q_n and q. Together with $\|J_q(P_n - P)\|_{\mathcal{H}} = O_p(t_n)$, this yields

$$\|J_q(P_m - P)\|_{\mathcal{H}} = o_p(1).$$

Since $\widetilde{k}_q$ metrizes weak convergence on probability measures, $P_m \rightsquigarrow P$ in probability. Thus

$$\int f\, d\Delta_n \longrightarrow 0 \quad \text{in probability}$$

for every fixed $f \in C(\mathcal{X})$.

The identity

$$t_n^{-1} J_q \Delta_n = t_n^{-1} J_{q_n}\Delta_n + K(a_n\, d\Delta_n), \qquad a_n := \frac{q_n - q}{t_n q q_n}, \tag{3.29}$$

holds exactly. The first term on the right is $o_p(1)$, while a_n is tight in $C(\mathcal{X})$ by Lemma 3.29 and the uniform lower bounds on the scalings.

Fix $\eta > 0$ and choose a compact set $\mathcal{A}_\eta \subset C(\mathcal{X})$ such that $P(a_n \in \mathcal{A}_\eta) \ge 1 - \eta$ for all sufficiently large n. On this event,

$$\begin{aligned}\|K(a_n\, d\Delta_n)\|_{\mathcal{H}} &= \sup_{\|h\|_{\mathcal{H}} \le 1} \left| \int a_n h\, d\Delta_n \right| \\ &\le \sup_{f \in \mathcal{F}_\eta} \left| \int f\, d\Delta_n \right|,\end{aligned}$$

where

$$\mathcal{F}_\eta := \{ah : a \in \mathcal{A}_\eta,\ h \in \mathcal{H}_1\}.$$

The closure of $\mathcal{F}_\eta$ is compact in $C(\mathcal{X})$ because the closure of $\mathcal{H}_1$ is compact and multiplication is continuous in the uniform norm. A finite-net argument, using $\|\Delta_n\|_{\mathrm{tv}} \le 2$ and the preceding weak convergence, gives

$$\sup_{f \in \mathcal{F}_\eta} \left| \int f \, d\Delta_n \right| = o_p(1).$$

Letting $\eta \downarrow 0$ shows $\|K(a_n \, d\Delta_n)\|_{\mathcal{H}} = o_p(1)$. Equation (3.29) therefore yields

$$\|J_q(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}).$$

□

Proof of Theorem 3.5. Lemma 3.31 gives

$$\mathrm{MMD}_{\widetilde{k}_{q_n}}(P_m, P_n) = \|J_{q_n}(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}).$$

Applying Lemma 3.4 gives

$$\mathrm{MMD}_{\widetilde{k}_q}(P_m, P_n) = \|J_q(P_m - P_n)\|_{\mathcal{H}} = o_p(n^{-1/2}).$$

For the final conclusion, put

$$H_n := P_n - P, \qquad \Delta_n := P_m - P_n.$$

Then $\|J_q H_n\|_{\mathcal{H}} = O_p(n^{-1/2})$ by Lemma 3.29, while $\|J_q \Delta_n\|_{\mathcal{H}} = o_p(n^{-1/2})$. Using (3.8) and boundedness of R_P,

$$\begin{aligned} &\left| \ddot{D}_P(H_n + \Delta_n) - \ddot{D}_P(H_n) \right| \\ &\quad \le \varepsilon \|\mathsf{R}_P\| \|J_q H_n\|_{\mathcal{H}} \|J_q \Delta_n\|_{\mathcal{H}} + \frac{\varepsilon}{2} \|\mathsf{R}_P\| \|J_q \Delta_n\|_{\mathcal{H}}^2 \\ &\quad = o_p(n^{-1}). \end{aligned}$$

Since $H_n + \Delta_n = P_m - P$, the population quadratic form is preserved up to $o_p(n^{-1})$. □

Proof of Lemma 3.5. Let $\{e_{j,n}\}$ be an orthonormal eigenbasis of $q_n^\perp$ for T_{P_n}, with corresponding eigenvalues $1 > \lambda_{1,n} \ge \lambda_{2,n} \ge \cdots \ge 0$. Since z is orthogonal to the first r retained eigenvectors, write

$$z = \sum_{j > r} a_j e_{j,n}.$$

Then

$$\begin{aligned}
&\left\langle z, (I - \mathsf{T}_{P_n}^2)^{-1} z \right\rangle_{\mathcal{H}} - \|z\|_{\mathcal{H}}^2 \\
&\quad = \sum_{j>r} \frac{\lambda_{j,n}^2}{1 - \lambda_{j,n}^2} a_j^2 \\
&\quad \le \frac{\lambda_{r+1,n}^2}{1 - \lambda_{r+1,n}^2} \sum_{j>r} a_j^2,
\end{aligned}$$

where the inequality follows from monotonicity of $x^2/(1-x^2)$ on $[0,1)$. Since $\sum_{j>r} a_j^2 = \|z\|_{\mathcal{H}}^2$, multiplication by $\varepsilon/2$ gives the stated bound. □

Proof of Theorem 3.4. Set $\gamma^1 = 0$ and $\gamma^2 = \gamma$ in Theorem 3.7. Then

$$\ddot{D}_P(\gamma) = \frac{\varepsilon}{2} \left\langle J_q \gamma, \mathsf{R}_P J_q \gamma \right\rangle_{\mathcal{H}}.$$

For a finite signed measure $\gamma \in \mathcal{E}_q^0$,

$$\begin{aligned}
\int \Phi_q^0(x)\, d\gamma(x) &= \Pi_{q^\perp} \int \frac{k_\varepsilon(\cdot, x)}{q(x)}\, d\gamma(x) \\
&= \Pi_{q^\perp} J_q \gamma = J_q \gamma,
\end{aligned}$$

because $\langle J_q\gamma, q\rangle_{\mathcal{H}} = \gamma(\mathbf{1}) = 0$. Consequently,

$$\begin{aligned}
\iint g_{S,P}(x,y)\, d\gamma(x) d\gamma(y) &= \frac{\varepsilon}{2} \left\| \mathsf{R}_P^{1/2} J_q \gamma \right\|_{\mathcal{H}}^2 \\
&= \ddot{D}_P(\gamma).
\end{aligned}$$

The identity extends to all of $\mathcal{E}_q^0$ by continuity and also shows that $g_{S,P}$ is positive definite.

By Lemma 3.27, on $q^\perp$,

$$I \preceq \mathsf{R}_P \preceq \frac{1}{1-\rho^2} I.$$

Therefore

$$\frac{\varepsilon}{2} \|J_q\gamma\|_{\mathcal{H}}^2 \le \ddot{D}_P(\gamma) \le \frac{\varepsilon}{2(1-\rho^2)} \|J_q\gamma\|_{\mathcal{H}}^2.$$

Lemma 3.10 gives $\|J_q\gamma\|_{\mathcal{H}} = \|\gamma\|_{\mathcal{E}_q}$, proving the claimed equivalence. □

Proof of Theorem 3.2. Theorem 3.5 gives

$$\sqrt{n}\, \|J_q(P_m - P_n)\|_{\mathcal{H}} = o_p(1).$$

Together with Lemma 3.29, this implies

$$\sqrt{n}(P_m - P) = \sqrt{n}(P_n - P) + o_p(1)$$

in $\mathcal{E}_q^0$. Applying the second-order functional delta method to Theorem 3.7, with the first marginal fixed at P, gives

$$S(P, P_n) = \ddot{D}_P(P_n - P) + o_p(n^{-1}),$$
$$S(P, P_m) = \ddot{D}_P(P_m - P) + o_p(n^{-1}).$$

Theorem 3.5 shows that the two quadratic terms differ by $o_p(n^{-1})$. Hence

$$S(P, P_m) = S(P, P_n) + o_p(n^{-1}).$$

Moreover, boundedness of R_P and $\|J_q(P_n - P)\|_{\mathcal{H}} = O_p(n^{-1/2})$ give

$$S(P, P_n) = O_p(n^{-1}), \qquad S(P, P_m) = O_p(n^{-1}).$$

For fixed oversampling parameter θ, Lemma 3.2 gives computational complexity $O(n^2 m + m^3 \log(n/m))$. □

3.F Results for maximum mean discrepancy

In this subsection, we consider the "kernel thinning" problem as a special case of data compression, where we seek to construct P_m such that $Q_K(P_n - P_m) = O(n^{-1})$, for $K : \mathcal{P}(\mathcal{X}) \to \mathcal{F}$, $K\mu = \int k(\cdot, x)\, d\mu(x)$, the kernel mean embedding associated with a generic RKHS kernel k. Our analysis concerns the decomposition

$$Q_K(P_m - P) \le 2Q_K(P_m - P_n) + 2Q_K(P_n - P),$$

which follows from the basic inequality $(a+b)^2 \le 2(a^2 + b^2)$. The latter term on the right is $O_p(n^{-1})$ by classical V-statistic theory.

Lemma 3.32 (Parametric empirical MMD rate)**.** *If k is continuous, $Q_K(P_n - P) = O_p(n^{-1})$.*

Proof. Since k is continuous on the compact set $\mathcal{X}^2$, it is bounded, and hence

$$\mathbb{E}[k(X, Y)^2] < \infty, \qquad \mathbb{E}[k(X, X)] < \infty.$$

The result follows from the limiting distribution of V-statistics Serfling, 1980, Chapter 6.4, Theorem B. □

Our main bound is taken with minor adjustments from (Hayakawa et al., 2022). In the following, we decompose our kernel $K = K_0 + K_1$, hence we denote the corresponding RKHS norms by $\|\cdot\|_{K_1}$.

Lemma 3.33 (Residual-kernel bound for MMD compression)**.** *Let $K = K_0 + K_1$, k, k_0, k_1 the corresponding kernel functions, $K, K_0, K_1 \succeq 0$, and assume P_m is such that $K_0(P_n - P_m) = 0$ and $\int k_1(x,x)(dP_n - dP_m) \geq 0$. Then,*

$$Q_K(P_m - P_n) \leq 4\,\mathrm{tr}(K_1/n).$$

Proof. Let $f = \sum_{i=1}^n a_i k(X_i, \cdot) =: Ka$. Plugging in definitions and using the assumption that $K_0(P_m - P_n) = 0$, we have

$$\int f dP_m - \int f dP_n = (P_m - P_n)Ka = (P_m - P_n)[K_1 + K_0]a = (P_m - P_n)K_1 a,$$

giving us a corresponding element $\tilde{f}$ of the RKHS corresponding to k_1. Observe further that $\|\tilde{f}\|_{K_1}^2 = a^\top K_1 a \leq a^\top K a = \|f\|_K^2$. Applying these observations,

$$\begin{aligned}
(P_m - P_n)K(P_m - P_n) &= \sup_{\|f\|_K \leq 1} \left(\int f dP_m - \int f dP_n \right)^2 \\
&\leq \sup_{\|f\|_{K_1} \leq 1} \left(\int f dP_m - \int f dP_n \right)^2 \\
&= \sup_{\|f\|_{K_1} \leq 1} \left(\int \langle f, k_1(x,\cdot)\rangle dP_m - \int \langle f, k_1(x,\cdot)\rangle dP_n \right)^2 \\
&\leq \left(\int \sqrt{k_1(x,x)} dP_m + \int \sqrt{k_1(x,x)} dP_n \right)^2 \qquad \text{(Cauchy-Schwarz)} \\
&\leq 2 \int k_1(x,x) dP_m + 2 \int k_1(x,x) dP_n \\
&\leq 4 \int k_1(x,x) dP_n = \frac{4}{n}\,\mathrm{tr}(K_1),
\end{aligned}$$

where the final inequality holds by an assumption of the lemma. □

These requirements can be enforced by recombination once a rank-r positive-semidefinite approximation has been obtained: the r retained features together with the constant function

yield a probability measure supported on at most $r+1$ observations. Combining this observation with the random-projection Nyström approximation gives Lemma 3.2.

Proof of Lemma 3.2. Condition on the observed sample and selected kernel, and let

$$\widehat{K}_n := \big[\widehat{k}_n(X_i, X_j)\big]_{i,j=1}^n = n\widehat{\mathbf{K}}_n$$

be the unnormalized Gram matrix. Algorithm 3 and Theorem 4.1 of Tropp et al. (2017), applied with target rank r and sketch size θr, produce a positive-semidefinite rank-r Nyström approximation $K_0 \preceq \widehat{K}_n$ satisfying

$$\begin{aligned}\mathbb{E}\big[\|\widehat{K}_n - K_0\|_1 \mid X_1, \dots, X_n, \widehat{k}_n\big] &\le \left(1 + \frac{r}{r(\theta-1)-1}\right) \|\widehat{K}_n - \widehat{K}_{n,[r]}\|_1 \\ &= n\left(1 + \frac{r}{r(\theta-1)-1}\right) \mathrm{Tail}_r(\widehat{\mathbf{K}}_n),\end{aligned}$$

where $\widehat{K}_{n,[r]}$ is the best rank-r approximation of $\widehat{K}_n$. The recombination algorithm then constructs a probability measure P_{r+1} satisfying the conditions of Lemma 3.33 for the decomposition $\widehat{K}_n = K_0 + (\widehat{K}_n - K_0)$. Applying that lemma and taking the conditional expectation gives

$$\mathbb{E}\left[\mathrm{MMD}^2_{\widehat{k}_n}(P_n, P_{r+1}) \,\middle|\, X_1, \dots, X_n, \widehat{k}_n\right] \le 4\left(1 + \frac{r}{r(\theta-1)-1}\right) \mathrm{Tail}_r(\widehat{\mathbf{K}}_n).$$

The stated complexity is the combined cost of the Nyström and recombination steps. □

Chapter 4

ENTROPIC SELF-TRANSPORT IN THE SMALL-ε LIMIT: SCORES, DENSITIES, DIFFUSION MAPS, AND FAST CORESETS

Abstract. Entropic optimal self-transport has recently been connected to the estimation of scores, diffusion processes, and manifold learning. We analyze the small-ε symmetric Schrödinger equation and derive a second-order expansion of the self-EOT potential under smooth bounded-score conditions. The expansion makes precise the relationship between self-EOT quantities and classical KDE estimators of density, score, and diffusion. A key observation is that an EOT regularization parameter ε naturally matches a Gaussian smoothing parameter $\varepsilon/2$; at this scale the leading terms agree, while the first discrepancies are explicit. These expressions yield examples in which KDEs outperform self-EOT and others in which self-EOT performs better. They also give a converse computational perspective: self-EOT quantities can themselves be approximated by simple Gaussian-smoothing expressions. In particular, two symmetric Sinkhorn updates already recover the full first-order self-EOT scaling, with $o(\varepsilon)$ error relative to the converged scaling.

Keywords. Entropic Optimal Transport, Diffusion Maps, Score Estimation, Nyström approximation

4.1 Introduction

4.1.1 Background

Entropic self-transport concerns the entropic optimal transport (EOT) map from a probability distribution μ to itself, an object of much recent interest. Ramdas et al. (2015) and Genevay et al. (2018) first highlighted this map, observing that standard EOT lacks the properties of a statistical divergence. To address this, they motivated its centering, the Sinkhorn divergence, which Feydy et al. (2019b) later studied in more depth.

Although transporting a distribution to itself is trivial in the unregularized problem, recent work has shown that its entropically regularized counterpart contains nontrivial geometric and statistical information, with connections to diffusion processes, score estimation, and manifold learning (Marshall and Coifman, 2019; Wormell and Reich, 2021; Cheng and Landa, 2024; Mordant, 2024; Agarwal et al., 2025). We discuss the closest related work in Section 4.1.3. Our goal is to give a direct small-ε analysis of the self-EOT Schrödinger equation and to use the resulting expansion to clarify the relationship between these quantities and their classical counterparts based on Gaussian smoothing (Davis et al., 2011; Parzen, 1962; Fukunaga and Hostetler, 1975; Coifman and Lafon, 2006b).

4.1.2 Overview of Main Findings

More formally, the self-EOT problem we consider is

$$\mathrm{OT}_\varepsilon(\mu,\mu) := \min_{\pi\in\Pi(\mu,\mu)} \left\{ \int \|x-y\|^2 \, d\pi(x,y) + \varepsilon \, \mathrm{KL}(\pi \parallel \mu\otimes\mu) \right\}, \tag{4.1}$$

where $\Pi(\mu,\mu)$ is the set of joint distributions with marginals μ. For $\varepsilon > 0$, the minimizer π_ε has density

$$\xi_\varepsilon(x,y) := \frac{d\pi_\varepsilon}{d(\mu\otimes\mu)}(x,y) = \exp\left(\frac{f_\varepsilon(x)+f_\varepsilon(y)}{\varepsilon}\right) k_\varepsilon(x,y), \tag{4.2}$$

where

$$k_\varepsilon(x,y) := \exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right)$$

is a Gaussian kernel and the potential f_ε solves the Schrödinger equation

$$f_\varepsilon(x) = -\varepsilon \log \int \exp[f_\varepsilon(y)/\varepsilon] \, k_\varepsilon(x,y) \, d\mu(y). \tag{4.3}$$

For the strictly positive Gaussian kernel, a symmetric solution is unique whenever it exists. Theorem 4.1 constructs this solution for all sufficiently small ε under the assumptions used below.

A main finding of this chapter is that, when μ has a Lebesgue density

$$h^2 := \frac{d\mu}{d\lambda}$$

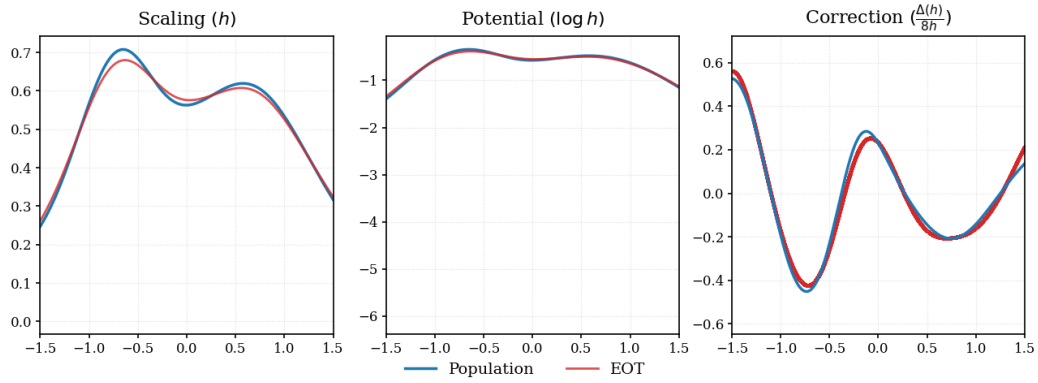


Figure 4.1: Data is sampled from a mixture of smoothed Laplace distributions, $\propto \exp(-\lambda\sqrt{1+(x-c)^2})$. In red, from left to right, we plot $h_\varepsilon := c_\varepsilon^{1/2} e^{-f_\varepsilon/\varepsilon}$, $\log h_\varepsilon = f_\varepsilon/\varepsilon - \frac{1}{2}\log c_\varepsilon$, and $(\log h_\varepsilon - \log h)/\varepsilon$, using their empirical analogues for $N = 50{,}000$ samples and $\varepsilon = 0.1$.

on $\mathbb{R}^d$ whose logarithm is sufficiently smooth, the potential satisfies the expansion

$$f_\varepsilon = -\varepsilon\frac{d}{4}\log(\pi\varepsilon) - \varepsilon\log h - \varepsilon^2\frac{\Delta h}{8h} + o(\varepsilon^2). \tag{4.4}$$

Here π is the mathematical constant, Δ denotes the Laplace operator, and the little-oh term converges to zero as $\varepsilon \to 0$ in the $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ norm introduced below. Figure 4.1 illustrates the three levels of the expansion on a one-dimensional mixture of smoothed-Laplace components. We compare the empirical self-EOT root scaling, its logarithm, and its rescaled first-order residual with the corresponding population quantities. A formal statement is given in Theorem 4.1. Our methodology stems from that developed in Goldfeld et al. (2022), Gonzalez-Sanz et al. (2022), and Kokot and Luedtke (2025) to derive limiting results in the fixed $\varepsilon > 0$ case via linearization of the Schrödinger equations.

Our expansion, being explicit in terms of h, its derivatives, and the heat operator, can be compared to classical estimators such as kernel density estimators (KDEs). A recurring feature of these comparisons is a factor-of-two in the natural Gaussian scale. This can be seen from the barycentric formulas:

$$-\frac{2\nabla f_\varepsilon}{\varepsilon} = \frac{4}{\varepsilon}\big(\mathbb{E}_{\pi_\varepsilon}[Y \mid X] - X\big), \qquad \nabla \log P_\delta(h^2) = \frac{2}{\delta}\big(\mathbb{E}[Y \mid X_\delta] - X_\delta\big).$$

Matching the displacement normalization gives

$$\frac{4}{\varepsilon} = \frac{2}{\delta} \quad \Longrightarrow \quad \boxed{\delta = \varepsilon/2}.$$

This leads to the following results developed in Section 4.3:

1. **Density estimation:** As a first step, we highlight that $c_\varepsilon e^{-2f_\varepsilon/\varepsilon} = h^2 + \frac{\varepsilon}{4} h\Delta h + o(\varepsilon)$, where $c_\varepsilon = (\pi\varepsilon)^{-d/2}$. We show that the **kernel density estimator (KDE)** $\boldsymbol{P_{\varepsilon/2}(h^2)}$ matches the self-EOT estimator up to an explicit $O(\varepsilon)$ correction.

2. **Score estimation:** Since the leading constant in (4.4) vanishes under differentiation, $\nabla \log h = -\nabla f_\varepsilon/\varepsilon - \varepsilon\nabla\left(\frac{\Delta h}{8h}\right) + o(\varepsilon)$, which gives a first-order characterization of the score in terms of self-EOT. We give the corresponding first-order expansion for the **Gaussian KDE score** and identify its explicit bias relative to self-EOT.

3. **Langevin Diffusion:** The bi-stochastic Laplacian $L_\varepsilon(f) = \frac{1}{\varepsilon}[f - \int \xi_\varepsilon(x,y) f(y)\, d\mu(y)] = \frac{1}{4}\left(\Delta f + 2\nabla \log h \cdot \nabla f\right) + O(\varepsilon)$, matching the **symmetric normalized Laplacian**, common in the Diffusion Maps literature, up to $O(\varepsilon)$ terms.

These results make precise the extent to which the density, score, and diffusion objects induced by self-EOT reproduce familiar Gaussian estimators, while also identifying the first terms on which the constructions differ. Our potential expansion also recovers the small-ε expansion of the entropic self-transport cost from (Conforti and Tamanini, 2021; Agarwal et al., 2025). We record the calculation in Appendix 4.B.

Using these observations, in Section 4.4 we reverse this perspective and construct KDE-based approximations to self-EOT quantities. We prove that two symmetric Sinkhorn steps reproduce the first-order population scaling, and utilize this procedure for fast coreset selection.

4.1.3 Related Work

Recent asymptotic work on entropic optimal transport has largely considered two regimes: perturbations of the marginal distributions at fixed $\varepsilon > 0$, and the vanishing-regularization

limit $\varepsilon \downarrow 0$. For fixed ε, (Mena and Weed, 2019; Barrio et al., 2022) established sample-complexity bounds and central limit theorems for entropic transport costs. Analogous results for the potentials and other related quantities were developed in Goldfeld et al. (2022) and Gonzalez-Sanz et al. (2022) via linearization of the Schrödinger system of equations.

In the vanishing-regularization setting, Mordant (2024) studies self-EOT with decreasing regularization and shows that a rescaling of its barycentric projection yields a score estimator, while also deriving limiting distributions for empirical potentials and plans. Agarwal et al. (2025) connect the same-marginal Schrödinger bridge to Langevin diffusion, showing that the derivative at zero of its conditional-expectation operators is the Langevin generator. Related diffusion approximations for Schrödinger bridges are developed by Mulcahy and Pal (2025).

Our proof draws on perspectives from both the fixed and vanishing regularization literature. We analyze the Schrödinger system directly, deriving limiting behavior as $\varepsilon \downarrow 0$ via an implicit function theorem.

Recent work in the diffusion-maps literature draws attention to these same self-EOT quantities. Starting from the local-kernel construction of Coifman and Lafon (2006b), Marshall and Coifman (2019) relate bi-stochastically normalized Gaussian kernels to heat operators on manifolds. Wormell and Reich (2021) study a Sinkhorn normalization that approximates Langevin diffusion, while Cheng and Landa (2024) establish convergence of bi-stochastically normalized graph Laplacians to weighted manifold Laplacians and analyze their robustness to high-dimensional noise.

A few further connections arise through the computational implications of our results. Altschuler, Weed, et al. (2018) and Altschuler, Bach, et al. (2019) study finite Sinkhorn iteration and Nyström approximations for scalable entropic OT, while related work uses kernel discrepancies for distribution compression and coreset construction (Li et al., 2024b; Kokot and Luedtke, 2025). Our two-step sinkhorn iteration and coreset selection guarantees bring these ideas into the setting of vanishing regularization.

4.2 Expansion via Implicit Function Theorem

4.2.1 Overview of Proof Strategy

For $\varepsilon > 0$, set

$$c_\varepsilon := (\pi\varepsilon)^{-d/2}, \qquad p_\varepsilon(z) := c_\varepsilon e^{-|z|^2/\varepsilon}, \qquad P_\varepsilon u(x) := \int_{\mathbb{R}^d} p_\varepsilon(x-y)u(y)\,dy. \tag{4.5}$$

Thus $P_\varepsilon = e^{(\varepsilon/4)\Delta}$ and define $P_0 := I$. The symbol c_ε will always denote the normalizing constant in (4.5); in particular, uppercase C is reserved for unspecified bounds.

Writing $d\mu = h^2\,dx$, the Schrödinger equation (4.3) becomes

$$\frac{f_\varepsilon}{\varepsilon} + \log\Big[c_\varepsilon^{-1} P_\varepsilon\Big(e^{f_\varepsilon/\varepsilon}h^2\Big)\Big] = 0. \tag{4.6}$$

Since $P_\varepsilon \to I$ as $\varepsilon \to 0$, this parameterization suggests that

$$f_\varepsilon/\varepsilon \approx -2\log\Big[c_\varepsilon^{-1/2}h\Big] - f_\varepsilon/\varepsilon \iff f_\varepsilon/\varepsilon \approx -\log(c_\varepsilon^{-1/2}h).$$

This motivates the study of the residual

$$\phi_\varepsilon := \frac{f_\varepsilon}{\varepsilon} + \log\Big(c_\varepsilon^{-1/2}h\Big). \tag{4.7}$$

For h appropriately regular, we will show that $\phi_\varepsilon \to 0$, and derive an explicit $O(\varepsilon)$ leading term.

Define the h-transform

$$P_\varepsilon^h u := \frac{P_\varepsilon(hu)}{h}. \tag{4.8}$$

Our primary interest will be to study the 0s of the map ψ defined as follows:

$$\psi(\varepsilon,\phi) := \phi + \log P_\varepsilon^h(e^\phi), \qquad \psi(0,\phi) := 2\phi. \tag{4.9}$$

A direct substitution in (4.6) shows that $\psi(\varepsilon,\phi_\varepsilon) = 0$. To study the roots of this map, we analyze the fixed points of $\mathcal{T}_\varepsilon$:

$$A_\varepsilon := D_\phi\psi(\varepsilon,0), \qquad \mathcal{T}_\varepsilon(\phi) := \phi - A_\varepsilon^{-1}\psi(\varepsilon,\phi). \tag{4.10}$$

4.2.2 Warm-Up: Multivariate Gaussian

We illustrate our proof strategy for a multivariate Gaussian. We focus on $\mu = N(0, I_d)$ for notational simplicity—the analysis extends to any nondegenerate Gaussian with the obvious modifications. We show that

$$f_\varepsilon(x) = \underbrace{-\varepsilon \frac{d}{4} \log(\pi\varepsilon) - \varepsilon \log h(x) - \varepsilon^2 \frac{\Delta(h)(x)}{8h(x)}}_{\text{dominant terms on RHS of (4.4)}} + (1 + |x|^2)\, o(\varepsilon^2) \tag{4.11}$$

with the $o(\varepsilon^2)$ a scalar that does not depend on x. Hence, (4.4) holds with the $o(\varepsilon^2)$ term shrinking to 0 for all derivatives of any fixed order, both in $L^p(\mu)$ for $p < \infty$ and uniformly on compacts. This differs slightly from our main result to follow, where instead that term shrinks to 0 in $C^s_{\text{ub}}(\mathbb{R}^d)$.

We use the search space $\Phi := \{\phi_{a,b} : a, b \in \mathbb{R}\}$ with $\phi_{a,b}(x) = a|x|^2 + b$. Working with this search space is reasonable given that $\log h$ and $\Delta(h)/h$ both belong to Φ when $\mu = N(0_d, I_d)$, and so (4.4) can only possibly hold if $\phi_\varepsilon \in \Phi$. A direct calculation also establishes that $\psi(\varepsilon, \Phi) \subseteq \Phi$. Indeed, in Appendix 4.A.1, we show that

$$\psi(\varepsilon, \phi_{a,b})(x) =: F(\varepsilon, a, b)^\top \begin{bmatrix} |x|^2 \\ 1 \end{bmatrix} \tag{4.12}$$

for $F : [0, \infty) \times \mathbb{R}^2 \to \mathbb{R}^2$ a smooth map. Thus, $\psi(\varepsilon, \phi_{a,b}) = 0$ is equivalent to the root-finding condition $F(\varepsilon, a, b) = 0$. While the full form of F is somewhat involved (Eq. S1), it admits a smooth extension to negative ε whose behavior near $\varepsilon = 0$ is simple:

1. $F(0, a, b) = (2a, 2b)$, and so the Jacobian of $(a, b) \mapsto F(0, a, b)$ has inverse $\frac{1}{2} I_2$;

2. $\partial_\varepsilon F(\varepsilon, 0, 0)|_{\varepsilon=0} = \frac{1}{16}(1, -2d)$.

