Statistical and algorithmic perspectives on estimating optimal transport maps

by

Aram-Alexandre Pooladian

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Data Science

New York University

August, 2025

Professor Jonathan Niles-Weed

"Tell me one good thing about those people Eliot helps."

"I can't."

"I thought not."

"It's a secret thing," she said, forced to argue, pleading for the argument to stop right there.

Without any notion of how merciless he was being, the Senator pressed on. "You're among friends now—suppose you tell us what this great secret is."

"The secret is that they're human," said Sylvia.

—Kurt Vonnegut, *God Bless You, Mr. Rosewater (or, Pearls Before Swine)*

Dedication

To my grandparents.

Acknowledgments

I want to begin by expressing my deepest thanks to my advisor, Jonathan Niles-Weed. I am so incredibly grateful that you took a chance on me—a kid who just wanted to live in New York City because he watched too much *Seinfeld* growing up. I will miss the enthusiasm you brought to every meeting, the gabbing about Broadway shows and books, and all the (mathematical) wisdom you often shared with me. Thank you, Jon, for giving me the appropriate freedoms, guidance, patience, kindness, and—most importantly—honesty throughout this long journey.

I would like to extend my gratitude to my committee members: Joan Bruna, Marco Cuturi, Yanjun Han, and Daniel Lacker. They have individually impacted me through their generosity, shared research interests, and brilliance. Working with Marco was one of the highlights early on in my PhD (which later turned into many arguments about notation conventions) and I always look forward to catching up at conferences together.

Aleks Donev had a profound impact on the way I present my work. I will always continue to think of him whenever I'm preparing slides or writing a paper. I'm sure he's got something snarky, yet constructive, to say from the world beyond.

The last few years have been extraordinarily fun, and this is in large part due to my long list of collaborators, including Pierre Ablin, Brandon Amos, Ricardo Baptista, Heli Ben-Hamu, Michael Brennan, Sinho Chewi, Marco Cuturi, Vincent Divol, Carles Domingo-Enrich, Doron Haviv, Yiheng Jiang, Parnian Kassraie, Michal Klein, Yaron Lipman, Youssef Marzouk, Eugene Ndiaye, Dana Pe'er, Ricky Tian Qi Chen, and James Thornton. I'm grateful that I get to call some

of you my dear friends. From this long list, I want to specifically thank Vincent Divol and Sinho Chewi without whom this thesis would not have been possible. Both are exemplary, meticulous, brilliant academics who have greatly shaped the way I approach research problems, and I am deeply thankful for having received their kindness and support over the years. I also want to thank Yiheng Jiang, for making my first mentoring experience incredibly rewarding and reminding me what this is really all about.

I have made so many wonderful friends over the years, from Marianopolis, to McGill, and now from NYU and around the world. There's too many of you to name without possibly forgetting someone—this is a good problem to have. I thank you all for lending an ear whenever I needed it, for laughing with me, and for helping me grow.

Kaylee, thank you for showing me the beauty in all the places I never thought to look and for helping me be more silly. I love you.

While I blame them for making me incredibly neurotic, I am immensely thankful for my family. First, my parents, Maral and Shahen, and younger brother, Ari, continue to motivate and inspire me more than they realize; I will always look up to them. I was also raised by my aunt Lori and uncle Raffi, who continue to demonstrate what it means to be unconditionally loving and supportive. And finally, a heartfelt thank you to my grandmother Astghik for showing me what it means to be kind.

Living in New York City the past few years has been nothing short of life-changing. I would like to thank every artist, busker, comedian, dollar-slice, euphoric-feeling-when-the-subway-finally-arrived, freelance writer, game of sidewalk-chess, "Hey boss", improv class, jazz musician, killer view, line of cocaine (kidding!), MTA employee, NYT crossword, open-mic night, pigeon, queer-bar, rat-infested park, subway dancer, theatre show, unbearably loud intersection, vinyl record store, "Watch where ya goin'!", xenomorph art exhibit,[1] yuppie-shamer, and zookeeper that, collectively, make this city one of the most amazing places on Earth.

[1] This actually exists.

Abstract

Optimal transport maps, or Brenier maps, have become widely adopted in data-driven domains as they provide a canonical transformation between independent datasets. While many existing methods aim for optimal statistical performance, they often fall short in practical regimes of interest, such as when the data is high-dimensional or when the sample size is large. In this thesis, we analyze principled algorithms for estimating optimal transport maps in precisely these regimes. Our first contribution is the introduction of the entropic Brenier map, an estimator of the Brenier map based on *entropic* optimal transport, which harnesses the computational efficiency of Sinkhorn's matrix scaling algorithm (Sinkhorn, 1967). We prove the first finite-sample guarantees for estimating optimal transport maps using this estimator, demonstrate that it is minimax optimal in the semi-discrete setting, and make further connections to the statistical estimation of Schrödinger bridge between two distributions. Next, we further derive new theoretical properties of the entropic Brenier map, such as bounds on the Lipschitz constant of the map as well as its stability with respect to the target measures; these results also yield new insights for the unregularized optimal transport map. For our final contribution, we propose a new optimization framework for functionals defined over a suitable family of optimal transport maps. As an application, we develop the first gradient-based algorithm for mean-field variational inference that comes with end-to-end convergence guarantees.

Contents

List of Figures

List of Appendices

1 Introduction

Optimal transport theory, first conceptualized by Gaspard Monge (Monge, 1781) and later formalized by Leonid Kantorovich (Kantorovitch, 1942), has emerged as a powerful tool to address mathematical, statistical, and computational questions that arise over the space of probability distributions.

In its basic form, the optimal transport problem is defined as follows: For two probability measures P, Q over $\mathbb{R}^d$, let $\Pi(P, Q)$ be the set of joint measures with left-marginal P and right-marginal Q, called the set of couplings. For a given cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, Kantorovitch (1942) proposed to minimize the average cost of displacement between the two marginals, resulting in the following optimization problem:

$$W_c(P, Q) \coloneqq \inf_{\pi \in \Pi(P,Q)} \iint c(x, y) \, \mathrm{d}\pi(x, y) \, . \tag{1.1}$$

As an application, Kantorovich considers the case where P and Q are discrete probability measures, resulting in a standard resource allocation problem in which W_c represents the total "work" in displacing, say, goods from factories to stores. Perhaps most striking is that (1.1) is the first instance of a *linear program*. For his contributions, Kantorovich emerged as a pioneer of modern mathematical programming and was awarded the Nobel Memorial Prize in Economics in 1975.

The impact of Kantorovich's work extends far beyond problems in resource allocation. For instance, taking $c(x, y) = \|x - y\|^p$ (p-powers of the Euclidean norm for $p \geq 1$), (1.1) becomes the

p-Wasserstein distance

$$W_p(P, Q) := \left(\inf_{\pi \in \Pi(P,Q)} \iint \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p}, \tag{1.2}$$

which metrizes weak convergence over the space of probability measures (with finite p-moments). This observation has allowed the Wasserstein distance to not only emerge as a powerful theoretical tool in mathematics and statistics, but due to recent computational advances, a methodological one.

The following scenario is one application of the Wasserstein distance for statistical purposes: A practitioner has two independent sets of data which they model as coming from two distributions, i.e., they have independent and identically distributed samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_n \sim Q$. In order to approximate the 2-Wasserstein distance between P and Q from samples, a naive estimator consists in plugging the empirical measures P_n and Q_n (with $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and $Q_n = n^{-1} \sum_{j=1}^n \delta_{Y_j}$) into (1.1). This results in the following linear program over the set of doubly stochastic matrices, where the rows and columns sum to $1/n$:

$$W_2^2(P_n, Q_n) = \min_{\Pi \in \mathbb{R}_+^{n \times n}} \langle C, \Pi \rangle, \quad \text{s.t.} \quad \Pi \mathbf{1}_n = \tfrac{1}{n} \mathbf{1}_n, \quad \Pi^\top \mathbf{1}_n = \tfrac{1}{n} \mathbf{1}_n, \tag{1.3}$$

where $C_{ij} = \|X_i - Y_j\|^2$ is an $n \times n$ cost matrix. To actually compute the estimator—an aspect of the problem which was not covered in Kantorovich's original work—the modern statistician has numerous algorithms at their disposal.[1] For example, the Hungarian algorithm can optimize (1.3) with a runtime complexity of roughly $O(n^3)$ and a space complexity of $O(n^2)$ to store the cost matrix into memory (Peyré and Cuturi, 2019). So, the practitioner is able to compute their estimator—but how accurate is it? Under mild assumptions, a line of work demonstrates that estimating p-Wasserstein distances with empirical measures suffers from what is called the *curse of dimensionality* (Chizat et al., 2020; Dudley, 1969; Manole and Niles-Weed, 2024), implying the

[1] Indeed, it was not until later that solvers were developed to solve linear programs (e.g., the simplex method due to Dantzig (1951)).

following rate of estimation for $p = 2$ and $d \geq 5$:

$$|W_2^2(P, Q) - W_2^2(P_n, Q_n)| \lesssim n^{-2/d}\,. \tag{1.4}$$

If the right-hand side of (1.4) is to be as small as possible, the practitioner requires a procedure that leverages as many samples as possible. The Hungarian algorithm, with an $O(n^2)$ storage capacity (and relatively slow runtime), is not amenable to scaling the number of samples past $n \asymp 10^4$. Thus, it appears that the effectiveness of the practitioner's estimator is stifled due to computational burden.

To circumvent this issue, Marco Cuturi proposed to incorporate *entropic regularization* to the Wasserstein objective in order to solve an approximation to (1.3) in the large n regime, resulting in the following *strongly convex* optimization objective

$$\min_{\Pi \in \mathbb{R}_+^{n \times n}} \langle C, \Pi \rangle + \varepsilon \sum_{i,j=1}^{n} \Pi_{ij} \log(\Pi_{ij})\,, \quad \text{s.t.} \quad \Pi \mathbf{1}_n = \mathbf{1}_n \tfrac{1}{n}\,, \quad \Pi^\top \mathbf{1}_n = \mathbf{1}_n \tfrac{1}{n}\,, \tag{1.5}$$

where $\varepsilon > 0$ denotes the regularization strength (Cuturi, 2013). This formulation is commonly known as *entropic optimal transport*, due to the entropy penalization term. To solve (1.5), Cuturi used a matrix-scaling algorithm due to Richard Sinkhorn (Sinkhorn, 1967). Sinkhorn's algorithm, which consists of only a few lines of code, has several benefits: (1) it does not require storing the cost matrix, which eliminates the $O(n^2)$ storage complexity, (2) it can take advantage of GPU-computation, and (3) it has a runtime of $\tilde{O}(n^2/\varepsilon)$ (Altschuler et al., 2017), which is significantly faster than the Hungarian algorithm for n large. Due to the computational prowess of Sinkhorn's algorithm, the procedure in (1.5) has emerged as the *de facto* approach to estimating optimal transport costs, even with statistical guarantees comparable to (1.4) as long as $\varepsilon = \varepsilon(n)$ is chosen in a particular manner (where $\varepsilon(n) \searrow 0$ as $n \nearrow \infty$); see Chizat et al. (2020) for example. Thus, from both a computational and statistical perspective, the statistician appears satiated in their

task of estimating the optimal transport cost on the basis of samples.

1.1 Going beyond distances

Over the last few years, modern statistical learning problems have experienced a considerable shift. The statistician is no longer merely interested in estimating distances between (empirical) distributions, but also transformations between them. We say that $T : \mathbb{R}^d \to \mathbb{R}^d$ lies in the set of *transport maps* between P and Q, denoted $T \in \mathcal{T}(P, Q)$, if for $X \sim P$, $T(X) \sim Q$. Transport maps arise in wide array of problems, ranging from machine learning and generative modeling (Arjovsky et al., 2017; Finlay et al., 2020a; Genevay et al., 2018; Grathwohl et al., 2018; Huang et al., 2021a; Salimans et al., 2018), computer graphics (Feydy et al., 2017; Solomon et al., 2015; 2016), economics and statistics (Carlier et al., 2016; Chernozhukov et al., 2017; Gunsilius and Xu, 2021; Torous et al., 2024), to the applied sciences (Bunne et al., 2022; Moriel et al., 2021; Schiebinger et al., 2019; Yang et al., 2020). *A priori*, there are infinitely many possible transport maps between two distributions. Once more, optimal transport theory can help the practitioner.

Focusing on the squared Euclidean cost, if P and Q have finite second moment, the following infinite-dimensional but non-convex optimization problem defines an *optimal transport map* between the two marginals:

$$T_0 \coloneqq \operatorname*{argmin}_{T \in \mathcal{T}(P,Q)} \int \|x - T(x)\|^2 \, \mathrm{d}P(x) \,. \tag{1.6}$$

Unlike (1.1), this formulation may not always have a minimizer.[2] Brenier (1991) showed that if P has a density, then there always exists $T_0 = \nabla\varphi_0$, where φ_0 is a convex function called a Brenier potential; and so, (1.6) is the same as (1.2) for $p = 2$. We will interchangeably refer to T_0 as both the optimal transport map and Brenier map.

Thus, a widely studied analogue to the previous statistical problem is the following: Given

[2]Consider moving a discrete probability measure with one atom to another with two atoms—there is no map.

$X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$, how can we estimate the optimal transport map from P to Q on the basis of samples? Importantly, for these various inference problems, the practitioner wants to know how to transport an *out-of-sample*, or new data-point, from the source to the target. In other words, the statistician wants to compute an estimator $\hat{T}_n$ with good statistical properties under the following notion of risk:

$$\mathbb{E}\|\hat{T}_n - T_0\|^2_{L^2(P)}\,. \tag{1.7}$$

There are two (broad) families of estimators which have been studied thusfar. The first are plug-in estimators at the level of the Brenier potential φ_0 (Divol et al., 2022; Hütter and Rigollet, 2021; Vacher et al., 2024). These works make use of the fact that optimal transport maps are gradients of Brenier potentials, which are known to be minimizers of the following *semidual* functional:

$$\varphi_0 \in \operatorname*{argmin}_{\varphi \in L^1(P)} \int \varphi \,\mathrm{d}P + \int \varphi^* \,\mathrm{d}Q\,, \tag{1.8}$$

where $h \mapsto h^*$ is the convex conjugate operator. Of course, we cannot optimize over all $L^1(P)$ functions (nor all of P or Q), so one can resort to some approximating family wherein the optimization is tractable. Thus, the practitioner optimizes

$$\hat{\varphi}_\Theta \in \operatorname*{argmin}_{\varphi \in \mathcal{F}_\Theta} \int \varphi \,\mathrm{d}P_n + \int \varphi^* \,\mathrm{d}Q_n\,, \tag{1.9}$$

where $\mathcal{F}_\Theta$ can (in principle) be any class of smooth functions where (1.8) can be evaluated (and recall P_n and Q_n are the empirical measures). The final estimator for the optimal transport map is then $\nabla\hat{\varphi}_\Theta$.

The presence of the conjugate operator appearing in the objective function (1.9) makes this optimization problem quite difficult. For instance, Hütter and Rigollet (2021) consider a (large

parametric) family of convex functions with bounded wavelet expansion, i.e., $\mathcal{F}_\Theta = \mathcal{F}_W$. They prove that their estimator attains the following (near-)minimax risk

$$\mathbb{E}\|\nabla\hat{\varphi}_W - T_0\|^2_{L^2(P)} \lesssim_{\log(n)} n^{-\frac{2(s-1)}{2s+d-4}}.$$

While their result is elegant from a statistical perspective, their approach is realistically intractable for any applications where $d \geq 3$, as it requires a gridding scheme. What if $\mathcal{F}_\Theta$ is something else? Divol et al. (2022) expand upon their work by considering *general function spaces*, which can consist of (large) parametric families, quadratics, Barron spaces, Reproducing Kernel Hilbert Spaces, and more. While they too prove minimax optimal estimation rates (and recover the results of Hütter and Rigollet (2021) in the process), their results are underwhelming from a computational point of view. Indeed, polynomial-type conjugate oracles (required to optimize the semidual functional) are unlikely to exist for general function classes. Some families will allow an efficient gradient-descent-type scheme to compute $\hat{\varphi}_\Theta$; for instance, optimizing over $\lambda \mapsto \sum_j^J \lambda_j\varphi_j$ for a fixed, finite collection of smooth, strongly convex functions $\{\varphi_j\}_{j=1}^J$. These estimating classes are rather restrictive, and the resulting computational complexity is rather infeasible for modern tasks involving optimal transport.

A second approach to estimating optimal transport maps consists of plug-in estimators at the level of the densities P and Q, where we perform the estimation through a two-stage process (Deb et al., 2021; Manole et al., 2024a). First, we use the data (recall $X_1, \ldots, X_n \sim P$, and $Y_1, \ldots, Y_n \sim Q$) to construct estimators $\hat{P}_n$ and $\hat{Q}_n$ of the densities that more sophisticated than empirical measures, such as kernel density estimators (KDEs). Then, from the estimated densities, one could draw as many samples as desired (that is, fresh samples $X'_1, \ldots, X'_n \sim \hat{P}_n$ and also $Y'_1, \ldots, Y'_n \sim \hat{Q}_n$) and find the optimal matching in the sense of (1.3). The resulting map is the final estimator. This approach comes with several drawbacks, namely that sampling from high-dimensional KDEs is non-trivial, and that, in order to benefit from the smoothness properties of these estimators, an exorbitant

number of samples must be drawn which exceeds the (n^2) storage complexity. So, despite being a central object in many applications, most existing estimators of optimal transport maps are, while statistically optimal, nearly impossible to compute when $d \gg 1$ or when $n \gg 10^3$.

Collectively, these observations motivate the following question which drives our thesis:

> How do we develop procedures for estimating optimal transport maps which enjoy favorable computational and statistical guarantees?

1.2 Contributions of this thesis

We now provide a chapter-by-chapter summary of this thesis.

Statistical estimation of optimal transport maps and beyond

Chapter 2: Entropic estimation of optimal transport maps

In this first chapter, we develop and study a tractable, scalable non-parametric estimator of the optimal transport map based on the *entropic optimal transport* problem

$$\mathrm{OT}_\varepsilon(P, Q) := \min_{\pi \in \Pi(P,Q)} \iint \|x - y\|^2 \, \mathrm{d}\pi(x, y) + \varepsilon \, \mathrm{KL}(\pi \| P \otimes Q) \,, \tag{1.10}$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence, or relative entropy, between π and the product measure; the minimizer is called the optimal entropic coupling, denoted π_ε. (Note that (1.10) is the population analogue to (1.5).)

Our first contribution is an entropic analogue of Brenier's theorem. We show that our estimator—the *barycentric projection* of the optimal entropic coupling, denoted $T_\varepsilon(x) = \mathbb{E}_{\pi_\varepsilon}[Y|X = x]$—can be characterized as a gradient field of entropic potentials. Moreover, on the basis of samples, our estimator is easy to compute using Sinkhorn's algorithm, and extends to out-of-sample

points.

As a result, unlike current approaches for map estimation, which are slow to evaluate when the dimension or number of samples is large, our approach is parallelizable and extremely efficient even for massive data sets. Under smoothness assumptions on the optimal map, we show that our estimator enjoys comparable statistical performance to other estimators in the literature, but with much lower computational cost.

We showcase the efficacy of our proposed estimator through numerical examples, even ones not explicitly covered by our assumptions. By virtue of Lepski's method, we propose a modified version of our estimator that is adaptive to the smoothness of the underlying optimal transport map. Our proofs are based on a modified duality principle for entropic optimal transport and on a method for approximating optimal entropic plans due to Pal (2024).

It is worth mentioning that in the years since we developed this estimator, there has been a flurry of follow-up works and applications surrounding the entropic Brenier map in both statistical and theoretical circles. For instance, in the case of $\varepsilon > 0$, the community has established various central limit theorems (Goldfeld et al., 2024a;b; Sadhu et al., 2024; 2025) for the entropic Brenier map, estimators for different costs or applications (Baptista et al., 2024; Cuturi et al., 2022; Klein et al., 2024; Masud et al., 2023; Werenski et al., 2023), conceived more practical estimators (Kassraie et al., 2024), made use of them in generative modeling (Haviv et al., 2024), and more.

This chapter is based off

> "Entropic estimation of optimal transport maps", in submission (2021), with Jonathan Niles-Weed.

Chapter 3: Minimax estimation of discontinuous optimal transport maps: The semi-discrete case

The analysis of the previous chapter for estimating optimal transport maps from data (and essentially all prior work, such as Deb et al. (2021); Divol et al. (2022); Hütter and Rigollet (2021); Manole et al. (2024a); Vacher et al. (2024)) heavily relies on the assumption that the underlying optimal transport map is Lipschitz. In particular, this assumption excludes any examples where the Brenier map may be discontinuous (which is likely the case in real-world scenarios, in view of the so-called manifold hypothesis (Brown et al., 2022)). As a first step towards developing estimation procedures for discontinuous maps, we now consider the important special case where the data distribution Q is a discrete measure supported on a finite number of points in $\mathbb{R}^d$. We revisit the entropic Brenier map estimator from Chapter 2 and demonstrate that it converges at the *minimax-optimal* rate of $n^{-1/2}$ in the semidiscrete case, where the rate is independent of dimension. We stress that other standard map estimation techniques both lack finite-sample guarantees in this setting and provably *suffer* from the curse of dimensionality. We confirm these results in numerical experiments, and provide experiments for other settings, not covered by our theory, which indicate that the entropic estimator is a promising methodology in the general discontinuous setting.

The contents of this chapter follow from

> "Minimax estimation of discontinuous optimal transport maps: The semi-discrete case", in the *40th International Conference on Machine Learning* (ICML 2023), with Vincent Divol and Jonathan Niles-Weed

Chapter 4: Plug-in estimation of Schrödinger bridges

In this chapter, we consider another family of transport maps based on *dynamic transport*. These methods (such as flow matching (Albergo and Vanden-Eijnden, 2022; Lipman et al., 2022;

Liu et al., 2022b) and denoising diffusion probabilistic models (Song et al., 2020)) have quickly emerged as powerful approaches to perform generative modeling on complex, high-dimensional distributions. Among this class of algorithms is the *Schrödinger bridge* between two distributions (Léonard, 2014; Schrödinger, 1932), which is essentially the "optimal" diffusion path between two probability measures.

We propose and analyze an estimator for the Schrödinger bridge between two probability distributions. Unlike existing approaches (De Bortoli et al., 2021), our method does not require iteratively simulating forward and backward diffusions or training neural networks to fit unknown drifts. Instead, we show that the potentials obtained from solving the static entropic optimal transport problem between the source and target samples can be modified to yield a natural plug-in estimator of the time-dependent drift that defines the bridge between two measures. Under minimal assumptions, we show that our proposal, which we call the *Sinkhorn bridge*, provably estimates the Schrödinger bridge with a rate of convergence that depends on the *intrinsic dimensionality* of the target measure. Our approach combines results from the areas of sampling, and theoretical and statistical entropic optimal transport.

This chapter is based on the following article

> "Plug-in estimation of Schrödinger bridges", to appear in *SIAM Journal of Mathematics and Data Science* (2025), with Jonathan Niles-Weed.

Interlude: Theoretical properties of entropic Brenier maps

A major contribution of this thesis is the introduction of the *entropic* transport map, or *entropic Brenier map*, from Chapter 2:

$$x \mapsto T_\varepsilon(x) := \mathbb{E}_{\pi_\varepsilon}[Y|X = x],$$

where recall π_ε is the optimal entropic coupling between two measures P and Q. Unlike the optimal transport map, the entropic Brenier map always uniquely exists under mild conditions (e.g., if the marginals have finite second moment), is defined for all $x \in \mathbb{R}^d$, and is automatically *analytic* in the interior of the domain of the source measure.

In this part of the thesis, we ask how we can exploit this newfound regularity of the entropic Brenier map, and simultaneously address theoretical questions pertaining to its unregularized counterpart (in the limit $\varepsilon \searrow 0$ regime).

Chapter 5: An entropic generalization of Caffarelli's contraction theorem via covariance inequalities

Many applications of the optimal transport map hinge on its regularity properties, such as its Lipschitz constant. Though, there are only a few instances when the Lipschitz constant of $\nabla\varphi_0$ can be precisely gleaned from the source and target measures. One such result is due to Caffarelli (2000): If $P \propto \exp(-V)$ and $Q \propto \exp(-W)$ with $\nabla^2 V \preceq \beta_V I$ and $\nabla^2 W \succeq \alpha_W I \succeq 0$, then

$$\|\nabla^2\varphi_0\|_{\mathrm{op}} \le \sqrt{\beta_V/\alpha_W}\,. \tag{1.11}$$

These types of Lipschitz estimates for transport maps have their use in transferring functional inequalities. As an example, suppose P satisfies what is known as a Poincaré inequality: there exists a constant C_P such that for any smooth function f

$$\mathrm{Var}_P(f) \le \mathrm{C}_P \mathbb{E}_{X\sim P}\|\nabla f(X)\|^2\,.$$

If we write $(\nabla\varphi_0)_\sharp P = Q$, with $\nabla\varphi_0$ uniformly L-Lipschitz, then one can easily show that Q satisfies a Poincaré inequality with constant $\mathrm{C}_Q \le L^2\mathrm{C}_P$.

The usual proof of Caffarelli's contraction theorem follows PDE-style arguments; see the

following survey by Kolesnikov (2011). In this chapter, we provide another (shorter) proof, based on the entropic Brenier map. We show that, under the usual Caffarelli assumptions stated above,

$$\|\nabla^2 \varphi_\varepsilon\|_{\mathrm{op}} \leq \frac{1}{2}\left(\sqrt{4\beta_V/\alpha_W + \beta_V^2 \varepsilon^2} - \beta_V \varepsilon\right). \tag{1.12}$$

The bound in (1.12) is tight as it is realized by Gaussians. Our proof of this result is a few lines and relies on two twin covariance inequalities: the Brascamp–Lieb inequality and Cramér–Rao inequalities. Taking the $\varepsilon \searrow 0$ limit, we recover Caffarelli's seminal result; to our knowledge, this is the shortest proof of this result. As an application, we prove a generalization of Caffarelli's statement a result due to Valdimarsson.

It is worth mentioning that in the years since our result was first made available, there have been numerous extensions. Conforti (2024) obtains results of a similar flavor to (1.12) but under weaker assumptions than strong log-concavity (though, they are unable to take the $\varepsilon \searrow 0$ limit in these cases) using techniques from stochastic calculus. More recently, Gozlan and Sylvestre (2025) have strengthened our technique to encompass more general conditions on the measures (for instance, they prove global Hölder estimates instead of Lipschitz estimates, or estimates when P is Cauchy), and a further improvement of our generalization of Valdimarsson's result.

The content of this chapter is based off the following article

> "An entropic generalization of Caffarelli's contraction theorem via covariance inequalities", in *Comptes Rendues Mathématique* (2023), with Sinho Chewi.

Chapter 6: Tight stability bounds for entropic Brenier maps

We now turn our attention to another long-standing question in the optimal transport community: for a fixed source measure ρ, is the mapping $\mu \mapsto T_0^\mu$ Hölder continuous with respect to the 2-Wasserstein distance? In other words, do there exist constants $C, \beta > 0$ such that for all

probability measures μ, ν with finite second moments,

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \leq C W_2^\beta(\mu, \nu)\,? \tag{1.13}$$

Since the inequality $W_2(\mu, \nu) \leq \|T_0^\mu - T_0^\nu\|_{L^2(\rho)}$ always holds, (1.13) would imply that the mapping $\mu \mapsto T_0^\mu$ is a bi-Hölder embedding of the Wasserstein space into $L^2(\rho)$. We call such an inequality a *stability bound*. A result of this type was first proven in the article by Gigli (2011), and has since received much attention in the optimal transport community (Delalande and Mérigot, 2023; Letrouit and Mérigot, 2024; Manole et al., 2024a; Mérigot et al., 2020).

The goal of this chapter is two-fold. First, we prove analogous stability results for the embedding given by entropic Brenier maps i.e., $\mu \mapsto T_\varepsilon^\mu$. A second, more ambitious question, is to see if stability bounds for the entropic Brenier map can yield new results for Brenier maps (much like in the Caffarelli setting). To this end, a corollary of our main result is the following stability bound for entropic Brenier maps between ρ, μ, ν which are assumed to lie in $B(0, R)$:

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq (1 + 2R^2/\varepsilon) W_2(\mu, \nu)\,,$$

Moreover, we give an example which demonstrates that, in generality, this result is tight. Armed with this result, we then prove the following stability bound for the *unregularized* Brenier maps

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \lesssim W_2^{1/3}(\mu, \nu)\,,$$

where we restrict μ, ν to be finitely supported in a ball of radius R with lower-bounded weights.

The content of this chapter is based off the following article

> "Tight stability for entropic Brenier maps", in *International Mathematics Research Notices* (2025), with Vincent Divol and Jonathan Niles-Weed

Optimization over the Wasserstein space

Chapter 7: Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space

In this final chapter, we study optimization problems that take place over (subsets of) the Wasserstein space: the metric space of (absolutely continuous) probability measures over $\mathbb{R}^d$ endowed with the 2-Wasserstein distance.

First, we develop a theory of finite-dimensional polyhedral subsets over the Wasserstein space and optimization of functionals over them via first-order methods. As an application of our theory, we turn to a widely studied infinite-dimensional optimization problem over the space of probability distributions: mean-field variational inference (MFVI) (Blei et al., 2017; Wainwright and Jordan, 2008). In MFVI, the practitioner seeks to approximate an unnormalized posterior density π over $\mathbb{R}^d$ by the closest product measure in the sense of the Kullback–Leibler divergence:

$$\pi^\star = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R})^{\otimes d}} \mathrm{KL}(\mu \| \pi) \,.$$

When π is strongly log-concave and log-smooth, we provide (1) approximation rates certifying that $\pi^\star$ is close to the minimizer $\pi^\star_\diamond$ of the KL divergence over a *polyhedral* set $\mathcal{P}_\diamond$,

$$\pi^\star_\diamond = \operatorname*{argmin}_{\mu \in \mathcal{P}_\diamond} \mathrm{KL}(\mu \| \pi) \,.$$

and (2) an algorithm for minimizing $\mathrm{KL}(\cdot \| \pi)$ over $\mathcal{P}_\diamond$ based on accelerated gradient descent over $\mathbb{R}^d$. As a byproduct of our analysis, we obtain the first end-to-end analysis for gradient-based algorithms for MFVI. We also discuss the implementation of our algorithm, with code available here.

These results are from

"Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space", to appear in *Foundations of Computational Mathematics* (2025+) and a preliminary abstract was accepted to the *Conference on Learning Theory* (COLT 2024), with Roger Jiang and Sinho Chewi.

Additional contributions

I had the (immense!) pleasure of taking part of many other collaborations during my PhD which, in the interest of preserving my sanity, did not make it into this thesis:

- "Wasserstein flow matching: Generative modeling over families of distributions" in the *Fourty-second International Conference on Machine Learning* (ICML 2025), with Doron Haviv, Dana Pe'er, and Brandon Amos;
- "Conditional simulation via entropic optimal transport: Toward non-parametric estimation of conditional Brenier maps", in the *28th International Conference on Artificial Intelligence and Statistics* (AISTATS 2025), with Ricardo Baptista, Michael Brennan, Youssef Marzouk, and Jonathan Niles-Weed;
- "Estimation of optimal transport maps in general function spaces", to appear in the *Annals of Statistics* (2025), with Vincent Divol and Jonathan Niles-Weed;
- "Progressive entropic optimal transport solvers", in *the 38th Conference on Neural Information Processing Systems* (NeurIPS 2024), with Parnian Kassraie, James Thornton, Jonathan Niles-Weed, and Marco Cuturi;
- "Learning costs for structured Monge displacements", in the *38th Conference on Neural Information Processing Systems* (NeurIPS 2024), with Michal Klein, Pierre Ablin, Eugene Ndiaye, Jonathan Niles-Weed, and Marco Cuturi;

- "Neural optimal transport with Lagrangian costs", in the *40th International Conference on Uncertainty in Artificial Intelligence* (UAI 2024), with Carles Domingo-Enrich, Ricky Tian-Qi Chen, and Brandon Amos;
- "Multisample flow matching: Straightening flows with minibatch couplings", in the *40th International Conference on Machine Learning* (ICML 2023), with Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky Tian-Qi Chen;
- "An explicit expansion of the Kullback–Leibler divergence along its Fisher–Rao gradient flow", in *Transactions on Machine Learning Research* (2023), with Carles Domingo-Enrich;
- "Debiaser beware: Pitfalls of centering regularized transport maps", in the *39th International Conference on Machine Learning* (ICML 2022), with Jonathan Niles-Weed and Marco Cuturi.

All remaining errors are my own.

1.3 Background

We henceforth denote the space of probability measures with finite second moments by $\mathcal{P}_2(\mathbb{R}^d)$. The class of such measures with densities (with respect to Lebesgue measure) are denoted by $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. The support of a probability measure μ is given by $\mathrm{supp}(\mu)$.

1.3.1 Optimal transport for the quadratic cost

For $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$, we define the set of *couplings* between P and Q by

$$\Pi(P,Q) \coloneqq \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \mid \pi(A \times \mathbb{R}^d) = P(A), \pi(\mathbb{R}^d \times A) = Q(A)\}. \tag{1.14}$$

The optimal transport distance under the squared-Euclidean cost, or the 2-*Wasserstein distance*, between P (the source measure) and Q (the target measure) is given by the following optimization

problem

$$\tfrac{1}{2}W_2^2(P,Q) \coloneqq \inf_{\pi\in\Pi(P,Q)} \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y)\,. \tag{1.15}$$

This optimization problem is commonly known as the Kantorovich formulation of optimal transport (Kantorovitch, 1942). As P, Q are assumed to have finite-second moments, a minimizer to (1.15) is always guaranteed to exist (Villani, 2009). We call this minimizer the *optimal (transport) coupling* between P and Q, and is denoted by π_0. Though, importantly, uniqueness of the minimizer cannot be asserted from the sole assumption that the marginals have finite-second moments.

It is natural to view view (1.15) as an convex program (albeit infinite-dimensional). Thus, we can obtain a "dual" optimization problem,

$$\tfrac{1}{2}W_2^2(P,Q) = \sup_{(f,g)\in\mathcal{F}} \mathcal{D}_0^{PQ}(f,g)\,, \tag{1.16}$$

where $\mathcal{F} \coloneqq \{(f,g) : f \in L^1(P), g \in L^1(Q)\}$, and

$$\mathcal{D}_0^{PQ}(f,g) \coloneqq \int f \,\mathrm{d}P + \int g \,\mathrm{d}Q \quad \text{s.t.} \quad f(x)+g(y) \le \tfrac{1}{2}\|x-y\|^2\,,$$

the constraint holds $P \otimes Q$ almost everywhere. If the marginals have finite-second moments, then there exists a maximizing pair of *Kantorovich potentials* (f_0, g_0) to (1.16) (see Villani, 2009).[3]

We require a final formulation of the 2-Wasserstein distance based on transport maps, which are vector-valued functions $T : \mathbb{R}^d \to \mathbb{R}^d$ with $T \in \mathcal{T}(P,Q)$ such that they satisfy the *pushforward property*.[4] The *optimal transport map* between P and Q is the solution to the following optimization

[3] These potentials are defined up to a translation: for $c \in \mathbb{R}$, $(f_0 + c, g_0 - c)$ gives the same objective value in (1.16) as the original pair (f_0, g_0).

[4] We say that $T \in \mathcal{T}(P,Q)$ if for $X \sim P$, then $T(X) \sim Q$

problem, called the *Monge problem* (Monge, 1781)

$$T_0 \coloneqq \operatorname*{argmin}_{T \in \mathcal{T}(P,Q)} \int \tfrac{1}{2}\|x - T(x)\|^2 \, \mathrm{d}P(x) \,. \tag{1.17}$$

In contrast to the primal Kantorovich formulation of the 2-Wasserstein distance (recall (1.15)): (1) (1.17) is a non-convex optimization problem and (2) a minimizer T_0 may not even be defined for arbitrary measures with finite second moment, whereas π_0 will at least exist (though possibly not unique).

The following theorem, due to Brenier (1991), unifies the solutions to the Kantorovich primal and dual problems, and the Monge problem by assuming that the source measure has a density.

Theorem 1.1 (Brenier's theorem). *For $P \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $Q \in \mathcal{P}(\mathbb{R}^d)$, let (f_0, g_0) denote the optimal Kantorovich potentials which solve (1.16), and define*

$$\varphi_0 \coloneqq \tfrac{1}{2}\|\cdot\|^2 - f_0 \,, \quad \psi_0 \coloneqq \tfrac{1}{2}\|\cdot\|^2 - g_0 \,, \tag{1.18}$$

to be corresponding Brenier *potentials. Then the (P-a.e. unique) optimal transport map T_0 between P and Q exists (P-a.e.) and is given by the gradient of a convex function φ_0. In other words, T_0 minimizes (1.17) and is given by*

$$T_0 \coloneqq \nabla \varphi_0 \,, \tag{1.19}$$

Moreover, we can write the optimal plan as $\mathrm{d}\pi_0(x, y) = \mathrm{d}P(x)\delta_{\{T_0(x)\}}(y)$.

If Q also has a density, we can similarly write the optimal transport map from Q to P (or the inverse *optimal transport map) as*

$$(T_0)^{-1} = \nabla \psi_0 \,, \tag{1.20}$$

which holds Q-a.e., and ψ_0 is also convex. Analogously, the optimal transport coupling can be expressed as $\mathrm{d}\pi_0(x,y) = \mathrm{d}Q(y)\delta_{\{(T_0)^{-1}(y)\}}(x)$.

From their definition and (1.16), it can be shown that the (forward) Brenier potential φ_0 minimizes the following version of the Kantorovich dual objective

$$\begin{aligned} \tfrac{1}{2}W_2^2(P,Q) &= \tfrac{1}{2}M_2(P+Q) - \min_{\varphi\in L^1(P)} \mathcal{S}_0^{PQ}(\varphi) \\ &\coloneqq \tfrac{1}{2}M_2(P+Q) - \Big(\min_{\varphi\in L^1(P)} \int \varphi \,\mathrm{d}P + \int \varphi^* \,\mathrm{d}Q\Big), \end{aligned} \tag{1.21}$$

where φ^* is the convex conjugate operator. In fact, φ_0 and ψ_0 are convex conjugates of one another in the sense that

$$\varphi_0(x) = \sup_y \{\langle x,y\rangle - \psi_0(y)\}, \qquad \psi_0(y) = \sup_x \{\langle x,y\rangle - \varphi_0(x)\}. \tag{1.22}$$

Thus, when all quantities are well-defined, we can write $\nabla\psi_0 = \nabla\varphi_0^* = (\nabla\varphi_0)^{-1}$ by standard results in convex analysis.

1.3.2 Entropic optimal transport for the quadratic cost

Let $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$. For a fixed parameter $\varepsilon > 0$, the *entropic optimal transport* problem between P and Q is

$$\mathrm{OT}_\varepsilon(P,Q) \coloneqq \inf_{\pi\in\Pi(P,Q)} \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi\|P\otimes Q), \tag{1.23}$$

where we define the *Kullback–Leibler divergence* as

$$\mathrm{KL}(\pi\|P\otimes Q) \coloneqq \int \log\Big(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}P(x)\,\mathrm{d}Q(y)}\Big)\,\mathrm{d}\pi(x,y),$$

whenever π admits a density with respect to $P \otimes Q$, and $+\infty$ otherwise. Note that when $\varepsilon = 0$, (1.23) reduces to (1.15). The entropic optimal transport problem was introduced to the machine learning community by Cuturi (2013) as a numerical scheme for approximating the 2-Wasserstein distance on the basis of samples.

An important consequence of the added regularization is that (1.23) is a strictly convex problem, and thus always admits a unique minimizer whenever P and Q have finite second moments (Genevay, 2019). We call this minimizer the *optimal entropic plan*, written $\pi_\varepsilon \in \Pi(P, Q)$.

As with the unregularized Kantorovich problem (1.16), a dual formulation of (1.23) exists

$$\mathrm{OT}_\varepsilon(P, Q) = \sup_{(f,g)\in\mathcal{F}} \mathcal{D}_\varepsilon^{PQ}(f, g) \tag{1.24}$$

where

$$\mathcal{D}_\varepsilon^{PQ}(f, g) := \int f \,\mathrm{d}P + \int g \,\mathrm{d}Q - \varepsilon \iint \left(e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} - 1\right) \mathrm{d}P(x)\,\mathrm{d}Q(y)\,. \tag{1.25}$$

Interestingly, as $\varepsilon \to 0$, we see that $\mathcal{D}_\varepsilon^{PQ}$ converges to the objective in (1.16), where the limit of the third term above becomes the hard constraint on the potentials. As with the primal problem, the assumption that P and Q have finite second moments ensures that there exists a unique maximizing pair $(f_\varepsilon, g_\varepsilon)$ to (1.24), which we call *entropic* Kantorovich potentials; see Genevay (2019); Nutz (2021) for more details on this point. As with the maximizers to (1.16), these functions are defined up to a constant translation.

The primal and dual optima are intimately connected through the following relationship due to Csiszár (1975):

$$\mathrm{d}\pi_\varepsilon(x, y) = \exp\left(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x - y\|^2}{\varepsilon}\right) \mathrm{d}P(x)\,\mathrm{d}Q(y)\,. \tag{1.26}$$

Finally, we mention that, although f_ε and g_ε are only *a priori* defined almost everywhere on the

support of P and Q, they can be extended to all of $\mathbb{R}^d$ (see Mena and Niles-Weed (2019); Nutz and Wiesel (2021)) via the optimality conditions

$$1 = \int e^{(f_\varepsilon(x)+g_\varepsilon(y)-\|x-y\|^2/2)/\varepsilon} \, \mathrm{d}Q(y) \quad \forall x \in \mathbb{R}^d \,, \tag{1.27}$$

$$1 = \int e^{(f_\varepsilon(x)+g_\varepsilon(y)-\|x-y\|^2/2)/\varepsilon} \, \mathrm{d}P(x) \quad \forall y \in \mathbb{R}^d \,. \tag{1.28}$$

Since $(f_\varepsilon, g_\varepsilon)$ are only unique up to adding a constant to f_ε and subtracting the same constant from g_ε. Unless specified otherwise, we will always assume the normalization convention $\int f_\varepsilon \, \mathrm{d}P = \int g_\varepsilon \, \mathrm{d}Q$.

1.3.3 Other notation

Basic definitions. The square-root of the determinant of a matrix is $J(\cdot) := \sqrt{\det(\cdot)}$. For $x \in \mathbb{R}^d$ and $r > 0$, we write $B_r(x)$ for the Euclidean ball of radius r centered at x. We denote the maximum and minimum of a and b by $a \vee b$ and $a \wedge b$, respectively. We use the symbols c and C to denote positive constants whose value may change from line to line, and write $a \lesssim b$ and $a \asymp b$ if there exists constants $c, C > 0$ such that $a \leq Cb$ and $cb \leq a \leq Cb$, respectively.

Function classes. For $\alpha \geq 0$ and a closed set Ω, we write $h \in C^\alpha(\Omega)$ if there exists an open set $U \supseteq \Omega$ and a function $g : U \to \mathbb{R}$ such that $g|_\Omega = h$ and such that g possesses $\lfloor \alpha \rfloor$ continuous derivatives and whose $\lfloor \alpha \rfloor$th derivative is $(\alpha - \lfloor \alpha \rfloor)$-Hölder smooth.

We write third total derivative of f at x in the direction $y \in \mathbb{R}^d$ as

$$d^3 f(x; y) := \sum_{i,j,k=1}^{d} \frac{\partial^3 f(x)}{\partial y_i \partial y_j \partial y_k} y_i y_j y_k \,.$$

Space of probability measures and divergences. For a function f and a probability measure ρ, we write $\|f\|^2_{L^2(\rho)} := \mathbb{E}_{X\sim\rho}\|f(X)\|^2$. Similarly, we write $\mathrm{Var}_\rho(f) := \mathbb{E}_{X\sim\rho}[(f(X) - \mathbb{E}_{X\sim\rho}[f(X)])^2]$

for the variance of f with respect to ρ.

A probability measure is called σ^2-subGaussian if for some $\sigma^2 > 0$,

$$\mathbb{E}\exp(\lambda^\top(Y - \mathbb{E}Y)) \leq \exp(\|\lambda\|^2\sigma^2/2)\,, \qquad \text{for all } \lambda \in \mathbb{R}^d\,.$$

If a measure ρ possesses a density with respect to the Lebesgue measure, we denote its differential entropy by $\mathcal{H}(\rho) = \int \log(\mathrm{d}\rho)\,\mathrm{d}\rho$.

We will use several divergences throughout this thesis apart from the Kullback–Leibler divergence. For instance, the total variation distance, as well as the χ-squared divergence and (squared) Hellinger distance, between two probability measures $P \ll Q$ are given by

$$\mathrm{TV}(P, Q) = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |P(A) - Q(A)|\,, \tag{1.29}$$

$$\chi^2(P\|Q) = \int \left(1 - \frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q\,, \tag{1.30}$$

$$\mathrm{H}^2(P, Q) = \tfrac{1}{2}\int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2. \tag{1.31}$$

Part I

Statistical estimation of optimal transport maps and beyond

2 Entropic estimation of optimal transport maps

2.1 Introduction

The goal of optimal transport is to find a map between two probability distributions that minimizes the squared Euclidean transportation cost. This formulation leads to what is known as the *Monge problem* (Monge, 1781):

$$\min_{T \in \mathcal{T}(P,Q)} \int \|x - T(x)\|^2 \, \mathrm{d}P(x)\,, \tag{2.1}$$

where $\mathcal{T}(P, Q)$ is the family of admissible transport maps from P to Q, i.e., for $X \sim P$, $T(X) \sim Q$. Due to their versatility and mathematical simplicity, optimal transport maps have found a wide range of uses in statistics and machine learning, (Arjovsky et al., 2017; Carlier et al., 2016; Chernozhukov et al., 2017; Courty et al., 2014; 2017; Finlay et al., 2020a; Huang et al., 2021a; Makkuva et al., 2020; Onken et al., 2021; Wang et al., 2010), computer graphics (Feydy et al., 2017; Solomon et al., 2015; 2016), and computational biology (Schiebinger et al., 2019; Yang et al., 2020), among other fields.

Of course, in these applied works, rarely are P and Q known exactly but rather the practitioner deals with samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_n \sim Q = (T_0)_\sharp P$, and the goal is to estimate the

optimal transport map T_0 on the basis of the data. Hütter and Rigollet (2021) first investigated this question and proposed an estimator $\hat{T}_n$ which achieves

$$\mathbb{E}\|\hat{T}_n - T_0\|^2_{L^2(P)} \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} \log^3(n)\,, \tag{2.2}$$

if $T_0 \in C^\alpha$, P and Q are compactly supported, and satisfy additional technical assumptions. Moreover, they showed that the rate in (2.2) is minimax optimal up to logarithmic factors. Though statistically optimal, their estimator is impractical to compute if $d > 3$, since it relies on a gridding scheme whose computational cost scales exponentially in the dimension. Recently, Deb et al. (2021) and Manole et al. (2024a) proposed plugin estimators that also achieve the minimax estimation rate. Though simpler to compute than the estimator of Hütter and Rigollet (2021), these estimators require at least $O(n^3)$ time to compute and cannot easily be parallelized, making them an unfavorable choice when the number of samples is large.

In this chapter, we adopt a different approach by leveraging recent advances in computational optimal transport based on entropic regularization (Peyré and Cuturi, 2019), which replaces (2.1) by

$$\inf_{\pi\in\Pi(P,Q)} \iint \frac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi\|P\otimes Q)\,, \tag{2.3}$$

where $\Pi(P,Q)$ denotes the set of couplings between P and Q and $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. This approach, which was popularized by Cuturi (2013), has been instrumental in the adoption of optimal transport methods in the machine learning community because it leads to a problem that can be solved by Sinkhorn's algorithm Sinkhorn (1967), whose time complexity scales *quadratically* in the number of samples (Altschuler et al., 2017). Moreover, Sinkhorn's algorithm is amenable to parallel implementation on GPUs, making it very attractive for large-scale problems (Altschuler et al., 2019; Feydy et al., 2020; 2019; Genevay et al., 2018).

2.1.1 Contributions

The efficiency and popularity of Sinkhorn's algorithm raise the tantalizing question of whether it is possible to use this practical technique to develop estimators of optimal transport maps with convergence guarantees. We develop such a procedure.

Under suitable technical assumptions on P and Q, we show that our estimator $\hat{T}$ enjoys the rate

$$\mathbb{E}\|\hat{T} - T_0\|^2_{L_2(P)} \lesssim n^{-\frac{(\alpha+1)}{2(d+\alpha+1)}} \log n$$

if the inverse map T_0^{-1} is C^α and $\alpha \in (1, 3]$. This rate is worse than that in (2.2), but our empirical results show that our estimator nevertheless outperforms all other estimators proposed in the literature in terms of *both* computational and statistical performance. The estimator we analyze was originally suggested by Seguy et al. (2018), who also showed consistency of the entropic plan in the large-n limit if the regularization parameter is taken to zero sufficiently fast. However, to our knowledge, our work offers the first finite-sample convergence guarantees for this proposal.

Our estimator is defined as the barycentric projection (Ambrosio et al., 2008) of the entropic optimal coupling between the empirical measures arising from the samples. The barycentric projection has been leveraged in other works on map estimation as a straightforward way of obtaining a function from a coupling between two probability measures (Deb et al., 2021). However, in the context of entropic optimal transport, this operation has a more canonical interpretation in light of Brenier's theorem (Brenier, 1991). Brenier's result says that the optimal transport map $T_0 = \nabla\varphi_0$ can be realized as the gradient of the function which solves the dual problem to (2.1).

We show in Proposition 2.3 that the barycentric projection of the entropic optimal coupling is the gradient of the function which solves the dual problem to (2.3). In addition to providing a connection to the classical theory of optimal transport, this observation provides a canonical extension $\hat{T}$ to out-of-sample points. Moreover, since Sinkhorn's algorithm computes solutions to the dual of (2.3), this interpretation shows that computing $\hat{T}$ is no more costly than solving (2.3).

Moreover, we propose a variant of our estimator that is *adaptive* in the sense that the smoothness parameter need not be explicitly known to the practitioner.

We analyze $\hat{T}$ by employing a strategy pioneered by Pal (2024) for understanding the structure of the optimal entropic coupling. This technique compares the solution to (2.3) to a coupling whose conditional laws are Gaussian, with mean and covariance characterized by the solution to (1.17). To leverage this comparison, we employ a duality principle in conjuction with an upper bound reminiscent of the short-time expansions of the value of (2.3) developed by Conforti and Tamanini (2021) and Chizat et al. (2020) (see Theorem 2.2).

2.1.2 Notation

A constant is a quantity whose value may depend on the smoothness parameters appearing in assumptions **(E1)** to **(E3)**, the set Ω, and the dimension, but on no other quantities. We denote the maximum and minimum of a and b by $a \vee b$ and $a \wedge b$, respectively. We use the symbols c and C to denote positive constants whose value may change from line to line, and write $a \lesssim b$ and $a \asymp b$ if there exists constants $c, C > 0$ such that $a \leq Cb$ and $cb \leq a \leq Cb$, respectively.

Our proofs based on empirical process theory will consider suprema over uncountable collections of random variables; however, since all the processes in question are separable, these suprema are still measurable (Giné and Nickl, 2021, Section 2.1).

2.1.3 Remaining background on entropic optimal transport

Throughout this chapter, we make use of the existing notation and conventions from Section 1.3.2. However, our proofs rely on a modified version of the duality relation given in (1.24), in which the supremum is taken over a larger set of functions. Though it is a straightforward consequence of Fenchel's inequality, we have not encountered this statement explicitly in the literature, so we highlight it here.

Proposition 2.1. *Assume P and Q possess finite second moments, and let π_ε be the optimal entropic plan for P and Q. Then*

$$\mathrm{OT}_\varepsilon(P,Q) = \sup_{\eta \in L^1(\pi_\varepsilon)} \int \eta \, \mathrm{d}\pi_\varepsilon - \varepsilon \iint e^{(\eta(x,y) - \frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}P(x) \, \mathrm{d}Q(y) + \varepsilon \,. \tag{2.4}$$

Comparing this proposition with (1.24), we see that we can always take $\eta(x,y) = f(x) + g(y)$, in which case (2.4) reduces to (1.24). The novelty in Proposition 2.1 therefore arises in showing that the quantity on the right side of (2.4) is still bounded above by $\mathrm{OT}_\varepsilon(P,Q)$. We give the short proof of Proposition 2.1 in Appendix A.3.

Several recent works have bridged the regularized and unregularized optimal transport regimes, with particular interest in the setting where $\varepsilon \to 0$. Convergence of π_ε to π_0 was studied by Carlier et al. (2017) and Léonard (2012), and recent work has quantified the convergence of the plans (Bernton et al., 2022; Ghosal et al., 2022; Hundrieser et al., 2024a; Klatt et al., 2020) and the potentials (Altschuler et al., 2022; Masud et al., 2023; Nutz and Wiesel, 2021; Rigollet and Stromme, 2022) in certain settings. Convergence of $\mathrm{OT}_\varepsilon(P,Q)$ to $\frac{1}{2}W_2^2(P,Q)$ has attracted significant research interest: under mild conditions, Pal (2024) proves a first-order convergence result for general convex costs (replacing $\frac{1}{2}\|\cdot\|^2$), and a second order expansion was subsequently obtained by Chizat et al. (2020) and Conforti and Tamanini (2021). We also rely on the following bound which we provide a short proof of in Appendix A.1.

Theorem 2.2. *Suppose P and Q have bounded densities with compact support. Then*

$$\mathrm{OT}_\varepsilon(P,Q) - \frac{1}{2}W_2^2(P,Q) + \varepsilon \log((2\pi\varepsilon)^{d/2}) \le -\frac{\varepsilon}{2}\left(\mathcal{H}(P) + \mathcal{H}(Q)\right) + \frac{\varepsilon^2}{8} I_0(P,Q) \,, \tag{2.5}$$

where $I_0(P,Q)$ is the integrated Fisher information along Wasserstein geodesics*, given by*

$$I_0(P,Q) := \int_0^1 \int \|\nabla \log P_t(x)\|^2 \, \mathrm{d}P_t(x) \, \mathrm{d}t \,, \tag{2.6}$$

where $P_t := ((1-t)\mathrm{id} + tT_0)_\sharp P$.