Using the above facts, an implicit function theorem guarantees the existence of a unique smooth solution curve $(a_\varepsilon, b_\varepsilon)$ in a neighborhood of $\varepsilon = 0$ with $(a_0, b_0) = (0, 0)$ and $F(\varepsilon, a_\varepsilon, b_\varepsilon) = 0$. In light of (4.12), the latter of these properties is equivalent to $\psi(\varepsilon, \phi_{a_\varepsilon, b_\varepsilon}) = 0$, and so $\phi_\varepsilon = \phi_{a_\varepsilon, b_\varepsilon}$. Beyond existence of the smooth curve $(a_\varepsilon, b_\varepsilon)$, the implicit

function theorem also shows its tangent satisfies $(\dot{a}_0, \dot{b}_0) := \partial_\varepsilon(a_\varepsilon, b_\varepsilon)|_{\varepsilon=0} = -\frac{1}{2}\partial_\varepsilon F(\varepsilon, 0, 0)|_{\varepsilon=0}$, yielding the Taylor approximation

$$(a_\varepsilon, b_\varepsilon) = (a_0, b_0) + \varepsilon(\dot{a}_0, \dot{b}_0) + o(\varepsilon) = \frac{\varepsilon}{32}(-1, 2d) + o(\varepsilon).$$

Hence, for ε small enough, $\phi_\varepsilon(x) = \phi_{a_\varepsilon, b_\varepsilon} = \frac{\varepsilon}{32}(-|x|^2 + 2d) + o(\varepsilon)$. Combining this with the fact that $\log h(x) = -\frac{d}{4}\log(2\pi) - \frac{|x|^2}{4}$ and $\frac{\Delta h(x)}{h(x)} = \frac{|x|^2}{4} - \frac{d}{2}$ yields (4.11).

Remark 4.1 (Exact expansion)**.** Our implicit function theorem argument only characterizes the potential f_ε in a neighborhood of 0, and only does so approximately. In Appendix 4.A.2, we use the closed-form expression for Gaussian self-EOT plans from Bunne et al., 2023 to provide a global characterization: $f_\varepsilon = \alpha_\varepsilon|x|^2 + \beta_\varepsilon$ for all $\varepsilon > 0$, with the coefficients $\alpha_\varepsilon, \beta_\varepsilon$ taking known closed forms. Taylor expanding those coefficients yields a strengthening of (4.11), with the remainder replaced by $(1+|x|^2)O(\varepsilon^4)$. While this strengthening is interesting in its own right, it heavily relies on the known form of Gaussian transport plans. It therefore does not appear to help with the general sort of expansion that we consider.

4.2.3 Main Result: Distributions with Smooth Scores

For a nonnegative integer s, write

$$C^s_{\mathrm{ub}}(\mathbb{R}^d) := \left\{u \in C^s_b(\mathbb{R}^d) : \lim_{z\to 0} \|u(\cdot + z) - u\|_{C^s_b} = 0\right\}, \qquad \|u\|_{C^s_b} := \sum_{|\alpha|\le s} \|\partial^\alpha u\|_\infty. \quad (4.13)$$

We denote by $\|\cdot\|$ the C^s_b norm unless stated otherwise, and $|\cdot|$ to denote the Euclidean norm on $\mathbb{R}^d$.

Lemma 4.1 (Regularity, marginal characterization, and uniqueness)**.** *Let $h > 0$, $\int h^2 = 1$, and suppose $\nabla \log h \in C^{s+1}_{\mathrm{ub}}(\mathbb{R}^d)$. For $\varepsilon > 0$ and $\phi \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$, define*

$$f_{\varepsilon,\phi} := -\varepsilon \log(c_\varepsilon^{-1/2} h) + \varepsilon\phi, \quad d\pi_{\varepsilon,\phi}(x, y) := p_\varepsilon(x - y)h(x)h(y)e^{\phi(x)+\phi(y)}\, dx\, dy.$$

Then $\psi(\varepsilon, \phi) = 0$ is equivalent to $\pi_{\varepsilon,\phi} \in \Pi(\mu, \mu)$, $d\mu = h^2\, dx$. For every $\varepsilon > 0$, the equation $\psi(\varepsilon, \cdot) = 0$ has at most one root in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$. Moreover, there exists $\varepsilon_0 > 0$ such that, for every $0 < \varepsilon \le \varepsilon_0$, this root exists. Whenever the relevant entropy integrals are finite, $\pi_{\varepsilon,\phi}$ is also the unique minimizer of the entropic transport objective over $\Pi(\mu, \mu)$.

Proof. We evaluate the marginal of $d\pi_{\varepsilon,\phi}$ to be

$$h(x)e^{\phi(x)}P_\varepsilon(he^\phi)(x).$$

It equals the density $h(x)^2$ precisely when

$$e^{\phi(x)}\frac{P_\varepsilon(he^\phi)(x)}{h(x)} = 1,$$

which, by positivity, is equivalent to

$$\phi + \log\left(\frac{P_\varepsilon(he^\phi)}{h}\right) = \psi(\varepsilon,\phi) = 0.$$

Symmetry gives the second marginal. Existence of a root $\phi_\varepsilon \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$ for all sufficiently small ε follows directly from Lemma S12; its hypotheses are verified by Lemmas S1, S3, S5, S8, and S11.

To prove global uniqueness in C^s_{ub}, let ϕ_0, ϕ_1 be roots, set $\delta = \phi_0 - \phi_1$, and denote the corresponding couplings by π_0, π_1. Since they have the same marginals,

$$\mathrm{KL}(\pi_0\|\pi_1) = 2\int \delta\, d\mu, \qquad \mathrm{KL}(\pi_1\|\pi_0) = -2\int \delta\, d\mu.$$

As the KL divergence is nonnegative this implies $\delta = 0$ hence $\pi_0 = \pi_1$. Finally, for every $\gamma \in \Pi(\mu,\mu)$ for which the objectives are finite,

$$J_\varepsilon(\gamma) - J_\varepsilon(\pi_{\varepsilon,\phi}) = \varepsilon\,\mathrm{KL}(\gamma\|\pi_{\varepsilon,\phi}) \geq 0,$$

with equality only when $\gamma = \pi_{\varepsilon,\phi}$.

□

Remark 4.2 (Choice of function space)**.** The assumption on $\nabla \log h$ is tailored to the h-transformed heat operator

$$P^h_\varepsilon u(x) = \int p_\varepsilon(y)\frac{h(x+y)}{h(x)}u(x+y)\,dy.$$

Indeed, bounded and uniformly continuous derivatives of $\nabla \log h$ control the ratios $h(x+y)/h(x) = \exp(\log h(x+y) - \log h(x))$, ensuring that P^h_ε acts continuously on C^s_{ub} and tends strongly to the identity as $\varepsilon \downarrow 0$. This topology is also stable under multiplication and smooth composition, which is needed for the logarithms and exponentials in the Schrödinger equation. The extra derivative in $\nabla \log h \in C^{s+1}_{\mathrm{ub}}$ is used to obtain the first-order C^s_{ub} expansion of P^h_ε.

Theorem 4.1 (Second-order expansion of the self-EOT potential)**.** *Let $h > 0$ satisfy $\int h^2 = 1$, put $\ell := \log h$, and suppose $\nabla \ell \in C^{s+1}_{\mathrm{ub}}(\mathbb{R}^d)$ for some integer $s \geq 0$. Then there is $\varepsilon_0 > 0$ such that, for $0 < \varepsilon \leq \varepsilon_0$, the symmetric Schrödinger equation has a unique potential f_ε, and*

$$\left\| f_\varepsilon + \varepsilon \frac{d}{4} \log(\pi\varepsilon) + \varepsilon \log h + \varepsilon^2 \frac{\Delta h}{8h} \right\| = o(\varepsilon^2). \tag{4.14}$$

Equivalently, the residual in (4.7) *satisfies*

$$\frac{\phi_\varepsilon}{\varepsilon} \longrightarrow -\frac{1}{8}\frac{\Delta h}{h} \qquad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{4.15}$$

Proof. By Lemma S12, there are $\varepsilon_0, r > 0$ and a unique $\phi_\varepsilon \in \overline{B}_r(0) \subset C^s_{\mathrm{ub}}(\mathbb{R}^d)$, for every $0 \leq \varepsilon \leq \varepsilon_0$, such that

$$\psi(\varepsilon, \phi_\varepsilon) = 0. \tag{1}$$

Moreover $\varepsilon \mapsto \phi_\varepsilon$ is continuous and $\phi_\varepsilon \to 0$ in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ as $\varepsilon \downarrow 0$. By Lemma 4.1 this root corresponds to a symmetric Schrödinger potential.

To obtain the expansion, apply the fundamental theorem of calculus to $\phi \mapsto \psi(\varepsilon, \phi)$:

$$0 = \psi(\varepsilon, 0) + \left[\int_0^1 D_\phi \psi(\varepsilon, t\phi_\varepsilon)\, dt \right] \phi_\varepsilon. \tag{2}$$

Denote the averaged differential in brackets by $\overline{A}_\varepsilon$. Thus we have $\phi_\varepsilon = \overline{A}_\varepsilon^{-1} \psi(\varepsilon, 0)$, and we aim to convert this into a first order expansion. Indeed, Lemma S13 shows that $\overline{A}_\varepsilon$ is invertible for small ε, its inverse is uniformly bounded, and

$$\overline{A}_\varepsilon^{-1} v \longrightarrow \frac{1}{2} v \quad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d) \quad \text{for each fixed } v \in C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{3}$$

At $\phi = 0$,

$$\psi(\varepsilon, 0) = \log\left(\frac{P_\varepsilon h}{h} \right).$$

Lemma S7, applied with $u = 1$, gives

$$\frac{P_\varepsilon h}{h} = 1 + \frac{\varepsilon}{4}\frac{\Delta h}{h} + o(\varepsilon).$$

Because $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ is a Banach algebra and the expression in parentheses is bounded away from zero, the Taylor formula for the logarithm yields

$$\frac{\psi(\varepsilon, 0)}{\varepsilon} \longrightarrow \frac{1}{4}\frac{\Delta h}{h} \quad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{4}$$

Set

$$v_\varepsilon := \frac{\psi(\varepsilon, 0)}{\varepsilon}, \qquad v := \frac{1}{4}\frac{\Delta h}{h}.$$

By (3)–(4) and the uniform inverse bound,

$$\left\| \overline{A}_\varepsilon^{-1} v_\varepsilon - \frac{1}{2} v \right\| \le \|\overline{A}_\varepsilon^{-1}\| \, \|v_\varepsilon - v\| + \left\| \overline{A}_\varepsilon^{-1} v - \frac{1}{2} v \right\|$$
$$\longrightarrow 0.$$

Combining this with (2) gives

$$\frac{\phi_\varepsilon}{\varepsilon} = -\overline{A}_\varepsilon^{-1} v_\varepsilon \longrightarrow -\frac{1}{8}\frac{\Delta h}{h}, \tag{5}$$

and

$$f_\varepsilon = -\varepsilon \log(c_\varepsilon^{-1/2} h) + \varepsilon \phi_\varepsilon = -\varepsilon \log(c_\varepsilon^{-1/2} h) - \frac{\varepsilon^2}{8}\frac{\Delta h}{h} + o(\varepsilon^2).$$

□

An analogous expansion holds for smooth positive densities on compact Riemannian manifolds embedded in Euclidean space; see Theorem S1 in Appendix 4.D. In that setting, the Euclidean Laplacian is replaced by the Laplace–Beltrami operator, with an additional explicit curvature correction arising from the ambient Gaussian kernel.

4.3 Connections to Classical Estimators

Theorem 4.1 expresses the small-ε behavior of the self-EOT potential in terms of the root density h and the heat operator. We now use this expansion to compare the density factor, score, and diffusion process estimators induced by self-EOT with their classical Gaussian-kernel counterparts. The common pattern is that, after matching the Gaussian scales, the leading terms agree and the first difference is explicit. Throughout this section, denote

$$q_\varepsilon := c_\varepsilon e^{-2f_\varepsilon/\varepsilon}.$$

All little-oh terms are taken in the indicated $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ norm, and the Gaussian normalization is that of (4.5).

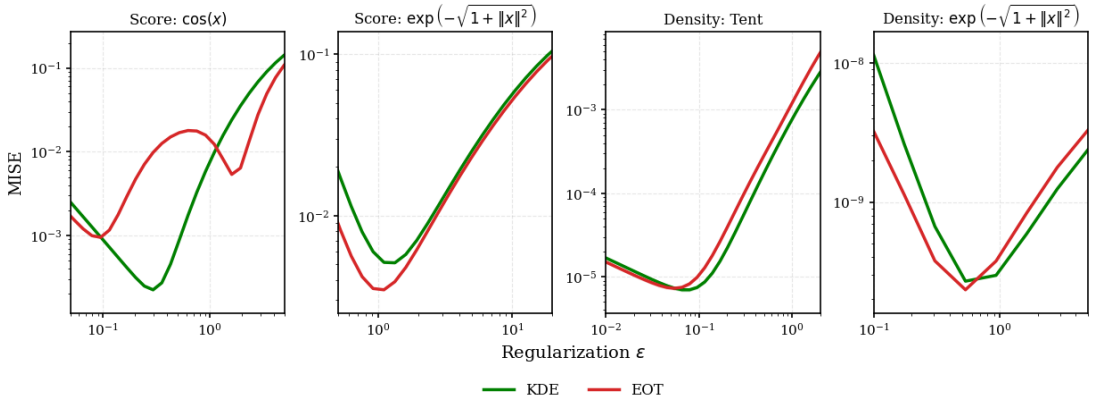


Figure 4.2: Comparison of self-EOT and Gaussian KDE at the matched smoothing parameter $\delta = \varepsilon/2$. The left panels report score error and the right panels density error. The cosine-score and tent-density examples are constructed so that the leading KDE bias vanishes on the interior evaluation region, while the smoothed-Laplace examples provides a control.

4.3.1 Density estimation

Although q_ε is defined implicitly through the Schrödinger equation, its first-order expansion can be simply expressed as the root density h multiplied by its convolution with a Gaussian.

Corollary 4.1 (First-order scaling and density expansion)**.** *Under the assumptions of Theorem 4.1,*

$$q_\varepsilon = hP_\varepsilon h + o(\varepsilon) = h^2 + \frac{\varepsilon}{4} h\Delta h + o(\varepsilon). \tag{4.16}$$

Proof. By the definition of the density factor and the normalization used in the theorem,

$$q_\varepsilon = c_\varepsilon e^{-2f_\varepsilon/\varepsilon} = h^2 e^{-2\phi_\varepsilon}. \tag{1}$$

Theorem 4.1 gives

$$-2\phi_\varepsilon = \frac{\varepsilon}{4}\frac{\Delta h}{h} + o(\varepsilon). \tag{2}$$

The exponential map is twice continuously Fréchet differentiable on bounded subsets of the Banach algebra $C^s_{\mathrm{ub}}(\mathbb{R}^d)$. Hence

$$e^{-2\phi_\varepsilon} = 1 - 2\phi_\varepsilon + O(\|\phi_\varepsilon\|^2) = 1 + \frac{\varepsilon}{4}\frac{\Delta h}{h} + o(\varepsilon),$$

because $\|\phi_\varepsilon\| = O(\varepsilon)$. Multiplying by h^2 in (1) gives

$$q_\varepsilon = h^2 + \frac{\varepsilon}{4} h\Delta h + o(\varepsilon).$$

The same heat expansion gives $hP_\varepsilon h = h^2 + (\varepsilon/4)h\Delta h + o(\varepsilon)$. □

As motivated in Section 4.2.1, we compare self-EOT at regularization ε with Gaussian smoothing at parameter $\delta = \varepsilon/2$. The next result corroborates this scaling, as it also aligns the common first-order density term.

Lemma 4.2 (KDE comparison)**.** *Under the assumptions of Theorem 4.1,*

$$P_{\varepsilon/2}(h^2) = h^2 + \frac{\varepsilon}{8}\Delta(h^2) + o(\varepsilon) = q_\varepsilon + \frac{\varepsilon}{4}|\nabla h|^2 + o(\varepsilon). \tag{4.17}$$

Proof. The heat-semigroup expansion at parameter $\varepsilon/2$ gives

$$P_{\varepsilon/2}(h^2) = h^2 + \frac{\varepsilon}{8}\Delta(h^2) + o(\varepsilon). \tag{1}$$

The product rule yields

$$\Delta(h^2) = 2h\Delta h + 2|\nabla h|^2. \tag{2}$$

Combining (1)–(2),

$$P_{\varepsilon/2}(h^2) = h^2 + \frac{\varepsilon}{4}\big(h\Delta h + |\nabla h|^2\big) + o(\varepsilon).$$

Subtracting Corollary 4.1 proves the claim. □

The right panels of Figure 4.2 illustrate this comparison. The first example is deliberately constructed to favor KDE. Let

$$w_0(x) = \begin{cases} x/4, & 0 \le x \le 2, \\ (4-x)/4, & 2 < x \le 4, \\ 0, & \text{otherwise}, \end{cases}$$

be the tent density, and let

$$w_\varepsilon := w_0 * p_\varepsilon, \qquad \varepsilon = 0.01.$$

The unsmoothed density p_0 is linear on either side of its mode, so $p_0'' = 0$ away from the kink. Hence the leading matched-KDE correction $(\varepsilon/8)p''$ vanishes there, whereas the corresponding self-EOT correction $(\varepsilon/4)hh''$, with $h = \sqrt{p}$, need not. We use $N = 100{,}000$ observations and evaluate the error only on an interior region away from the kink and boundary. For comparison, the second density experiment uses

$$w(x) \propto \exp\Big(-\sqrt{1+|x|^2}\Big), \qquad x \in \mathbb{R}^5,$$

for which neither first-order correction is forced to vanish. We use $N = 100{,}000$ training observations and evaluate the squared error against the normalized population density on 5,000 independently drawn points with $|x| < 6$.

4.3.2 Score estimation

Differentiating the potential expansion allows us to compare self-EOT and Gaussian estimators of the score. These correspond respectively to barycentric projection and mean shift/Gaussian denoising (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Vincent, 2011).

Corollary 4.2 (Self-EOT score expansion)**.** *Under the assumptions of Theorem 4.1,*

$$\left\| \phi_\varepsilon - \frac{1}{2}\log\frac{h}{P_\varepsilon h} \right\| = o(\varepsilon). \tag{4.18}$$

If $s \geq 1$, then in $C^{s-1}_{\mathrm{ub}}(\mathbb{R}^d)$,

$$\frac{2\nabla f_\varepsilon}{\varepsilon} = -\nabla \log(hP_\varepsilon h) + o(\varepsilon), \tag{4.19}$$

$$= -\nabla \log h^2 - \frac{\varepsilon}{4}\nabla\left(\frac{\Delta h}{h}\right) + o(\varepsilon). \tag{4.20}$$

Proof. Lemma S7 gives $\log(P_\varepsilon h/h) = (\varepsilon/4)(\Delta h/h) + o(\varepsilon)$, while Theorem 4.1 gives $\phi_\varepsilon = -(\varepsilon/8)(\Delta h/h) + o(\varepsilon)$. This proves (4.18). Differentiating (4.7) and using (4.18) yields (4.19); expanding the logarithm of $P_\varepsilon h/h$ yields (4.20). □

Lemma 4.3 (KDE score comparison)**.** *Under the assumptions of Theorem 4.1, if $s \geq 1$, then in $C^{s-1}_{\mathrm{ub}}(\mathbb{R}^d)$,*

$$\nabla \log P_\varepsilon(h^2) = \nabla \log h^2 + \frac{\varepsilon}{4}\nabla\left(\frac{\Delta(h^2)}{h^2}\right) + o(\varepsilon). \tag{4.21}$$

At the matched parameter $\varepsilon/2$,

$$\nabla \log P_{\varepsilon/2}(h^2) = -\frac{2\nabla f_\varepsilon}{\varepsilon} + \frac{\varepsilon}{4}\nabla|\nabla \log h|^2 + o(\varepsilon). \tag{4.22}$$

Proof. Set $r_\varepsilon := P_\varepsilon(h^2)/h^2$. Applying Lemma S7 with log-root 2ℓ gives

$$r_\varepsilon = 1 + \frac{\varepsilon}{4}\frac{\Delta(h^2)}{h^2} + o(\varepsilon).$$

Since $\log P_\varepsilon(h^2) = \log h^2 + \log r_\varepsilon$, composition with log near 1 and differentiation prove (4.21). At parameter $\varepsilon/2$, subtract (4.20) and use

$$\frac{\Delta(h^2)}{h^2} = 2\frac{\Delta h}{h} + 2|\nabla \log h|^2.$$

This proves (4.22). □

Example 4.1 (Matched denoising representations)**.** Let $(X, Y) \sim \pi_\varepsilon$. Barycentric projection gives

$$\nabla f_\varepsilon(x) = 2\big(x - \mathbb{E}_{\pi_\varepsilon}[Y \mid X = x]\big),$$

and therefore

$$-\frac{2\nabla f_\varepsilon(x)}{\varepsilon} = \frac{4}{\varepsilon}\big(\mathbb{E}_{\pi_\varepsilon}[Y \mid X = x] - x\big). \tag{4.23}$$

For Gaussian smoothing at parameter δ, let $Y \sim \mu$ and $X_\delta \mid Y = y \sim N(y, (\delta/2)I)$. Then

$$\nabla \log P_\delta(h^2)(x) = \frac{2}{\delta}\big(\mathbb{E}[Y \mid X_\delta = x] - x\big).$$

At the matched choice $\delta = \varepsilon/2$, this becomes

$$\nabla \log P_{\varepsilon/2}(h^2)(x) = \frac{4}{\varepsilon}\big(\mathbb{E}[Y \mid X_{\varepsilon/2} = x] - x\big). \tag{4.24}$$

Thus the self-EOT and KDE scores use exactly the same displacement scaling, differing only through the conditional law.

The left panels of Figure 4.2 give the analogous comparison for score estimation. The first example is again chosen so that the leading KDE bias vanishes. We take

$$s(x) = \frac{1}{2}\cos x, \qquad x \in [-\pi/2, \pi/2],$$

whose score is

$$\nabla \log q(x) = -\tan x.$$

On the interior of the support,

$$\frac{p''}{p} = -1, \qquad \text{and hence} \qquad \left(\frac{p''}{p}\right)' = 0.$$

Thus the matched KDE score has zero first-order bias there, whereas the self-EOT correction need not vanish. We draw $N = 25{,}000$ observations and report the error only on $|x| \le 1$, away from the boundary singularities of the score. The second panel uses the five-dimensional smoothed-Laplace density

$$p(x) \propto \exp\left(-\sqrt{1+|x|^2}\right), \quad \nabla \log p(x) = -\frac{x}{\sqrt{1+|x|^2}},$$

as a control in which the leading terms are nonzero for both estimators. This experiment demonstrates improved performance of the KDE in the former example, and self-EOT in the latter where neither estimator vanishes at leading order.

4.3.3 Diffusion approximation

We next compare the associated Markov operators. Let

$$d_\varepsilon := P_\varepsilon(h^2)$$

be the Gaussian KDE. The symmetrically normalized diffusion-map operator, followed by row normalization, is

$$\mathsf{D}_\varepsilon u := \frac{d_\varepsilon^{-1/2} P_\varepsilon\left(h^2 d_\varepsilon^{-1/2} u\right)}{d_\varepsilon^{-1/2} P_\varepsilon\left(h^2 d_\varepsilon^{-1/2}\right)}. \tag{4.25}$$

The self-EOT transition operator is

$$\mathsf{T}_\varepsilon u(x) := \int \xi_\varepsilon(x,y) u(y)\, d\mu(y) = q_\varepsilon(x)^{-1/2} P_\varepsilon\left(h^2 q_\varepsilon^{-1/2} u\right)(x). \tag{4.26}$$

Our diffusion calculation follows the general blueprint of Landa and Cheng (2023a) and Cheng and Landa (2024), leveraging limiting behavior of the scalings to achieve the result. In Landa and Cheng (2023a), the required first-order control is imposed as an additional

assumption. Cheng and Landa (2024) instead analyzes a modified, approximately bi-stochastic scaling problem with an additional uniform lower-bound constraint on the scaling factors. Thus, their argument does not directly establish the expansion for the self-EOT solution. We bridge this gap in Theorem 4.1.

Proposition 4.1 (Diffusion-map/self-EOT approximation)**.** *Fix an integer $s \geq 0$ and assume the hypotheses of Theorem 4.1 with s there replaced by $s+2$. Then, for every $u \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d)$,*

$$\mathsf{T}_\varepsilon u = u + \frac{\varepsilon}{4}\left(\Delta u + 2\nabla \log h \cdot \nabla u\right) + o_{C^s_{\mathrm{ub}}(\mathbb{R}^d)}(\varepsilon), \tag{4.27}$$

$$\mathsf{D}_\varepsilon u = u + \frac{\varepsilon}{4}\left(\Delta u + 2\nabla \log h \cdot \nabla u\right) + o_{C^s_{\mathrm{ub}}(\mathbb{R}^d)}(\varepsilon). \tag{4.28}$$

In particular,

$$\|\mathsf{D}_\varepsilon u - \mathsf{T}_\varepsilon u\| = o(\varepsilon), \tag{4.29}$$

and the corresponding rescaled generators differ by $o(1)$ on every fixed $u \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d)$.

Proof. We first record a calculation that applies to both normalizations. Suppose

$$\rho_\varepsilon = 1 + \varepsilon a + o_{C^{s+2}_{\mathrm{ub}}}(\varepsilon), \qquad a \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d),$$

and define

$$\mathsf{K}^\rho_\varepsilon u := \rho_\varepsilon^{-1/2} P^h_\varepsilon\left(\rho_\varepsilon^{-1/2} u\right).$$

Smooth composition gives

$$\rho_\varepsilon^{-1/2} = 1 - \frac{\varepsilon}{2} a + o_{C^{s+2}_{\mathrm{ub}}}(\varepsilon).$$

Using linearity of P^h_ε, the fixed-input expansion from Lemma S7, strong continuity on the fixed function au, and uniform boundedness on the $o(\varepsilon)$ remainder,

$$P^h_\varepsilon\left(\rho_\varepsilon^{-1/2} u\right) = u - \frac{\varepsilon}{2} au + \frac{\varepsilon}{4} h^{-1}\Delta(hu) + o(\varepsilon).$$

Multiplying by the outer factor yields

$$\mathsf{K}^\rho_\varepsilon u = u + \frac{\varepsilon}{4} h^{-1}\Delta(hu) - \varepsilon a u + o(\varepsilon). \tag{4.30}$$

Setting $u = 1$ gives

$$\mathsf{K}^\rho_\varepsilon 1 = 1 + \varepsilon\left(\frac{1}{4}\frac{\Delta h}{h} - a\right) + o(\varepsilon).$$

Hence row normalization cancels the multiplication term:

$$\frac{\mathsf{K}_\varepsilon^\rho u}{\mathsf{K}_\varepsilon^\rho 1} = u + \frac{\varepsilon}{4}\left(\Delta u + 2\nabla \log h \cdot \nabla u\right) + o(\varepsilon), \tag{4.31}$$

where we used

$$h^{-1}\Delta(hu) = \Delta u + 2\nabla \log h \cdot \nabla u + \frac{\Delta h}{h} u.$$

For the self-EOT factor,

$$\rho_\varepsilon := \frac{q_\varepsilon}{h^2} = e^{-2\phi_\varepsilon} = 1 + \frac{\varepsilon}{4}\frac{\Delta h}{h} + o_{C^{s+2}_{\mathrm{ub}}}(\varepsilon).$$

Moreover,

$$\mathsf{K}_\varepsilon^\rho = \mathsf{T}_\varepsilon, \qquad \mathsf{T}_\varepsilon 1 = 1$$

exactly, so (4.31) gives (4.27).

For the diffusion-map factor,

$$\rho_\varepsilon := \frac{d_\varepsilon}{h^2} = \frac{P_\varepsilon(h^2)}{h^2} = 1 + \frac{\varepsilon}{4}\frac{\Delta(h^2)}{h^2} + o_{C^{s+2}_{\mathrm{ub}}}(\varepsilon).$$

By definition,

$$\mathsf{D}_\varepsilon u = \frac{\mathsf{K}_\varepsilon^\rho u}{\mathsf{K}_\varepsilon^\rho 1},$$

so (4.31) gives (4.28). Subtracting the two expansions proves (4.29). □

The common first-order generator is

$$\mathcal{L}_h u := \frac{1}{4}\left(\Delta u + \nabla \log h^2 \cdot \nabla u\right),$$

the Langevin generator with invariant density h^2 under the present time normalization. Thus the two one-step population operators agree through the full first-order term.

4.4 Approximating Self-EOT

Most small-ε analyses are motivated by self-EOT approximations of population quantities such as scores or diffusion processes. Our results reveal a useful converse: self-EOT quantities can themselves be approximated by simple Gaussian-smoothing formulas. We show that at the natural $O(\varepsilon)$ scale, two symmetric updates already achieve $o(\varepsilon)$ error.

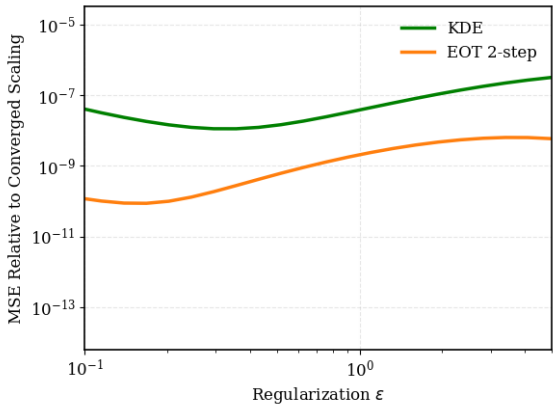


Figure 4.3: Approximation error of Gaussian KDE and two symmetric Sinkhorn updates relative to the fully converged self-EOT scaling, as a function of the regularization parameter ε. The two-step approximation substantially reduces the error of the initial KDE approximation, consistent with the first-order guarantee of Lemma 4.4.

4.4.1 Two symmetric Sinkhorn updates

The factor q_ε satisfies the marginal equation

$$q_\varepsilon^{1/2} = P_\varepsilon\left(h^2 q_\varepsilon^{-1/2}\right). \tag{4.32}$$

We compute this fixed point using a population version of the symmetric Sinkhorn iteration of Feydy et al. (2019b). In dual coordinates, their update averages the current potential with its Sinkhorn update. Writing this iteration in terms of the factor $q_{\varepsilon,t}$ gives

$$q_{\varepsilon,0} := 1, \qquad q_{\varepsilon,t+1} := \sqrt{q_{\varepsilon,t}}\, P_\varepsilon\left(\frac{h^2}{\sqrt{q_{\varepsilon,t}}}\right). \tag{4.33}$$

Its positive fixed points are exactly the solutions of (4.32). The first two iterates are

$$q_{\varepsilon,1} = P_\varepsilon(h^2), \qquad q_{\varepsilon,2} = \sqrt{P_\varepsilon(h^2)}\, P_\varepsilon\left(\frac{h^2}{\sqrt{P_\varepsilon(h^2)}}\right).$$

Thus the first iterate is exactly a Gaussian KDE, while the second uses this KDE to locally reweight the density before applying Gaussian smoothing again.