2.2 Estimator and main results

Given the optimal entropic plan π_ε between P and Q, we define its barycentric projection to be

$$T_\varepsilon(x) := \int y \, \mathrm{d}\pi_\varepsilon^x(y) = \mathbb{E}_{\pi_\varepsilon}[Y \mid X = x] \,. \tag{2.7}$$

A priori, this map is only defined P-almost everywhere, making it unsuitable for evaluation outside the support of P. In particular, since we will study the barycentric projection obtained from the optimal entropic plan between empirical measures, this definition does not extend outside the sample points. However, the duality relations (recall (1.27) and (1.28)) implies that we may define a version of the conditional density of Y given $X = x$ for *all* $x \in \mathbb{R}^d$ by

$$\mathrm{d}\pi_\varepsilon^x(y) = e^{\frac{1}{\varepsilon}(f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q(y) = \frac{e^{\frac{1}{\varepsilon}(g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q(y)}{\int e^{\frac{1}{\varepsilon}(g_\varepsilon(y')-\frac{1}{2}\|x-y'\|^2)} \, \mathrm{d}Q(y')} \,,$$

where $(f_\varepsilon, g_\varepsilon)$ are the optimal entropic potentials. This furnishes an extension of T_ε to all of $\mathbb{R}^d$ by

$$T_\varepsilon(x) := \frac{\int y e^{\frac{1}{\varepsilon}(g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q(y)}{\int e^{\frac{1}{\varepsilon}(g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q(y)} \,.$$

We call T_ε the *entropic map* between P and Q, though we stress that $(T_\varepsilon)_\sharp P \neq Q$ in general. This natural definition is motivated by the following observation, which shows that the entropic map can also be defined as the map obtained by replacing the optimal potential in Brenier's theorem by its entropic counterpart. To this end, write

$$(\varphi_\varepsilon, \psi_\varepsilon) := \left(\tfrac{1}{2}\|\cdot\|^2 - f_\varepsilon, \tfrac{1}{2}\|\cdot\|^2 - g_\varepsilon\right) \tag{2.8}$$

to be *entropic Brenier potentials*, and observe that

$$T_\varepsilon(x) := \frac{\int y e^{\frac{1}{\varepsilon}(x^\top y - \psi_\varepsilon(y))}\, \mathrm{d}Q(y)}{\int e^{\frac{1}{\varepsilon}(x^\top y - \psi_\varepsilon ps(y))}\, \mathrm{d}Q(y)}\,. \tag{2.9}$$

Proposition 2.3. *Let $(\varphi_\varepsilon, \psi_\varepsilon)$ be optimal entropic Brenier potentials in the sense of (2.8), and let T_ε be the entropic map. Then $T_\varepsilon = \nabla\varphi_\varepsilon$.*

Proof. The dual optimality conditions (1.27) implies

$$f_\varepsilon(x) = -\varepsilon \log \int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\, \mathrm{d}Q(y)\,.$$

Taking the gradient of this expression yields

$$\begin{aligned}\nabla f_\varepsilon(x) &= -\varepsilon \frac{\int (-(x-y)/\varepsilon) e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\, \mathrm{d}Q(y)}{\int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\, \mathrm{d}Q(y)} \\ &= x - \frac{\int y e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\, \mathrm{d}Q(y)}{\int e^{(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon}\, \mathrm{d}Q(y)} = x - T_\varepsilon(x)\,.\end{aligned}$$

□

We write $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for the empirical distributions corresponding to the samples from P and Q, respectively. Our proposed estimator is $T_{\varepsilon,(n,n)}$, the entropic map between P_n and Q_n, which can be written explicitly as

$$T_{\varepsilon,(n,n)}(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}{\frac{1}{n}\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}\,, \tag{2.10}$$

where $g_{\varepsilon,(n,n)}$ is the optimal entropic potential corresponding to Q_n in the optimal entropic plan between P_n and Q_n, which can be obtained as part of the output of Sinkhorn's algorithm (see Peyré and Cuturi, 2019). In other words, once the optimal entropic potential is found, the map $T_{\varepsilon,(n,n)}(x)$ can therefore be evaluated in linear time. We discuss these computational aspects thoroughly in

Section 2.4. As in standard nonparametric estimation (Tsybakov, 2009), the optimal choice of ε will be dictated by the smoothness of the target function.

Remark 2.4. We briefly take a moment to discuss the applicability of our estimator in a wider statistical context. A body of work (e.g., Chernozhukov et al., 2017; Hallin et al., 2021) studies the estimation of multivariate ranks and quantiles through *inverse* optimal transport maps. For this purpose, it is important that estimators of transport maps be invertible. We remark that the entropic map as defined above has this property since it is strongly monotone, in the sense that $(T_\varepsilon(x) - T_\varepsilon(y))^\top (x - y) > 0$ (see Rigollet and Stromme, 2022, Proposition 10). However, our procedure also gives rise to an even simpler estimator for the inverse transport map, namely the map $T_\varepsilon^{\mathrm{inv}} := \mathrm{id} - \nabla g_\varepsilon$. By interchanging the roles of P and Q in our assumptions, we can provide both computational and statistical guarantees for this map as well.

To prove quantitative rates of convergence for $T_{\varepsilon,(n,n)}$, we make the following regularity assumptions on P and Q:

(E1) $P, Q \in \mathcal{P}_{ac}(\Omega)$ for a compact set Ω, with densities satisfying $p(x), q(x) \leq M$ and $q(x) \geq m > 0$ for all $x \in \Omega$,

(E2) $\varphi_0 \in C^2(\Omega)$ and $\varphi_0^* \in C^{\alpha+1}(\Omega)$ for $\alpha > 1$,

(E3) $T_0 = \nabla\varphi_0$, with $\mu I \preceq \nabla^2\varphi_0(x) \preceq LI$ for $\mu, L > 0$ for all $x \in \Omega$,

In what follows, all constants may depend on the dimension, the set Ω, M, m, μ, L, and $\|\varphi_0^*\|_{C^{\alpha+1}}$.

The above assumptions are qualitatively similar to those that have appeared in previous works on the estimation of optimal transport maps.

(E1) is a standard assumption in the statistical analysis of optimal transport map estimation. (It is present in the works of, e.g., Deb et al. (2021); Hütter and Rigollet (2021); Manole et al. (2024a); Vacher et al. (2024).) All of these works require that P and Q be compactly supported. Some of the tools we employ extend beyond the compact support setting; for example, Conforti

and Tamanini (2021) show that the expansion presented in Theorem 2.2 continues to hold for unbounded measures under suitable moment assumptions. However, our proofs require strong *a priori* bounds on the optimal transport map as well as on the entropic coupling for the random empirical measures P_n and Q_n, which do not have clear analogues in the non-compact setting.

(E3) is also standard, and in prior work it has often been assumed implicitly as a consequence of a strengthened form of **(E1)**. Caffarelli's regularity theory (Caffarelli, 1992) guarantees that if we assume that the set Ω in **(E1)** is *convex* and that the density p is also bounded below, then T_0 is continuous; if we further assume that $p, q \in C^\beta(\Omega)$ for any $\beta > 0$, then **(E3)** holds. **(E3)** can therefore be viewed as being only slightly stronger than **(E1)**, so long as Ω is convex. **(E3)** plays a crucial role in this and prior work, since, as was originally noticed by Ambrosio (see Gigli, 2011), this assumption guarantees stability of the optimal transport map, as a function of the source and target measures.

Our most unusual assumption is **(E2)**. Prior work analyzes estimators for T_0 under the assumption that $\varphi_0 \in C^{\alpha+1}(\Omega)$ for $\alpha > 1$, with rates that depend on α. For technical reasons, our proofs require a Laplace expansion in the "target space" corresponding to the dual Brenier potential φ_0^*. Consequently, we instead assume that $\varphi_0^* \in C^{\alpha+1}(\Omega)$, so that our rates depend on the smoothness of the *inverse* map T_0. We elaborate on this point further in the discussions surrounding Lemma 2.9.

Our main result is the following.

Theorem 2.5. *Under assumptions **(E1)** to **(E3)**, the entropic map $\hat{T} = T_{\varepsilon,(n,n)}$ from P_n to Q_n with regularization parameter $\varepsilon \asymp n^{-\frac{1}{d+\bar{\alpha}+1}}$ satisfies*

$$\mathbb{E}\|\hat{T} - T_0\|^2_{L^2(P)} \lesssim (1 + I_0(P, Q))n^{-\frac{(\bar{\alpha}+1)}{2(d+\bar{\alpha}+1)}} \log n\,,$$

where $\bar{\alpha} = \alpha \wedge 3$.

When $d \to \infty$ and $\alpha \to 1$, we formally obtain the rate $n^{-(1+o(1))/d}$. By contrast, Hütter and Rigollet (2021) show that, up to logarithmic factors, the rate $n^{-2(1+o(1))/d}$ is minimax optimal in

this setting. Theorem 2.5 therefore falls short of the minimax rate by a factor of approximately 2 in the exponent; however, our numerical experiments in Section 2.4 show that $\hat{T}$ is competitive with minimax-optimal estimators in practice.

To analyze our estimator, we adopt a two-step approach. We first consider the one-sample setting and show that the entropic map $T_{\varepsilon,n}$ between P and Q_n is close to T_0 in expectation. We prove the following.

Theorem 2.6. *Under assumptions **(E1)** to **(E3)** there exists a constant $\varepsilon_0 > 0$ such that for $\varepsilon \leq \varepsilon_0$, the entropic map $T_{\varepsilon,n}$ between P and Q_n satisfies*

$$\mathbb{E}\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \lesssim \varepsilon^{1-d/2}\log(n)n^{-1/2} + \varepsilon^{(\bar{\alpha}+1)/2} + \varepsilon^2 I_0(P,Q)\,,$$

with $\bar{\alpha} = \alpha \wedge 3$. Choosing $\varepsilon \asymp n^{-\frac{1}{d+\bar{\alpha}-1}}$, we get the one-sample estimation rate

$$\mathbb{E}\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \lesssim (1 + I_0(P,Q))n^{-\frac{\bar{\alpha}+1}{2(d+\bar{\alpha}-1)}}\,.$$

Remark 2.7. It can happen that $I_0(P,Q)$ is infinite, so the bounds of Theorem 2.5 and 2.6 are sometimes vacuous. However, Chizat et al. (2020) prove that $I_0(P,Q) \leq C$ for a positive constant C when P and Q satisfy **(E1)** to **(E3)** for $\alpha \geq 2$. Therefore, in this range for α, we obtain the rates in the theorems above without additional restrictions.

As a corollary to Theorem 2.6, we have the following population-level estimate between T_ε and T_0, which is potentially of independent interest.

Corollary 2.8. *Assume **(E1)** to **(E3)**, then*

$$\|T_\varepsilon - T_0\|^2_{L^2(P)} = \|\nabla\varphi_\varepsilon - \nabla\varphi_0\|^2_{L^2(P)} \lesssim \varepsilon^2 I_0(P,Q) + \varepsilon^{(\bar{\alpha}+1)/2}, \tag{2.11}$$

where $\bar{\alpha} = 3 \wedge \alpha$.

The proof of Theorem 2.6 is technical, and our approach is closely inspired by Pal (2024) and empirical process theory arguments developed by Genevay et al. (2019) and Mena and Niles-Weed (2019). We give a summary of our argument here, and carry out the details in the following section.

Following Pal (2024), we define the *divergence* $D[y|x^*] := -x^\top y + \varphi_0(x) + \varphi_0^*(y)$, where φ_0 solves the semidual (1.21). Though this quantity is a function of x and y, it is notationally convenient to write it in a way that highlights its dependence on $x^* := T_0(x)$. Indeed, we rely throughout on the following fact

Lemma 2.9. *Under assumptions **(E2)** and **(E3)**, for any $x \in \mathrm{supp}(P)$, we have*

$$D[y|x^*] = \frac{1}{2}(y - x^*)^\top \nabla^2 \varphi_0^*(x^*)(y - x^*) + o(\|y - x^*\|^2) \quad \text{as } y \to x^*, \tag{2.12}$$

as well as the non-asymptotic bound

$$\frac{1}{2L}\|y - x^*\|^2 \le D[y|x^*] \le \frac{1}{2\mu}\|y - x^*\|^2. \tag{2.13}$$

Proof. This follows directly from Taylor's theorem and the fact that $\nabla \varphi_0^*(x^*) = T_0^{-1}(x^*) = x$. □

We then define a conditional probability density in terms of this divergence:

$$q_\varepsilon^x(y) = \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} e^{\frac{-1}{\varepsilon} D[y|x^*]}, \quad Z_\varepsilon(x) := \frac{1}{\Lambda_\varepsilon} \int \exp\left(-\frac{1}{\varepsilon} D[y|x^*]\right) dy, \tag{2.14}$$

for $\Lambda_\varepsilon = (2\pi\varepsilon)^{d/2}$. By virtue of (2.12), if φ_0^* is sufficiently smooth, then q_ε^x will be approximately Gaussian with mean x^* and covariance $\varepsilon \nabla^2 \varphi_0^*(x^*)^{-1} = \varepsilon \nabla^2 \varphi_0(x)$. We quantify this approximation via Laplace's method; details appear in Appendix A.2. Using variational arguments, reminiscent of those employed by Bobkov and Götze (1999) in the study of transportation inequalities, we then compare the measure $\pi_{\varepsilon,n}$ to the measure $q_\varepsilon^x(y)\,dy\,dP(x)$ and compute accurate estimates of $T_{\varepsilon,n}$ via Laplace's method.

A similar but much simpler argument establishes the following bound in the two-sample case.

Theorem 2.10. *Let $T_{\varepsilon,(n,n)}$ be the entropic map from P_n to Q_n, and let $T_{\varepsilon,n}$ be as in Theorem 2.6. Under assumptions **(E1)** to **(E3)**, for $\varepsilon \leq 1$, $T_{\varepsilon,(n,n)}$ satisfies*

$$\mathbb{E}\|T_{\varepsilon,(n,n)} - T_{\varepsilon,n}\|^2_{L^2(P)} \lesssim \varepsilon^{-d/2} \log(n) n^{-1/2} .$$

Combining Theorem 2.6 and 2.10 yields our main result.

Proof of Theorem 2.5. We have

$$\begin{aligned}\mathbb{E}\|T_{\varepsilon,(n,n)} - T_0\|^2_{L^2(P)} &\lesssim \mathbb{E}\|T_{\varepsilon,(n,n)} - T_{\varepsilon,n}\|^2_{L^2(P)} + \mathbb{E}\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \\ &\lesssim \varepsilon^{-d/2} \log(n) n^{-1/2} + \varepsilon^{(\tilde{\alpha}+1)/2} + \varepsilon^2 I_0(P,Q) .\end{aligned}$$

Choosing $\varepsilon \asymp n^{-\frac{1}{d+\tilde{\alpha}+1}}$ yields the bound. □

2.2.1 One-sample estimates

In this section, we prove Theorem 2.6, which relates T_0 to the entropic map between P and Q_n:

$$T_{\varepsilon,n}(x) = \frac{\int y e^{\frac{1}{\varepsilon}(g_{\varepsilon,n}(y) - \frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q_n(y)}{\int e^{\frac{1}{\varepsilon}(g_{\varepsilon,n}(y) - \frac{1}{2}\|x-y\|^2)} \, \mathrm{d}Q_n(y)} = \int y \, \mathrm{d}\pi^x_{\varepsilon,n}(y) ,$$

where $\pi_{\varepsilon,n}$ is the optimal entropic plan for P and Q_n. We stress that since $T_{\varepsilon,n}$ is based on the entropic map from P to Q_n, the second equality holds for P-almost every x.

Our main tool is the following inequality, which allows us to compare $\pi_{\varepsilon,n}$ to the measure constructed from the conditional densities q^x_ε. The proof relies crucially on Proposition 2.1 and on the second order-expansion provided in Theorem 2.2.

Proposition 2.11. *Assume **(E1)** to **(E3)**, and let $a \in [L\varepsilon, 1]$ for $\varepsilon \le 1$. Then*

$$\mathbb{E}\Big\{ \sup_{h:\Omega\to\mathbb{R}^d} \iint \big(h(x)^\top (y - T_0(x)) - a\|h(x)\|^2\big)\, \mathrm{d}\pi_{\varepsilon,n}(x,y) \\ - \iint (e^{h(x)^\top (y-T_0(x)) - a\|h(x)\|^2} - 1) q_\varepsilon^x(y)\, \mathrm{d}y\, \mathrm{d}P(x)\Big\} \\ \lesssim \varepsilon I_0(P,Q) + \varepsilon^{(\bar\alpha-1)/2} + \varepsilon^{-d/2}\log(n) n^{-1/2},$$

where the supremum is taken over all $h \in L^2(P)$.

Proof. Given $h \in L^2(P)$, write

$$j_h(x,y) = h(x)^\top (y - T_0(x)) - a\|h(x)\|^2 .$$

Choosing $\eta(x,y) = \varepsilon(j_h(x,y) + \log(q_\varepsilon^x(y)/q(y))) + \|x-y\|^2/2$ and applying Proposition 2.1 with the measures P and Q_n, we obtain

$$\sup_{h:\Omega\to\mathbb{R}^d} \int j_h\, \mathrm{d}\pi_{\varepsilon,n} + \int \log \frac{q_\varepsilon^x(y) e^{\frac{1}{2\varepsilon}\|x-y\|^2}}{q(y)}\, \mathrm{d}\pi_{\varepsilon,n}(x,y) \\ - \iint e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)}\, \mathrm{d}Q_n(y)\, \mathrm{d}P(x) + 1 \le \varepsilon^{-1} \mathrm{OT}_\varepsilon(P, Q_n) .$$

We first expand $\iint \log \frac{q_\varepsilon^x(y) e^{\frac{1}{2\varepsilon}\|x-y\|^2}}{q(y)}\, \mathrm{d}\pi_{\varepsilon,n}(x,y)$, where we use the fact that $\pi_{\varepsilon,n}$ has marginals P and Q_n:

$$\begin{aligned} &\iint \log \frac{q_\varepsilon^x(y) e^{\frac{1}{2\varepsilon}\|x-y\|^2}}{q(y)}\, \mathrm{d}\pi_{\varepsilon,n}(x,y) \\ &\quad = \frac{1}{\varepsilon} \iint \left[f_0(x) + g_0(y) + \varepsilon \log\left(\frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon}\right) - \varepsilon \log(q(y)) \right] \mathrm{d}\pi_{\varepsilon,n}(x,y) \\ &\quad = \frac{1}{\varepsilon}\Big(\int f_0(x)\, \mathrm{d}P(x) + \int g_0(y)\, \mathrm{d}Q_n(y) \Big) - \log(\Lambda_\varepsilon) \\ &\qquad - \int \log(Z_\varepsilon(x))\, \mathrm{d}P(x) - \int \log(q(y))\, \mathrm{d}Q_n(y) , \end{aligned}$$

where (f_0, g_0) solve (1.16). Replacing Q_n by Q yields

$$\iint \log \frac{q_\varepsilon^x(y) e^{\frac{1}{2\varepsilon}\|x-y\|^2}}{q(y)} \,\mathrm{d}\pi_{\varepsilon,n}(x,y) = \frac{1}{2\varepsilon} W_2^2(P,Q) - \log(\Lambda_\varepsilon) - \int \log(Z_\varepsilon(x)) \,\mathrm{d}P(x) \\ - \mathcal{H}(Q) + \int (g_0/\varepsilon - \log(q))(\mathrm{d}Q_n - \mathrm{d}Q).$$

A change of variables (see Pal, 2024, Lemma 3(iv)) implies

$$\frac{\mathcal{H}(Q) - \mathcal{H}(P)}{2} = \int \log J(\nabla^2 \varphi_0^*(x^*)) \,\mathrm{d}P(x)\,,$$

where we recall that $x^* = T_0(x)$. Substituting this identity into the preceding expression yields

$$\begin{aligned}\iint \log \frac{q_\varepsilon^x(y) e^{\frac{1}{2\varepsilon}\|x-y\|^2}}{q(y)} \,\mathrm{d}\pi_{\varepsilon,n}(x,y) &= \frac{1}{2\varepsilon} W_2^2(P,Q) - \log(\Lambda_\varepsilon) - \frac{1}{2}(\mathcal{H}(Q) + \mathcal{H}(P)) \\ &\quad + \int (g_0/\varepsilon - \log(q))(\mathrm{d}Q_n - \mathrm{d}Q) \\ &\quad - \int \log(Z_\varepsilon(x) J(\nabla^2 \varphi_0^*(x^*))) \,\mathrm{d}P(x)\,.\end{aligned}$$

We therefore obtain

$$\sup_{h:\Omega\to\mathbb{R}^d} \int j_h \,\mathrm{d}\pi_{\varepsilon,n} - \iint e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)} \,\mathrm{d}Q_n(y) \,\mathrm{d}P(x) + 1 \\ \leq \varepsilon^{-1}\Big(\mathrm{OT}_\varepsilon(P,Q_n) - \frac{1}{2} W_2^2(P,Q) + \varepsilon \log(\Lambda_\varepsilon) + \frac{\varepsilon}{2}(\mathcal{H}(Q) + \mathcal{H}(P)) \Big) + \Delta_1\,,$$

where $\Delta_1 := \int (g_0/\varepsilon - \log(q))(\mathrm{d}Q - \mathrm{d}Q_n) + \int \log(Z_\varepsilon(x) J(\nabla^2 \varphi_0^*(x^*))) \,\mathrm{d}P(x)$. Applying Theorem 2.2, we may further bound

$$\sup_{h:\Omega\to\mathbb{R}^d} \int j_h \,\mathrm{d}\pi_{\varepsilon,n} - \iint e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)} \,\mathrm{d}Q_n(y) \,\mathrm{d}P(x) + 1 \leq \frac{\varepsilon}{8} I_0 + \Delta_1 + \Delta_2\,,$$

where $\Delta_2 := \varepsilon^{-1}(\mathrm{OT}_\varepsilon(P,Q_n) - \mathrm{OT}_\varepsilon(P,Q))$. Now we turn our attention to the second term on the

left side. Since

$$\iint e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)} \,\mathrm{d}Q(y)\,\mathrm{d}P(x) = \iint_{\mathrm{supp}(Q)} e^{j_h(x,y)} q_\varepsilon^x(y)\,\mathrm{d}y\,\mathrm{d}P(x) \\ \leq \iint e^{j_h(x,y)} q_\varepsilon^x(y)\,\mathrm{d}y\,\mathrm{d}P(x)\,,$$

we have

$$\sup_{h:\Omega\to\mathbb{R}^d} \int j_h \,\mathrm{d}\pi_{\varepsilon,n} - \iint (e^{j_h(x,y)} - 1) q_\varepsilon^x(y)\,\mathrm{d}y\,\mathrm{d}P(x) \leq \frac{\varepsilon}{8} I_0 + \Delta_1 + \Delta_2 + \Delta_3\,,$$

where

$$\Delta_3 := \sup_{h:\Omega\to\mathbb{R}^d} \iint e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)} \,\mathrm{d}P(x)(\mathrm{d}Q_n - \mathrm{d}Q)(y)$$

and where we have used the fact that $q_\varepsilon^x(y)$ is a probability density.

It therefore remains only to show that

$$\mathbb{E}[\Delta_1 + \Delta_2 + \Delta_3] \lesssim \varepsilon^{(\bar{\alpha}-1)/2} + \varepsilon^{-d/2} \log(n) n^{-1/2}\,.$$

First, a Laplace expansion (Corollary A.3) implies

$$\mathbb{E}\Delta_1 = \int \log(Z_\varepsilon(x) J(\nabla^2 \varphi_0^*(x^*)))\,\mathrm{d}P(x) \lesssim \varepsilon^{(\bar{\alpha}-1)/2}$$

Second, known results on the finite-sample convergence of the Sinkhorn divergence (Corollary A.10) yield

$$\mathbb{E}\Delta_2 \lesssim (\varepsilon^{-1} + \varepsilon^{-d/2}) \log(n) n^{-1/2}\,,$$

It therefore remains to bound Δ_3, which an empirical process theory argument (Proposition A.5) shows

$$\mathbb{E}\Delta_3 \lesssim \varepsilon^{-d/2} n^{-1/2}$$

as long as $a \in [L\varepsilon, 1]$.

We obtain that

$$\mathbb{E}[\Delta_1 + \Delta_2 + \Delta_3] \lesssim \varepsilon^{(\bar{\alpha}-1)/2} + (\varepsilon^{-1} + \varepsilon^{-d/2}) \log(n) n^{-1/2} + \varepsilon^{-d/2} n^{-1/2},$$

and since $\varepsilon \leq 1$, we obtain the bound

$$\mathbb{E}[\Delta_1 + \Delta_2 + \Delta_3] \lesssim \varepsilon^{(\bar{\alpha}-1)/2} + \varepsilon^{-d/2} n^{-1/2} \log(n),$$

as desired. □

To exploit Proposition 2.11, we show that we can choose a function h for which the left side of the above expression scales like $\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)}$.

We first establish three lemmas, whose proofs are deferred.

Lemma 2.12. *Fix $x \in \operatorname{supp}(P)$, and write $\bar{y}^x = \int y q^x_\varepsilon(y)\, dy$. There exists a positive constant C, independent of x, such that for all $\varepsilon \in (0,1)$ and $\|v\|_2 \leq 1$,*

$$\int e^{(v^\top (y - \bar{y}^x))^2/(C\varepsilon)} q^x_\varepsilon(y)\, dy \leq 2.$$

In probabilistic language, Lemma 2.12 implies that if Y^x is a random variable with density q^x_ε, then $\varepsilon^{-1/2}(Y^x - \mathbb{E}Y^x)$ is subgaussian (Vershynin, 2018). By applying standard moment bounds for subgaussian random variables, we then arrive at the following result.

Lemma 2.13. *There exists a positive constant C such that if $a \geq C\varepsilon$, then for any $h : \mathbb{R}^d \to \mathbb{R}^d$ we have*

$$\iint e^{h(x)^\top (y - T_0(x)) - a\|h(x)\|^2} q^x_\varepsilon(y)\, dy\, dP(x) \leq \int e^{\frac{1}{4\varepsilon}\|\bar{y}^x - T_0(x)\|^2}\, dP(x).$$

Finally, we show by an application of Laplace's method that $\bar{y}^x$ is close to $T_0(x)$.

Lemma 2.14. *Assume **(E1)** to **(E3)**. For all $x \in \mathrm{supp}(P)$,*

$$\|\bar{y}^x - T_0(x)\|^2 \lesssim \varepsilon^{\alpha \wedge 2}.$$

With these lemmas in hand, we can complete the proof.

Proof of Theorem 2.6. We may assume $\varepsilon_0 \le 1$. Since $e^t - 1 \le 2t$ for $t \in [0,1]$, Lemma 2.14 implies that as long as ε_0 is sufficiently small, for $\varepsilon \le \varepsilon_0$,

$$e^{\frac{1}{4\varepsilon}\|\bar{y}^x - T_0(x)\|^2} - 1 \lesssim \varepsilon^{(\alpha-1)\wedge 1} \le \varepsilon^{(\bar{\alpha}-1)/2},$$

where the last inequality holds for $\alpha \ge 1$ and $\varepsilon \le 1$. Combining this fact with Lemma 2.13, we obtain that for any $h : \mathbb{R}^d \to \mathbb{R}^d$ and $a \ge C\varepsilon$,

$$\iint (e^{h(x)^\top (y - T_0(x)) - a\|h(x)\|^2} - 1) q_\varepsilon^x(y)\, \mathrm{d}y\, \mathrm{d}P(x) \lesssim \varepsilon^{(\bar{\alpha}-1)/2}.$$

For a sufficiently small constant ε_0, the interval $[C\varepsilon, 1]$ is non-empty for $\varepsilon \le \varepsilon_0$, so combining this fact with Proposition 2.11 yields that for $a \in [C\varepsilon, 1]$ and $\varepsilon \le \varepsilon_0$,

$$\mathbb{E} \sup_{h:\Omega \to \mathbb{R}^d} \iint \left(h(x)^\top (y - T_0(x)) - a\|h(x)\|^2\right) \mathrm{d}\pi_{\varepsilon,n}(x,y) \lesssim \varepsilon I_0 + \varepsilon^{(\bar{\alpha}-1)/2} + \varepsilon^{-d/2} \log(n) n^{-1/2}. \tag{2.15}$$

If we pick $h(x) = \frac{1}{2a}(T_{\varepsilon,n}(x) - T_0(x))$, the integral on the left side equals

$$\frac{1}{2a}\mathbb{E} \iint \left((T_{\varepsilon,n}(x) - T_0(x))^\top (y - T_0(x)) - \frac{1}{2}\|T_{\varepsilon,n}(x) - T_0(x)\|^2\right) \mathrm{d}\pi_{\varepsilon,n}(x,y) \tag{2.16}$$

By definition, $T_{\varepsilon,n}(x) = \int y\, \mathrm{d}\pi_{\varepsilon,n}^x(y)$, so disintegrating $\pi_{\varepsilon,n}(x,y)$ and recalling that the first marginal

of $\pi_{\varepsilon,n}$ is P yields

$$\iint \left(T_{\varepsilon,n}(x) - T_0(x) \right)^\top (y - T_0(x)) - \frac{1}{2} \|T_{\varepsilon,n}(x) - T_0(x)\|^2 \Big) \, \mathrm{d}\pi_{\varepsilon,n}(x,y) \\ = \int \frac{1}{2} \|T_{\varepsilon,n}(x) - T_0(x)\|^2 \, \mathrm{d}P(x) = \frac{1}{2} \|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} .$$

Combining this with (2.15) and (2.16) and picking $a = C\varepsilon$ yields

$$\mathbb{E}\|T_{\varepsilon,n} - T_0\|^2_{L^2(P)} \lesssim \varepsilon^2 I_0 + \varepsilon^{(\bar{\alpha}+1)/2} + \varepsilon^{1-d/2} \log(n) n^{-1/2} ,$$

as desired. □

2.2.2 Two-sample estimates

We now turn our attention to the two-sample case. Let $\pi_{\varepsilon,(n,n)}$ be the optimal entropic plan between P_n and Q_n and $(f_{\varepsilon,(n,n)}, g_{\varepsilon,(n,n)})$ the corresponding entropic potentials. We aim to show that

$$\mathbb{E}\|T_{\varepsilon,(n,n)} - T_{\varepsilon,n}\|^2_{L^2(P)} \lesssim (\varepsilon^{-1} + \varepsilon^{-d/2}) \log(n) n^{-1/2} .$$

As in Section 2.2.1, we proceed via duality arguments, but our task is considerably simplified by the fact that the measure Q_n remains fixed in passing from $T_{\varepsilon,(n,n)}$ to $T_{\varepsilon,n}$. Let us write

$$\gamma(x,y) = e^{\frac{1}{\varepsilon}(f_{\varepsilon,(n,n)}(x) + g_{\varepsilon,(n,n)}(y) - \frac{1}{2}\|x-y\|^2)} = \frac{e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(y) - \frac{1}{2}\|x-y\|^2)}}{\frac{1}{n}\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}$$

for the $P_n \otimes Q_n$ density of $\pi_{\varepsilon,(n,n)}$, where the second equality holds $P_n \otimes Q_n$ almost everywhere and furnishes an extension of γ to all $x \in \mathbb{R}^d$.

We employ the following analogue of Proposition 2.11, which does not require the full force of assumptions **(E1)** to **(E3)**.

Proposition 2.15. *The support of P and Q lies in Ω, then*

$$\mathbb{E}\Big\{ \sup_{\chi:\Omega\times\Omega\to\mathbb{R}} \iint \chi(x,y)\,\mathrm{d}\pi_{\varepsilon,n}(x,y) - \iint (e^{\chi(x,y)}-1)\gamma(x,y)\,\mathrm{d}P(x)\,\mathrm{d}Q_n(y)\Big\} \\ \lesssim (\varepsilon^{-1}+\varepsilon^{-d/2})\log(n)n^{-1/2}\,,$$

where the supremum is taken over all $\chi \in L^1(\pi_{\varepsilon,n})$.

The proof of Theorem 2.10 is now straightforward.

Proof. As in the proof of Theorem 2.6, consider

$$\chi(x,y) = h(x)^\top (y - T_{\varepsilon,(n,n)}(x)) - a\|h(x)\|^2$$

for h and a to be specified. By definition of $T_{\varepsilon,(n,n)}$, we have

$$\int h(x)^\top (y - T_{\varepsilon,(n,n)}(x))\gamma(x,y)\,\mathrm{d}Q_n(y) = h(x)^\top \left(\int y\gamma(x,y)\,\mathrm{d}Q_n(y) - T_{\varepsilon,(n,n)}(x)\right) \\ = h(x)^\top \left(\frac{\frac{1}{n}\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i)-c(x,Y_i))}}{\frac{1}{n}\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i)-c(x,Y_i))}} - T_{\varepsilon,(n,n)}(x)\right) = 0$$

for all $x \in \mathbb{R}^d$. Moreover, since Ω is compact, by the Cauchy-Schwarz inequality, there exists a constant C such that

$$|h(x)^\top (y - T_{\varepsilon,(n,n)}(x))| \le C\|h(x)\| \quad \forall y \in \Omega\,.$$

Hoeffding's inequality therefore implies that if $a \ge C^2/2$, then this choice of χ satisfies

$$\iint (e^{\chi(x,y)}-1)\gamma(x,y)\,\mathrm{d}Q_n(y)\,\mathrm{d}P(x) \le 0\,.$$

Choosing $h(x) = \frac{1}{2a}(T_{\varepsilon,n}(x) - T_{\varepsilon,(n,n)}(x))$, we conclude as in the proof of Theorem 2.6 that for

$\varepsilon \leq 1$,

$$\frac{1}{4a}\mathbb{E}\|T_{\varepsilon,n} - T_{\varepsilon,(n,n)}\|^2_{L^2(P)} \lesssim (\varepsilon^{-1} + \varepsilon^{-d/2})\log(n)n^{-1/2} \lesssim \varepsilon^{-d/2}\log(n)n^{-1/2},$$

and picking a to be a sufficiently large constant yields the claim. □

2.3 Adaptive estimation

In Theorems 2.5 and 2.6, the optimal choice of the regularization parameter ε depends on n, d, and α. Although the number of samples and dimension are obviously known to the practitioner, the smoothness of the transport map is often not known *a priori*. However, Lepski's method (see Birgé, 2001) can be used to obtain a data-driven method of choosing ε, which gives rise to an estimator that adapts to the unknown smoothness parameter α.

For notational convenience, for any $\alpha > 1$, let $s := \alpha + 1$ be the smoothness of the conjugate Brenier potential φ_0^*. We assume that $s \in [2 + \iota, 4]$ for some $\iota > 0$ sufficiently small and fixed. Let $\mathcal{S}$ be the following discrete subset

$$\mathcal{S} := \{2 + \iota = s_{\min} = s_1 < s_2 < \cdots < s_N = s_{\max} = 4\},$$

where $s_j - s_{j-1} \asymp (\log n)^{-1}$, and set

$$\varepsilon_s = (n/\log n)^{-1/2(d+s)}, \quad \psi_n(s) = (\varepsilon_s)^s = (n/\log n)^{-s/2(d+s)}. \tag{2.17}$$

To calibrate our choice of ε, we rely on sample splitting. Let $\mathbb{D} := \{(X_i, Y_i)\}_{i=1}^n$ denote our initial dataset, and let $\mathbb{D}'$ denote an independent copy of $\mathbb{D}$. Denote by P_n' and Q_n' the empirical measures arising from $\mathbb{D}'$. Our choice of smoothness parameter is given by the following rule:

$$\hat{s} := \max\{s \in \mathcal{S} \,:\, \|\hat{T}_{\varepsilon_s} - \hat{T}_{\varepsilon_{s'}}\|^2_{L^2(P_n')} \leq K\psi_n(s'), \forall\, s' \leq s, s' \in \mathcal{S}\}, \tag{2.18}$$

for a positive constant K. The following theorem shows that choosing $\varepsilon = \varepsilon_{\hat{s}}$ gives rise to an adaptive estimator.

Theorem 2.16. *Suppose **(E1)** to **(E3)** holds, with $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$, resulting in $\mathbb{D} = \{(X_i, Y_i)\}_{i=1}^{\lfloor n/2 \rfloor}$ and a hold-out set $\mathbb{D}'$. Suppose $\hat{s}$ is chosen according to (2.18) for K sufficiently large, with $\varepsilon = \varepsilon_{\hat{s}}$ chosen as in (2.17). The resulting estimator $\hat{T}_{\varepsilon_{\hat{s}}}$ exhibits a risk in $L^2(P)$ that matches Theorem 2.5 up to log factors.*

The proof of Theorem 2.16 uses standard ideas and is deferred to Appendix A.5.

2.4 Computational aspects

Our reason for studying the entropic map as an optimal transport map estimator arises from its strong computational benefits, which are a consequence of the efficiency of Sinkhorn's algorithm for entropic optimal transport (see Peyré and Cuturi, 2019). In this section, we compare the computational complexity of the entropic map to the estimators of Hütter and Rigollet (2021), Deb et al. (2021), and Manole et al. (2024a) in the two-sample setting. Finally, we perform several experiments that demonstrate the computational advantages of our procedure. Throughout this section, we use $\tilde{O}$ to hide poly-logarithmic factors in the sample size n.

2.4.1 Estimator complexities from prior work

We first describe the wavelet-based estimator proposed by Hütter and Rigollet (2021). Recall that this estimator is minimax optimal for all $\alpha > 1$. The implementation of this estimator requires various discretization and approximation schemes. The authors of that work use a numerical implementation of the Daubechies wavelets to approximate the optimal Brenier potential, and then compute its convex conjugate by means of a discrete Legendre transform on a discrete grid. The gradient of the resulting potential is then obtained using finite differences, and this is extended

to data outside the grid by linear interpolation. Though computing this estimator takes time that scales only linearly in the sample size n, the main bottleneck of this approach from a computational standpoint is the computation of the Legendre transform on the grid, which requires at least cN^d operations, where N denotes the resolution of the grid. Since this resolution needs to be chosen fine enough to be negligible, the exponential dependence in d makes this approach prohibitively expensive in most applications.

Another estimator recently analyzed in the literature by Manole et al. (2024a) is the "1-Nearest Neighbor" estimator, which we denote by $\hat{T}^{\mathrm{1NN}}_{(n,n)}$, which achieves the minimax rate when T_0 is bi-Lipschitz (i.e., $\alpha = 1$ and **(E3)** is satisfied) over a compact domain Ω. The estimator takes the form

$$\hat{T}^{\mathrm{1NN}}_{(n,n)}(x) = \sum_{i,j=1}^{n} (n\hat{\pi}_{ij})\mathbf{1}_{V_i}(x)Y_j, \tag{2.19}$$

where $\mathbf{1}$ is the indicator function for a set, and $(V_i)_{i=1}^n$ are the Voronoi regions generated by $(X_i)_{i=1}^n$, i.e.,

$$V_i = \{x \in \Omega \ : \ \|x - X_i\| \leq \|x - X_j\|, \ \forall j \neq i\},$$

and $\hat{\pi}$ is the optimal coupling that solves (1.15) when the measures are the empirical measures P_n and Q_n. Solving for $\hat{\pi}$ can be done through the Hungarian algorithm, and has time complexity $O(n^3)$. However, unlike the wavelet estimator described above, computing this estimator does not require constructing a grid whose size scales exponentially with dimension.

For the $\alpha > 1$ case, both Manole et al. (2024a) and Deb et al. (2021) propose estimators based on density estimation. For these approaches, the idea is to construct nonparametric density estimates of the measures P and Q, resample points from these densities, and finally perform the appropriate matching using the Hungarian algorithm once again. Though tractable in low dimensions, this approach is limited by the difficulty of sampling from nonparametric density estimates, which typically requires time scaling exponentially in the dimension d.

In short, prior estimators proposed in the literature either have runtime scaling exponentially

in d (in the case of the wavelet estimator or estimators based on nonparametric density estimation) or cubicly in n (in the case of the 1NN estimator). By contrast, in the following section, we show that our estimator can be computed in nearly $O(n^2)$ time.

2.4.2 Computational complexity of the entropic map

We now turn to the computational analysis of our estimator, which has the closed-form representation

$$\hat{T}_{\varepsilon,(n,n)}(x) = \frac{\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}{\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}. \tag{2.20}$$

The computational burden of our estimator falls on computing the optimal entropic potential evaluated at the data $g_{\varepsilon,(n,n)}(Y_i)$. Indeed, once we have this potential, it is clear that the remainder of (2.20) can be computed in $O(n)$ time.

The leading approach to compute optimal entropic potentials in practice is *Sinkhorn's algorithm* (Peyré and Cuturi, 2019; Sinkhorn, 1967), an alternating minimization algorithm that computes approximations of the entropic potentials by iteratively updating f and g so that they satisfy one of the two dual optimality conditions given in (1.27) and (1.28). Explicitly, defining $f^{(0)} = 0$, Sinkhorn's algorithm performs the updates

$$g^{(k)}(y) = -\varepsilon \log \frac{1}{n} \sum_{i=1}^n e^{\frac{1}{\varepsilon}(f^{(k)}(X_i) - \frac{1}{2}\|X_i-y\|^2)}$$
$$f^{(k+1)}(x) = -\varepsilon \log \frac{1}{n} \sum_{j=1}^n e^{\frac{1}{\varepsilon}(g^{(k)}(Y_j) - \frac{1}{2}\|x-Y_j\|^2)}.$$

until termination. Since it is only necessary to compute $f^{(k)}$ and $g^{(k)}$ on the support of P_n and Q_n, respectively, each iteration can be implemented in $O(n^2)$ time.

Note that this update rule guarantees that

$$\int e^{\frac{1}{\varepsilon}(f^{(k)}(x) + g^{(k)}(y) - \frac{1}{2}\|x-y\|^2)} \, \mathrm{d}P_n(x) = 1$$

for all y at each iteration. By contrast, the other optimality condition (1.27) is *not* satisfied at each iteration, though Sinkhorn (1967) showed that

$$\int e^{\frac{1}{\varepsilon}(f^{(k)}(x)+g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)}\,\mathrm{d}Q_n(y) \to 1$$

as $k \to \infty$, and therefore that the iterates of Sinkhorn's algorithm converge to optimal entropic potentials.

To analyze the running time of our estimator, we will leverage recent analyses of the convergence rate of Sinkhorn's algorithm (Altschuler et al., 2017; Cuturi, 2013; Dvurechensky et al., 2018) to explicitly quantify the error incurred by terminating after a finite number of steps. For $k \geq 0$, we consider the entropic map estimator obtained after k iterates of Sinkhorn's algorithm:

$$T^{(k)}(x) = \frac{\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}(g^{(k)}(Y_i)-\frac{1}{2}\|x-Y_i\|^2)}}{\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g^{(k)}(Y_i)-\frac{1}{2}\|x-Y_i\|^2)}}\,. \tag{2.21}$$

Despite the fact that $g^{(k)}$ is *not* an entropic potential for the original problem, the following theorem shows that $T^{(k)}$ is nevertheless an acceptable estimator if k is sufficiently large.

Theorem 2.17. *Suppose assumptions **(E1)** to **(E3)** hold, and we choose ε as in Theorem 2.5. Then for any $k \gtrsim n^{7/(d+\bar{\alpha}+1)} \log n$,*

$$\mathbb{E}\|T^{(k)} - T_0\|^2_{L^2(P)} \lesssim (1 + I_0(P,Q))n^{-\frac{(\bar{\alpha}+1)}{2(d+\bar{\alpha}+1)}} \log n\,,$$

where $\bar{\alpha} = 3 \wedge \alpha$. In particular, an estimator achieving the same rate as the estimator in Theorem 2.5 can be computed in $\tilde{O}(n^{2+7/(d+\bar{\alpha}+1)}) = n^{2+o_d(1)}$ time.

Proof. We begin by decomposing the error and applying Theorem 2.6:

$$\begin{aligned}\mathbb{E}\|T^{(k)} - T_0\|_{L^2(P)}^2 &\lesssim \mathbb{E}\|T^{(k)} - T_{\varepsilon,n}\|_{L^2(P)}^2 + \mathbb{E}\|T_{\varepsilon,n} - T_0\|_{L^2(P)}^2 \\ &\lesssim \mathbb{E}\|T^{(k)} - T_{\varepsilon,n}\|_{L^2(P)}^2 + \varepsilon^{1-d/2}\log(n)n^{-1/2} + \varepsilon^{(\bar{\alpha}+1)/2} + \varepsilon^2 I_0(P,Q)\,.\end{aligned}$$

We proceed almost exactly as in Theorem 2.10, and consider

$$\chi(x,y) = h(x)^\top \left(y - T^{(k)}(x)\right) - a\|h(x)\|^2,$$

for $h : \mathbb{R}^d \to \mathbb{R}^d$ and a to be specified. For $x \in \mathbb{R}^d$, $y \in \operatorname{supp}(Q_n)$, define

$$\tilde{\gamma}(x,y) = \frac{\exp\left(\frac{1}{\varepsilon}(g^{(k)}(y) - \frac{1}{2}\|x-y\|^2)\right)}{\frac{1}{n}\sum_{i=1}^n \exp\left(\frac{1}{\varepsilon}(g^{(k)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)\right)}\,. \tag{2.22}$$

By construction, $\int \tilde{\gamma}(x,y)\,\mathrm{d}Q_n(y) = 1$ for all $x \in \mathbb{R}^d$, and $T^{(k)}(x) = \int y\tilde{\gamma}(x,y)\,\mathrm{d}Q_n(y)$. Therefore, for any $h : \mathbb{R}^d \to \mathbb{R}^d$,

$$\int h(x)^\top \left(y - T^{(k)}(x)\right)\tilde{\gamma}(x,y)\,\mathrm{d}Q_n(y) = 0$$

for all $x \in \mathbb{R}^d$. Moreover, since Ω is compact, there exists a constant C such that

$$|h(x)^\top(y - T^{(k)}(x))| \le C\|h(x)\| \quad \forall x, y \in \Omega\,.$$

Hoeffding's inequality therefore implies that for a sufficiently large, this choice of χ satisfies

$$\iint (e^{\chi(x,y)} - 1)\tilde{\gamma}(x,y)\,\mathrm{d}Q_n\,\mathrm{d}P(x) \le 0\,.$$

Now, define a probability measure $\tilde{P}$ with the same support as P_n by setting

$$\frac{\mathrm{d}\tilde{P}(x)}{\mathrm{d}P_n(x)} = \int e^{\frac{1}{\varepsilon}(f^{(k)}(x)+g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)}\,\mathrm{d}Q_n(y)\,, \tag{2.23}$$

and let

$$f^{(k+1)}(x) = -\varepsilon \log \frac{1}{n}\sum_{i=1}^{n} \exp\left(\varepsilon^{-1}(g^{(k)}(Y_i) - \tfrac{1}{2}\|x-Y_i\|^2)\right)\,. \tag{2.24}$$

We claim that $\tilde{\gamma}(x,y) = \exp(\frac{1}{\varepsilon}(f^{(k+1)}(x)+g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)$ is the $\tilde{P}\otimes Q_n$ density of the optimal entropic plan between $\tilde{P}$ and Q_n. We have already observed that $\int \tilde{\gamma}(x,y)\,\mathrm{d}Q_n(y) = 1$ for all $x\in\mathbb{R}^d$ by construction, so it suffices to note that for all $y\in\mathrm{supp}(Q_n)$,

$$\begin{aligned}
\int \tilde{\gamma}(x,y)\,\mathrm{d}\tilde{P}(x) &= \int \frac{e^{\varepsilon^{-1}(g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)}}{\int e^{\varepsilon^{-1}(g^{(k)}(y')-\frac{1}{2}\|x-y'\|^2)}\,\mathrm{d}Q_n(y')}\,\mathrm{d}\tilde{P}(x)\\
&= \int \frac{e^{\varepsilon^{-1}(f^{(k)}(x)+g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)}}{\int e^{\varepsilon^{-1}(f^{(k)}(x)+g^{(k)}(y')-\frac{1}{2}\|x-y'\|^2)}\,\mathrm{d}Q_n(y')}\,\mathrm{d}\tilde{P}(x)\\
&= \int e^{\varepsilon^{-1}(f^{(k)}(x)+g^{(k)}(y)-c(x,y))}\,\mathrm{d}P_n(x) = 1\,.
\end{aligned}$$

Therefore $(f^{(k+1)}, g^{(k)})$ satisfy (1.28), so $\tilde{\gamma}$ is indeed the $\tilde{P}\otimes Q_n$ density of the optimal entropic plan between the two measures.

Applying Proposition A.4, we obtain for any $\varepsilon\le 1$

$$\mathbb{E}\sup_{h:\mathbb{R}^d\to\mathbb{R}^d} \iint h(x)^\top\left(y - T^{(k)}(x)\right) - a\|h(x)\|^2\,\mathrm{d}\pi_{\varepsilon,n} \lesssim \varepsilon^{-1}\delta + \varepsilon^{-d/2}\log(n)n^{-1/2}\,, \tag{2.25}$$

where $\delta := \mathrm{TV}(\tilde{P}, P_n)$. Choosing $h(x) = \frac{1}{2a}\left(T_{\varepsilon,n}(x) - T^{(k)}(x)\right)$, we conclude as in Theorem 2.10, resulting in

$$\mathbb{E}\|T^{(k)} - T_{\varepsilon,n}\|^2_{L^2(P)} \lesssim \varepsilon^{-1}\delta + \varepsilon^{-d/2}\log(n)n^{-1/2}\,.$$

All together, we have

$$\mathbb{E}\|T^{(k)} - T_0\|^2_{L^2(P)} \lesssim \varepsilon^{-1}\delta + \varepsilon^{-d/2}\log(n)n^{-1/2} + \varepsilon^{(\bar{\alpha}+1)/2} + \varepsilon^2 I_0(P,Q)\,.$$

The first term will be negligible if $\delta \lesssim \varepsilon^3$.

By definition, $\tilde{P}$ is the first marginal of the joint distribution with density

$$e^{\frac{1}{\varepsilon}(f^{(k)}(x)+g^{(k)}(y)-\frac{1}{2}\|x-y\|^2)}\,.$$

with respect to $P_n \otimes Q_n$. By Altschuler et al. (2017, Theorem 2), if k satisfies

$$k \gtrsim \delta^{-2}\log(n \cdot \max_{i,j} e^{\frac{1}{2\varepsilon}\|x_i-y_j\|^2}) \gtrsim \delta^{-2}\varepsilon^{-1}\log n\,,$$

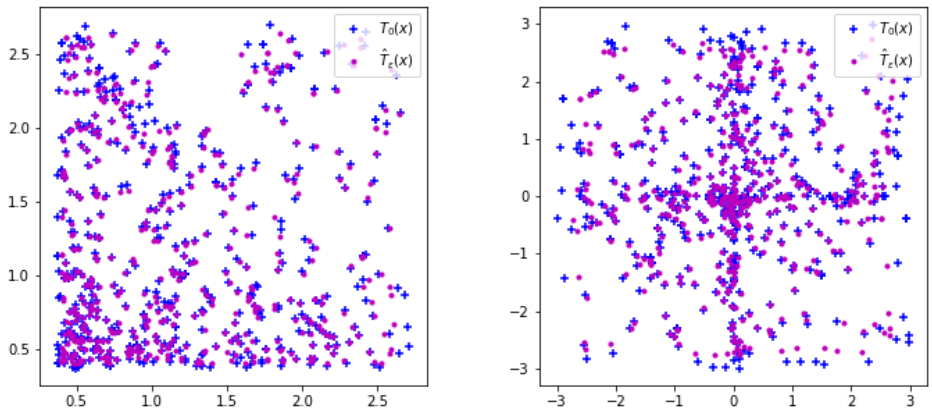
then $\mathrm{TV}(\tilde{P}, P_n) \leq \delta$. Choosing $\delta = \varepsilon^3 \asymp n^{-3/(d+\bar{\alpha}+1)}$ yields the claim. □

Remark 2.18. A surprising feature of Theorem 2.17 is that the necessary number of iterations decreases with the dimension d. This reflects the fact that when d is large, the optimal choice of ε is also larger, and it is well established both theoretically and empirically that the performance of Sinkhorn's algorithm improves considerably as ε increases (Altschuler et al., 2017; Cuturi, 2013).

2.4.3 Empirical performance

We test two implementations of Sinkhorn's algorithm, one from the Python Optimal Transport (POT) library (Flamary et al., 2021), and an implementation that uses the KeOps library optimized for GPUs. Both implementations employ log-domain stabilization to avoid numerical overflow issues arising from the small choice of ε.

For simplicity, we employ the same experimental setup as Hütter and Rigollet (2021). We generate i.i.d. samples from a source distribution P, which we always take to be $[-1,1]^d$, and from



(a) $T_0(x) = \exp(x)$ coordinate-wise **(b)** $T_0(x) = 3x^2\text{sign}(x)$ coordinate-wise

Figure 2.1: Visualization of $\hat{T}_\varepsilon$ and $T_0(x)$ in 2 dimensions.

a target distribution $Q = (T_0)_\sharp P$, where we define $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ to be an optimal transport map obtained by applying a monotone scalar function coordinate-wise.[1]

In Figure 2.1, we visualize the output of our estimator in $d = 2$. The figures depict the effect of evaluating the estimator $\hat{T}_\varepsilon$ and the true map T_0 on additional test points $X'_1, \ldots, X'_m$ drawn i.i.d. from P.

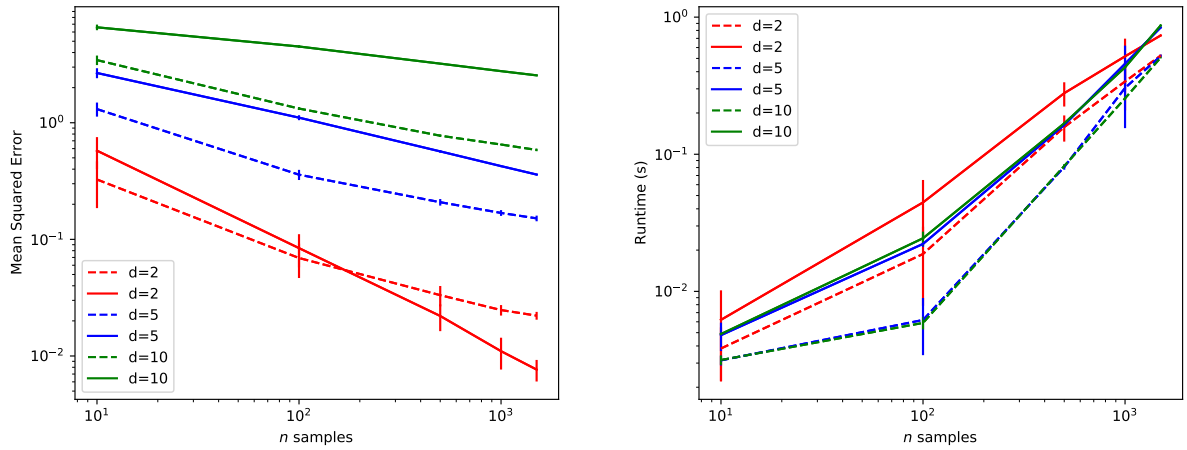
2.4.3.1 Comparison to a tractable minimax estimator

Among the previously discussed estimators, the 1-Nearest Neighbor estimator analyzed in Manole et al. (2024a) is the most tractable, and the only one remotely comparable to our method. As discussed in Section 2.4.1, this approach uses the Hungarian algorithm which has a runtime of $O(n^3)$. However, since it is not parallelizable, we compare its performance to the non-parallel CPU implementation of Sinkhorn's algorithm from the POT library.

We perform a simple experiment comparing our approach to theirs: let $P = [-1, 1]^d$ and let $T_0(x) = \exp(x)$, acting coordinate-wise. We vary d and n, and track runtime performance of

[1]Note that any component-wise monotone function is the gradient of a convex function.

both estimators, as well as the Mean Squared Error (MSE) of the map estimate[2], averaged over 20 runs. For our estimator, we choose ε as suggested by Theorem 2.5. We observe that in $d = 2$,



(a) MSE comparison

(b) Runtime comparison

Figure 2.2: Dashed lines are our estimator, solid lines are $\hat{T}^{1\text{NN}}$, and $T_0(x) = \exp(x)$

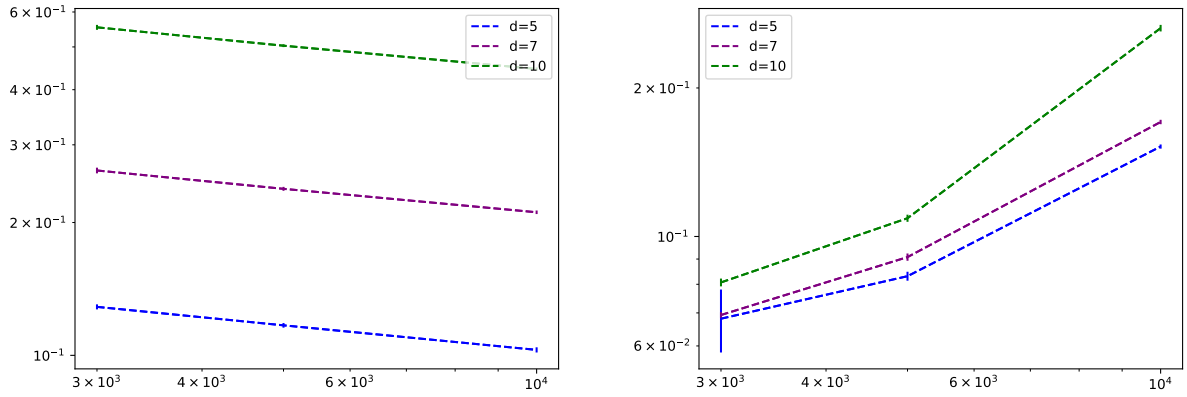
the MSE of the two estimators are comparable, though our error deteriorates for large n, which reflects our slightly sub-optimal estimation rate. However, as d increases to moderate dimensions, our estimator consistently outperforms $\hat{T}^{1\text{NN}}$ in both MSE and runtime with the choice of ε in Theorem 2.5. For both estimators, the CPU runtime begins to become significant (on the order of seconds) when n exceeds 1500.

2.4.3.2 Parallel estimation on massive data sets

Figure 2.2 makes clear that computation of both estimators slows for $n \gg 10^3$ when implemented on a CPU. However, Sinkhorn's algorithm can be easily parallelized. Unlike the 1-Nearest Neighbor estimator—and all other transport map estimators of which we are aware—our proposal therefore runs extremely efficiently on GPUs. We again average performance over 20 runs, and choose ε as in the previous example, with T_0 again as the exponential map (coordinate-wise). We

[2]We calculate MSE by performing Monte Carlo integration over the space $[-1, 1]^d$.

see in Figure 2.3 that even when $n = 10^4$ and $d = 10$, it takes roughly a third of a second to perform the optimization.



(a) MSE comparison

(b) Runtime comparison

Figure 2.3: Performance of a parallel implementation of our estimator on large data sets.

3 Minimax estimation of discontinuous optimal transport maps: The semidiscrete case

3.1 Introduction

In this chapter, we revisit the task of estimating optimal transport maps, i.e., minimizers of the following non-convex, infinite-dimensional optimization problem[1]

$$\nabla\varphi_0 = \operatorname*{argmin}_{T\in\mathcal{T}(P,Q)} \int \|x - T(x)\|^2 \,\mathrm{d}P(x)\,,$$

on the basis of fixed data $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$.

Recall that the first finite-sample analysis of this problem was performed by Hütter and Rigollet (2021), who proposed an estimator for $\nabla\varphi_0$ under the assumption that φ_0 is $s+1$-times continuously differentiable, for $s > 1$. They showed that a wavelet-based estimator $\hat{\varphi}_{\mathrm{W}}$ satisfies

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathrm{W}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim n^{-\frac{2s}{2s+d-2}} \log^2(n)\,,$$

[1] $\mathcal{T}(P,Q)$ is the set of transport maps from P to Q.

and that this rate is minimax optimal up to logarithmic factors. Their analysis requires that P and Q have bounded densities with compact support $\Omega \subseteq \mathbb{R}^d$, and that φ_0 be both strongly convex and smooth. Implementing the estimator $\hat{\varphi}_{\mathrm{W}}$ is computationally challenging even in moderate dimensions, and is practically infeasible for $d > 3$. Follow up works (Deb et al., 2021; Divol et al., 2022; Manole et al., 2024a; Pooladian and Niles-Weed, 2021; Vacher et al., 2024) have proposed alternative estimators which improve upon $\hat{\varphi}_{\mathrm{W}}$ either in computational efficiency or in the generality in which they apply. Though these subsequent works go significantly beyond the setting considered by Hütter and Rigollet (2021), none have eliminated the crucial assumption that φ_0 is smooth, i.e., that the transport map $\nabla\varphi_0$ is Lipschitz.

There are two estimators proposed in this line of work that are particularly practical and worth highlighting. Manole et al. (2024a) study the 1-Nearest Neighbor estimator $\hat{T}_{1\mathrm{NN}}$. This estimator is obtained by solving the empirical optimal transport problem between the samples, which is then extended to a function defined on $\mathbb{R}^d$ using a projection scheme; see Section 3.3 for more details. Given n samples from the source and target measures in $\mathbb{R}^d$, $\hat{T}_{1\mathrm{NN}}$ has a runtime of $O(n^3)$ via the Hungarian Algorithm (see Peyré and Cuturi, 2019, Chapter 3), and, for $d \geq 5$, achieves the rate

$$\mathbb{E}\|\hat{T}_{1\mathrm{NN}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim n^{-\frac{2}{d}} \tag{3.1}$$

whenever the optimal Brenier potential φ_0 is smooth and strongly convex, and under mild regularity conditions on P. Recall that in the previous chapter, we conducted a statistical analysis of an estimator originally proposed by Seguy et al. (2018) based on entropic optimal transport. The efficiency of Sinkhorn's algorithm for large-scale problems (Cuturi, 2013; Peyré and Cuturi, 2019) makes this estimator attractive from a computational perspective, and we also gave statistical guarantees, though these fall short of being minimax-optimal.

Despite this progress, none of the aforementioned results can be applied in situations where $\nabla\varphi_0$ is not Lipschitz. And in practice, even requiring the *continuity* of the transport map can be

far too stringent. It is indeed too much to hope for that an underlying data distribution (e.g. over the space of images) has one single connected component; this is supported by recent work that stipulates that the underlying data distribution is the union of *disjoint* manifolds of varying intrinsic dimension (Brown et al., 2022). In such a setting, the transport map $\nabla\varphi_0$ will not be continuous, demonstrating the need of considering the problem of the statistical estimation of *discontinuous* transport maps to get closer to real-world situations.

As a first step, we choose to focus on the case where the target distribution $Q = \sum_{j=1}^{J} q_j \delta_{y_j}$ is discrete while the source measure P has full support, often called the *semi-discrete* setting in the optimal transport literature. In this setting, the optimal transport map $\nabla\varphi_0$ is constant over regions known as Laguerre cells (each cell corresponding to a different atom of the discrete measure), while displaying discontinuities on their boundaries (see Section 3.1.3 for more details). Figure 3.1 provides such an example. Semi-discrete optimal transport therefore provides a natural class of discontinuous transport maps. We focus on this setting for two reasons. First, it has garnered a

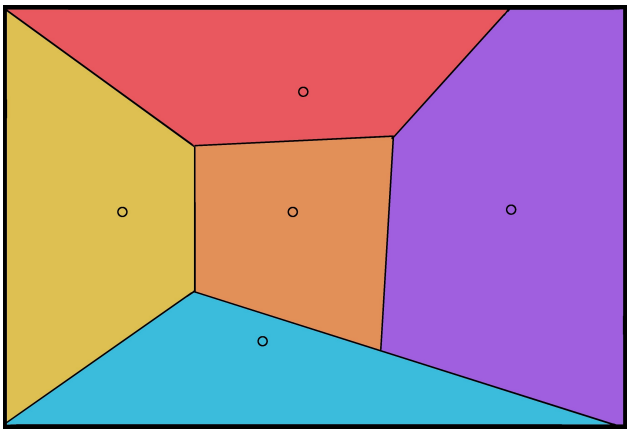


Figure 3.1: An illustration of a semi-discrete optimal transport map. The support of P, the whole rectangle, is partitioned into regions, each of which is transported to one of the atoms of the discrete target measure Q. The resulting map is discontinuous at the boundaries of each cell.

lot of attention in recent years, in both computational and theoretical circles (see, e.g., Altschuler et al., 2022; Chen et al., 2022a; Mérigot et al., 2021), due in particular to its connection with the

quantization problem (Graf and Luschgy, 2007). Second, the semi-discrete setting is intriguing from a statistical perspective: existing results show that statistical estimation problems involving semi-discrete optimal transport can escape the curse of dimensionality (del Barrio et al., 2022a; del Barrio and Loubes, 2019; Forrow et al., 2019; Hundrieser et al., 2024b). For example, Hundrieser et al. (2024b, Theorem 3.2) show that if P_n and Q_n are empirical measures consisting of i.i.d. samples from P and Q, then the semi-discrete assumption implies

$$\mathbb{E}|W_2^2(P,Q) - W_2^2(P_n,Q_n)| \lesssim n^{-1/2}\,.$$

These results offer the tantalizing possibility that semi-discrete transport maps can be estimated at the rate $n^{-1/2}$, in sharp contrast to the dimension-dependent rates obtained in bounds such as (3.1). However, the optimal rates of estimation for semi-discrete transport maps are not known, and no estimators with finite-sample convergence guarantees exist.

3.1.1 Main Contributions

We show that the computationally efficient estimator $\hat{T}_\varepsilon$ from Chapter 2 provably estimates discontinuous semi-discrete optimal transport maps at the optimal rate. More precisely, our contributions are the following:

1. For Q discrete and P with full support on a compact, convex set, we show that $\hat{T}_\varepsilon$ achieves the following *dimension-independent* convergence rate to the optimal transport map (see Theorem 3.2)

$$\mathbb{E}\|\hat{T}_\varepsilon - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim n^{-1/2}\,, \tag{3.2}$$

   when the regularization parameter $\varepsilon \asymp n^{-1/2}$. We further show (Proposition 3.13) that this rate is minimax optimal.

2. As a by-product of our analysis, we give new *parametric* rates of convergence to the entropic

Brenier map T_ε, a result which improves exponentially on prior work in the dependence on ε (see Theorem 3.8 and Remark 3.9).

3. Our proof technique requires several new results, including a novel stability bound for the entropic Brenier maps (Proposition 3.10), and a new stability result for the entropic dual Brenier potentials in the semi-discrete case (Proposition 3.12).

4. We show that, unlike $\hat{T}_\varepsilon$, the 1-Nearest-Neighbor estimator is provably suboptimal in the semi-discrete setting (see Proposition 3.14) by exhibiting a discrete measure Q such that the risk suffers from the curse of dimensionality:

$$\mathbb{E}\|\hat{T}_{1\mathrm{NN}} - \nabla\varphi_0\|^2_{L^2(P)} \gtrsim n^{-1/d}\,.$$

5. In Section 3.3, we verify our theoretical findings on synthetic experiments. We also show by simulation that the entropic estimator appears to perform well even outside the semi-discrete setting, suggesting it as a promising choice for estimating other types of discontinuous maps.

3.1.2 Notation

The Euclidean ball centered at a with radius $r > 0$ is written as $B(a; r)$. The symbols C and c denote positive constants whose value may change from line to line. Write $a \lesssim b$ and $a \asymp b$ if there exist constants $c, C > 0$ such that $a \le Cb$ and $cb \le a \le Cb$, respectively. For an integer $N \in \mathbb{N}$, we let $[N] := \{1, \dots, N\}$. For a function f and a probability measure ρ, we write $\|f\|^2_{L^2(\rho)} := \mathbb{E}_{X\sim\rho}\|f(X)\|^2$. Similarly, we write $\mathrm{Var}_\rho(f) := \mathbb{E}_{X\sim\rho}[(f(X) - \mathbb{E}_{X\sim\rho}[f(X)])^2]$ for the variance of f with respect to ρ.

3.1.3 Background on optimal transport

3.1.3.1 Semi-discrete optimal transport

In optimal transport, the semi-discrete setting refers to the case where P has as density with respect to the Lebesgue measure on $\mathbb{R}^d$, and Q is a discrete measure supported on finitely many points. The following theorem characterizes the optimal transport map in this situation, which exhibits a particular structure compared to the general results in the previous section. Let $[J] \coloneqq \{1, \dots, J\}$.

Proposition 3.1 (Aurenhammer et al., 1998). *If $P \in \mathcal{P}_{ac}(\Omega)$ and Q is a discrete measure supported on the points $y_1, \dots, y_J$, then the optimal transport map $\nabla\varphi_0$ is given by*

$$\nabla\varphi_0(x) \coloneqq \operatorname*{argmax}_{j \in [J]} \{\langle x, y_j \rangle - \psi_0(y_j)\}, \tag{3.3}$$

where ψ_0 is the convex dual to φ_0 in the sense of (1.22).

Here, the optimal dual Brenier potential ψ_0 can be identified with a *vector* in $\mathbb{R}^J$, defined by the number of atoms, and the optimal Brenier potential is consequently given by

$$\varphi_0 \coloneqq \max_{j \in [J]} \{\langle x, y_j \rangle - \psi_0(y_j)\}.$$

Although φ_0 is not differentiable, only subdifferentiable, we still use the gradient notation as $\nabla\varphi_0$ is well-defined P-almost everywhere.

The map $\nabla\varphi_0$ partitions the space into J convex polytopes $L_j \coloneqq \nabla\varphi_0^{-1}(\{y_j\})$ called *Laguerre cells*; recall Figure 3.1. From this definition, it is clear that for a given $x \in L_j$, $x \mapsto \nabla\varphi_0(x) = y_j$ is the optimal transport mapping. The difficulty in finding this map lies in determining the cells L_j, or equivalently the dual variables $\psi_0(y_j)$.

When we want to place emphasis on the underlying measures, we will write $\varphi_0 = \varphi_0^{P \to Q}$, $\psi_0 = \psi_0^{P \to Q}$ and $T_0 = T_0^{P \to Q}$.

3.1.3.2 Reminders for entropic optimal transport

We include some brief reminders for entropic optimal transport. For two measures P and Q with finite second moment, recall that the primal entropic optimal transport problem is

$$\mathrm{OT}_\varepsilon(P,Q) := \min_{\pi\in\Gamma(P,Q)} \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi\|P\otimes Q)\,, \tag{3.4}$$

where $\mathrm{KL}(\mu\|\nu) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu$ when $\mu \in \mathcal{P}(\Omega)$ is absolutely continuous with respect to $\nu \in \mathcal{P}(\Omega)$, and $\varepsilon > 0$ is a positive number.

Equation (3.4) admits the following semi-dual formulation, which is now an unconstrained optimization problem (Genevay, 2019; Marino and Gerolin, 2020)

$$\begin{aligned}\mathrm{OT}_\varepsilon(P,Q) = \tfrac{1}{2}M_2(P) + \tfrac{1}{2}M_2(Q) - \inf_{\varphi,\psi}\Bigg(&\int \varphi\,\mathrm{d}P + \int \psi\,\mathrm{d}Q \\ &+ \varepsilon\iint (e^{(\langle x,y\rangle-\varphi(x)-\psi(y))/\varepsilon} - 1)\,\mathrm{d}P(x)\,\mathrm{d}Q(y)\Bigg),\end{aligned} \tag{3.5}$$

where $(\varphi,\psi) \in L_1(P)\times L_1(Q)$. When P and Q have finite second moments, (3.4) admits a *unique* minimizer, π_ε and we have the existence of minimizers to (3.5), which we denote as $(\varphi_\varepsilon,\psi_\varepsilon)$. We call π_ε the *entropic optimal plan* and $(\varphi_\varepsilon,\psi_\varepsilon)$ are called *entropic Brenier potentials*. The following optimality relation further relates these primal and dual solutions (Csiszár, 1975):

$$\mathrm{d}\pi_\varepsilon(x,y) := e^{(\langle x,y\rangle-\varphi_\varepsilon(x)-\psi_\varepsilon(y))/\varepsilon}\,\mathrm{d}P(x)\,\mathrm{d}Q(y)\,.$$

If $(X,Y)\sim\pi_\varepsilon$, we may define the conditional probability π_ε^x of Y given that $X = x$, with density

$$\frac{\mathrm{d}\pi_\varepsilon^x}{\mathrm{d}Q}(y) \propto \exp\left((\langle x,y\rangle - \psi_\varepsilon(y))/\varepsilon\right)\,. \tag{3.6}$$

The barycentric projection of the optimal entropic coupling π_ε, or *entropic Brenier map*, is a central object of study in several works (see e.g., del Barrio et al. (2022b); Goldfeld et al. (2024a); Pooladian

and Niles-Weed (2021); Rigollet and Stromme (2022)), defined as

$$T_\varepsilon(x) = \int y \, \mathrm{d}\pi_\varepsilon^x(y) = \nabla \varphi_\varepsilon(x)\,, \tag{3.7}$$

where π_ε^x is as in (3.6). Note that this quantity is well defined for all $x \in \mathbb{R}^d$ as long as the source and target measures have compact support; in particular, it applies to both discrete and continuous measures. The second equality follows from (1.27) and the dominated convergence theorem. As in the unregularized case, we will write $\varphi_\varepsilon = \varphi_\varepsilon^{P\to Q}$, $\psi_\varepsilon = \psi_\varepsilon^{P\to Q}$ and $T_\varepsilon = T_\varepsilon^{P\to Q}$ when we want to emphasize on the dependency with respect to the underlying measures.

3.1.3.3 Related work

Characterizing the convergence of entropic objects (e.g. potentials, cost, plans) to their unregularized counterparts in the $\varepsilon \to 0$ regime has been a topic of several works in recent years. Convergence of the costs OT_ε to W_2^2 with precise rates was investigated by Chizat et al. (2020); Conforti and Tamanini (2021); Pal (2024). The works of Bernton et al. (2022); Carlier et al. (2017); Ghosal et al. (2022); Léonard (2012) study the convergence of the minimizers π_ε to π_0 under varying assumptions. Convergence of the potentials in a very general setting was established by Nutz and Wiesel (2021), though without a rate of convergence in ε. In the semi-discrete case, this gap was closed by Altschuler et al. (2022) followed closely by Delalande (2022), which gave non-asymptotic rates. The Sinkhorn Divergence, a non-negative, symmetric version of OT_ε, was introduced by Genevay et al. (2018), was statistically analysed by Goldfeld et al. (2024a) and also del Barrio et al. (2022b); Gonzalez-Sanz et al. (2022), and was connected to the entropic Brenier map by Pooladian et al. (2022). The recent work by Rigollet and Stromme (2022) proved parametric rates of estimation between the empirical entropic Brenier map and its population counterpart, though with an exponentially poor dependence on the regularization parameter (see Remark 3.9). Entropic optimal transport has also come into contact with the area of deep generative modeling through

the following works by De Bortoli et al. (2021); Finlay et al. (2020a), among others.

3.2 Statistical performance of the entropic estimator in the semi-discrete setting

Let P_n and Q_n be the empirical measures associated with two n-samples from P and Q. We make the following regularity assumptions on P, already introduced by Delalande (2022).

(S1) The measure P has a compact convex support $\Omega \subseteq B(0;R)$, with a density p satisfying $0 < p_{\min} \leq p \leq p_{\max} < \infty$ for positive constants $p_{\min}$, $p_{\max}$ and R.

For example, P can be the uniform distribution over Ω, or a truncated Gaussian distribution. Furthermore, we will need the following assumption on Q.

(S2) The discrete probability measure $Q = \sum_{j=1}^{J} q_j \delta_{y_j}$ is such that $q_j \geq q_{\min} > 0$ and $y_j \in B(0;R)$ for all $j \in [J]$.

The goal of this section is to prove the following theorem:

Theorem 3.2. *Let P satisfy **(S1)** and let Q satisfy **(S2)**. Let $\hat{T}_\varepsilon = T_\varepsilon^{P_n \to Q_n}$ be the entropic Brenier map defined from the finite-samples. Then, for $\varepsilon \asymp n^{-1/2}$ and n large enough,*

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)} \lesssim n^{-1/2}\,. \tag{3.8}$$

Remark 3.3. We remark that the hidden constants in Theorem 3.8 and related results depend on $J, p_{\min}, p_{\max}, q_{\min}$ and R.

Remark 3.4 (Fixing the support via rounding). At present, the entropic map need not necessarily map exactly to one of $\{y_1, \dots, y_J\}$. In fact, $\hat{T}_\varepsilon : \mathbb{R}^d \to \mathrm{conv}(\{Y_1, \dots, Y_n\})$, where $\mathrm{conv}(A)$ is the convex hull for some set A. In turn, the support of the entropic map does not in general match that

of Q. However, this can be readily fixed with a rounding scheme. We can replace our estimator by $\bar{T}_\varepsilon$ which is obtained by mapping the output of $\hat{T}_\varepsilon$ to its nearest neighbor in the support of Q—this projection step is easy to compute, given that we essentially know the support of Q via samples. By viewing this as a projection onto an appropriate set (namely, the set of transport maps with codomain equal to the support of Q), and applying the triangle inequality, it holds that

$$\mathbb{E}\|\bar{T}_\varepsilon - T_0\|^2_{L^2(P)} \leq 2\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)}$$

but $\bar{T}_\varepsilon$ matches the support of Q.

Let $T_\varepsilon = T_\varepsilon^{P\to Q}$ denote the entropic Brenier map associated to P and Q. Our proof relies on the following bias-variance decomposition:

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)} \lesssim \mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} + \|T_\varepsilon - T_0\|^2_{L^2(P)}.$$

Following the next two results (Theorem 3.5 and Theorem 3.8) and the preceding decomposition, the proof of Theorem 3.2 is merely a balancing act in the regularization parameter ε.

Theorem 3.5. *Let P satisfy **(S1)** and let Q satisfy **(S2)**. Then, for ε small enough,*

$$\|T_\varepsilon - T_0\|^2_{L^2(P)} \lesssim \varepsilon\,. \tag{3.9}$$

The proof of Theorem 3.5 relies on the following qualitative picture: if a point x belongs to some Laguerre cell L_j, and is far away from the boundary of L_j, then the entropic optimal plan π_ε will send almost all of its mass towards the point $y_j = T_0(x)$, sending an exponentially small amount of mass to the other points y_j. Such a picture is correct as long as x is at distance at least ε from the boundary of the Laguerre cell L_j, incurring a total error of order ε. A rigorous proof of Theorem 3.5 can be found in Appendix B.2.

Note that this rate is slower than the rate appearing in Corollary 2.8 in the continuous-to-

continuous case. The following example shows that the dependency in ε is optimal in Theorem 3.5, indicating that the presence of discontinuities necessarily affects the approximation properties of the entropic Brenier map.

Example 3.6. Let P be a probability measure on $\mathbb{R}$ having a symmetric bounded density p continuous at 0, and let $Q = \frac{1}{2}(\delta_{-1} + \delta_1)$. Following Altschuler et al. (2022, Section 3), one can check that the entropic Brenier map in this setting is the following scaled sigmoidal function

$$T_\varepsilon(x) = \tanh(2x/\varepsilon)\,,$$

whereas the optimal transport map $T_0(x) = \operatorname{sign}(x)$. Then, performing a computation

$$\begin{aligned}\|T_\varepsilon - T_0\|^2_{L^2(P)} &= 2\int_0^\infty (1 - \tanh(2x/\varepsilon))^2 p(x)\,\mathrm{d}x \\ &= \varepsilon \int_0^\infty (1 - \tanh(u))^2 p(u\varepsilon/2)\,\mathrm{d}u \\ &= \varepsilon p(0)(\log(4) - 1) + o(\varepsilon)\,,\end{aligned}$$

where in the last step we invoked the dominated convergence theorem, and computed the limiting integral.

Remark 3.7. Assumption **(S1)** can be relaxed for Theorem 3.5 to hold. More precisely, it can be replaced by Assumptions 2.2 and 2.9 of Altschuler et al. (2022), that hold for unbounded measures such as the normal distribution.

Finally, we present the sample-complexity result:

Theorem 3.8. *Let P satisfy **(S1)** and let Q satisfy **(S2)**. Then, for $0 < \varepsilon \le 1$ such that $\log(1/\varepsilon) \lesssim n/\log(n)$*

$$\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} \lesssim \varepsilon^{-1} n^{-1}\,. \tag{3.10}$$

Remark 3.9. Rigollet and Stromme (2022) show that if P and Q are merely compactly supported with $\mathrm{supp}(P), \mathrm{supp}(Q) \subseteq B(0;R)$, then

$$\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} \lesssim e^{cR^2/\varepsilon}\varepsilon^{-1}n^{-1}\,, \tag{3.11}$$

where $c > 0$ is some absolute positive constant. Thus, under the additional structural assumptions of the semi-discrete formulation, we are able to significantly improve the rate of convergence between the empirical and population entropic Brenier maps.

The proof of Theorem 3.8 relies on a novel stability result, reminiscent of Manole et al. (2024a, Theorem 6), which is of independent interest. We provide the proof in Appendix B.3.

Proposition 3.10. *Let μ, ν, μ', ν' be four probability measures supported in $B(0;R)$. Then the entropic maps $T_\varepsilon^{\mu\to\nu}$ and $T_\varepsilon^{\mu'\to\nu'}$ satisfy*

$$\frac{\varepsilon}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu'\to\nu'}\|^2_{L^2(\mu)} \le \int (\varphi_\varepsilon^{\mu'\to\nu'} - \varphi_\varepsilon^{\mu\to\nu})\,\mathrm{d}\mu + \int (\psi_\varepsilon^{\mu'\to\nu'} - \psi_\varepsilon^{\mu\to\nu})\,\mathrm{d}\nu + \varepsilon\mathrm{KL}(\nu\|\nu')$$

Remark 3.11. The right side of the bound in Proposition 3.10 is equal to

$$S_\varepsilon(\mu,\nu) - S_\varepsilon(\mu',\nu') + \int f_\varepsilon^{\mu'\to\nu'}\,\mathrm{d}(\mu'-\mu) + \int g_\varepsilon^{\mu'\to\nu'}\,\mathrm{d}(\nu'-\nu) + \varepsilon\mathrm{KL}(\nu\|\nu')\,,$$

where $f_\varepsilon^{\mu'\to\nu'} = \frac{1}{2}\|\cdot\|^2 - \varphi_\varepsilon^{\mu'\to\nu'}$ and $g_\varepsilon^{\mu'\to\nu'} = \frac{1}{2}\|\cdot\|^2 - \psi_\varepsilon^{\mu'\to\nu'}$. Proposition 3.10 is therefore the entropic analogue of the stability bounds of Manole et al. (2024a, Theorem 6) and Ghosal and Sen (2022, Lemma 5.1). Unlike those results, Proposition 3.10 allows both the source and target measure to be modified, and does not require any smoothness assumptions.

3.2.1 Proof sketch of Theorem 3.8

To prove Theorem 3.8, we first consider the *one-sample setting*, where we assume that we only have access to samples $Y_1, \dots, Y_n \sim Q$, but we have full access to P. We then consider the one-sample entropic estimator $T_\varepsilon^{P \to Q_n}$. We apply Proposition 3.10 with $\mu = \mu' \coloneqq P$, $\nu \coloneqq Q_n$ and $\nu' \coloneqq Q$, yielding (see Corollary B.3 for details)

$$\frac{\varepsilon}{8R^2}\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|^2_{L^2(\mu)} \le \mathbb{E}\Big(\int (\psi_\varepsilon - \psi_\varepsilon^{P\to Q_n})\,\mathrm{d}(Q_n - Q) + \varepsilon \mathrm{KL}(Q_n\|Q)\Big)\,.$$

Let $\chi^2(P\|Q)$ denote the χ^2-divergence between probability measure. Young's inequality (see Lemma B.15) and the inequality $\mathrm{KL}(Q_n\|Q) \le \chi^2(Q_n\|Q)$ yield the following bound:

$$\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|^2_{L^2(P)} \le \frac{8R^2}{\varepsilon}\Big(\frac{\mathbb{E}[\mathrm{Var}_Q(\psi_\varepsilon^{P\to Q_n} - \psi_\varepsilon)]}{2} + \frac{\mathbb{E}[\chi^2(Q_n\|Q)]}{2}\Big) + 8R^2\mathbb{E}[\chi^2(Q_n\|Q)]\,.$$

To complete our proof sketch, we use a new stability result on the entropic dual Brenier potentials, catered for the semi-discrete setting.

Proposition 3.12. *Let μ be a measure that satisfies **(S1)**. Let ν, ν' be two discrete probability measures supported on $\{y_1, \dots, y_J\}$, with $\nu' \ge \lambda\nu$ for some $\lambda > 0$. Then, for $0 < \varepsilon \le 1$,*

$$\mathrm{Var}_\nu(\psi_\varepsilon^{\mu\to\nu'} - \psi_\varepsilon^{\mu\to\nu}) \le \frac{C}{\lambda^2}\chi^2(\nu'\|\nu), \tag{3.12}$$

where C depends on R, $p_{\min}$ and $p_{\max}$.

Moreover, a computation provided in Lemma B.16 shows that $\mathbb{E}[\chi^2(Q_n\|Q)] = \frac{J-1}{n}$, which is enough to conclude the proof of the one-sample case, see Appendix B.5 for details.

The two-sample setting is tackled using similar reasoning, where we ultimately prove in

Section B.6 that the risk $\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon^{P\to Q_n}\|^2_{L^2(P)}$ is upper bounded by

$$\frac{8R^2}{\varepsilon}\mathbb{E}\int (\varphi_\varepsilon^{P\to Q_n} - \varphi_\varepsilon^{P_n\to Q_n})\,\mathrm{d}(P_n - P)\,.$$

Such a quantity can again be related to the estimation of the dual potentials $\psi_\varepsilon^{P\to Q_n}$ and $\psi_\varepsilon^{P_n\to Q_n}$. Using the same reasoning as before, we expect a parametric rate of convergence for this term as well. Merging the two results completes the proof of Theorem 3.8. We refer to Appendix B.6 for full details.

3.3 Comparing against the 1NN estimator

3.3.1 Rate optimality of the entropic Brenier map

The upper bound of Theorem 3.8 shows that our estimator achieves the $n^{-1/2}$ rate. In fact, the following simple proposition tells us that this rate is optimal in the semi-discrete case.

Proposition 3.13. *Let P be the uniform distribution on $[-1/2, 1/2]^d$ and for any $J \geq 2$, let $\mathcal{Q}_J$ denote the space of of probability measures with at most J atoms, supported on $[-1/2, 1/2]^d$. Define the minimax rate of estimation*

$$\mathcal{R}_n(\mathcal{Q}_J) = \inf_{\hat{T}} \sup_{Q\in\mathcal{Q}_J} \mathbb{E}_{Q^n}[\|\hat{T} - T_0^{P\to Q}\|^2_{L_2(P)}]\,.$$

Then, it holds that $\mathcal{R}_n(\mathcal{Q}_J) \geq n^{-1/2}/64$.

Proof. Let e be a vector of the canonical basis of $\mathbb{R}^d$, scaled by $1/2$. Fix $0 < r < 1/2$ and let $Q_0 = \frac{1}{2}\delta_{-e} + \frac{1}{2}\delta_e$ and $Q_1 = (\frac{1}{2} - r)\delta_{-e} + (\frac{1}{2} + r)\delta_e$. A computation gives $\|T_0^{P\to Q_0} - T_0^{P\to Q_1}\|^2_{L_2(P)} = r$. Therefore, by Le Cam's lemma (see, e.g., Wainwright, 2019, Chapter 15),

$$\mathcal{R}_n(\mathcal{Q}_{J,R}) \geq \frac{r}{8}(1 - \mathrm{TV}(Q_0^n, Q_1^n)). \tag{3.13}$$

Let $\mathrm{H}^2(Q_0, Q_1)$ denote the (squared) Hellinger distance between measures. We have

$$\mathrm{TV}(Q_0^n, Q_1^n)^2 \le \mathrm{H}^2(Q_0^n, Q_1^n) \le n\mathrm{H}^2(Q_0, Q_1)\,.$$

Furthermore, a computation gives

$$\mathrm{H}^2(Q_0, Q_1) = \left(\sqrt{\tfrac{1}{2} - r} - \sqrt{\tfrac{1}{2}}\right)^2 + \left(\sqrt{\tfrac{1}{2} + r} - \sqrt{\tfrac{1}{2}}\right)^2 = 2 - (\sqrt{1+2r} + \sqrt{1-2r}) \le 4r^2.$$

We obtain the conclusion by picking $r = n^{-1/2}/4$. □

3.3.2 The 1NN estimator is proveably suboptimal

The 1-Nearest-Neighbor estimator, henceforth denoted $\hat{T}_{1\mathrm{NN}}$, was proposed by Manole et al. (2024a) as a computational surrogate for estimating optimal transport maps in the low smoothness regime. Written succinctly, their estimator is $\hat{T}_{1\mathrm{NN}}(x) = \sum_{i=1}^n \mathbf{1}_{V_i}(x) Y_{\hat{\pi}(i)}$, where $(V_i)_{i=1}^n$ are Voronoi regions i.e.,

$$V_i \coloneqq \{x \in \mathbb{R}^d \;:\; \|x - X_i\| \le \|x - X_k\|\,, \forall\, k \ne i\}\,,$$

and $\hat{\pi}$ is the optimal transport plan between the empirical measures P_n and Q_n, which amounts to a permutation. Computing the closest X_i to a new sample x has runtime $O(n\log(n))$, though the complexity of this estimator is determined by computing the plan $\hat{\pi}$, which takes $O(n^3)$ time via, e.g., the Hungarian Algorithm (see Peyré and Cuturi, 2019, Chapter 3).

When φ_0 is smooth and strongly convex, Manole et al. (2024a) showed that, for $d \ge 5$,

$$\mathbb{E}\|\hat{T}_{1\mathrm{NN}} - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim n^{-2/d}\,.$$

In contrast to the rate optimality of the entropic Brenier map, we now show that $\hat{T}_{1\mathrm{NN}}$ is

proveably suboptimal in the semi-discrete setting. Not only does it fail to recover the minimax rate obtained by the entropic Brenier map, but its performance in fact degrades in comparison to the smooth case. A proof appears in Appendix B.7.

Proposition 3.14. *There exist a measure P satisfying **(S1)** and a discrete measure Q satisfying **(S2)** such that for $d \geq 3$*

$$\mathbb{E}\|\hat{T}_{1NN} - T_0^{P\to Q}\|^2_{L^2(P)} \gtrsim n^{-1/d}\,.$$

3.3.3 Experiments

We briefly verify our theoretical findings on synthetic experiments. To create the following plots, we draw two sets of n i.i.d. points from P, $(X_1, \dots, X_n)$ and $(X'_1, \dots, X'_n)$, and create target points $Y_i = T_0(X'_i)$, where T_0 is known to us in advance in order to generate the data. Our estimators are computed on the data $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$, and we evaluate the Mean-Squared error criterion

$$\mathrm{MSE}(\hat{T}) = \|\hat{T} - T_0\|^2_{L^2(P)}$$

of a given map estimator $\hat{T}$ using Monte Carlo integration, using 50000 newly sampled points from P. We plot the means across 10 repeated trials, accompanied by their standard deviations.

3.3.3.1 Semi-discrete example #1

First consider $P = \mathrm{Unif}([0,1]^d)$ and create atoms $\{y_1, \dots, y_J\}$ by partitioning the points along the first coordinate for all $j \in [J]$:

$$(y_j)[1] = \frac{(j - 1/2)}{J}\,, \quad (y_j)[2] = \cdots = (y_j)[d] = 0.5\,.$$

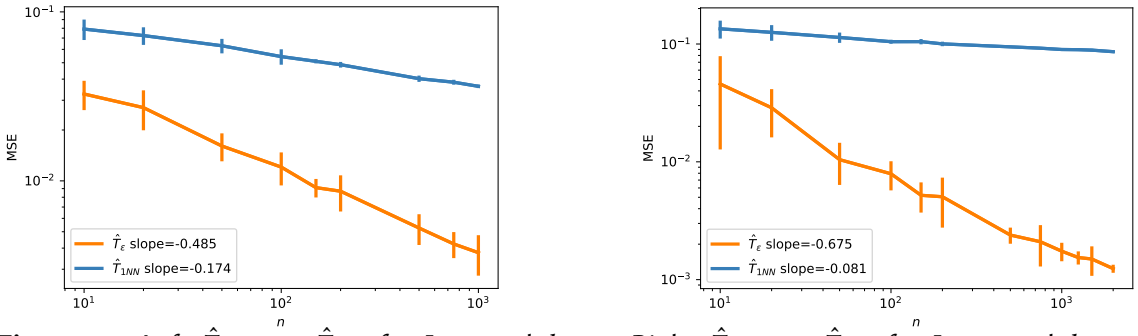


Figure 3.2: Left: $\hat{T}_\varepsilon$ versus $\hat{T}_{1\mathrm{NN}}$ for $J = 2$ and $d = 10$. Right: $\hat{T}_\varepsilon$ versus $\hat{T}_{1\mathrm{NN}}$ for $J = 10$ and $d = 50$.

We choose uniform $q_j = 1/J$ for $j \in [J]$. In this case, it is easy to see that the optimal transport map $T_0(x)$ is uniquely defined by the first coordinate of x_1. Figure 3.2 illustrates the rate-optimal performance of the entropic Brenier map, and the proveably suboptimal performance of the 1-Nearest-Neighbor estimator.

3.3.3.2 Semi-discrete example #2

We now consider a synthetic experiment with far less symmetry. Let $P = \mathrm{Unif}([0,1]^d)$, and fix $J \in \mathbb{N}$. We randomly generate $y_1, \ldots, Y_J \in [0,1]^d$, and also randomly generate $\psi_0 \in \mathbb{R}^J$, and consider the optimal transport map $T_0(x) = \mathrm{argmin}_{j\in[J]}\{x^\top y_j - (\psi_0)_j\}$. We define $Q = (T_0)_\sharp P$, leading to the same setup as before, but with a less structured optimal transport map. We consider $J = 5$ and $d = 50$, and repeat the procedure of the preceding section to generate our data, and the resulting estimator. Figure 3.3 contains plots the MSE as a function of n, where again we see a log-linear slope of around -0.5, which agrees with our theory.

3.3.3.3 Discontinuous example

We turn our attention to a discontinuous transport map, where for $x \in \mathbb{R}^d$, all the coordinates are fixed except for the first one

$$T_0(x) = 2\mathrm{sign}(x[1]) \otimes x[2] \otimes \cdots \otimes x[d]\,.$$

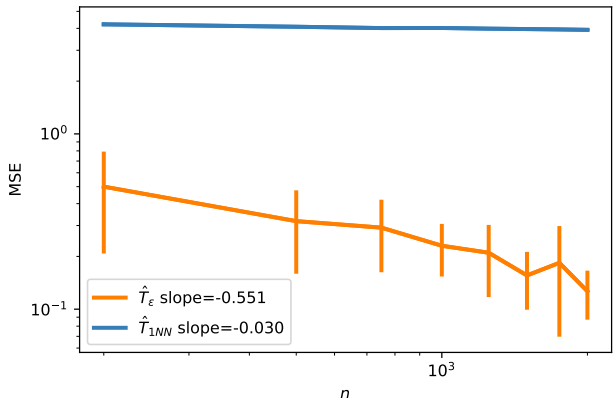


Figure 3.3: $\hat{T}_\varepsilon$ versus $\hat{T}_{1\mathrm{NN}}$ for with ψ_0 random in $d = 50$

We choose $P = \mathrm{Unif}([-1,1]^d)$ to exhibit a discontinuity in the data. Focusing on $d = 10$, we see in Figure 3.4 that the entropic map estimator avoids the curse of dimensionality and enjoys a faster convergence rate, with better constants.

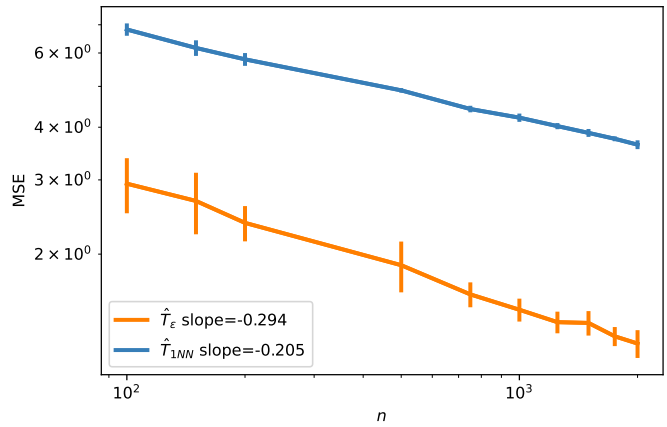


Figure 3.4: $\hat{T}_\varepsilon$ versus $\hat{T}_{1\mathrm{NN}}$ for $d = 10$

4 | Plug-in estimation of Schrödinger bridges

4.1 Introduction

Modern statistical learning tasks often involve not merely the comparison of two unknown probability distributions but also the estimation of *transformations* from one distribution to another. Estimating such transformations is necessary when we want to generate new samples, infer trajectories, or track the evolution of particles in a dynamical system. In these applications, we want to know not only how "close" two distributions are, but also how to "go" between them.

Optimal transport theory defines objects that are well suited for both of these tasks (Santambrogio, 2015; Villani, 2009). The 2-Wasserstein distance is a popular tool for comparing probability distributions for data analysis in statistics (Carlier et al., 2016; Chernozhukov et al., 2017; Ghosal and Sen, 2022), machine learning (Salimans et al., 2018), and the applied sciences (Bunne et al., 2023b; Manole et al., 2024b). Recall that under suitable conditions, the two probability measures that we want to compare (say, μ and ν) induce an optimal transport map: the uniquely defined vector-valued function which acts as a transport map between μ and ν such that the distance traveled is minimal in the L^2 sense (Brenier, 1991). Despite being a central object in many applications, the optimal transport map is difficult to compute and suffers from poor statistical estimation guarantees in high dimensions; see Divol et al. (2022); Hütter and Rigollet (2021); Manole et al. (2024a).

These drawbacks of the optimal transport map suggest that other approaches for defining a transport between two measures may often be more appropriate. For example, *flow based* or iterative approaches have recently begun to dominate in computational applications—these methods sacrifice the L^2-optimality of the optimal transport map to place greater emphasis on the tractability of the resulting transport. The work of Chen et al. (2018) proposed continuous normalizing flows (CNFs), which use neural networks to model the vector field in an ordinary differential equation (ODE). This machinery was exploited by several groups simultaneously (Albergo and Vanden-Eijnden, 2022; Lipman et al., 2022; Liu et al., 2022b) for the purpose of developing tractable constructions of vector fields that satisfy the *continuity equation* (recall Section 4.2.1.1), and whose flow maps therefore yield valid transports between source and target measures.

An increasingly popular alternative method for iterative transport is based on the Fokker–Planck equation (see Section 4.2.1 for a definition). This formulation incorporates a diffusion term, and the resulting dynamics follow a *stochastic* differential equation (SDE). Though there exist many stochastic dynamics that give rise to valid transports, a canonical role is played by the *Schrödinger bridge* (SB). Just as the optimal transport map minimizes the L^2 distance in transporting between two distributions, the SB minimizes the *relative entropy* of the diffusion process, and therefore has an interpretation as the "simplest" stochastic process bridging the two distributions—indeed, the SB originates as a *Gedankenexperiment* (or "thought experiment") of Erwin Schrödinger in modeling the large deviations of diffusing gasses (Schrödinger, 1932). There are many equivalent formulations of the SB problem (see Section 4.2.2), though for the purposes of transport, its most important property is that it gives rise to a pair of SDEs that interpolate between two measures μ and ν:

$$\mathrm{d}X_t = b_t^\star(X_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,, \quad X_0 \sim \mu, X_1 \sim \nu\,, \tag{4.1}$$

$$\mathrm{d}Y_t = d_t^\star(Y_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,, \quad Y_0 \sim \nu, Y_1 \sim \mu\,, \tag{4.2}$$

where $\varepsilon > 0$ plays the role of thermal noise.[1] Concretely, (4.1) indicates that samples from ν can be obtained by drawing samples from μ and simulating an SDE with drift $b_t^\star$, and (4.2) shows how this process can be performed in reverse. Though these dynamics are of obvious use in generating samples, the difficulty lies in obtaining estimators for the drifts.

Nearly a century later, Schrödinger's thought experiment has been brought to reality, having found applications in the generation of new images, protein structures, and more (Kawakita et al., 2022; Lee et al., 2024; Liu et al., 2022a; Nusken et al., 2022; Shi et al., 2022; Thornton et al., 2022). The foundation for these advances is the work of De Bortoli et al. (2021), who propose to train two neural networks to act as the forward and backward drifts, which are iteratively updated to ensure that each diffusion yields samples from the appropriate distribution. This is reminiscent of the iterative proportion fitting procedure of Fortet (1940), and can be interpreted as a version of Sinkhorn's matrix-scaling algorithm (Cuturi, 2013; Sinkhorn, 1967) on path space.

While the framework of De Bortoli et al. (2021) is popular from a computational perspective, it is worth emphasizing that this method is relatively costly, as it necessitates the undesirable task of simulating an SDE at each training iteration. Moreover, despite the recent surge in applications, current methods do not come with statistical guarantees to quantify their performance. In short, existing work leaves open the problem of developing tractable, statistically rigorous estimators for the Schrödinger bridge.

4.1.1 Contributions

We propose and analyze a computationally efficient estimator of the Schrödinger bridge which we call the *Sinkhorn Bridge*. Our main insight is that it is possible to estimate the *time-dependent* drifts in (4.1) and (4.2) by solving a *single, static* entropic optimal transport problem between samples from the source and target measures. Our approach is to compute the potentials $(\hat{f}, \hat{g})$ obtained by running Sinkhorn's algorithm on the data $X_1, \dots, X_m \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$ and plug

[1] We assume throughout our work that the reference process is Brownian motion with volatility ε; see Section 4.2.2.

these estimates into a simple formula for the drifts. For example, in the forward case, our estimator reads

$$\hat{b}_t(z) := (1-t)^{-1}\Big(-z + \frac{\sum_{j=1}^n Y_j \exp\big((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon\big)}{\sum_{j=1}^n \exp\big((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon\big)}\Big).$$

See Section 4.3.1 for a detailed motivation for the choice of $\hat{b}_t$. Once the estimated potential $\hat{g}$ is obtained from a single use of Sinkhorn's algorithm on the source and target data at the beginning of the procedure, computing $\hat{b}_t(z)$ for any $z \in \mathbb{R}^d$ and any $t \in (0,1)$ is trivial.

We show that the solution to a discretized SDE implemented with the estimated drift $\hat{b}_t$ closely tracks the law of the solution to (4.1) on the whole interval $[0,\tau]$, for any $\tau \in [0,1)$. Indeed, writing $\mathrm{P}^\star_{[0,\tau]}$ for the law of the process solving (4.1) on $[0,\tau]$ and $\hat{\mathrm{P}}_{[0,\tau]}$ for the law of the process obtained by initializing from a fresh sample $X_0 \sim \mu$ and solving a discrete-time SDE with drift $\hat{b}_t$, we prove bounds on the risk

$$\mathbb{E}[\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \mathrm{P}^\star_{[0,\tau]})]$$

that imply that, for fixed $\varepsilon > 0$ and $\tau \in [0,1)$, the Schrödinger bridge can be estimated at the *parametric* rate. Moreover, though it is well known that such bounds must diverge as $\varepsilon \to 0$ or $\tau \to 1$, we demonstrate that the rate of growth depends on the *intrinsic* dimension k of the target measure rather than the ambient dimension d. When $\mathrm{k} \ll d$, this gives strong justification for the use of the Sinkhorn Bridge estimator in high-dimensional problems.

To give a particular example in a special case, our results provide novel estimation rates for the *Föllmer bridge*, an object which has also garnered interest in the machine learning community (Chen et al., 2024; Huang, 2024; Vargas et al., 2023). In this setting, the source measure is a Dirac mass, and we suppose the target measure ν is supported on a ball of radius R contained within a k-dimensional smooth submanifold of $\mathbb{R}^d$. Taking the volatility level to be unity, we show that the

Föllmer bridge up to time $\tau \in [0, 1)$ can be estimated in total variation with precision ϵ_{TV} using n samples and N SDE-discretization steps, where

$$n \asymp R^2(1-\tau)^{-\mathrm{k}-2}\epsilon_{\mathrm{TV}}^{-2}, \quad N \lesssim dR^4(1-\tau)^{-4}\epsilon_{\mathrm{TV}}^{-2}.$$

As advertised, for fixed $\tau \in [0, 1)$, these bounds imply parametric scaling on the number of samples (which matches similar findings in the entropic optimal transport literature, see, e.g., Stromme (2024)) and exhibit a "curse of dimensionality" only with respect to the *intrinsic* dimension of the target, k. As our main theorem shows, these phenomena are not unique to the Föllmer bridge, and hold for arbitrary volatility levels and general source measures. Moreover, by tuning τ appropriately, we show how these estimation results yield guarantees for sampling from the target measure ν, see Section 4.4.3. These guarantees also suffer only from a "curse of intrinsic dimensionality." Since the drifts arising from the Föllmer bridge can be viewed as the score of a kernel density estimator of ν with a Gaussian kernel (see (4.26)), this benign dependence on the ambient dimension is a significant improvement over guarantees recently obtained for such estimators in the context of denoising diffusion probabilistic models (Wibisono et al., 2024). Our improved rates are due to the intimate connection between the SB problem and entropic optimal transport in which intrinsic dimensionality plays a crucial role (Groppe and Hundrieser, 2024; Stromme, 2024). We expound on this connection in the main text.

We are not the first to notice the simple connection between the static entropic potentials and the SB drift. Finlay et al. (2020a) first proposed to exploit this connection to simulate the SB by learning static potentials via a neural network-based implementation of Sinkhorn's algorithm; however, due to some notational inaccuracies and implementation errors, the resulting procedure was not scalable. This work shows the theoretical soundness of their approach, with a much simpler, tractable algorithm and with rigorous statistical guarantees.

4.1.1.1 Notation

We denote the space of probability measures over $\mathbb{R}^d$ with finite second moment by $\mathcal{P}_2(\mathbb{R}^d)$. We write $B(x, R) \subseteq \mathbb{R}^d$ to indicate the (Euclidean) ball of radius $R > 0$ centered at $x \in \mathbb{R}^d$. We denote the maximum of a and b by $a \vee b$. We write $a \lesssim b$ (resp. $a \asymp b$) if there exists constants $C > 0$ (resp. $c, C > 0$ such that $a \leq Cb$ (resp. $cb \leq a \leq Cb$). We let $\text{path} := C([0,1], \mathbb{R}^d)$ be the space of paths with $X_t : \text{path} \to \mathbb{R}^d$ given by the canonical mapping $X_t(h) = h_t$ for any $h \in \text{path}$ and any $t \in [0,1]$. For a path measure $\mathsf{P} \in \mathcal{P}(\text{path})$ and any $t \in [0,1]$, we write $\mathsf{P}_t := (X_t)_\sharp \mathsf{P} \in \mathcal{P}(\mathbb{R}^d)$ for the t^{th} marginal of P_t. Similarly, for $s, t \in [0,1]$, we can define the joint probability measure $\mathsf{P}_{st} := (X_s, X_t)_\sharp \mathsf{P}$. We write $\mathsf{P}_{[0,t]}$ for the restriction of the P to $C([0,t], \mathbb{R}^d)$. Since path is a Polish space, we can define regular conditional probabilities for the law of a path given its value at time t, which we denote $\mathsf{P}_{|t}$. For any $s > 0$, we write $\Lambda_s := (2\pi s)^{-d/2}$ for the normalizing constant of the density of the Gaussian distribution $\mathcal{N}(0, sI)$.

4.1.2 Related work

On Schrödinger bridges. Connections between entropic optimal transport and the Schrödinger bridge (SB) problem are well studied; see the comprehensive survey by Léonard (2014). We were also inspired by the works of Ripani (2019), Gentil et al. (2020), as well as Chen et al. (2016; 2021b) (which cover these topics from the perspective of optimal control), and the more recent article by Kato (2024) (which revisits the large-deviation perspective of this problem). The special case of the Föllmer bridge and its variants has been a topic of recent study in theoretical communities (Eldan et al., 2020; Mikulincer and Shenfeld, 2024).

Interest in computational methods for SBs has been explosive in over the last few years, see Bunne et al. (2023a); Chen et al. (2024); De Bortoli et al. (2021); Shi et al. (2024; 2022); Tong et al. (2023); Vargas et al. (2023); Yim et al. (2023) for recent developments in deep learning. The works by Bernton et al. (2019); Pavon et al. (2021); Vargas et al. (2021) use more traditional

statistical methods to estimate the SB, with various goals in mind. For example, Bernton et al. (2019) propose a sampling scheme based on trajectory refinements using a approximate dynamic programming approach. Pavon et al. (2021) and Vargas et al. (2021) propose methods to compute the (intermediate) density directly based on maximum likelihood-type estimators: Pavon et al. (2021) directly model the densities of interest and devise a scheme to update the weights; Vargas et al. (2021) use Gaussian processes to model the forward and backward drifts, and update them via a maximum-likelihood type loss.

On entropic optimal transport. Our work is closely related to the growing literature on statistical entropic optimal transport, specifically on the developments surrounding the *entropic transport map*. This object was introduced by Pooladian and Niles-Weed (2021) as a computationally friendly estimator for optimal transport maps in the regime $\varepsilon \to 0$; see also Pooladian et al. (2023) for minimax estimation rates in the semi-discrete regime. When ε is fixed, the theoretical properties of the entropic maps have been analyzed (Chewi and Pooladian, 2023; Chiarini et al., 2022; Conforti, 2024; Conforti et al., 2023; Divol et al., 2025) as well as their statistical properties (del Barrio et al., 2022b; Goldfeld et al., 2024a;b; Gonzalez-Sanz et al., 2022; Rigollet and Stromme, 2022; Werenski et al., 2023). Ghosal et al. (2022); Nutz and Wiesel (2021) study the stability of entropic potentials and plans in a qualitative sense under minimal regularity assumptions. Most recently, Stromme (2024) and Groppe and Hundrieser (2024) established the connections between statistical entropic optimal transport and intrinsic dimensionality (for both maps and costs). Daniels et al. (2021) investigates sampling using entropic optimal transport couplings combined with neural networks. Closely related are the works by Chizat et al. (2022) and Lavenant et al. (2024), which highlight the use of entropic optimal transport for trajectory inference. A more flexible alternative to the entropic transport map was recently developed by Kassraie et al. (2024), who proposed a transport that progressively displaces the source measure to the target measure by computing a new entropic transport map at each step to approximate the *McCann interpolation* (McCann, 1997).

4.2 Background

4.2.1 Preliminaries on entropic optimal transport

We require a slightly different change in notation this chapter. Recall the primal and dual entropic optimal transport problems are

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \inf_{\pi\in\Pi(P,Q)} \iint \tfrac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi\|\mu\otimes\nu) \tag{4.3a}$$

$$= \sup_{(f,g)\in\mathcal{F}} \mathcal{D}_\varepsilon^{\mu\nu}(f,g) \tag{4.3b}$$

where $\mathcal{F} = L^1(\mu)\times L^1(\nu)$ and

$$\mathcal{D}_\varepsilon^{\mu\nu}(f,g) \coloneqq \int f\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu - \varepsilon \iint \left(\Lambda_\varepsilon e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} - 1\right)\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\,, \tag{4.4}$$

where $\Lambda_\varepsilon = (2\pi\varepsilon)^{-d/2}$. Solutions to both problems are guaranteed to exist when $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$. The minimizer to (4.3a) is called the *optimal entropic plan*, written $\pi^\star\in\Pi(\mu,\nu)$, and the dual optimizers the *optimal entropic (Kantorovich) potentials*, written $(f^\star,g^\star)$.[2]

Csiszár (1975) showed that the primal and dual optima are intimately connected through the following relationship:[3]

$$\mathrm{d}\pi^\star(x,y) = \Lambda_\varepsilon \exp\left(\frac{f^\star(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\,. \tag{4.5}$$

Though $f^\star$ and $g^\star$ are *a priori* defined almost everywhere on the support of μ and ν, they can be extended to all of $\mathbb{R}^d$ (see Mena and Niles-Weed (2019); Nutz and Wiesel (2021)) via the optimality

[2] Though $\pi^\star$ and the other objects discussed in this section depend on ε, we will omit this dependence for the sake of readability, though we will track the dependence on ε in our bounds.