Lemma 4.4 (Two-step approximation)**.** *Under the assumptions of Theorem 4.1,*

$$q_{\varepsilon,2} = hP_\varepsilon h + o(\varepsilon) = q_\varepsilon + o(\varepsilon). \tag{4.34}$$

Proof. Set $r_\varepsilon := P_\varepsilon(h^2)/h^2$. Applying Lemma S7 with ℓ replaced by 2ℓ gives

$$r_\varepsilon = 1 + \frac{\varepsilon}{4}\frac{\Delta(h^2)}{h^2} + o(\varepsilon).$$

The smooth scalar maps $z \mapsto z^{\pm 1/2}$ near 1 therefore yield

$$\sqrt{P_\varepsilon(h^2)} = h + \frac{\varepsilon}{8}\frac{\Delta(h^2)}{h} + o(\varepsilon),$$
$$\frac{h^2}{\sqrt{P_\varepsilon(h^2)}} = h - \frac{\varepsilon}{8}\frac{\Delta(h^2)}{h} + o(\varepsilon).$$

Using $P_\varepsilon h = h + (\varepsilon/4)\Delta h + o(\varepsilon)$ and strong continuity of P_ε on the fixed first-order coefficient,

$$P_\varepsilon\left(\frac{h^2}{\sqrt{P_\varepsilon(h^2)}}\right) = h + \frac{\varepsilon}{4}\Delta h - \frac{\varepsilon}{8}\frac{\Delta(h^2)}{h} + o(\varepsilon).$$

Multiplication cancels the two $\Delta(h^2)$ terms and gives $q_{\varepsilon,2} = h^2 + (\varepsilon/4)h\Delta h + o(\varepsilon)$. The conclusion follows from Corollary 4.1. □

Lemma 4.4 shows that the complete first-order expansion of the converged factor is already present after two updates. On compact sets, the approximation transfers directly to the potential. Figure 4.3 illustrates this approximation numerically. We compare the fully converged self-EOT scaling with (i) the matched Gaussian KDE and (ii) the scaling obtained after two symmetric Sinkhorn updates. Across the displayed regularization parameters, the two-step approximation is substantially closer to the fully converged scaling than the one-step KDE approximation.

Corollary 4.3 (Two-step potential approximation on compact sets)**.** *Let $K \subset \mathbb{R}^d$ be compact and define*

$$f_{\varepsilon,2} := -\frac{\varepsilon}{2}\log\big(c_\varepsilon^{-1} q_{\varepsilon,2}\big).$$

Then

$$\|f_{\varepsilon,2} - f_\varepsilon\|_{C(K)} = o(\varepsilon^2). \tag{4.35}$$

Proof. By Lemma 4.4, $q_{\varepsilon,2} - q_\varepsilon = o(\varepsilon)$ uniformly on K. Both factors converge uniformly to h^2, which is bounded away from zero on K. The logarithm is therefore uniformly Lipschitz on their common range, so

$$\| \log q_{\varepsilon,2} - \log q_\varepsilon \|_{C(K)} = o(\varepsilon).$$

Since $f_\varepsilon = -(\varepsilon/2)\log(c_\varepsilon^{-1} q_\varepsilon)$, multiplication by $\varepsilon/2$ proves the claim. □

As a computational application, we follow Altschuler, Bach, et al. (2019) and use low-rank Nyström approximations of the Gaussian kernel to accelerate the symmetric Sinkhorn updates. Since our approximation result requires only two updates, we benchmark the low-rank methods directly against the dense two-step potential. We draw $X_i \sim N(0, I_d/d)$ for $d \in \{1, 3, 5\}$, fix $\varepsilon = 0.5$, and vary n from 5,000 to 300,000. For each setting, adaptive RP-Cholesky and uniform Nyström ranks are chosen to reproduce the dense two-step potential to relative L^2 error below 1%. The ratio decreases with n in each displayed dimension, illustrating the increasing computational advantage of the low-rank approximation at larger sample sizes.

4.4.2 Coreset approximation

The scaling approximation also transfers to the CO2 coreset objective of Kokot and Luedtke (2025). A uniform approximation of the rescaled kernel controls the quadratic objective over the entire sparse feasible set, so a minimizer based on the two-step kernel is asymptotically near-optimal for the fully converged objective.

Given observations $x_1, \ldots, x_n$ with empirical measure $\mu_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$, let

$$\mathcal{W}_{m,n} := \left\{ w \in [0,1]^n : \mathbf{1}^\top w = 1,\ \|w\|_0 \le m \right\}.$$

Each $w \in \mathcal{W}_{m,n}$ determines the coreset measure $\nu_w := \sum_{i=1}^n w_i \delta_{x_i}$.

Recall

$$\xi_\varepsilon(x,y) = \frac{p_\varepsilon(x-y)}{\sqrt{q_\varepsilon(x) q_\varepsilon(y)}}. \tag{4.36}$$

Define the two-step kernel

$$\xi_{\varepsilon,2}(x,y) := \frac{p_\varepsilon(x-y)}{\sqrt{q_{\varepsilon,2}(x) q_{\varepsilon,2}(y)}}.$$

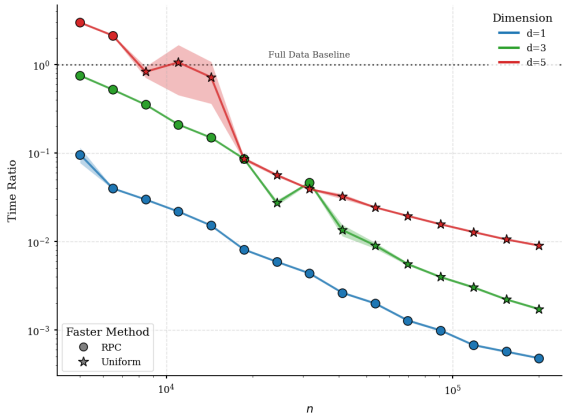


Figure 4.4: Runtime ratio of the faster low-rank two-step approximation to the dense two-step symmetric Sinkhorn computation, for $d \in \{1, 3, 5\}$ and $\varepsilon = 0.5$. Both low-rank methods are calibrated to relative L^2 potential error below 1%. Circles denote settings in which adaptive RP-Cholesky is faster and stars settings in which uniform Nyström is faster. The horizontal line at one is the dense two-step baseline; shaded regions show one standard error across runs.

Factoring out the common scaling c_ε, put

$$\bar{\xi}_\varepsilon := c_\varepsilon^{-1}\xi_\varepsilon = \frac{k_\varepsilon}{\sqrt{q_\varepsilon \otimes q_\varepsilon}}, \qquad \bar{\xi}_{\varepsilon,2} := c_\varepsilon^{-1}\xi_{\varepsilon,2} = \frac{k_\varepsilon}{\sqrt{q_{\varepsilon,2} \otimes q_{\varepsilon,2}}}. \tag{4.37}$$

Write

$$\bar{\Xi}_\varepsilon := [\bar{\xi}_\varepsilon(x_i, x_j)]_{i,j=1}^n, \qquad \bar{\Xi}_{\varepsilon,2} := [\bar{\xi}_{\varepsilon,2}(x_i, x_j)]_{i,j=1}^n,$$

and, with $a(w) := n^{-1}\mathbf{1} - w$, set

$$\mathcal{J}_\varepsilon(w) := a(w)^\top \bar{\Xi}_\varepsilon a(w), \qquad \mathcal{J}_{\varepsilon,2}(w) := a(w)^\top \bar{\Xi}_{\varepsilon,2} a(w). \tag{4.38}$$

Thus $\mathcal{J}_\varepsilon(w) = c_\varepsilon^{-1}\,\mathrm{MMD}^2_{\xi_\varepsilon}(\mu_n, \nu_w)$, and analogously for the two-step kernel.

Corollary 4.4 (Transfer of the two-step coreset)**.** *Assume the hypotheses of Theorem 4.1. Let all sample points belong to a fixed compact set $K \subset \mathbb{R}^d$. Then*

$$\sup_{w \in \mathcal{W}_{m,n}} |\mathcal{J}_{\varepsilon,2}(w) - \mathcal{J}_\varepsilon(w)| = o(\varepsilon). \tag{4.39}$$

Consequently, if

$$\widehat{w}_{\varepsilon,2} \in \operatorname*{argmin}_{w \in \mathcal{W}_{m,n}} \mathcal{J}_{\varepsilon,2}(w),$$

then

$$\mathcal{J}_\varepsilon(\widehat{w}_{\varepsilon,2}) \le \inf_{w \in \mathcal{W}_{m,n}} \mathcal{J}_\varepsilon(w) + o(\varepsilon). \tag{4.40}$$

Proof. Lemma 4.4 gives

$$\|q_{\varepsilon,2} - q_\varepsilon\|_{C(K)} = o(\varepsilon). \tag{4.41}$$

The same lemma and Corollary 4.1 give $q_{\varepsilon,2} \to h^2$ and $q_\varepsilon \to h^2$ uniformly on K. Since h is strictly positive and K is compact,

$$m_K := \min_{x \in K} h(x)^2 > 0.$$

Thus, for all sufficiently small ε,

$$q_\varepsilon(x) \ge \frac{m_K}{2}, \qquad q_{\varepsilon,2}(x) \ge \frac{m_K}{2}, \qquad x \in K.$$

The map $(r,s) \mapsto (rs)^{-1/2}$ is continuously differentiable on every compact subset of $(0,\infty)^2$. The mean-value theorem and (4.41) therefore imply

$$\sup_{x,y \in K} \left| \frac{1}{\sqrt{q_{\varepsilon,2}(x) q_{\varepsilon,2}(y)}} - \frac{1}{\sqrt{q_\varepsilon(x) q_\varepsilon(y)}} \right| = o(\varepsilon).$$

Since $0 < k_\varepsilon(x,y) \le 1$, multiplication by the Gaussian factor preserves this bound. Hence, with

$$\delta_\varepsilon := \sup_{x,y \in K} |\bar{\xi}_{\varepsilon,2}(x,y) - \bar{\xi}_\varepsilon(x,y)|,$$

we have $\delta_\varepsilon = o(\varepsilon)$.

For any feasible w, both $n^{-1}\mathbf{1}$ and w are probability vectors, so

$$\|a(w)\|_1 = \|n^{-1}\mathbf{1} - w\|_1 \le 2.$$

Consequently,

$$\begin{aligned} |\mathcal{J}_{\varepsilon,2}(w) - \mathcal{J}_\varepsilon(w)| &= \left| \sum_{i,j=1}^n a_i(w) a_j(w) \big(\bar{\Xi}_{\varepsilon,2} - \bar{\Xi}_\varepsilon\big)_{ij} \right| \\ &\le \delta_\varepsilon \left(\sum_{i=1}^n |a_i(w)| \right)^2 \le 4\delta_\varepsilon. \end{aligned}$$

The bound is independent of w, m, and n, proving (4.39).

The feasible set $\mathcal{W}_{m,n}$ is a finite union of compact simplices, so both minima are attained. Let $w_\varepsilon^\star$ minimize $\mathcal{J}_\varepsilon$. Then

$$\begin{aligned}\mathcal{J}_\varepsilon(\widehat{w}_{\varepsilon,2}) &\le \mathcal{J}_{\varepsilon,2}(\widehat{w}_{\varepsilon,2}) + 4\delta_\varepsilon \\ &\le \mathcal{J}_{\varepsilon,2}(w_\varepsilon^\star) + 4\delta_\varepsilon \\ &\le \mathcal{J}_\varepsilon(w_\varepsilon^\star) + 8\delta_\varepsilon.\end{aligned}$$

Since $\delta_\varepsilon = o(\varepsilon)$, this is (4.40). Multiplying either objective by the common positive constant c_ε does not alter its minimizers; the rescaling in (4.37) is used only to state the additive approximation at its natural $o(\varepsilon)$ scale. □

Figure 4.5 gives a qualitative illustration of the resulting coreset selections. We represent the 300×300 astronaut image as 90,000 points in five-dimensional location–color space,

$$(y/H, x/W, R, G, B),$$

and construct weighted coresets of size 2,500. We compare coresets formed using diffusion-map normalization, self-EOT scaling, and its two-step approximation. Marker color records the original pixel color and marker size the selected coreset weight. In the resulting compressed images, we see a stronger correspondence between the algorithms for smaller ε, aligning with our approximation guarantees.

4.5 Discussion

Previous works demonstrate that self-EOT potentials, barycentric projections, and normalized kernels contain meaningful score and diffusion information. Our results address the complementary goal of assessing the difference between these quantities and classical Gaussian estimators. Kernel density estimation and gradient-based mode estimation have long used Gaussian smoothing to estimate densities and their scores (Davis et al., 2011; Parzen, 1962; Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002). At the appropriate matched Gaussian scale, we demonstrate that the leading behavior agrees, and the expansion identifies the first term separating the two constructions. This interpretation is closely tied to the

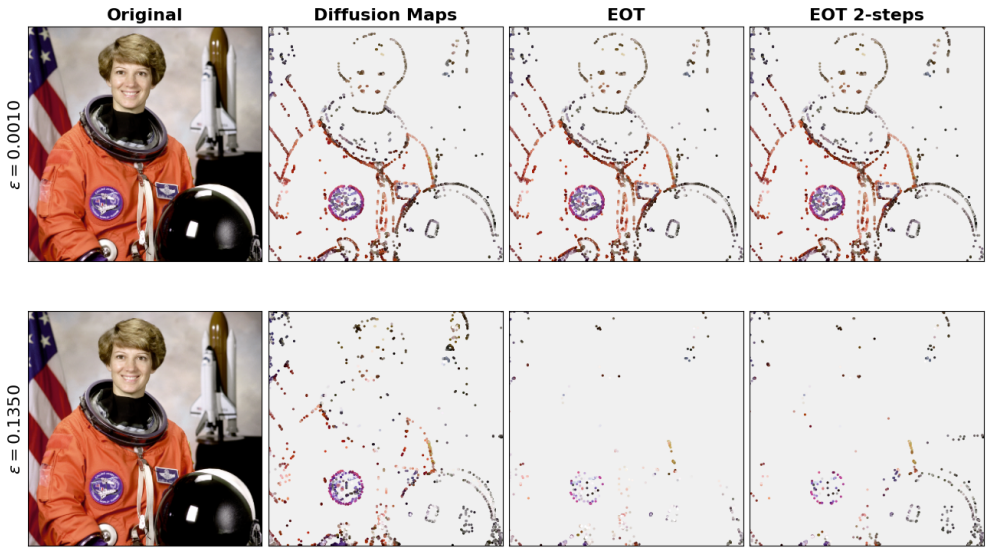


Figure 4.5: Weighted coresets of size 2,500 for the 90,000-pixel astronaut image, represented in five-dimensional location–RGB feature space. Columns compare diffusion-map scaling, self-EOT scaling, and the two-step self-EOT approximation; rows correspond to two Gaussian kernel scales. Candidate features are obtained from a rank-5,000 RP-Cholesky factorization before weighted subset selection. Marker color is inherited from the original pixel and marker size is proportional to its coreset weight.

proof strategy. Working directly with the Schrödinger equation and its h-transformed heat operator keeps the linearization in explicit form, allowing for a direct comparison to be made.

The results in this chapter concern population bias expansions, hence further research is needed to analyze the variance of these estimators to establish empirical estimation rates. Our comparisons should also not be interpreted as suggesting that either self-EOT or Gaussian KDE is optimal from the standpoint of population bias. Under the present Gaussian normalization, the spatial bandwidth is of order $\sqrt{\varepsilon}$, so the $O(\varepsilon)$ bias terms identified above correspond to the usual second-order smoothing bias. Higher-order local polynomial or bias corrected kernel methods can eliminate these leading terms and, under sufficient smoothness, achieve bias of smaller order than ε. The comparison here is instead between self-EOT and its natural Gaussian-smoothing counterpart, with the goal of identifying where their leading behavior agrees and differs. Natural next steps include joint (n, ε) asymptotics, data-driven selection of ε, and extensions under weaker regularity or support conditions.

Appendices

4.A Derivations for the Multivariate Gaussian Case (Section 4.2.2)

4.A.1 Derivation of quadratic form in (4.12)

We will show that $F(\varepsilon, a, b) = (F_1(\varepsilon, a), F_2(\varepsilon, a, b))$ for appropriately specified coordinate functions F_1, F_2. Recall that $\psi(\varepsilon, \phi) = \phi + \log[h^{-1} P_\varepsilon(he^\phi)]$. With the Gaussian root-density $h(x) \propto e^{-|x|^2/4}$ and the ansatz $\phi_{a,b}(x) = a|x|^2 + b$, the input to the heat operator is a Gaussian:

$$h(y)e^{\phi_{a,b}(y)} \propto e^b \exp\left[-\left(\tfrac{1}{4} - a\right)|y|^2\right].$$

Let $\gamma := \frac{1}{4} - a$. The heat operator P_ε convolves this input with the kernel $k_\varepsilon(x, y) \propto \exp(-\|x - y\|^2/\varepsilon)$. Evaluating the Gaussian integral $\int \exp\big(-\frac{\|x-y\|^2}{\varepsilon} - \gamma|y|^2\big)dy$ via square completion yields

$$P_\varepsilon(he^{\phi_{a,b}})(x) \propto e^b \, (1 + \varepsilon\gamma)^{-d/2} \exp\left(-\frac{\gamma}{1 + \varepsilon\gamma}|x|^2\right).$$

Applying the h-transform multiplication by $h^{-1}(x) \propto e^{|x|^2/4}$ and taking the logarithm, we obtain

$$\begin{aligned}\log P^h_\varepsilon(e^{\phi_{a,b}})(x) &= b - \frac{d}{2}\log(1 + \varepsilon\gamma) + \left(\frac{1}{4} - \frac{\gamma}{1 + \varepsilon\gamma}\right)|x|^2 \\ &= b - \frac{d}{2}\log\big(1 + \varepsilon(\tfrac{1}{4} - a)\big) + \frac{a + \varepsilon(\frac{1}{16} - \frac{a}{4})}{1 + \varepsilon(\frac{1}{4} - a)}|x|^2.\end{aligned}$$

Finally, adding $\phi_{a,b}(x) = a|x|^2 + b$ to this expression yields

$$\psi(\varepsilon, \phi_{a,b})(x) = \left(a + \frac{a + \varepsilon(\frac{1}{16} - \frac{a}{4})}{1 + \varepsilon(\frac{1}{4} - a)}\right)|x|^2 + \left(2b - \frac{d}{2}\log\big(1 + \varepsilon(\tfrac{1}{4} - a)\big)\right) \tag{S1}$$

$$=: F_1(\varepsilon, a)|x|^2 + F_2(\varepsilon, a, b). \tag{S2}$$

4.A.2 Derivations for Remark 4.1

Derivation of exact form of potentials for all $\varepsilon > 0$. We will show that

$$f_\varepsilon(x) = \alpha_\varepsilon|x|^2 + \beta_\varepsilon \quad \forall \varepsilon > 0, \tag{S3}$$

with $\alpha_\varepsilon := 1 - \rho_\varepsilon$ and $\beta_\varepsilon := -\frac{\varepsilon d}{4}\log(1 - \rho_\varepsilon^2)$ for $\rho_\varepsilon := [(16 + \varepsilon^2)^{1/2} - \varepsilon]/4$. Using Eq. 3 in Bunne et al., 2023, the self-EOT plan is

$$\pi_\varepsilon = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} I_d & \rho_\varepsilon I_d \\ \rho_\varepsilon I_d & I_d \end{bmatrix}\right) \quad \text{for} \quad \rho_\varepsilon := \frac{\sqrt{16 + \varepsilon^2} - \varepsilon}{4}. \tag{S4}$$

The gradient of the potential is given by barycentric projection as $\nabla f_\varepsilon(x) = 2(x - \mathbb{E}_{\pi_\varepsilon}[Y|X = x]) = 2(1 - \rho_\varepsilon)x$. Integrating this field yields

$$f_\varepsilon(x) = f_\varepsilon(0) + \int_0^1 \langle \nabla f_\varepsilon(tx), x\rangle \, dt = f_\varepsilon(0) + (1 - \rho_\varepsilon)|x|^2,$$

solving for $f_\varepsilon(0)$ by plugging (S4) into (4.2) with $(x, y) = (0, 0)$ yields the exact potential:

$$f_\varepsilon(x) = (1 - \rho_\varepsilon)|x|^2 - \varepsilon\frac{d}{4}\log(1 - \rho_\varepsilon^2) =: \alpha_\varepsilon|x|^2 + \beta_\varepsilon. \tag{S5}$$

Tightening the $o(\varepsilon^2)$ scalar term in (4.11) to $O(\varepsilon^4)$. We show that knowing the explicit form of the potentials from (S5) makes it possible to tighten the expansion derived from the implicit function theorem (Eq. 4.11). Concretely, we will show that

$$f_\varepsilon(x) = \underbrace{-\varepsilon\frac{d}{4}\log(\pi\varepsilon) - \varepsilon\log h(x) - \varepsilon^2\frac{\Delta(h)(x)}{8h(x)}}_{\text{dominant terms on RHS of (4.4)}} + (1+|x|^2)\,O(\varepsilon^4), \tag{S6}$$

To do this, we expand (S5) as $\varepsilon \to 0$. Noting that $\rho_\varepsilon = 1 - \varepsilon/4 + \varepsilon^2/32 + O(\varepsilon^4)$, we find $(1-\rho_\varepsilon) = \varepsilon/4 - \varepsilon^2/32 + O(\varepsilon^4)$. Similarly, expanding the intercept term yields $-\frac{\varepsilon d}{4}\log(1-\rho_\varepsilon^2) = -\frac{\varepsilon d}{4}\log(\varepsilon/2) + \frac{\varepsilon^2 d}{16} + O(\varepsilon^4)$. Plugging these coefficients back into (S5) recovers

$$f_\varepsilon(x) = \left(-\frac{\varepsilon d}{4}\log\left(\frac{\varepsilon}{2}\right) + \frac{\varepsilon^2 d}{16}\right) + \left(\frac{\varepsilon}{4} - \frac{\varepsilon^2}{32}\right)|x|^2 + (1+|x|^2)\,O(\varepsilon^4).$$

Finally, since h is the root-density of $\mu = N(0_d, I_d)$, we have that $\log h(x) = -\frac{d}{4}\log(2\pi) - \frac{|x|^2}{4}$ and $\frac{\Delta h(x)}{h(x)} = \frac{|x|^2}{4} - \frac{d}{2}$. Plugging this in above yields (S6).

4.B Entropic Cost Expansion

As a brief aside, we demonstrate how our expansions recover the result of Conforti and Tamanini (2021).

Corollary S1 (Self-EOT cost expansion)**.** *Assume the hypotheses of Theorem 4.1 with $s = 0$ and, in addition, $\log h \in L^1(\mu)$. Define*

$$\mathrm{Ent}(\mu) := -\int \log(h^2)\,d\mu, \qquad \mathcal{I}(\mu) := \int |\nabla \log(h^2)|^2\,d\mu = 4\int |\nabla h|^2\,dx. \tag{S7}$$

Then

$$\mathrm{OT}_\varepsilon(\mu,\mu) = -\varepsilon\frac{d}{2}\log(\pi\varepsilon) + \varepsilon\,\mathrm{Ent}(\mu) + \frac{\varepsilon^2}{16}\mathcal{I}(\mu) + o(\varepsilon^2). \tag{S8}$$

Proof. At the optimum, (4.2) gives $\|x-y\|^2 + \varepsilon\log\xi_\varepsilon(x,y) = f_\varepsilon(x) + f_\varepsilon(y)$. Integrating against π_ε yields the exact identity $\mathrm{OT}_\varepsilon(\mu,\mu) = 2\int f_\varepsilon\,d\mu$. Integrating Theorem 4.1 gives

$$2\int f_\varepsilon\,d\mu = -\varepsilon\frac{d}{2}\log(\pi\varepsilon) - 2\varepsilon\int \log h\,d\mu - \frac{\varepsilon^2}{4}\int \frac{\Delta h}{h}\,d\mu + o(\varepsilon^2).$$

The assumptions give $\nabla h = h\nabla\log h \in L^2$ and $\Delta h = h(\Delta\log h + |\nabla\log h|^2) \in L^2$. The standard Sobolev integration-by-parts identity therefore yields

$$\int \frac{\Delta h}{h}\,d\mu = \int h\Delta h\,dx = -\int |\nabla h|^2\,dx = -\frac{1}{4}\mathcal{I}(\mu).$$

Together with $-2\int \log h\,d\mu = \mathrm{Ent}(\mu)$, this proves (S8). □

4.C Supporting Arguments for Theorem 4.1

Throughout this appendix, $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ and $\|\cdot\|$ are as in (4.13), $\ell := \log h$, M_a denotes multiplication by a, and $\tau_y u(x) := u(x+y)$. We set $P_0 = I$ and $P_0^h = I$. For $\varepsilon > 0$,

$$p_\varepsilon(y) = (\pi\varepsilon)^{-d/2} e^{-|y|^2/\varepsilon}, \qquad P_\varepsilon u(x) = \int_{\mathbb{R}^d} p_\varepsilon(y) u(x+y)\, dy.$$

Lemma S1 (Elementary properties of $C^s_{\mathrm{ub}}(\mathbb{R}^d)$)**.** *For every integer $s \geq 0$, translations act isometrically on $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ and $\|\tau_y u - u\| \to 0$ as $y \to 0$. Moreover, there is a constant $C_s < \infty$ such that*

$$\|uv\| \leq C_s \|u\| \|v\|, \qquad \|M_u\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \leq C_s \|u\|. \tag{S9}$$

In particular, $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ is a Banach algebra.

Proof. Translations commute with every derivative and preserve the supremum norm, so $\|\tau_y u\| = \|u\|$. The convergence $\|\tau_y u - u\| \to 0$ is exactly the uniform-translation condition in the definition of $C^s_{\mathrm{ub}}(\mathbb{R}^d)$.

For a multi-index α, Leibniz's rule gives

$$\partial^\alpha (uv) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial^\beta u)(\partial^{\alpha-\beta} v).$$

Taking supremum norms, summing over $|\alpha| \leq s$, and absorbing the finitely many binomial coefficients into C_s proves the first estimate in (S9). It also proves the multiplier estimate. Each term in the Leibniz formula is uniformly continuous because it is a product of bounded uniformly continuous functions, so $uv \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$. Finally, $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ is closed in C^s_b under the norm $\|\cdot\|$: if $u_n \to u$ in C^s_b, then $\|\tau_y u - u\| \leq 2\|u - u_n\| + \|\tau_y u_n - u_n\|$. For any $\varepsilon > 0$, choose n such that $\|u - u_n\| \leq \varepsilon/4$, and then choose $r_n > 0$ so that $\|\tau_y u_n - u_n\| \leq \varepsilon/2$ whenever $\|y\| \leq r_n$. This verifies that $u \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$. Since $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ is a closed subspace of the Banach space $C^s_b(\mathbb{R}^d)$, it is complete. □

Lemma S2 (Regularity inherited from the log-gradient)**.** *Let $h > 0$ satisfy $h \in L^2(\mathbb{R}^d)$ and put $\ell = \log h$. If $\nabla \ell \in C^m_{\mathrm{ub}}(\mathbb{R}^d)$, then $h \in C^{m+1}_{\mathrm{ub}}(\mathbb{R}^d)$. In particular, under the hypotheses of Theorem 4.1,*

$$h, h^2 \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d), \qquad \frac{\Delta h}{h} = \Delta \ell + |\nabla \ell|^2 \in C^s_{\mathrm{ub}}(\mathbb{R}^d), \qquad \frac{\Delta(h^2)}{h^2} = 2\Delta\ell + 4|\nabla\ell|^2 \in C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{S10}$$

Proof. Set $L := \|\nabla \ell\|_\infty$. For $y \in B(x,1)$, the mean-value theorem gives $\ell(y) \geq \ell(x) - L$, and hence

$$\|h\|_{L^2}^2 \geq \int_{B(x,1)} h(y)^2 \, dy \geq |B(0,1)| e^{-2L} h(x)^2.$$

Thus h is bounded, with a bound depending only on L and $\|h\|_{L^2}$. In particular, $\nabla h = h \nabla \ell$ is bounded, so h is uniformly continuous.

For every nonzero multi-index α, repeated differentiation of $h = e^\ell$ gives

$$\partial^\alpha h = h \, \mathcal{P}_\alpha\big((\partial^\beta \ell)_{0<|\beta|\leq|\alpha|}\big), \tag{S11}$$

where $\mathcal{P}_\alpha$ is a universal polynomial with no constant term. This is proved inductively: differentiating the right side either produces $\partial_j h = h \partial_j \ell$ or differentiates one of the entries of the polynomial. Since $\nabla \ell \in C^m_{\mathrm{ub}}(\mathbb{R}^d)$, all derivatives of ℓ of orders $1, \ldots, m+1$ are bounded and uniformly continuous. This verifies $h \in C^{m+1}_{\mathrm{ub}}(\mathbb{R}^d)$, and $h^2 \in C^{m+1}_{\mathrm{ub}}(\mathbb{R}^d)$ by Lemma S1.

More generally, (S11) implies that

$$\frac{\partial^\alpha h}{h} \in C^{m+1-|\alpha|}_{\mathrm{ub}}(\mathbb{R}^d), \qquad 1 \leq |\alpha| \leq m+1.$$

Indeed, the ratio is a polynomial in derivatives of ℓ of orders at most $|\alpha|$, and taking a further k derivatives requires derivatives of ℓ only through order $|\alpha| + k$. In particular,

$$\frac{\nabla h}{h} = \nabla \ell, \qquad \frac{\Delta h}{h} = \Delta \ell + |\nabla \ell|^2.$$

Similarly, because $h^2 = e^{2\ell}$,

$$\frac{\Delta(h^2)}{h^2} = 2\Delta \ell + 4|\nabla \ell|^2.$$

If $\nabla \ell \in C^{s+1}_{\mathrm{ub}}(\mathbb{R}^d)$, all derivatives appearing after applying up to s further derivatives to the last two identities remain bounded and uniformly continuous. □

4.C.1 Composition and heat-kernel estimates

Lemma S3 (Smooth composition on $C^s_{\mathrm{ub}}(\mathbb{R}^d)$)**.** *Let $J \subset \mathbb{R}$ be open and let $F \in C^{s+2}(J)$. On every bounded subset of $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ whose ranges lie in a fixed compact subset of J, composition $F_\circ(u) := F \circ u$ is continuously Fréchet differentiable, with*

$$D(F_\circ)(u)[v] = F'(u)v.$$

Its differential is locally Lipschitz in operator norm. In particular, this applies to $F(z) = e^z$ on bounded sets and to $F(z) = \log z$ and $F(z) = z^{-1}$ on sets bounded away from zero.

Proof. For $|\alpha| \le s$, the multivariate chain rule expresses $\partial^\alpha(F(u))$ as a finite sum of terms

$$F^{(k)}(u) \prod_{j=1}^{k} \partial^{\alpha_j} u, \qquad \alpha_1 + \cdots + \alpha_k = \alpha, \quad |\alpha_j| \ge 1.$$

On the indicated set, every derivative of F through order $s+2$ is bounded on a common compact interval. Lemma S1 therefore bounds $\|F(u)\|$ uniformly on bounded sets. The same formula, together with translation continuity of the factors, shows that $F(u) \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$.

The following Taylor identities are applied pointwise in the scalar target. For $\|v\|$ sufficiently small, every pointwise segment $u(x) + tv(x)$, $0 \le t \le 1$, remains in one fixed compact subset of J. Pointwise Taylor expansion with integral remainder gives

$$F(u+v) - F(u) - F'(u)v = v^2 \int_0^1 (1-t) F''(u+tv)\, dt.$$

Since $F \in C^{s+2}(J)$, the functions $F''(u+tv)$ are uniformly bounded in C^s_{ub} on this neighborhood. Lemma S1 therefore gives

$$\|F(u+v) - F(u) - F'(u)v\| \le C\|v\|^2.$$

This proves Fréchet differentiability. We now seek to prove that the differential is locally Lipschitz, i.e. that $\|M_{F'(u+v)} - M_{F'(u)}\| \le C\|v\|$, for all v in a neighborhood of u. By Lemma S1, it again suffices to show $\|F'(u+v) - F'(u)\| \le C\|v\|$. Similarly, Taylor expansion yields

$$F'(u+v) - F'(u) = v \int_0^1 F''(u+tv)\, dt.$$

and the claim follows similarly. □

Lemma S4 (Heat semigroup on $C^s_{\mathrm{ub}}(\mathbb{R}^d)$)**.** *For every integer $s \ge 0$ and every $\varepsilon \ge 0$, $P_\varepsilon : C^s_{\mathrm{ub}}(\mathbb{R}^d) \to C^s_{\mathrm{ub}}(\mathbb{R}^d)$ and*

$$\|P_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \le 1. \tag{S12}$$

The family is strongly continuous at zero and operator-norm continuous on $(0, \infty)$. If $u \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d)$, then

$$\frac{P_\varepsilon u - u}{\varepsilon} \longrightarrow \frac{1}{4}\Delta u \qquad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{S13}$$

Proof. Differentiation under the integral gives $\partial^\alpha P_\varepsilon u = P_\varepsilon \partial^\alpha u$ for $|\alpha| \le s$, hence differentiation is given by convolution of derivatives against the Gaussian density p_ε, implying contractivity in sup-norm. Moreover,

$$\|\tau_z P_\varepsilon u - P_\varepsilon u\| \le \|\tau_z u - u\|,$$

so $P_\varepsilon u \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$. We further have

$$|\partial^\alpha P_\varepsilon u(x) - \partial^\alpha u(x)| = \left| \int p_\varepsilon(y) \partial^\alpha u(x+y) - \partial^\alpha u(x)\, dy \right| \le \int p_\varepsilon(y) |\partial^\alpha \tau_y u(x) - \partial^\alpha u(x)|\, dy.$$

Hence, for $\omega_u(r) := \sup_{|y| \le r} \|\tau_y u - u\|$, we have

$$\begin{aligned} \|P_\varepsilon u - u\| \le \int p_\varepsilon(y) \|\tau_y u - u\|\, dy &\le \omega_u(r) + \int_{|y|>r} p_\varepsilon(y) \|\tau_y u - u\|\, dy \\ &\le \omega_u(r) + 2\|u\| \int_{|y|>r} p_\varepsilon(y)\, dy. \end{aligned}$$

Hence,

$$\limsup_{\varepsilon \to 0} \|P_\varepsilon u - u\| \le \omega_u(r) \overset{r \to 0}{\longrightarrow} 0.$$

If $\varepsilon, \delta > 0$, then

$$|\partial^\alpha (P_\varepsilon - P_\delta) u| = \left| \int [p_\varepsilon(y) - p_\delta(y)] \partial^\alpha \tau_y u\, dy \right| \le \|\partial^\alpha u\|_\infty \int |p_\varepsilon(y) - p_\delta(y)|\, dy.$$

This yields

$$\|P_\varepsilon - P_\delta\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \le \|p_\varepsilon - p_\delta\|_{L^1},$$

and the right side tends to zero as $\varepsilon \to \delta$, proving operator-norm continuity away from zero.

For the generator statement, fix $|\alpha| \le s$ and apply Taylor's formula to $v = \partial^\alpha u$:

$$v(x+y) = v(x) + \nabla v(x) \cdot y + \frac{1}{2} y^\top D^2 v(x) y + R_v(x,y),$$

where

$$R_v(x,y) = \int_0^1 (1-t) y^\top \big(D^2 v(x+ty) - D^2 v(x) \big) y\, dt.$$

Thus $\sup_x |R_v(x,y)| \le |y|^2 \omega_{D^2 v}(|y|)$, where the modulus tends to zero at the origin. The Gaussian is centered and has covariance $(\varepsilon/2) I_d$, so integration gives

$$\partial^\alpha (P_\varepsilon u - u) = \frac{\varepsilon}{4} \partial^\alpha \Delta u + \int p_\varepsilon(y) R_v(\cdot, y)\, dy.$$

After the change of variables $y = \sqrt{\varepsilon} z$, the last term divided by ε is bounded by

$$\int p_1(z)|z|^2 \omega_{D^2 v}(\sqrt{\varepsilon}|z|)\, dz,$$

which tends to zero by dominated convergence. Summing over $|\alpha| \le s$ proves (S13). □

4.C.2 The h-transform

Lemma S5 (Uniform h-transform bounds and operator approximation)**.** *Suppose* $\nabla \ell \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$. *For every* $\varepsilon_0 > 0$*:*

1. $P^h_\varepsilon : C^s_{\mathrm{ub}}(\mathbb{R}^d) \to C^s_{\mathrm{ub}}(\mathbb{R}^d)$ *is bounded uniformly for* $0 < \varepsilon \le \varepsilon_0$*;*

2. $\eta_\varepsilon := P^h_\varepsilon 1$ *is bounded in* $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ *and there is* $c_{\varepsilon_0} > 0$ *such that* $\inf_x \eta_\varepsilon(x) \ge c_{\varepsilon_0}$ *for all* $0 < \varepsilon \le \varepsilon_0$*;*

3. $$\|P^h_\varepsilon - P_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \longrightarrow 0 \qquad (\varepsilon \downarrow 0). \tag{S14}$$

Consequently, $P^h_\varepsilon u \to u$ *in* $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ *for every fixed* $u \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$*, and* $\eta_\varepsilon \to 1$ *in* $C^s_{\mathrm{ub}}(\mathbb{R}^d)$.

Proof. For $y \in \mathbb{R}^d$, put

$$R_y(x) := \frac{h(x+y)}{h(x)} = e^{\ell(x+y)-\ell(x)}.$$

If $L := \|\nabla \ell\|_\infty$, the mean-value theorem gives

$$e^{-L|y|} \le R_y(x) \le e^{L|y|} \qquad (x, y \in \mathbb{R}^d). \tag{S15}$$

For every nonzero multi-index β, Faà di Bruno's formula gives

$$\partial^\beta_x R_y(x) = R_y(x)\, \mathcal{Q}_\beta\big((\partial^\gamma \ell(x+y) - \partial^\gamma \ell(x))_{0<|\gamma|\le|\beta|}\big), \tag{S16}$$

where $\mathcal{Q}_\beta$ is a universal polynomial with no constant term. Indeed, every positive-order derivative of $\ell(x+y) - \ell(x)$ is exactly such a difference. Define

$$b_s(y) := \|R_y - 1\| := \|R_y - 1\|_\infty + \sum_{0<|\beta|\le s} \|\partial^\beta R_y\|_\infty,$$

with the sum omitted when $s = 0$. Since the derivatives $\partial^\gamma \ell$, $1 \le |\gamma| \le s$, are bounded and uniformly continuous, it follows from (S16) that

$$\|\partial^\beta R_y\|_\infty \le C_\beta \|R_y\|_\infty \|\tau_y \nabla \ell - \nabla \ell\| \to 0.$$

In combination with (S15) we have

$$b_s(y) \longrightarrow 0 \quad (y \to 0), \qquad b_s(y) \le C_s e^{L|y|} \quad (y \in \mathbb{R}^d), \tag{S17}$$

where we bound the zeroth-order term by $|R_y - 1| \le e^{L|y|} - 1$.

We now seek to translate these regularity properties of the ratio to the integral operator P_ε^h. We can express

$$P_\varepsilon^h u(x) = \int p_\varepsilon(y) R_y(x) u(x+y)\, dy. \tag{S18}$$

Differentiating under the integral and applying Leibniz's rule yields, for $|\alpha| \le s$,

$$\partial^\alpha P_\varepsilon^h u(x) = \sum_{\beta \le \alpha} \binom{\alpha}{\beta} \int p_\varepsilon(y) \partial^\beta R_y(x) \partial^{\alpha-\beta} u(x+y)\, dy. \tag{S19}$$

The bounds above justify differentiation by dominated convergence. They also give

$$\|P_\varepsilon^h u\| \le C_s \|u\| \int p_\varepsilon(y) e^{L|y|}\, dy = C_s \|u\| \int p_1(z) e^{L\sqrt{\varepsilon}|z|}\, dz \le C_{s,L,\varepsilon_0} \|u\|.$$

It remains to verify uniform continuity of $P_\varepsilon^h u$ and its derivatives through order s. Define

$$\Omega_s(r) := \max_{1 \le |\gamma| \le s+1} \sup_{|z| \le r} \|\tau_z \partial^\gamma \ell - \partial^\gamma \ell\|_\infty,$$

with only $\nabla \ell$ used when $s = 0$. Then $\Omega_s(r) \to 0$ as $r \downarrow 0$.

To control the zeroth-order ratio, put

$$A := \ell(x+y+z) - \ell(x+z), \qquad B := \ell(x+y) - \ell(x).$$

Both $|A|$ and $|B|$ are bounded by $L|y|$, and

$$A - B = \int_0^1 \big[\nabla \ell(x+z+ty) - \nabla \ell(x+ty)\big]^\top y\, dt.$$

Hence

$$\|\tau_z R_y - R_y\|_\infty \le C e^{L|y|} |y| \Omega_s(|z|). \tag{S20}$$

For $0 < |\beta| \le s$, write

$$D_\gamma(x,y) := \partial^\gamma \ell(x+y) - \partial^\gamma \ell(x).$$

The vectors $(D_\gamma)_{0<|\gamma|\le|\beta|}$ range in a fixed bounded set. Since the polynomial $\mathcal{Q}_\beta$ is Lipschitz on bounded sets and

$$\|\tau_z D_\gamma(\cdot,y) - D_\gamma(\cdot,y)\|_\infty \le 2\Omega_s(|z|),$$

combining (S16) with (S20) gives, also for $\beta = 0$,

$$\|\tau_z \partial^\beta R_y - \partial^\beta R_y\|_\infty \le C e^{L|y|}(1+|y|)\Omega_s(|z|), \qquad |\beta| \le s. \tag{S21}$$

Applying this estimate in (S19), and also translating the factor $\partial^{\alpha-\beta}u(x+y)$, yields

$$\|\tau_z \partial^\alpha P_\varepsilon^h u - \partial^\alpha P_\varepsilon^h u\|_\infty \le C \int p_\varepsilon(y) e^{L|y|}(1+|y|)\,dy\big[\Omega_s(|z|)\|u\| + \|\tau_z u - u\|\big].$$

The Gaussian moment is finite and the bracket tends to zero as $z \to 0$. Thus every derivative in (S19) is uniformly continuous. Hence $P_\varepsilon^h : C^s_{\mathrm{ub}}(\mathbb{R}^d) \to C^s_{\mathrm{ub}}(\mathbb{R}^d)$ and the first assertion holds.

Similarly to (S19), we have

$$\|(P_\varepsilon^h - P_\varepsilon)u\| = \left\| \int p_\varepsilon(y)(R_y(\cdot) - 1)u(\cdot + y)\,dy \right\| \le C_s \|u\| \int p_\varepsilon(y) b_s(y)\,dy. \tag{1}$$

After the change of variables $y = \sqrt{\varepsilon} z$, the integral in (1) becomes

$$\int p_1(z) b_s(\sqrt{\varepsilon} z)\,dz.$$

It tends to zero by (S17) and dominated convergence, because $p_1(z)e^{L\sqrt{\varepsilon_0}|z|}$ is integrable. This proves (S14).

Finally, (S15) gives the pointwise estimates

$$\int p_\varepsilon(y) e^{-L|y|}\,dy \le \eta_\varepsilon(x) \le \int p_\varepsilon(y) e^{L|y|}\,dy.$$

For $0 < \varepsilon \le \varepsilon_0$, the left side has a strictly positive uniform lower bound and the right side a finite uniform upper bound. The $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ bound follows from the first assertion with $u = 1$. Moreover, $P_\varepsilon 1 = 1$, so (1) with $u = 1$ gives $\|\eta_\varepsilon - 1\| \to 0$. Combining (S14) with strong continuity of P_ε proves $P_\varepsilon^h u \to u$ for each fixed u. □

Lemma S6 (Continuity in the heat parameter)**.** *Under the assumptions of Lemma S5, the map $\varepsilon \mapsto P_\varepsilon^h$ is operator-norm continuous on $(0,\infty)$. At 0 it is strongly continuous.*

Proof. Fix $0 < a < b < \infty$. The map $\varepsilon \mapsto p_\varepsilon(y)$ is continuously differentiable, with

$$\partial_\varepsilon p_\varepsilon(y) = \left(-\frac{d}{2\varepsilon} + \frac{|y|^2}{\varepsilon^2}\right) p_\varepsilon(y).$$

For $\varepsilon \in [a,b]$,

$$|\partial_\varepsilon p_\varepsilon(y)| \le C_{a,b}(1+|y|^2)p_b(y).$$

Using (S19) and the bounds from Lemma S5,

$$\|P_\varepsilon^h - P_\delta^h\|_{\mathcal{L}(C^s_{\mathrm{ub}})} \le C \int_{\mathbb{R}^d} |p_\varepsilon(y) - p_\delta(y)| e^{L|y|}\, dy.$$

The fundamental theorem of calculus in ε therefore gives

$$\|P_\varepsilon^h - P_\delta^h\|_{\mathcal{L}(C^s_{\mathrm{ub}})} \le C_{a,b}|\varepsilon - \delta| \int_{\mathbb{R}^d} (1+\|y\|^2) p_b(y) e^{L\|y\|}\, dy.$$

The final integral is finite, proving operator-norm continuity on $(0,\infty)$. Strong continuity at zero was proved in Lemma S5. □

Lemma S7 (First-order h-transform expansion)**.** *Suppose $\nabla \ell \in C^{s+1}_{\mathrm{ub}}(\mathbb{R}^d)$. For every $u \in C^{s+2}_{\mathrm{ub}}(\mathbb{R}^d)$,*

$$\frac{P_\varepsilon^h u - u}{\varepsilon} \longrightarrow \frac{1}{4}\frac{\Delta(hu)}{h} = \frac{1}{4}\left[\Delta u + 2\nabla \ell \cdot \nabla u + (\Delta \ell + |\nabla \ell|^2)u\right] \quad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{S22}$$

In particular,

$$\frac{\psi(\varepsilon,0)}{\varepsilon} = \frac{\log(P_\varepsilon h/h)}{\varepsilon} \longrightarrow \frac{1}{4}\frac{\Delta h}{h} \quad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d). \tag{S23}$$

Proof. Throughout the proof we will make use of many of the same bounds and arguments as in Lemma S5. The goal of this argument is to convert a leading order Taylor expansion of the integrand into the desired limiting result. Define

$$F_u(x,y) := R_y(x)u(x+y) = \frac{h(x+y)u(x+y)}{h(x)}, \qquad G_\alpha(x,y) := \partial_x^\alpha F_u(x,y), \quad |\alpha| \le s.$$

Then

$$P_\varepsilon^h u(x) = \int_{\mathbb{R}^d} p_\varepsilon(y) F_u(x,y)\, dy.$$

We first control the second y-derivatives of G_α. Repeated product and chain rules, together with

$$\partial_{x_j} R_y(x) = R_y(x)\big[\partial_j \ell(x+y) - \partial_j \ell(x)\big], \qquad \partial_{y_j} R_y(x) = R_y(x)\partial_j \ell(x+y),$$

show that, for multi-indices α, κ,

$$\partial_y^\kappa \partial_x^\alpha F_u(x,y) = R_y(x)\, \mathcal{Q}_{\alpha,\kappa} \begin{pmatrix} (\partial^\gamma \ell(x))_{1\le|\gamma|\le|\alpha|}, \\ (\partial^\gamma \ell(x+y))_{1\le|\gamma|\le|\alpha|+|\kappa|}, \\ (\partial^\nu u(x+y))_{|\nu|\le|\alpha|+|\kappa|} \end{pmatrix}. \tag{S24}$$

where $\mathcal{Q}_{\alpha,\kappa}$ is a universal polynomial whose monomials are linear in the derivatives of u.

Let $L := \|\nabla \ell\|_\infty$. Since

$$|R_y(x)| \le e^{L|y|},$$

and all positive-order derivatives of ℓ through order $s+2$ are bounded, taking $|\kappa| = 2$ gives

$$\sup_x |D_y^2 G_\alpha(x,y)| \le C\|u\|_{C^{s+2}_{\mathrm{ub}}} e^{L|y|}. \tag{S25}$$

Combining (S24) with uniform continuity of the derivatives of ℓ and u, and with $\|R_y - 1\|_\infty \le e^{L|y|} - 1$, implies that

$$\omega_\alpha(r) := \sup_{\substack{x\in\mathbb{R}^d \\ |y|\le r}} |D_y^2 G_\alpha(x,y) - D_y^2 G_\alpha(x,0)| \longrightarrow 0 \qquad (r \downarrow 0). \tag{S26}$$

Taylor's formula in the y variable gives

$$G_\alpha(x,y) = G_\alpha(x,0) + \nabla_y G_\alpha(x,0)^\top y + \frac{1}{2} y^\top D_y^2 G_\alpha(x,0) y + \mathcal{R}_\alpha(x,y),$$

where

$$\mathcal{R}_\alpha(x,y) = \int_0^1 (1-t) y^\top \big[D_y^2 G_\alpha(x,ty) - D_y^2 G_\alpha(x,0)\big] y \, dt.$$

Consequently,

$$\sup_x |\mathcal{R}_\alpha(x,y)| \le \min\left\{ \frac{1}{2}|y|^2 \omega_\alpha(|y|), C\|u\|_{C^{s+2}_{\mathrm{ub}}} |y|^2 e^{L|y|} \right\}. \tag{S27}$$

Differentiating under the integral,

$$\partial^\alpha P_\varepsilon^h u(x) = \int_{\mathbb{R}^d} p_\varepsilon(y) G_\alpha(x,y)\, dy.$$

The Gaussian is centered and has covariance $(\varepsilon/2)I_d$, so

$$\int y\, p_\varepsilon(y)\, dy = 0, \qquad \int y_i y_j p_\varepsilon(y)\, dy = \frac{\varepsilon}{2}\delta_{ij}.$$

Moreover, $G_\alpha(x,0) = \partial^\alpha u(x)$, and, since

$$F_u(x,y) = \frac{(hu)(x+y)}{h(x)},$$

we have

$$\Delta_y G_\alpha(x,0) = \partial_x^\alpha \left(\frac{\Delta(hu)}{h}\right)(x).$$

It follows that

$$\partial^\alpha \left[P_\varepsilon^h u - u - \frac{\varepsilon}{4}\frac{\Delta(hu)}{h}\right](x) = \int_{\mathbb{R}^d} p_\varepsilon(y)\mathcal{R}_\alpha(x,y)\, dy. \tag{S28}$$

To control the remainder, set $y = \sqrt{\varepsilon}\, z$. Since $p_\varepsilon(y)\, dy = p_1(z)\, dz$,

$$\frac{1}{\varepsilon}\int p_\varepsilon(y)\mathcal{R}_\alpha(x,y)\, dy = \int p_1(z)\frac{\mathcal{R}_\alpha(x,\sqrt{\varepsilon}z)}{\varepsilon}\, dz.$$

For each fixed z, the first estimate in (S27) gives

$$\sup_x \frac{|\mathcal{R}_\alpha(x,\sqrt{\varepsilon}z)|}{\varepsilon} \le \frac{1}{2}|z|^2\omega_\alpha(\sqrt{\varepsilon}|z|) \longrightarrow 0.$$

For $0 < \varepsilon \le \varepsilon_0$, the second estimate gives

$$\sup_x \frac{|\mathcal{R}_\alpha(x,\sqrt{\varepsilon}z)|}{\varepsilon} \le C\|u\|_{C^{s+2}_{\mathrm{ub}}}|z|^2 e^{L\sqrt{\varepsilon_0}|z|}.$$

The resulting dominating function

$$p_1(z)|z|^2 e^{L\sqrt{\varepsilon_0}|z|}$$

is integrable. Dominated convergence therefore shows that the right-hand side of (S28), divided by ε, converges uniformly in x to zero. Summing over $|\alpha| \le s$ proves

$$\frac{P_\varepsilon^h u - u}{\varepsilon} \longrightarrow \frac{1}{4}\frac{\Delta(hu)}{h} \qquad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d).$$

Finally,

$$\frac{\Delta(hu)}{h} = \Delta u + 2\nabla\ell^\top\nabla u + \big(\Delta\ell + |\nabla\ell|^2\big)u,$$

which proves (S22).

Taking $u = 1$ yields

$$\eta_\varepsilon = 1 + \frac{\varepsilon}{4}\frac{\Delta h}{h} + o_{C^s_{\mathrm{ub}}}(\varepsilon).$$

Thus $\|\eta_\varepsilon - 1\|_{C^s_{\mathrm{ub}}} = O(\varepsilon)$, and η_ε is bounded away from zero for sufficiently small ε. Lemma S3 gives

$$\frac{\log \eta_\varepsilon}{\varepsilon} \longrightarrow \frac{1}{4}\frac{\Delta h}{h} \qquad \text{in } C^s_{\mathrm{ub}}(\mathbb{R}^d),$$

concluding the proof. □

4.C.3 Properties of ψ

Lemma S8 (Differential of ψ and its near-origin form)**.** *For $\varepsilon \geq 0$, the map $\phi \mapsto \psi(\varepsilon, \phi)$ is continuously Fréchet differentiable on $C^s_{\mathrm{ub}}(\mathbb{R}^d)$, with*

$$D_\phi \psi(\varepsilon, \phi)u = u + \mathcal{K}_{\varepsilon,\phi} u, \qquad \mathcal{K}_{\varepsilon,\phi} u := \frac{P^h_\varepsilon(e^\phi u)}{P^h_\varepsilon(e^\phi)}. \tag{S29}$$

For every fixed $\varepsilon_0 > 0$, there are $r, C > 0$ such that, whenever $0 \leq \varepsilon \leq \varepsilon_0$ and $\|\phi\|, \|\widetilde{\phi}\| \leq r$,

$$\|D_\phi \psi(\varepsilon, \phi) - D_\phi \psi(\varepsilon, \widetilde{\phi})\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \leq C\|\phi - \widetilde{\phi}\|. \tag{S30}$$

Moreover, there is a scalar function $\omega(\varepsilon) \downarrow 0$ such that

$$\|D_\phi \psi(\varepsilon, \phi) - (I + P_\varepsilon)\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \leq \omega(\varepsilon) + C\|\phi\|, \qquad \|\phi\| \leq r. \tag{S31}$$

Proof. For $\varepsilon > 0$, Lemmas S3 and S5, followed by the chain rule, give (S29). At $\varepsilon = 0$, the definition $\psi(0, \phi) = 2\phi$ gives the same formula with $\mathcal{K}_{0,\phi} = I$.

Write

$$q_{\varepsilon,\phi} := P^h_\varepsilon(e^\phi), \qquad \eta_\varepsilon = q_{\varepsilon,0}.$$

If $\|\phi\| \leq r$, then $e^\phi \geq e^{-r}$ pointwise, and hence

$$q_{\varepsilon,\phi} \geq e^{-r}\eta_\varepsilon.$$

Lemma S5 therefore gives a positive lower bound uniform in small ε and ϕ. The same lemmas give uniform $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ bounds for e^ϕ, $q_{\varepsilon,\phi}$, and their reciprocals. Furthermore,

$$\|e^\phi - e^{\widetilde{\phi}}\| \leq C\|\phi - \widetilde{\phi}\|, \qquad \|q_{\varepsilon,\phi} - q_{\varepsilon,\widetilde{\phi}}\| \leq C\|\phi - \widetilde{\phi}\|,$$

and local Lipschitz continuity of inversion yields

$$\|q_{\varepsilon,\phi}^{-1} - q_{\varepsilon,\widetilde{\phi}}^{-1}\| \le C\|\phi - \widetilde{\phi}\|.$$

Now

$$\mathcal{K}_{\varepsilon,\phi} = M_{q_{\varepsilon,\phi}^{-1}} P_\varepsilon^h M_{e^\phi}.$$

Adding and subtracting $M_{q_{\varepsilon,\phi}^{-1}} P_\varepsilon^h M_{e^{\widetilde{\phi}}}$ and applying the multiplier and operator bounds proves (S30).

At the origin,

$$\mathcal{K}_{\varepsilon,0} - P_\varepsilon = M_{\eta_\varepsilon^{-1}}(P_\varepsilon^h - P_\varepsilon) + (M_{\eta_\varepsilon^{-1}} - I)P_\varepsilon. \tag{1}$$

By Lemma S5, $P_\varepsilon^h - P_\varepsilon \to 0$ in operator norm and $\eta_\varepsilon \to 1$ in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$. Smoothness of inversion and the multiplier estimate therefore make the norm of (1) tend to zero. Combining this with (S30) at $\widetilde{\phi} = 0$ proves (S31). □

4.C.4 Preconditioned Fixed Point Theorem

Set

$$\eta_\varepsilon := P_\varepsilon^h 1, \qquad \mathcal{K}_{\varepsilon,0} := M_{\eta_\varepsilon^{-1}} P_\varepsilon^h, \qquad A_\varepsilon := D_\phi\psi(\varepsilon,0) = I + \mathcal{K}_{\varepsilon,0}, \tag{S32}$$

with $A_0 = 2I$.

Lemma S9 (Stable inversion under small perturbations)**.** *Let X be a Banach space. Suppose that $A_\varepsilon, E_\varepsilon \in \mathcal{L}(X)$, that every A_ε is invertible, and that*

$$\sup_\varepsilon \|A_\varepsilon^{-1}\| \le C_0, \qquad \|E_\varepsilon\| \longrightarrow 0.$$

Then $A_\varepsilon + E_\varepsilon$ is invertible for all sufficiently small ε, its inverse is uniformly bounded, and

$$\|(A_\varepsilon + E_\varepsilon)^{-1} - A_\varepsilon^{-1}\| \longrightarrow 0.$$

Consequently, if $A_\varepsilon^{-1} v \to Lv$ for every fixed $v \in X$, then

$$(A_\varepsilon + E_\varepsilon)^{-1} v \longrightarrow Lv$$

for every fixed $v \in X$.

Proof. For sufficiently small ε, $\|A_\varepsilon^{-1}E_\varepsilon\| \le 1/2$, and

$$A_\varepsilon + E_\varepsilon = A_\varepsilon(I + A_\varepsilon^{-1}E_\varepsilon).$$

The Neumann series gives

$$\|(A_\varepsilon + E_\varepsilon)^{-1}\| \le 2C_0.$$

The resolvent identity

$$(A_\varepsilon + E_\varepsilon)^{-1} - A_\varepsilon^{-1} = -(A_\varepsilon + E_\varepsilon)^{-1}E_\varepsilon A_\varepsilon^{-1}$$

then proves operator-norm convergence of the inverses. The final assertion follows by applying this operator-norm convergence to each fixed v. □

Lemma S10 (Uniform inverse for $I + P_\varepsilon$)**.** *For every $s \ge 0$,*

$$\sup_{\varepsilon \ge 0} \|(I + P_\varepsilon)^{-1}\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} < \infty. \tag{S33}$$

Moreover, $(I + P_\varepsilon)^{-1} \to \frac{1}{2}I$ strongly as $\varepsilon \downarrow 0$.

Proof. For $\varepsilon > 0$, the Fourier multiplier of P_ε is $e^{-\varepsilon|\xi|^2/4}$. Hence the multiplier of the inverse is

$$\frac{1}{1 + e^{-\varepsilon|\xi|^2/4}} = 1 - a(\sqrt{\varepsilon}\,\xi), \qquad a(\xi) := \frac{1}{1 + e^{|\xi|^2/4}}.$$

The function a is Schwartz. Let $r = \mathcal{F}^{-1}a \in \mathcal{S}$ and $r_\varepsilon(x) = \varepsilon^{-d/2}r(x/\sqrt{\varepsilon})$. Then $\|r_\varepsilon\|_{L^1} = \|r\|_{L^1}$ and

$$(I + P_\varepsilon)^{-1} = I - r_\varepsilon * . \tag{S34}$$

To justify the identity without taking Fourier transforms of an arbitrary bounded function, note that the Fourier multiplier identity implies the corresponding equality between the L^1 convolution kernels. Fubini's theorem then verifies (S34) on every bounded function. Convolution commutes with derivatives, so

$$\|(I + P_\varepsilon)^{-1}u\| \le (1 + \|r\|_{L^1})\|u\|,$$

uniformly in $\varepsilon > 0$. At $\varepsilon = 0$ the inverse is $\frac{1}{2}I$, proving (S33).