[3] The normalization factor Λ_ε is not typically used in the computational optimal transport literature, but it simplifies some formulas in what follows. Since the procedure we propose is invariant under translation of the optimal entropic potentials, this normalization factor does not affect either our algorithm or its analysis.

conditions

$$f^\star(x) = -\varepsilon \log\left(\Lambda_\varepsilon \int e^{(g^\star(y)-\|x-y\|^2/2)/\varepsilon}\,\mathrm{d}\nu(y)\right),\ g^\star(y) = -\varepsilon \log\left(\Lambda_\varepsilon \int e^{(f^\star(x)-\|x-y\|^2/2)/\varepsilon}\,\mathrm{d}\mu(x)\right).$$

At times, it will be convenient to work with *entropic Brenier potentials*, defined as

$$(\varphi^\star, \psi^\star) \coloneqq (\tfrac{1}{2}\|\cdot\|^2 - f^\star, \tfrac{1}{2}\|\cdot\|^2 - g^\star).$$

Note that the gradients of the entropic Brenier potentials[4] are related to barycentric projections of the optimal entropic coupling

$$\nabla\varphi^\star(x) = \mathbb{E}_{\pi^\star}[Y|X=x], \qquad \nabla\psi^\star(y) = \mathbb{E}_{\pi^\star}[X|Y=y].$$

For a proof of this fact, see Pooladian and Niles-Weed (2021, Proposition 2). By analogy with the unregularized optimal transport problem, these are called *entropic Brenier maps*. The following relationships can also be readily verified:

$$\nabla^2\varphi^\star(x) = \varepsilon^{-1}\mathrm{Cov}_{\pi^\star}[Y|X=x], \qquad \nabla^2\psi^\star(y) = \varepsilon^{-1}\mathrm{Cov}_{\pi^\star}[X|Y=y]. \tag{4.7}$$

4.2.1.1 A dynamic formulation via the continuity equation

Entropic optimal transport can also be understood from a dynamical perspective. Let $(\mathrm{p}_t)_{t\in[0,1]}$ be a family of measures in $\mathcal{P}_2(\mathbb{R}^d)$, and let $(v_t)_{t\in[0,1]}$ be a family of vector fields. We say that the pair satisfies the *continuity equation*, written $(\mathrm{p}_t, v_t) \in \mathfrak{C}$, if $\mathrm{p}_0 = \mu$, $\mathrm{p}_1 = \nu$, and, for $t \in [0,1]$,

$$\partial_t \mathrm{p}_t + \nabla\cdot(v_t \mathrm{p}_t) = 0. \tag{4.8}$$

[4] Passing the gradient under the integral is permitted via dominated convergence under suitable tail conditions on μ and ν.

Solutions to (4.8) are understood to hold in the weak sense (that is, with respect to suitably smooth test functions).

The continuity equation can be viewed as the analogue of the marginal constraints being satisfied (i.e., the set $\Pi(\mu, \nu)$ above): it represents both the conservation of mass and the requisite end-point constraints for the path $(\mathrm{p}_t)_{t\in[0,1]}$. With this, we can cite a clean expression of the dynamic formulation of the entropic optimal transport problem (see Conforti and Tamanini (2021) or Chizat et al. (2020)) if μ and ν are absolutely continuous and have finite entropy:

$$\mathrm{OT}_\varepsilon(\mu, \nu) + C_0(\varepsilon, \mu, \nu) = \inf_{(\mathrm{p}_t, v_t)\in\mathfrak{C}} \int_0^1 \int \left(\frac{1}{2}\|v_t(x)\|^2 + \frac{\varepsilon^2}{8}\|\nabla \log \mathrm{p}_t(x)\|^2\right) \mathrm{d}\mathrm{p}_t(x)\, \mathrm{d}t\,, \tag{4.9}$$

where $C_0(\varepsilon, \mu, \nu) \coloneqq \varepsilon \log(\Lambda_\varepsilon) + \frac{\varepsilon}{2}(\mathcal{H}(\mu) + \mathcal{H}(\nu))$ is an additive constant, with $\mathcal{H}(\mu) \coloneqq \int \log(\mathrm{d}\mu)\, \mathrm{d}\mu$, similarly for $\mathcal{H}(\nu)$.

The case $\varepsilon = 0$ reduces to the celebrated Benamou–Brenier formulation of optimal transport (Benamou and Brenier, 2000).

4.2.1.2 A stochastic formulation via the Fokker–Planck equation

We now revisit the dynamic formulation of entropic optimal transport, this time based on the *Fokker–Planck equation*. This equation is said to be satisfied by a pair $(\mathrm{p}_t, b_t) \in \mathfrak{F}$ if $\mathrm{p}_0 = \mu$, $\mathrm{p}_1 = \nu$, and, for $t \in [0, 1]$,

$$\partial_t \mathrm{p}_t + \nabla \cdot (b_t \mathrm{p}_t) = \frac{\varepsilon}{2} \Delta \mathrm{p}_t\,.$$

Then, under the same conditions as above,

$$\mathrm{OT}_\varepsilon(\mu, \nu) + C_1(\varepsilon, \mu) = \inf_{(\mathrm{p}_t, b_t)} \int_0^1 \int \frac{1}{2}\|b_t(x)\|^2\, \mathrm{d}\mathrm{p}_t(x)\, \mathrm{d}t\,, \tag{4.10}$$

where $C_1(\varepsilon, \mu) = \varepsilon \log(\Lambda_\varepsilon) + \varepsilon \mathcal{H}(\mu)$. The equivalence between the objective functions (4.9) and (4.10), as well as the continuity equation and Fokker–Planck equations, is classical. For completeness, we provide details of these computations in Appendix C.1. A key property of this equivalence is the following relationship which relates the optimizers of (4.9), written $(\mathrm{p}_t^\star, v_t^\star)$ and (4.10), written $(\mathrm{p}_t^\star, b_t^\star)$:

$$b_t^\star = v_t^\star + \frac{\varepsilon}{2} \nabla \log \mathrm{p}_t^\star \,.$$

We stress that the minimizer $\mathrm{p}_t^\star$ is the same for both (4.9) and (4.10).

4.2.2 The Schrödinger Bridge problem and the Fokker–Planck equation

We will now briefly develop the required machinery to understand the Schrödinger bridge problem. We will largely following the expositions of Gentil et al. (2020); Léonard (2012; 2014); Ripani (2019).

For $\varepsilon > 0$, we let $\mathrm{R} \in \mathcal{P}(\mathrm{path})$ denote the law of the reversible Brownian motion on $\mathbb{R}^d$ with volatility ε, with the Lebesgue measure as the initial distribution.[5] We write the joint distribution of the initial and final positions under R by $\mathrm{R}_{01}(\mathrm{d}x, \mathrm{d}y) = \Lambda_\varepsilon \exp(-\frac{1}{2}\|x - y\|^2/\varepsilon)\, \mathrm{d}x\, \mathrm{d}y$.

With the above, we arrive at Schrödinger's bridge problem over path measures:

$$\min_{\mathrm{P} \in \mathcal{P}(\mathrm{path})} \varepsilon\, \mathrm{KL}(\mathrm{P} \| \mathrm{R}) \quad \text{s.t.} \quad \mathrm{P}_0 = \mu\,, \mathrm{P}_1 = \nu\,, \tag{4.11}$$

where $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and are absolutely continuous with finite entropy. Let $\mathrm{P}^\star$ be the unique solution to (4.11), which exists as the problem is strictly convex. Léonard (2014) shows that there

[5] The problem below remains well posed even though R is not a probability measure; see Léonard (2014) for complete discussions.

exist two non-negative functions $\mathfrak{f}^\star, \mathfrak{g}^\star : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\mathsf{P}^\star = \mathfrak{f}^\star(X_0)\mathfrak{g}^\star(X_1)\mathsf{R}\,, \tag{4.12}$$

where $\mathrm{Law}(X_0) = \mu$ and $\mathrm{Law}(X_1) = \nu$.

A further connection can be made: if we apply the chain-rule for the KL divergence by conditioning on times $t = 0, 1$, the objective function (4.11) decomposes into

$$\varepsilon\,\mathrm{KL}(\mathsf{P}\|\mathsf{R}) = \varepsilon\,\mathrm{KL}(\mathsf{P}_{01}\|\mathsf{R}_{01}) + \varepsilon\,\mathbb{E}_{\mathsf{P}}\mathrm{KL}(\mathsf{P}_{|01}\|\mathsf{R}_{|01})\,.$$

Under the assumption that μ and ν have finite entropy, it can be shown that the first term on the right-hand side is equivalent to the objective for the entropic optimal transport problem in (4.3a). Moreover, the second term vanishes if we choose the measure P so that the conditional measure $\mathsf{P}_{|01}$ is the same as $\mathsf{R}_{|01}$, i.e., a Brownian bridge. Therefore, the objective function in (4.11) is minimized when $\mathsf{P}^\star_{01} = \pi^\star$ and when P writes as a mixture of Brownian bridges with the distribution of initial and final points given by $\pi^\star$:

$$\mathsf{P}^\star = \iint \mathsf{R}(\cdot|X_0 = x_0, X_1 = x_1)\pi^\star(\mathrm{d}x_0, \mathrm{d}x_1)\,. \tag{4.13}$$

Much of the discussion above assumed that μ and ν are absolutely continuous with finite entropy; indeed, the manipulations in this section as well as in Section 4.2.1.1 and 4.2.1.2 are not justified if this condition fails. Though the finite entropy conditioned is adopted liberally in the literature on Schrödinger bridges, in this work we will have to consider bridges between measures that may not be absolutely continuous (for example, empirical measures). Noting that the entropic optimal transport problem (4.3a) has a unique solution for *any* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we leverage this fact to use (4.13) as the *definition* of the Schrödinger bridge between two probability measures: for any pair of probability distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, their *Schrödinger bridge* is the mixture of Brownian

bridges given by (4.13), where $\pi^\star$ is the solution to the entropic optimal transport problem (4.3a).

4.3 Proposed estimator: The Sinkhorn bridge

Our goal is to efficiently estimate the Schrödinger bridge (SB) on the basis of samples. Let $\mathsf{P}^\star$ denote the SB between μ and ν, and define the the time-marginal flow of the bridge by

$$\mathsf{p}_t^\star := \mathsf{P}_t^\star, \qquad t \in [0,1]\,. \tag{4.14}$$

This choice of notation is deliberate: when μ and ν have finite entropy, the t-marginals of $\mathsf{P}^\star$ for $t \in [0,1]$ solve the dynamic formulations (4.9) and (4.10) (Léonard, 2014, Proposition 4.1). In the existing literature, $\mathsf{p}_t^\star$ is sometimes called the the *entropic interpolation* between μ and ν. See Gentil et al. (2020); Léonard (2012; 2014); Ripani (2019) for interesting properties of entropic interpolations (for example, their relation to functional inequalities). Our goal is to provide an estimator $\hat{\mathsf{P}}$ such that $\mathbb{E}[\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})]$ is small for all $\tau < 1$. In particular, this marginals of our estimator $\hat{\mathsf{P}}$ are estimators $\hat{\mathsf{p}}_t$ of $\mathsf{p}_t^\star$ for all $t \in [0,1)$.[6]

We call our estimator the *Sinkhorn bridge*, and we outline its construction below. Our main observation involves revisiting some finer properties of entropic interpolations as a function of the static entropic potentials. Once everything is concretely expressed, a natural plug-in estimator will arise which is amenable to both computational and statistical considerations.

[6]For reasons that will be apparent in the next section, time $\tau = 1$ must be excluded from the analysis.

4.3.1 From Schrödinger to Sinkhorn and back

We outline two crucial observations from which our estimator naturally arises. First, we note that $\mathrm{p}_t^\star$ can be explicitly expressed as the following density (Léonard, 2014, Theorem 3.4)

$$\mathrm{p}_t^\star(\mathrm{d}z) \coloneqq \mathsf{H}_{(1-t)\varepsilon}[\exp(g^\star/\varepsilon)\nu](z)\mathsf{H}_{t\varepsilon}[\exp(f^\star/\varepsilon)\mu](z)\,\mathrm{d}z\,, \tag{4.15}$$

where H_s is the *heat semigroup*, which acts on a measure Q via

$$Q \mapsto \mathsf{H}_s[Q](z) \coloneqq \Lambda_s \int e^{-\frac{1}{2s}\|x-z\|^2} Q(\mathrm{d}x)\,.$$

This expression for the marginal of distribution $\mathrm{p}_t^\star$ follows directly from (4.13):

$$\begin{aligned}
\mathrm{p}_t^\star(z) &\coloneqq \iint \mathsf{R}_t(z|X_0 = x_0, X_1 = x_1)\pi^\star(\mathrm{d}x_0, \mathrm{d}x_1) \\
&= \iint \mathcal{N}(z|ty + (1-t)x, t(1-t)\varepsilon)\pi^\star(\mathrm{d}x, \mathrm{d}y) \\
&= \Lambda_\varepsilon \iint e^{((f^\star(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon)}\mathcal{N}(z|ty + (1-t)x, t(1-t)\varepsilon)\mu(\mathrm{d}x)\nu(\mathrm{d}y) \\
&= \int e^{g^\star(y)/\varepsilon}\mathcal{N}(z|y, (1-t)\varepsilon)\nu(\mathrm{d}y) \int e^{f^\star(x)/\varepsilon}\mathcal{N}(z|x, t\varepsilon)\mu(\mathrm{d}x) \\
&= \mathsf{H}_{1-t}[\exp(g^\star/\varepsilon)\nu](z)\mathsf{H}_t[\exp(f^\star/\varepsilon)\mu](z)
\end{aligned}$$

where throughout we use $\mathcal{N}(z|m, \sigma^2)$ to denote the Gaussian density with mean m and covariance $\sigma^2 I$, and the fourth equality follows from computing the explicit density of the product of two Gaussians.

Also, Léonard (2014, Proposition 4.1) shows that when μ and ν have finite entropy, the optimal drift in (4.10) is given by

$$b_t^\star(z) = \varepsilon\nabla \log \mathsf{H}_{(1-t)\varepsilon}[\exp(g^\star/\varepsilon)\nu](z)\,,$$

whence the pair $(\mathrm{p}_t^\star, b_t^\star)$ satisfies the Fokker–Planck equation. This fact implies that if X_t solves

$$\mathrm{d}X_t = b_t^\star(X_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,, \qquad X_0 \sim \mu\,, \tag{4.16}$$

then $\mathrm{p}_t^* = \mathrm{Law}(X_t)$. In fact, more is true: the SDE (4.16) give rise to a path measure, which exactly agrees with the Schrödinger bridge. Though Léonard (2014) derives these facts for μ and ν with finite entropy, we show in Proposition 4.1, below, that they hold in more generality.

Further developing the expression for $b_t^\star$, we obtain

$$b_t^\star(z) = (1-t)^{-1}\Bigg(-z + \frac{\int y e^{(g^\star(y) - \frac{1}{2(1-t)}\|z-y\|^2)/\varepsilon}\,\mathrm{d}\nu(y)}{\int e^{(g^\star(y) - \frac{1}{2(1-t)}\|z-y\|^2)/\varepsilon}\,\mathrm{d}\nu(y)}\Bigg) =: (1-t)^{-1}(-z + \nabla\varphi_{1-t}^\star(z))\,. \tag{4.17}$$

Thus, our final expression for the SDE that yields the Schrödinger bridge is

$$\mathrm{d}X_t = (-(1-t)^{-1}X_t + (1-t)^{-1}\nabla\varphi_{1-t}^\star(X_t))\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,. \tag{4.18}$$

Once again, we emphasize that our choice of notation here is deliberate: the drift is expressed as a function of a particular entropic Brenier map, namely, the entropic Brenier map between $\mathrm{p}_t^\star$ and ν with regularization parameter $(1-t)\varepsilon$.

We summarize this collection of crucial properties in the following proposition; see Appendix C.2 for proofs. We note that this result avoids the finite entropy requirements of analogous results in the literature (Léonard, 2014; Shi et al., 2024).

Proposition 4.1. *Let π be a* probability *measure of the form*

$$\pi(\mathrm{d}x_0, \mathrm{d}x_1) = \Lambda_\varepsilon \exp((f(x_0) + g(x_1) - \tfrac{1}{2}\|x_0 - x_1\|^2)/\varepsilon)\mu_0(\mathrm{d}x_0)\mu_1(\mathrm{d}x_1)\,, \tag{4.19}$$

for any *measurable f and g and* any *probability measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Let M the path measure*

given by a mixture of Brownian bridges with respect to (4.19) *as in* (4.13)*, with* t*-marginals* m_t *for* $t \in [0,1]$*. The following hold:*

1. *The path measure* M *is Markov;*

2. *The marginal* m_t *is given by*

$$\mathsf{m}_t(\mathrm{d}z) = \mathsf{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z)\mathsf{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z)\mathrm{d}z\,;$$

3. M *is the law of the solution to the SDE*

$$\mathrm{d}X_t = \varepsilon\nabla \log \mathsf{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](X_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,, \quad X_0 \sim \mu_0\,;$$

4. *The drift above can be expressed as* $b_t(z) = (1-t)^{-1}(z - \nabla\varphi_{1-t}(z))$*, where* $\nabla\varphi_{1-t}$ *is the entropic Brenier map between* m_t *and* ρ *with regularization strength* $(1-t)\varepsilon$*, where*

$$\rho(\mathrm{d}x_1) = \mu_1(\mathrm{d}x_1)\exp\big(g(x_1)/\varepsilon + \log \mathsf{H}_\varepsilon[e^{f/\varepsilon}\mu_0](x_1)\big)\,.$$

   If (4.19) *is the* optimal *entropic coupling between* μ_0 *and* μ_1*, then* $\rho \equiv \mu_1$*.*

4.3.2 Defining the estimator

In light of (4.17), it is easy to define an estimator on the basis of samples. Let $X_1, \ldots, X_m \sim \mu$ and $Y_1, \ldots, Y_n \sim \nu$, and let $\mu_m := m^{-1}\sum_{i=1}^m \delta_{X_i}$, and similarly $\nu_n := n^{-1}\sum_{j=1}^n \delta_{Y_j}$. Let $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ be the optimal entropic potentials associated with $\mathrm{OT}_\varepsilon(\mu_m, \nu_n)$, which can be computed efficiently via Sinkhorn's algorithm (Cuturi, 2013; Peyré and Cuturi, 2019) with a runtime of $O(mn/\varepsilon)$ (Altschuler

et al., 2017). A natural plug-in estimator for the optimal drift is thus

$$\begin{aligned}\hat{b}_t(z) &:= \varepsilon \nabla \log \mathsf{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n] \\ &= (1-t)^{-1}\Big(-z + \frac{\sum_{j=1}^n Y_j \exp\big((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon\big)}{\sum_{j=1}^n \exp\big((\hat{g}_j - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon\big)}\Big) \\ &=: (1-t)^{-1}(-z + \nabla\hat{\varphi}_{1-t}(z))\end{aligned} \tag{4.20}$$

Further discussions on the numerical aspects of our estimator are deferred to Section 4.5. Since we want to estimate the path given by $\mathsf{P}^\star$, our estimator is given by the solution to the following SDE:

$$\mathrm{d}\hat{X}_t = (-(1-k\eta)^{-1}\hat{X}_{k\eta} + (1-k\eta)^{-1}\nabla\hat{\varphi}_{1-k\eta}(\hat{X}_{k\eta}))\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,, \tag{4.21}$$

for $t \in [k\eta, (k+1)\eta]$, where $\eta \in (0,1)$ is some step-size, and k is the iteration number. Though it is convenient to write the drift in terms of a time-varying entropic Brenier map, (4.20) shows that for all $t \in (0,1)$, our estimator is a simple function of the potential $\hat{g}$ obtained from a single call to Sinkhorn's algorithm.

Remark 4.2. To the best of our knowledge, the idea of using static potentials to estimate the SB drift was first explored by Finlay et al. (2020a). However, their proposal had some inconsistencies. For example, they assume a finite entropy condition on the source and target measures, and perform a standard Gaussian convolution on $\mathbb{R}^d$ instead of our proposed convolution $\mathsf{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n]$. The former leads to a computationally intractable estimator, whereas, as we have shown above, the former has a simple form that is trivial to compute.

Remark 4.3. An alternative approach to computing the Schrödinger bridge is due to Stromme (2023): Given n samples from the source and target measure, one can efficiently compute the in-sample entropic optimal coupling $\hat{\pi}$ on the basis of samples via Sinkhorn's algorithm. Resampling a pair $(X', Y') \sim \hat{\pi}$ and computing the Brownian bridge between X' and Y' yields an approximate sample from the Schrödinger bridge. We remark that the computational complexity of our approach is

significantly lower than that of Stromme (2023). While both methods use Sinkhorn's algorithm to compute an entropic optimal coupling between the source and target measures, Stromme's estimator necessitates n *fresh* samples from μ and ν to obtain a single approximate sample from the SB. By contrast, having used our method to estimate the drifts, fresh samples from μ can be used to generate unlimited approximate samples from the SB.

4.4 Main results and proof sketch

We now present the proof sketches to our main result. We first present a sketch focusing purely on the statistical error incurred by our estimator, and later, using standard tools (Chen et al., 2022b; Lee et al., 2023), we incorporate the additional time-discretization error. All omitted proofs in this section are deferred to Appendix C.3.

4.4.1 Statistical analysis

We restrict our analysis to the one-sample estimation task, as it is the closest to real-world applications where the source measure is typically known (e.g., the standard Gaussian) and the practitioner is given finitely many samples from a distribution of interest (e.g., images). Thus, we assume full access to μ and access to ν through i.i.d. data, and let $(\hat{f}, \hat{g})$ correspond to the optimal entropic potentials solving $\mathrm{OT}_\varepsilon(\mu, \nu_n)$, which give rise to an optimal entropic plan π_n. Formally, this corresponds to the $m \to \infty$ limit of the setting described in Section 4.3.2; the estimator for the drift (4.20) is unchanged.

Let $\tilde{\mathrm{P}}$ be the Markov measure associated with the mixture of Brownian bridges defined with respect to π_n. By Proposition 4.1, the t-marginals are given by

$$\tilde{\mathrm{p}}_t(z) = \mathsf{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n](z)\mathsf{H}_{t\varepsilon}[\exp(\hat{f}/\varepsilon)\mu](z)\,, \tag{4.22}$$

and the one-sample empirical drift is equal to

$$\hat{b}_t(z) = \varepsilon \nabla \log \mathsf{H}_{(1-t)\varepsilon}[\exp(\hat{g}/\varepsilon)\nu_n](z)\,.$$

Thus, $\tilde{\mathsf{P}}$ is the law of the following process with $\tilde{X}_0 \sim \mu$

$$\mathrm{d}\tilde{X}_t = \hat{b}_t(\tilde{X}_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t\,. \tag{4.23}$$

Note that this agrees with our estimator in (4.21), but without discretization. This process is not technically implementable, but forms an important theoretical tool in our analysis.

Our main result of this section is the following theorem.

Theorem 4.4 (One-sample estimation; no discretization). *Suppose both $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and ν is supported on a k-dimensional smooth submanifold of $\mathbb{R}^d$ whose support is contained in a ball of radius $R > 0$. Let $\tilde{\mathsf{P}}$ (resp. P) be the path measure corresponding to (4.23) (resp. (4.17)). Then it holds that, for any $\tau \in [0, 1)$,*

$$\mathbb{E}[\mathrm{TV}^2(\tilde{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \Big(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2\varepsilon^{-\mathsf{k}}}{(1-\tau)^{\mathsf{k}+2}n}\Big)\,.$$

As mentioned in the introduction, the parametric rates will not be surprising given the proof sketch below, which incorporates ideas from entropic optimal transport. The rates diverge exponentially in k as $\tau \to 1$; this is a consequence of the fact that the estimated drift $\hat{b}_t$ enforces that the samples exactly collapse onto the training data at terminal time, which is far from the true target measure.

The proof of Theorem 4.4 uses key ideas from Stromme (2024): We introduce the following entropic plan

$$\bar{\pi}_n(x, y) := \Lambda_\varepsilon \exp\big((\bar{f}(x) + g^\star(y) - \tfrac{1}{2}\|x-y\|^2)/\varepsilon\big)\mu(\mathrm{d}x)\nu_n(\mathrm{d}y)\,, \tag{4.24}$$

where $g^\star$ is the optimal entropic potential for the population measures (μ, ν), and where we call $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ a *rounded* potential, defined as

$$\bar{f}(x) \coloneqq -\varepsilon \log\Big(\Lambda_\varepsilon \cdot n^{-1} \sum_{j=1}^{n} \exp((g^\star(Y_j) - \tfrac{1}{2}\|x - Y_j\|^2)/\varepsilon)\Big)\,.$$

Note that $\bar{f}$ can be viewed as the Sinkhorn update involving the potential $g^\star$ and measure ν_n, and that $\bar{\pi}_n \in \Gamma(\mu, \bar{\nu}_n)$, where $\bar{\nu}_n$ is a rescaled version of ν_n. We again exploit Proposition 4.1. Consider the path measure associated to the mixture of Brownian bridges with respect to $\bar{\pi}_n$, denoted $\bar{\mathrm{P}}$ (with t-marginals $\bar{\mathrm{p}}_t$), which corresponds to an SDE with drift

$$\begin{aligned} \bar{b}_t(z) &= \varepsilon \nabla \log \mathrm{H}_{1-t}[\exp(g^\star/\varepsilon)\nu_n](z) \\ &= (1-t)^{-1}\Big(-z + \frac{\sum_{j=1}^N Y_j \exp((g^\star(Y_j) + \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)}{\sum_{j=1}^N \exp((g^\star(Y_j) + \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)}\Big)\,. \end{aligned} \tag{4.25}$$

Introducing the path measure $\bar{\mathrm{P}}_{[0,\tau]}$ into the bound via triangle inequality and then applying Pinsker's inequality, we arrive at

$$\begin{aligned} \mathbb{E}[\mathrm{TV}^2(\tilde{\mathrm{P}}_{[0,\tau]}, \mathrm{P}^\star_{[0,\tau]})] &\lesssim \mathbb{E}[\mathrm{TV}^2(\tilde{\mathrm{P}}_{[0,\tau]}, \bar{\mathrm{P}}_{[0,\tau]})] + \mathbb{E}[\mathrm{TV}^2(\bar{\mathrm{P}}_{[0,\tau]}, \mathrm{P}^\star_{[0,\tau]})] \\ &\lesssim \mathbb{E}[\mathrm{KL}(\tilde{\mathrm{P}}_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] + \mathbb{E}[\mathrm{KL}(\mathrm{P}^\star_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})]\,, \end{aligned}$$

We analyse the two terms separately, each term involving proof techniques developed by Stromme (2024). We summarize the results in the following propositions, which yield the proof of Theorem 4.4.

Proposition 4.5. *Assume the conditions of Theorem 4.4, then for any* $\tau \in [0, 1)$

$$\mathbb{E}[\mathrm{KL}(\tilde{\mathrm{P}}_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] \le \frac{1}{\varepsilon}\mathbb{E}[\mathrm{OT}_\varepsilon(\mu, \nu_n) - \mathrm{OT}_\varepsilon(\mu, \nu)] \le \varepsilon^{-(\mathrm{k}/2+1)} n^{-1/2}\,.$$

Proposition 4.6. *Assume the conditions of Theorem 4.4, then*

$$\mathbb{E}[\mathrm{KL}(\mathrm{P}^{\star}_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] \leq \frac{R^2 \varepsilon^{-\mathrm{k}}}{n} (1-\tau)^{-\mathrm{k}-2} .$$

4.4.2 Completing the results

We now incorporate the discretization error. Letting $\hat{\mathrm{P}}$ denote the path measure induced by the dynamics of (4.21), we use the triangle inequality to introduce the path measure $\tilde{\mathrm{P}}$:

$$\mathbb{E}[\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \mathrm{P}^{\star}_{[0,\tau]})] \lesssim \mathbb{E}[\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \tilde{\mathrm{P}}_{[0,\tau]})] + \mathbb{E}[\mathrm{TV}^2(\tilde{\mathrm{P}}_{[0,\tau]}, \mathrm{P}^{\star}_{[0,\tau]})] .$$

The second term is precisely the statistical error, controlled by Theorem 4.4. For the first term, we employ a now-standard discretization argument (see e.g., Chen et al. (2022b)) which bounds the total variation error as a function of the step-size parameter η and the Lipschitz constant of the empirical drift, which can be easily bounded in our setting.

Proposition 4.7. *Suppose $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Denoting L_τ for the Lipschitz constant of $\hat{b}_\tau$ (recall (4.20)) for $t \in [0,1)$ and η the step-size of the SDE discretization, it holds that*

$$\mathbb{E}[\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \tilde{\mathrm{P}}_{[0,\tau]})] \lesssim (\varepsilon+1) L_\tau^2 d\eta .$$

In particular, if $\mathrm{supp}(\nu) \subseteq B(0;R)$, then

$$\mathbb{E}[\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \tilde{\mathrm{P}}_{[0,\tau]})] \lesssim (\varepsilon+1)(1-\tau)^{-2} d\eta (1 \vee R^4 (1-\tau)^{-2} \varepsilon^{-2}) .$$

We now aggregate the statistical and approximation error into one final result.

Theorem 4.8. *Suppose $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathrm{supp}(\nu) \subseteq B(0,R) \subseteq \mathcal{M}$, where $\mathcal{M}$ is a k-dimensional submanifold of $\mathbb{R}^d$. Given n i.i.d. samples from ν, the one-sample Sinkhorn bridge $\hat{\mathrm{P}}$ estimates the*

Schrödinger bridge $\mathsf{P}^\star$ *with the following error*

$$\mathbb{E}[\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \Big(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2\varepsilon^{-\mathsf{k}}}{(1-\tau)^{\mathsf{k}+2}n}\Big) + (\varepsilon+1)(1-\tau)^{-2}d\eta(1 \vee R^4(1-\tau)^{-2}\varepsilon^{-2})\,.$$

Assuming $R \geq 1$ *and* $\varepsilon = 1$*, the Schrödinger bridge can be estimated in total variation distance to accuracy* ϵ_{TV} *with* n *samples and* N *Euler–Maruyama steps, where*

$$n \asymp \frac{R^2}{(1-\tau)^{\mathsf{k}+2}\epsilon_{\mathrm{TV}}^2} \vee \epsilon_{\mathrm{TV}}^{-4}, \quad N \lesssim \frac{dR^4}{(1-\tau)^4\epsilon_{\mathrm{TV}}^2}\,.$$

Note that our error rates improve as $\varepsilon \to \infty$; since this is also the regime in which Sinkhorn's algorithm terminates rapidly, it is natural to suppose that ε should be large in practice. This is misleading, however: as ε grows, the Schrödinger bridge becomes less and less informative,[7] and the marginal $\mathsf{p}^\star_\tau$ only resembles ν when τ becomes very close to 1. We elaborate on the use of the SB for sampling in the following section.

4.4.3 Application: Sampling with the Föllmer bridge

Theorem 4.8 does not immediately imply guarantees for sampling from the target distribution ν. Obtaining such guarantees requires arguing that simulating the Sinkhorn bridge on a suitable interval $[0, \tau]$ for τ close to 1 yields samples close to the true density (without completely collapsing onto the training data). We provide such a guarantee in this section, for the special case of the Föllmer bridge. We adopt this setting only for concreteness; similar arguments apply more broadly.

The Föllmer bridge is a special case of the Schrödinger bridge due to Hans Föllmer (Föllmer, 1985). In this setting, $\mu = \delta_a$ for any $a \in \mathbb{R}^d$, and our estimator takes a particularly simple form:

$$\hat{b}^{\mathsf{F}}_t(z) = (1-t)^{-1}\Big(-z + \frac{\sum_{j=1}^n Y_j \exp\big((\frac{1}{2}\|Y_j\|^2 - \frac{1}{2(1-t)}\|z-Y_j\|^2)/\varepsilon\big)}{\sum_{j=1}^n \exp\big((\frac{1}{2}\|Y_j\|^2 - \frac{1}{2(1-t)}\|z-Y_j\|^2)/\varepsilon\big)}\Big), \tag{4.26}$$

[7] In other words, the transport path is more and more volatile.

Note that in this special case, calculating the drift does not require the use of Sinkhorn's algorithm, and the drift, in fact, corresponds to the score of a kernel density estimator applied to ν_n. We provide a calculation of these facts in Appendix C.4 for completeness.

We then have the following guarantee.

Corollary 4.9. *Consider the assumptions of Theorem 4.8, further suppose that $\mu = \delta_0$ and $\varepsilon = 1$ and that the second moment of ν is bounded by d. Suppose we use n samples from ν to estimate the Föllmer drift, and simulate the resulting SDE using N Euler–Maruyama iterations until time $\tau = 1 - \epsilon_{\mathrm{W}_2}^2/d$, with*

$$n \asymp \frac{R^2 d^{\mathrm{k}+2}}{\epsilon_{\mathrm{W}_2}^{2k+4}\epsilon_{\mathrm{TV}}^2} \vee \epsilon_{\mathrm{TV}}^{-4} \qquad N \lesssim \frac{R^4 d^5}{\epsilon_{\mathrm{W}_2}^8 \epsilon_{\mathrm{TV}}^2}\,.$$

Then the density given by the Sinkhorn bridge at time τ iterations will be ϵ_{TV}-close in total variation to a measure which is ϵ_{W_2}-close to ν in the 2-Wasserstein distance.

Note that the choice $\varepsilon = 1$ was merely out of convenience. If instead the practitioner was willing to pay the computational price of solving Sinkhorn's algorithm for small ε and large n, then the number of requisite iterations N would decrease. Finally, notice that the number of samples scales exponentially in the intrinsic dimension $\mathrm{k} \ll d$ instead of the ambient dimension d. This is, of course, unavoidable, but improves upon recent work that uses kernel density estimators to prove a similar result for denoising diffusion probabilistic models (Wibisono et al., 2024).

Remark 4.10. Recently, Huang (2024) also proposed (4.26) to estimate the Föllmer drift. They provide no statistical estimation guarantees of the drift, nor any sampling guarantees; their contributions are largely empirical, demonstrating that the proposed estimator is tractable for high-dimensional tasks. The work of Huang et al. (2021b) also proposes an estimator for the Föllmer bridge based on having partial access to the log-density ratio of the target distribution (without the normalizing constant).

Algorithm 1: Sinkhorn bridges

Input: Data $\{X_i\}_{i=1}^m \sim \mu$, $\{Y_j\}_{j=1}^n \sim \nu$, parameters $\varepsilon > 0$, $\tau \in (0,1)$, and $N \geq 1$

Compute: Sinkhorn potentials $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$; // Using POT or OTT

Initialize: $x^{(0)} = x \sim \mu$, $k = 0$, stepsize $\eta = \tau/N$

while $k \leq N - 1$ **do**

  $x^{(k+1)} = x^{(k)} + \eta \hat{b}_{k\eta}(x^{(k)}) + \sqrt{\eta\varepsilon}\xi$; // $\xi \sim \mathcal{N}(0, I)$

  $k \leftarrow k + 1$

end

return $x^{(N)}$

4.5 Numerical performance

Our approach is summarized in Algorithm 1, and open-source code for replicating our experiments is available at https://github.com/APooladian/SinkhornBridge.[8]

For a fixed regularization parameter $\varepsilon > 0$, the runtime of computing $(\hat{f}, \hat{g})$ on the basis of samples has complexity $O(mn/(\varepsilon\delta_{\text{tol}}))$, where δ_{tol} is a required tolerance parameter that measures how closely the the marginal constraints are satisfied (Altschuler et al., 2022; Cuturi, 2013; Peyré and Cuturi, 2019). Once these are computed, the evaluation of $\hat{b}_{k\eta}$ is $O(n)$, with the remaining runtime being the number of iteration steps, denoted by N. In all our experiments, we take $m = n$, thus the total runtime complexity of the algorithm is a fixed cost of $O(n^2/(\varepsilon\delta_{\text{tol}}))$, followed by $O(nN)$ for each new sample to be generated (which can be parallelized).

4.5.1 Qualitative illustration

As a first illustration, we consider standard two-dimensional datasets from the machine learning literature. For all examples, we use $n = 2000$ training points from both the source and target measure, and run Sinkhorn's algorithm with $\varepsilon = 0.1$. For generation, we set $\tau = 0.9$, and consider $N = 50$ Euler–Maruyama steps. Figure 4.1 contains the resulting simulations, starting from out-of-sample points. We see reasonable performance in each case.

[8] Our estimator is implemented in both the POT and OTT-JAX frameworks.

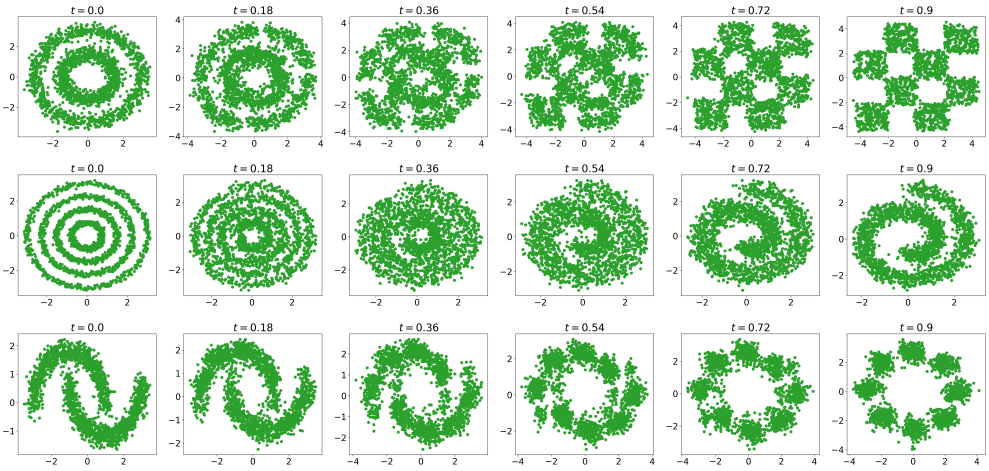


Figure 4.1: Schrödinger bridges on the basis of samples from toy datasets.

4.5.2 Quantitative illustrations

We quantitatively assess the performance of our estimator using synthetic examples from the deep learning literature (Bunne et al., 2023a; Gushchin et al., 2023).

The Gaussian case We first demonstrate that we are indeed learning the drift and that the claimed rates are empirically justified. As a first step, we consider the simple case where $\mu = \mathcal{N}(a, A)$ and $\nu = \mathcal{N}(b, B)$ for two positive-definite $d \times d$ matrices A and B and arbitrary vectors $a, b \in \mathbb{R}^d$. In this regime, the optimal drift $b_\tau^\star$ and $\mathrm{p}_\tau^\star$ has been computed in closed-form by Bunne et al. (2023a); see equations (25)-(29) in their work.

To verify that we are indeed learning the drift, we first draw n samples from μ and ν, and compute our estimator, $\hat{b}_\tau$ for any $\tau \in [0, 1)$. We then evaluate the mean-squared error

$$\mathrm{MSE}(n, \tau) = \|\hat{b}_\tau - b_\tau^\star\|^2_{L^2(\mathrm{p}_\tau^\star)},$$

by a Monte Carlo approximation, with $n_{\mathrm{MC}} = 10000$. For simplicity, with $d = 3$, we choose $A = I$ and randomly generate a positive-definite matrix B, and center the Gaussians. We fix $\varepsilon = 1$ and vary n used to define our estimator, and perform the simulation ten times to generate error bars

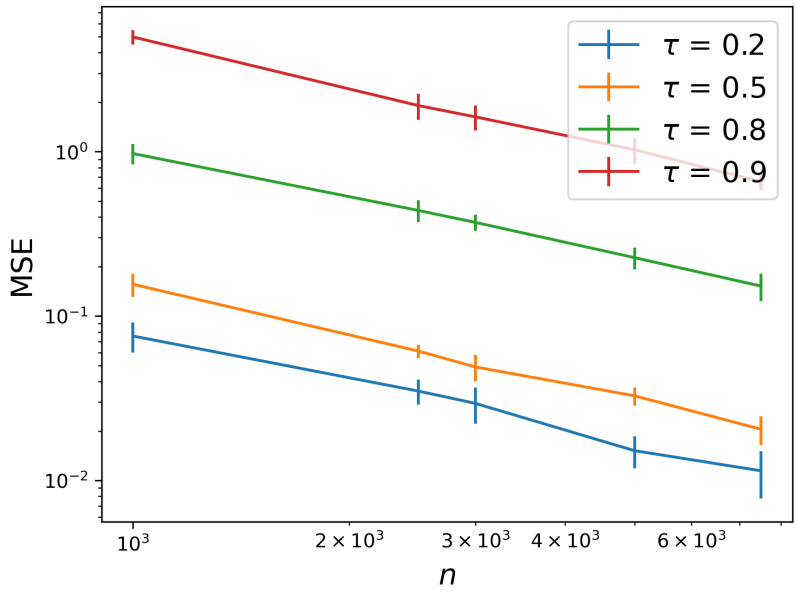


Figure 4.2: MSE for estimating the Gaussian drift as (n, τ) vary, averaged over 10 trials.

across various choices of $\tau \in [0, 1)$; see Figure 4.2.

It is clear from the plot that the *constant* associated to the rate of estimation gets worse as $\tau \to 1$, but the overall rate of convergence appears unchanged, which hovers around n^{-1} for all choices of τ shown in the plot, as expected from e.g., Proposition 4.5.

MULTIMODAL MEASURES WITH CLOSED-FORM DRIFT The next setting is due to Gushchin et al. (2023); they devised a drift that defines the Schrödinger bridge between a Gaussian and a more complicated measure with multiple modes. This explicit drift allowed them to benchmark multiple neural network based methods for estimating the Schrödinger bridge for non-trivial couplings (e.g., beyond the Gaussian to Gaussian setting). We briefly remark that the approaches discussed in their work fall under the "continuous estimation" paradigm, where researchers assume they can endlessly sample from the distributions when training (using new samples per training iteration).

We consider the same pre-fixed drift as found in their publicly available code, which transports the standard Gaussian to a distribution with four modes. We consider the case $d = 64$ and $\varepsilon = 1$, as these hyperparameters are most extensively studied in their work, where they provide the most

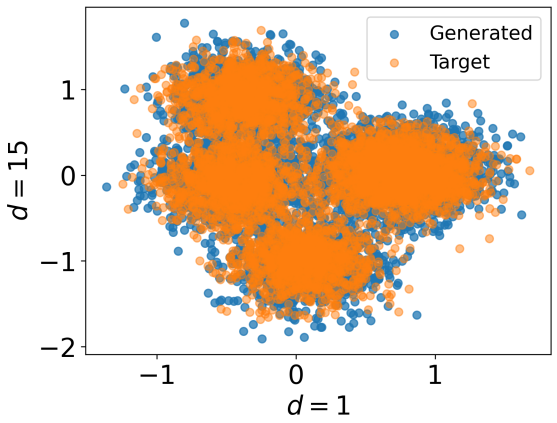


Figure 4.3: Plotting generated and resampled target data in $d = 64$.

Method	BW-UVP
Ours	**0.41 ± 0.03**
MLE-SB	0.56
EgNOT	0.85
FB-SDE-A	0.65

Table 4.1: Comparison to neural network approaches in BW-UVP for $d = 64$.

details on the other models. We use $n = 4096$ training samples from the source and target data they construct (which is significantly less than the total number of samples required for any of the neural network based models) and perform our estimation procedure, and we take $N = 100$ discretization steps (which is half as many as most of the works they consider) to simulate to time $\tau = 0.99$. To best illustrate the four mixture components, Figure 4.3 contains a scatter plot of the first and fifteenth dimension, containing fresh target samples and our generated samples.

We compare to the ground-truth samples using the unexplained variance percentage (UVP) based on the Bures–Wasserstein distance (Bures, 1969):

$$\mu \mapsto \text{BW-UVP}_\nu(\mu) := 100 \frac{\text{BW}^2(\mathcal{N}_\mu, \mathcal{N}_\nu)}{0.5 \cdot \text{Var}(\nu)},$$

where $\mathcal{N}_\mu = \mathcal{N}(\mathbb{E}_\mu[X], \text{Cov}_\mu(X))$, and same for $\mathcal{N}_\nu$.[9] While seemingly ad hoc, the BW-UVP is widely used in the machine learning literature as a means of quantifying the quality of the generated samples (see e.g., Daniels et al. (2021)). We compute the BW-UVP with 10^4 generated samples from the target and our approach, averaged over 5 trials, and used the results of Gushchin et al. (2023) for the remaining methods (MLE-SB is by Vargas et al. (2021), EgNOT is by Mokrov et al.

[9]For us, these quantities are computed on the basis of samples.

(2023), and FB-SDE-A is by Chen et al. (2021a)). We see that the Sinkhorn bridge has significantly lower BW-UVP compared to the other approaches while requiring less compute resources and training data.

Part II

Interlude: Theoretical properties of entropic Brenier maps

5 An entropic generalization of Caffarelli's contraction theorem via covariance inequalities

5.1 Introduction

The following seminal result is due to Caffarelli (2000).

Theorem 5.1 (Caffarelli's contraction theorem)**.** *Let $P = \exp(-V)$ and $Q = \exp(-W)$ have smooth densities on $\mathbb{R}^d$, with $\nabla^2 V \preceq \beta_V I$ and $\nabla^2 W \succeq \alpha_W I \succ 0$. Then, the optimal transport map $\nabla\varphi_0$ from P to Q is $\sqrt{\beta_V/\alpha_W}$-Lipschitz.*

Here, $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$ is a convex function, known as a *Brenier potential*. The optimal transport map $\nabla\varphi_0 : \mathbb{R}^d \to \mathbb{R}^d$ pushes forward P to Q, in the sense that if X is a random variable with law P, then $\nabla\varphi_0(X)$ is a random variable with law Q. See Section 1.3.1 and the textbook by Villani (2021) for background on optimal transport.

Caffarelli's contraction theorem can be used to transfer functional inequalities, such as a Poincaré inequality, from the standard Gaussian measure on $\mathbb{R}^d$ to other probability measures (Bakry et al., 2014). Towards this end, recent works have also constructed and studied alternative Lipschitz transport maps (e.g. Kim and Milman (2012); Mikulincer and Shenfeld (2023;

2024); Neeman (2022)), but still the properties of the original optimal transport map remain of fundamental interest, with many questions unresolved (Colombo et al., 2017; Valdimarsson, 2007).

Indeed, besides the application to functional inequalities, the structural properties of optimal transport maps play a fundamental role in theoretical and methodological advances in optimal transport, such as the control of the curvature of the Wasserstein space through the notion of extendible geodesics (Ahidar-Coutrix et al., 2020; Le Gouic et al., 2022), the stability of Wasserstein barycenters (Chewi et al., 2020), and the statistical estimation of optimal transport maps (Hütter and Rigollet, 2021).

In applied domains, however, the inauspicious computational and statistical burden of solving the original optimal transport problem has instead led practitioners to consider *entropically regularized* optimal transport, as pioneered by Cuturi (2013). In addition to its practical merits, entropic optimal transport enjoys a rich mathematical theory, rooted in its connection to the classical Schrödinger bridge problem (Léonard, 2014), which has led to powerful applications to high-dimensional probability (Fathi et al., 2020; Gentil et al., 2020; Ledoux, 2018). As such, it is natural to study the properties of the entropic analogue of the optimal transport map.

5.1.1 Contributions

In this chapter, we prove a generalization of Caffarelli's contraction theorem to the setting of entropic optimal transport. Namely, we study the Hessian of the *entropic Brenier potential* which admits a representation as a covariance matrix (Lemma 5.4). By applying two well-known inequalities for covariance matrices (the Brascamp–Lieb inequality and the Cramér–Rao inequality), we quickly deduce a sharp upper bound on the operator norm of the Hessian which holds for any value $\varepsilon > 0$ of the regularization parameter.

As a byproduct of our analysis, by sending $\varepsilon \searrow 0$ and appealing to recent convergence results for the entropic Brenier potentials (Nutz and Wiesel, 2021), we obtain the shortest proof of Caffarelli's contraction theorem to date. Notably, our argument allows us to sidestep the regularity

of the optimal transport map, which is a key obstacle in Caffarelli's original proof.

Recently, Fathi et al. (2020) gave a proof of Caffarelli's theorem using a surprising equivalence between Theorem 5.1 and a statement about Wasserstein projections, which was discovered through the theory of weak optimal transport (Gozlan and Juillet, 2020). In order to verify the latter, their proof also used ideas from entropic optimal transport. In comparison, we note that our argument is more direct and also allows us to handle the case of non-zero regularization ($\varepsilon > 0$).

To further demonstrate the applicability of our technique, in Section 5.4 we prove a generalization of Caffarelli's result: if $\nabla^2 V \preceq A^{-1}$ and $\nabla^2 W \succeq B^{-1}$, where A and B are arbitrary commuting positive definite matrices, then the Hessian of the Brenier potential from P to Q is pointwise upper bounded (in the PSD ordering) by $A^{-1/2}B^{1/2}$. This result implies a remarkable extremal property of optimal transport maps between Gaussian measures, namely: the optimal transport map from $\mathcal{N}(0, A)$ to $\mathcal{N}(0, B)$ maximizes the Hessian of the Brenier potential at any point among all possible measures P and Q satisfying our assumptions. To the best of our knowledge, this result is new.

5.2 Background

5.2.1 Assumptions

Henceforth, we say that the pair (P, Q) satisfies our regularity conditions if:

1. P has full support on $\mathbb{R}^d$ and Q is supported on a convex subset of $\mathbb{R}^d$. Let Ω_Q denote the interior of the support of Q, so that Ω_Q is a convex open set.

2. P and Q admit positive Lebesgue densities on $\mathbb{R}^d$ and Ω_Q, which we can therefore be written $\exp(-V)$ and $\exp(-W)$ respectively for functions $V, W : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. We abuse notation and identify the measures with their densities, thus writing $P = \exp(-V)$ and $Q = \exp(-W)$.

3. We assume that V and W are twice continuously differentiable on $\mathbb{R}^d$ and Ω_Q respectively.

Some of these assumptions can be eventually relaxed, but they suffice for the purposes of this work. Throughout the rest of the paper and for the sake of simplicity, these regularity assumptions are assumed to hold for the probability measures under consideration.

5.2.2 Optimal transport without regularization

Let P and Q be probability measures with finite second moment. The *optimal transport problem* is the optimization problem

$$\underset{\pi \in \Pi(P,Q)}{\text{minimize}} \quad \iint \tfrac{1}{2} \, \|x - y\|^2 \, \mathrm{d}\pi(x, y) \tag{5.1}$$

where $\Pi(P, Q)$ is the set of joint probability measures with marginals P and Q. The following fundamental result characterizes the optimal solution to (5.1).

Theorem 5.2 (Brenier's theorem)**.** *Suppose that P admits a density with respect to Lebesgue measure. Then, there exists a proper, convex, lower semicontinuous function $\varphi_0 : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that the optimal transport plan in (5.1) can be written $\pi_0 = (\mathrm{id}, \nabla\varphi_0)_\sharp P$. The function φ_0 is called the* Brenier potential*, and the mapping $\nabla\varphi_0$ is called the* optimal transport map *from P to Q. Moreover, the optimal transport map $\nabla\varphi_0$ is unique up to P-almost everywhere equality.*

The Brenier potential φ_0 is obtained as the solution to the dual problem

$$\underset{\varphi \in \Gamma_0}{\text{maximize}} \quad \int \Big(\frac{\|\cdot\|^2}{2} - \varphi \Big) \, \mathrm{d}P + \int \Big(\frac{\|\cdot\|^2}{2} - \varphi^* \Big) \, \mathrm{d}Q \,, \tag{5.2}$$

where φ^* is the convex conjugate to φ, and Γ_0 is the set of proper, convex, lower semicontinuous functions on $\mathbb{R}^d$.

We refer to Villani (2021) for further background.

5.2.3 Optimal transport with entropic regularization

Entropic optimal transport is the problem that arises when we add the Kullback–Liebler (KL) divergence, denoted $\mathrm{KL}(\cdot \,\|\, \cdot)$, as a regularizer to (5.1):

$$\underset{\pi \in \Pi(P,Q)}{\text{minimize}} \quad \iint \tfrac{1}{2} \, \|x - y\|^2 \, \mathrm{d}\pi(x,y) + \varepsilon \, \mathrm{KL}(\pi \| P \otimes Q) \,. \tag{5.3}$$

The following statement characterizes the solution to (5.3) (Csiszár, 1975; Nutz and Wiesel, 2021; Peyré and Cuturi, 2019).

Theorem 5.3 (Entropic optimal transport)**.** *Let P and Q be probability measures on $\mathbb{R}^d$ and fix $\varepsilon > 0$. Then there exists a unique solution $\pi_\varepsilon \in \Pi(P,Q)$ to (5.3). Moreover, π_ε has the form*

$$\pi_\varepsilon(\mathrm{d}x, \mathrm{d}y) = \exp\Big(\frac{f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2} \, \|x - y\|^2}{\varepsilon} \Big) \, P(\mathrm{d}x) \, Q(\mathrm{d}y) \,, \tag{5.4}$$

where $(f_\varepsilon, g_\varepsilon)$ are maximizers for the dual problem

$$\underset{(f,g) \in L^1(P) \times L^1(Q)}{\text{maximize}} \quad \int f \, \mathrm{d}P + \int g \, \mathrm{d}Q - \varepsilon \iint e^{(f(x) + g(y) - \frac{1}{2} \|x-y\|^2)/\varepsilon} \, \mathrm{d}P(x) \, \mathrm{d}Q(y) + \varepsilon \,. \tag{5.5}$$

The constraint that π_ε has marginals P and Q implies the following dual optimality conditions for $(f_\varepsilon, g_\varepsilon)$ (see Mena and Niles-Weed (2019); Nutz and Wiesel (2021)):

$$f_\varepsilon(x) = -\varepsilon \log \int e^{(g_\varepsilon(y) - \frac{1}{2} \|x-y\|^2)/\varepsilon} \, \mathrm{d}Q(y) \qquad (x \in \mathbb{R}^d) \,, \tag{5.6}$$

$$g_\varepsilon(y) = -\varepsilon \log \int e^{(f_\varepsilon(x) - \frac{1}{2} \|x-y\|^2)/\varepsilon} \, \mathrm{d}P(x) \qquad (y \in \mathbb{R}^d) \,. \tag{5.7}$$

In particular, f_ε and g_ε are smooth. In this work, it is more convenient to work with the *entropic*

Brenier potentials, defined as

$$(\varphi_\varepsilon, \psi_\varepsilon) := (\tfrac{1}{2}\|\cdot\|^2 - f_\varepsilon,\ \tfrac{1}{2}\|\cdot\|^2 - g_\varepsilon)\,. \tag{5.8}$$

Since $(f_\varepsilon, g_\varepsilon)$ are only unique up to adding a constant to f_ε and subtracting the same constant from g_ε, we fix the normalization convention $\int f_\varepsilon \,\mathrm{d}P = \int g_\varepsilon \,\mathrm{d}Q$. Under this condition, it was shown by Nutz and Wiesel (2021) that we have convergence to the Brenier potential $\varphi_\varepsilon \longrightarrow \varphi_0$ as $\varepsilon \searrow 0$.

Adopting this new notation, with $P = \exp(-V)$ and $Q = \exp(-W)$, we can rewrite the entropic optimal plan as

$$\pi_\varepsilon(\mathrm{d}x, \mathrm{d}y) = \exp\Big(-\frac{\varphi_\varepsilon(x) + \psi_\varepsilon(y) - \langle x, y\rangle}{\varepsilon} - V(x) - W(y)\Big)\,\mathrm{d}x\,\mathrm{d}y\,.$$

The entropic Brenier potentials were first introduced to develop a computationally tractable estimator of the optimal transport map $\nabla\varphi_0$ (Pooladian et al., 2022; Pooladian and Niles-Weed, 2021; Seguy et al., 2018). Indeed, this is motivated by the following observation, which acts as an entropic version of Brenier's theorem. Write $\pi_\varepsilon^{Y|X=x}$ for the conditional distribution of Y given $X = x$ for $(X, Y) \sim \pi_\varepsilon$, and similarly define $\pi_\varepsilon^{X|Y=y}$. Then, by Pooladian and Niles-Weed (2021, Proposition 2), $\nabla\varphi_\varepsilon$ is the barycentric projection

$$\nabla\varphi_\varepsilon(x) = \int y \,\mathrm{d}\pi_\varepsilon^{Y|X=x}(y)\,. \tag{5.9}$$

For clarity of exposition, we abuse notation and abbreviate $\pi_\varepsilon^{Y|X=x}$ by π_ε^x and $\pi_\varepsilon^{X|Y=y}$ by π_ε^y when there is no danger of confusion.

The following lemma is a straightforward computation using (5.4), (5.6), and (5.7).

Lemma 5.4. *It holds that*

$$\nabla^2\varphi_\varepsilon(x) = \varepsilon^{-1}\operatorname{Cov}_{Y\sim\pi_\varepsilon^x}(Y)\,, \qquad \textit{and} \qquad \nabla^2\psi_\varepsilon(y) = \varepsilon^{-1}\operatorname{Cov}_{X\sim\pi_\varepsilon^y}(X)\,.$$

In particular, both φ_ε and ψ_ε are convex. Moreover, under our regularity conditions,

$$\nabla_y^2 \log(1/\pi_\varepsilon^x)(y) = \varepsilon^{-1}\,\nabla^2\psi_\varepsilon(y) + \nabla^2 W(y)\,,$$
$$\nabla_x^2 \log(1/\pi_\varepsilon^y)(x) = \varepsilon^{-1}\,\nabla^2\varphi_\varepsilon(x) + \nabla^2 V(x)\,.$$

5.2.4 Two covariance inequalities

In our proofs, we make use of the following key inequalities.

Lemma 5.5. *Let $P = \exp(-V)$ be a probability measure on $\mathbb{R}^d$ and assume that V is twice continuously differentiable on the interior of its domain. Then, the following hold.*

1. *(Brascamp–Lieb inequality) If in addition we assume that P is strictly log-concave, then it holds that $\mathrm{Cov}_{X\sim P}(X) \preceq \mathbb{E}_{X\sim P}[(\nabla^2 V(X))^{-1}]$.*
2. *(Cramér–Rao inequality) $\mathrm{Cov}_{X\sim P}(X) \succeq (\mathbb{E}_{X\sim P}[\nabla^2 V(X)])^{-1}$.*

The Brascamp–Lieb inequality is classical, and we refer readers to Bakry et al. (2014); Bobkov and Ledoux (2000); Cordero-Erausquin (2017) for several proofs. To make our exposition more self-contained, we provide a proof of the Cramér–Rao inequality in the appendix.

5.3 Main result and proof

We now state and prove our main theorem.

Theorem 5.6. *Let $P = \exp(-V)$ and $Q = \exp(-W)$.*

1. *Suppose that (P, Q) satisfy our regularity assumptions, as well as*

$$\nabla^2 V \preceq \beta_V I\,, \qquad \text{and} \qquad \nabla^2 W \succeq \alpha_W I \succ 0\,.$$

Then, for every $\varepsilon > 0$ and all $x \in \mathbb{R}^d$, the Hessian of the entropic Brenier potential satisfies

$$\nabla^2 \varphi_\varepsilon(x) \preceq \frac{1}{2}\left(\sqrt{4\beta_V/\alpha_W + \varepsilon^2 \beta_V^2} - \varepsilon\beta_V\right) I\,.$$

2. *Suppose that (Q, P) satisfy our regularity assumptions, as well as*

$$\nabla^2 V \succeq \alpha_V I \succ 0\,, \qquad \text{and} \qquad \nabla^2 W \preceq \beta_W I\,.$$

Then, for every $\varepsilon > 0$ and all $x \in \Omega_P := \mathrm{int}(\mathrm{supp}(P))$, the Hessian of the entropic Brenier potential satisfies

$$\nabla^2 \varphi_\varepsilon(x) \succeq \frac{1}{2}\left(\sqrt{4\alpha_V/\beta_W + \varepsilon^2 \alpha_V^2} - \varepsilon\alpha_V\right) I\,.$$

Observe that as $\varepsilon \searrow 0$, we formally expect the following bounds on the Brenier potential:

$$\sqrt{\alpha_V/\beta_W}\, I \preceq \nabla^2 \varphi_0(x) \preceq \sqrt{\beta_V/\alpha_W}\, I\,.$$

In particular, this recovers Caffarelli's contraction theorem (Theorem 5.1). We make this intuition rigorous below by appealing to convergence results for the entropic potentials as the regularization parameter ε tends to zero.

Proof of Theorem 5.6. **Upper bound.** Fix $x \in \mathbb{R}^d$. Recall from Lemma 5.4 that

$$\nabla^2 \varphi_\varepsilon(x) = \varepsilon^{-1}\,\mathrm{Cov}_{Y \sim \pi_\varepsilon^x}(Y)\,.$$

By an application of the Brascamp–Lieb inequality, this results in the upper bound

$$\begin{aligned}\nabla^2\varphi_\varepsilon(x) &= \varepsilon^{-1}\operatorname{Cov}_{Y\sim\pi_\varepsilon^x}(Y)\\ &\preceq \varepsilon^{-1}\mathbb{E}_{Y\sim\pi_\varepsilon^x}\left[\left(\varepsilon^{-1}\nabla^2\psi_\varepsilon(Y)+\nabla^2 W(Y)\right)^{-1}\right]\\ &\preceq \mathbb{E}_{Y\sim\pi_\varepsilon^x}\left[\left(\nabla^2\psi_\varepsilon(Y)+\varepsilon\alpha_W I\right)^{-1}\right],\end{aligned}$$

where in the last inequality we also used the lower bound on the spectrum of $\nabla^2 W$. Next, using Lemma 5.4 and the Cramér–Rao inequality (Lemma 5.5), we obtain the lower bound

$$\begin{aligned}\nabla^2\psi_\varepsilon(Y) &= \varepsilon^{-1}\operatorname{Cov}_{X\sim\pi_\varepsilon^Y}(X)\\ &\succeq \varepsilon^{-1}\left(\mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\varepsilon^{-1}\nabla^2\varphi_\varepsilon(X)+\nabla^2 V(X)\right]\right)^{-1}\\ &\succeq \left(\mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\nabla^2\varphi_\varepsilon(X)+\varepsilon\beta_V I\right]\right)^{-1},\end{aligned}$$

where we used the upper bound on the spectrum of $\nabla^2 V$. Combining these inequalities,

$$\nabla^2\varphi_\varepsilon(x) \preceq \mathbb{E}_{Y\sim\pi_\varepsilon^x}\left[\left(\left(\mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\nabla^2\varphi_\varepsilon(X)+\varepsilon\beta_V I\right]\right)^{-1}+\varepsilon\alpha_W I\right)^{-1}\right].$$

Now, define the quantity

$$L_\varepsilon := \sup_{x\in\mathbb{R}^d}\lambda_{\max}\left(\nabla^2\varphi_\varepsilon(x)\right).$$

Then, we have shown

$$\lambda_{\max}\left(\nabla^2\varphi_\varepsilon(x)\right) \leqslant \left((L_\varepsilon+\varepsilon\beta_V)^{-1}+\varepsilon\alpha_W\right)^{-1}.$$

Taking the supremum over $x \in \mathbb{R}^d$,

$$L_\varepsilon \leqslant \left((L_\varepsilon + \varepsilon\beta_V)^{-1} + \varepsilon\alpha_W\right)^{-1}.$$

Solving the inequality yields

$$L_\varepsilon \leq \frac{1}{2}\left(\sqrt{4\beta_V/\alpha_W + \varepsilon^2\beta_V^2} - \varepsilon\beta_V\right). \tag{5.10}$$

Lower bound. The lower bound argument is symmetric, but we give the details for completeness. Using Lemma 5.4 and the Cramér–Rao inequality (Lemma 5.5),

$$\begin{aligned}
\nabla^2\varphi_\varepsilon(x) &= \varepsilon^{-1}\operatorname{Cov}_{Y\sim\pi_\varepsilon^x}(Y) \\
&\geq \varepsilon^{-1}\left(\mathbb{E}_{Y\sim\pi_\varepsilon^x}\left[\varepsilon^{-1}\nabla^2\psi_\varepsilon(Y) + \nabla^2 W(Y)\right]\right)^{-1} \\
&\geq \left(\mathbb{E}_{Y\sim\pi_\varepsilon^x}\left[\nabla^2\psi_\varepsilon(Y) + \varepsilon\beta_W I\right]\right)^{-1}.
\end{aligned}$$

Applying Lemma 5.4 and the Brascamp–Lieb inequality (Lemma 5.5),

$$\begin{aligned}
\nabla^2\psi_\varepsilon(Y) &= \varepsilon^{-1}\operatorname{Cov}_{X\sim\pi_\varepsilon^Y}(X) \\
&\preceq \varepsilon^{-1}\mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\left(\varepsilon^{-1}\nabla^2\varphi_\varepsilon(X) + \nabla^2 V(X)\right)^{-1}\right] \\
&\preceq \mathbb{E}_{X\sim\pi_\varepsilon^Y}\left[\left(\nabla^2\varphi_\varepsilon(X) + \varepsilon\alpha_V I\right)^{-1}\right].
\end{aligned}$$

Combining the two inequalities and setting

$$\ell_\varepsilon := \inf_{x\in\Omega_P} \lambda_{\min}\left(\nabla^2\varphi_\varepsilon(x)\right),$$

we deduce that

$$\ell_\varepsilon \geqslant \left((\ell_\varepsilon + \varepsilon\alpha_V)^{-1} + \varepsilon\beta_W\right)^{-1}.$$

On the other hand, from Lemma 5.4, we know that $\ell_\varepsilon \geqslant 0$. Solving the inequality then yields

$$\ell_\varepsilon \geqslant \frac{1}{2}\left(\sqrt{4\alpha_V/\beta_W + \varepsilon^2\alpha_V^2} - \varepsilon\alpha_V\right).$$

□

Next, we rigorously deduce Caffarelli's contraction theorem from Theorem 5.6.

Proof of Caffarelli's contraction (Theorem 5.1). For every $\varepsilon > 0$, by Theorem 5.6, we have proven that $\nabla^2\varphi_\varepsilon \preceq L_\varepsilon I$, with L_ε as in (5.10). Equivalently, this can be reformulated as saying that $\frac{L_\varepsilon\|\cdot\|^2}{2} - \varphi_\varepsilon$ is convex. Fix some $\delta > 0$; in particular, for ε sufficiently small, $\frac{(\sqrt{\beta_V/\alpha_W}+\delta)\|\cdot\|^2}{2} - \varphi_\varepsilon$ is convex.

Upon passing to a sequence $\varepsilon_k \searrow 0$, existing results on the convergence of entropic optimal transport potentials show that $\varphi_{\varepsilon_k} \to \varphi_0$ in $L^1(P)$ (see Nutz and Wiesel (2021)). Passing to a further subsequence, we obtain $\varphi_{\varepsilon_k} \to \varphi_0$ (P-almost surely). It follows that $\frac{(\sqrt{\beta_V/\alpha_W}+\delta)\|\cdot\|^2}{2} - \varphi_0$ is convex for every $\delta > 0$ (see the remark after Rockafellar (1997, Theorem 25.7)), and thus for $\delta = 0$ as well. □

Remark 5.7. Our main theorem provides both upper and lower bounds for $\nabla^2\varphi_\varepsilon$. In the case when $\varepsilon = 0$, the lower bound follows from the upper bound. Indeed, if φ_0 is the Brenier potential for the optimal transport from P to Q, then the convex conjugate φ_0^* is the Brenier potential for the optimal transport from Q to P. By applying Caffarelli's contraction theorem to φ_0^* and appealing to convex duality, it yields a lower bound on $\nabla^2\varphi_0$. However, we are not aware of a method of deducing the lower bound from the upper bound for positive values of ε.

Remark 5.8. In Appendix D.2, by inspecting the Gaussian case, we show that Theorem 5.6 is sharp for every $\varepsilon > 0$.

An inspection of the proof of the upper bound in Theorem 5.6 reveals the following more general pair of inequalities.

Proposition 5.9. *Let (P, Q) be probability measures satisfying our regularity conditions. Then, for all $x \in \mathbb{R}^d$ and $y \in \Omega_Q$,*

$$\nabla^2 \varphi_\varepsilon(x) \preceq \mathbb{E}_{Y \sim \pi_\varepsilon^x}\left[\left(\nabla^2 \psi_\varepsilon(Y) + \varepsilon\, \nabla^2 W(Y)\right)^{-1}\right],$$
$$\nabla^2 \psi_\varepsilon(y) \succeq \left(\mathbb{E}_{X \sim \pi_\varepsilon^y}\left[\nabla^2 \varphi_\varepsilon(X) + \varepsilon\, \nabla^2 V(X)\right]\right)^{-1}.$$

In the next section, we use these inequalities to prove a generalization of Caffarelli's theorem.

5.4 A generalization to commuting positive definite matrices

In the next result, we replace the main assumptions of Caffarelli's contraction theorem, namely $\nabla^2 V \preceq \beta_V I$ and $\nabla^2 W \succeq \alpha_W I$, by the conditions

$$\nabla^2 V \preceq A^{-1} \qquad \text{and} \qquad \nabla^2 W \succeq B^{-1}, \tag{5.11}$$

where A and B are commuting positive definite matrices. Recall that the Hessian of the Brenier potential between the Gaussian distributions $\mathcal{N}(0, A)$ and $\mathcal{N}(0, B)$ is the matrix $A^{-1/2}B^{1/2}$ (Gelbrich, 1990). In light of this observation, the following theorem is sharp for every pair of commuting positive definite (A, B), and shows that the Brenier potential between Gaussians achieves the largest possible Hessian among all source and target measures obeying the constraint (5.11).

Theorem 5.10. *Let (P, Q) satisfy our regularity conditions as well as the condition (5.11). Then, the*

Hessian of the Brenier potential satisfies the uniform bound: for all $x \in \mathbb{R}^d$, it holds that

$$\nabla^2 \varphi_0(x) \preceq A^{-1/2} B^{1/2} .$$

As in Theorem 5.6, the proof technique also yields a lower bound on $\nabla^2 \varphi_0$ under appropriate assumptions. We omit this result because it is straightforward.

Proof. Let C_ε be the smallest constant $C \geqslant 0$ such that $\nabla^2 \varphi_\varepsilon(x) \preceq A^{-1/2} B^{1/2} + CI$ for all $x \in \mathbb{R}^d$. In light of Theorem 5.6, C_ε is well-defined and finite. Equivalently,

$$C_\varepsilon = \sup_{x \in \mathbb{R}^d} \sup_{e \in \mathbb{R}^d,\ \|e\|=1} \left\langle e, [\nabla^2 \varphi_\varepsilon(x) - A^{-1/2} B^{1/2}]\, e \right\rangle .$$

Let (x, e) achieve the above supremum. (If the supremum is not attained, then the rest of the proof goes through with minor modifications.)

Using our assumptions and Proposition 5.9, we obtain

$$C_\varepsilon = \left\langle e, [\nabla^2 \varphi_\varepsilon(x) - A^{-1/2} B^{1/2}]\, e \right\rangle \leqslant \left\langle e, \left[\left((A^{-1/2} B^{1/2} + C_\varepsilon I + \varepsilon A^{-1})^{-1} + \varepsilon B^{-1} \right)^{-1} - A^{-1/2} B^{1/2} \right] e \right\rangle .$$

From our assumptions and Theorem 5.6, we know that the spectrum of $M_\varepsilon := A^{-1/2} B^{1/2} + C_\varepsilon I$ is bounded away from zero and infinity as $\varepsilon \searrow 0$, which justifies the Taylor expansion

$$\begin{aligned} \left((M_\varepsilon + \varepsilon A^{-1})^{-1} + \varepsilon B^{-1} \right)^{-1} &= \left(M_\varepsilon^{-1} - \varepsilon M_\varepsilon^{-1} A^{-1} M_\varepsilon^{-1} + \varepsilon B^{-1} + O(\varepsilon^2) \right)^{-1} \\ &= M_\varepsilon + \varepsilon A^{-1} - \varepsilon M_\varepsilon B^{-1} M_\varepsilon + O(\varepsilon^2) . \end{aligned}$$

Hence,

$$\begin{aligned} C_\varepsilon &\leqslant \left\langle e, \left[M_\varepsilon + \varepsilon A^{-1} - \varepsilon M_\varepsilon B^{-1} M_\varepsilon + O(\varepsilon^2) - A^{-1/2} B^{1/2}\right] e\right\rangle \\ &\leqslant C_\varepsilon + \varepsilon \left\langle e, \left[A^{-1} - M_\varepsilon B^{-1} M_\varepsilon\right] e\right\rangle + O(\varepsilon^2) \\ &= C_\varepsilon + \varepsilon \left\langle e, \left[C_\varepsilon A^{-1/2} B^{-1/2} + C_\varepsilon^2 B^{-1}\right] e\right\rangle + O(\varepsilon^2)\,. \end{aligned}$$

This shows that $\lim_{\varepsilon \searrow 0} C_\varepsilon = 0$ (otherwise $(C_\varepsilon)_{\varepsilon > 0}$ would have a strictly positive cluster point which would contradict the above inequality for small enough $\varepsilon > 0$).

By combining this fact with convergence of the entropic Brenier potentials as in the proof of Theorem 5.1, we deduce the result. □

Next, we show how our theorem recovers and extends a result of Valdimarsson (Valdimarsson, 2007). Valdimarsson proves that if:

- $\bar{A}$, $\bar{B}$, and G are positive definite matrices;
- $\bar{A} \preceq G$ and $\bar{B}$ commutes with G;
- $P = \mathcal{N}(0, \bar{B}G^{-1}) * \mu$ where $*$ denotes convolution and μ is an arbitrary probability measure on $\mathbb{R}^d$; and
- $Q = \exp(-W)$ with $\nabla^2 W \geq \bar{B}^{-1/2} \bar{A}^{-1} \bar{B}^{-1/2}$;

then the Brenier potential satisfies $\nabla^2 \varphi_0 \preceq G$. This result was then used to derive new forms of the Brascamp–Lieb inequality.[1]

To prove this result, we first check that convolution with any probability measure only makes the density more log-smooth.

[1] This is a different Brascamp–Lieb inequality than the one in Lemma 5.5.

Lemma 5.11. *Let $\widetilde{P} \propto \exp(-\widetilde{V})$ be a probability measure, where $\widetilde{V} : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Let $P := \widetilde{P} * \mu = \exp(-V)$ where μ is any probability measure on $\mathbb{R}^d$. Suppose that for some positive definite matrix A^{-1}, we have $\nabla^2 \widetilde{V} \preceq A^{-1}$. Then, $\nabla^2 V \preceq A^{-1}$ as well.*

Proof. An elementary computation shows that if we define the probability measure

$$\nu_y(\mathrm{d}x) := \frac{\exp(-\widetilde{V}(y-x))\,\mu(\mathrm{d}x)}{\int \exp(-\widetilde{V}(y-x'))\,\mu(\mathrm{d}x')}$$

then

$$\nabla^2 V(y) = \mathbb{E}_{X\sim\nu_y}\left[\nabla^2 \widetilde{V}(y-X)\right] - \mathrm{Cov}_{X\sim\nu_y}\left(\nabla \widetilde{V}(y-X)\right),$$

from which the result follows. □

From the lemma, we deduce that under Valdimarsson's assumptions, for $P = \exp(-V)$, we have $\nabla^2 V \preceq \bar{B}^{-1} G$. Also, $\nabla^2 W \succeq \bar{B}^{-1/2} \bar{A}^{-1} \bar{B}^{-1/2} \succeq \bar{B}^{-1} G^{-1}$. By Theorem 5.10, the Brenier potential φ_0 satisfies $\nabla^2 \varphi_0 \preceq G$. However, it is seen that our argument yields much more. For example, rather than requiring P to be a convolution with a Gaussian measure, we can allow P to be a convolution with any measure $\exp(-\widetilde{V})$ satisfying $\nabla^2 \widetilde{V} \preceq \bar{B}^{-1} G$.

Remark 5.12. It is natural to ask whether Theorem 5.10 can be obtained by first applying Caffarelli's contraction theorem to show that the optimal transport map $\widetilde{T}_0$ between the measures $(A^{-1/2})_\sharp P$ and $(B^{-1/2})_\sharp Q$ is 1-Lipschitz, and then considering the mapping $T_0(x) := B^{1/2} \widetilde{T}_0(A^{-1/2} x)$. Although T_0 is indeed a valid transport mapping from P to Q, under our assumptions ∇T_0 is not guaranteed to be symmetric, so it does not make sense to ask whether or not $\nabla T_0 \preceq A^{-1/2} B^{1/2}$.

In Valdimarsson's application to Brascamp–Lieb inequalities, it is crucial that the transport map T_0 is chosen so that ∇T_0 is a symmetric positive definite matrix. Symmetry of ∇T_0 implies that T_0 is the gradient $\nabla \varphi_0$ of a function $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$, and positive definiteness implies that φ_0 is convex. By Brenier's theorem, the unique gradient of a convex function that pushes forward P to

Q is the optimal transport map. Thus, it is crucial that we consider the *optimal* transport map here; in particular, alternative maps such as the ones by Kim and Milman (2012); Mikulincer and Shenfeld (2024) cannot be applied.

6 | Tight stability bounds for entropic Brenier maps

6.1 Introduction

The theory of optimal transport defines a geometry over probability measures via the 2-*Wasserstein distance*: for a source measure ρ and a target measure μ with finite second moments, their Wasserstein distance is given by

$$W_2^2(\rho, \mu) \coloneqq \min_{T : T_\sharp \rho = \mu} \int \|x - T(x)\|^2 \, \mathrm{d}\rho(x) \,, \tag{6.1}$$

where the constraint $T_\sharp \rho = \mu$ means that for $X \sim \rho$, $T(X) \sim \mu$, i.e., T is a transport map. The minimizer to (6.1), when it exists, is called an optimal transport map, which we denote by T_0^μ. A seminal result by Brenier (1991) states that a unique optimal transport map between ρ and μ exists whenever ρ has a density, and moreover $T_0^\mu = \nabla \varphi_0^\mu$, where φ_0^μ is some convex function. We will henceforth refer to optimal transport maps as *Brenier maps*, and the corresponding convex functions that generate them as *Brenier potentials*.

A long-standing question in the optimal transport community is the following: is the mapping $\mu \mapsto T_0^\mu$ Hölder continuous with respect to the 2-Wasserstein distance? In other words, do there

exist constants $C, \beta > 0$ such that for all probability measures μ, ν with finite second moments,

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \leq C W_2^\beta(\mu, \nu)\,? \tag{6.2}$$

Since the inequality $W_2(\mu, \nu) \leq \|T_0^\mu - T_0^\nu\|_{L^2(\rho)}$ always holds, (6.2) would imply that the mapping $\mu \mapsto T_0^\mu$ is a bi-Hölder embedding of the Wasserstein space into $L^2(\rho)$. We call such an inequality a *stability bound*.

The unique structure of the one-dimensional optimal transport problem shows that when ρ, μ, and ν are probability measures on $\mathbb{R}$, the bound (6.2) holds with $C = \beta = 1$—that is, the mapping $\mu \mapsto T_0^\mu$ is an isometry (see, e.g., Panaretos and Zemel, 2020, Chapter 2). On the other hand, Andoni et al. (2015) showed that if $d \geq 3$, then (6.2) cannot hold uniformly over all probability measures μ and ν on $\mathbb{R}^d$ with finite second moment. In fact, their main statement is significantly stronger and rules out the possibility of embedding the Wasserstein space into any L^p space, even in a very weak sense. Nevertheless, as we describe further below, a stability bound such as (6.2) can hold if further conditions are imposed on μ and ν, for instance, if they are compactly supported.

An early investigation in this direction is due to Gigli (2011), who showed that even when μ and ν are compactly supported, the exponent in (6.2) cannot be better than $\beta = \frac{1}{2}$. However, in the same paper, the author reports a simple proof due to Ambrosio that shows that if one of the Brenier maps, say T_0^ν, is Λ-Lipschitz, then $\beta = \frac{1}{2}$ is achievable, with $C = 2\sqrt{\Lambda R}$, where R is the diameter of the support of ρ; see also Mérigot et al. (2020, Theorem 2.3) for a precise statement and proof of this result. More recently, Manole et al. (2024a) showed that if T_0^ν is Λ-Lipschitz and its inverse is $1/\lambda$-Lipschitz, then $\beta = 1$ is achievable, with constant $C = \sqrt{\Lambda/\lambda}$.

Though these positive results are encouraging, requiring *a priori* smoothness bounds on one of the two Brenier potentials excludes many cases of practical interest, for instance, the case of discontinuous Brenier maps. Such maps arise commonly in applications of optimal transport to machine learning, where it is natural to consider probability measures that lie on a union

of manifolds of different intrinsic dimension (Brown et al., 2022). There has therefore been significant recent interest in obtaining stability bounds without such assumptions; see Berman (2021); Delalande and Mérigot (2023); Mérigot et al. (2020). The results of Delalande and Mérigot (2023) are the most recent. They show that if ρ has a (uniformly upper and lower bounded) density supported on a convex set $\mathcal{X}$, with μ and ν also supported on a compact set $\mathcal{Y}$, then

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \leq C_{d,\mathcal{X},\mathcal{Y},\rho} W_2^{1/6}(\mu,\nu)\,.$$

In fact, the authors prove this bound for the W_1 distance. Their proof technique relies on applications of the Brascamp–Lieb and Prekopa–Leindler inequalities.

In this chapter, we study analogous stability bounds for *entropic Brenier maps*. As entropic optimal transport is a natural smoothed analogue to the optimal transport problem, and it is reasonable to hope that techniques developed for entropic optimal transport can give insights into the structure of the unregularized problem.

Despite the importance of entropic Brenier maps, much less is known about their stability properties. The first result in this area is due to Carlier et al. (2024), who showed that if ρ, μ, and ν are compactly supported, then

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq C_\varepsilon W_2(\mu,\nu)\,, \tag{6.3}$$

where C_ε is a constant that grows exponentially as ε tends to zero.

This striking result reveals that entropic Brenier maps automatically enjoy better stability properties than unregularized Brenier maps when $\varepsilon > 0$. However, if (6.3) is to be used to extract either practical bounds for entropic Brenier maps or insights about unregularized Brenier maps in the $\varepsilon \to 0$ limit, it is crucial to obtain sharp bounds on the constant C_ε.

6.1.1 Contributions

The goal of this chapter is to improve the Lipschitz constant for the embedding $\mu \mapsto T_\varepsilon^\mu$ as a function of ε. Our main theorem is technical, but it readily implies results in the following three scenarios of interest.

First, if the source and target measures are merely supported in the Euclidean ball of radius R, then

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq \big(1 + 2R^2/\varepsilon\big) W_2(\mu, \nu)\,,$$

see Corollary 6.5. We stress that none of the measures here require densities, and so, a priori, Brenier maps may not exist, while their entropic counterparts do. Moreover, up to universal constants, we show that this bound is tight; see Remark 6.6. This is an exponential improvement on the bounds provided by Carlier et al. (2024).

As in the unregularized case, the preceding bounds can be improved under smoothness assumptions on the entropic Brenier potentials. Such assumptions are arguably more reasonable than in the unregularized case, since it is sometimes possible to obtain *a priori* smoothness bounds for entropic Brenier potentials via elementary tools (see, e.g., Chewi and Pooladian, 2023). If one of the entropic Brenier maps, say T_ε^ν, is Λ-Lipschitz, we show that the previous bound can be improved to

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq \big(1 + 2\sqrt{R\Lambda/\varepsilon}\big) W_2(\mu, \nu)\,,$$

Going further, if the backward entropic Brenier map S_ε^ν (see Section 6.2 for a precise definition) is $1/\lambda$-Lipschitz, then the bound becomes independent of the regularization parameter:

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq \big(1 + 2\sqrt{\Lambda/\lambda}\big) W_2(\mu, \nu)\,.$$

See Corollary 6.7 for these last two results. In particular, up to constants, this result is analogous to the stability bound established by Manole et al. (2024a).

As a novel application, we turn to the *semi-discrete* setting of optimal transport, where μ and ν are both supported on finitely many atoms and ρ has a sufficiently well-behaved density. In this setting, we partially close the gap left by Gigli and others, where we prove that

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \lesssim W_2^{1/3}(\mu, \nu)\,, \tag{6.4}$$

where μ, ν satisfy appropriate regularity conditions, as does the source measure ρ, and the suppressed constant depends on these regularity assumptions. While our results do not allow for arbitrary discrete measures, they hold for a wide class of discrete measures and do not require the support of the atoms to be the same. The proof starts from the following application of the triangle inequality

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \le \|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)} + \|T_0^\nu - T_\varepsilon^\nu\|_{L^2(\rho)} + C_\varepsilon W_2(\mu, \nu)\,.$$

Under appropriate assumptions on ρ and the two discrete measures μ and ν, we are able to control the first two terms using existing techniques, and the third term can be controlled via Corollary 6.5. Balancing the resulting terms as a function of ε, we obtain the final bound that appears in (6.4). Our identification of the sharp constant C_ε is crucial to obtaining the result. See Section 6.4 for more details.

6.2 Background

6.2.1 Entropic optimal transport and notation

For two probability measures $\rho, \nu \in \mathcal{P}_2$, the entropic optimal transport objective (Cuturi, 2013) is defined as

$$\mathrm{OT}_\varepsilon(\rho, \nu) \coloneqq \min_{\pi \in \Pi(\rho,\nu)} \iint \tfrac{1}{2}\|x - z\|^2 \,\mathrm{d}\pi(x, z) + \varepsilon \,\mathrm{KL}(\pi \| \rho \otimes \nu)\,, \tag{6.5}$$

for some $\varepsilon > 0$, and $\mathrm{KL}(\pi \| \rho \otimes \nu)$ is the Kullback–Leibler divergence, defined as

$$\mathrm{KL}(\pi \| \rho \otimes \nu) \coloneqq \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\rho \otimes \mathrm{d}\nu}\right) \mathrm{d}\pi$$

when π is absolutely continuous with respect to $\rho \otimes \nu$, and $+\infty$ otherwise. Note that due to the regularization term, the problem is strictly convex with a unique minimizer π_ε^ν, the *optimal entropic (transport) coupling*.[1]

The entropic optimal transport problem also admits a dual formulation (see, e.g., Genevay, 2019):

$$\mathrm{OT}_\varepsilon(\rho, \nu) \coloneqq \tfrac{1}{2} M_2(\rho + \nu) - \min_{\varphi \in L^1(\rho)} \int \varphi \,\mathrm{d}\rho + \int \Phi_\varepsilon^\rho[\varphi] \,\mathrm{d}\nu\,, \tag{6.6}$$

where Φ_ε^ν is the following operator

$$\forall z \in \mathbb{R}^d,\ \Phi_\varepsilon^\rho[\varphi](z) \coloneqq \varepsilon \log \int e^{(\langle x, z\rangle - \varphi(x))/\varepsilon} \,\mathrm{d}\rho(x)\,,$$

which should be thought of as the entropic analogue to the convex conjugate operator. Indeed, notice that as $\varepsilon \to 0$, $\Phi_\varepsilon^\rho[\varphi](z)$ converges to the ρ-essential supremum of the function $x \mapsto \langle x, z\rangle - \varphi(x)$. We write the minimizer to (6.6) as φ_ε^ν, from which we obtain the minimizing pair of

[1] Since ρ is fixed throughout, we use the superscript ν (respectively, μ) to indicate objects that correspond to the entropic optimal transport problem between ρ and ν (respectively, ρ and μ).

entropic Brenier potentials

$$(\varphi_\varepsilon^\nu, \psi_\varepsilon^\nu) \coloneqq (\varphi_\varepsilon^\nu, \Phi_\varepsilon^\rho[\varphi_\varepsilon^\nu]) = (\Phi_\varepsilon^\nu[\psi_\varepsilon^\nu], \psi_\varepsilon^\nu)\,,$$

where Φ_ε^ν is defined analogously to Φ_ε^ρ. Again, this pair is unique up to constant shifts.

Moreover, by the dual optimality conditions, we can define versions of the entropic Brenier potentials taking values in the extended reals, for all $x \in \mathbb{R}^d$ and $z \in \mathbb{R}^d$, respectively. Thus, we freely write

$$\begin{aligned} \varphi_\varepsilon^\nu(x) &\coloneqq \varepsilon \log \int e^{(\langle x,z\rangle - \psi_\varepsilon^\nu(z))/\varepsilon}\, \mathrm{d}\nu(z) \quad (x \in \mathbb{R}^d) \\ \psi_\varepsilon^\nu(z) &\coloneqq \varepsilon \log \int e^{(\langle x,z\rangle - \varphi_\varepsilon^\nu(x))/\varepsilon}\, \mathrm{d}\rho(x) \quad (z \in \mathbb{R}^d)\,. \end{aligned} \tag{6.7}$$

See Mena and Niles-Weed (2019); Nutz and Wiesel (2021) for more details. Note that $\varphi_\varepsilon^\nu : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ (resp. ψ_ε^ν) is a convex function which is analytic on the interior of its domain $\mathrm{dom}(\varphi_\varepsilon^\nu)$ (resp. $\mathrm{dom}(\psi_\varepsilon^\nu)$),

An important feature of the entropic optimal transport problem is that the optimal solutions to (6.5) and (6.6) satisfy the following primal-dual relationship (Csiszár, 1975):

$$\mathrm{d}\pi_\varepsilon^\nu(x,z) \coloneqq \gamma_\varepsilon^\nu(x,z)\, \mathrm{d}\rho(x)\, \mathrm{d}\nu(z) \coloneqq e^{(\langle x,z\rangle - \psi_\varepsilon^\nu(z) - \varphi_\varepsilon^\nu(x))/\varepsilon}\, \mathrm{d}\rho(x)\, \mathrm{d}\nu(z)\,.$$

Concretely, γ_ε^ν, the density of π_ε^ν with respect to $\rho \otimes \nu$, can be written explicitly in terms of the entropic Brenier potentials $(\varphi_\varepsilon^\nu, \psi_\varepsilon^\nu)$.

Let (X, Z) be a pair of random variables with distribution π_ε^ν. For a given $x \in \mathrm{dom}(\varphi_\varepsilon^\nu)$, we abuse notation and define the conditional probability of $Z|X = x$ as

$$\mathrm{d}\pi_\varepsilon^\nu(z|x) = e^{(\langle x,z\rangle - \varphi_\varepsilon^\nu(x) - \psi_\varepsilon^\nu(z))/\varepsilon}\, \mathrm{d}\nu(z)\,.$$

Similarly we denote $\pi_\varepsilon^\nu(\cdot|z) := \pi_\varepsilon^\nu(\cdot|Z = z)$ whenever $z \in \mathrm{dom}(\psi_\varepsilon^\nu)$. Likewise, if (X, Y) are distributed according to the optimal entropic coupling π_ε^μ between ρ and μ, we will write $\pi_\varepsilon^\mu(\cdot|x)$ and $\pi_\varepsilon^\mu(\cdot|y)$ for the conditional distributions of $Y|X = x$ and $X|Y = y$, respectively. We will adopt the convention throughout that X, Y, and Z always refer to random variables with marginal distributions ρ, μ, and ν, respectively.

Following e.g., Pooladian and Niles-Weed (2021); Rigollet and Stromme (2022), we define, respectively, the *forward* and *backward entropic Brenier maps* from ρ to ν to be barycentric projections of π_ε^ν (Ambrosio et al., 2008, Definition 5.4.2): for $x \in \mathrm{dom}(\varphi_\varepsilon^\nu)$ and $z \in \mathrm{dom}(\psi_\varepsilon^\nu)$, we define

$$T_\varepsilon^\nu(x) := \int z \, \mathrm{d}\pi_\varepsilon^\nu(z|x)\,, \quad S_\varepsilon^\nu(z) := \int x \, \mathrm{d}\pi_\varepsilon^\nu(x|z)$$

whenever the integrals are well-defined. Unlike the unregularized case, $(S_\varepsilon^\nu)^{-1} \neq T_\varepsilon^\nu$. Note that by Jensen's inequality, $T_\varepsilon^\nu \in L^2(\rho)$ with $\|T_\varepsilon^\nu\|_{L^2(\rho)}^2 \leq M_2(\nu)$ (resp. $S_\varepsilon^\nu \in L^2(\nu)$ with $\|S_\varepsilon^\nu\|_{L^2(\nu)}^2 \leq M_2(\rho)$). Also note that by the dominated convergence theorem, the gradient of φ_ε^ν (resp. ψ_ε^ν) from (6.7) has a natural interpretation as the forward (resp. backward) entropic Brenier map: whenever x is in the interior of $\mathrm{dom}(\varphi_\varepsilon^\nu)$ and z is in the interior of $\mathrm{dom}(\psi_\varepsilon^\nu)$,

$$\nabla\varphi_\varepsilon^\nu(x) = T_\varepsilon^\nu(x)\,, \quad \nabla\psi_\varepsilon^\nu(z) = S_\varepsilon^\nu(z)\,. \tag{6.8}$$

Under the same condition, a similar expression holds for the Hessians of the entropic Brenier potentials (see, e.g., Chewi and Pooladian, 2023, Lemma 1):

$$\nabla^2\varphi_\varepsilon^\nu(x) = \varepsilon^{-1}\mathrm{Cov}_{\pi_\varepsilon^\nu}(Z|X = x)\,, \quad \nabla^2\psi_\varepsilon^\nu(z) = \varepsilon^{-1}\mathrm{Cov}_{\pi_\varepsilon^\nu}(X|Z = z)\,. \tag{6.9}$$

Throughout, we will write $(\varphi_\varepsilon^\nu, \psi_\varepsilon^\nu)$ for the entropic Brenier potentials associated to $\mathrm{OT}_\varepsilon(\rho, \nu)$, T_ε^ν for the forward entropic Brenier map, and S_ε^ν for the backward entropic Brenier map (and the same for for $\mathrm{OT}_\varepsilon(\rho, \mu)$).

Similarly, we will simply write φ_0^ν, ψ_0^ν and T_0^ν to refer to the quantities associated to the unregularized optimal transport problem $W_2^2(\rho, \nu)$ (as for $W_2^2(\rho, \mu)$).

6.2.2 Related work in entropic optimal transport

Fixed regularization. The initial motivation for studying (6.5) in the machine learning literature was its significant computational benefits compared to the standard optimal transport problem (Altschuler et al., 2017; Cuturi, 2013). As a result, the study of entropic objects for a fixed $\varepsilon > 0$ regularization parameter has been of great interest in a number of fields. For example, del Barrio et al. (2022b); Goldfeld et al. (2024a); Gonzalez-Sanz et al. (2022) studied statistical limit theorems for entropic optimal transport. Conforti et al. (2023); Greco et al. (2023); Nutz and Wiesel (2023) study the convergence of Sinkhorn's algorithm to the optimal Brenier potentials at the population level. The works by Klein et al. (2024); Masud et al. (2023); Pooladian et al. (2022); Rigollet and Stromme (2022); Stromme (2024); Werenski et al. (2023) studied additional computational or statistical properties of entropic Brenier maps. As previously mentioned, Carlier et al. (2024) initiated the study of the stability properties of entropic Brenier maps under variations of the target measure, though their techniques differ significantly from ours.

Vanishing regularization. Theoretical properties of entropic optimal transport for vanishing regularization parameter are widely studied in both statistical and theoretical works. For example, convergence of the regularized to unregularized couplings was studied by Bernton et al. (2022); Carlier et al. (2017); Ghosal et al. (2022); Léonard (2012), and convergence of the transport costs by Chizat et al. (2020); Conforti and Tamanini (2021); Eckstein and Nutz (2023); Pal (2024). Nutz and Wiesel (2021) established convergence of the entropic to non-entropic Brenier potentials under minimal assumptions; this convergence was improved in the case of semi-discrete optimal transport by Altschuler et al. (2022) and Delalande (2022). Chewi and Pooladian (2023) established a short proof of Caffarelli's contraction theorem (Caffarelli, 2000) via covariance inequalities and

entropic optimal transport, which was subsequently generalized by Conforti (2024). Statistical convergence of entropic Brenier maps to unregularized Brenier maps was established by Pooladian et al. (2023); Pooladian and Niles-Weed (2021), the latter paper focusing on the semi-discrete setting.

6.2.3 Key ingredient: A transport inequality for conditional entropic couplings

At the core of our approach is the use of a specific transport inequality which has been developed for other purposes in the study of sampling and functional inequalities (Anari et al., 2021a;b; Bauerschmidt et al., 2023; Chen and Eldan, 2022). We refer to Bauerschmidt et al. (2023, Section 3.7) for more details, and briefly overview the necessary inequalities and notation here.

Let $q \in \mathcal{P}_2$ be a probability measure with finite moment generating function whose covariance is denoted by $\mathrm{Cov}(q)$. For $h \in \mathbb{R}^d$, we define the tilt $\mathcal{T}_h q$ of q as the probability measure satisfying

$$\forall z \in \mathbb{R}^d, \quad \frac{\mathrm{d}\mathcal{T}_h q}{\mathrm{d}q}(z) := \frac{\exp(\langle h, z\rangle)}{\mathbb{E}_{Z\sim q}[\exp(\langle h, Z\rangle)]}\,.$$

We say that q is *tilt-stable*[2] if for all $h \in \mathbb{R}^d$, $\mathrm{Cov}(\mathcal{T}_h q) \preceq C_{\mathsf{T}} I$ for some $C_{\mathsf{T}} > 0$. If q is tilt-stable with constant C_{T}, then for all probability measures $p \in \mathcal{P}_2$,

$$\|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|^2 \le 2C_{\mathsf{T}}\mathrm{KL}(p\|q)\,,$$

see Bauerschmidt et al. (2023, Lemma 3.21).

Our main observation is that conditional entropic couplings are tilt-stable, with a constant that can be written in terms of the entropic Brenier potentials. For an entropic potential φ_ε^ν whose

[2]This name is not standard, but we introduce it here because the standard name for this concept (entropic stability) is likely to cause confusion in the context of our main results.

domain is all of $\mathbb{R}^d$, we write

$$H_{\max}(\varphi_\varepsilon^\nu) \coloneqq \sup_{u\in\mathbb{R}^d} \|\mathrm{Cov}_{\pi_\varepsilon^\nu}(Z|X=u)\|_{\mathrm{op}} = \varepsilon \sup_{u\in\mathbb{R}^d} \|\nabla^2\varphi_\varepsilon^\nu(u)\|_{\mathrm{op}}\,, \tag{6.10}$$

and define $H_{\max}(\psi_\varepsilon^\nu)$ analogously. The second equality in (6.10) is justified by the fact that (6.9) holds everywhere when $\mathrm{dom}(\varphi_\varepsilon^\nu) = \mathbb{R}^d$. If either potential is not finite on all of $\mathbb{R}^d$, we adopt the convention that $H_{\max} = +\infty$.

Lemma 6.1. *Let $x \in \mathbb{R}^d$ and let $\pi_\varepsilon^\nu(\cdot|x)$ be a conditional entropic coupling between two probability measures $\rho, \nu \in \mathcal{P}_2$. Assume that $\mathrm{dom}(\varphi_\varepsilon^\nu) = \mathbb{R}^d$. Then for any $h \in \mathbb{R}^d$,*

$$\mathcal{T}_h\pi_\varepsilon^\nu(\cdot|x) = \pi_\varepsilon^\nu(\cdot|x+\varepsilon h)\,.$$

Corollary 6.2. *The conditional entropic coupling $\pi_\varepsilon^\nu(\cdot|x)$ (resp. $\pi_\varepsilon^\nu(\cdot|z)$) is tilt-stable with constant $H_{\max}(\varphi_\varepsilon^\nu)$ (resp. $H_{\max}(\psi_\varepsilon^\nu)$).*

6.3 Main results

We now present our general stability result for entropic Brenier maps.

Theorem 6.3 (Stability of entropic Brenier maps)**.** *Suppose ρ, μ, ν have finite second moment. Then*

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le \left(1 + \frac{2(H_{\max}(\varphi_\varepsilon^\nu)H_{\max}(\psi_\varepsilon^\nu))^{1/2}}{\varepsilon}\right) W_2(\mu,\nu)\,.$$

Remark 6.4. Note that if the potentials $(\varphi_\varepsilon^\nu, \psi_\varepsilon^\nu)$ are not finite everywhere, the quantities $H_{\max}(\varphi_\varepsilon^\nu)$ and $H_{\max}(\psi_\varepsilon^\nu)$ are infinite by convention, and the inequality becomes vacuous. The potentials are finite everywhere whenever ρ and ν have moment-generating functions that are finite everywhere (including the important case of bounded supports), but also when ρ and ν have support equal to $\mathbb{R}^d$, without additional tail assumptions.

From here, we can prove the results highlighted in the introduction as special cases.

Corollary 6.5 (Entropic stability for bounded measures)**.** *Suppose ρ and ν are supported in $B(0;R)$, and μ has finite second moment. Then*

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \leq \left(1 + \frac{2R^2}{\varepsilon}\right) W_2(\mu,\nu)\,.$$

Proof. Since ρ and ν are supported in $B(0;R)$, (6.9) implies that both $H_{\max}(\varphi_\varepsilon^\nu)$ and $H_{\max}(\psi_\varepsilon^\nu)$ are smaller than R^2, which completes the proof. □

Note that Corollary 6.5 is *entirely general* in its requirements, and does not rely on smoothness of maps, nor do any of the measures require densities.

Remark 6.6 (Tightness of Corollary 6.5). We now demonstrate that Corollary 6.5 is tight for general bounded probability measures. Fix $R > 0$, and let $p_\theta := \frac{1}{2}\delta_{Re_\theta} + \frac{1}{2}\delta_{-Re_\theta}$ with $e_\theta = (\cos(\theta), \sin(\theta))$, for $\theta \in [0, \frac{\pi}{2}]$. Let $\rho := p_{\pi/2}$ and $\varepsilon > 0$, and let π_ε^θ denote the entropic optimal coupling between ρ and p_θ for $\theta \in [0, \frac{\pi}{2})$. One can deduce that the optimal entropic coupling is symmetric for any such θ:

$$\pi_\varepsilon^\theta(x,y) = \pi_\varepsilon^\theta(-x,-y)\,,$$

Following the calculations in Altschuler et al. (2022, Section 3), one can choose $\psi_\varepsilon^\theta(e_\theta) = \psi_\varepsilon^\theta(-e_\theta) = 0$, so that by (6.7), for all $x \in \mathbb{R}^d$,

$$\varphi_\varepsilon^\theta(x) = \varepsilon \log\left(\tfrac{1}{2}e^{R\langle x, e_\theta\rangle/\varepsilon} + \tfrac{1}{2}e^{R\langle x,-e_\theta\rangle/\varepsilon}\right).$$

Then we compute

$$T_\varepsilon^\theta(x) = Re_\theta\left(\pi_\varepsilon^\theta(x,e_\theta) - \pi_\varepsilon^\theta(x,-e_\theta)\right) = Re_\theta \tanh(R\langle x, e_\theta\rangle/\varepsilon)\,.$$

Let $\mu := p_0$ and $\nu := p_\theta$. Following the above calculations, it is clear that

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} = R\|e_\theta\|\sqrt{\tanh^2(R\sin(\theta)/\varepsilon)} = \frac{R^2\theta}{\varepsilon} + O(\theta^4)\,.$$

It is also easy to verify that $W_2(p_0, p_\theta) \asymp \theta$, since the optimal transport map from p_0 to p_θ is the standard 2×2 rotation matrix acting on the dirac masses. This example shows that for θ small the dependence $R^2\varepsilon^{-1}$ in Corollary 6.5 is tight.

The following example provides the entropic analogue of Theorem 6 from Manole et al. (2024a); their result is formally recovered in the $\varepsilon \to 0$ limit.

Corollary 6.7 (Improved stability under smoothness)**.** *Suppose T_ε^ν is uniformly Λ-Lipschitz. If ρ is supported in $B(0;R)$, then*

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le \big(1 + 2\sqrt{\Lambda R^2/\varepsilon}\big)W_2(\mu,\nu)\,.$$

If instead S_ε^ν is uniformly $1/\lambda$-Lipschitz, then

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le (1 + 2\sqrt{\Lambda/\lambda})W_2(\mu,\nu)\,.$$

Proof. The first claim follows from the bounds $H_{\max}(\varphi_\varepsilon^\nu) \le \Lambda\varepsilon$, which follows from (6.10), and $H_{\max}(\psi_\varepsilon^\nu) \le R^2$. For the second, we instead use that $H_{\max}(\psi_\varepsilon^\nu) \le \varepsilon/\lambda$. □

6.3.1 Proof of Theorem 6.3

Our proof relies on three propositions. To continue, we require the following objects. Let $\tau \in \Pi(\mu,\nu)$ be a fixed (though not necessarily unique) optimal transport coupling between μ and ν. For $z \in \mathbb{R}^d$, let $\tau(\cdot|z)$ be associated (regular) conditional measure (see Bogachev, 2007, Chapter

10), so that for all measurable $f : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$

$$\iint f(y,z)\,\mathrm{d}\tau(y,z) = \int \left(\int f(y,z)\,\mathrm{d}\tau(y|z) \right) \mathrm{d}\nu(z)\,.$$

For $x \in \mathbb{R}^d$, let $Q(\cdot|x)$ be the probability measure with

$$\forall z \in \mathbb{R}^d,\ \frac{\mathrm{d}Q(\cdot|x)}{\mathrm{d}\nu}(z) \coloneqq \int \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y|z)\,, \tag{6.11}$$

where $\gamma_\varepsilon^\mu(x,y)$ is the density of π_ε^μ w.r.t. $\rho \otimes \mu$. Note that this indeed defines a density as we have the relation

$$\int \left(\int \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y|z) \right) \mathrm{d}\nu(z) = \iint \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y,z) = \int \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\mu(y) = 1\,.$$

We also define the conditional Kullback–Leibler divergence:

$$I \coloneqq \int \mathrm{KL}(Q(\cdot|x) \| \pi_\varepsilon^\nu(\cdot|x))\,\mathrm{d}\rho(x)\,, \tag{6.12}$$

We are now in a position to proceed with the proof. First, we decompose the difference of forward entropic Brenier maps into a $W_2(\mu,\nu)$ term, plus a term depending on I.

Proposition 6.8. *Suppose ρ, μ, ν have finite second moment. Then*

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le W_2(\mu,\nu) + (2H_{\max}(\varphi_\varepsilon^\nu) I)^{1/2}\,.$$

In (6.12), we defined the conditional relative entropy between $Q(\cdot|x)$ and the conditional entropic coupling $\pi_\varepsilon^\nu(\cdot|x)$. We now turn to directly bounding the quantity I. Note that for all

$z \in \mathbb{R}^d$

$$\frac{\mathrm{d}Q(\cdot|x)}{\mathrm{d}\pi_\varepsilon^\nu(\cdot|x)}(z) = \frac{\int \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y|z)}{\gamma_\varepsilon^\nu(x,z)}\,.$$

An application of Jensen's inequality then yields that

$$I \le \bar{I} \coloneqq \iiint \log\Big(\frac{\gamma_\varepsilon^\mu(x,y)}{\gamma_\varepsilon^\nu(x,z)}\Big)\gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y,z)\,\mathrm{d}\rho(x)\,. \tag{6.13}$$

Expanding the densities $\gamma_\varepsilon^\mu(x,y)$ and $\gamma_\varepsilon^\nu(x,z)$ and performing the integration, we obtain

$$\begin{aligned}\varepsilon\bar{I} &= \iiint \langle x, y-z\rangle \gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\tau(y,z)\,\mathrm{d}\rho(x) + \int \varphi_\varepsilon^\nu\,\mathrm{d}\rho + \int \psi_\varepsilon^\nu\,\mathrm{d}\nu - \int \varphi_\varepsilon^\mu\,\mathrm{d}\rho - \int \psi_\varepsilon^\mu\,\mathrm{d}\mu \\ &= \iint \langle S_\varepsilon^\mu(y), y-z\rangle\,\mathrm{d}\tau(y,z) + \int \varphi_\varepsilon^\nu\,\mathrm{d}\rho + \int \psi_\varepsilon^\nu\,\mathrm{d}\nu - \int \varphi_\varepsilon^\mu\,\mathrm{d}\rho - \int \psi_\varepsilon^\mu\,\mathrm{d}\mu\end{aligned}$$

where we use the equality $S_\varepsilon^\mu(y) = \int x\gamma_\varepsilon^\mu(x,y)\,\mathrm{d}\rho(x)$ in the last line. If we define $\tilde{I}$ as a symmetric analogue to $\bar{I}$, namely,

$$\tilde{I} \coloneqq \iiint \log\Big(\frac{\gamma_\varepsilon^\nu(x,z)}{\gamma_\varepsilon^\mu(x,y)}\Big)\gamma_\varepsilon^\nu(x,z)\,\mathrm{d}\tau(y,z)\,\mathrm{d}\rho(x)\,,$$

then, since $0 \le \tilde{I}$, a symmetric calculation immediately yields the following:

Proposition 6.9. *Suppose ρ, μ, ν have finite second moment. Then*

$$\varepsilon\bar{I} \le \varepsilon(\bar{I}+\tilde{I}) = \iint \langle S_\varepsilon^\mu(y) - S_\varepsilon^\nu(z), y-z\rangle\,\mathrm{d}\tau(y,z)\,. \tag{6.14}$$

All in all, our bound currently reads

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le W_2(\mu,\nu) + (2\varepsilon^{-1}H_{\max}(\varphi_\varepsilon^\nu))^{1/2}\Big(\iint \langle S_\varepsilon^\mu(y) - S_\varepsilon^\nu(z), y-z\rangle\,\mathrm{d}\tau(y,z)\Big)^{1/2}.$$

The second term depends on the two backward entropic Brenier maps, S_ε^μ and S_ε^ν. The next proposition shows that tilt stability can again be used to bound the difference between these backward maps by $\bar{I}$.

Proposition 6.10. *Suppose ρ, μ, ν have finite second moment. Then*

$$\iint \|S_\varepsilon^\mu(y) - S_\varepsilon^\nu(z)\|^2 \, \mathrm{d}\tau(y,z) \leq 2H_{\max}(\psi_\varepsilon^\nu)\bar{I} \,.$$

Finally, with these results in hand, we can prove our main result.

Proof of Theorem 6.3. From (6.14), we apply Cauchy-Schwarz, resulting in

$$\varepsilon\bar{I} \leq \iint \langle S_\varepsilon^\mu(y) - S_\varepsilon^\nu(z), y - z\rangle \, \mathrm{d}\tau(y,z) \leq W_2(\mu,\nu)(2H_{\max}(\psi_\varepsilon^\nu)\bar{I})^{1/2} \,.$$

This ultimately implies

$$I^{1/2} \leq \bar{I}^{1/2} \leq \varepsilon^{-1} W_2(\mu,\nu)(2H_{\max}(\psi_\varepsilon^\nu))^{1/2} \,,$$

where we recall the first inequality from (6.13). Together with Proposition 6.8, the proof is complete. □

6.4 Application: Improved quantitative stability of semi-discrete optimal transport maps

As an application of our new stability results for entropic Brenier maps, we turn to proving quantitative stability results for *unregularized* optimal transport maps. As highlighted in the introduction, our proof technique for proving quantitative stability for optimal transport maps is

based on the following decomposition:

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \le \|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)} + \|T_0^\nu - T_\varepsilon^\nu\|_{L^2(\rho)} + \|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \,. \tag{6.15}$$

Recall that Corollary 6.5 takes care of the last term under virtually no assumptions other than boundedness of the measures. It remains to control the first two terms in the above decomposition, also known as *bias* terms.