Finally,

$$(I + P_\varepsilon)^{-1}u - \tfrac{1}{2}u = \tfrac{1}{2}(I + P_\varepsilon)^{-1}(I - P_\varepsilon)u.$$

The inverse is uniformly bounded and $P_\varepsilon u \to u$ in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$ by Lemma S4, which proves strong convergence. □

Lemma S11 (Uniform inversion at 0)**.** *There is $\varepsilon_0 > 0$ such that A_ε is invertible for $0 \le \varepsilon \le \varepsilon_0$ and*

$$\sup_{0\le\varepsilon\le\varepsilon_0} \|A_\varepsilon^{-1}\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} < \infty. \tag{S35}$$

Furthermore, $A_\varepsilon^{-1} \to \frac{1}{2}I$ strongly.

Proof. Let $B_\varepsilon := I + P_\varepsilon$ and $E_\varepsilon := A_\varepsilon - B_\varepsilon$. Equation (S31) at $\phi = 0$ gives

$$\|E_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \longrightarrow 0. \tag{S36}$$

More explicitly, the error is exactly

$$E_\varepsilon = M_{\eta_\varepsilon^{-1}}(P^h_\varepsilon - P_\varepsilon) + (M_{\eta_\varepsilon^{-1}} - I)P_\varepsilon,$$

and each term tends to zero by Lemma S5 and smoothness of inversion in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$.

Lemma S10 gives a uniform bound for B_ε^{-1} and the strong limit $B_\varepsilon^{-1} \to \frac{1}{2}I$. Applying Lemma S9 with base operator B_ε and perturbation E_ε proves that A_ε is invertible for all sufficiently small ε, that its inverse is uniformly bounded, and that

$$\|A_\varepsilon^{-1} - B_\varepsilon^{-1}\| \longrightarrow 0.$$

This proves (S35), including $A_0^{-1} = \frac{1}{2}I$, and also gives $A_\varepsilon^{-1} \to \frac{1}{2}I$ strongly. □

Lemma S12 (Preconditioned contraction and solution continuity)**.** *Suppose $\nabla\ell \in C^s_{\mathrm{ub}}(\mathbb{R}^d)$. There exist $r > 0$, $\varepsilon_0 > 0$, and $q \in (0, 1)$ such that, for $0 \le \varepsilon \le \varepsilon_0$, the map*

$$\mathcal{T}_\varepsilon(\phi) := \phi - A_\varepsilon^{-1}\psi(\varepsilon, \phi) \tag{S37}$$

is a contraction with constant q from $\overline{B}_r(0) \subset C^s_{\mathrm{ub}}(\mathbb{R}^d)$ into itself. Its fixed point ϕ_ε is the unique root of $\psi(\varepsilon, \cdot)$ in that ball, the path $\varepsilon \mapsto \phi_\varepsilon$ is continuous, and

$$\|\phi_\varepsilon\| \le \frac{1}{1-q}\|A_\varepsilon^{-1}\psi(\varepsilon, 0)\| \longrightarrow 0. \tag{S38}$$

Proof. Differentiating (S37) in ϕ gives

$$D\mathcal{T}_\varepsilon(\phi) = I - A_\varepsilon^{-1} D_\phi \psi(\varepsilon, \phi) = A_\varepsilon^{-1}\big(A_\varepsilon - D_\phi \psi(\varepsilon, \phi)\big). \tag{S39}$$

In particular, $D\mathcal{T}_\varepsilon(0) = 0$ for every ε. By Lemmas S11 and S8, there are constants M, C independent of small ε such that

$$\|A_\varepsilon^{-1}\| \le M, \qquad \|D_\phi \psi(\varepsilon, \phi) - A_\varepsilon\| \le C\|\phi\| \quad (\|\phi\| \le r_0).$$

Choose $0 < r \le r_0$ so that $q := MCr < 1$. For $\phi, \widetilde{\phi} \in \overline{B}_r$, the line segment between them lies in the ball, and the Banach-space mean-value formula together with (S39) gives

$$\|\mathcal{T}_\varepsilon(\phi) - \mathcal{T}_\varepsilon(\widetilde{\phi})\| \le q\|\phi - \widetilde{\phi}\|. \tag{S40}$$

Next,

$$\mathcal{T}_\varepsilon(0) = -A_\varepsilon^{-1} \psi(\varepsilon, 0).$$

By Lemma S5, $\psi(\varepsilon, 0) = \log \eta_\varepsilon \to 0$ in $C^s_{\mathrm{ub}}(\mathbb{R}^d)$, and the inverse family is uniformly bounded. Hence $\|\mathcal{T}_\varepsilon(0)\| \to 0$. Decrease ε_0 so that $\|\mathcal{T}_\varepsilon(0)\| \le (1-q)r$. Then for $\phi \in \overline{B}_r$,

$$\|\mathcal{T}_\varepsilon(\phi)\| \le q\|\phi\| + \|\mathcal{T}_\varepsilon(0)\| \le qr + (1-q)r = r.$$

Thus the ball is invariant. Banach's fixed-point theorem supplies a unique fixed point ϕ_ε in the ball. Since A_ε is invertible, the fixed-point equation is equivalent to $\psi(\varepsilon, \phi_\varepsilon) = 0$. Applying (S40) with $\widetilde{\phi} = 0$ gives

$$\|\phi_\varepsilon\| \le q\|\phi_\varepsilon\| + \|\mathcal{T}_\varepsilon(0)\|,$$

which is exactly (S38).

It remains to prove continuity. At zero, it follows directly from (S38). Fix now $\varepsilon_* > 0$. By Lemma S6, $P^h_\varepsilon \to P^h_{\varepsilon_*}$ in operator norm as $\varepsilon \to \varepsilon_*$. The uniform composition and lower-bound estimates in the proof of Lemma S8 then imply

$$\sup_{\phi \in \overline{B}_r} \|\psi(\varepsilon, \phi) - \psi(\varepsilon_*, \phi)\| \to 0.$$

The same operator-norm continuity gives $A_\varepsilon \to A_{\varepsilon_*}$; the resolvent identity and uniform inverse bounds give $A_\varepsilon^{-1} \to A_{\varepsilon_*}^{-1}$ in operator norm. Consequently,

$$\sup_{\phi \in \overline{B}_r} \|\mathcal{T}_\varepsilon(\phi) - \mathcal{T}_{\varepsilon_*}(\phi)\| \to 0.$$

Using the fixed-point equations and (S40),

$$(1-q)\|\phi_\varepsilon - \phi_{\varepsilon_*}\| \le \sup_{\phi \in \overline{B}_r} \|\mathcal{T}_\varepsilon(\phi) - \mathcal{T}_{\varepsilon_*}(\phi)\|,$$

which proves continuity at every positive ε_*. □

Lemma S13 (Averaged differential)**.** *Let ϕ_ε be the branch from Lemma S12 and define*

$$\overline{A}_\varepsilon := \int_0^1 D_\phi \psi(\varepsilon, t\phi_\varepsilon)\, dt. \tag{S41}$$

Then

$$\|\overline{A}_\varepsilon - A_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(\mathbb{R}^d))} \longrightarrow 0. \tag{S42}$$

For all sufficiently small ε, $\overline{A}_\varepsilon$ is invertible, $\sup_\varepsilon \|\overline{A}_\varepsilon^{-1}\| < \infty$, and $\overline{A}_\varepsilon^{-1} \to \frac{1}{2} I$ strongly.

Proof. By (S30),

$$\|\overline{A}_\varepsilon - A_\varepsilon\| \le \int_0^1 \|D_\phi \psi(\varepsilon, t\phi_\varepsilon) - D_\phi \psi(\varepsilon, 0)\|\, dt \le \frac{C}{2}\|\phi_\varepsilon\| \longrightarrow 0,$$

which proves (S42). Apply Lemma S9 with base operator A_ε and perturbation $\overline{A}_\varepsilon - A_\varepsilon$. Lemma S11 then implies that $\overline{A}_\varepsilon$ is invertible for all sufficiently small ε, its inverse is uniformly bounded, and

$$\|\overline{A}_\varepsilon^{-1} - A_\varepsilon^{-1}\| \longrightarrow 0.$$

Since $A_\varepsilon^{-1} \to \frac{1}{2} I$ strongly, the same is true of $\overline{A}_\varepsilon^{-1}$.

Finally, the fundamental theorem of calculus in the ϕ variable gives

$$0 = \psi(\varepsilon, \phi_\varepsilon) = \psi(\varepsilon, 0) + \overline{A}_\varepsilon \phi_\varepsilon,$$

and hence the exact identity

$$\phi_\varepsilon = -\overline{A}_\varepsilon^{-1} \psi(\varepsilon, 0).$$

□

4.D Compact Manifolds with Ambient Gaussian Kernels

Let (M, g) be a closed, connected, smooth Riemannian manifold of dimension d, and let

$$\iota : M \hookrightarrow \mathbb{R}^D$$

be a smooth isometric embedding. Write dV for Riemannian volume and Δ_M for the Laplace–Beltrami operator, with $-\Delta_M$ nonnegative. For $\varepsilon > 0$, define

$$p_\varepsilon^{\mathrm{amb}}(x, y) := c_\varepsilon \exp\left(-\frac{\|\iota(x) - \iota(y)\|^2}{\varepsilon}\right), \qquad c_\varepsilon := (\pi\varepsilon)^{-d/2}, \tag{S43}$$

and

$$P_\varepsilon^M u(x) := \int_M p_\varepsilon^{\mathrm{amb}}(x, y) u(y)\, dV(y), \qquad P_0^M := I. \tag{S44}$$

The degree and the symmetric degree normalization are

$$m_\varepsilon := P_\varepsilon^M 1, \qquad \widehat{P}_\varepsilon^M := M_{m_\varepsilon^{-1/2}} P_\varepsilon^M M_{m_\varepsilon^{-1/2}}, \qquad m_0 := 1, \quad \widehat{P}_0^M := I, \tag{S45}$$

where M_a denotes multiplication by a. Thus

$$\widehat{p}_\varepsilon^M(x, y) = \frac{p_\varepsilon^{\mathrm{amb}}(x, y)}{\sqrt{m_\varepsilon(x) m_\varepsilon(y)}}.$$

For comparison, write

$$H_\varepsilon^M := e^{(\varepsilon/4)\Delta_M}, \qquad H_0^M := I. \tag{S46}$$

For an integer $s \geq 0$, put

$$C^s_{\mathrm{ub}}(M) := C^s(M), \qquad \|u\|_s := \sum_{j=0}^{s} \|\nabla^j u\|_\infty.$$

The subscript "ub" is retained to match the Euclidean notation, although boundedness and uniform continuity are automatic on the compact manifold M.

The argument separates into two parts. Lemmas S18–S22 contain all geometric and heat-kernel input. The product-equation arguments in Sections 4.D.2 and 4.D.3 are then standard Banach-space arguments of the same type as those used in Appendix B.

4.D.1 Geometric Kernel Estimates

Lemma S14 (Bounded C^s calculus)**.** *For every integer $s \geq 0$, $C^s_{\mathrm{ub}}(M)$ is a Banach algebra:*

$$\|uv\|_s \leq C_s \|u\|_s \|v\|_s. \tag{S47}$$

If $J \subset \mathbb{R}$ is open and $F \in C^{s+2}(J)$, then composition $u \mapsto F \circ u$ is continuously Fréchet differentiable on every bounded set whose ranges lie in a fixed compact subset of J, with

$$D(F_\circ)(u)[v] = F'(u)v.$$

Its differential is locally Lipschitz in operator norm.

Proof. The covariant Leibniz rule expresses $\nabla^j(uv)$ as a finite sum of contractions of $\nabla^k u$ and $\nabla^{j-k} v$, proving (S47). Completeness follows from equivalence of $\|\cdot\|_s$ with the usual chartwise C^s norms on a finite atlas. Repeated covariant differentiation of $F(u)$ gives the same finite Faà di Bruno sums as in Lemma S3. The identities below are applied pointwise in the scalar target. After restricting to a sufficiently small $C^s_{\mathrm{ub}}(M)$ neighborhood, every pointwise line segment remains in the same compact subset of J. The algebra bound and the pointwise identities

$$F(u+v) - F(u) - F'(u)v = v^2 \int_0^1 (1-t) F''(u+tv)\, dt$$

and

$$F'(u) - F'(\widetilde{u}) = (u - \widetilde{u}) \int_0^1 F''\big(\widetilde{u} + t(u-\widetilde{u})\big)\, dt$$

give the asserted differentiability and local Lipschitz estimates. □

We next record the elementary mapping estimate that will be used to translate local heat-kernel asymptotics into operator bounds.

Lemma S15 (Mapping estimate for heat-type kernels)**.** *Fix an integer $s \geq 0$ and let $\rho \geq 0$. Let*

$$K_\varepsilon \in C^\infty(M \times M), \qquad 0 < \varepsilon \leq \varepsilon_0,$$

and define

$$T_\varepsilon u(x) := \int_M K_\varepsilon(x,y) u(y)\, dV(y).$$

Assume that there is a neighborhood W of the diagonal $\Delta_M \subset M \times M$ such that:

(i) *For every $0 \le a \le s$ and every $N < \infty$,*

$$\sup_{(x,y)\notin W} |\nabla_x^a K_\varepsilon(x,y)| \le C_{a,N}\varepsilon^N. \tag{S48}$$

(ii) *After a standard finite localization of W to product coordinate charts, each localized coordinate kernel, including the Riemannian volume density, has the form*

$$A_{\varepsilon,j}(\xi,\eta) = \varepsilon^{-d/2+\rho/2} a_j\left(\sqrt{\varepsilon}, \frac{\xi-\eta}{\sqrt{\varepsilon}}, \eta\right). \tag{S49}$$

For every multi-index $|\beta| \le s$, there is $g_{j,\beta} \in L^1(\mathbb{R}^d)$ such that

$$\sup_{\substack{0<r\le\sqrt{\varepsilon_0}\\ \eta\in\mathbb{R}^d}} \left|\partial_\eta^\beta a_j(r,z,\eta)\right| \le g_{j,\beta}(z), \qquad z \in \mathbb{R}^d. \tag{S50}$$

Then

$$\|T_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \le C_s \varepsilon^{\rho/2}, \qquad 0 < \varepsilon \le \varepsilon_0. \tag{S51}$$

Proof. Choose product coordinate charts

$$\kappa_j : U_j \longrightarrow V_j \subset \mathbb{R}^d$$

and functions

$$\chi_j \in C_c^\infty(U_j \times U_j), \qquad j = 1, \dots, J,$$

such that

$$\sum_{j=1}^J \chi_j = 1 \quad \text{on } W.$$

Set

$$\chi_0 := 1 - \sum_{j=1}^J \chi_j$$

and decompose

$$T_\varepsilon = T_{\varepsilon,0} + \sum_{j=1}^J T_{\varepsilon,j},$$

where $T_{\varepsilon,j}$ has kernel $\chi_j K_\varepsilon$.

For the off-diagonal term, χ_0 vanishes on W. By the covariant Leibniz rule and (S48), for every $0 \le a \le s$ and every N,

$$\sup_{x,y} \left|\nabla_x^a\big(\chi_0 K_\varepsilon\big)(x,y)\right| \le C_{a,N}\varepsilon^N.$$

Since M has finite volume,

$$\|\nabla^a T_{\varepsilon,0} u\|_\infty \le C_{a,N} \varepsilon^N \|u\|_\infty.$$

Choosing $N \ge \rho/2$ and summing over $a \le s$ gives

$$\|T_{\varepsilon,0} u\|_{C^s_{\mathrm{ub}}(M)} \le C_s \varepsilon^{\rho/2} \|u\|_{C^s_{\mathrm{ub}}(M)}. \tag{S52}$$

Fix one near-diagonal piece $T_{\varepsilon,j}$. Write

$$\xi = \kappa_j(x), \qquad \eta = \kappa_j(y).$$

If

$$dV(y) = q_j(\eta)\, d\eta,$$

then the localized coordinate kernel is

$$A_{\varepsilon,j}(\xi,\eta) := \chi_j(x,y) K_\varepsilon(x,y) q_j(\eta).$$

This is the quantity represented in (S49).

Multiplying u by a fixed cutoff equal to one on the second projection of $\operatorname{supp} \chi_j$, its coordinate representative may be extended by zero to a function $u_j \in C^s(\mathbb{R}^d)$ satisfying

$$\|u_j\|_{C^s(\mathbb{R}^d)} \le C_{j,s} \|u\|_{C^s_{\mathrm{ub}}(M)}. \tag{S53}$$

Thus

$$(T_{\varepsilon,j} u)(\kappa_j^{-1}(\xi)) = \int_{\mathbb{R}^d} A_{\varepsilon,j}(\xi,\eta) u_j(\eta)\, d\eta.$$

Put

$$r := \sqrt{\varepsilon}, \qquad z := \frac{\xi - \eta}{r}.$$

Then $\eta = \xi - rz$ and $d\eta = r^d\, dz$. Substitution of (S49) gives

$$(T_{\varepsilon,j} u)(\kappa_j^{-1}(\xi)) = \varepsilon^{\rho/2} \int_{\mathbb{R}^d} a_j(r,z,\xi - rz) u_j(\xi - rz)\, dz. \tag{S54}$$

For a multi-index $|\alpha| \le s$, differentiation in ξ gives

$$\begin{aligned} &\partial_\xi^\alpha (T_{\varepsilon,j} u)(\kappa_j^{-1}(\xi)) \\ &\quad = \varepsilon^{\rho/2} \sum_{\beta \le \alpha} \binom{\alpha}{\beta} \int_{\mathbb{R}^d} (\partial_\eta^\beta a_j)(r,z,\xi - rz)\, \partial^{\alpha-\beta} u_j(\xi - rz)\, dz. \end{aligned} \tag{S55}$$

This differentiation is justified by (S50), since the absolute value of each differentiated integrand is bounded by

$$g_{j,\beta}(z)\|u_j\|_{C^s(\mathbb{R}^d)},$$

which is integrable in z.

Taking the supremum in ξ and using (S53),

$$\begin{aligned}\left\|\partial_\xi^\alpha\left[(T_{\varepsilon,j}u)\circ\kappa_j^{-1}\right]\right\|_\infty &\le C_\alpha\varepsilon^{\rho/2}\|u_j\|_{C^s(\mathbb{R}^d)}\sum_{\beta\le\alpha}\|g_{j,\beta}\|_{L^1}\\ &\le C_{j,s}\varepsilon^{\rho/2}\|u\|_{C^s_{\mathrm{ub}}(M)}.\end{aligned}$$

Summing over $|\alpha|\le s$, and using equivalence of coordinate and intrinsic C^s norms on the fixed support of the localized term, gives

$$\|T_{\varepsilon,j}u\|_{C^s_{\mathrm{ub}}(M)} \le C_{j,s}\varepsilon^{\rho/2}\|u\|_{C^s_{\mathrm{ub}}(M)}.$$

There are finitely many near-diagonal pieces. Summing their bounds and using (S52) proves

$$\|T_\varepsilon u\|_{C^s_{\mathrm{ub}}(M)} \le C_s\varepsilon^{\rho/2}\|u\|_{C^s_{\mathrm{ub}}(M)}.$$

□

This lemma is the operator-norm consequence of the local kernel class used in the heat calculus of Grieser (2004, Definition 2.1 and Lemma 2.3). We included the short proof to make clear which part of that calculus is needed here.

Lemma S16 (Continuity and generator limit for the intrinsic heat semigroup)**.** *For every integer $s\ge 0$, the family*

$$H_\varepsilon^M := e^{(\varepsilon/4)\Delta_M}$$

is uniformly bounded and strongly continuous on $C^s_{\mathrm{ub}}(M)$, and is operator-norm continuous for $\varepsilon>0$. Moreover, for every $u\in C^{s+2}_{\mathrm{ub}}(M)$,

$$\frac{H_\varepsilon^M u-u}{\varepsilon}\longrightarrow\frac{1}{4}\Delta_M u \qquad \text{in } C^s_{\mathrm{ub}}(M).$$

Proof. These are standard properties of the heat semigroup on a closed Riemannian manifold; see, e.g., Grieser (2004, Theorem 1.1 and Section 2.4) for the heat-kernel construction and

short-time convergence, and Engel and Nagel (2000, Theorem II.4.29) for positive-time norm continuity of an immediately compact C_0-semigroup. The heat semigroup is immediately smoothing, and hence compact on $C^s_{\mathrm{ub}}(M)$ for every positive time.

Its generator on $C^s_{\mathrm{ub}}(M)$ contains $C^{s+2}_{\mathrm{ub}}(M)$ and acts there as Δ_M. Thus, with the present time normalization,

$$\frac{H^M_\varepsilon u - u}{\varepsilon} = \frac{e^{(\varepsilon/4)\Delta_M} u - u}{\varepsilon} \longrightarrow \frac{1}{4}\Delta_M u.$$

□

Lemma S17 (Uniform inverse for intrinsic heat)**.** *For every integer $s \geq 0$,*

$$\sup_{0\leq\varepsilon\leq\varepsilon_0} \left\|(I + H^M_\varepsilon)^{-1}\right\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} < \infty. \tag{S56}$$

Moreover, for every fixed $v \in C^s_{\mathrm{ub}}(M)$,

$$(I + H^M_\varepsilon)^{-1} v \longrightarrow \frac{1}{2} v \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S57}$$

Proof. Put $A := -\Delta_M/4$, so that $H^M_\varepsilon = e^{-\varepsilon A}$. Choose $a \in \mathcal{S}(\mathbb{R})$ such that

$$a(r) = \frac{1}{1 + e^r}, \qquad r \geq 0,$$

and set $S_\varepsilon := I - a(\varepsilon A)$. On a Laplace eigenfunction e_j satisfying $Ae_j = \lambda_j e_j$, $\lambda_j \geq 0$,

$$S_\varepsilon e_j = \frac{1}{1 + e^{-\varepsilon\lambda_j}} e_j, \qquad S_\varepsilon (I + H^M_\varepsilon) e_j = e_j.$$

The semiclassical functional calculus on a closed manifold, applied with $h = \sqrt{\varepsilon}$ to $a(h^2 A)$, gives a uniformly controlled semiclassical smoothing family; see Küster (2017, Theorem 2.6). Its localized kernels have the heat-scale form of Lemma S15 with $\rho = 0$. Hence

$$\sup_{0<\varepsilon\leq\varepsilon_0} \|a(\varepsilon A)\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} < \infty,$$

and the same is true of S_ε.

Finite linear combinations of Laplace eigenfunctions are dense in $C^s_{\mathrm{ub}}(M)$. Indeed, first approximate v by $H^M_\delta v$ in C^s. For fixed $\delta > 0$, approximate $H^M_{\delta/2} v$ in $L^2(M)$ by a finite eigenfunction sum and apply the bounded smoothing map $H^M_{\delta/2} : L^2(M) \to C^s_{\mathrm{ub}}(M)$. Since S_ε

and $I + H_\varepsilon^M$ are bounded on $C^s_{\mathrm{ub}}(M)$, the two inverse identities verified on finite eigenfunction sums extend by density. Thus

$$S_\varepsilon = (I + H_\varepsilon^M)^{-1},$$

and (S56) follows.

Finally,

$$(I + H_\varepsilon^M)^{-1} v - \frac{1}{2} v = \frac{1}{2}(I + H_\varepsilon^M)^{-1}(I - H_\varepsilon^M)v.$$

The inverse family is uniformly bounded and $H_\varepsilon^M v \to v$ strongly, proving (S57). □

Let II denote the second fundamental form of the embedding and let

$$\mathbf{H} := \mathrm{tr}_g \, \mathrm{II}$$

be its mean-curvature vector. Define

$$\omega_M := \frac{2\|\mathrm{II}\|^2 - \|\mathbf{H}\|^2}{16}. \tag{S58}$$

Here $\mathbf{H}$ is the unnormalized mean-curvature vector, and all tensor norms are pointwise norms induced by g and the Euclidean metric on the normal bundle.

Lemma S18 (Ambient Gaussian expansion)**.** *For every integer $s \geq 0$, the family P_ε^M is uniformly bounded on $C^s_{\mathrm{ub}}(M)$ for small ε, strongly continuous at zero, and operator-norm continuous for $\varepsilon > 0$. For every $u \in C^{s+2}_{\mathrm{ub}}(M)$,*

$$\frac{P_\varepsilon^M u - u}{\varepsilon} \longrightarrow \frac{1}{4}\Delta_M u + \omega_M u \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S59}$$

In particular,

$$m_\varepsilon = 1 + \varepsilon \omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon). \tag{S60}$$

Proof. We first prove the expansion. Choose

$$0 < \rho_1 < \rho_2 < \mathrm{inj}(M)$$

small enough that

$$\|\iota(\exp_x v) - \iota(x)\|^2 \geq c|v|^2 \tag{S61}$$

uniformly for $x \in M$ and $|v| \le \rho_2$, for some $c > 0$. Such a choice is possible because ι is isometric and M is compact.

Choose a smooth cutoff χ such that

$$\chi(x,y) = 1 \quad \text{if } d_M(x,y) \le \rho_1, \qquad \chi(x,y) = 0 \quad \text{if } d_M(x,y) \ge \rho_2.$$

The cutoff may be chosen smooth because $d_M(x,y)^2$ is smooth on a neighborhood of the diagonal of radius smaller than the injectivity radius. Write

$$P_\varepsilon^M = P_{\varepsilon,\mathrm{loc}}^M + P_{\varepsilon,\mathrm{far}}^M$$

for the operators obtained by multiplying the kernel by χ and $1-\chi$, respectively.

The off-diagonal term. The support of $1-\chi$ is a compact subset of $M \times M$ disjoint from the diagonal. Since ι is injective, there is $c_0 > 0$ such that

$$\|\iota(x) - \iota(y)\| \ge c_0$$

on this support. For every $0 \le a \le s$, covariant differentiation in x gives a finite sum of terms bounded by

$$C_a \varepsilon^{-N_a} \exp\left(-\frac{c_0^2}{\varepsilon}\right)$$

for some $N_a < \infty$. Since

$$\varepsilon^{-N_a} e^{-c_0^2/\varepsilon} = O(\varepsilon^N) \qquad \text{for every } N < \infty,$$

we obtain

$$\|P_{\varepsilon,\mathrm{far}}^M u\|_{C^s(M)} = O(\varepsilon^N)\|u\|_\infty \tag{S62}$$

for every N. Thus the off-diagonal part is negligible at order ε.

The local expansion. It suffices to work on one member of a finite cover on which TM admits a smooth orthonormal frame. We identify T_xM with $\mathbb{R}^d$ using this frame; all estimates below are uniform over the finite cover.

Put

$$r := \sqrt{\varepsilon}, \qquad y = \exp_x(rz).$$

Since

$$d_M(x, \exp_x(rz)) = r|z|$$

within the injectivity radius, the localized operator can be written as

$$P^M_{\varepsilon,\mathrm{loc}} u(x) = \int_{\mathbb{R}^d} \pi^{-d/2} \chi_r(z) e^{-Q_r(x,z)} u(\exp_x(rz)) J_x(rz)\, dz, \tag{S63}$$

where

$$\chi_r(z) := \chi(x, \exp_x(rz)),$$
$$Q_r(x,z) := \frac{\|\iota(\exp_x(rz)) - \iota(x)\|^2}{r^2},$$

and $J_x(v)$ is the volume density in normal coordinates:

$$dV(\exp_x v) = J_x(v)\, dv.$$

The integral in (S63) is extended by zero outside $r|z| < \rho_2$.

The standard normal-coordinate expansions, uniformly in x, are

$$\|\iota(\exp_x v) - \iota(x)\|^2 = |v|^2 - \frac{1}{12}\|\mathrm{II}_x(v,v)\|^2 + O(|v|^5), \tag{S64}$$
$$J_x(v) = 1 - \frac{1}{6}\,\mathrm{Ric}_x(v,v) + O(|v|^3). \tag{S65}$$

The same expansions hold after taking any fixed number of derivatives in the base point x, with uniform remainder bounds on the compact manifold.

After substituting $v = rz$, the first expansion gives

$$Q_r(x,z) = |z|^2 - \frac{r^2}{12}\|\mathrm{II}_x(z,z)\|^2 + O(r^3|z|^5), \tag{S66}$$

while

$$J_x(rz) = 1 - \frac{r^2}{6}\,\mathrm{Ric}_x(z,z) + O(r^3|z|^3). \tag{S67}$$

Taylor's formula along the geodesic $t \mapsto \exp_x(trz)$ gives

$$u(\exp_x(rz)) = u(x) + r\, du_x[z] + \frac{r^2}{2}\nabla^2 u_x[z,z] + r^2 R_{r,u}(x,z), \tag{S68}$$

where, for every fixed z,

$$\|R_{r,u}(\cdot, z)\|_{C^s} \longrightarrow 0 \qquad (r \downarrow 0). \tag{S69}$$

Moreover, repeated differentiation in x of the integral Taylor remainder gives, for some integer N_s,

$$\|R_{r,u}(\cdot,z)\|_{C^s} \le C_s\|u\|_{C^{s+2}(M)}(1+|z|)^{N_s} \tag{S70}$$

whenever $r|z| \le \rho_2$.

We now combine (S66)– (S68). The following form of the resulting expansion is what will be used below:

$$\begin{aligned}\chi_r(z)e^{-Q_r(x,z)}u(\exp_x(rz))J_x(rz)&\\ = e^{-|z|^2}\Bigg[u(x) + r\,du_x[z] + \frac{r^2}{2}\nabla^2 u_x[z,z] - \frac{r^2}{6}\operatorname{Ric}_x(z,z)u(x)&\\ + \frac{r^2}{12}\|\mathrm{II}_x(z,z)\|^2 u(x)\Bigg] + r^2\mathcal{R}_r(x,z),&\end{aligned} \tag{S71}$$

where

$$\|\mathcal{R}_r(\cdot,z)\|_{C^s} \longrightarrow 0 \qquad \text{for every fixed } z, \tag{S72}$$

and, for some $c_1 > 0$ and some integer L_s,

$$\|\mathcal{R}_r(\cdot,z)\|_{C^s} \le C_s\|u\|_{C^{s+2}(M)}(1+|z|)^{L_s}e^{-c_1|z|^2}. \tag{S73}$$

For completeness, we justify these two remainder properties. After taking up to s derivatives in the base point and dividing by r^2, the finitely many omitted terms have the form

$$r^{\kappa}p(x,z), \qquad \kappa > 0,$$

where $p(x,z)$ and its relevant base-point derivatives have polynomial growth in z. Choose $\theta_s > 0$ sufficiently small that every such term tends uniformly to zero on

$$|z| \le r^{-\theta_s}.$$

On this region $r|z| \to 0$, so the remainders in (S66)– (S68), together with Taylor's formula

$$e^a = 1 + a + O(a^2e^{|a|}),$$

give (S72), uniformly after up to s base-point derivatives.