To the best of our knowledge, bounds on the bias of entropic Brenier maps are known only under strong assumptions. For example, Pooladian and Niles-Weed (2021) showed that (see their Corollary 1)

$$\|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)}^2 \lesssim \varepsilon^2 I_0(\rho, \mu) \,,$$

where $I_0(\rho, \mu)$ is the integrated Fisher information along the Wasserstein geodesic between ρ and μ, where ρ and μ have upper and lower bounded densities over compact domains. Such assumptions, while essential for estimating optimal transport maps on the basis of samples, are too restrictive for our purposes.[3]

With this in mind, our goal is to establish quantitative control on the bias of entropic Brenier maps under less restrictive regularity conditions. Specifically, we turn to the *semi-discrete* setting, where ρ has a density, and μ and ν are both discrete measures. As we will shortly see, this setting allows for (6.15) to be used to obtain meaningful bounds on the stability of two semi-discrete optimal transport maps when the discrete measures themselves have favorable properties.

We briefly recall some fundamental notions from semi-discrete optimal transport: let $\mu = \sum_{j=1}^J \mu_j \delta_{y_j}$ be a discrete probability measure with atoms located at the points $\{y_j\}_{j=1}^J$ with corre-

[3] Indeed, under these assumptions, it is well-known via Caffarelli regularity theory (Caffarelli, 1992; 1996) that the corresponding optimal transport map is Lipschitz, so the results of Gigli (2011) already imply a stability bound.

sponding weights $\mu_j > 0$. In this setting, the Brenier potential is given explicitly by

$$\varphi_0^\mu(x) = \max_{j \in \{1,\dots,J\}} \langle x, y_j \rangle - (\psi_0^\mu)_j \,,$$

where $\psi_0^\mu \in \mathbb{R}^J$ is the dual potential. Note that φ_0^μ is ρ-almost everywhere differentiable, and so the Brenier map $T_0^\mu = \nabla \varphi_0^\mu$ is well-defined. The inverse transport map is now set-valued, where for a given target atom y_j, we define the *Laguerre cell* $L_j := (T_0^\mu)^{-1}(y_j)$. These cells partition the support of ρ. Consequently, for $x \in L_j$, the optimal transport mapping is $x \mapsto T_0^\mu(x) = y_j$.

With these notions in hand, we are ready to present the following result on the convergence of entropic Brenier maps to their unregularized counterpart; its proof is located in Appendix E.5. This is a slightly different version of Theorem 3.5 by Pooladian et al. (2023), which based off the results of Delalande (2022).

Proposition 6.11 (Quantitative bias in the semi-discrete setting)**.** *Let ρ be a compactly supported probability distribution with a density over $\mathbb{R}^d$ and μ be a discrete measure, written $\mu = \sum_{j=1}^J \mu_j \delta_{y_j}$. Then for all $\varepsilon > 0$,*

$$\|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)}^2 \le e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon} \varepsilon \sum_{i,j} \frac{\|y_i - y_j\|}{2} \int_0^\infty h_{ij}(u\varepsilon)\left(1 + e^{u/2}\right)^{-1} \mathrm{d}u \,, \tag{6.16}$$

where $0 \le h_{ij}(\cdot)$ measures the amount of overlap between L_i and L_j weighted against the source measure ρ (see (E.1) in the appendix for precise details), and $(\tilde{\psi}_\varepsilon^\nu)_j = (\psi_\varepsilon^\nu)_j - \varepsilon \log \mu_j$ for $j \in \{1, \dots, J\}$.

In addition, suppose that

- ***(T1)*** *the density of ρ has convex support in $B(0; R)$, is α-Hölder continuous for $\alpha \in (0, 1]$, and there exist $\rho_{\min}, \rho_{\max}$ such that $0 < \rho_{\min} \le \rho(x) \le \rho_{\max}$,*
- ***(T2)*** *the support of μ lies in $B(0; R)$, and all the weights are uniformly lower-bounded i.e., $\mu_j \ge \mu_{\min} > 0$ for all $j \in \{1, \dots, J\}$.*

Then it holds that, for all $\varepsilon > 0$

$$\|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)}^2 \leq C_0 e^{C_1 \varepsilon^\alpha} \varepsilon\,, \tag{6.17}$$

where the constants depend on $\rho_{\min}, \rho_{\max}, d, R, \mu_{\min}, J, \min_{i\neq j} \|y_i - y_j\|$, *and on the maximum angle formed by three non aligned points among the atoms* $\{y_j\}_{j=1}^J$.

Remark 6.12. Following the asymptotic results of Altschuler et al. (2022), we can take a limit of (6.16), resulting in the computation

$$\limsup_{\varepsilon\to\infty} \varepsilon^{-1}\|T_0^\mu - T_\varepsilon^\mu\|_{L^2(\rho)}^2 = \sum_{i,j} \frac{\|y_i - y_j\| h_{ij}(0)}{2} \int_0^\infty \left(1 + e^{u/2}\right)^{-1} \mathrm{d}u = \sum_{i,j} \|y_i - y_j\| h_{ij}(0) \log(2)\,,$$

where we used that $t \mapsto h_{ij}(t)$ is continuous at $t = 0$ (which holds if, for instance, ρ has an upper bounded density with compact support). We conjecture that this quantity is uniformly bounded for all discrete measures.

Combined with (6.15) and Corollary 6.5, we can state and prove our main theorem for this section. While the conditions do not allow for arbitrary discrete measures, we stress that they permit a wide class of discrete measures, and in particular measures supported on different masses. To our knowledge, this is the first general improvement to the stability bound of Delalande and Mérigot (2023), even in the semi-discrete case.

Theorem 6.13 (Near-tight stability in the semi-discrete setting)**.** *Suppose* ρ *satisfies* ***(T1)****, and both* μ *and* ν *each independently satisfy* ***(T2)*** *(with possibly all different parameters). Then*

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \lesssim W_2^{1/3}(\mu, \nu)\,,$$

where the underlying constants depend on those from Proposition 6.11.

Proof. First, we note that if $W_2(\mu, \nu) \geq 1$, then $\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \leq 2R \leq 2RW_2(\mu, \nu)^{1/3}$, so the only

case of interest is when $W_2(\mu, \nu) \leq 1$. Then, using the decomposition from (6.15) with Corollary 6.5 and two applications of (6.17), we obtain

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)}^2 \leq 4 \max\{C_0(\mu)e^{C_1(\mu)\varepsilon^\alpha}, C_0(\nu)e^{C_1(\nu)\varepsilon^\alpha}\}\varepsilon + (2 + 4R^4/\varepsilon^2)W_2^2(\mu, \nu)\,.$$

Choosing $\varepsilon = W_2^{2/3}(\mu, \nu) \leq 1$, we obtain our desired rate with a prefactor scaling like $C_0 e^{C_1} + 1$, where we choose the worse constant arising from the bias terms between μ and ν. □

Remark 6.14. Closest to this result is that of Bansil and Kitagawa (2022): when μ and ν are supported on the *same* atoms, the following bound holds[4]

$$\|T_0^\mu - T_0^\nu\|_{L^2(\rho)}^2 \leq (J-1)\mathrm{diam}(\Omega)^2 \mathrm{TV}(\mu, \nu)\,, \tag{6.18}$$

where $\mathrm{TV}(\cdot, \cdot)$ is the total variation distance, and J is the number of atoms in the support of μ and ν. Theorem 6.13 implies meaningful bounds in some situations in which (6.18) fails to do so. For example, consider the simple setting where $\rho = \mathrm{Unif}(B(0,1))$ and $\mu_\theta = \frac{1}{2}(\delta_{e_\theta} + \delta_{-e_\theta})$ with $e_\theta = (\cos(\theta), \sin(\theta))$ for $0 < \theta \ll \pi/2$. It is easy to verify that $W_2(\mu_0, \mu_\theta) \asymp \theta$, so Theorem 6.13 gives $\|T_0^\mu - T_0^\nu\|_{L^2(\rho)} \lesssim \theta^{1/3}$. On the other hand, since $\mathrm{TV}(\mu_0, \mu_\theta) = 1$, the bound (6.18) is vacuous.

[4]This bound is not explicitly written in the paper but it can be extracted from their Theorem 1.3.

Part III

Optimization over the Wasserstein space

7 Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space

7.1 Introduction

This chapter develops a framework for optimizing over *polyhedral* subsets of the Wasserstein space, with accompanying guarantees. Our main application is to provide the first end-to-end computational guarantees for mean-field variational inference (Blei et al., 2017; Wainwright and Jordan, 2008) under standard tractability assumptions on the posterior distribution. We now contextualize our work with respect to the broader literature.

Optimization over (subsets of) the Wasserstein space (the metric space of probability measures over $\mathbb{R}^d$ endowed with the 2-Wasserstein distance, see Section 7.2) has found diverse and effective applications in modern machine learning. Notable examples include distributionally robust optimization (Kuhn et al., 2019; Yue et al., 2022), the computation of barycenters (Altschuler et al., 2021; Backhoff-Veraguas et al., 2022; Chewi et al., 2020; Cuturi and Doucet, 2014; Zemel and Panaretos, 2019), sampling (Chewi, 2024; Jordan et al., 1998; Wibisono, 2018), and variational

inference (see below). The development of optimization algorithms over this space, however, has been hindered by significant implementation challenges stemming from its infinite-dimensional nature and the curse of dimensionality which impedes efficient representation of high-dimensional distributions.

To alleviate these hurdles, a popular approach is to restrict the optimization to tractable subfamilies of probability distributions, such as finite-dimensional parametric families. Note that this is in contrast to Euclidean optimization, in which constraint sets are typically imposed as part of the problem (e.g., affine constraints in operations research). Here we view the use of a constraint set in the Wasserstein space as a *design choice*, with the end goals of flexibility, interpretability, and computational tractability.

An important motivating example is that of variational inference (VI), which seeks the best approximation to a probability measure π over $\mathbb{R}^d$ in the sense of KL divergence over some subset of probability measures C:

$$\pi^\star \in \operatorname*{argmin}_{\mu \in C} \mathrm{KL}(\mu \| \pi) = \operatorname*{argmin}_{\mu \in C} \int \log\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big)\, \mathrm{d}\mu\,. \tag{7.1}$$

For example, C could be taken to be the class of non-degenerate Gaussian distributions, in which case (7.1) is known as Gaussian VI. Recently, by leveraging the rich theory of gradient flows over the Wasserstein space, Diao et al. (2023); Lambert et al. (2022) provided algorithmic guarantees for Gaussian VI under standard tractability assumptions, i.e., strong log-concavity and log-smoothness of π.

We instead study the problem of *mean-field* VI, in which C is taken to be the class of product measures over $\mathbb{R}^d$, written $\mathcal{P}(\mathbb{R})^{\otimes d}$. In this context, the works by Lacker (2023); Yao and Yang (2022); Zhang and Zhou (2020) have also developed algorithms based on Wasserstein gradient flows, although computational guarantees for VI are still nascent (see Section 7.5.1 for further details and comparison with the literature).

The main result of our work is to provide computational guarantees under the usual tractability assumptions for π. Our approach is to replace the set of product measures by a smaller, "*polyhedral*" subset $\mathcal{P}_\diamond$ which we prove is an accurate approximation to $\mathcal{P}(\mathbb{R})^{\otimes d}$, in the sense that the minimizer $\pi^\star$ of (7.1) is in fact close to the KL minimizer $\pi^\star_\diamond$ over $\mathcal{P}_\diamond$ with quantifiable approximation rates. This motivates our development of a theory of polyhedral optimization over the Wasserstein space which, when applied to the mean-field VI problem, furnishes algorithms for minimization of the KL divergence over $\mathcal{P}_\diamond$ with theoretical (even accelerated) guarantees. More broadly, we are hopeful that the success of polyhedral optimization for mean-field VI will encourage the further use of polyhedral constraint sets to model other problems of interest.

We discuss the implementation of our algorithm in Section 7.5.5.1, with code available here. Below, we describe our contributions in more detail.

7.1.1 Main contributions

Polyhedral optimization in the Wasserstein space. We study parametric sets of the following form:

$$\operatorname{cone}(\mathcal{M})_\sharp\rho := \left\{ \left(\textstyle\sum_{T\in\mathcal{M}} \lambda_T T\right)_\sharp \rho \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \right\},$$

where $\mathbb{R}_+^{|\mathcal{M}|}$ is the non-negative orthant, $\mathcal{M}$ is a family of user-chosen optimal transport maps, and ρ is a fixed, known, reference measure. To our knowledge, such sets have not previously appeared in the literature.

Before proceeding, however, we must dispel a potential source of confusion: although $\operatorname{cone}(\mathcal{M})$ is a convex subset of the space of optimal transport maps at ρ—in other words, a convex subset of the *tangent space* to Wasserstein space at ρ—the set $\operatorname{cone}(\mathcal{M})_\sharp\rho$ is *not always* a convex subset of the Wasserstein space itself, in the sense of being closed under Wasserstein geodesics. For this to hold, we impose a further condition on $\mathcal{M}$, known as *compatibility* (Boissard et al., 2015).

Although compatibility is restrictive, it is nevertheless powerful enough to capture our application to mean-field VI described below. We refer to the set $\mathrm{cone}(\mathcal{M})_\sharp\rho$, for a compatible family $\mathcal{M}$, as a *polyhedral* subset of the Wasserstein space (or more specifically, a cone).

The assumption of compatibility entails strong consequences: we show that in fact, the set $(\mathrm{cone}(\mathcal{M})_\sharp\rho, W_2)$ is isometric to $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$, where $\|\cdot\|_Q$ is a *Euclidean* norm. This isometry allows us to optimize functionals over $\mathrm{cone}(\mathcal{M})_\sharp\rho$ via lightweight first-order algorithms for Euclidean optimization in lieu of Wasserstein optimization, which often requires computationally burdensome approximation schemes such as interacting particle systems. In particular, we can apply projected gradient descent or incorporate faster, *accelerated* methods. Moreover, under the isometry, convex subsets of $\mathrm{cone}(\mathcal{M})$ map to convex subsets of $\mathrm{cone}(\mathcal{M})_\sharp\rho$, giving rise to a bevy of geodesically convex constraint sets over which tractable optimization is feasible. This includes Wasserstein analogues of polytopes, to which we can apply the projection-free Frank–Wolfe algorithm. We show that as soon as the objective functional $\mathcal{F}$ is geodesically convex and smooth, these algorithms inherit the usual rates of convergence from the convex optimization literature.

Application to mean-field VI. We next turn toward mean-field VI as a compelling application of our theory of polyhedral optimization. Throughout, we only assume that π satisfies the standard assumptions of strong log-concavity and log-smoothness. By leveraging the structure of the mean-field VI solution and establishing regularity bounds for optimal transport maps between well-conditioned product measures, we first prove an approximation result which shows that the solution $\pi^\star$ to mean-field VI in (7.1) is well-approximated by the minimizer $\pi^\star_\diamond$ of the KL divergence over a suitable polyhedral approximation $\mathcal{P}_\diamond$ of the space of product measures. Importantly, our approximation rates, owing to the coordinate-wise decomposability of mean-field VI, do not incur the curse of dimensionality.

Next, we establish the geodesic strong convexity and geodesic smoothness of the KL divergence over $\mathcal{P}_\diamond$. Consequently, bringing to bear the full force of the Euclidean–Wasserstein equivalence,

we obtain, to the best of our knowledge, the first *end-to-end* convergence rates for mean-field VI.

7.1.2 Related work

To the best of our knowledge, our introduction of polyhedral sets and theory of polyhedral optimization over the Wasserstein space are novel. A special case of our set is

$$\operatorname{conv}(\mathcal{M})_\sharp \rho := \left\{ \left(\textstyle\sum_{T \in \mathcal{M}} \lambda_T T \right)_\sharp \rho \;\middle|\; \lambda \in \Delta_{|\mathcal{M}|} \right\},$$

where $\Delta_{|\mathcal{M}|}$ is the $|\mathcal{M}|$-simplex. Such a constraint set is used by Boissard et al. (2015); Gunsilius et al. (2024); Werenski et al. (2022), and is usually studied in the context of Wasserstein barycenters. The work of Bonneel et al. (2016) also considers $\operatorname{conv}(\mathcal{M})_\sharp \rho$, but makes no assumptions on the maps, and they tackle the problem from a computational angle via Sinkhorn's algorithm (Cuturi, 2013; Peyré and Cuturi, 2019), albeit without convergence guarantees. Albergo et al. (2024) use the same set, but without incorporating any optimal transport theory.

Our approach to mean-field VI, which parameterizes the variational family as the pushforward of a reference measure via transport maps, has its roots in the literature on generative modeling and normalizing flows (Chen et al., 2018; Finlay et al., 2020a;b; Huang et al., 2021a). We provide further background information and literature on mean-field VI in Section 7.5.1, and omit it here to avoid redundancies.

Finally, we mention that our work falls under the category of *linearized optimal transport* (Wang et al., 2013), which we closely address in Section 7.3.2.

7.2 Background on optimal transport

In this section, we provide background on optimal transport relevant to our work and refer to Santambrogio (2015); Villani (2009) for details. Throughout, we assume that all probability

measures admit a density function with respect to Lebesgue measure. We let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of probability measures with density over $\mathbb{R}^d$ with finite second moment.

For $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, the squared 2-Wasserstein distance is written as

$$W_2^2(\rho, \mu) = \inf_{T : T_\sharp \rho = \mu} \int \|x - T(x)\|_2^2 \, \mathrm{d}\rho(x) \,, \tag{7.2}$$

where the collection $\{T : T_\sharp \rho = \mu\}$ is the set of all valid transport maps: for $X \sim \rho$, $T(X) \sim \mu$.

Since we assumed ρ has a density, Brenier's theorem (Brenier, 1991) states that there exists a unique minimizer to (7.2), called the *optimal transport map* $T_\star$ between ρ and μ. Further, $T_\star = \nabla \varphi_\star$ for some convex function $\varphi_\star$, called a Brenier potential.

Additionally, since μ also has a density, then there exists an optimal transport map between μ and ρ, given by $\nabla \varphi_\star^* = (T_\star)^{-1}$, where $\varphi_\star^*(y) \coloneqq \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \varphi_\star(x)\}$ is the Fenchel conjugate of $\varphi_\star$. For more information on (differentiable) convex functions and conjugacy, we suggest Hiriart-Urruty and Lemaréchal (2004); Rockafellar (1997).

Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is *m-strongly convex* in some norm $\|\cdot\|$ if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|x - y\|^2 \,, \qquad x, y \in \mathbb{R}^d \,,$$

and *M-smooth* in some norm $\|\cdot\|$ if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|x - y\|^2 \,, \qquad x, y \in \mathbb{R}^d \,,$$

where $m, M > 0$.

For two probability measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, let $\nabla \varphi^{0 \to 1}$ denote the optimal transport map from μ_0 to μ_1. The *(unique) constant-speed geodesic* between μ_0 and μ_1 is given by the curve

$(\mu_t)_{t\in[0,1]}$, with

$$\mu_t = (\nabla\varphi_t)_\sharp\mu_0 := (\mathrm{id} + t\,(\nabla\varphi^{0\to 1} - \mathrm{id}))_\sharp\mu_0\,. \tag{7.3}$$

If we equip $\mathcal{P}_2(\mathbb{R}^d)$, the space of probability distributions with finite second moment over $\mathbb{R}^d$, with the 2-Wasserstein distance, we obtain a metric space $\mathbb{W} := (\mathcal{P}_2(\mathbb{R}^d), W_2)$ (Villani, 2021, Theorem 7.3), which we call the *Wasserstein space*. In fact, it can be formally viewed as a Riemannian manifold over which one can define gradient flows of functionals (Otto, 2001). We refer the interested reader to consult the background sections of Altschuler et al. (2021) or to Chewi (2024) for a light exposition and further details.

The Riemannian structure of the Wasserstein space is crucial for the development of optimization over this space, as it furnishes appropriate Wasserstein analogues of basic concepts from Euclidean optimization, such as the gradient mapping, convexity, and smoothness. In particular, we say that a subset C of the Wasserstein space is *geodesically convex* if it is closed under taking geodesics (7.3). Also, a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is *geodesically (strongly) convex* (resp. *geodesically smooth*) if the map $[0,1] \to \mathbb{R}$, $t \mapsto \mathcal{F}(\mu_t)$ is (strongly) convex (resp. smooth) along every constant-speed geodesic $(\mu_t)_{t\in[0,1]}$.

7.3 Polyhedral sets in the Wasserstein space

In this section, we establish properties of the constraint set

$$\mathrm{cone}(\mathcal{M})_\sharp\rho := \left\{ \left(\textstyle\sum_{T\in\mathcal{M}} \lambda_T T\right)_\sharp\rho \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \right\}, \tag{7.4}$$

with respect to the known base measure ρ and a fixed set of optimal transport maps $\mathcal{M}$. Typically, we have in mind finite $\mathcal{M}$, in which case (7.4) is valid. Otherwise, (7.4) should be modified to range only over λ with finitely many non-zero coordinates, or in other words, $\mathrm{cone}(\mathcal{M})$ is the

smallest set containing all conic combinations of maps in $\mathcal{M}$.

Despite its simplicity, we argue that the geometry of $\text{cone}(\mathcal{M})_\sharp\rho$ is surprisingly deceptive. Most strikingly, it is *not* always a geodesically convex set. Consider $T_1(x) = x$, $T_2(x) = A^{1/2}x$, and $T_3(x) = B^{1/2}x$, with $\rho = \mathcal{N}(0, I)$, the standard Gaussian in $\mathbb{R}^d$, and $A, B \succ 0$. In this setting, $\text{cone}(\mathcal{M})_\sharp\rho$ is the following set of Gaussians:

$$\text{cone}(\mathcal{M})_\sharp\rho = \left\{\mathcal{N}(0,\ (\lambda_1 I + \lambda_2 A^{1/2} + \lambda_3 B^{1/2})^2) \;\middle|\; \lambda \in \mathbb{R}^3_+\right\}. \tag{7.5}$$

One can check with virtually any randomly generated positive definite matrices A and B that, as long as all three matrices I, A, B are not mutually diagonalizable, the geodesic between $\mathcal{N}(0, A)$ and $\mathcal{N}(0, B)$ does not lie in (7.5). This simple example illustrates that some care is required in order to define convex constraint sets in the Wasserstein space.

7.3.1 Compatible families of transport maps

In the Gaussian example above, geodesic convexity of $\text{cone}(\mathcal{M})_\sharp\rho$ is recovered if we additionally assume that I, A, and B are mutually diagonalizable. This reflects a certain property of the maps T_1, T_2, T_3, which can be generalized to a property known as *compatibility* (Boissard et al., 2015). We recall its definition and basic properties in the sequel. As always, we assume that ρ admits a density with respect to Lebesgue measure.

Let $\mathcal{M}$ be a set of bijective vector-valued maps, given by gradients of convex functions. We call the set of maps $\mathcal{M}$ *compatible* if

$$\text{for all } T_1, T_2 \in \mathcal{M}, \quad T_1 \circ (T_2)^{-1} \text{ is the gradient of a convex function.}$$

Compatibility is a fundamental notion which lies at the heart of numerous other works (see Bigot et al., 2017; Boissard et al., 2015; Cazelles et al., 2018; Chewi et al., 2021; Panaretos and Zemel,

2016; Werenski et al., 2022). See Panaretos and Zemel (2020) for details.

The main motivation for compatibility is the following theorem.

Theorem 7.1 (Compatibility induces geodesic convexity)**.** *Suppose that $\mathcal{M}$ is compatible. Then,* $\operatorname{cone}(\mathcal{M})_\sharp\rho$ *is a geodesically convex set. Moreover, for any convex subset* $\mathcal{K} \subseteq \operatorname{cone}(\mathcal{M})$*, the set* $\mathcal{K}_\sharp\rho$ *is a geodesically convex set.*

Although this result is not difficult to prove, we were unable to find it in the existing literature. In fact, it follows as a direct consequence of the isometry established in Section 7.3.2, which will show that $\operatorname{cone}(\mathcal{M})_\sharp\rho$ is isometric to a convex subset of a Hilbert space.

Motivated by this theorem, we propose the following definition.

Definition 7.2. Let $\mathcal{M}$ be a compatible and finite family of optimal transport maps. We refer to $\operatorname{cone}(\mathcal{M})_\sharp\rho$ as a *polyhedral set* in the Wasserstein space.

More generally, a polyhedral set in the Wasserstein space is a set of the form $\mathcal{K}_\sharp\rho$ where $\mathcal{K} \subseteq \operatorname{cone}(\mathcal{M})$ is polyhedral and $\mathcal{M}$ is a compatible family.

The next sequence of lemmas furnish important examples of compatible families, which we prove in Appendix F.1.

Lemma 7.3 (Mutually diagonalizable linear maps)**.** *Let $\mathcal{M}$ be a family of mutually diagonalizable and positive definite linear maps $\mathbb{R}^d \to \mathbb{R}^d$. Then, $\mathcal{M}$ is a compatible family.*

Lemma 7.4 (Radial maps)**.** *Let*

$$\mathcal{M} := \{x \mapsto g(\|x\|_2)\, x \mid g : \mathbb{R}_+ \to \mathbb{R}_+ \text{ is continuous and strictly increasing}\}\,.$$

Then, $\mathcal{M}$ is a compatible family.

Lemma 7.5 (One-dimensional maps). *Let $\mathcal{M}$ denote the family of continuous and increasing*[1] *functions $\mathbb{R} \to \mathbb{R}$. Then, $\mathcal{M}$ is a compatible family.*

Lemma 7.6 (Direct sum). *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be compatible families of maps on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively. Then, $\mathcal{M} \coloneqq \{(x_1, x_2) \mapsto (T_1(x_1), T_2(x_2)) \mid T_1 \in \mathcal{M}_1,\ T_2 \in \mathcal{M}_2\}$ is a compatible family of maps on $\mathbb{R}^{d_1+d_2}$.*

Lemma 7.7 (Adding the identity). *Let $\mathcal{M}$ be a compatible family. Then, $\mathcal{M} \cup \{\mathrm{id}\}$ is a compatible family.*

Lemma 7.8 (Adding translations). *Let $\mathcal{M}$ be a compatible family of maps on $\mathbb{R}^d$. Then, $\{x \mapsto T(x) + v \mid T \in \mathcal{M},\ v \in \mathbb{R}^d\}$ is a compatible family of maps.*

Lemma 7.9 (Cones). *Let $\mathcal{M}$ be a compatible family. Then, $\mathrm{cone}(\mathcal{M})$ is a compatible family.*

In the sequel, we will use these results in order to build rich compatible families, especially with an eye toward approximating coordinate-wise separable maps which arise in mean-field VI (see Section 7.5.3). In particular, Lemma 7.9 is the starting point for the development of our theory of polyhedral optimization in the Wasserstein space.

Remark 7.10. In our applications of interest, $\mathrm{cone}(\mathcal{M})$ is typically constructed as follows: let $\mathcal{M}_1, \ldots, \mathcal{M}_d$ be univariate compatible families (Lemma 7.5). We then take $\mathrm{cone}(\mathcal{M})$ to be the cone generated by the direct sum of $\mathcal{M}_1, \ldots, \mathcal{M}_d$ via Lemma 7.6. It is easy to see that a generating family of this cone is the set of maps $x \mapsto (0, \ldots, 0, T_i(x_i), 0, \ldots, 0)$, where $T_i \in \mathcal{M}_i$. This is a finite family of size $\sum_{i=1}^{d} |\mathcal{M}_i|$.

[1] Technically, $\mathcal{M}$ does not consist of *bijective* maps, which we required in the definition of compatibility. In one dimension, however, the notion of compatibility still makes sense once we replace the inverse function with the quantile function.

7.3.2 Isometry with Euclidean geometry

A key consequence of compatibility is that the Wasserstein distance equals the *linearized optimal transport distance* with respect to ρ, i.e., for $T, \tilde{T} \in \mathcal{M}$,

$$\mathrm{d}^2_{\mathrm{LOT},\rho}(T_\sharp\rho, \tilde{T}_\sharp\rho) := \|\tilde{T} - T\|^2_{L^2(\rho)} = \|\tilde{T} \circ T^{-1} - \mathrm{id}\|^2_{L^2(T_\sharp\rho)} = W_2^2(T_\sharp\rho, \tilde{T}_\sharp\rho)\,, \tag{7.6}$$

where we applied compatibility in the last equality to argue that $\tilde{T} \circ T^{-1}$ is the optimal transport map from $T_\sharp\rho$ to $\tilde{T}_\sharp\rho$. This equality shows that for compatible $\mathcal{M}$, the geometry of $\mathrm{cone}(\mathcal{M})_\sharp\rho$ is in a sense trivial, being isometric to a convex subset of the Hilbert space $L^2(\rho)$. This fundamental property lies at the heart of the widespread usage of one-dimensional optimal transport in applications, see Basu et al. (2014); Cai et al. (2020); Khurana et al. (2023); Kolouri and Rohde (2015); Kolouri et al. (2016); Park and Thorpe (2018); Wang et al. (2013) for applications.

Next, we consider a family of the form $\mathrm{cone}(\mathcal{M})$, where $\mathcal{M}$ is finite. By its very definition, $\mathrm{cone}(\mathcal{M})$ is naturally parameterized by the non-negative orthant. Henceforth, we write

$$T^\lambda := \sum_{T\in\mathcal{M}} \lambda_T T\,, \qquad \mu_\lambda := (T^\lambda)_\sharp\rho\,.$$

We can therefore consider the induced metric on $\mathbb{R}_+^{|\mathcal{M}|}$. A straightforward calculation reveals:

$$\mathrm{d}^2_{\mathrm{LOT},\rho}(\mu_\eta, \mu_\lambda) = \|\textstyle\sum_{T\in\mathcal{M}}(\eta_T - \lambda_T)\,T\|^2_{L^2(\rho)} = (\eta-\lambda)^\top Q(\eta-\lambda) = \|\eta-\lambda\|^2_Q\,,$$

where the matrix Q has entries $Q_{T,\tilde{T}} := \langle T, \tilde{T}\rangle_{L^2(\rho)}$ for $T, \tilde{T} \in \mathcal{M}$. Here, Q is nothing more than a Gram matrix, which is always positive semi-definite. This collection of observations proves the following result.

Theorem 7.11. *Let $\mathcal{M}$ be a finite family of optimal transport maps with Q defined as the Gram matrix with entries $Q_{T,\tilde{T}} = \langle T, \tilde{T}\rangle_{L^2(\rho)}$ for $T, \tilde{T} \in \mathcal{M}$. Then, $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$ is always isometric to*

$(\mathrm{cone}(\mathcal{M})_\sharp\rho, \mathrm{d}_{\mathrm{LOT},\rho})$. *If, in addition,* $\mathcal{M}$ *is a compatible family (i.e.,* $\mathrm{cone}(\mathcal{M})_\sharp\rho$ *is polyhedral), then* $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$ *is isometric to* $(\mathrm{cone}(\mathcal{M})_\sharp\rho, W_2)$.

As we develop in the next section, Theorem 7.11 paves the way for the application of scalable first-order Euclidean optimization algorithms for minimization problems over polyhedral subsets of the Wasserstein space.

7.4 Polyhedral optimization in the Wasserstein space

Let $\mathrm{cone}(\mathcal{M})_\sharp\rho$ be polyhedral and recall the Gram matrix Q from Theorem 7.11, with entries given by $Q_{T,\tilde{T}} = \langle T, \tilde{T}\rangle_{L^2(\rho)}$. We now turn toward the problem of minimizing a functional $\mathcal{F}$ over $\mathrm{cone}(\mathcal{M})_\sharp\rho$. Henceforth, we assume that Q is positive definite, so that Q^{-1} exists. The positive definiteness of Q follows if the maps $T \in \mathcal{M}$ are linearly independent in $L^2(\rho)$.

7.4.1 Continuous-time gradient flow

The isometry of Section 7.3.2 implies that the constrained Wasserstein gradient flow of $\mathcal{F}$ is equivalent to the gradient flow of the functional $\lambda \mapsto \mathcal{F}(\mu_\lambda)$ with respect to the Q-geometry.[2] The latter gradient flow can be written explicitly as

$$\dot{\lambda}(t) = -Q^{-1}\,\nabla_\lambda \mathcal{F}(\mu_{\lambda(t)})\,. \tag{7.7}$$

Then, geodesic strong convexity over $\mathbb{W}$ translates to strong convexity of $\lambda \mapsto \mathcal{F}(\mu_\lambda)$ over $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$ for free. The following theorem[3] establishes convergence rates for this continuous-time flow; see Lambert et al. (2022, Appendix D) for a proof.

[2] See Nesterov (2018, §4.2.1) for a thorough discussion on optimization over general Euclidean spaces.

[3] In the case where we further constrain the gradient flow to lie in a convex set, (7.7) should be replaced by a differential inclusion. Since this is not relevant to the subsequent developments, we omit a fuller discussion of this point.

Theorem 7.12. *Suppose $\mathcal{F}$ is geodesically m-strongly convex over $\mathbb{W}$, for $m \geq 0$. Let $\mathrm{cone}(\mathcal{M})_\sharp\rho$ be polyhedral. Then, $\mathcal{F}$ is geodesically m-strongly convex over $\mathrm{cone}(\mathcal{M})_\sharp\rho$. Moreover, if $\mu_\star \equiv \mu_{\lambda^\star} \in \mathrm{cone}(\mathcal{M})_\sharp\rho$ is a minimizer of $\mathcal{F}$ over $\mathrm{cone}(\mathcal{M})_\sharp\rho$, the following convergence rates hold for the gradient flow (7.7).*

1. *If $m = 0$, then $\mathcal{F}(\mu_{\lambda(t)}) - \mathcal{F}(\mu_\star) \leq \frac{1}{2t} W_2^2(\mu_{\lambda(0)}, \mu_\star)$.*
2. *If $m > 0$, then:*
   (a) $W_2^2(\mu_{\lambda(t)}, \mu_\star) \leq \exp(-2mt)\, W_2^2(\mu_{\lambda(0)}, \mu_\star)$.
   (b) $\mathcal{F}(\mu_{\lambda(t)}) - \mathcal{F}(\mu_\star) \leq \exp(-2mt)\, (\mathcal{F}(\mu_{\lambda(0)}) - \mathcal{F}(\mu_\star))$.

7.4.2 Time-discretization made easy

Appealing to the isometry in Section 7.3.2, optimization of a geodesically convex and geodesically smooth functional $\mathcal{F}$ over a polyhedral set $\mathrm{cone}(\mathcal{M})_\sharp\rho$ boils down to a finite-dimensional, convex, smooth, *Euclidean* optimization problem of the form

$$\min_{\lambda \in \mathbb{R}_+^{|\mathcal{M}|}} \mathcal{F}(\mu_\lambda)\,. \tag{7.8}$$

More generally, we consider optimization over arbitrary convex subsets $K \subseteq \mathbb{R}_+^{|\mathcal{M}|}$, and we let $\mathcal{K} \coloneqq \{T^\lambda \mid \lambda \in K\}$ denote the corresponding subset of $\mathrm{cone}(\mathcal{M})$. It leads to the problem

$$\min_{\lambda \in K} \mathcal{F}(\mu_\lambda)\,.$$

Our consideration of general constraint sets K is not purely for the sake of generality, as we in fact use the full power of polyhedral optimization in our application to mean-field VI (in particular, see Theorem 7.24 and Appendix F.3.3).

We consider accelerated projected gradient descent (Beck, 2017), as well as stochastic projected gradient descent which is useful when only a stochastic gradient is available (as in Section 7.5.5.2). Moreover, when restricted to any *polytope* in the non-negative orthant, we also consider the projection-free Frank–Wolfe algorithm (Frank and Wolfe, 1956). We briefly describe the algorithms and state their corresponding convergence guarantees. Note that we could also port over guarantees for other Euclidean optimization algorithms in a similar manner, but we omit them for brevity.

7.4.2.1 Accelerated projected gradient descent

Starting at an initial point $\lambda^{(0)} \in K$, we can solve (7.8) by applying a projected variant of Nesterov's accelerated gradient descent method (Nesterov, 1983), a well-known extrapolation technique that improves upon the convergence rate for projected gradient descent and is optimal for smooth convex optimization (Nemirovski and Yudin, 1983). The algorithm is given as Algorithm 2. Here, $\mathrm{proj}_{K,Q}(\cdot)$ is the orthogonal projection operator onto K with respect to the $\|\cdot\|_Q$ norm.

We summarize the following well-known convergence results for accelerated projected gradient descent (APGD) below; see Beck (2017, Chapter 10) for proofs.

Theorem 7.13 (Convergence results for APGD)**.** *Let* $\mathrm{cone}(\mathcal{M})_\sharp\rho$ *be polyhedral and* $\mathcal{K} \subseteq \mathrm{cone}(\mathcal{M})$ *be convex. Suppose that* $\mathcal{F}$ *is geodesically* m*-strongly convex and* M*-smooth over* $\mathcal{K}_\sharp\rho$ *and let* $\mu_\star$ *denote a minimizer over this set. Let* $(\lambda^{(t)} : t = 0, 1, 2, 3 \ldots)$ *denote the iterates of Algorithm 2.*

1. *If* $m = 0$*, then* $\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \lesssim Mt^{-2}\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star)$.

2. *If* $m > 0$*, then for* $\kappa := M/m$*,*

   (a) $W_2^2(\mu_{\lambda^{(t)}}, \mu_\star) \lesssim \kappa \exp(-t/\sqrt{\kappa})\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star)$.

   (b) $\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \leq (1 - 1/\sqrt{\kappa})^t \left(\mathcal{F}(\mu_{\lambda^{(0)}}) - \mathcal{F}(\mu_\star) + \frac{m}{2}\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star)\right)$.

Algorithm 2: Accelerated projected gradient descent over cone($\mathcal{M}$)

Input: $\lambda^{(0)} \in K$, functional $\mathcal{F}$ (m-convex and M-smooth in W_2), compatible family $\mathcal{M}$
Initialize: $\eta^{(0)} = \lambda^{(0)}$, $\kappa \leftarrow M/m$ if $m > 0$, and $\gamma_{(0)} = 1$ if $m = 0$
for $t = 0, 1, 2, 3, \ldots$ **do**
  $\lambda^{(t+1)} \leftarrow \mathrm{proj}_{K,Q}(\eta^{(t)} - \frac{1}{M} Q^{-1} \nabla_\lambda \mathcal{F}(\mu_{\eta^{(t)}}))$
  if $m > 0$ **then**
    $\eta^{(t+1)} \leftarrow \lambda^{(t+1)} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\lambda^{(t+1)} - \lambda^{(t)})$
  else
    $\gamma_{(t+1)} \leftarrow \frac{1+\sqrt{1+4\gamma_{(t)}^2}}{2}$
    $\eta^{(t+1)} \leftarrow \lambda^{(t+1)} + \left(\frac{\gamma_{(t)}-1}{\gamma_{(t+1)}}\right)(\lambda^{(t+1)} - \lambda^{(t)})$
  end
end

7.4.2.2 Stochastic projected gradient descent

In some situations, the full gradient $\nabla_\lambda \mathcal{F}(\mu_\lambda)$ cannot be computed, usually due to high computational costs. Instead, *stochastic* first-order methods alleviate this issue by instead allowing for the use of an unbiased stochastic gradient oracle, written $\hat{\nabla}_\lambda \mathcal{F}(\mu_\lambda)$.[4] The decreased computational overhead has contributed to the widespread use of stochastic gradient methods as a pillar of modern machine learning (Bubeck, 2015). We limit our discussions to the case where $\mathcal{F}$ is smooth and strongly convex, as this setting will be the most relevant later. Other settings readily generalize, though we omit them for brevity.

We provide a description of stochastic projected gradient descent (SPGD) in Algorithm 3, and convergence analysis in Theorem 7.14 which requires the following standard assumption on the variance of the unbiased estimator:

(VB) There exist constants $c_0, c_1 \geq 0$ such that for any $\lambda \in K$, the gradient estimate satisfies

$$\mathbb{E}[\|Q^{-1}(\hat{\nabla}_\lambda \mathcal{F}(\mu_\lambda) - \nabla_\lambda \mathcal{F}(\mu_\lambda))\|_Q^2] \leq c_0 + c_1 \mathbb{E}[W_2^2(\mu_\lambda, \mu_\star)] .$$

[4] An unbiased estimator of the gradient is one which $\mathbb{E}_{\mu_\lambda}[\hat{\nabla}_\lambda \mathcal{F}(\mu_\lambda)] = \nabla_\lambda \mathcal{F}(\mu_\lambda)$.

Algorithm 3: Stochastic projected gradient descent over $\mathrm{cone}(\mathcal{M})$

Input: $\lambda^{(0)} \in K$, functional $\mathcal{F}$ (m-convex and M-smooth in W_2), compatible family $\mathcal{M}$, fixed step-size $h > 0$, and unbiased stochastic gradient oracle $\hat{\nabla}_\lambda \mathcal{F}(\cdot)$

for $t = 0, 1, 2, 3, \ldots$ **do**

$\quad \lambda^{(t+1)} \leftarrow \mathrm{proj}_{K,Q}(\lambda^{(t)} - h\, Q^{-1}\, \hat{\nabla}_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}))$

end

Note that c_0, c_1 in (VB) will typically depend on the smoothness and strong convexity parameters of $\mathcal{F}$, and possibly the dimension of the problem.

Theorem 7.14 (Convergence results for SPGD). *Let* $\mathrm{cone}(\mathcal{M})_\sharp \rho$ *be polyhedral and* $\mathcal{K} \subseteq \mathrm{cone}(\mathcal{M})$ *be convex. Suppose that* $\mathcal{F}$ *is geodesically* m*-strongly convex and* M*-smooth over* $\mathcal{K}_\sharp \rho$*, let* $\mu_\star$ *denote a minimizer over this set, and suppose that* (VB) *holds. Let* $(\lambda^{(t)} : t = 0, 1, 2, 3 \ldots)$ *denote the iterates of Algorithm 3 and let* $\varepsilon > 0$ *be sufficiently small. If we choose step size* $h \asymp \frac{m\varepsilon^2}{c_0} \leq \frac{m}{2c_1} \wedge \frac{1}{2\kappa M}$*, and the number of iterations is at least*

$$t \gtrsim \frac{c_0}{m^2 \varepsilon^2} \log(W_2(\mu_{\lambda^{(0)}}, \mu_\star)/\varepsilon)\,,$$

then $\mathbb{E}[W_2^2(\mu_{\lambda^{(t)}}, \mu_\star)] \leq \varepsilon^2$.

For completeness, we provide a short proof of Theorem 7.14 in Appendix F.2.

7.4.2.3 Frank–Wolfe

In this section, we consider optimization over a *polytope*, i.e., a set of the form

$$\mathrm{conv}(\mathcal{M})_\sharp \rho := \left\{ \left(\textstyle\sum_{T \in \mathcal{M}} \lambda_T T \right)_\sharp \rho \;\middle|\; \lambda \in \Delta_{|\mathcal{M}|} \right\},$$

where $\mathcal{M}$ is a finite family of compatible maps and $\Delta_{|\mathcal{M}|}$ denotes the $|\mathcal{M}|$-dimensional simplex. Note that $\mathrm{conv}(\mathcal{M}) \subseteq \mathrm{cone}(\mathcal{M})$, where $\mathrm{cone}(\mathcal{M})_\sharp \rho$ is polyhedral, so that $\mathrm{conv}(\mathcal{M})$ is an example of a convex constraint set $\mathcal{K}$ considered in the previous subsection. The convergence guarantees

for accelerated projection gradient descent in Theorem 7.13 therefore apply to optimization over $\mathrm{conv}(\mathcal{M})_\sharp\rho$.

In this setting, however, there is a popular alternative to projected gradient descent known as *conditional gradient descent* or the *Frank–Wolfe* (FW) algorithm (Frank and Wolfe, 1956; Jaggi, 2013). In this scheme, we find a descent direction that ensures our iterates remain within the constraint set. This direction $\eta^{(t)}$ is found at each iterate $\lambda^{(t)}$ by solving the following linear sub-problem:

$$\eta^{(t)} = \operatorname*{argmin}_{\eta \in \Delta_{|\mathcal{M}|}} \langle \nabla_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}), \eta - \lambda^{(t)} \rangle \,. \tag{7.9}$$

Finding this direction can be substantially cheaper than the projection step in Algorithm 2. Indeed, the sub-problem (7.9) does not depend on the matrix Q. It is not hard to see that the minimizer $\eta^{(t)}$ must be attained at one of the $|\mathcal{M}|$ vertices of the simplex.

The full algorithm is presented in Algorithm 4. Known results provide sublinear convergence of the objective gap, which does not improve under strong convexity assumptions; see Beck (2017, Chapter 13) for proofs and discussions.

Theorem 7.15 (Convergence results for FW)**.** *Suppose that $\mathcal{F}$ is geodesically convex and M-smooth over $\mathrm{conv}(\mathcal{M})_\sharp\rho$, and let $\mu_\star$ be a minimizer of $\mathcal{F}$ over this set. Let $(\lambda^{(t)} : t = 0, 1, 2, 3 \ldots)$ denote the iterates of Algorithm 4, with step size $\alpha^{(t)} = 2/(t+2)$. Then,*

$$\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \lesssim M t^{-1} \operatorname{diam}\big(\mathrm{conv}(\mathcal{M})_\sharp\rho\big)^2 \,. \tag{7.10}$$

Via the isometry in Section 7.3.2, $\operatorname{diam}(\mathrm{conv}(\mathcal{M})_\sharp\rho)$ equals the diameter of $\mathrm{conv}(\mathcal{M})$ in the Q-norm. In terms of the matrix Q, this is at most $2 \max_{T \in \mathcal{M}} \sqrt{Q_{T,T}}$.

Remark 7.16. We are not the first to consider applying FW over the Wasserstein space. Kent et al. (2021) use FW to optimize functionals over the constraint set $\{W_2(\cdot, \pi) \le \delta\}$ for some

Algorithm 4: Frank–Wolfe over $\operatorname{conv}(\mathcal{M})$

Input: $\lambda^{(0)} \in \Delta_{|\mathcal{M}|}$, functional $\mathcal{F}$, and compatible family $\mathcal{M}$ with $|\mathcal{M}| = J$
for $t = 0, 1, 2, 3, \ldots$ **do**
  $j^* \leftarrow \operatorname{argmin}_{j \in [J]} \langle \nabla_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}), e_j - \lambda^{(t)} \rangle$
  $\lambda^{(t+1)} \leftarrow (1 - \alpha^{(t)})\, \lambda^{(t)} + \alpha^{(t)}\, e_{j^*}$ `// ` $\alpha^{(t)} = \frac{2}{t+2}$ `is a standard step size choice`
end

$\delta > 0$ and some fixed probability measure π. In their work, the optimization truly occurs in an infinite-dimensional space. The authors prove various discrete-time rates of convergence under noisy gradient oracles and Hölder smoothness of the objective function, among other general properties. The core difference between our works is the constraint set of interest, resulting in our algorithm being simpler. Indeed, our setup is purely parametric.

7.4.3 Enriching the family of compatible maps

When applying our polyhedral optimization framework to specific problems of interest, it is sometimes useful to first enrich the compatible family. For example, one notable advantage of doing so is that it increases the expressive power of the constraint set. Another example is that for our application to mean-field VI in Section 7.5, it will be necessary for us to ensure a uniform lower bound on the Jacobian derivatives of the maps in our family (i.e., they are gradients of *strongly convex* potentials).

The second issue can be addressed by adding $\alpha\,\mathrm{id}$ to each member of the family. Indeed, by Lemma 7.9 and Lemma 7.7, $\operatorname{cone}(\mathcal{M} \cup \{\mathrm{id}\})$ is a compatible family, and then we can restrict to the convex subset $\mathcal{K} \subseteq \operatorname{cone}(\mathcal{M} \cup \{\mathrm{id}\})$ corresponding to λ for which the coefficient λ_{id} in front of id is α. The guarantees of Section 7.4.2.1 then apply directly to optimization over $\mathcal{K}_\sharp \rho$. However, we prefer to handle the $\alpha\,\mathrm{id}$ term separately, and so we define the *cone generated by* $\mathcal{M}$

with tip $\alpha\,\mathrm{id}$ to be the family

$$\mathrm{cone}(\mathcal{M};\, \alpha\,\mathrm{id}) \coloneqq \alpha\,\mathrm{id} + \mathrm{cone}(\mathcal{M})\,.$$

Similarly, to address the first issue, we would like to enrich our compatible family by adding translations, via Lemma 7.8. To this end, we define our *augmented cone*, $\underline{\mathrm{cone}}(\mathcal{M})$ for short, to be

$$\underline{\mathrm{cone}}(\mathcal{M}) \coloneqq \left\{ \textstyle\sum_{T\in\mathcal{M}} \lambda_T T + v \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|},\ v \in \mathbb{R}^d \right\}.$$

Similarly, we define

$$\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id}) \coloneqq \alpha\,\mathrm{id} + \underline{\mathrm{cone}}(\mathcal{M})\,.$$

The augmented cone is parameterized by $(\lambda, v) \in \mathbb{R}_+^{|\mathcal{M}|} \times \mathbb{R}^d$. We may assume that each of the maps $T \in \mathcal{M}$ has mean zero under ρ, since this does not affect the augmented cone. Under this assumption, it is easy to see (c.f. the proof of Theorem 7.28) that we still obtain an isometry with a Euclidean metric: $W_2^2(\mu_{\eta,u}, \mu_{\lambda,v}) = \|\eta - \lambda\|_Q^2 + \|u - v\|^2$. In this setting, the first-order algorithms must be modified to compute the gradient and projection steps with respect to this metric.

Remark 7.17 (Broader impact of our framework). We now pause to briefly discuss the broader impact of polyhedral sets. We want to stress that, even without compatibility, our framework can be used to optimize functionals over any convex subset of the tangent space, provided that the functional is convex with respect to the *linearized* optimal transport distance. In turn, this is equivalent to requiring that the functional is convex along generalized geodesics (see Ambrosio et al., 2008, §9.2), which is typically the case when the functional is convex in the Wasserstein geometry; for example, it holds for the KL divergence with respect to a log-concave measure. This substantially expands the scope of applications as it allows for optimization over *any* convex subset of the tangent space, not just compatible ones.

7.5 Application to mean-field variational inference

As our main application of polyhedral optimization over the Wasserstein space, we turn to variational inference (VI) (Blei et al., 2017). In this framework, we are given access to an unnormalized probability measure, known as the posterior, written $\pi \propto \exp(-V)$, from which we wish to obtain samples for downstream tasks. In principle, one can draw samples from π via Markov chain Monte Carlo methods, but these have computational drawbacks, such as potentially long burn-in times. Instead, VI suggests to minimize the Kullback–Leibler (KL) divergence over a constraint set to obtain a proxy measure that is easy to sample from. Commonly used constraint sets in the literature include the space of non-degenerate Gaussians, location-scale families, mixtures of Gaussians, and the space of product measures.

For a general constraint set C, the VI optimization problem reads

$$\pi_C^\star \in \operatorname*{argmin}_{\mu \in C} \mathrm{KL}(\mu \| \pi) \coloneqq \operatorname*{argmin}_{\mu \in C} \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z\,, \tag{7.11}$$

where Z, the unknown normalizing constant of $\pi \propto \exp(-V)$ plays no part in the optimization problem. The following assumption, which will play a crucial role in our analyses in Section 7.5.3 and Section 7.5.4, is standard in the literature on log-concave sampling (Chewi, 2024):

(WC) π is ℓ_V-strongly log-concave and L_V-log-smooth, i.e., $\ell_V I \preceq \nabla^2 V \preceq L_V I$ for $\ell_V, L_V > 0$.

In brief, we say that π is *well-conditioned*. We denote by $\kappa \coloneqq L_V/\ell_V$ the condition number.

The following lemma allows us to refer to the unique minimizer of the VI problem, which follows from the strong geodesic convexity of the KL divergence (see the discussions around Proposition 7.27).

Lemma 7.18. *Suppose C is a geodesically convex subset of $\mathcal{P}_2(\mathbb{R}^d)$, and suppose that π is strongly log-concave. Then, there is a unique minimizer of $\mathrm{KL}(\cdot\|\pi)$ over C.*

Despite the widespread use of variational inference in numerous settings (see for example Blei et al., 2017; Wainwright and Jordan, 2008), explicit guarantees have only recently been established for a few constraint families. Recently, Diao et al. (2023); Lambert et al. (2022) obtained computational guarantees for Gaussian VI by way of constrained Wasserstein gradient flows. Domke (2020); Domke et al. (2023); Kim et al. (2023) considered VI for location-scale families and provided algorithmic guarantees, though they abstained from the gradient flow formalism. Subsequent work by Yi and Liu (2023) made this connection precise.

In the sequel, we develop end-to-end computational guarantees for mean-field variational inference. This is done in five stages:

1. Transfer assumptions on the posterior π, namely (WC), to the mean-field solution $\pi^\star$ (see Proposition 7.19 in Section 7.5.1).

2. Use the properties of $\pi^\star$ to obtain regularity properties of the optimal transport map $T^\star$ from the standard Gaussian measure to $\pi^\star$, via Caffarelli's contraction theorem and the Monge–Ampère equation (see Theorem 7.21 in Section 7.5.2).

3. Show that polyhedral sets in the Wasserstein space can approximate mean-field measures arbitrarily well, making use of the regularity properties of the optimal transport map, approximation theory, and Wasserstein calculus (see Theorem 7.23, Theorem 7.24, and Theorem 7.26 in Section 7.5.3).

4. Provide convergence guarantees for optimizing the KL divergence over these polyhedral sets (see Theorem 7.29 in Section 7.5.4).

5. Describe implementation details for our final algorithm (see Section 7.5.5.1).

7.5.1 Mean-field variational inference

In mean-field VI, the constraint set is the space of product measures over $\mathbb{R}^d$, written $\mathcal{P}(\mathbb{R})^{\otimes d}$. Thus, the optimization problem is

$$\pi^\star \in \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R})^{\otimes d}} \mathrm{KL}(\mu \| \pi)\,, \tag{7.12}$$

where, by design, the constraint set forces the minimizers to be of the form

$$\pi^\star(x_1, \ldots, x_d) = \bigotimes_{i=1}^{d} \pi_i^\star(x_i)\,. \tag{7.13}$$

Mean-field VI has a rich history in the realm of statistical inference; see Section 2.3 in Blei et al. (2017) for a brief historical introduction. Despite being widely used, computational and statistical guarantees have only recently emerged. A standard algorithm to solve (7.12) is Coordinate Ascent VI (CAVI) (see Blei et al., 2017, Section 2.4), the updates for which can be implemented for certain conjugate models. Guarantees for CAVI were provided recently in Bhattacharya et al. (2023) under a generalized correlation condition for π; see also Arnese and Lacker (2024) and Lavenant and Zanella (2024).

More closely related to our work is the use of Wasserstein gradient flows. The work of Lacker (2023) connects mean-field VI to constrained Wasserstein gradient flows, providing continuous-time guarantees via projected log-Sobolev inequalities but without a concrete algorithmic implementation; see also Lacker et al. (2024). Also, in the context of a Bayesian latent variable models, convergence guarantees for a Wasserstein gradient flow under a well-conditioned assumption at the population level was established by Yao and Yang (2022). Toward the issue of implementation, they suggested two strategies based on particle approximation combined with either Langevin sampling or optimization over transport maps respectively, but they did not analyze the error arising from the particle approximation. Zhang and Zhou (2020) study the theoretical and computational

properties of mean-field variational inference in the context of community detection. Despite the promising nature of these works, implementation remains a challenge in complete generality.

MEAN-FIELD EQUATIONS. Via calculus of variations, one can readily derive the following system of *mean-field equations* from (7.12): for $i \in [d]$,

$$\pi_i^\star(x_i) \propto \exp\Big(-\int_{\mathbb{R}^{d-1}} V(x_1, \dots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\Big)\,. \tag{7.14}$$

These are also sometimes called *self-consistency equations*; we give a derivation in Appendix F.3.1. From the structure of $\pi^\star$, we can prove the following result.

Proposition 7.19. *Suppose that π is well-conditioned (WC). Then, (7.12) admits a* unique *minimizer of the form (7.13), where each $\pi_i^\star$ is well-conditioned (WC) with the same parameters ℓ_V, L_V as π.*

Uniqueness of the minimizer follows as a corollary of Lemma 7.18 and Lacker (2023, Proposition 3.2), which shows that $\mathcal{P}(\mathbb{R})^{\otimes d}$ is a geodesically convex subset of the Wasserstein space, and the individual $\pi_i^\star$ measures being well-conditioned is immediate from (7.14).

OUR APPROACH. We approach solving mean-field VI by optimizing over a suitably rich family of compatible maps. To this end, we want to relate (7.12) to

$$\pi_\diamond^\star := (T_\diamond^\star)_\sharp \rho \in \operatorname*{argmin}_{\mu \in \mathcal{P}_\diamond} \mathrm{KL}(\mu \| \pi)\,, \tag{7.15}$$

where $\mathcal{P}_\diamond$ is a polyhedral subset of the Wasserstein space (Definition 7.2). Recall that polyhedral subsets of the Wasserstein space are geodesically convex (see Theorem 7.1). Combined with Lemma 7.18, the following corollary is immediate.

Corollary 7.20. *Suppose that $\mathcal{P}_\diamond$ is a polyhedral subset of the Wasserstein space, and that π is well-conditioned (WC). Then the minimizer to (7.15) is* unique*, denoted by $\pi_\diamond^\star$.*

Borrowing inspiration from the existing literature combining normalizing flows and optimal transport (Chen et al., 2018; Finlay et al., 2020a;b; Huang et al., 2021a), our goal is to transfer the difficulty of estimating the measure $\pi^\star$ to estimating an appropriate optimal transport map. Indeed, $\mathcal{P}_\diamond$ is a collection of pushforwards of a base measure ρ via optimal transport maps. Going forward, we will provide a systematic way of choosing both ρ, the base measure, and $\mathcal{M}$, the set of optimal transport maps which generates $\mathcal{P}_\diamond$.

A natural candidate for the base density ρ is the standard Gaussian distribution in $\mathbb{R}^d$. Beyond its naturality, this choice is justified by powerful regularity results, described in the next section, for the optimal transport map $T^\star$ from ρ to the mean-field solution $\pi^\star$. This regularity result, in turn, will feed into the approximation theory of Section 7.5.3.

7.5.2 Regularity of optimal transport maps between well-conditioned product measures

In this section, we study the regularity of the optimal transport map from the standard Gaussian to the mean-field solution $\pi^\star$. More generally, our regularity bounds hold for the optimal transport map from the Gaussian to any well-conditioned product measure, or between any two well-conditioned product measures μ and ν (either by writing $T^{\mu\to\nu}$ as $T^{\rho\to\nu} \circ (T^{\rho\to\mu})^{-1}$ and directly applying the results of this section, or by repeating the arguments thereof).

Theorem 7.21. *Let $\rho = \mathcal{N}(0, I)$ and suppose that π is well-conditioned (WC). Then, there exists a unique, coordinate-wise separable optimal transport map from ρ to $\pi^\star$, the minimizer to (7.12), written $T^\star(x) = (T_1^\star(x_1), \dots, T_d^\star(x_d))$. Each map $T_i^\star$ satisfies*

$$\sqrt{1/L_V} \le (T_i^\star)' \le \sqrt{1/\ell_V}\,.$$

Moreover, we have the higher-order regularity bounds

$$|(T_i^\star)''(x)| \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,(1+|x|)\,, \qquad \textit{and} \qquad |(T_i^\star)'''(x)| \lesssim \frac{\kappa^2}{\sqrt{\ell_V}}\,(1+|x|^2)\,. \tag{7.16}$$

The bounds on $(T_i^\star)'$ in Theorem 7.21 in fact follow immediately from two landmark results in optimal transport, and Proposition 7.19. First, since ρ admits a density, then Brenier's theorem (Brenier, 1991) states that there always exists a unique optimal transport map from ρ to any target measure, in this case, $\pi^\star$. Obviously, since both ρ and $\pi^\star$ are product measures, the corresponding optimal transport map is coordinate-wise separable. Then, Caffarelli's contraction theorem (Caffarelli, 2000) yields tight lower and upper bounds on the derivatives of each component of $T^\star$ as a function of the strong log-concavity and log-smoothness parameters of ρ and $\pi^\star$. See, e.g., Chewi and Pooladian (2023, Theorem 4) for a precise statement of the contraction theorem, and a short proof based on entropic optimal transport.

On the other hand, we have not seen the bounds (7.16) in the literature. In general, regularity theory for optimal transport is notoriously challenging due to the fully non-linear nature of the associated Monge–Ampère PDE; see Villani (2021, Section 4.2.2) for an exposition to Caffarelli's celebrated regularity theory. Here, we can avoid difficult arguments by exploiting the coordinate-wise separability of the transport map and straightforward computations with the Monge–Ampère equation. See Appendix F.3.2 for the proof.

The regularity we obtain is essentially optimal, since we started with information on the derivatives of $\pi^\star$ up to order two, and we obtain regularity bounds for the Kantorovich potential (of which $T^\star$ is the gradient) up to order *four*. Such higher-order regularity bounds are not only useful for obtaining sharper approximation results, but are in fact essential for establishing the key result Theorem 7.26 in Section 7.5.3.

Remark 7.22. In prior works that statistically estimate optimal transport maps on the basis of samples (such as Deb et al. (2021); Divol et al. (2022); Hütter and Rigollet (2021); Manole et al.

(2024a); Pooladian and Niles-Weed (2021)), bounds on the Jacobian of the optimal transport map of interest are necessary and standard. In contrast, here these bounds hold as a consequence of our problem setting (in particular, from (WC)).

7.5.3 Approximating the mean-field solution with compatible maps

So, Theorem 7.21 tells us that we can view $\pi^\star = (T^\star)_\sharp \rho$, where $T^\star$ obeys desirable regularity properties. The goal of this section is to demonstrate that we can prescribe a class of maps $\mathcal{M}$ such that the minimizer of the KL divergence over $\mathcal{P}_\diamond := \underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp \rho$,

$$\pi_\diamond^\star \in \operatorname*{argmin}_{\mu \in \mathcal{P}_\diamond} \mathrm{KL}(\mu \| \pi)\,,$$

is close to $\pi^\star$ in the Wasserstein distance. Then, Section 7.5.4 will provide guarantees for computing $\pi_\diamond^\star$ via KL minimization over this set.

The first step is to prove an approximation theorem: there *exists* an element $\hat{\pi}_\diamond \in \mathcal{P}_\diamond$ such that $\pi^\star$ is close to $\hat{\pi}_\diamond$. We state this as the following general result. Here, we write $\|D(\bar{T} - \hat{T})\|^2_{L^2(\rho)}$ for the quantity $\int \|D(\bar{T} - \hat{T})\|_{\mathrm{F}}^2 \,\mathrm{d}\rho$.

Theorem 7.23. *Let $\rho = \mathcal{N}(0, I)$. For any $\varepsilon > 0$, there exists a compatible family $\mathcal{M}$ of optimal transport maps of size $\widetilde{O}(\kappa^{1/2} d^{5/4} / \varepsilon^{1/2})$, with the following property. For any coordinate-wise separable map $\bar{T} : \mathbb{R}^d \to \mathbb{R}^d$ with Jacobian satisfying the first and second derivative bounds of Theorem 7.21, there exists $\hat{T} \in \underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})$, with $\alpha = 1/\sqrt{L_V}$, such that $W_2(\bar{T}_\sharp \rho, \hat{T}_\sharp \rho) = \|\bar{T} - \hat{T}\|_{L^2(\rho)} \leqslant \varepsilon / \ell_V^{1/2}$ and $\|D(\bar{T} - \hat{T})\|_{L^2(\rho)} \lesssim \kappa^{1/2} d^{1/4} \varepsilon^{1/2} / \ell_V^{1/2}$.*

Approximation theory has a large literature which aims at proving uniform rates of approximation over various function classes by linear combinations of well-chosen basis elements. Typical choices of basis functions include polynomials, splines, wavelets, etc., with more recent literature investigating approximations via neural networks.

While resting on standard techniques, the most important departure of our result from the literature is the coordinate-wise structure of $\bar{T}$, which allows for approximation rates that do not incur the curse of the dimensionality, in the sense that the cardinality of $|\mathcal{M}|$ does not depend exponentially on the dimension d. Observe the presence of a structural constraint: in one dimension, the problem essentially boils down to approximating a monotonically increasing function via conic combinations of the generating set $\mathcal{M}$.

Our construction is described as follows. Let $R > 0$ denote a truncation parameter, and let $\delta > 0$ denote a mesh size. We partition the interval $[-R, +R]$ into sub-intervals of size δ. Then, $\mathcal{M}$ consists of all functions of the form $x \mapsto (0, \dots, 0, \psi(\delta^{-1}(x_i - a)), 0, \dots, 0)$, where only the i-th coordinate of the output is non-zero, $I = [a, a + \delta]$ is a sub-interval of size δ, and $\psi : \mathbb{R} \to \mathbb{R}$ is piecewise linear, defined via $\psi(x) \coloneqq 1 \wedge x_+$. Proofs are given in Appendix F.3.3.

This piecewise linear construction exploits the smoothness of $\bar{T}$ up to order two, but no further. On the other hand, from Theorem 7.21 we see that $T^\star$ also obeys a bound on its *third* derivative, so we can expect to obtain better approximation rates through a smoother dictionary. This is indeed the case, but the approximating set becomes more complicated (in particular, it is no longer the pushforward of a pointed cone, but a general polyhedral set), so we defer the details to Appendix F.3.3.

Theorem 7.24 (Higher-order smoothness)**.** *There exists a polyhedral set $\mathcal{P}_\diamond$ with an explicit generating family of size $\widetilde{O}(\kappa^{2/3} d^{7/6} / \varepsilon^{1/3})$ with the following property. In the setting of Theorem 7.23, assume also that each component $\bar{T}_i$ of $\bar{T}$ obeys the third derivative bound in Theorem 7.21. Then, there exists $\hat{T} \in \mathcal{P}_\diamond$ such that $W_2(\bar{T}_\sharp \rho, \hat{T}_\sharp \rho) = \|\bar{T} - \hat{T}\|_{L^2(\rho)} \leqslant \varepsilon / \ell_V^{1/2}$ and $\|D(\bar{T} - \hat{T})\|_{L^2(\rho)} \lesssim \kappa^{2/3} d^{1/6} \varepsilon^{2/3} / \ell_V^{1/2}$.*

Remark 7.25. We note that the worse dependence on κ for the smoother dictionary is due to our derivative bounds for the optimal transport map; we have no reason to believe it is fundamental.

The two preceding results show that we can, with prior knowledge of $\pi^\star$, *construct* some $\hat{\pi}_\diamond \in \mathcal{P}_\diamond$ which is close to $\pi^\star$, but it does not guarantee that we can *find* $\hat{\pi}_\diamond$ easily. The next

result addresses this issue by showing that $\pi^\star$ is close to the *minimizer* $\pi^\star_\diamond$ of the KL divergence over $\mathcal{P}_\diamond$, and hence can be computed using the algorithms in Section 7.5.4. As the proof reveals, establishing this statement is related to a geodesic *smoothness* property for the KL divergence, which is quite non-trivial since the entropy is non-smooth over the full Wasserstein space (see the further discussion in the next section). We are able to verify this smoothness property on the geodesic connecting $\hat{\pi}_\diamond$ to $\pi^\star$ using the bounds on $\|D(\hat{T}_\diamond - T^\star)\|_{L^2(\rho)}$ in our approximation results (Theorem 7.23 and Theorem 7.24). The proof of Theorem 7.26 is also found in Appendix F.3.3.

Theorem 7.26. *The mean-field solution $\pi^\star$ is close to the* minimizer $\pi^\star_\diamond$ *of the KL divergence over $\mathcal{P}_\diamond$ with corresponding generating family $\mathcal{M}$, in the sense that $\sqrt{\ell_V}\, W_2(\pi^\star_\diamond, \pi^\star) \leqslant \varepsilon$, in the following two cases.*

1. *For the piecewise linear construction of Theorem 7.23, the size of the family is bounded by $|\mathcal{M}| \leqslant \widetilde{O}(\kappa^2 d^{3/2}/\varepsilon)$.*

2. *For the higher-order approximation scheme of Theorem 7.24, the size of the family is bounded by $|\mathcal{M}| \leqslant \widetilde{O}(\kappa^{3/2} d^{5/4}/\varepsilon^{1/2})$.*

7.5.4 Computational guarantees for mean-field VI

Having identified polyhedral subsets $\mathcal{P}_\diamond$ of the Wasserstein space over which the KL minimizer $\pi^\star_\diamond$ is close to the desired mean-field VI solution $\pi^\star$, we are now in a position to apply our theory of polyhedral optimization and thereby obtain novel computational guarantees for mean-field VI. Recall that $\pi \propto \exp(-V)$, and

$$\mathrm{KL}(\mu\|\pi) = \mathcal{V}(\mu) + \mathcal{H}(\mu) := \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z\,, \tag{7.17}$$

where $Z > 0$ is the normalizing constant of π.

For concreteness, we focus our discussion on the setting in which $\mathcal{P}_\diamond = \underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})_\sharp \rho$, such

as in Theorem 7.23, although the discussion below can be adapted to more general polyhedral sets. As in Section 7.4.2, we apply Euclidean optimization algorithms over the parameterization of $\underline{\text{cone}}(\mathcal{M}; \alpha\,\mathrm{id})$; see Section 7.5.5.1 for a discussion of implementation.

In order to apply the algorithmic guarantees from Section 7.4.2, we must verify the strong geodesic convexity and geodesic smoothness of the KL divergence over the set $\mathcal{P}_\diamond$. Strong convexity follows from the celebrated fact that the KL divergence with respect to an ℓ_V-strongly log-concave measure π is ℓ_V-strongly geodesically convex (see Villani, 2009, Particular Case 23.15), together with the geodesic convexity of $\mathcal{P}_\diamond$ (Theorem 7.1 and Section 7.4.3).

Proposition 7.27 (Strong convexity of the KL divergence over geodesically convex sets)**.** *Assume that π is well-conditioned (WC). Then, the KL divergence $\mathrm{KL}(\cdot\|\pi)$ is ℓ_V-strongly geodesically convex over any geodesically convex subset of the Wasserstein space.*

Smoothness of the KL divergence, however, is more subtle, owing to the non-smoothness of the entropy $\mathcal{H}$ over the full Wasserstein space; see Diao et al. (2023) for further discussion of this point. Prior works therefore established smoothness over restricted subsets of the Wasserstein space (e.g., Lambert et al., 2022), or utilized proximal methods which succeed in the absence of smoothness (e.g., Diao et al., 2023). We adopt the former approach, and for this we require a further property of the family $\mathcal{M}$ of generating maps.

First, without loss of generality, we may assume that each $T \in \mathcal{M}$ has mean zero under ρ: $\int T \,\mathrm{d}\rho = 0$. Indeed, subtracting the means from the maps in the generating set does not affect $\underline{\text{cone}}(\mathcal{M}; \alpha\,\mathrm{id})$, since $\underline{\text{cone}}(\mathcal{M}; \alpha\,\mathrm{id})$ is augmented by translations. Assuming now that $\mathcal{M}$ is centered, we recall the Gram matrix Q with entries $Q_{T,\tilde{T}} := \langle T, \tilde{T}\rangle_{L^2(\rho)}$. We also form the Gram matrix of the Jacobians, $Q^{(1)}$, with entries $Q^{(1)}_{T,\tilde{T}} := \langle DT, D\tilde{T}\rangle_{L^2(\rho)} := \int \langle DT, D\tilde{T}\rangle \,\mathrm{d}\rho$. Our main assumption on $\mathcal{M}$ is an upper bound on $Q^{(1)}$ in terms of Q. We refer to families $\mathcal{M}$ satisfying this condition as *regular*.

(Υ) There exists $\Upsilon > 0$ such that for the Gram matrices associated with a centered family $\mathcal{M}$, it

holds that $Q^{(1)} \preceq \Upsilon Q$.

We remark that when $\mathcal{M}$ is constructed as a direct sum of univariate families via Lemma 7.6 (c.f. Remark 7.10), as in our approximation results (Section 7.5.3), the matrices Q and $\tilde{Q}$ have a $d \times d$ block diagonal structure; see Section 7.5.5.1 for details. Consequently, the regularity Υ of the family $\mathcal{M}$ is the same as the regularity parameter for the *univariate* family used to construct $\mathcal{M}$, and is therefore nominally "dimension-free".[5]

We can now establish our geodesic smoothness result for the KL divergence over the augmented and pointed cone $\underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp\rho$, where $\mathcal{M}$ is a regular generating family.

Proposition 7.28 (Smoothness of the KL divergence over $\underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp\rho$)**.** *Assume that π is well-conditioned (WC) and that $\mathcal{M}$ is regular (Υ). Then, $\mathrm{KL}(\cdot\|\pi)$ is M-geodesically smooth over $\underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp\rho$, with smoothness constant bounded by*

$$M \leqslant L_V + \Upsilon/\alpha^2\,.$$

From Theorem 7.21, we know that the optimal transport map $T^\star$ from ρ to the mean-field solution $\pi^\star$ is the gradient of a $1/\sqrt{L_V}$-strongly convex potential, so we take $\alpha = 1/\sqrt{L_V}$. The smoothness constant for the KL divergence then becomes $(1 + \Upsilon)\, L_V$. With these results in hand, we can state our accelerated convergence guarantees for mean-field VI, which follow directly from the previous propositions, and Theorem 7.13.

Theorem 7.29 (Accelerated mean-field VI)**.** *Assume that π is well-conditioned (WC) and that $\mathcal{M}$ is regular (Υ). Let $\pi^\star_\diamond$ denote the unique minimizer of $\mathrm{KL}(\cdot\|\pi)$ over the polyhedral set $\mathcal{P}_\diamond = \underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp\rho$ with $\alpha = 1/\sqrt{L_V}$. Then, the iterates of accelerated projected gradient descent yield*

[5] However, if one wishes to maintain the same quality of approximation in high dimension, our approximation results in Section 7.5.3 require taking the size of the univariate family to scale mildly with the dimension, and in this case the parameter Υ may indeed scale with the dimension.

a measure $\mu_{(t)}$ with the guarantee $W_2(\mu_{(t)}, \pi^\star_\diamond) \leq \varepsilon$, with a number of iterations bounded by

$$t = O\Big(\sqrt{\kappa\,(1+\Upsilon)}\,\log\big(\sqrt{\kappa}\,W_2(\mu_{(0)}, \pi^\star_\diamond)/\varepsilon\big)\Big)\,,$$

where $\kappa \coloneqq L_V/\ell_V$ is the condition number of π.

By combining Theorem 7.29 with our approximation result in Theorem 7.26, which provides a bound on $W_2(\pi^\star_\diamond, \pi^\star)$ for explicit choices of $\mathcal{P}_\diamond$ with corresponding bounds on the size of $|\mathcal{M}|$, we can then ensure that the iterate $\mu_{(t)}$ is close to $\pi^\star$ in the Wasserstein distance. Namely, we ensure that $W_2(\mu_{(t)}, \pi^\star) \leqslant \varepsilon$ provided that we use either of the dictionaries in Theorem 7.26 and we take the number of iterations t as in Theorem 7.29; here, t is the iteration complexity, whereas the bounds on the size of the dictionary in Theorem 7.26 govern the per-iteration cost.

As previously mentioned, our analysis is made possible by bypassing the non-differentiability of entropy over $\mathcal{P}(\mathbb{R})$ and instead optimizing over pointed cones characterized by a univariate compatible family $\mathcal{M}$. Thus, the constant $\Upsilon > 0$ should blow up as the polyhedral set approaches $\mathcal{P}(\mathbb{R})$. This fact is summarized in the following lemma for the piecewise-linear family.

Lemma 7.30. *Let $\mathcal{M}$ be the piecewise linear construction of Theorem 7.23. Then, $\Upsilon \lesssim |\mathcal{M}_1|^2$, where $\mathcal{M}_1$ is the generating family in a single dimension. Since $|\mathcal{M}_1| = J$, the bound is equivalently $\Upsilon \lesssim J^2$.*

As a corollary, we can fully characterize the runtime of solving the MFVI problem.

Corollary 7.31 (End-to-end guarantees for MFVI)**.** *Consider the setting of Theorem 7.29. Then the required runtime to compute $\pi^\star_\diamond$ becomes*

$$t = O(J\kappa^{1/2}\log(\sqrt{\kappa}d/\varepsilon))\,.$$

If $\pi^\star_\diamond$ is meant to approximation $\pi^\star$, then we can use the approximation guarantees from Theorem 7.26

to obtain the complete convergence guarantee of

$$t = O(\kappa^{5/2} d^{1/2} \varepsilon^{-1} \log(\sqrt{\kappa} d/\varepsilon)) \,.$$

As we demonstrate below, the regime of $J = O(1)$ appears to suffice numerically. To the best of our knowledge, this constitutes the first *accelerated* and *end-to-end* convergence result for mean-field VI. See Section 7.5.1 for comparisons with the literature.

7.5.5 Algorithms for mean-field VI

In this section, we discuss implementation details for our proposed mean-field VI algorithm, which includes an analysis of stochastic gradient descent over our polyhedral sets.

7.5.5.1 Implementation details

Recall that the goal is to compute a product measure approximation to π which has density proportional to $\exp(-V)$ on $\mathbb{R}^d$.

Building the family of maps. The first step is to build a family $\mathcal{M}_1$ of increasing maps $\mathbb{R} \to \mathbb{R}$. The specification of these maps is left to the user; in Section 7.5.3, we have provided an example of a family of maps with favorable approximation properties. For later purposes, it is also important to center the maps to ensure that they have mean zero under ρ; this is done by computing the expectations of the maps via one-dimensional Gaussian quadrature and subtracting the means.

Let J denote the size of $|\mathcal{M}_1|$ and write $\mathcal{M}_1 = \{T_1, \ldots, T_J\}$.

Parameterization of the cone. As discussed in Section 7.4.3, it is useful to augment the cone with translations. Once the one-dimensional family $\mathcal{M}_1$ has been specified, it generates the d-dimensional augmented cone of maps parameterized by $(\lambda, v) \in \mathbb{R}_+^{Jd} \times \mathbb{R}^d$: the corresponding map $T^{\lambda,v}$ is given by $T^{\lambda,v}(x) = \alpha x + \sum_{i=1}^d \sum_{j=1}^J \lambda_{i,j} T_j(x_i)\, e_i + v$.

Construction of the Q matrix. For concreteness, let us fix the reference measure ρ to be the standard Gaussian $\mathcal{N}(0, I_d)$. We must compute the $Jd \times Jd$ matrix Q, with entries $Q_{(i,j);(i',j')} := \int \langle T_j(x_i)\, e_i, T_{j'}(x_{i'})\, e_{i'} \rangle\, \rho(\mathrm{d}x)$. From this expression, it is clear that Q is block diagonal; in fact, if we let $Q^{\mathcal{M}_1}$ denote the matrix corresponding to the one-dimensional family, with entries $Q^{\mathcal{M}_1}_{j,j'} := \int T_j T_{j'}\, \mathrm{d}\rho_1$ (here ρ_1 is the one-dimensional standard Gaussian), then $Q = I_d \otimes Q^{\mathcal{M}_1}$, and hence the full matrix Q never has to be stored in memory.

The entries of the $J \times J$ matrix $Q^{\mathcal{M}_1}$ can be precomputed, either via Monte Carlo sampling from ρ_1, or via one-dimensional Gaussian quadrature.

Computation of the gradient and projection. In order to apply the algorithms in Section 7.4, we must specify the gradient of $\mathrm{KL}((T^{\lambda,v})_\sharp \rho \| \pi)$ w.r.t. (λ, v) and the projection operator w.r.t. the Q-norm, $\|\cdot\|_Q$. Recall that we compute the gradients and projections for the λ variable w.r.t. $\|\cdot\|_Q$, and for the v variable in the standard Euclidean norm.

Using the change of variables formula,

$$\mathrm{KL}((T^{\lambda,v})_\sharp \rho \| \pi) = \int \left[V(T^{\lambda,v}(x)) - \log \det DT^\lambda(x) \right] \rho(\mathrm{d}x) + \int \log \rho \, \mathrm{d}\rho + \log Z \,.$$

The partial derivatives are therefore computed to be

$$\begin{aligned} \partial_{\lambda_{i,j}} \mathrm{KL}((T^{\lambda,v})_\sharp \rho \| \pi) &= \int \left[\partial_i V(T^{\lambda,v}(x))\, T_j(x_i) - \langle e_i, (DT^\lambda)^{-1}(x)\, e_i \rangle\, T_j'(x_i) \right] \rho(\mathrm{d}x)\,, \\ \nabla_v \mathrm{KL}((T^{\lambda,v})_\sharp \rho \| \pi) &= \int \nabla V(T^{\lambda,v}(x))\, \rho(\mathrm{d}x)\,. \end{aligned} \tag{7.18}$$

For the terms explicitly involving V, one can draw Monte Carlo samples from the Gaussian ρ and approximate them via empirical averages (assuming access to evaluations of the partial derivatives of V).

To compute the second term, note that DT^λ is diagonal:

$$DT^\lambda(x) = \alpha I_d + \operatorname{diag}\left(\sum_{j=1}^{J} \lambda_{i,j} T_j'(x_i)\right)_{i=1}^{d}.$$

Hence, inversion of $DT^\lambda(x)$ is very fast, requiring only $O(Jd)$ time to compute $DT^\lambda(x)$ and then $O(d)$ time to invert it. Moreover, the (i,i)-entry of $(DT^\lambda)^{-1}(x)$ only depends on x_i, so the second term reduces to a *one-dimensional integral*:

$$\int \langle e_i, (DT^\lambda)^{-1}(x)\, e_i \rangle\, T_j'(x_i)\, \rho(\mathrm{d}x) = \int \frac{T_j'(x_i)}{\alpha + \sum_{j'=1}^{J} \lambda_{i,j'} T_{j'}'(x_i)}\, \rho_1(\mathrm{d}x_i).$$

In turn, this one-dimensional integral can be computed rapidly via Gaussian quadrature.

To summarize: the gradient of the potential energy term (the term involving V) can be approximated via Monte Carlo sampling, and the gradient of the entropy term decomposes along the coordinates and can therefore be dealt with via standard quadrature rules. Note that many of these steps can be parallelized. In Section 7.5.5.2, we control the variance of the stochastic gradient, thereby obtaining guarantees for SPGD.

To compute the projection of a point $\eta \in \mathbb{R}^{Jd}$ onto the non-negative orthant $\mathbb{R}_+^{Jd}$ w.r.t. $\|\cdot\|_Q$, one must solve the following optimization problem:

$$\min_{\lambda \in \mathbb{R}_+^{Jd}} \langle \lambda - \eta, Q(\lambda - \eta) \rangle.$$

Again, due to the block diagonal structure of Q, this is equivalent to solving d independent projection problems: in each one, we must project a point in $\mathbb{R}^J$ onto $\mathbb{R}_+^J$ in the $Q^{\mathcal{M}_1}$-norm. This is a smooth, convex problem that can itself be solved via, e.g., projected gradient descent, or L-BFGS-B (Zhu et al., 1997), or any standard quadratic program solver.

7.5.5.2 Convergence for stochastic mean-field VI

In Section 7.5.5.1, we noted that in general, the gradient of the KL divergence involves an integral over ρ, which can be approximated via Monte Carlo sampling. This leads to a *stochastic* projected gradient algorithm for mean-field VI, and this section is devoted to obtaining convergence guarantees for SPGD.

Our goal here is not to conduct a comprehensive study, but rather to show how such guarantees can be obtained, and hence we impose a number of simplifying assumptions. We do not work with the cone augmented by translations, so that the maps are parameterized solely by $\lambda \in \mathbb{R}_+^{|\mathcal{M}|}$ (the v-component is easier to handle and only introduces extra notational burden into the proofs). Also, we consider a stochastic approximation of the gradient of the potential term via a single sample drawn from ρ at each iteration, and we assume that the gradient of the entropy is computed exactly. As discussed in Section 7.5.5.1, the gradient of the entropy can be handled via one-dimensional quadrature.

Even with these simplifications, the variance bound is somewhat involved. Motivated by the piecewise linear construction of Theorem 7.23, in which all maps $T \in \mathcal{M}$ can be taken to be *bounded*, we impose the following assumption.

(Ξ) There exists $\Xi > 0$ such that for the Gram matrix $Q^{\mathcal{M}_1}$ associated with the centered *univariate* family $\mathcal{M}_1$, we have the pointwise bound $\langle Q^{-1}, \bar{Q}(x)\rangle \le \Xi J$ for all $x \in \mathbb{R}$, where $\bar{Q}_{T,\tilde{T}}(x) = T(x)\tilde{T}(x)$ for $T, \tilde{T} \in \mathcal{M}_1$. Here, $J := |\mathcal{M}_1|$.

Similarly to Lemma 7.30, we can also quantify Ξ for the piecewise linear dictionary.

Lemma 7.32. *Let $\mathcal{M}$ be the piecewise linear construction of Theorem 7.23. Then, $\Xi \lesssim |\mathcal{M}_1|^2$, where $\mathcal{M}_1$ is the generating family in a single dimension. Since $|\mathcal{M}_1| = J$, the bound is equivalently $\Xi \lesssim J^2$.*

The following lemma established a variance bound of the type (VB) which, when combined with Theorem 7.14, proves Theorem 7.34.

Lemma 7.33 (Variance bound for stochastic mean-field VI)**.** *Assume π is well-conditioned (WC) and that $\mathcal{M}$ is generated from a univariate family $\mathcal{M}_1$ satisfying (Ξ). Let $Q^{-1}\,\hat{\nabla}_\lambda \mathrm{KL}(\cdot\|\pi)$ denote the stochastic gradient (see Appendix F.3.5). Let $\pi^\star_\diamond$ denote the unique minimizer of $\mathrm{KL}(\cdot\|\pi)$ over $\mathrm{cone}(\mathcal{M};\, \alpha\,\mathrm{id})_\sharp\rho$ with $\alpha = 1/\sqrt{L_V}$. Then, the following second moment bound holds:*

$$\mathbb{E}[\operatorname{tr}\operatorname{Cov}(Q^{-1/2}\,\hat{\nabla}_\lambda \mathrm{KL}(\mu_\lambda\|\pi))] \le 2L_V^2 \Xi J\, W_2^2(\mu_\lambda, \pi^\star_\diamond) + 4L_V \Xi J\,(L_V\, W_2^2(\pi^\star_\diamond, \pi^\star) + \kappa d)\,.$$

Let us assume that the κd term is larger than $L_V W_2^2(\pi^\star_\diamond, \pi^\star)$; this can be guaranteed via the approximation result in Section 7.5.3. The next theorem follows immediately from Theorem 7.14 and the previous lemma.

Theorem 7.34 (Convergence of stochastic mean-field VI)**.** *Assume that π is well-conditioned (WC) and that $\mathcal{M}$ is regular (ϒ) and generated by a univariate family satisfying (Ξ). Let $\pi^\star_\diamond$ denote the unique minimizer of $\mathrm{KL}(\cdot\|\pi)$ over $\mathrm{cone}(\mathcal{M};\, \alpha\,\mathrm{id})_\sharp\rho$ with $\alpha = 1/\sqrt{L_V}$. Then, for all sufficiently small ε, the iterates of stochastic projected gradient descent yield a measure $\mu_{(t)}$ with the guarantee $\sqrt{\ell_V}\,\mathbb{E}[W_2(\mu_{(t)}, \pi^\star_\diamond)] \le \varepsilon$, with a number of iterations bounded by*

$$t \gtrsim \frac{\Xi \kappa^2 J d}{\varepsilon^2} \log(\sqrt{\ell_V}\, W_2(\mu_{(0)}, \pi^\star_\diamond)/\varepsilon)\,,$$

and step size $h \asymp \varepsilon^2/(L_V \Xi \kappa J d)$.

As with Theorem 7.29, we can state the following corollary given that Ξ is uniformly bounded via Lemma 7.32.

Corollary 7.35 (End-to-end guarantees with SPGD)**.** *Consider the setting of Theorem 7.34. Then the required runtime to estimate $\pi^\star_\diamond$ becomes*

$$t = O(d J^3 \kappa^2 \varepsilon^{-2} \log(\sqrt{\ell_V} d/\varepsilon))\,.$$

If $\pi_\diamond^\star$ is meant to approximation $\pi^\star$, then we can use the approximation guarantees from Theorem 7.26 to obtain the complete convergence guarantee of

$$t = O(\kappa^8 d^{5/2} \varepsilon^{-5} \log(\sqrt{\ell_V} d/\varepsilon)) \,.$$

7.6 Numerical experiments

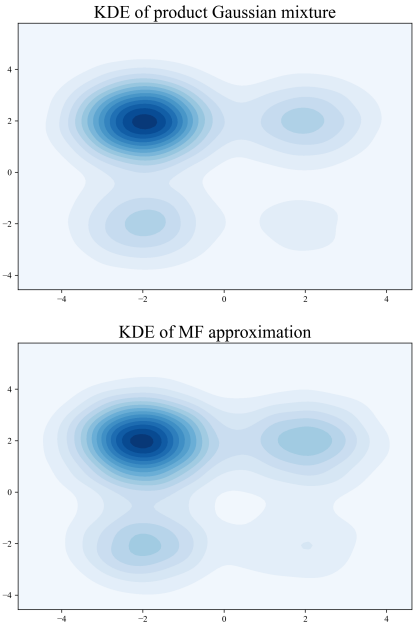


Figure 7.1: KDEs for the optimal product Gaussian mixture and our algorithm.

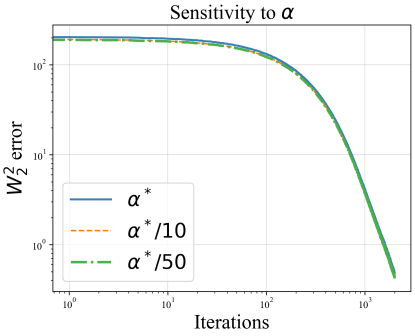


Figure 7.2: Our algorithm is robust to the choice of α.

We showcase our proposed MFVI algorithm on numerical experiments. Experimental details are deferred to Section F.4, and the code to reproduce the experiments is available here. Across all experiments, which include low- and high-dimensional settings, we took the piecewise linear dictionary (Theorem 7.23) with the same value for the size $J = |\mathcal{M}_1| = 28$ of the univariate family (hence $|\mathcal{M}| = Jd$), and we ran stochastic gradient descent (without acceleration) with a batch size of 2000 samples per iteration.

7.6.1 Product Gaussian mixture

In our first experiment, the target is a mixture of four Gaussians in $\mathbb{R}^2$ which is itself a product measure. Despite the non-log-concavity, our algorithm correctly recovers the correct target. Though, we note that this approach is sensitive to the initialization, but this is expected as the landscape is non-convex.

7.6.2 Non-isotropic Gaussian

Next, we computed the mean-field approximation of a randomly generated centered and non-isotropic Gaussian in dimension $d = 5$. Letting Σ denote the covariance matrix, the mean-field approximation is also a Gaussian with diagonal covariance and entries $(\Sigma_{\mathrm{MF}})_{i,i} = 1/(\Sigma^{-1})_{i,i}$ (see Section F.4.2 for a calculation of this fact).

In Figure 7.2, we plot the W_2^2 error between the covariance matrix of our algorithm iterate (computed from samples) and Σ_{MF}, which is a lower bound on the true W_2^2 distance (cf. Cuesta-Albertos et al., 1996).

In this case, the optimal parameter choice α^* is known, though this is rarely the case in practice. We ran our algorithm for various choices of α, fixing all other parameters to be the same. We see that our algorithm does not depend heavily on the choice of hyperparameter α, and the practitioner can safely choose a small value of α without sacrificing performance.

7.6.3 Synthetic Bayesian logistic regression

As a final example, we turn to Bayesian logistic regression on synthetic data; precise details are deferred to Section F.4.3. In summary, we are given i.i.d. data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$, (where $d = 20$ and $n = 100$) from which we want to recover a parameter θ. When assuming an

improper (Lebesgue) prior on θ, the posterior is given by

$$V(\theta) = \sum_{i=1}^{n} \left[\log(1 + \exp(\theta^\top X_i)) - Y_i \, \theta^\top X_i \right].$$

Note that V is not strongly convex as $V(\theta)$ behaves like a linear function as $\|\theta\| \to +\infty$. With V and ∇V in hand, our algorithm is fully implementable. As we considered an improper prior, a comparison to CAVI is not possible. Instead, we compared against standard Langevin Monte Carlo (LMC). The final histograms were generated using 2000 samples from both the mean-field VI algorithm and LMC. Figure 7.3 contains the 20 marginals for both our approach and LMC, which are closely aligned.

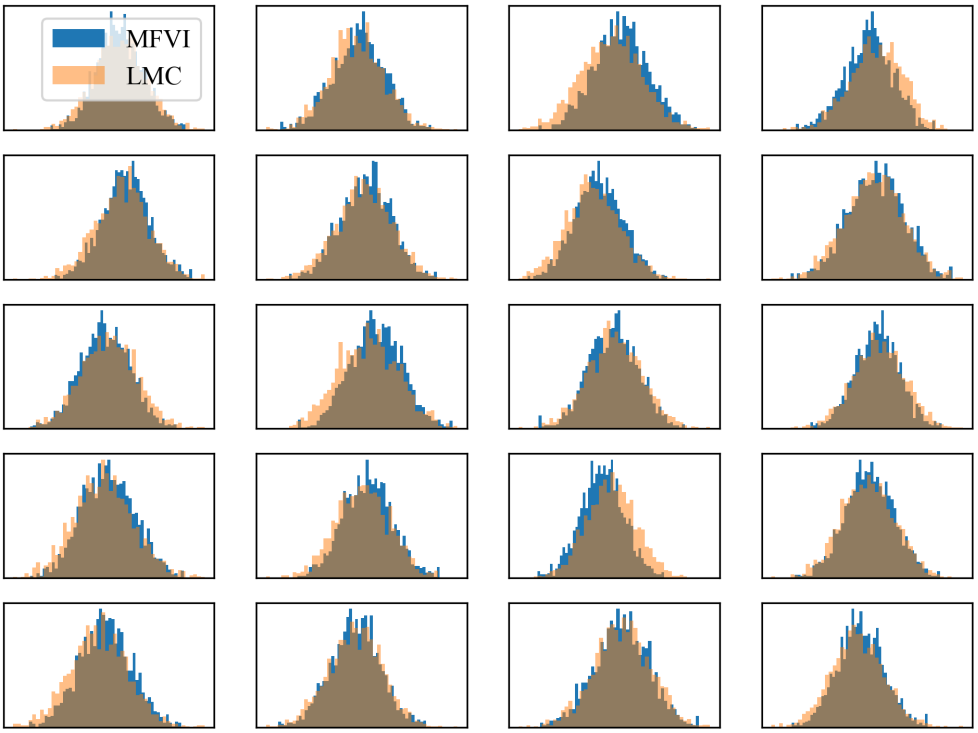


Figure 7.3: Histograms of the first ten marginals computed via our mean-field VI algorithm vs. Langevin Monte Carlo for a 20-dimensional Bayesian logistic regression example.

7.7 Extension to mixtures of product measures

In this section, we extend our methodology to approximations via mixtures of product measures. The motivation is simply that many more measures can be approximated via mixtures of (approximate) product measures, e.g., Gibbs distributions with small gradient complexity (Austin, 2019; Eldan, 2018; Eldan and Gross, 2018; Jain et al., 2019).

In Section 7.5.4, we minimized $\mathrm{KL}(\cdot\|\pi)$ over $\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})_\sharp\rho$, where $\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})$ is parameterized by the pair $(\lambda, v) \in \mathcal{M} := \mathbb{R}_+^{|\mathcal{M}|} \times \mathbb{R}^d$, equipped with the norm $\|\cdot\|_{Q\oplus I_d}$. In this section, following Lambert et al. (2022), a *mixture of product measures* is specified by a mixing measure $P \in \mathcal{P}(\mathcal{M})$ and corresponds to the measure $\mu_P := \int (T^{\lambda,v})_\sharp\rho\, P(\mathrm{d}\lambda, \mathrm{d}v)$. We can now equip the space $\mathcal{P}(\mathcal{M})$ with the Wasserstein geometry (with respect to $\|\cdot\|_{Q\oplus I_d}$), and we shall derive the Wasserstein gradient flow of the functional $P \mapsto \mathrm{KL}(\mu_P\|\pi)$.

This approach to mixture modelling is inspired by the distance on Gaussian mixtures proposed in Chen et al. (2019); Delon and Desolneux (2020); see Bing et al. (2023) for a statistical perspective.

In this section, we again use the abstract parameterization $T^{\lambda,v} = \alpha\,\mathrm{id} + \sum_{T\in\mathcal{M}} \lambda_T T + v$. Proofs are given in Appendix F.5.

Theorem 7.36. *The Wasserstein gradient flow of $P \mapsto \mathrm{KL}(\mu_P\|\pi)$ is the flow $(P^{(t)})_{t\geqslant 0}$ specified as follows. For each $t \geqslant 0$, $P^{(t)}$ is the law of $(\lambda^{(t)}, v^{(t)})$, where*

$$\dot\lambda_T^{(t)} = -\int \left\langle \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)},v^{(t)}}, T \right\rangle \mathrm{d}\rho\,, \qquad \text{for } T \in \mathcal{M}\,,$$
$$\dot v^{(t)} = -\int \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)},v^{(t)}}\, \mathrm{d}\rho\,.$$

In practice, we use a finite number K of mixture components, in which case

$$P = \frac{1}{K}\sum_{k=1}^K \delta_{(\lambda[k],v[k])}\,, \qquad \mu_P = \frac{1}{K}\sum_{k=1}^K (T^{\lambda[k],v[k]})_\sharp\rho\,. \tag{7.19}$$

The system of ODEs above then becomes an interacting particle system:

$$\begin{aligned}
\dot{\lambda}_T^{(t)}[k] &= -\int \left\langle \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}, T \right\rangle \mathrm{d}\rho\,, && \text{for } T \in \mathcal{M}\,,\\
\dot{v}^{(t)}[k] &= -\int \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho\,.
\end{aligned}$$

The particles interact through the common term $\log \mu_{P^{(t)}}$. More explicitly, by the change of variables formula,

$$\mu_P = \frac{1}{K} \sum_{k=1}^{K} \frac{\rho \circ (T^{\lambda[k], v[k]})^{-1}}{\det DT^{\lambda[k], v[k]} \circ (T^{\lambda[k], v[k]})^{-1}}\,.$$

Note that computing $\nabla \log \mu_P$ now requires taking the second derivative of the transport maps, which hinders implementation. In this case, a smooth family $\mathcal{M}$ is required.

The dynamics (7.19) maintains equal weights for each of the particles at each time. We can similarly derive the gradient flow with respect to the Wasserstein–Fisher–Rao (or Hellinger–Kantorovich) geometry, which captures unbalanced optimal transport (Chizat et al., 2018; Liero et al., 2016; 2018). The use of this geometry for sampling was pioneered in Lu et al. (2019).

Theorem 7.37. *The Wasserstein–Fisher–Rao gradient flow of $P \mapsto \mathrm{KL}(\mu_P \| \pi)$, initialized at $P^{(0)} = \sum_{k=1}^{K} w^{(0)}[k]\, \delta_{(\lambda^{(0)}[k], v^{(0)}[k])}$ with $\sum_{k=1}^{K} w^{(0)}[k] = 1$, can be described as follows. For each time $t \geqslant 0$, let $P^{(t)} = \sum_{k=1}^{K} w^{(t)}[k]\, \delta_{(\lambda^{(t)}[k], v^{(t)}[k])}$ and $r^{(t)}[k] \coloneqq \sqrt{w^{(t)}[k]}$. Then,*

$$\begin{aligned}
\dot{\lambda}_T^{(t)}[k] &= -\int \left\langle \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}, T \right\rangle \mathrm{d}\rho\,, && \textit{for } T \in \mathcal{M}\,,\\
\dot{v}^{(t)}[k] &= -\int \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho\,,\\
\dot{r}^{(t)}[k] &= -\left(\int \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho - \int \log \frac{\mu_{P^{(t)}}}{\pi}\, \mathrm{d}\mu_{P^{(t)}} \right) r^{(t)}[k]\,.
\end{aligned}$$

We leave it as an open question to obtain convergence rates for this flow.

A | Supplement to Chapter 2

A.1 Second-order error estimate

In this section, we outline a short proof of Theorem 2.2.

Theorem A.1. *Suppose P and Q have bounded densities with compact support. Then*

$$\mathrm{OT}_\varepsilon(P,Q) - \frac{1}{2}W_2^2(P,Q) + \varepsilon\log((2\pi\varepsilon)^{d/2}) \le -\frac{\varepsilon}{2}\left(\mathcal{H}(P)+\mathcal{H}(Q)\right) + \frac{\varepsilon^2}{8}I_0(P,Q), \tag{A.1}$$

where $I_0(P,Q)$ is the integrated Fisher information along the Wasserstein geodesic between P and Q.

The proof hinges on the *dynamic* formulations of $W_2^2(P,Q)$ and $\mathrm{OT}_\varepsilon(P,Q)$ (Benamou and Brenier, 2000; Chizat et al., 2020; Conforti and Tamanini, 2021). We begin with the former:

$$\frac{1}{2}W_2^2(P,Q) = \inf_{\rho,v}\int_0^1\int_{\mathbb{R}^d}\frac{1}{2}\|v(t,x)\|_2^2\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t, \tag{A.2}$$

subject to $\partial_t\rho + \nabla\cdot(\rho v) = 0$, called the *continuity equation*, with $\rho(0,\cdot) = p(\cdot)$ and $\rho(1,\cdot) = q(\cdot)$. We let (ρ_0, v_0) denote the joint minimizers to (A.2) satisfying these conditions.

Similarly, there exists a dynamic formulation for OT_ε (see Chizat et al., 2020; Conforti and Tamanini, 2021, for more information): for two measures with bounded densities and compact

support,

$$\mathrm{OT}_\varepsilon(P,Q) + \varepsilon \log(\Lambda_\varepsilon) = \inf_{\rho,v} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{1}{2}\|v(t,x)\|_2^2 + \frac{\varepsilon^2}{8}\|\nabla_x \log(\rho(t,x))\|_2^2 \right) \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t \qquad \text{(A.3)}$$
$$- \frac{\varepsilon}{2}(\mathcal{H}(P) + \mathcal{H}(Q),$$

subject to the same conditions as (A.2), where $\Lambda_\varepsilon = (2\pi\varepsilon)^{d/2}$.

If we plug in the minimizers from (A.2) into (A.3), we get exactly the result of (A.1) by optimality

$$\begin{aligned}\mathrm{OT}_\varepsilon(P,Q) + \varepsilon \log(\Lambda_\varepsilon) &\leq \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2}\|v_0(t,x)\|_2^2 \rho_0(t,x)\,\mathrm{d}x\,\mathrm{d}t + \frac{\varepsilon^2}{8} I_0(P,Q) - \frac{\varepsilon}{2}(\mathcal{H}(P) + \mathcal{H}(Q)), \\ &= \frac{1}{2} W_2^2(P,Q) + \frac{\varepsilon^2}{8} I_0(P,Q) - \frac{\varepsilon}{2}(\mathcal{H}(P) + \mathcal{H}(Q)),\end{aligned}$$

where we identify $I_0(P,Q) = \int_0^1 \int_{\mathbb{R}^d} \|\nabla_x \log \rho_0(t,x)\|_2^2 \rho_0(t,x)\,\mathrm{d}x\,\mathrm{d}t$.

A.2 Laplace's method proof

In this section, we prove a quantitative approximation to the integral

$$I(\varepsilon) := \frac{1}{\Lambda_\varepsilon} \int \exp\left(-\frac{1}{\varepsilon} f(x)\right) \mathrm{d}x\,, \qquad \text{(A.4)}$$

when $\varepsilon \to 0$, with f convex and sufficiently regular and where $\Lambda_\varepsilon = (2\pi\varepsilon)^{d/2}$. This approximation relies on expanding f around its global minimum; assuming that f is twice-differentiable, the behavior of f near its minimum will be quadratic, so that (A.4) will resemble a Gaussian integral for ε sufficiently small.

Recall that for a positive definite matrix S, we define $J(S) := \sqrt{\det(S)}$.

In what follows, we write $\mathrm{d}^2 f(0,y), \mathrm{d}^3 f(0,y)$ for the second and third total derivative of f at

x, respectively. That is, for $y \in \mathbb{R}^d$

$$\mathrm{d}^2 f(x,y) := y^\top \nabla^2 f(x) y, \;\; \mathrm{d}^3 f(x,y) := \sum_{i,j,k=1}^{d} \frac{\partial^3 f(x)}{\partial y_i \partial y_j \partial y_k} y_i y_j y_k \,.$$

We also define the set $B_r(a) := \{y \in \mathbb{R}^d \mid \|y - a\| \le r\}$, for some $r > 0$ and $a \in \mathbb{R}^d$.

Theorem A.2. *Let $I(\varepsilon)$ be as in (A.4), with $f \in C^{\alpha+1}$, m-strongly convex, M-smooth, and $\alpha > 1$. Assume f has a global minimum at x^*. Then there exist positive constants c and C depending on m, M, α, d, and $\|f\|_{C^{\alpha+1}}$ such that for all $\varepsilon \in (0,1)$,*

$$c \le J(\nabla^2 f(x^*)) I(\varepsilon) \le 1 + C(\varepsilon^{(\alpha-1)/2 \wedge 1}) \,. \tag{A.5}$$

Proof. Without loss of generality, we may assume that $x^* = 0$. For the remainder of the proof, we let $A := \nabla^2 f(0)$. Let $\tau = C_{m,M,d,\alpha}\sqrt{\log(2\varepsilon^{-1})}$, where the constant is to be decided later. We split the desired integral into two parts:

$$I(\varepsilon) = \frac{1}{\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(0)} e^{\frac{-1}{\varepsilon} f(y)} \,\mathrm{d}y + \frac{1}{\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(0)^c} e^{\frac{-1}{\varepsilon} f(y)} \,\mathrm{d}y =: I_1(\varepsilon) + I_2(\varepsilon) \,.$$

Lower bounds Note that $I_2(\varepsilon) \ge 0$, so it suffices to only prove $I_1(\varepsilon) \ge \frac{c}{\sqrt{\det(A)}}$ for some constant $c > 0$.

Since $f \in C^{\alpha+1}$, we have the following Taylor expansion

$$-f(y) \ge -\frac{1}{2} y^\top A y - C\|y\|^{(\alpha+1)\wedge 3} \ge -\frac{M}{2}\|y\|^2 - C\|y\|^{(\alpha+1)\wedge 3}$$

for some constant $C > 0$. Using this expansion, we arrive at

$$\begin{aligned} I_1(\varepsilon) &= \frac{1}{\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(0)} \exp\left[-\frac{M}{2\varepsilon}\|y\|^2 - \frac{C}{\varepsilon}\|y\|^{(\alpha+1)\wedge 3}\right] \mathrm{d}y \\ &\geq \frac{1}{\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(0)} \exp\left[-\frac{M}{2\varepsilon}\|y\|^2 - \frac{C}{\varepsilon}(\tau\sqrt{\varepsilon})^{(\alpha+1)\wedge 3}\right] \mathrm{d}y\,. \end{aligned}$$

Performing a change of measure and rearranging, we get

$$\begin{aligned} J(A)I_1(\varepsilon) &\geq e^{-C(\tau\sqrt{\varepsilon})^{(\alpha+1)\wedge 3}/\varepsilon} \frac{J(A)}{(2M\pi)^{d/2}} \int_{B_{\tau\sqrt{M}(0)}} e^{-\frac{1}{2}\|y\|^2} \mathrm{d}y \\ &\gtrsim e^{-C(\tau\sqrt{\varepsilon})^{(\alpha+1)\wedge 3}/\varepsilon} J(A)\mathbb{P}(\|Y\| \leq \tau\sqrt{M})\,, \end{aligned}$$

where $Y \sim N(0, I_d)$. Since $\alpha > 1$, the quantity $C(\tau\sqrt{\varepsilon})^{(\alpha+1)\wedge 3}/\varepsilon$ is bounded as $\varepsilon \to 0$, so we may bound $e^{-C(\tau\sqrt{\varepsilon})^{(\alpha+1)\wedge 3}/\varepsilon}$ from below by a constant. Since $J(A)$ and $\mathbb{P}(\|Y\| \leq \tau\sqrt{M})$ are both also bounded from below, we obtain that $J(A)I_1(\varepsilon) \geq c > 0$, as desired.

Upper bounds We first show that the contribution from $I_2(\varepsilon)$ is negligible. The strong convexity of f implies

$$f \geq \frac{m}{2}\|y\|^2,$$

leading us to the upper bound

$$\begin{aligned} I_2(\varepsilon) &\leq \frac{1}{\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(0)^c} e^{-\frac{m}{2\varepsilon}\|y\|^2} \mathrm{d}y \\ &= \frac{1}{(2m\pi)^{d/2}} \int_{B_\tau(0)^c} e^{-\frac{1}{2}\|y\|^2} \mathrm{d}y \\ &\leq \frac{1}{(2m\pi)^{d/2}} e^{-\frac{1}{4}\tau^2} \int e^{-\frac{1}{4}\|y\|^2} \mathrm{d}y \\ &\lesssim e^{-\frac{1}{4}\tau^2}, \end{aligned}$$

where in the penultimate inequality we have used the fact that $e^{-\frac{1}{4}\|y\|^2} \leq e^{-\frac{1}{4}\tau^2}$ on $B_\tau(0)^c$. Taking $C_{m,M,d,\alpha}$ sufficiently large in the definition of τ, we can make this term smaller than ε.

For upper bounds on $I_1(\varepsilon)$, we proceed in a similar fashion. If $f \in C^{\alpha+1}$ for $\alpha \in (1, 2]$, then we employ the bound

$$-f(y) \leq -\frac{1}{2}y^\top A y + C\|y\|^{\alpha+1},$$

yielding

$$I_1(\varepsilon) = \frac{1}{\Lambda_\varepsilon}\int_{B_{\tau\sqrt{\varepsilon}}(0)} e^{-\frac{1}{\varepsilon}f(y)}\,\mathrm{d}y \leq \frac{1}{\Lambda_\varepsilon}\int_{B_{\tau\sqrt{\varepsilon}}(0)} e^{-\frac{1}{2\varepsilon}y^\top A y + \frac{C}{\varepsilon}\|y\|^{\alpha+1}}\,\mathrm{d}y\,.$$

Performing the change of variables $u = \sqrt{1/\varepsilon}y$, we arrive at

$$I_1(\varepsilon) \leq \frac{1}{(2\pi)^{d/2}}\int_{B_\tau(0)} e^{\frac{-1}{2}u^\top A u} e^{C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}}\,\mathrm{d}u$$

Since $\alpha > 1$, the term $C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}$ is bounded above on $B_\tau(0)$, so that there exists a positive constant C' such that

$$e^{C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}} \leq 1 + C'\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1} \quad \forall u \in B_\tau(0)\,.$$

We obtain

$$\begin{aligned} I_1(\varepsilon) &\leq \frac{1}{(2\pi)^{d/2}}\int_{B_\tau(0)} e^{\frac{-1}{2}u^\top A u}(1 + C'\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1})\,\mathrm{d}u \\ &\leq \frac{1}{(2\pi)^{d/2}}\int e^{\frac{-1}{2}u^\top A u}(1 + C'\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1})\,\mathrm{d}u. \end{aligned}$$

Performing another change of variables yields

$$I_1(\varepsilon) \leq \frac{1}{(2\pi)^{d/2}J(A)}\int (1 + C'\varepsilon^{(\alpha-1)/2}\|A^{-1/2}u\|^{\alpha+1})e^{-\frac{1}{2}\|u\|^2}\,\mathrm{d}u$$

We obtain

$$J(A)I_1(\varepsilon) \leq 1 + C''\varepsilon^{(\alpha-1)/2}\,.$$

Combining this with the bound on $J(A)I_2(\varepsilon)$ yields the bound for $\alpha \leq 2$.