For domination, (S61) gives

$$e^{-Q_r(x,z)} \le e^{-c|z|^2}$$

on the support of χ_r, while all differentiated geometric factors grow at most polynomially in z. Hence the remainder on $|z| \le r^{-\theta_s}$ is bounded by a polynomial times $e^{-c_1|z|^2}$.

On the complementary region $|z| > r^{-\theta_s}$, both the exact integrand and every retained term satisfy the same Gaussian-polynomial bound. Any negative power of r produced by division by r^2 or by a fixed number of derivatives can be absorbed into a slightly weaker Gaussian: for every $N, m < \infty$ and $0 < c_1 < c$,

$$r^{-N}(1+|z|)^m e^{-c|z|^2} \le C_{N,m} e^{-c_1|z|^2}, \qquad |z| > r^{-\theta_s},$$

for sufficiently small r. This proves (S73).

Finally, the replacement of the cutoff χ_r by 1 in the explicit terms of (S71) introduces only a rapidly small error. Indeed, $\chi_r(z) \ne 1$ implies

$$|z| \ge \frac{\rho_1}{r},$$

and therefore, for every polynomial p and every $N < \infty$,

$$\int_{|z| \ge \rho_1/r} |p(z)| e^{-c|z|^2}\, dz = O(r^N). \tag{S74}$$

Substituting (S71) into (S63), using (S74), and applying dominated convergence to (S72)–(S73), we obtain in $C^s_{\mathrm{ub}}(M)$

$$\begin{aligned} P^M_\varepsilon u = \int_{\mathbb{R}^d} \pi^{-d/2} e^{-|z|^2} \Bigg[& u + r\, du[z] + \frac{r^2}{2} \nabla^2 u[z,z] \\ & - \frac{r^2}{6} \operatorname{Ric}(z,z) u + \frac{r^2}{12} \|\mathrm{II}(z,z)\|^2 u \Bigg] dz + o(r^2). \end{aligned} \tag{S75}$$

It remains to evaluate the Gaussian moments. Under the probability density

$$\pi^{-d/2} e^{-|z|^2},$$

we have

$$\mathbb{E}[z_i] = 0, \qquad \mathbb{E}[z_i z_j] = \frac{1}{2}\delta_{ij}, \tag{S76}$$

and

$$\mathbb{E}[z_i z_j z_k z_\ell] = \frac{1}{4}\left(\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}\right).$$

Thus the linear term vanishes and

$$\frac{1}{2}\mathbb{E}[\nabla^2 u[z,z]] = \frac{1}{4}\Delta_M u.$$

Similarly,

$$-\frac{1}{6}\mathbb{E}[\mathrm{Ric}(z,z)]u = -\frac{1}{12}\,\mathrm{Scal}\; u.$$

Writing

$$\mathrm{II}_{ij} := \mathrm{II}(e_i, e_j)$$

in an orthonormal basis,

$$\begin{aligned}\mathbb{E}\|\mathrm{II}(z,z)\|^2 &= \sum_{i,j,k,\ell} \langle \mathrm{II}_{ij}, \mathrm{II}_{k\ell}\rangle \mathbb{E}[z_i z_j z_k z_\ell] \\ &= \frac{1}{4}\left(\|\mathbf{H}\|^2 + 2\|\mathrm{II}\|^2\right).\end{aligned}$$

Hence the chordal correction is

$$\frac{1}{48}\left(\|\mathbf{H}\|^2 + 2\|\mathrm{II}\|^2\right) u.$$

By the Gauss identity,

$$\mathrm{Scal} = \|\mathbf{H}\|^2 - \|\mathrm{II}\|^2.$$

Therefore

$$\begin{aligned}-\frac{1}{12}\,\mathrm{Scal} + \frac{1}{48}\left(\|\mathbf{H}\|^2 + 2\|\mathrm{II}\|^2\right) &= \frac{2\|\mathrm{II}\|^2 - \|\mathbf{H}\|^2}{16} \\ &= \omega_M.\end{aligned}$$

Since $r^2 = \varepsilon$, (S75) becomes

$$P_\varepsilon^M u = u + \varepsilon\left(\frac{1}{4}\Delta_M u + \omega_M u\right) + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon),$$

which proves (S59).

We next record the operator-continuity properties. The same near-diagonal rescaling and (S61) show that the localized ambient kernel has the heat-type form of Lemma S15 with no additional power of $\sqrt{\varepsilon}$. Together with (S62), that lemma gives

$$\sup_{0<\varepsilon\le\varepsilon_0} \|P_\varepsilon^M\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} < \infty. \tag{S77}$$

The expansion already shows that

$$P_\varepsilon^M u \longrightarrow u$$

in $C^s_{\mathrm{ub}}(M)$ for $u \in C^{s+2}_{\mathrm{ub}}(M)$. Since $C^\infty(M)$ is dense in $C^s_{\mathrm{ub}}(M)$, the uniform bound (S77) extends this convergence to every $u \in C^s_{\mathrm{ub}}(M)$. Thus the family is strongly continuous at zero.

Finally, fix $0 < a < b < \infty$. Put

$$\Phi(x,y) := \|\iota(x) - \iota(y)\|^2.$$

For $\varepsilon > 0$,

$$\partial_\varepsilon p_\varepsilon^{\mathrm{amb}}(x,y) = \left(-\frac{d}{2\varepsilon} + \frac{\Phi(x,y)}{\varepsilon^2}\right) p_\varepsilon^{\mathrm{amb}}(x,y).$$

Since $[a,b] \times M \times M$ is compact, this kernel and all of its x-derivatives through order s are uniformly bounded there. Consequently,

$$\sup_{\varepsilon\in[a,b]} \left\|\partial_\varepsilon P_\varepsilon^M\right\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} < \infty.$$

The fundamental theorem of calculus therefore gives

$$\|P_\varepsilon^M - P_\delta^M\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \le C_{a,b,s}|\varepsilon - \delta|, \qquad \varepsilon,\delta \in [a,b],$$

proving operator-norm continuity on $(0,\infty)$.

Taking $u = 1$ in (S59) and using $\Delta_M 1 = 0$ gives

$$m_\varepsilon = P_\varepsilon^M 1 = 1 + \varepsilon\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon),$$

which proves (S60). □

Lemma S19 (Cancellation under symmetric degree normalization)**.** *For every integer $s \geq 0$, the family $\widehat{P}_\varepsilon^M$ is uniformly bounded on $C^s_{\mathrm{ub}}(M)$ for small ε, strongly continuous at zero, and operator-norm continuous for $\varepsilon > 0$. For every $u \in C^{s+2}_{\mathrm{ub}}(M)$,*

$$\frac{\widehat{P}_\varepsilon^M u - u}{\varepsilon} \longrightarrow \frac{1}{4}\Delta_M u \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S78}$$

Proof. Recall that

$$\widehat{P}_\varepsilon^M = M_{a_\varepsilon} P_\varepsilon^M M_{a_\varepsilon}, \qquad a_\varepsilon := m_\varepsilon^{-1/2}.$$

We first derive the first-order expansion. Apply (S60) with s replaced by $s+2$:

$$m_\varepsilon = 1 + \varepsilon\omega_M + r_\varepsilon, \qquad \|r_\varepsilon\|_{s+2} = o(\varepsilon).$$

Since $m_\varepsilon \to 1$ in $C^{s+2}(M)$, it is bounded away from zero for sufficiently small ε. Lemma S14, applied to $z \mapsto z^{-1/2}$ in a fixed neighborhood of 1, therefore gives

$$a_\varepsilon = 1 - \frac{\varepsilon}{2}\omega_M + \widetilde{r}_\varepsilon, \qquad \|\widetilde{r}_\varepsilon\|_{s+2} = o(\varepsilon). \tag{S79}$$

Fix $u \in C^{s+2}_{\mathrm{ub}}(M)$ and put

$$b := \frac{1}{2}\omega_M u.$$

By the Banach-algebra estimate,

$$a_\varepsilon u = u - \varepsilon b + q_\varepsilon, \qquad \|q_\varepsilon\|_{s+2} = o(\varepsilon). \tag{S80}$$

Applying P_ε^M gives

$$P_\varepsilon^M(a_\varepsilon u) = P_\varepsilon^M u - \varepsilon P_\varepsilon^M b + P_\varepsilon^M q_\varepsilon.$$

The ambient expansion of Lemma S18 yields

$$P_\varepsilon^M u = u + \varepsilon\left(\frac{1}{4}\Delta_M u + \omega_M u\right) + o_{C^s_{\mathrm{ub}}}(\varepsilon).$$

Moreover, $b \in C^s_{\mathrm{ub}}(M)$ is fixed, so strong continuity of P_ε^M gives

$$P_\varepsilon^M b = b + o_{C^s_{\mathrm{ub}}}(1).$$

Finally, uniform boundedness of P_ε^M on $C^s_{\mathrm{ub}}(M)$ and $\|q_\varepsilon\|_s = o(\varepsilon)$ imply

$$\|P_\varepsilon^M q_\varepsilon\|_s = o(\varepsilon).$$

Hence

$$P_\varepsilon^M(a_\varepsilon u) = u + \varepsilon\left(\frac{1}{4}\Delta_M u + \omega_M u - b\right) + o_{C^s_{\mathrm{ub}}}(\varepsilon)$$
$$= u + \varepsilon\left(\frac{1}{4}\Delta_M u + \frac{1}{2}\omega_M u\right) + o_{C^s_{\mathrm{ub}}}(\varepsilon). \tag{S81}$$

Multiplying (S81) by the outer factor a_ε from (S79), and using the Banach-algebra estimate once more, gives

$$\widehat{P}_\varepsilon^M u = \left(1 - \frac{\varepsilon}{2}\omega_M + o_{C^s_{\mathrm{ub}}}(\varepsilon)\right)\left[u + \varepsilon\left(\frac{1}{4}\Delta_M u + \frac{1}{2}\omega_M u\right) + o_{C^s_{\mathrm{ub}}}(\varepsilon)\right]$$
$$= u + \frac{\varepsilon}{4}\Delta_M u + o_{C^s_{\mathrm{ub}}}(\varepsilon).$$

The two multiplication terms involving $\omega_M u/2$ cancel. This proves (S78).

We next establish the operator properties. From $m_\varepsilon \to 1$ in $C^s_{\mathrm{ub}}(M)$ and smooth composition,

$$a_\varepsilon \longrightarrow 1 \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S82}$$

In particular,

$$\sup_{0<\varepsilon\le\varepsilon_0} \|a_\varepsilon\|_s < \infty.$$

The multiplier estimate and uniform boundedness of P_ε^M therefore give

$$\|\widehat{P}_\varepsilon^M u\|_s = \|a_\varepsilon P_\varepsilon^M(a_\varepsilon u)\|_s$$
$$\le C_s\|a_\varepsilon\|_s\|P_\varepsilon^M\|_{\mathcal{L}(C^s_{\mathrm{ub}})}\|a_\varepsilon u\|_s$$
$$\le C_s'\|u\|_s,$$

uniformly for sufficiently small ε. Thus $\widehat{P}_\varepsilon^M$ is uniformly bounded.

For strong continuity at zero, write

$$\widehat{P}_\varepsilon^M u - u = (a_\varepsilon - 1)P_\varepsilon^M(a_\varepsilon u) + P_\varepsilon^M\big((a_\varepsilon - 1)u\big) + (P_\varepsilon^M u - u). \tag{S83}$$

The first term tends to zero in $C^s_{\mathrm{ub}}(M)$ by (S82), the multiplier estimate, and the uniform bounds on P_ε^M and a_ε. The second tends to zero for the same reason, and the third tends to zero by strong continuity of P_ε^M. Hence

$$\widehat{P}_\varepsilon^M u \longrightarrow u \qquad \text{for every } u \in C^s_{\mathrm{ub}}(M).$$

Finally, fix $\varepsilon_* > 0$. By Lemma S18,

$$P_\varepsilon^M \longrightarrow P_{\varepsilon_*}^M \qquad \text{in } \mathcal{L}(C^s_{\mathrm{ub}}(M))$$

as $\varepsilon \to \varepsilon_*$. Since

$$m_\varepsilon = P_\varepsilon^M 1,$$

we also have

$$m_\varepsilon \longrightarrow m_{\varepsilon_*} \qquad \text{in } C^s_{\mathrm{ub}}(M).$$

The function m_{ε_*} is strictly positive on compact M, because the ambient Gaussian kernel is strictly positive. Therefore m_ε remains uniformly bounded away from zero for ε in a sufficiently small neighborhood of ε_*, and smooth composition gives

$$a_\varepsilon \longrightarrow a_{\varepsilon_*} \qquad \text{in } C^s_{\mathrm{ub}}(M).$$

Equivalently,

$$M_{a_\varepsilon} \longrightarrow M_{a_{\varepsilon_*}} \qquad \text{in } \mathcal{L}(C^s_{\mathrm{ub}}(M)).$$

Using

$$\widehat{P}_\varepsilon^M = M_{a_\varepsilon} P_\varepsilon^M M_{a_\varepsilon},$$

add and subtract

$$M_{a_{\varepsilon_*}} P_\varepsilon^M M_{a_\varepsilon} \quad \text{and} \quad M_{a_{\varepsilon_*}} P_{\varepsilon_*}^M M_{a_\varepsilon}.$$

The resulting three terms converge to zero in operator norm by the preceding multiplier convergence, operator-norm continuity of P_ε^M, and local uniform boundedness of all three families. Thus

$$\widehat{P}_\varepsilon^M \longrightarrow \widehat{P}_{\varepsilon_*}^M \qquad \text{in } \mathcal{L}(C^s_{\mathrm{ub}}(M)),$$

which proves positive-time operator-norm continuity. □

The preceding lemma identifies the infinitesimal generator on each fixed smooth input. For the inverse argument below, we also need a zeroth-order comparison in operator norm.

Lemma S20 (Comparison of heat-type kernels)**.** *Let*

$$A, B \in \Psi_H^{-1}(M)$$

in the notation of Grieser (2004, Definition 2.1), and let A_t, B_t denote the corresponding integral operators,

$$A_t u(x) := \int_M A(t,x,y)u(y)\,dV(y), \qquad B_t u(x) := \int_M B(t,x,y)u(y)\,dV(y).$$

Suppose that their leading terms agree:

$$\Phi_{-1}(A) = \Phi_{-1}(B).$$

Then, for every integer $s \geq 0$,

$$\|A_t - B_t\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \leq C_s\sqrt{t} \tag{S84}$$

for all sufficiently small $t > 0$.

Proof. Set

$$R := A - B.$$

Since $\Psi_H^{-1}(M)$ is a vector space,

$$R \in \Psi_H^{-1}(M).$$

The leading-term map is linear, so the assumption gives

$$\Phi_{-1}(R) = \Phi_{-1}(A) - \Phi_{-1}(B) = 0.$$

By Grieser (2004, Lemma 2.4(a)),

$$R \in \Psi_H^{-3/2}(M). \tag{S85}$$

We verify that this is precisely the additional $\sqrt{t}$ factor needed to apply Lemma S15. By Grieser (2004, Definition 2.1), membership $R \in \Psi_H^{\alpha}(M)$ means that, locally near the diagonal,

$$R(t,x,y) = t^{-(d+2)/2-\alpha}\widetilde{R}\left(\sqrt{t}, \frac{\xi-\eta}{\sqrt{t}}, \eta\right),$$

where ξ, η are local coordinates and $\widetilde{R}$ is smooth down to $t = 0$ and rapidly decreasing in the rescaled variable, uniformly with all derivatives. The same definition also requires R and all of its derivatives to be $O(t^N)$ away from the diagonal for every N.

Taking $\alpha = -3/2$ in this formula gives

$$-\frac{d+2}{2} - \alpha = -\frac{d+2}{2} + \frac{3}{2} = -\frac{d}{2} + \frac{1}{2}.$$

Thus every localized piece of R has the form

$$R(t,x,y) = t^{-d/2+1/2}\widetilde{R}\left(\sqrt{t}, \frac{\xi-\eta}{\sqrt{t}}, \eta\right). \tag{S86}$$

After multiplication by the standard localization cutoffs and absorption of the smooth Riemannian volume density, the rescaled amplitude still has the same uniform rapid-decay bounds. Hence (S86), together with the off-diagonal decay above, verifies the hypotheses of Lemma S15 with $\rho = 1$.

Applying that lemma to the integral operator with kernel $R = A - B$ gives

$$\|A_t - B_t\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \le C_s t^{1/2},$$

which proves (S84). □

Lemma S21 (Comparison with intrinsic heat)**.** *For every integer $s \ge 0$,*

$$\left\|\widehat{P}^M_\varepsilon - H^M_\varepsilon\right\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} = O(\sqrt{\varepsilon}). \tag{S87}$$

In particular,

$$\left\|\widehat{P}^M_\varepsilon - H^M_\varepsilon\right\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \longrightarrow 0 \qquad (\varepsilon \downarrow 0).$$

Proof. We first compare the raw ambient Gaussian with intrinsic heat. Put

$$t := \frac{\varepsilon}{4},$$

so that

$$H^M_\varepsilon = e^{t\Delta_M}$$

and

$$p^{\mathrm{amb}}_{4t}(x,y) = (4\pi t)^{-d/2}\exp\left(-\frac{\|\iota(x)-\iota(y)\|^2}{4t}\right). \tag{S88}$$

We claim that

$$p^{\mathrm{amb}}_{4t} \in \Psi^{-1}_H(M)$$

and that its leading term is

$$\Phi_{-1}(p^{\mathrm{amb}})(X,y) = (4\pi)^{-d/2} \exp\left(-\frac{\|X\|^2_{g(y)}}{4}\right). \tag{S89}$$

The off-diagonal condition in Grieser (2004, Definition 2.1) follows immediately from (S88). Indeed, outside any fixed neighborhood of the diagonal, compactness and injectivity of ι give

$$\|\iota(x) - \iota(y)\| \geq c > 0.$$

Any fixed number of t-, x-, and y-derivatives of p^{amb}_{4t} is therefore bounded by

$$Ct^{-N_0} e^{-c^2/(4t)} = O(t^N)$$

for every $N < \infty$.

It remains to verify the local heat-scale representation. Fix $y_0 \in M$ and choose a coordinate chart

$$\kappa : U \longrightarrow V \subset \mathbb{R}^d$$

around y_0. Shrinking the chart if necessary, we may take V convex and work on a smaller relatively compact coordinate neighborhood. Write

$$F := \iota \circ \kappa^{-1}, \qquad \eta := \kappa(y), \qquad r := \sqrt{t}.$$

For x near y, write

$$\kappa(x) = \eta + rX.$$

Then

$$F(\eta + rX) - F(\eta) = r \int_0^1 DF_{\eta+\theta rX}[X]\, d\theta.$$

Thus

$$G(r, X, \eta) := \frac{F(\eta + rX) - F(\eta)}{r} = \int_0^1 DF_{\eta+\theta rX}[X]\, d\theta$$

extends smoothly to $r = 0$, with

$$G(0, X, \eta) = DF_\eta[X]. \tag{S90}$$

Consequently,

$$Q(r, X, \eta) := \|G(r, X, \eta)\|^2$$

is smooth in (r, X, η) down to $r = 0$.

Choose a smooth cutoff in the coordinate neighborhood, equal to one on a smaller neighborhood of y_0. As in the coordinate-invariance argument of Grieser (2004, Lemma 2.3), this allows the local rescaled amplitude to be extended smoothly to all X. On the region where the cutoff is active,

$$p_{4t}^{\mathrm{amb}}(x, y) = t^{-d/2}\widetilde{p}(r, X, \eta),$$

where

$$\widetilde{p}(r, X, \eta) = (4\pi)^{-d/2} \exp\left(-\frac{1}{4}Q(r, X, \eta)\right) \tag{S91}$$

up to the fixed cutoff factor.

After shrinking the coordinate neighborhood once more if necessary, the smooth embedding F is uniformly bi-Lipschitz there. Hence

$$\|F(\xi) - F(\eta)\| \geq c\|\xi - \eta\|$$

for some $c > 0$, and therefore

$$Q(r, X, \eta) = \frac{\|F(\eta + rX) - F(\eta)\|^2}{r^2} \geq c^2\|X\|^2.$$

Repeated differentiation of (S91) produces a polynomial in X and derivatives of Q, multiplied by $e^{-Q/4}$. Smoothness of F on the fixed compact coordinate neighborhood bounds those derivatives by polynomials in $\|X\|$. Thus, for every collection of derivatives and every $N < \infty$,

$$\left|D_{r,X,\eta}^{\gamma}\widetilde{p}(r, X, \eta)\right| \leq C_{\gamma,N}(1 + \|X\|)^{-N}.$$

This is precisely the rapid-decay condition in Grieser (2004, Definition 2.1). Hence

$$p_{4t}^{\mathrm{amb}} \in \Psi_H^{-1}(M).$$

By (S90) and the fact that ι is isometric,

$$Q(0, X, \eta) = \|DF_\eta[X]\|_{\mathbb{R}^D}^2 = \|X\|_{g(y)}^2.$$

Evaluating the rescaled amplitude at $r = 0$ therefore gives (S89).

On the other hand, the intrinsic heat kernel $h_t^M(x,y)$ belongs to $\Psi_H^{-1}(M)$ and has leading term

$$\Phi_{-1}(h^M)(X,y) = (4\pi)^{-d/2} \exp\left(-\frac{\|X\|_{g(y)}^2}{4}\right); \tag{S92}$$

see Grieser (2004, Section 2.4.1). Hence

$$\Phi_{-1}(p^{\mathrm{amb}}) = \Phi_{-1}(h^M).$$

Lemma S20 now yields

$$\|P_{4t}^M - e^{t\Delta_M}\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \le C_s\sqrt{t}.$$

Since $t = \varepsilon/4$,

$$\|P_\varepsilon^M - H_\varepsilon^M\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \le C_s\sqrt{\varepsilon}. \tag{S93}$$

We next control the effect of degree normalization. Put

$$a_\varepsilon := m_\varepsilon^{-1/2}.$$

By (S60) and Lemma S14,

$$a_\varepsilon = 1 - \frac{\varepsilon}{2}\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon). \tag{S94}$$

In particular,

$$\|a_\varepsilon - 1\|_s = O(\varepsilon), \qquad \sup_{\varepsilon\le\varepsilon_0} \|a_\varepsilon\|_s < \infty.$$

Using

$$\widehat{P}_\varepsilon^M = M_{a_\varepsilon} P_\varepsilon^M M_{a_\varepsilon},$$

we have the exact decomposition

$$\widehat{P}_\varepsilon^M - P_\varepsilon^M = (M_{a_\varepsilon} - I)P_\varepsilon^M M_{a_\varepsilon} + P_\varepsilon^M(M_{a_\varepsilon} - I).$$

The multiplier estimate from Lemma S14 gives

$$\|M_{a_\varepsilon} - I\|_{\mathcal{L}(C^s_{\mathrm{ub}})} \le C_s\|a_\varepsilon - 1\|_s = O(\varepsilon),$$

while M_{a_ε} and P_ε^M are uniformly bounded. Therefore

$$\|\widehat{P}_\varepsilon^M - P_\varepsilon^M\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} = O(\varepsilon). \tag{S95}$$

Combining (S93) and (S95),

$$\begin{aligned}\|\widehat{P}_\varepsilon^M - H_\varepsilon^M\| &\le \|\widehat{P}_\varepsilon^M - P_\varepsilon^M\| + \|P_\varepsilon^M - H_\varepsilon^M\| \\ &= O(\varepsilon) + O(\sqrt{\varepsilon}) = O(\sqrt{\varepsilon}),\end{aligned}$$

in $\mathcal{L}(C^s_{\mathrm{ub}}(M))$. This proves (S87). □

Lemma S22 (Uniform inverse inherited by the normalized ambient kernel)**.** *For all sufficiently small $\varepsilon \ge 0$, $I + \widehat{P}_\varepsilon^M$ is invertible on $C^s_{\mathrm{ub}}(M)$, with*

$$\sup_{0\le\varepsilon\le\varepsilon_0} \left\|(I + \widehat{P}_\varepsilon^M)^{-1}\right\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} < \infty. \tag{S96}$$

Moreover, for every fixed $v \in C^s_{\mathrm{ub}}(M)$,

$$(I + \widehat{P}_\varepsilon^M)^{-1} v \longrightarrow \frac{1}{2} v \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S97}$$

Proof. Put

$$R_\varepsilon := \widehat{P}_\varepsilon^M - H_\varepsilon^M.$$

Lemma S21 gives $\|R_\varepsilon\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \to 0$. Apply Lemma S9 with

$$A_\varepsilon := I + H_\varepsilon^M, \qquad E_\varepsilon := R_\varepsilon.$$

The uniform inverse bound and strong inverse limit for A_ε are provided by Lemma S17. Hence $I + \widehat{P}_\varepsilon^M = A_\varepsilon + E_\varepsilon$ is invertible for all sufficiently small ε, its inverse is uniformly bounded, and

$$\|(I + \widehat{P}_\varepsilon^M)^{-1} - (I + H_\varepsilon^M)^{-1}\|_{\mathcal{L}(C^s_{\mathrm{ub}}(M))} \longrightarrow 0.$$

Combining this operator-norm convergence with

$$(I + H_\varepsilon^M)^{-1} v \longrightarrow \frac{1}{2} v$$

for each fixed v proves (S97). At $\varepsilon = 0$, $\widehat{P}_0^M = I$, so the inverse is $\frac{1}{2} I$. □

4.D.2 Product equations and the entropic coupling

Fix

$$h \in C^{s+2}_{\mathrm{ub}}(M), \qquad h > 0, \qquad \int_M h^2 \, dV = 1. \tag{S98}$$

Compactness implies that h is bounded away from zero and hence that $h^{-1} \in C^{s+2}_{\mathrm{ub}}(M)$. Define

$$\mathcal{G}^{\mathrm{raw}}_{\varepsilon}(g) := gP^{M}_{\varepsilon}g - h^2, \tag{S99}$$

$$\mathcal{G}_{\varepsilon}(u) := u\widehat{P}^{M}_{\varepsilon}u - h^2. \tag{S100}$$

Lemma S23 (Exact degree conjugacy and marginal equation)**.** *Let $\varepsilon > 0$, let $g > 0$, and put*

$$u := m_{\varepsilon}^{1/2} g.$$

Then

$$\mathcal{G}^{\mathrm{raw}}_{\varepsilon}(g) = \mathcal{G}_{\varepsilon}(u). \tag{S101}$$

Moreover, the measure

$$d\pi_{\varepsilon,g}(x,y) := p^{\mathrm{amb}}_{\varepsilon}(x,y)g(x)g(y)\,dV(x)\,dV(y) \tag{S102}$$

can equivalently be written as

$$d\pi_{\varepsilon,g}(x,y) = \widehat{p}^{M}_{\varepsilon}(x,y)u(x)u(y)\,dV(x)\,dV(y).$$

The following are equivalent:

$$\mathcal{G}^{\mathrm{raw}}_{\varepsilon}(g) = 0, \qquad \mathcal{G}_{\varepsilon}(u) = 0, \qquad \pi_{\varepsilon,g} \in \Pi(\mu,\mu), \quad d\mu := h^2 dV.$$

Proof. Since

$$\widehat{P}^{M}_{\varepsilon} = M_{m_{\varepsilon}^{-1/2}} P^{M}_{\varepsilon} M_{m_{\varepsilon}^{-1/2}},$$

we have

$$u\widehat{P}^{M}_{\varepsilon}u = m_{\varepsilon}^{1/2} g\, m_{\varepsilon}^{-1/2} P^{M}_{\varepsilon} g = gP^{M}_{\varepsilon}g,$$

which proves (S101). The kernel identity follows from $u = m_{\varepsilon}^{1/2} g$ and the definition of $\widehat{p}^{M}_{\varepsilon}$.

Integrating (S102) in y gives first marginal density

$$g(x)P^{M}_{\varepsilon}g(x).$$

Thus the first marginal equals $h^2 dV$ exactly when $\mathcal{G}^{\mathrm{raw}}_{\varepsilon}(g) = 0$. Symmetry of the kernel gives the same condition for the second marginal. The equivalence with the normalized equation follows from (S101). □

Lemma S24 (Uniqueness of a positive product root)**.** *For each $\varepsilon > 0$, the equation*

$$u\widehat{P}_\varepsilon^M u = h^2$$

has at most one strictly positive continuous solution. Consequently, the raw equation

$$gP_\varepsilon^M g = h^2$$

also has at most one strictly positive continuous solution.

Proof. Suppose $u, \widetilde{u} > 0$ are two normalized roots and set

$$r := \log(u/\widetilde{u}).$$

Dividing their equations gives

$$1 = e^{r(x)} \frac{\widehat{P}_\varepsilon^M(e^r\widetilde{u})(x)}{\widehat{P}_\varepsilon^M\widetilde{u}(x)}. \tag{S103}$$

Let $M_r := \max_M r$ and $m_r := \min_M r$. At a maximizer x_M, the quotient in (S103) is a weighted average of e^r with strictly positive weights, and is therefore at least e^{m_r}. Hence

$$1 \geq e^{M_r+m_r}.$$

At a minimizer x_m, the same weighted average is at most e^{M_r}, so

$$1 \leq e^{m_r+M_r}.$$

Thus $M_r + m_r = 0$. Equality at x_M forces a strictly positive weighted average of e^r to equal its minimum. Since $\widehat{p}_\varepsilon^M(x_M, y) > 0$ and $\widetilde{u}(y) > 0$ for every y, this is possible only if e^r is constant. Substitution into (S103) then shows that the constant is one. Therefore $u = \widetilde{u}$. Exact conjugacy gives uniqueness for the raw equation. □

Lemma S25 (Gibbs potential and entropic optimality)**.** *Suppose $g_\varepsilon > 0$ solves the raw product equation. Then the coupling $\pi_{\varepsilon,g_\varepsilon}$ from* (S102) *is the unique minimizer of*

$$\gamma \longmapsto \int_{M\times M} \|\iota(x) - \iota(y)\|^2 \, d\gamma(x, y) + \varepsilon \, \mathrm{KL}(\gamma\|\mu \otimes \mu)$$

over $\Pi(\mu, \mu)$. Its symmetric potential is

$$f_\varepsilon = \varepsilon \log\left(\frac{c_\varepsilon^{1/2} g_\varepsilon}{h^2}\right). \tag{S104}$$

Proof. By Lemma S23, $\pi_{\varepsilon,g_\varepsilon} \in \Pi(\mu,\mu)$. Formula (S104) gives

$$\frac{d\pi_{\varepsilon,g_\varepsilon}}{d(\mu\otimes\mu)}(x,y) = \exp\left(\frac{f_\varepsilon(x)+f_\varepsilon(y)-\|\iota(x)-\iota(y)\|^2}{\varepsilon}\right).$$

For any $\gamma \in \Pi(\mu,\mu)$ with finite objective, expansion of the relative entropy with respect to $\pi_{\varepsilon,g_\varepsilon}$ cancels the two potential terms because the marginals agree, and gives

$$J_\varepsilon(\gamma) - J_\varepsilon(\pi_{\varepsilon,g_\varepsilon}) = \varepsilon\,\mathrm{KL}\big(\gamma\|\pi_{\varepsilon,g_\varepsilon}\big) \geq 0.$$

Equality holds only when the two couplings coincide. All quantities are finite for the candidate coupling because M is compact and h, g_ε are strictly positive and continuous. □

4.D.3 Functional analysis of the normalized product equation

In this subsection,

$$X := C^s_{\mathrm{ub}}(M), \qquad K_\varepsilon := \widehat{P}^M_\varepsilon,$$

and $\mathcal{G}_\varepsilon$ is the normalized functional in (S100).

Lemma S26 (Differential of the product functional)**.** *For every $\varepsilon \geq 0$, $\mathcal{G}_\varepsilon : X \to X$ is continuously Fréchet differentiable, with*

$$D\mathcal{G}_\varepsilon(u)[v] = vK_\varepsilon u + uK_\varepsilon v. \tag{S105}$$

There is $C < \infty$, independent of sufficiently small ε, such that

$$\|D\mathcal{G}_\varepsilon(u) - D\mathcal{G}_\varepsilon(\widetilde{u})\|_{\mathcal{L}(X)} \leq C\|u-\widetilde{u}\|_s. \tag{S106}$$

At the base point h, define

$$B_\varepsilon := D\mathcal{G}_\varepsilon(h) = M_{K_\varepsilon h} + M_h K_\varepsilon. \tag{S107}$$

Then

$$B_\varepsilon = M_h(I+K_\varepsilon) + M_{K_\varepsilon h - h}. \tag{S108}$$

Proof. Recall that

$$X = C^s_{\mathrm{ub}}(M)$$

and, by definition,

$$\mathcal{G}_\varepsilon(u) = uK_\varepsilon u - h^2.$$

By Lemma S14, X is a Banach algebra: there is $C_s < \infty$ such that

$$\|fg\|_s \le C_s \|f\|_s \|g\|_s, \qquad f, g \in X. \tag{S109}$$

In particular, multiplication by $f \in X$ defines a bounded operator on X, with

$$\|M_f\|_{\mathcal{L}(X)} \le C_s \|f\|_s.$$

Moreover, by Lemma S19, $K_\varepsilon : X \to X$ is bounded for every ε, and, for some $\varepsilon_0 > 0$,

$$C_K := \sup_{0 \le \varepsilon \le \varepsilon_0} \|K_\varepsilon\|_{\mathcal{L}(X)} < \infty. \tag{S110}$$

Thus $\mathcal{G}_\varepsilon(u) \in X$ whenever $u \in X$.

Fix $\varepsilon \ge 0$ and $u, v \in X$. Since K_ε is linear,

$$\begin{aligned} \mathcal{G}_\varepsilon(u+v) - \mathcal{G}_\varepsilon(u) &= (u+v)K_\varepsilon(u+v) - uK_\varepsilon u \\ &= vK_\varepsilon u + uK_\varepsilon v + vK_\varepsilon v. \end{aligned} \tag{S111}$$

Define

$$L_{\varepsilon,u} v := vK_\varepsilon u + uK_\varepsilon v.$$

This map is linear in v. It is also bounded: by (S109),

$$\begin{aligned} \|L_{\varepsilon,u} v\|_s &\le C_s \|v\|_s \|K_\varepsilon u\|_s + C_s \|u\|_s \|K_\varepsilon v\|_s \\ &\le 2C_s \|K_\varepsilon\|_{\mathcal{L}(X)} \|u\|_s \|v\|_s. \end{aligned}$$

Hence $L_{\varepsilon,u} \in \mathcal{L}(X)$.

The final term in (S111) is quadratic in the increment. Indeed,

$$\|vK_\varepsilon v\|_s \le C_s \|K_\varepsilon\|_{\mathcal{L}(X)} \|v\|_s^2.$$

Therefore

$$\frac{\|\mathcal{G}_\varepsilon(u+v) - \mathcal{G}_\varepsilon(u) - L_{\varepsilon,u} v\|_s}{\|v\|_s} \le C_s \|K_\varepsilon\|_{\mathcal{L}(X)} \|v\|_s \longrightarrow 0$$

as $\|v\|_s \to 0$, proving (S105).

We next control the dependence of the differential on the base point. From (S105), for any $v \in X$,

$$[D\mathcal{G}_\varepsilon(u) - D\mathcal{G}_\varepsilon(\widetilde{u})]\, v = vK_\varepsilon(u-\widetilde{u}) + (u-\widetilde{u})K_\varepsilon v.$$

Using again the algebra estimate gives

$$\|[D\mathcal{G}_\varepsilon(u) - D\mathcal{G}_\varepsilon(\widetilde{u})]\, v\|_s \le 2C_s\|K_\varepsilon\|_{\mathcal{L}(X)}\|u-\widetilde{u}\|_s\|v\|_s.$$

Taking the supremum over $\|v\|_s \le 1$ yields

$$\|D\mathcal{G}_\varepsilon(u) - D\mathcal{G}_\varepsilon(\widetilde{u})\|_{\mathcal{L}(X)} \le 2C_s\|K_\varepsilon\|_{\mathcal{L}(X)}\|u-\widetilde{u}\|_s. \tag{S112}$$

For each fixed ε, this proves continuity of $u \mapsto D\mathcal{G}_\varepsilon(u)$ and hence $\mathcal{G}_\varepsilon \in C^1(X;X)$. Using the uniform bound (S110), the constant on the right-hand side of (S112) can be chosen independently of $0 \le \varepsilon \le \varepsilon_0$, proving (S106).

Finally, setting $u = h$ in (S105) gives

$$B_\varepsilon v = vK_\varepsilon h + hK_\varepsilon v = M_{K_\varepsilon h}v + M_h K_\varepsilon v,$$

so

$$B_\varepsilon = M_{K_\varepsilon h} + M_h K_\varepsilon,$$

which proves (S107). Adding and subtracting M_h gives

$$\begin{aligned} B_\varepsilon &= M_h + M_h K_\varepsilon + \big(M_{K_\varepsilon h} - M_h\big) \\ &= M_h(I + K_\varepsilon) + M_{K_\varepsilon h - h}, \end{aligned}$$

proving (S108). □

Lemma S27 (Uniform inverse of the base differential)**.** *For all sufficiently small $\varepsilon \ge 0$, B_ε is invertible on X, with*

$$\sup_\varepsilon \|B_\varepsilon^{-1}\|_{\mathcal{L}(X)} < \infty. \tag{S113}$$

For every fixed $v \in X$,

$$B_\varepsilon^{-1} v \longrightarrow \frac{v}{2h} \qquad \text{in } X. \tag{S114}$$

Proof. Set

$$L_\varepsilon := M_h(I + K_\varepsilon), \qquad E_\varepsilon := M_{K_\varepsilon h - h}.$$

Then $B_\varepsilon = L_\varepsilon + E_\varepsilon$. By Lemma S22,

$$L_\varepsilon^{-1} = (I + K_\varepsilon)^{-1} M_{h^{-1}}$$

is uniformly bounded. Moreover, for each fixed $v \in X$,

$$L_\varepsilon^{-1} v = (I + K_\varepsilon)^{-1}(h^{-1}v) \longrightarrow \frac{v}{2h}.$$

Strong continuity of K_ε at the fixed function h gives

$$\|K_\varepsilon h - h\|_s \longrightarrow 0,$$

and hence, by the multiplier estimate,

$$\|E_\varepsilon\|_{\mathcal{L}(X)} \longrightarrow 0.$$

Lemma S9, applied with base operator L_ε and perturbation E_ε, now proves that B_ε is invertible for all sufficiently small ε, its inverse is uniformly bounded, and

$$B_\varepsilon^{-1} v \longrightarrow \frac{v}{2h}$$

for every fixed $v \in X$. □

Lemma S28 (Preconditioned contraction and solution branch)**.** *There exist $r > 0$, $\varepsilon_0 > 0$, and $q \in (0, 1)$ such that*

$$T_\varepsilon(u) := u - B_\varepsilon^{-1}\mathcal{G}_\varepsilon(u) \tag{S115}$$

is a contraction with constant q from

$$B_r(h) := \{u \in X : \|u - h\|_s \le r\}$$

into itself for every $0 \le \varepsilon \le \varepsilon_0$. Its fixed point u_ε is the unique positive solution of

$$uK_\varepsilon u = h^2.$$

The path $\varepsilon \mapsto u_\varepsilon$ is continuous in X, and

$$\|u_\varepsilon - h\|_s \le \frac{1}{1-q} \left\|B_\varepsilon^{-1}\mathcal{G}_\varepsilon(h)\right\|_s \longrightarrow 0. \tag{S116}$$

Proof. By the preceding uniform-inverse lemma for B_ε, there exists $\varepsilon_1 > 0$ such that B_ε is invertible for $0 \le \varepsilon \le \varepsilon_1$ and

$$C_{\text{inv}} := \sup_{0 \le \varepsilon \le \varepsilon_1} \|B_\varepsilon^{-1}\|_{\mathcal{L}(X)} < \infty. \tag{S117}$$

For each such ε, B_ε^{-1} is independent of the variable u. Since $\mathcal{G}_\varepsilon$ is continuously Fréchet differentiable by Lemma S26, differentiation of (S115) gives

$$\begin{aligned} DT_\varepsilon(u) &= I - B_\varepsilon^{-1} D\mathcal{G}_\varepsilon(u) \\ &= B_\varepsilon^{-1} \left[B_\varepsilon - D\mathcal{G}_\varepsilon(u) \right], \end{aligned} \tag{S118}$$

because $B_\varepsilon = D\mathcal{G}_\varepsilon(h)$. In particular,

$$DT_\varepsilon(h) = 0$$

for every $0 \le \varepsilon \le \varepsilon_1$.

By (S106), there is a constant C, independent of sufficiently small ε, such that

$$\|B_\varepsilon - D\mathcal{G}_\varepsilon(u)\|_{\mathcal{L}(X)} = \|D\mathcal{G}_\varepsilon(h) - D\mathcal{G}_\varepsilon(u)\|_{\mathcal{L}(X)} \le C\|u - h\|_s.$$

Combining this with (S117) and (S118) gives

$$\|DT_\varepsilon(u)\|_{\mathcal{L}(X)} \le C_{\text{inv}} C \|u - h\|_s. \tag{S119}$$

Since $h > 0$ and M is compact,

$$h_{\min} := \inf_{x \in M} h(x) > 0.$$

Choose $r > 0$ sufficiently small that

$$q := C_{\text{inv}} C r < 1, \qquad r < \frac{1}{2} h_{\min}.$$

Because the C^s_{ub} norm controls the supremum norm, every $u \in B_r(h)$ satisfies

$$u(x) \ge h(x) - \|u - h\|_\infty \ge h_{\min} - r > \frac{1}{2} h_{\min} > 0.$$

Thus every function in $B_r(h)$ is strictly positive.

Moreover, $B_r(h)$ is convex. Hence for $u, \widetilde{u} \in B_r(h)$, the segment

$$u_t := \widetilde{u} + t(u - \widetilde{u}), \qquad 0 \le t \le 1,$$

also lies in $B_r(h)$. The Banach-space fundamental theorem of calculus gives

$$T_\varepsilon(u) - T_\varepsilon(\widetilde{u}) = \int_0^1 DT_\varepsilon(u_t)[u - \widetilde{u}]\, dt.$$

By (S119), $\|DT_\varepsilon(u_t)\| \le C_{\mathrm{inv}} C r = q$, and therefore

$$\|T_\varepsilon(u) - T_\varepsilon(\widetilde{u})\|_s \le q\|u - \widetilde{u}\|_s. \tag{S120}$$

Thus T_ε has a common contraction constant $q < 1$ on $B_r(h)$.

It remains to show that this ball is invariant. At its center,

$$T_\varepsilon(h) - h = -B_\varepsilon^{-1}\mathcal{G}_\varepsilon(h).$$

Since

$$\mathcal{G}_\varepsilon(h) = hK_\varepsilon h - h^2 = h(K_\varepsilon h - h),$$

strong continuity of K_ε at zero on the fixed function h, together with the Banach-algebra estimate, gives

$$\|\mathcal{G}_\varepsilon(h)\|_s \longrightarrow 0.$$

The uniform inverse bound (S117) therefore implies

$$\|T_\varepsilon(h) - h\|_s = \|B_\varepsilon^{-1}\mathcal{G}_\varepsilon(h)\|_s \longrightarrow 0. \tag{S121}$$

Decrease ε_1, if necessary, and denote the resulting upper bound by ε_0, so that

$$\|T_\varepsilon(h) - h\|_s \le (1 - q)r, \qquad 0 \le \varepsilon \le \varepsilon_0.$$

Then for $u \in B_r(h)$,

$$\begin{aligned}
\|T_\varepsilon(u) - h\|_s &\le \|T_\varepsilon(u) - T_\varepsilon(h)\|_s + \|T_\varepsilon(h) - h\|_s \\
&\le q\|u - h\|_s + (1 - q)r \\
&\le qr + (1 - q)r = r.
\end{aligned}$$

Thus

$$T_\varepsilon(B_r(h)) \subseteq B_r(h).$$

By Lemma S14, X is a Banach space, so the closed ball $B_r(h)$ is complete. Banach's fixed-point theorem therefore gives a unique $u_\varepsilon \in B_r(h)$ satisfying

$$T_\varepsilon(u_\varepsilon) = u_\varepsilon.$$

Because B_ε is invertible,

$$T_\varepsilon(u_\varepsilon) = u_\varepsilon \quad \Longleftrightarrow \quad \mathcal{G}_\varepsilon(u_\varepsilon) = 0,$$

and hence

$$u_\varepsilon K_\varepsilon u_\varepsilon = h^2.$$

The choice of r implies $u_\varepsilon > 0$. For $\varepsilon > 0$, Lemma S24 shows that this is the unique positive root of the product equation. At $\varepsilon = 0$, $K_0 = I$, so the equation becomes

$$u^2 = h^2,$$

whose unique positive solution is $u = h$. In particular, $u_0 = h$.

Applying (S120) to u_ε and h gives

$$\begin{aligned} \|u_\varepsilon - h\|_s &= \|T_\varepsilon(u_\varepsilon) - h\|_s \\ &\le \|T_\varepsilon(u_\varepsilon) - T_\varepsilon(h)\|_s + \|T_\varepsilon(h) - h\|_s \\ &\le q\|u_\varepsilon - h\|_s + \|B_\varepsilon^{-1}\mathcal{G}_\varepsilon(h)\|_s. \end{aligned}$$

Therefore

$$(1-q)\|u_\varepsilon - h\|_s \le \|B_\varepsilon^{-1}\mathcal{G}_\varepsilon(h)\|_s,$$

which proves (S116). Its right-hand side tends to zero by (S121). Since $u_0 = h$, this also proves continuity of $\varepsilon \mapsto u_\varepsilon$ at $\varepsilon = 0$.

It remains to prove continuity at positive ε. Fix $\varepsilon_* \in (0, \varepsilon_0]$. By the previously established operator-norm continuity of K_ε for positive ε,

$$\|K_\varepsilon - K_{\varepsilon_*}\|_{\mathcal{L}(X)} \longrightarrow 0 \qquad (\varepsilon \to \varepsilon_*).$$

For $u \in B_r(h)$,

$$\mathcal{G}_\varepsilon(u) - \mathcal{G}_{\varepsilon_*}(u) = u(K_\varepsilon - K_{\varepsilon_*})u.$$

Since

$$\sup_{u \in B_r(h)} \|u\|_s \le \|h\|_s + r,$$

the Banach-algebra estimate gives

$$\sup_{u \in B_r(h)} \|\mathcal{G}_\varepsilon(u) - \mathcal{G}_{\varepsilon_*}(u)\|_s \longrightarrow 0. \tag{S122}$$

Similarly, from

$$B_\varepsilon = M_{K_\varepsilon h} + M_h K_\varepsilon,$$

the multiplier estimate gives

$$\|B_\varepsilon - B_{\varepsilon_*}\|_{\mathcal{L}(X)} \longrightarrow 0.$$

Since the inverses are uniformly bounded, the resolvent identity

$$B_\varepsilon^{-1} - B_{\varepsilon_*}^{-1} = B_\varepsilon^{-1}(B_{\varepsilon_*} - B_\varepsilon)B_{\varepsilon_*}^{-1}$$

implies

$$\|B_\varepsilon^{-1} - B_{\varepsilon_*}^{-1}\|_{\mathcal{L}(X)} \longrightarrow 0. \tag{S123}$$

Finally,

$$\begin{aligned} T_\varepsilon(u) - T_{\varepsilon_*}(u) = & -\big(B_\varepsilon^{-1} - B_{\varepsilon_*}^{-1}\big)\mathcal{G}_\varepsilon(u) \\ & - B_{\varepsilon_*}^{-1}\big(\mathcal{G}_\varepsilon(u) - \mathcal{G}_{\varepsilon_*}(u)\big). \end{aligned}$$

The family $\mathcal{G}_\varepsilon(u)$ is uniformly bounded for $u \in B_r(h)$ and ε near ε_*, by the uniform bound for K_ε and the Banach-algebra estimate. Therefore (S122) and (S123) imply

$$\sup_{u \in B_r(h)} \|T_\varepsilon(u) - T_{\varepsilon_*}(u)\|_s \longrightarrow 0.$$

Using the two fixed-point equations,

$$\begin{aligned} \|u_\varepsilon - u_{\varepsilon_*}\|_s &= \|T_\varepsilon(u_\varepsilon) - T_{\varepsilon_*}(u_{\varepsilon_*})\|_s \\ &\le \|T_\varepsilon(u_\varepsilon) - T_\varepsilon(u_{\varepsilon_*})\|_s \\ &\quad + \|T_\varepsilon(u_{\varepsilon_*}) - T_{\varepsilon_*}(u_{\varepsilon_*})\|_s \\ &\le q\|u_\varepsilon - u_{\varepsilon_*}\|_s + \sup_{u \in B_r(h)} \|T_\varepsilon(u) - T_{\varepsilon_*}(u)\|_s. \end{aligned}$$

Hence

$$(1-q)\|u_\varepsilon - u_{\varepsilon_*}\|_s \le \sup_{u\in B_r(h)} \|T_\varepsilon(u) - T_{\varepsilon_*}(u)\|_s \longrightarrow 0.$$

Thus $\varepsilon \mapsto u_\varepsilon$ is continuous at every positive ε_*, completing the proof. □

Lemma S29 (Averaged differential)**.** *Let u_ε be the branch from Lemma S28, and define*

$$\overline{B}_\varepsilon := \int_0^1 D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big)\, dt. \tag{S124}$$

Then

$$\|\overline{B}_\varepsilon - B_\varepsilon\|_{\mathcal{L}(X)} \longrightarrow 0. \tag{S125}$$

For all sufficiently small ε, $\overline{B}_\varepsilon$ is invertible with uniformly bounded inverse, and, for each fixed $v \in X$,

$$\overline{B}_\varepsilon^{-1} v \longrightarrow \frac{v}{2h} \qquad \text{in } X. \tag{S126}$$

Moreover,

$$u_\varepsilon - h = -\overline{B}_\varepsilon^{-1}\mathcal{G}_\varepsilon(h). \tag{S127}$$

Proof. We first compare the averaged differential with the differential at the base point. Since X is a Banach space, $\mathcal{L}(X)$ is also Banach. Moreover, Lemma S26 shows that $u \mapsto D\mathcal{G}_\varepsilon(u)$ is continuous in operator norm. Hence

$$t \longmapsto D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big)$$

is continuous from $[0,1]$ into $\mathcal{L}(X)$, so the integral in (S124) is well defined as a Banach-space-valued integral.

Recall that

$$B_\varepsilon = D\mathcal{G}_\varepsilon(h).$$

Therefore

$$\overline{B}_\varepsilon - B_\varepsilon = \int_0^1 \big[D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big) - D\mathcal{G}_\varepsilon(h)\big]\, dt. \tag{S128}$$

By the uniform Lipschitz estimate (S106),

$$\big\|D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big) - D\mathcal{G}_\varepsilon(h)\big\|_{\mathcal{L}(X)} \le Ct\|u_\varepsilon - h\|_s.$$

Using the standard norm inequality for Banach-space-valued integrals in (S128) gives

$$\begin{aligned}\|\overline{B}_\varepsilon - B_\varepsilon\|_{\mathcal{L}(X)} &\le \int_0^1 Ct\|u_\varepsilon - h\|_s \, dt \\ &= \frac{C}{2}\|u_\varepsilon - h\|_s. \end{aligned} \tag{S129}$$

Lemma S28 gives $u_\varepsilon \to h$ in X, and therefore

$$\|\overline{B}_\varepsilon - B_\varepsilon\|_{\mathcal{L}(X)} \longrightarrow 0,$$

proving (S125).

Set

$$E_\varepsilon := \overline{B}_\varepsilon - B_\varepsilon.$$

By (S125), $\|E_\varepsilon\|_{\mathcal{L}(X)} \to 0$. Applying Lemma S9 with base operator B_ε and perturbation E_ε, and using Lemma S27, shows that $\overline{B}_\varepsilon$ is invertible for all sufficiently small ε, its inverse is uniformly bounded, and

$$\|\overline{B}_\varepsilon^{-1} - B_\varepsilon^{-1}\|_{\mathcal{L}(X)} \longrightarrow 0.$$

Consequently, for each fixed $v \in X$,

$$\overline{B}_\varepsilon^{-1} v \longrightarrow \frac{v}{2h},$$

which proves (S126). Here $v/(2h) \in X$ because h is bounded away from zero and $h^{-1} \in X$ by Lemma S14.

Finally, we derive the exact root identity. Since $\mathcal{G}_\varepsilon : X \to X$ is continuously Fréchet differentiable, the map

$$F_\varepsilon(t) := \mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big), \qquad 0 \le t \le 1,$$

is continuously differentiable as an X-valued map. By the chain rule,

$$F_\varepsilon'(t) = D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big)[u_\varepsilon - h].$$

The Banach-space fundamental theorem of calculus therefore gives

$$\begin{aligned}\mathcal{G}_\varepsilon(u_\varepsilon) - \mathcal{G}_\varepsilon(h) &= \int_0^1 D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big)[u_\varepsilon - h] \, dt \\ &= \left[\int_0^1 D\mathcal{G}_\varepsilon\big(h + t(u_\varepsilon - h)\big) \, dt\right](u_\varepsilon - h) \\ &= \overline{B}_\varepsilon(u_\varepsilon - h).\end{aligned}$$

Since u_ε is a root, $\mathcal{G}_\varepsilon(u_\varepsilon) = 0$, so

$$-\mathcal{G}_\varepsilon(h) = \overline{B}_\varepsilon(u_\varepsilon - h).$$

For sufficiently small ε, $\overline{B}_\varepsilon$ is invertible by the preceding argument. Applying its inverse yields

$$u_\varepsilon - h = -\overline{B}_\varepsilon^{-1}\mathcal{G}_\varepsilon(h),$$

which is (S127). □

4.D.4 Expansion and comparison with intrinsic heat

Theorem S1 (Ambient-Gaussian compact-manifold expansion)**.** *Assume* (S98)*. Then there exists $\varepsilon_0 > 0$ such that, for every $0 \le \varepsilon \le \varepsilon_0$, there are unique positive functions*

$$u_\varepsilon, g_\varepsilon \in C^s_{\mathrm{ub}}(M)$$

satisfying

$$u_\varepsilon \widehat{P}^M_\varepsilon u_\varepsilon = h^2, \qquad g_\varepsilon P^M_\varepsilon g_\varepsilon = h^2, \qquad u_\varepsilon = m_\varepsilon^{1/2} g_\varepsilon, \tag{S130}$$

with $u_0 = g_0 = h$. Both branches are continuous in $C^s_{\mathrm{ub}}(M)$, and

$$\frac{u_\varepsilon - h}{\varepsilon} \longrightarrow -\frac{1}{8}\Delta_M h, \tag{S131}$$

$$\frac{g_\varepsilon - h}{\varepsilon} \longrightarrow -\frac{1}{8}\Delta_M h - \frac{1}{2}\omega_M h \tag{S132}$$

in $C^s_{\mathrm{ub}}(M)$.

Equivalently, if

$$\phi_\varepsilon := \log\left(\frac{u_\varepsilon}{h}\right), \tag{S133}$$

then

$$\frac{\phi_\varepsilon}{\varepsilon} \longrightarrow -\frac{1}{8}\frac{\Delta_M h}{h} \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S134}$$

For every sufficiently small $\varepsilon > 0$, the coupling $\pi_{\varepsilon,g_\varepsilon}$ is the unique entropic self-transport plan for $\mu = h^2 dV$ under the squared ambient cost. Its symmetric potential satisfies

$$f_\varepsilon = -\frac{\varepsilon d}{4}\log(\pi\varepsilon) - \varepsilon \log h - \varepsilon^2\left[\frac{1}{8}\frac{\Delta_M h}{h} + \frac{1}{2}\omega_M\right] + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon^2). \tag{S135}$$

Proof. We first construct the normalized and raw solution branches. Lemma S28 gives, for all sufficiently small $\varepsilon \geq 0$, a positive solution u_ε of

$$u_\varepsilon \widehat{P}_\varepsilon^M u_\varepsilon = h^2,$$

with $u_0 = h$, and shows that $\varepsilon \mapsto u_\varepsilon$ is continuous in $C^s_{\mathrm{ub}}(M)$. For $\varepsilon > 0$, Lemma S24 makes this the unique positive solution; at $\varepsilon = 0$ the equation is $u^2 = h^2$, whose unique positive solution is $u = h$.

By the exact conjugacy of the raw and normalized product equations,

$$g_\varepsilon := m_\varepsilon^{-1/2} u_\varepsilon$$

is the unique positive solution of

$$g_\varepsilon P_\varepsilon^M g_\varepsilon = h^2,$$

and

$$u_\varepsilon = m_\varepsilon^{1/2} g_\varepsilon.$$

Since $m_\varepsilon \to 1$ in $C^s_{\mathrm{ub}}(M)$ as $\varepsilon \downarrow 0$, smooth composition gives $m_\varepsilon^{-1/2} \to 1$, so $g_\varepsilon \to h$ at zero. Positive-time continuity follows similarly from the positive-time operator-norm continuity of P_ε^M, since $m_\varepsilon = P_\varepsilon^M 1$. Thus both branches are continuous.

We next identify the first-order expansion of the normalized branch. Recall that

$$\mathcal{G}_\varepsilon(u) = u \widehat{P}_\varepsilon^M u - h^2.$$

At the base point h,

$$\mathcal{G}_\varepsilon(h) = h\big(\widehat{P}_\varepsilon^M h - h\big). \tag{S136}$$

Since $h \in C^{s+2}_{\mathrm{ub}}(M)$, the normalized-generator expansion from Lemma S19 gives

$$\frac{\widehat{P}_\varepsilon^M h - h}{\varepsilon} \longrightarrow \frac{1}{4}\Delta_M h \qquad \text{in } C^s_{\mathrm{ub}}(M).$$

Multiplication by the fixed function h is bounded on $C^s_{\mathrm{ub}}(M)$, and hence

$$v_\varepsilon := \frac{\mathcal{G}_\varepsilon(h)}{\varepsilon} \longrightarrow v := \frac{1}{4} h \Delta_M h \qquad \text{in } C^s_{\mathrm{ub}}(M). \tag{S137}$$

Lemma S29 gives the exact identity

$$u_\varepsilon - h = -\overline{B}_\varepsilon^{-1}\mathcal{G}_\varepsilon(h).$$

Therefore

$$\frac{u_\varepsilon - h}{\varepsilon} = -\overline{B}_\varepsilon^{-1} v_\varepsilon. \tag{S138}$$

The averaged inverse family is uniformly bounded, and for each fixed $w \in C^s_{\mathrm{ub}}(M)$,

$$\overline{B}_\varepsilon^{-1} w \longrightarrow \frac{w}{2h}.$$

Although v_ε varies with ε, we may write

$$\begin{aligned}\left\|\overline{B}_\varepsilon^{-1} v_\varepsilon - \frac{v}{2h}\right\|_s &\le \|\overline{B}_\varepsilon^{-1}\|_{\mathcal{L}(X)}\|v_\varepsilon - v\|_s \\ &\quad + \left\|\overline{B}_\varepsilon^{-1} v - \frac{v}{2h}\right\|_s.\end{aligned}$$

The first term tends to zero by (S137) and the uniform inverse bound, and the second by the strong inverse limit. Thus

$$\overline{B}_\varepsilon^{-1} v_\varepsilon \longrightarrow \frac{v}{2h} = \frac{1}{8}\Delta_M h.$$

Combining this with (S138) proves (S131).

We now convert from the normalized branch to the raw branch. The degree expansion and smooth composition give

$$m_\varepsilon^{-1/2} = 1 - \frac{\varepsilon}{2}\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon). \tag{S139}$$

Since $g_\varepsilon = m_\varepsilon^{-1/2} u_\varepsilon$, the exact identity

$$\frac{g_\varepsilon - h}{\varepsilon} = m_\varepsilon^{-1/2}\frac{u_\varepsilon - h}{\varepsilon} + \frac{m_\varepsilon^{-1/2} - 1}{\varepsilon}\, h \tag{S140}$$

holds. By (S131),

$$m_\varepsilon^{-1/2}\frac{u_\varepsilon - h}{\varepsilon} \longrightarrow -\frac{1}{8}\Delta_M h,$$

while (S139) gives

$$\frac{m_\varepsilon^{-1/2} - 1}{\varepsilon}\, h \longrightarrow -\frac{1}{2}\omega_M h.$$

Substitution into (S140) proves (S132).

We next pass to the logarithmic variable. Since M is compact and $h > 0$, h is bounded away from zero. By Lemma S14, $h^{-1} \in C^s_{\mathrm{ub}}(M)$. Set

$$w_\varepsilon := \frac{u_\varepsilon - h}{h}.$$

Then (S131) implies

$$\frac{w_\varepsilon}{\varepsilon} \longrightarrow -\frac{1}{8}\frac{\Delta_M h}{h}, \qquad \|w_\varepsilon\|_s = O(\varepsilon). \tag{S141}$$

Moreover,

$$\frac{u_\varepsilon}{h} = 1 + w_\varepsilon, \qquad \phi_\varepsilon = \log(1 + w_\varepsilon).$$

For sufficiently small ε, $1 + w_\varepsilon$ stays in a fixed compact subinterval of $(0, \infty)$. Taylor's formula in the Banach algebra therefore gives

$$\|\log(1 + w_\varepsilon) - w_\varepsilon\|_s \le C\|w_\varepsilon\|_s^2 = O(\varepsilon^2).$$

Dividing by ε and using (S141) proves (S134).

Finally, consider the entropic self-transport potential. For $\varepsilon > 0$, Lemma S25 identifies $\pi_{\varepsilon,g_\varepsilon}$ as the unique entropic self-transport plan and gives

$$f_\varepsilon = \varepsilon \log\left(\frac{c_\varepsilon^{1/2} g_\varepsilon}{h^2}\right).$$

Using

$$g_\varepsilon = m_\varepsilon^{-1/2} u_\varepsilon = m_\varepsilon^{-1/2} h e^{\phi_\varepsilon},$$

we obtain the exact decomposition

$$f_\varepsilon = -\varepsilon \log(c_\varepsilon^{-1/2} h) + \varepsilon\phi_\varepsilon - \frac{\varepsilon}{2}\log m_\varepsilon. \tag{S142}$$

Since $c_\varepsilon = (\pi\varepsilon)^{-d/2}$,

$$-\varepsilon \log(c_\varepsilon^{-1/2} h) = -\frac{\varepsilon d}{4}\log(\pi\varepsilon) - \varepsilon \log h.$$

Equation (S134) gives

$$\varepsilon\phi_\varepsilon = -\frac{\varepsilon^2}{8}\frac{\Delta_M h}{h} + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon^2).$$

Also, from

$$m_\varepsilon = 1 + \varepsilon\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon)$$

and smooth composition of the logarithm near 1,

$$\log m_\varepsilon = \varepsilon\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon).$$

Hence

$$-\frac{\varepsilon}{2}\log m_\varepsilon = -\frac{\varepsilon^2}{2}\omega_M + o_{C^s_{\mathrm{ub}}(M)}(\varepsilon^2).$$

Substituting these three expansions into (S142) proves (S135). □

Corollary S2 (First-order equivalence with the intrinsic heat problem)**.** *Let $\widetilde{u}_\varepsilon$ denote the local positive branch solving*

$$\widetilde{u}_\varepsilon H^M_\varepsilon \widetilde{u}_\varepsilon = h^2, \qquad \widetilde{u}_0 = h.$$

Then

$$\|u_\varepsilon - \widetilde{u}_\varepsilon\|_s = o(\varepsilon). \tag{S143}$$

Proof. Set

$$\widetilde{\mathcal{G}}_\varepsilon(u) := uH^M_\varepsilon u - h^2.$$

The proofs of Lemmas S26– S29 use only uniform boundedness and continuity of the kernel family, uniform invertibility and the strong inverse limit for $I + K_\varepsilon$, and the fixed-input generator expansion. These properties hold with $K_\varepsilon = H^M_\varepsilon$ by Lemmas S16 and S17. Hence there is a local positive branch $\widetilde{u}_\varepsilon \to h$ satisfying

$$\frac{\widetilde{u}_\varepsilon - h}{\varepsilon} \longrightarrow -\frac{1}{8}\Delta_M h \qquad \text{in } C^s_{\mathrm{ub}}(M).$$

Theorem S1 gives the same limit for u_ε. Therefore

$$\frac{u_\varepsilon - \widetilde{u}_\varepsilon}{\varepsilon} \longrightarrow 0$$

in $C^s_{\mathrm{ub}}(M)$, which is equivalent to (S143). □

BIBLIOGRAPHY

Aamari, E. and C. Levrard (2018). *Non-Asymptotic Rates for Manifold, Tangent Space, and Curvature Estimation*. arXiv: 1705.00989 [math.ST]. URL: https://arxiv.org/abs/1705.00989.

Abbe, E., E. Boix-Adsera, and T. Misiakiewicz (2024). *The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks*. arXiv: 2202.08658 [cs.LG]. URL: https://arxiv.org/abs/2202.08658.

Agarwal, M., Z. Harchaoui, G. Mulcahy, and S. Pal (2025). "Langevin Diffusion Approximation to Same Marginal Schrödinger Bridge". In: *arXiv preprint arXiv:2505.07647*.

Altschuler, J., F. Bach, A. Rudi, and J. Niles-Weed (2019). "Massively scalable Sinkhorn distances via the Nyström method". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/f55cadb97eaff2ba1980e001b0bd9842-Paper.pdf.

Altschuler, J., J. Weed, and P. Rigollet (2018). *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*. arXiv: 1705.09634 [cs.DS]. URL: https://arxiv.org/abs/1705.09634.

Ambrosio, L., N. Gigli, and G. Savare (2006). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel. ISBN: 9783764373092. URL: https://books.google.com/books?id=Hk_wNp0sc4gC.

Arnaboldi, L., Y. Dandi, F. Krzakala, L. Pesce, and L. Stephan (2024). "Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions". In: *arXiv preprint arXiv:2405.15459*.

Aswani, A., P. Bickel, and C. Tomlin (Feb. 2011). "Regression on manifolds: Estimation of the exterior derivative". In: *The Annals of Statistics* 39.1, pp. 48–81. DOI: 10.1214/10-AOS834.

Bakry, D., I. Gentil, and M. Ledoux (2013). *Analysis and geometry of Markov diffusion operators*. Vol. 348. Springer Science & Business Media.

Barla, A., F. Odone, and A. Verri (2002). "Hausdorff kernel for 3D object acquisition and detection". In: *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*. Springer, pp. 20–33.

Barrio, E. del, A. Gonzalez-Sanz, J.-M. Loubes, and J. Niles-Weed (2022). *An improved central limit theorem and fast convergence rates for entropic transportation costs*. arXiv: 2204.09105 [math.ST]. URL: https://arxiv.org/abs/2204.09105.

Bay, H., T. Tuytelaars, and L. Van Gool (2006). "Surf: Speeded up robust features". In: *European conference on computer vision*. Springer, pp. 404–417.

Beaglehole, D., D. Holzmüller, A. Radhakrishnan, and M. Belkin (2025). "xRFM: Accurate, scalable, and interpretable feature learning models for tabular data". In: *arXiv preprint arXiv:2508.10053*.

Beaglehole, D., A. Radhakrishnan, P. Pandit, and M. Belkin (2023). "Mechanism of feature learning in convolutional neural networks". In: *arXiv preprint arXiv:2309.00570*.

Belkin, M., S. Ma, and S. Mandal (2018). "To understand deep learning we need to understand kernel learning". In: *International Conference on Machine Learning*. PMLR, pp. 541–549.

Bengio, Y., A. Courville, and P. Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Berlinet, A. and C. Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bietti, A., J. Bruna, C. Sanford, and M. J. Song (2022). "Learning single-index models with shallow neural networks". In: *Advances in neural information processing systems* 35, pp. 9768–9783.

Bigoni, D., Y. Marzouk, C. Prieur, and O. Zahm (2022). "Nonlinear dimension reduction for surrogate modeling using gradient information". In: *Information and Inference: A Journal of the IMA* 11.4, pp. 1597–1639.

Bigot, J., E. Cazelles, and N. Papadakis (2019). *Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration*. arXiv: 1804.08962 [stat.ME].

Boix-Adsera, E., E. Littwin, E. Abbe, S. Bengio, and J. Susskind (2023). "Transformers learn through gradual rank increase". In: *Advances in Neural Information Processing Systems* 36, pp. 24519–24551.

Breiman, L. and W. S. Meisel (1976). "General estimates of the intrinsic variability of data in nonlinear regression models". In: *Journal of the American Statistical Association* 71.354, pp. 301–307.

Bunne, C., Y.-P. Hsieh, M. Cuturi, and A. Krause (2023). "The schrödinger bridge between gaussian measures has a closed form". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 5802–5833.

Bush, J. (2021). *C4L Image Dataset*. DOI: 10.21227/bc9m-f507. URL: https://dx.doi.org/10.21227/bc9m-f507.

Canas, G. and L. Rosasco (2012). "Learning probability measures with respect to optimal transport metrics". In: *Advances in Neural Information Processing Systems* 25.

Carrell, A. M., A. Gong, A. Shetty, R. Dwivedi, and L. Mackey (2025). *Low-Rank Thinning*. arXiv: 2502.12063 [stat.ML]. URL: https://arxiv.org/abs/2502.12063.

Chadebec, C. and S. Allassonnière (2022). *A Geometric Perspective on Variational Autoencoders*. arXiv: 2209.07370 [stat.ML]. URL: https://arxiv.org/abs/2209.07370.

Chapelle, O., P. Haffner, and V. N. Vapnik (1999). "Support vector machines for histogram-based image classification". In: *IEEE transactions on Neural Networks* 10.5, pp. 1055–1064.

Chen, Y., E. N. Epperly, J. A. Tropp, and R. J. Webber (2025). "Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations". In: *Communications on Pure and Applied Mathematics* 78.5, pp. 995–1041.

Chen, Y., Y. Wang, J. Su, et al. (n.d.). "Unified Framework for Coreset Selection and Dataset Distillation by Distribution Matching". In: ().

Chen, Y., M. Welling, and A. Smola (2012). "Super-samples from kernel herding". In: *arXiv preprint arXiv:1203.3472*.

Cheng, X. and B. Landa (2024). "Bi-stochastically normalized graph Laplacian: convergence to manifold Laplacian and robustness to outlier noise". In: *Information and Inference: A Journal of the IMA* 13.4, iaae026.

Chmiela, S., A. Tkatchenko, H. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller (Mar. 2017). "Machine Learning of Accurate Energy-Conserving Molecular Force Fields". In: *Science Advances*.

Claici, S., A. Genevay, and J. Solomon (2020). *Wasserstein Measure Coresets*. arXiv: 1805.07412 [stat.ML]. URL: https://arxiv.org/abs/1805.07412.

Coifman, R. R. and S. Lafon (2006a). "Diffusion maps". In: *Applied and computational harmonic analysis* 21.1, pp. 5–30.

— (2006b). "Diffusion maps". In: *Applied and Computational Harmonic Analysis* 21.1. Special Issue: Diffusion Maps and Wavelets, pp. 5–30. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2006.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1063520306000546.

Comaniciu, D. and P. Meer (2002). "Mean shift: a robust approach toward feature space analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5, pp. 603–619. DOI: 10.1109/34.1000236.

Conforti, G. and L. Tamanini (2021). "A formula for the time derivative of the entropic cost and applications". In: *Journal of Functional Analysis* 280.11, p. 108964.

Csiszár, I. (1975). "I-divergence geometry of probability distributions and minimization problems". In: *The annals of probability*, pp. 146–158.

Cui, T., X. Tong, and O. Zahm (2024). "Optimal Riemannian metric for Poincar {\'e} inequalities and how to ideally precondition Langevin dymanics". In: *arXiv preprint arXiv:2404.02554*.

Cuturi, M. (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26.

Dalal, N. and B. Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee, pp. 886–893.

Damian, A., E. Nichani, R. Ge, and J. D. Lee (2023). "Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models". In: *Advances in Neural Information Processing Systems* 36, pp. 752–784.

Das, P., M. Moll, H. Stamati, L. Kavraki, and C. Clementi (2006). "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction". In: *Proceedings of the National Academy of Sciences* 103.26, pp. 9885–9890.

Das, P., M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi (2006). "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction". In: *Proceedings of the National Academy of Sciences* 103.26, pp. 9885–9890.

Davis, R. A., K.-S. Lii, and D. N. Politis (2011). "Remarks on some nonparametric estimates of a density function". In: *Selected Works of Murray Rosenblatt*. Springer, pp. 95–100.

Dwivedi, R. and L. Mackey (2023). *Kernel Thinning*. arXiv: 2105.05842 [stat.ML].

— (2025). *Generalized Kernel Thinning*. arXiv: 2110.01593 [stat.ML]. URL: https://arxiv.org/abs/2110.01593.

El Karoui, N. (Feb. 2010). "The spectrum of kernel random matrices". In: *The Annals of Statistics* 38.1, pp. 1–50. DOI: 10.1214/09-AOS720.

Engel, K.-J. and R. Nagel (2000). *One-parameter semigroups for linear evolution equations*. Springer.

Federer, H. (1959). "Curvature measures". In: *Transactions of the American Mathematical Society* 93.3, pp. 418–491.

Fernholz, L. T. (1983). "Hadamard Differentiation". In: *von Mises Calculus For Statistical Functionals*. New York, NY: Springer New York, pp. 16–24. ISBN: 978-1-4612-5604-5. DOI: 10.1007/978-1-4612-5604-5_3. URL: https://doi.org/10.1007/978-1-4612-5604-5_3.

Feydy, J., T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré (2019a). "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690.

— (16–18 Apr 2019b). "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings

of Machine Learning Research. PMLR, pp. 2681–2690. URL: https://proceedings.mlr.press/v89/feydy19a.html.

Friedman, J. H. (1979). "A tree-structured approach to nonparametric multiple regression". In: *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop held in Heidelberg, April 2–4, 1979*. Springer, pp. 5–22.

Fukunaga, K. and L. Hostetler (1975). "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Transactions on information theory* 21.1, pp. 32–40.

Garreau, D., W. Jitkrittum, and M. Kanagawa (2018). *Large sample analysis of the median heuristic*. arXiv: 1707.07269 [math.ST]. URL: https://arxiv.org/abs/1707.07269.

Genevay, A., G. Peyre, and M. Cuturi (Sept. 2018). "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, pp. 1608–1617. URL: https://proceedings.mlr.press/v84/genevay18a.html.

Genovese, C. R., M. Perone-Pacifico, I. Verdinelli, and L. A. Wasserman (2012). "Minimax Manifold Estimation". In: *Journal of Machine Learning Research* 13, pp. 1263–1291. URL: http://dl.acm.org/citation.cfm?id=2343687.

Genton, M. G. (2001). "Classes of kernels for machine learning: a statistics perspective". In: *Journal of machine learning research* 2.Dec, pp. 299–312.

Goldfeld, Z., K. Kato, G. Rioux, and R. Sadhu (2022). "Limit theorems for entropic optimal transport maps and the Sinkhorn divergence". In: *arXiv preprint arXiv:2207.08683*.

Gong, A., K. Choi, and R. Dwivedi (2024). "Supervised Kernel Thinning". In: *arXiv preprint arXiv:2410.13749*.

Gonzalez-Sanz, A., J.-M. Loubes, and J. Niles-Weed (2022). "Weak limits of entropy regularized Optimal Transport; potentials, plans and divergences". In: arXiv: 2207.07427 [math.PR].

Gorishniy, Y., I. Rubachev, and A. Babenko (2023). *On Embeddings for Numerical Features in Tabular Deep Learning*. arXiv: 2203.05556 [cs.LG]. URL: https://arxiv.org/abs/2203.05556.

Gorishniy, Y., I. Rubachev, V. Khrulkov, and A. Babenko (2021). "Revisiting deep learning models for tabular data". In: *Advances in neural information processing systems* 34, pp. 18932–18943.

Gretton, A., K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola (2008). *A Kernel Method for the Two-Sample Problem*. arXiv: 0805.2368 [cs.LG].

Grieser, D. (2004). "Notes on heat kernel asymptotics". In: *Available on his website: http://www. staff. unioldenburg. de/daniel. grieser/wwwlehre/Schriebe/heat. pdf*.

Hayakawa, S., H. Oberhauser, and T. Lyons (2022). "Positively Weighted Kernel Quadrature via Subsampling". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 6886–6900. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/2dae7d1ccf1edf76f8ce7c282bdf4730-Paper-Conference.pdf.

— (2023). *Sampling-based Nyström Approximation and Kernel Quadrature*. arXiv: 2301.09517 [math.NA].

Heise, D. R. (1971). *Multivariate Model Building: The Validation of a Search Strategy*.

Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny (2001). "Structure adaptive approach for dimension reduction". In: *Annals of Statistics*, pp. 1537–1566.

Isserlis, L. (1918). "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables". In: *Biometrika* 12.1/2, pp. 134–139.

Jacot, A., F. Gabriel, and C. Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31.

Jank, W. (2005). "Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM". In: *Computational statistics & data analysis* 48.4, pp. 685–701.

Joachims, T. (1998). "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer, pp. 137–142.

Juditsky, A. B., O. V. Lepski, and A. B. Tsybakov (2009). "Nonparametric estimation of composite functions". In: *The Annals of Statistics* 37.3, pp. 1360–1404. DOI: 10.1214/08-AOS611.

Kiani, B., J. Wang, and M. Weber (2024). "Hardness of Learning Neural Networks Under the Manifold Hypothesis". In: *Advances in Neural Information Processing Systems* 37, pp. 5661–5696.

Koelle, S., H. Zhang, M. Meilă, and Y.-C. Chen (2022). "Manifold coordinates with physical meaning". In: *Journal of Machine Learning Research* 23.

Kokot, A. and A. Luedtke (2025). "Coreset selection for the Sinkhorn divergence and generic smooth divergences". In: *arXiv preprint arXiv:2504.20194*.

Kokot, A., O.-V. Murad, and M. Meila (2025). "The Noisy Laplacian: a Threshold Phenomenon for Non-Linear Dimension Reduction". In: *Forty-second International Conference on Machine Learning*.

Kondor, R. and T. Jebara (2003). "A kernel between sets of vectors". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 361–368.

Küster, B. (2017). "On the Semiclassical Functional Calculus for h-Dependent Functions". In: *Annals of Global Analysis and Geometry* 52.1, pp. 57–97. DOI: 10.1007/s10455-017-9549-1.

Landa, B. and X. Cheng (2023a). "Robust Inference of Manifold Density and Geometry by Doubly Stochastic Scaling". In: *SIAM Journal on Mathematics of Data Science* 5.3, pp. 589–614. DOI: 10.1137/22M1516968. eprint: https://doi.org/10.1137/22M1516968. URL: https://doi.org/10.1137/22M1516968.

— (2023b). "Robust inference of manifold density and geometry by doubly stochastic scaling". In: *SIAM Journal on Mathematics of Data Science* 5.3, pp. 589–614.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Lee, J. D., K. Oko, T. Suzuki, and D. Wu (2024). "Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit". In: *Advances in Neural Information Processing Systems* 37, pp. 58716–58756.

Lee, K.-Y., B. Li, and F. Chiaromonte (2013). "A general theory for nonlinear sufficient dimension reduction: Formulation and estimation". In: *The Annals of Statistics*, pp. 221–249.

Li, L., R. Dwivedi, and L. Mackey (2024a). *Debiased Distribution Compression*. arXiv: 2404.12290 [stat.ML]. URL: https://arxiv.org/abs/2404.12290.

— (2024b). "Debiased distribution compression". In: *arXiv preprint arXiv:2404.12290*.

Litterer, C. and T. Lyons (2012a). "HIGH ORDER RECOMBINATION AND AN APPLICATION TO CUBATURE ON WIENER SPACE". In: *The Annals of Applied Probability* 22.4, pp. 1301–1327. ISSN: 10505164. URL: http://www.jstor.org/stable/41713361 (visited on 07/16/2024).

Litterer, C. and T. Lyons (2012b). "High order recombination and an application to cubature on Wiener space". In.

Liu, H., M. Chen, T. Zhao, and W. Liao (2021). "Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks". In: *International Conference on Machine Learning*. PMLR, pp. 6770–6780.

Lowe, D. G. (1999). "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee, pp. 1150–1157.

Luedtke, A. (2024). *Simplifying debiased inference via automatic differentiation and probabilistic programming*. arXiv: 2405.08675 [stat.ME]. URL: https://arxiv.org/abs/2405.08675.

Marshall, N. F. and R. R. Coifman (2019). "Manifold learning with bi-stochastic kernels". In: *IMA Journal of Applied Mathematics* 84.3, pp. 455–482.

Meilă, M. and H. Zhang (2024). "Manifold learning: What, how, and why". In: *Annual Review of Statistics and Its Application* 11.1, pp. 393–417.

Mena, G. and J. Weed (2019). *Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem*. arXiv: 1905.11882 [math.ST]. URL: https://arxiv.org/abs/1905.11882.

Mordant, G. (2024). "The entropic optimal (self-) transport problem: Limit distributions for decreasing regularization with application to score function estimation". In: *arXiv preprint arXiv:2412.12007*.

Mousavi-Hosseini, A., S. Park, M. Girotti, I. Mitliagkas, and M. A. Erdogdu (2022). "Neural networks efficiently learn low-dimensional representations with sgd". In: *arXiv preprint arXiv:2209.14863*.

Mukherjee, S., Q. Wu, and D.-X. Zhou (2010). “Learning gradients on manifolds”. In.

Mulcahy, G. and S. Pal (2025). *Diffusion Approximations to Schrödinger Bridges on Manifolds*. arXiv: 2512.18867 [math.PR]. URL: https://arxiv.org/abs/2512.18867.

Nichani, E., A. Damian, and J. D. Lee (2023). “Provable guarantees for nonlinear feature learning in three-layer neural networks”. In: *Advances in Neural Information Processing Systems* 36, pp. 10828–10875.

Nouy, A. and A. Pasco (2025). “Surrogate to Poincar\’e inequalities on manifolds for dimension reduction in nonlinear feature spaces”. In: *arXiv preprint arXiv:2505.01807*.

Odone, F., A. Barla, and A. Verri (2005). “Building kernels from binary strings for image matching”. In: *IEEE Transactions on Image Processing* 14.2, pp. 169–180.

Osborne, M. (2010). “Bayesian Gaussian processes for sequential prediction, optimisation and quadrature”. PhD thesis. Oxford University, UK.

Parzen, E. (1962). “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3, pp. 1065–1076.

Pollard, D. (1982). “Quantization and the method of k-means”. In: *IEEE Transactions on Information theory* 28.2, pp. 199–205.

Pope, P., C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein (2021). “The intrinsic dimension of images and its impact on learning”. In: *arXiv preprint arXiv:2104.08894*.

Radhakrishnan, A., D. Beaglehole, P. Pandit, and M. Belkin (2022). “Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features”. In: *arXiv preprint arXiv:2212.13881*.

Radhakrishnan, A., M. Belkin, and D. Drusvyatskiy (2025). “Linear recursive feature machines provably recover low-rank matrices”. In: *Proceedings of the National Academy of Sciences* 122.13, e2411325122.

Ramdas, A., N. Garcia, and M. Cuturi (2015). *On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests*. arXiv: 1509.02237 [math.ST].

Rigollet, P. and J. Weed (2018). “Entropic optimal transport is maximum-likelihood deconvolution”. In: *Comptes Rendus. Mathématique* 356.11-12, pp. 1228–1235.

Römisch, W. (2004). “Delta method, infinite dimensional”. In.

Rüschendorf, L. and W. Thomsen (1993). "Note on the Schrödinger equation and I-projections". In: *Statistics & probability letters* 17.5, pp. 369–375.

Sachdeva, N. and J. McAuley (2023). *Data Distillation: A Survey*. arXiv: 2301.04272 [cs.LG]. URL: https://arxiv.org/abs/2301.04272.

Samarov, A. M. (1993). "Exploring regression structure using nonparametric functional estimation". In: *Journal of the American Statistical Association* 88.423, pp. 836–847.

Schmid, C. and R. Mohr (1997). "Local grayvalue invariants for image retrieval". In: *IEEE transactions on pattern analysis and machine intelligence* 19.5, pp. 530–535.

Schölkopf, B., P. Simard, A. Smola, and V. Vapnik (1997). "Prior knowledge in support vector kernels". In: *Advances in neural information processing systems* 10.

Schölkopf, B. and A. J. Smola (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Scornet, E. (2016). "Random forests and kernel methods". In: *IEEE Transactions on Information Theory* 62.3, pp. 1485–1500.

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. [Nachdr.] Wiley series in probability and mathematical statistics : Probability and mathematical statistics. New York, NY [u.a.]: Wiley. XIV, 371. ISBN: 0471024031. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy.

Shen, G., Y. Jiao, Y. Lin, J. L. Horowitz, and J. Huang (2021). "Deep quantile regression: Mitigating the curse of dimensionality through composition". In: *arXiv preprint arXiv:2107.04907*.

Sriperumbudur, B. K., K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet (2012). "On the empirical estimation of integral probability metrics". In: *Electronic Journal of Statistics* 6.none, pp. 1550–1599. DOI: 10.1214/12-EJS722. URL: https://doi.org/10.1214/12-EJS722.

Stein, C. M. (1981). "Estimation of the mean of a multivariate normal distribution". In: *The annals of Statistics*, pp. 1135–1151.

Steinley, D. (2006). "K-means clustering: a half-century synthesis". In: *British Journal of Mathematical and Statistical Psychology* 59.1, pp. 1–34.

Strocchi, M., M. A. Gsell, C. M. Augustin, O. Razeghi, C. H. Roney, A. J. Prassl, E. J. Vigmond, J. M. Behar, J. S. Gould, C. A. Rinaldi, et al. (2020). "Simulating ventricular systolic motion in a four-chamber heart model with spatially varying robin boundary conditions to model the effect of the pericardium". In: *Journal of biomechanics* 101, p. 109645.

Takeda, H., S. Farsiu, and P. Milanfar (2007). "Kernel regression for image processing and reconstruction". In: *IEEE Transactions on image processing* 16.2, pp. 349–366.

— (2008). "Deblurring using regularized locally adaptive kernel regression". In: *IEEE transactions on image processing* 17.4, pp. 550–563.

Tenenbaum, J. B., V. d. Silva, and J. C. Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500, pp. 2319–2323.

Trivedi, S., J. Wang, S. Kpotufe, and G. Shakhnarovich (2014). "A consistent estimator of the expected gradient outerproduct." In: *UAI*, pp. 819–828.

Tropp, J. A., A. Yurtsever, M. Udell, and V. Cevher (2017). *Fixed-Rank Approximation of a Positive-Semidefinite Matrix from Streaming Data*. arXiv: 1706.05736 [cs.NA].

Tsybakov, A. B. (2009). "Nonparametric estimators". In: *Introduction to Nonparametric Estimation*. New York, NY: Springer New York, pp. 1–76. ISBN: 978-0-387-79052-7. DOI: 10.1007/978-0-387-79052-7_1. URL: https://doi.org/10.1007/978-0-387-79052-7_1.

Vaart, A. W. van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Vaart, A. v. d. and J. A. Wellner (2023). In: *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Vacher, A., B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard (2021). "A dimension-free computational upper-bound for smooth optimal transport estimation". In: *Conference on Learning Theory*. PMLR, pp. 4143–4173.

Verdière, R., C. Prieur, and O. Zahm (2025). "Diffeomorphism-based feature learning using Poincaré inequalities on augmented input space". In: *Journal of Machine Learning Research* 26.139, pp. 1–31.

Vincent, P. (2011). "A Connection Between Score Matching and Denoising Autoencoders". In: *Neural Computation* 23.7, pp. 1661–1674. DOI: 10.1162/NECO_a_00142.

Vishwanathan, S. V. N., N. N. Schraudolph, R. Kondor, and K. M. Borgwardt (2010). "Graph kernels". In: *The Journal of Machine Learning Research* 11, pp. 1201–1242.

Wallraven, Caputo, and Graf (2003). "Recognition with local features: the kernel recipe". In: *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, pp. 257–264.

Weed, J. and F. Bach (2019). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In.

Weickert, J. (1999). "Coherence-enhancing diffusion filtering". In: *International journal of computer vision* 31.2, pp. 111–127.

Wormell, C. L. and S. Reich (2021). "Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization". In: *SIAM Journal on Numerical Analysis* 59.3, pp. 1687–1734.

Wu, Q., J. Guinney, M. Maggioni, and S. Mukherjee (2010). "Learning gradients: predictive models that infer geometry and statistical dependence". In: *The Journal of Machine Learning Research* 11, pp. 2175–2198.

Wu, Y. and M. Maggioni (2024). "Conditional regression for the Nonlinear Single-Variable Model". In: *arXiv preprint arXiv:2411.09686*.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). "An adaptive estimation of dimension reduction space". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 363–410.

Xiong, Z., N. Dalmasso, S. Sharma, F. Lecue, D. Magazzeni, V. K. Potluru, T. Balch, and M. Veloso (2024). *Fair Wasserstein Coresets*. arXiv: `2311.05436 [stat.ML]`. URL: `https://arxiv.org/abs/2311.05436`.

Yeh, Y.-R., S.-Y. Huang, and Y.-J. Lee (2008). "Nonlinear dimension reduction with kernel sliced inverse regression". In: *IEEE transactions on Knowledge and Data Engineering* 21.11, pp. 1590–1603.

Yin, H., Y. Qiu, and X. Wang (2025). "Wasserstein Coreset via Sinkhorn Loss". In: *Transactions on Machine Learning Research*.

Yuan, G., M. Xu, S. Kpotufe, and D. Hsu (2023). "Efficient estimation of the central mean subspace via smoothed gradient outer products". In: *arXiv preprint arXiv:2312.15469*.

Zhu, L., D. Davis, D. Drusvyatskiy, and M. Fazel (2025). "Iteratively reweighted kernel machines efficiently learn sparse functions". In: *arXiv preprint arXiv:2505.08277*.