When $\alpha > 2$, we use the same technique but expand to the third order, yielding

$$\begin{aligned} I_1(\varepsilon) &= \frac{1}{\Lambda_\varepsilon}\int_{B_{\tau\sqrt{\varepsilon}}(0)} e^{-\frac{1}{\varepsilon}f(y)}\,\mathrm{d}y \\ &\leq \frac{1}{\Lambda_\varepsilon}\int_{B_{\tau\sqrt{\varepsilon}}(0)} e^{-\frac{1}{2\varepsilon}y^\top Ay - \frac{1}{6\varepsilon}\,\mathrm{d}^3 f(0,y) + \frac{C}{\varepsilon}\|y\|^{\alpha+1}}\,\mathrm{d}y \\ &= \frac{1}{(2\pi)^{d/2}}\int_{B_\tau(0)} e^{-\frac{1}{2}u^\top Au - \frac{\varepsilon^{1/2}}{6}\,\mathrm{d}^3 f(0,u) + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}}\,\mathrm{d}u \end{aligned}$$

Since $-\frac{\varepsilon^{1/2}}{6}\,\mathrm{d}^3 f(0,u) + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}$ is bounded on $B_\tau(0)$, we have

$$e^{-\frac{\varepsilon^{1/2}}{6}\,\mathrm{d}^3 f(0,u) + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1}} \leq 1 - \frac{\varepsilon^{1/2}}{6}\,\mathrm{d}^3 f(0,u) + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1} + R(u)\,,$$

where R is a positive remainder term satisfying $R(u) \lesssim \varepsilon(\,\mathrm{d}^3 f(0,u))^2 + \varepsilon^{\alpha-1}\|u\|^{2(\alpha+1)}$. We obtain

$$I_1(\varepsilon) \leq \frac{1}{(2\pi)^{d/2}}\int_{B_\tau(0)}\left(1 - \frac{\varepsilon^{1/2}}{6}\,\mathrm{d}^3 f(0,u) + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1} + R(u)\right)e^{-\frac{1}{2}u^\top Au}\,\mathrm{d}u\,.$$

The symmetry of $B_\tau(0)$ and the fact that $\,\mathrm{d}^3 f(0,u)e^{-\frac{1}{2}u^\top Au}$ is an odd function of u imply

$$\int_{B_\tau(0)} \mathrm{d}^3 f(0,u)e^{-\frac{1}{2}u^\top Au}\,\mathrm{d}u = 0\,,$$

so

$$\begin{aligned} I_1(\varepsilon) &\le \frac{1}{(2\pi)^{d/2}} \int (1 + C\varepsilon^{(\alpha-1)/2}\|u\|^{\alpha+1} + R(u))e^{-\frac{1}{2}u^\top Au}\,\mathrm{d}u \\ &= \frac{1}{(2\pi)^{d/2}J(A)} \int (1 + C\varepsilon^{(\alpha-1)/2}\|A^{-1/2}u\|^{\alpha+1} + R(A^{-1/2}u))e^{-\frac{1}{2}\|u\|^2}\,\mathrm{d}u \\ &\le 1 + C''\varepsilon^{(\alpha-1)/2} + C''\varepsilon\,, \end{aligned}$$

which is the desired bound. □

Corollary A.3. *Assume **(E2)** and **(E3)**. For all $\alpha \in (1, 3]$, there exist positive constants c and C such that*

$$c \le J(\nabla^2\varphi_0^*(x^*))Z_\varepsilon(x) \le 1 + C\varepsilon^{(\alpha-1)/2}\,, \tag{A.6}$$

for all $\varepsilon \in (0, 1)$ and $x \in \mathrm{supp}(P)$.

Proof. Take $f(\cdot) = D[\cdot|x^*]$ which is $1/L$-strongly convex, and $1/\mu$-smooth, with minimizer x^* (see (2.13)). The claim now follows from Theorem A.2. □

A.3 Omitted proofs

A.3.1 Proof of Proposition 2.1

It suffices to show that

$$\mathrm{OT}_\varepsilon(P, Q) \ge \sup_{\eta \in L^1(\pi_\varepsilon)} \int \eta\,\mathrm{d}\pi_\varepsilon - \varepsilon \iint e^{(\eta(x,y)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}P(x)\,\mathrm{d}Q(y) + \varepsilon\,,$$

since the other direction follows from choosing $\eta(x, y) = f(x) + g(y)$ and using (1.24).

Write

$$\gamma(x,y) = e^{\frac{1}{\varepsilon}(f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2)}$$

for the $P \otimes Q$ density of π_ε. The inequality

$$a \log a \geq ab - e^b + a$$

holds for all $a \geq 0$ and $b \in \mathbb{R}$, as can be seen by noting that the right side is a concave function of b which achieves its maximum at $b = \log a$. Applying this inequality with $a = \gamma(x,y)$ and $b = \frac{1}{\varepsilon}(\eta(x,y) - \frac{1}{2}\|x-y\|^2)$ and integrating with respect to $P \otimes Q$ yields

$$\int \log \gamma \, \mathrm{d}\pi_\varepsilon \geq \frac{1}{\varepsilon}\left(\int \eta \, \mathrm{d}\pi_\varepsilon - \int \frac{1}{2}\|x-y\|^2 \, \mathrm{d}\pi_\varepsilon(x,y) \right) - \iint e^{(\eta(x,y)-\frac{1}{2}\|x-y\|^2)/\varepsilon} \, \mathrm{d}P(x) \, \mathrm{d}Q(y) + 1$$

Multiplying by ε and using the fact that

$$\int \varepsilon \log \gamma \, \mathrm{d}\pi_\varepsilon = \int (f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2) \, \mathrm{d}\pi_\varepsilon = \mathrm{OT}_\varepsilon(P,Q) - \int \frac{1}{2}\|x-y\|^2 \, \mathrm{d}\pi_\varepsilon(x,y)$$

yields the claim. □

A.3.2 Proof of Proposition 2.15

Proposition 2.15 follows from the following more general result by choosing $\tilde{P} = P_n$.

Proposition A.4. *Let P and Q be probability measures with support contained in Ω, and denote by P_n and Q_n corresponding empirical measures. If $\tilde{P}$ is a probability measure with support in Ω such*

that $\mathrm{TV}(\tilde{P}, P_n) \le \delta$ *for some* $\delta \ge 0$*, then*

$$\mathbb{E}\Big\{ \sup_{\chi:\Omega\times\Omega\to\mathbb{R}} \iint \chi(x,y)\,\mathrm{d}\pi_{\varepsilon,n}(x,y) - \iint (e^{\chi(x,y)} - 1)\tilde{\gamma}(x,y)\,\mathrm{d}P(x)\,\mathrm{d}Q_n(y)\Big\} \\ \lesssim \varepsilon^{-1}\delta + (\varepsilon^{-1} + \varepsilon^{-d/2})\log(n)n^{-1/2}\,,$$

where $\pi_{\varepsilon,n}$ *is the optimal entropic plan for* P *and* Q_n*,* $\tilde{\gamma}$ *is the* $\tilde{P}\otimes Q_n$ *density of the optimal entropic plan for* $\tilde{P}$ *and* Q_n*, and the supremum is taken over all* $\chi \in L^1(\pi_{\varepsilon,n})$*.*

Proof. Write $\tilde{f}_\varepsilon$ and $\tilde{g}_\varepsilon$ for the optimal entropic potentials for $\tilde{P}$ and Q_n, so that

$$\tilde{\gamma}(x,y) = \exp\left(\varepsilon^{-1}(\tilde{f}_\varepsilon(x) + \tilde{g}_\varepsilon(y) - \frac{1}{2}\|x-y\|^2)\right).$$

Plugging in $\eta(x,y) = \varepsilon\chi(x,y) + \tilde{f}_\varepsilon(x) + \tilde{g}_\varepsilon(y)$ into Proposition 2.1 gives

$$\sup_{\chi:\Omega\times\Omega\to\mathbb{R}} \iint \chi\,\mathrm{d}\pi_{\varepsilon,n} - \iint (e^{\chi(x,y)} - 1)\tilde{\gamma}(x,y)\,\mathrm{d}P(x)\,\mathrm{d}Q_n(y) \le \varepsilon^{-1}\Big(\mathrm{OT}_\varepsilon(P,Q_n) \\ - \int \tilde{f}_\varepsilon\,\mathrm{d}P - \int \tilde{g}_\varepsilon\,\mathrm{d}Q_n\Big),$$

where we have used that $\tilde{\gamma}$ is a probability density with respect to $P\otimes Q_n$ by the optimality conditions (1.27) and (1.28).

Let $f_{\varepsilon,n}$ and $g_{\varepsilon,n}$ be the optimal entropic dual potentials for P and Q_n. As in the proof of Lemma A.8, the optimality of $\tilde{f}_\varepsilon$ and $\tilde{g}_\varepsilon$ for the pair $(\tilde{P}, Q_n)$ implies

$$\int \tilde{f}_\varepsilon\,\mathrm{d}\tilde{P} + \int \tilde{g}_\varepsilon\,\mathrm{d}Q_n \ge \int f_{\varepsilon,n}\,\mathrm{d}\tilde{P} + \int g_{\varepsilon,n}\,\mathrm{d}Q_n - \varepsilon \iint e^{\frac{1}{\varepsilon}(f_{\varepsilon,n}(x)+g_{\varepsilon,n}(y)-\frac{1}{2}\|x-y\|^2)}\,\mathrm{d}\tilde{P}(x)\,\mathrm{d}Q_n(y) + \varepsilon \\ = \int f_{\varepsilon,n}\,\mathrm{d}\tilde{P} + \int g_{\varepsilon,n}\,\mathrm{d}Q_n\,,$$

since $\int e^{\frac{1}{\varepsilon}(f_{\varepsilon,n}(x)+g_{\varepsilon,n}(y)-\frac{1}{2}\|x-y\|^2)}\,\mathrm{d}Q_n(y) = 1$ by the dual optimality condition (1.27). Therefore

$$\begin{aligned}\mathrm{OT}_\varepsilon(P,Q_n) - \int \tilde{f}_\varepsilon\,\mathrm{d}P - \int \tilde{g}_\varepsilon\,\mathrm{d}Q_n &\le \int (f_{\varepsilon,n}-\tilde{f}_\varepsilon)(\mathrm{d}P-\mathrm{d}\tilde{P}) \\ &= \int (f_{\varepsilon,n}-\tilde{f}_\varepsilon)(\mathrm{d}P-\mathrm{d}P_n) + \int (f_{\varepsilon,n}-\tilde{f}_\varepsilon)(\mathrm{d}P_n-\mathrm{d}\tilde{P})\end{aligned}$$

By Genevay et al. (2019, Proposition 1), we may choose $f_{\varepsilon,n}$ and $\tilde{f}_\varepsilon$ to satisfy $\left\|f_{\varepsilon,n}\right\|_\infty, \|\tilde{f}_\varepsilon\|_\infty \lesssim 1$, so we may bound the second term as

$$\int (f_{\varepsilon,n}-\tilde{f}_\varepsilon)(\mathrm{d}P_n-\mathrm{d}\tilde{P}) \lesssim \mathrm{TV}(\tilde{P},P_n) \le \delta\,.$$

Also, since $f_{\varepsilon,n}$ is independent of P_n,

$$\mathbb{E} f_{\varepsilon,n}(\mathrm{d}P-\mathrm{d}P_n)(y) = 0\,.$$

Altogether, we obtain

$$\mathbb{E}\sup_{\chi:\Omega\times\Omega\to\mathbb{R}} \iint \chi\,\mathrm{d}\pi_{\varepsilon,n} - \iint (e^{\chi(x,y)}-1)\gamma(x,y)\,\mathrm{d}P(x)\,\mathrm{d}Q_n(y) \lesssim \varepsilon^{-1}\left(\delta + \mathbb{E}\int \tilde{f}_\varepsilon(\mathrm{d}P_n-\mathrm{d}P)\right).$$

We conclude by again appealing to Genevay et al. (2019, Proposition 1): since $\tilde{f}_\varepsilon$ is an optimal entropic potential for the pair of compactly distributed probability measures $(\tilde{P},Q_n)$, its derivatives up to order s are bounded by $C_{s,d,K}(1+\varepsilon^{1-s})$ on any compact set K for any $s \ge 0$. Taking K to be a suitably large ball containing Ω and applying Lemma A.8 with $s = d/2$ yields the claim. □

A.3.3 Proofs from Section 2.2.1

Proof of Lemma 2.12. Fix $x \in \mathrm{supp}(P)$ and let $x^* := T_0(x)$, and for notational convenience, write Y for the random variable with density q_ε^x, and denote its mean by $\bar{Y}$. It suffices to show the

existence of a constant K such that for any unit vector v,

$$\mathbb{E}e^{(v^\top(Y-x^*))^2/4L\varepsilon} \le K\,. \tag{A.7}$$

Indeed, by Young's and Jensen's inequalities, this implies

$$\mathbb{E}e^{(v^\top(Y-\bar{Y}))^2/8L\varepsilon} \le e^{(v^\top(\bar{Y}-x^*))^2/4L\varepsilon}\mathbb{E}e^{(v^\top(Y-x^*))^2/4L\varepsilon} \le K^2\,,$$

and hence by another application of Jensen's inequality that

$$\mathbb{E}e^{(v^\top(Y-\bar{Y}))^2/C\varepsilon} \le 2$$

for $C = 8LK^2$.

We prove (A.7) using the strong convexity of $D[y|x^*]$. By (2.13),

$$\begin{aligned}
\mathbb{E}e^{(v^\top(Y-x^*))^2/4L\varepsilon} &\le \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon}\int e^{-\frac{1}{\varepsilon}D[y|x^*]+\frac{1}{4L\varepsilon}\|y-x^*\|^2}\,\mathrm{d}y \\
&\le \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon}\int e^{-\frac{1}{4L\varepsilon}\|y-x^*\|^2}\,\mathrm{d}y \\
&= \frac{(2L)^{d/2}}{Z_\varepsilon(x)} \\
&\lesssim 1\,,
\end{aligned}$$

where the final inequality uses Corollary A.3.

□

Proof of Lemma 2.13. Let us first fix an $x \in \mathrm{supp}(P)$, and write Y for the random variable with density q_ε^x and $\bar{Y}$ for its mean, and write $x^* := T_0(x)$. Lemma 2.12 implies (see Vershynin, 2018,

Proposition 2.5.2) that there exists a positive constant C, independent of x, such for any $v \in \mathbb{R}^d$,

$$\mathbb{E}e^{(v^\top(Y-x^*))} = e^{v^\top(\bar{Y}-x^*)}\mathbb{E}e^{(v^\top(Y-\bar{Y}))} \le e^{v^\top(\bar{Y}-x^*)+C\varepsilon\|v\|^2} \le e^{\frac{1}{4\varepsilon}\|\bar{Y}-x^*\|^2+(C+1)\varepsilon\|v\|^2},$$

where the last step uses Young's inequality. Equivalently, for $a > (C+1)\varepsilon$, we have for all $x \in \operatorname{supp}(P)$ and $v \in \mathbb{R}^d$

$$\int e^{(v^\top(y-x^*))-a\|v\|^2} q_\varepsilon^x(y)\,\mathrm{d}y \le e^{\frac{1}{4\varepsilon}\|\bar{y}^x-x^*\|^2}.$$

Applying this inequality with $v = h(x)$ and integrating with respect to P yields the claim. □

Proof of Lemma 2.14. It suffices to prove the claim for $\alpha \in (1,2]$. Let us fix an $x \in \operatorname{supp}(P)$. Since $\varphi_0^* \in C^{\alpha+1}(\Omega)$, Taylor's theorem implies

$$D[y|x^*] = -x^\top y + \varphi_0(x) + \varphi_0^*(y) = \frac{1}{2}(y-x^*)^\top\nabla^2\varphi_0^*(x^*)(y-x^*) + R(y|x^*),$$

where the remainder satisfies

$$|R(y|x^*)| \lesssim \|y-x^*\|^{1+\alpha}. \tag{A.8}$$

We aim to bound

$$\|\bar{y}^x - x^*\| = \left\|\frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon}\int (y-x^*)e^{-\frac{1}{\varepsilon}D[y|x^*]}\,\mathrm{d}y\right\|$$

Let $\tau = C\sqrt{\log(2\varepsilon^{-1})}$ for a sufficiently large constant C. As in the proof of Theorem A.2, the

contribution to the integral from the set $B_{\tau\sqrt{\varepsilon}}(x^*)^c$ is negligible; indeed, (2.13) implies

$$\begin{aligned}
&\left\| \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(x^*)^c} (y-x^*) e^{-\frac{1}{\varepsilon}D[y|x^*]}\,\mathrm{d}y \right\| \\
&\quad \le \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(x^*)^c} \|y-x^*\| e^{-\frac{1}{2L\varepsilon}\|y-x^*\|^2}\,\mathrm{d}y \\
&\quad = \frac{\varepsilon^{(d+1)/2}}{Z_\varepsilon(x)\Lambda_\varepsilon} \int_{B_\tau(0)^c} \|y\| e^{-\frac{1}{2L}\|y\|^2}\,\mathrm{d}y \\
&\quad \le \frac{\varepsilon^{(d+1)/2}}{Z_\varepsilon(x)\Lambda_\varepsilon} \left(\int \|y\|^2 e^{-\frac{1}{2L}\|y\|^2}\,\mathrm{d}y \right)^{1/2} \left(\int_{B_\tau(0)^c} e^{-\frac{1}{2L}\|y\|^2}\,\mathrm{d}y \right)^{1/2} \\
&\quad \lesssim \varepsilon^{1/2} \mathbb{P}[\|Y\| \ge \tau]\,, \qquad Y \sim \mathcal{N}(0, I_d)\,,
\end{aligned}$$

and this quantity can be made smaller than ε by choosing the constant in the definition of τ sufficiently large.

It remains to bound

$$\begin{aligned}
&\left\| \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(x^*)} (y-x^*) e^{-\frac{1}{\varepsilon}R(y|x^*)} e^{-\frac{1}{2\varepsilon}(y-x^*)^\top \nabla^2 \varphi_0^*(x^*)(y-x^*)}\,\mathrm{d}y \right\| = \\
&\qquad \left\| \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon} \int_{B_{\tau\sqrt{\varepsilon}}(x^*)} (y-x^*) \left(e^{-\frac{1}{\varepsilon}R(y|x^*)} - 1 \right) e^{-\frac{1}{2\varepsilon}(y-x^*)^\top \nabla^2 \varphi_0^*(x^*)(y-x^*)}\,\mathrm{d}y \right\|, \quad \text{(A.9)}
\end{aligned}$$

where we have used that

$$\int_{B_{\tau\sqrt{\varepsilon}}(x^*)} (y-x^*) e^{-\frac{1}{2\varepsilon}(y-x^*)^\top \nabla^2 \varphi_0^*(x^*)(y-x^*)}\,\mathrm{d}y = 0\,.$$

By (A.8),

$$\frac{1}{\varepsilon}|R(y|x^*)| \lesssim \frac{1}{\varepsilon}\|y-x^*\|^{1+\alpha} \lesssim 1 \qquad \forall y \in B_{\tau\sqrt{\varepsilon}}(x^*)\,,$$

and since $|e^t - 1| \lesssim |t|$ for $|t| \lesssim 1$, we obtain that

$$\left| e^{-\frac{1}{\varepsilon}R(y|x^*)} - 1 \right| \lesssim \frac{1}{\varepsilon}\|y-x^*\|^{1+\alpha} \qquad \forall y \in B_{\tau\sqrt{\varepsilon}}(x^*)\,.$$

Therefore (A.9) is bounded above by

$$\frac{C}{Z_\varepsilon(x)\Lambda_\varepsilon\varepsilon}\int_{\mathbb{R}^d}\|y-x^*\|^{2+\alpha}\exp\big(-\frac{1}{2\varepsilon}(y-x^*)^\top\nabla^2\varphi_0^*(x^*)(y-x^*)\big)\,\mathrm{d}y \lesssim \varepsilon^{\alpha/2}\,,$$

where in the last step we have applied Corollary A.3. We therefore obtain

$$\|\mathbb{E}_{q_\varepsilon^x}(Y)-x^*\| \;\lesssim \varepsilon^{\alpha/2}+\varepsilon\,.$$

Taking squares, we get the desired result. □

A.4 Supplementary results

Proposition A.5. *For any $x\in\operatorname{supp}(P)$, if $a\in[L\varepsilon,1]$, then*

$$\mathbb{E}\sup_{h:\Omega\to\mathbb{R}^d}\int e^{j_h(x,y)}\frac{q_\varepsilon^x(y)}{q(y)}(\mathrm{d}Q_n-\mathrm{d}Q)(y)\lesssim(1+\varepsilon^{-d/2})n^{-1/2}\,,$$

where the implicit constant is uniform in x.

Proof. To bound this process, we employ the following two lemmas:

Lemma A.6. *If $a\geq L\varepsilon$, then for any $v\in\mathbb{R}^d$,*

$$v^\top(y-x^*)-a\|v\|^2-\frac{1}{\varepsilon}D[y|x^*]\leq-\frac{\varepsilon L}{2}\|v\|^2\,.$$

Proof. By (2.13), $D[y|x^*]\geq\frac{1}{2L}\|y-x^*\|^2$. Combining this fact with Young's inequality yields

$$v^\top(y-x^*)-a\|v\|^2-\frac{1}{\varepsilon}D[y|x^*]\leq\frac{\varepsilon L}{2}\|v\|^2+\frac{1}{2\varepsilon L}\|y-x^*\|^2-a\|v\|^2-\frac{1}{\varepsilon}D[y|x^*]\leq-\frac{\varepsilon L}{2}\|v\|^2\,,$$

as claimed. □

By slight abuse of notation, for any $v \in \mathbb{R}^d$, write $j_v : \Omega \to \mathbb{R}$ for the function

$$j_v(y) = v^\top (y - T_0(x)) - a\|v\|^2 .$$

Let

$$\mathcal{J}_\varepsilon = \{e^{j_v} \frac{q_\varepsilon^x(y)}{q(y)} : v \in \mathbb{R}^d\} \tag{A.10}$$

Lemma A.7. *If $a \in [L\varepsilon, 1]$, then*

$$\log N(\tau, \mathcal{J}_\varepsilon, \|\cdot\|_{L^\infty(Q)}) \lesssim d \log(K/\tau) ,$$

where $K \lesssim (1 + \varepsilon^{-d/2})$.

Proof. Fix $\delta \in (0, 1)$. Let $\mathcal{N}_\delta$ be a $\delta^{3/2}$-net with respect to the Euclidean metric of a ball of radius $\delta^{-1/2}$ in $\mathbb{R}^d$, and consider the set

$$\mathcal{G}_\delta := \left\{ e^{j_v} \frac{q_\varepsilon^x(y)}{q(y)} : v \in \mathcal{N}_\delta \right\} \cup \{e^{j_w} \frac{q_\varepsilon^x(y)}{q(y)}\} ,$$

where $w \in \mathbb{R}^d$ is an arbitrary vector of norm $\delta^{-1/2}$. By Lemma A.6, if $a > L\varepsilon$ and $\|v\| \geq R$, then

$$\sup_{y \in \mathrm{supp}(Q)} e^{j_v} \frac{q_\varepsilon^x(y)}{q(y)} \leq \sup_{y \in \mathrm{supp}(Q)} \frac{1}{Z_\varepsilon(x)\Lambda_\varepsilon q(y)} e^{-\frac{1}{2\varepsilon L}R^2} \leq \sup_{y \in \mathrm{supp}(Q)} \frac{2L}{Z_\varepsilon(x)\Lambda_\varepsilon q(y) R^2} .$$

Therefore, if $v \in \mathbb{R}^d$ satisfies $\|v\| \geq \delta^{-1/2}$, then

$$\sup_{y \in \mathrm{supp}(Q)} \left| e^{j_w(y)} \frac{q_\varepsilon^x(y)}{q(y)} - e^{j_v(y)} \frac{q_\varepsilon^x(y)}{q(y)} \right| \leq \sup_{y \in \mathrm{supp}(Q)} \frac{4\delta L}{Z_\varepsilon(x)\Lambda_\varepsilon q(y)} \leq K\delta$$

for $K = \sup_{y \in \mathrm{supp}(Q)} \frac{1+4L}{Z_\varepsilon(x)\Lambda_\varepsilon q(y)} \lesssim \varepsilon^{-d/2}$.

On the other hand, if $v \in \mathbb{R}^d$ satisfies $\|v\| \le \delta^{-1/2}$, pick $u \in \mathcal{N}_\delta$ satisfying $\|u - v\| \le \delta^{3/2}$. We then have

$$\left| e^{j_u(y)} \frac{q_\varepsilon^x(y)}{q(y)} - e^{j_v(y)} \frac{q_\varepsilon^x(y)}{q(y)} \right| \le \frac{|j_u(y) - j_v(y)|}{q(y)},$$

where we have used Lemma A.6 combined with the inequality

$$|e^a - e^b| \le |a - b| \qquad \forall a, b \le 0\,.$$

Since $\|u - v\| \le \delta^{3/2}$ and $\|u\| + \|v\| \le 2\delta^{-1/2}$, we have for any $y \in \Omega$,

$$|j_u(y) - j_v(y)| = |(u - v)^\top (y - T_0(x)) - a(\|u\|^2 - \|v\|^2)| \lesssim \delta^{3/2} + \delta a\,,$$

where we have used the fact that y and $T_0(x)$ lie in the compact set Ω. Therefore, as long as $a \le 1$, this quantity is bounded by $C\delta$ for a positive constant C.

All told, we obtain that for any $v \in \mathbb{R}^d$, there exists a $g \in \mathcal{G}_\delta$ such that

$$\left\| e^{j_v} \frac{q_\varepsilon^x(y)}{q(y)} - g \right\|_{L^\infty(Q)} \lesssim K\delta\,,$$

where $K \lesssim 1 + \varepsilon^{-d/2}$. Moreover, Lemma A.6 implies that, for any $g \in \mathcal{G}_\delta$,

$$\|g\|_{L^\infty(Q)} \le \sup_{y \in \mathrm{supp}(Q)} \frac{1}{Z_\varepsilon(x) \Lambda_\varepsilon q(y)} \le K.$$

By a volume argument, we may choose $\mathcal{N}_\delta$ such that it satisfies

$$\log |\mathcal{N}_\delta| \lesssim \log(1/\delta).$$

We therefore obtain for any $\tau \leq K$,

$$\log N(\tau, \mathcal{J}_\varepsilon, \|\cdot\|_{L^\infty(Q)}) \leq \log |\mathcal{G}_{\tau/K}| \lesssim \log(K/\tau)\,,$$

as claimed. □

Returning to the empirical process, we obtain by a chaining bound (Giné and Nickl, 2021, Theorem 3.5.1)

$$\begin{aligned}\mathbb{E} \sup_{h:\Omega\to\mathbb{R}^d} \int e^{j_h(x,y)} \frac{q_\varepsilon^x(y)}{q(y)} (\mathrm{d}Q_n - \mathrm{d}Q)(y) &= \mathbb{E} \sup_{j\in\mathcal{J}} \int j(y)(\mathrm{d}Q_n - \mathrm{d}Q)(y) \\ &\lesssim n^{-1/2} \int_0^K \sqrt{\log(K/\tau)}\,\mathrm{d}\tau \\ &\lesssim K n^{-1/2}\,.\end{aligned}$$

Recalling that $K \lesssim (1+\varepsilon^{-d/2})$ completes the proof. □

Lemma A.8. *For a convex, compact $K \subseteq \mathbb{R}^d$, for any real number $s \geq d/2$, and $M > 0$, let $C^s(K;M)$ be the set of s-Hölder smooth functions with Hölder norm bounded by M. For any probability measure ν with support contained in K and corresponding empirical measure ν_n, we have that*

$$\mathbb{E} \sup_{g\in C^s(K;M)} \int g(y)(\mathrm{d}\nu_n(y) - \mathrm{d}\nu(y)) \lesssim C_K M \log(n) n^{-1/2}\,.$$

Proof. We write $\mathcal{F}$ to be the set of functions in $C^s(K;1)$. A version of Dudley's chaining bound (see, e.g., von Luxburg and Bousquet, 2003, Theorem 16) therefore implies for any $\delta \geq 0$,

$$\mathbb{E} \sup_{g\in C^s(K;M)} \int g(y)(\mathrm{d}\nu_n(y) - \mathrm{d}\nu(y)) \lesssim M\left(\delta + n^{-1/2} \int_\delta^1 \sqrt{\log N(\tau, \mathcal{F}, \|\cdot\|_\infty)}\,\mathrm{d}\tau\right).$$

Letting $s \geq d/2$ and applying standard covering number bounds for Hölder spaces (Vaart and

Wellner, 1996, Theorem 2.7.1) implies

$$\mathbb{E} \sup_{g \in C^s(K;M)} \int g(y)(\mathrm{d}\nu_n(y) - \mathrm{d}\nu(y)) \lesssim C_K \inf_{\delta \geq 0} M \left(\delta + n^{-1/2} \int_\delta^1 \tau^{-1} \, \mathrm{d}\tau \right).$$

Taking $\delta = n^{-1/2}$ yields

$$\mathbb{E} \sup_{g \in C^s(K;M)} \int g(y)(\mathrm{d}\nu_n(y) - \mathrm{d}\nu(y)) \lesssim C_K M n^{-1/2}(1 - \log(n^{-1/2})) \lesssim C_K M n^{-1/2} \log n,$$

as claimed. □

Lemma A.9. *Let P and Q be compactly supported, and let $(f_\varepsilon, g_\varepsilon)$ denote the optimal dual potentials corresponding to $\mathrm{OT}_\varepsilon(P, Q)$. For any real number $s \geq 0$, the derivatives of $(f_\varepsilon, g_\varepsilon)$ up to order s are bounded by $C_{s,d,K}(1 + \varepsilon^{1-s})$ on any compact set K, where $C_{s,d,K} > 0$ is some constant independent of ε.*

Proof. It suffices to show the claim for f_ε. Let r be a positive integer, and let $\lambda \in [0, 1]$. By Genevay et al. (2019, Theorem 2), it holds that

$$\|f_\varepsilon\|_{C^r} = O(1 + \varepsilon^{1-r}).$$

For any $s \geq 0$, we can write $s = r + (1 - \lambda)$ for some $\lambda \in (0, 1)$ and $r \in \mathbb{N}$. Consequently, any such s can be written as $s = \lambda r + (1 - \lambda)(r + 1)$, from which we can now apply an interpolation inequality between the two integers (Lunardi, 2009):

$$\begin{aligned}
\|f_\varepsilon\|_{C^{\lambda r + (1-\lambda)(r+1)}} &\lesssim \|f_\varepsilon\|_{C^r}^{\lambda} \|f_\varepsilon\|_{C^{r+1}}^{1-\lambda} \\
&\lesssim (1 + \varepsilon^{1-r})^{\lambda} (1 + \varepsilon^{-r})^{1-\lambda} \\
&\leq 1 + \varepsilon^{(1-r)\lambda - r(1-\lambda)} \\
&= 1 + \varepsilon^{-r+\lambda} \\
&= 1 + \varepsilon^{1-s}.
\end{aligned}$$

Thus, $\|f_\varepsilon\|_{C^s} = O(1+\varepsilon^{1-s})$ for any $s \geq 0$, as desired. □

Corollary A.10. *If P and Q are compactly supported, then*

$$\mathbb{E}\,\mathrm{OT}_\varepsilon(P, Q_n) - \mathrm{OT}_\varepsilon(P, Q) \lesssim (1+\varepsilon^{1-d/2})\log(n)n^{-1/2}\,.$$

Proof. Let $(f_{\varepsilon,n}, g_{\varepsilon_n})$ be the optimal dual potentials for P and Q_n. Following Mena and Niles-Weed (2019, Proposition 2), observe that

$$\begin{aligned}\mathrm{OT}_\varepsilon(P, Q_n) - \mathrm{OT}_\varepsilon(P, Q) &= \int f_{(\varepsilon,n)}\,\mathrm{d}P + \int g_{(\varepsilon,n)}\,\mathrm{d}Q_n - \sup_{f,g}\Big\{\int f\,\mathrm{d}P + \int g\,\mathrm{d}Q \\ &\qquad - \varepsilon\iint e^{(f(x)+g(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}P(x)\,\mathrm{d}Q(y) + \varepsilon\Big\} \\ &\leq \int g_{(\varepsilon,n)}(y)(\mathrm{d}Q_n(y) - \mathrm{d}Q(y))\,,\end{aligned}$$

where the bound follows from choosing $(f_{(\varepsilon,n)}, g_{(\varepsilon,n)})$ in the supremum and using

$$\int e^{(f_{(\varepsilon,n)}(x)+g_{(\varepsilon,n)}(y)-\frac{1}{2}\|x-y\|^2)/\varepsilon}\,\mathrm{d}P(x) = 1 \quad \forall y \in \mathbb{R}^d$$

by the dual optimality condition (1.28).

We conclude by applying Lemma A.9: the derivatives of $g_{\varepsilon,n}$ up to order s are bounded by $C_{s,d,K}(1+\varepsilon^{1-s})$ on any compact set K for any $s \geq 0$, so we may take K to be a suitably large ball containing the support of P and Q and apply Lemma A.8 with $s = d/2$. □

A.5 Proof of Theorem 2.16

We recall the notation from the main text. For convenience, we consider $\alpha \geq 1+\iota$ for some $\iota > 0$ sufficiently small, but fixed. Let $s := \alpha + 1$, which defines the regularity of the conjugate Brenier potential φ_0^*, thus $s \in [2+\iota, 4]$ for our problem considerations, since smoothness is capped

at $\alpha = 3$. Let $\mathcal{S}$ be the following discrete subset

$$\mathcal{S} \coloneqq \{s_{\min} = s_1 < s_2 < \cdots < s_N = s_{\max}\},$$

where $s_{\min} = 2 + \iota$, $s_N = 4$, with increments $s_j - s_{j-1} \asymp (\log n)^{-1}$, and set

$$\varepsilon_s = (n/\log n)^{-1/2(d+s)}, \quad \psi_n(s) = (\varepsilon_s)^s = (n/\log n)^{-s/2(d+s)}.$$

Let $\mathbb{D}_n \coloneqq \{(X_i, Y_i)\}_{i=1}^n$ denote our initial dataset with hold-out dataset $\mathbb{D}'_n$. The latter gives rise to empirical measures P'_n and Q'_n. Our choice of smoothness parameter is given by the following rule:

$$\hat{s} \coloneqq \max\{s \in \mathcal{S} \ : \ \|\hat{T}_{\varepsilon_s} - \hat{T}_{\varepsilon_{s'}}\|^2_{L^2(P'_n)} \lesssim \psi_n(s'), \forall\, s' \le s, s' \in \mathcal{S}\}. \tag{A.11}$$

The proof closely follows an exposition of Lepski's method due to Hütter and Mao (2017).

For a given probability measure and its empirical counterpart from n samples, written ρ and ρ_n, we will frequently return to the empirical process over a given function class $\mathcal{M}$, written

$$\|\rho - \rho_n\|_{\mathcal{M}} \coloneqq \sup_{f \in \mathcal{M}} \left| \int f \, \mathrm{d}(\rho - \rho_n) \right|.$$

We will consider the following function classes: $\mathcal{F}_\varepsilon$ will denote the class of entropic Kantorovich potentials for a regularization parameter ε, and $\mathcal{J}_\varepsilon$ be the function class from (A.10). $\mathcal{H}_N$ will denote the random, P_n-measurable set of N^2 bounded functions of the form

$$\|\hat{T}_{s_i}(x) - \hat{T}_{s_j}(x)\|_2^2,$$

for $i, j \in \{1, 2, \ldots, N\}$, where we recall that N is the cardinality of $\mathcal{S}$.

Without loss of generality, we can assume $\varphi_0^* \in C^{s_i}$ for some $s_i \in \mathcal{S}$. We define the event $\mathcal{E}_j \coloneqq \{\hat{s} = s_j\}$ for all $j \in [N]$, and denote our estimator by $\hat{T}_{\hat{s}}$ (for clarity, we omit the explicit

dependence on ε). The ratio between the risk of $\hat{T}_{\hat{s}}$ and the oracle rate $\psi_n(s_i)$ can be written as

$$\mathbb{E}[\|\hat{T}_{\hat{s}} - T_0\|^2_{L^2(P)}\psi_n(s_i)^{-1}] = \sum_{j=1}^{i-1} \mathbb{E}[\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)}\psi_n(s_i)^{-1}\mathbf{1}(\mathcal{E}_j)] + \sum_{j=i}^{N} \mathbb{E}[\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)}\psi_n(s_i)^{-1}\mathbf{1}(\mathcal{E}_j)] \,.$$

Our goal is to show that the right-hand side is upper bounded by an absolute constant. We study the two terms above separately.

Let us first focus on the terms where $j \geq i$, i.e. our estimator of the smoothness of the optimal transport map is larger than the actual smoothness parameter. Inside the expectation, we can write via Young's inequality

$$\begin{aligned}
\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)} &\lesssim \|\hat{T}_{s_j} - \hat{T}_{s_i}\|^2_{L^2(P)} + \|\hat{T}_{s_i} - T_0\|^2_{L^2(P)} \\
&= \|\hat{T}_{s_j} - \hat{T}_{s_i}\|^2_{L^2(P'_n)} + \|\hat{T}_{s_i} - T_0\|^2_{L^2(P)} + \int \tilde{h}\,\mathrm{d}(P - P'_n) \\
&\leq \|\hat{T}_{s_j} - \hat{T}_{s_i}\|^2_{L^2(P'_n)} + \|\hat{T}_{s_i} - T_0\|^2_{L^2(P)} + \|P - P'_n\|_{\mathcal{H}_N} \,,
\end{aligned}$$

where $\tilde{h} = \|\hat{T}_{s_j} - \hat{T}_{s_i}\|_2^2$. We conclude by taking expectations. The first term on the right-hand side is bounded by $\psi_n(s_i)$: our estimator $\hat{s} = s_j$ under the event $\mathcal{E}_j$, and our criterion for $\hat{s}$, namely (A.11), and $s_i \leq s_j$ by assumption. For the second term: as $\varphi_0^* \in C^{s_i}$, our main theorem (Theorem 2.5) tells us that

$$\mathbb{E}\|\hat{T}_{s_i} - T_0\|^2_{L^2(P)} \lesssim \psi_n(s_i) \,.$$

The third term, by Hoeffding's inequality and a union bound, satisfies

$$\mathbb{E}\|P'_n - P\|_{\mathcal{H}_N} = \mathbb{E}[\mathbb{E}[\|P'_n - P\|_{\mathcal{H}_N} \mid P_n]] \lesssim \log\log(n)/\sqrt{n} \,,$$

where we used that $N \asymp \log n$. Note that the third term is in fact faster than any $\psi_n(s_i)$ for any

choice of $s_i \in \mathcal{S}$. Altogether, this gives the following bound

$$\sum_{j=i}^{N} \mathbb{E}[\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)} \psi_n(s_i)^{-1} \mathbf{1}(\mathcal{E}_j)] \lesssim (c_0^2 + \tilde{c}_0^2) \sum_{j=i}^{N} \mathbb{P}(\mathcal{E}_j) + \bar{c}_0^2 \leq C_0 \,,$$

for three different constants $c_0, \tilde{c}_0, \bar{c}_0 > 0$.

We now turn our attention to the case where $j < i$, which is more technical. Focusing on one term in the summand, we want to choose t_j to appropriately balance

$$\mathbb{E}[\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)} \psi_n(s_i)^{-1} \mathbf{1}(\mathcal{E}_j)] \leq t_j \mathbb{P}(\mathcal{E}_j) + \int_{t_j}^{\infty} \mathbb{P}(\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)} \psi_n(s_i)^{-1} \geq t) \, \mathrm{d}t \,.$$

By definition of the estimator, we can upper bound $\mathbb{P}(\mathcal{E}_j)$ by two events, leading to

$$\mathbb{P}(\mathcal{E}_j) \leq \sum_{l=1}^{i-1} \left(\mathbb{P}(\|\hat{T}_{s_i} - T_0\|^2_{L^2(P'_n)} \psi_n(s_l)^{-1} > c_0^2/4) + \mathbb{P}(\|\hat{T}_{s_l} - T_0\|^2_{L^2(P'_n)} \psi_n(s_l)^{-1} > c_0^2/4) \right) . \quad \text{(A.12)}$$

Indeed, since $s_j < s_i$ and since we are on the set $\mathcal{E}_j$, there must exist an $s_j < s' < s_i$ such that

$$\|\hat{T}_{s_i} - \hat{T}_{s'}\|^2_{L^2(P'_n)} \psi_n(s') > c_0.$$

By Young's inequality, we can break this up into two possible events, whereby summing over all possible s' gives the above bound (we replace s' by s_l). Finally, we note that we also have the inequality

$$\mathbb{P}(\|\hat{T}_{s_i} - T_0\|^2_{L^2(P'_n)} \psi_n(s_l)^{-1} > c_0^2/4) \leq \mathbb{P}(\|\hat{T}_{s_i} - T_0\|^2_{L^2(P'_n)} \psi_n(s_i)^{-1} > c_0^2/4) \,,$$

since $\psi_n(\cdot)$ is decreasing. It remains to bound these two tail probabilities across all $l < i$, where note the norm is measured in $L^2(P'_n)$. To continue, we require the following lemma.

Proposition A.11. *There exist absolute constants $c, C > 0$ such that for $t \geq c$,*

$$\mathbb{P}\left(\|\hat{T}_s - T_0\|^2_{L^2(P)} \psi_n(s)^{-1} \geq ct\right) \leq \exp\left(-\frac{t^2 \log(n)}{C}\right).$$

Proof. For any choice of $s \in (2, 4]$, it holds that

$$\|\hat{T}_s - T_0\|^2_{L^2(P)} \lesssim \varepsilon^{s/2} + \|Q_n - Q\|_{\mathcal{J}_\varepsilon} + \varepsilon^{-1}\|P_n - P\|_{\mathcal{F}_\varepsilon},$$

which stems from the calculations that appear between Theorem 2.6 and Theorem 2.10. Both $\|Q_n - Q\|_{\mathcal{J}_\varepsilon}$ and $\|P_n - P\|_{\mathcal{F}_\varepsilon}$ are subGaussian random variables via McDiarmid's inequality: for two constants $a, b > 0$, it holds that for t large enough

$$\mathbb{P}(\|Q_n - Q\|_{\mathcal{J}_\varepsilon} \geq (1+t)(\varepsilon^{-d} n^{-1})^{1/2}) \leq e^{-at^2/2},$$
$$\mathbb{P}(\varepsilon^{-1}\|P_n - P\|_{\mathcal{F}_\varepsilon} \geq (\varepsilon^{-1} n^{-1})^{1/2} t + \varepsilon^{-d/2} n^{-1/2}) \leq e^{-bt^2/2}.$$

Consequently, we can merge these via a union bound; taking the worst case constant, we have that for $t \geq c\varepsilon^{-d/2} n^{-1/2}$, for $c > 0$ sufficiently large, it holds that

$$\mathbb{P}(\|\hat{T}_s - T_0\|^2_{L^2(P)} \gtrsim \varepsilon^{s/2} + \varepsilon^{-d/2} n^{-1/2} t) \leq e^{-ct^2/2}.$$

Dividing through by $\psi_n(s) := (n/\log(n))^{-\frac{s}{2(d+s)}}$ completes the proof. □

We can also obtain tail bounds under $L^2(P_n')$ at virtually no cost. Indeed, for any $s \in \mathcal{S}$,

$$\|\hat{T}_s - T_0\|^2_{L^2(P_n')} \lesssim \|\hat{T}_s - T_0\|^2_{L^2(P)} + \|P_n' - P\|_{\mathcal{H}_N},$$

where the last term has expectation bounded above by $\log\log(n) n^{-1/2}$ up to a constant factor (indeed, since $T_0 = T_{s_i}$, this is perfectly fine at the cost of adding one more function to the set). By

employing a further union bound, we can state Proposition A.11 as

$$\mathbb{P}\left(\|\hat{T}_s - T_0\|^2_{L^2(P'_n)}\psi_n(s)^{-1} \geq ct\right) \leq 2\exp\left(-\frac{t^2\log(n)}{C}\right), \tag{A.13}$$

for any $s \in \mathcal{S}$, where the constants that appear are slightly different. Indeed, since $\log\log(n)/\sqrt{n} \ll \psi_n(s)$, nothing is lost by incorporating this additional term.

Returning to (A.12), we can take c_0 sufficiently large in both terms, we can employ (A.13) for all the terms in the summand, which results in

$$\mathbb{P}(\mathcal{E}_j) \leq n^{-c_0^2/(8C)}.$$

For the integrated tail, we use a similar argument, appealing to Proposition A.11 directly. Indeed, for $t \geq C\psi_n(s_j)/\psi_n(s_i)$, the following bound holds:

$$\mathbb{P}\left(\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)}\psi_n(s_i)^{-1} \geq t\right) \leq \exp\left(-\frac{t^2\log(n)}{C}\frac{\psi_n(s_i)}{\psi_n(s_j)}\right). \tag{A.14}$$

Choosing $t_j = c_1\sqrt{\psi_n(s_j)/\psi_n(s_i)}$, the tail can be upper bounded as

$$\begin{aligned}\int_{t_j}^{\infty} \exp\left(-\frac{t^2\log(n)}{C}\frac{\psi_n(s_i)}{\psi_n(s_j)}\right) \mathrm{d}t &\leq \left(\frac{\psi_n(s_j)C}{\psi_n(s_i)\log(n)}\right)\sqrt{\frac{\psi_n(s_i)}{\psi_n(s_j)c_1^2}}\exp\left(-\frac{c_1\log(n)}{C}\right) \\ &= \sqrt{\frac{\psi_n(s_j)}{\psi_n(s_i)}}\frac{C}{c_1\log(n)}\exp\left(-\frac{c_1\log(n)}{C}\right).\end{aligned}$$

Merging everything together, we obtain rather crudely that

$$\begin{aligned}\sum_{j=1}^{i-1} \mathbb{E}[\|\hat{T}_{s_j} - T_0\|^2_{L^2(P)} \psi_n(s_i)^{-1} \mathbf{1}(\mathcal{E}_j)] &\leq \sum_{j=1}^{i-1} \left(t_j n^{-c_0^2/(8C)} + \sqrt{\frac{\psi_n(s_j)}{\psi_n(s_i)}} \frac{C}{c_1 \log(n)} \exp\left(-\frac{c_1 \log(n)}{C}\right)\right) \\ &\leq \sum_{j=1}^{i-1} \frac{1}{\log n} \\ &\asymp 1\,,\end{aligned}$$

since there exist $N \asymp \log(n)$ terms. This completes the proof.

B | Supplement to Chapter 3

B.1 Reminders on semi-discrete entropic optimal transport

We recall in this section some known results on entropic optimal transport that will be needed later. Let $\mu, \nu \in \mathcal{P}(\Omega)$, where $\Omega \subset B(0;R)$ is a compact set.

Lemma B.1 (Genevay et al., 2019). *The entropic potential $(\varphi_\varepsilon^{\mu\to\nu}, \psi_\varepsilon^{\mu\to\nu})$ have a bounded amplitude, in the sense that*

$$\max_{x\in\Omega} \varphi_\varepsilon^{\mu\to\nu} - \min_{x\in\Omega} \varphi_\varepsilon^{\mu\to\nu} \leq cR \tag{B.1}$$

for some absolute constant c, and similarly for $\psi_\varepsilon^{\mu\to\nu}$.

Assume now that $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ is a discrete measure. In this situation, only the values of the dual potential $\psi_\varepsilon^{\mu\to\nu}$ on the points $y_1, \dots, y_J$ are relevant. We therefore consider $\psi_\varepsilon^{\mu\to\nu}$ as a vector in $\mathbb{R}^J$. The potentials $\varphi_\varepsilon^{\mu\to\nu}$ and $\psi_\varepsilon^{\mu\to\nu}$ are dual of one another, in the sense of the ε-Legendre transform. Given a finite measure ρ, the ε-Legendre transform of a function h with respect to ρ is given by

$$\Phi_\varepsilon^\rho(h)(x) = \varepsilon \log \int e^{(\langle x,y\rangle - h(x))/\varepsilon} \, \mathrm{d}\rho(x). \tag{B.2}$$

Modifying (1.27) and (1.28) for entropic Brenier maps tell us that $\varphi_\varepsilon^{\mu\to\nu} = \Phi_\varepsilon^\nu(\psi_\varepsilon^{\mu\to\nu})$ and vice-versa. In the semi-discrete setting, it is also convenient to introduce the ε-Legendre transform with

respect to the counting measure σ on $\{y_1, \dots, y_J\}$. For a vector $\psi \in \mathbb{R}^J$, we have

$$\Phi_\varepsilon(\psi)(x) \coloneqq \Phi_\varepsilon^\sigma(\psi)(x) = \varepsilon \log \sum e^{(\langle x, y_j \rangle - \psi(y_j))/\varepsilon}. \tag{B.3}$$

The Φ_ε transform and the Φ_ε^ν transform are linked through the relation

$$\Phi_\varepsilon^\nu(\psi) = \Phi_\varepsilon(\tilde{\psi}) \qquad \text{where} \qquad \tilde{\psi}(y_j) = \psi(y_j) - \varepsilon \log \nu_j\,, \tag{B.4}$$

where we call $\tilde{\psi}$ a *shifted* potential. With this notation, the optimality condition on the potentials can be rephrased. Let

$$F_\varepsilon^{\mu \to \nu} : \psi \in \mathbb{R}^J \to \int \Phi_\varepsilon(\psi)\, \mathrm{d}\mu + \int \psi\, \mathrm{d}\nu\,. \tag{B.5}$$

Then, the function $F_\varepsilon^{\mu \to \nu}$ is minimized at $\tilde{\psi}_\varepsilon^{\mu \to \nu}$. For $\psi \in \mathbb{R}^J$ and $x \in \mathbb{R}^d$, we introduce the probability measure supported on $\{y_1, \dots, y_J\}$ given by

$$\forall i \in [J], \quad \pi_\varepsilon^x[\psi](y_i) = \frac{e^{(\langle x, y_i \rangle - \psi(y_i))/\varepsilon}}{\sum_{j=1}^J e^{(\langle x, y_j \rangle - \psi(y_j))/\varepsilon}} = e^{(\langle x, y_i \rangle - \Phi_\varepsilon(\psi)(x) - \psi(y_i))/\varepsilon}. \tag{B.6}$$

A computation gives $\nabla F_\varepsilon^{\mu \to \nu}(\psi) = \int \pi_\varepsilon^x[\psi]\, \mathrm{d}\mu(x) - \nu$, so that at optimality, we have

$$\int \pi_\varepsilon^x[\tilde{\psi}_\varepsilon^{\mu \to \nu}]\, \mathrm{d}\mu(x) = \nu. \tag{B.7}$$

In this case, $\pi_\varepsilon^x = \pi_\varepsilon^x[\tilde{\psi}_\varepsilon^{\mu \to \nu}]$ is the conditional distribution of the second marginal of π_ε given that the first is equal to x. More generally, for any potential ψ, the first order condition implies that ψ is equal to $\tilde{\psi}_\varepsilon^{\mu \to \nu_\psi}$, the optimal dual potential between μ an $\nu_\psi = \int \pi_\varepsilon^x[\psi]\, \mathrm{d}\mu(x)$.

B.2 Bound on the approximation error

Proof of Theorem 3.5. Let $i, j \in [J]$. We define the jth slack at $x \in L_i$ by

$$\frac{1}{2}\Delta_{ij}(x) = -\langle x, y_j\rangle + \varphi_0(x) + \psi_0(y_j). \tag{B.8}$$

As φ_0 is the Legendre transform of ψ_0, we have $\Delta_{ij}(x) \geq 0$. If the cells L_i and L_j have a nonempty intersection, the set $H_{ij}(t) = \{x \in L_i : \Delta_{ij}(x) = t\}$ represents the trace on L_i of the hyperplane spanned by the boundary between L_i and L_j, shifted by t. It is stated by Altschuler et al. (2022) that for every nonnegative measurable function $f : \mathbb{R} \to \mathbb{R}_+$,

$$\int_{L_i} f(\Delta_{ij}(x))p(x)\,\mathrm{d}x = \frac{1}{2\|y_i - y_j\|}\int_0^\infty f(t)h_{ij}(t)\,\mathrm{d}t, \tag{B.9}$$

where $h_{ij}(t) = \int_{H_{ij}(t)} p(x)\,\mathrm{d}\mathcal{H}_{d-1}(x)$ and $\mathcal{H}_{d-1}$ is the $(d-1)$-dimensional Hausdorff measure. In particular, $w_{ij} = h_{ij}(0)$ is the (weighted) surface of the boundary between the i^{th} and j^{th} Laguerre cells (should it exist). Given $x \in L_i$, let $s(x) = \min_{j\neq i} \frac{1}{2}\Delta_{ij}(x)$. When the point x is sufficiently inside its Laguerre cell, the conditional probability π_ε^x becomes extremely concentrated around the point y_i, as the next lemma shows. Note that $\pi_0^x = \delta_{y_i}$ when $x \in L_i$.

Lemma B.2. *Let $x \in L_i$. For ε small enough, it holds that for every $j \in [J]$, $|\pi_\varepsilon^x(y_j) - \pi_0^x(y_j)| \leq ce^{-s(x)/\varepsilon}$, where c depends on J, the distances $\|y_i - y_j\|$ and on the quantities w_{ij}.*

Such a result was already stated in Delalande (2022, Corollary 2.2), although while requiring that the source measure P has a Hölder continuous density. Only assumption **(S1)** is needed here.

Proof. According to Altschuler et al. (2022, Proposition 4.6), for ε small enough,

$$\varepsilon^{-1}\|\tilde{\psi}_\varepsilon - \psi_0\|_\infty \leq C, \tag{B.10}$$

where $\tilde{\psi}_\varepsilon$ is the shifted version of ψ_ε (see (B.3)) and C depends on the distances $\|y_i - y_j\|$ and on the w_{ij}s. Following Delalande (2022, Proof of Corollary 2.2) and (B.6), we have for $j \neq i$

$$|\pi_\varepsilon^x(y_j) - \pi_0^x(y_j)| = \pi_\varepsilon^x(y_j) = \frac{e^{(\langle x,y_j\rangle - \tilde{\psi}_\varepsilon(y_j))/\varepsilon}}{\sum_{j'=1}^J e^{(\langle x,y_{j'}\rangle - \tilde{\psi}_\varepsilon(y_{j'}))/\varepsilon}} \leq e^{2C} \frac{e^{(\langle x,y_j\rangle - \psi_0(y_j))/\varepsilon}}{\sum_{j'=1}^J e^{(\langle x,y_{j'}\rangle - \psi_0(y_{j'}))/\varepsilon}} \leq e^{2C} e^{-s(x)/\varepsilon}.$$

A similar computation yields that $|\pi_\varepsilon^x(y_i) - \pi_0^x(y_i)| = |\pi_\varepsilon^x(y_i) - 1| \leq J e^{2C} e^{-s(x)/\varepsilon}$. □

We can bound for any $x \in L_i$,

$$\|T_\varepsilon(x) - T_0(x)\| = \| \sum_{j=1}^J y_j (\pi_\varepsilon^x(y_j) - \pi_0^x(y_j))\| \leq c \sum_{j=1}^J \|y_j\| e^{-s(x)/\varepsilon}. \tag{B.11}$$

Therefore, letting C' denote a constant, which may depend on J, whose value may change from line to line, we obtain

$$\|T_\varepsilon - T_0\|_{L_2(P)}^2 = \sum_{i=1}^J \int_{L_i} \|T_\varepsilon(x) - T_0(x)\|^2 \, \mathrm{d}P(x) \leq C' \sum_{i=1}^J \int_{L_i} \sum_{j=1}^J e^{-2s(x)/\varepsilon} \, \mathrm{d}P(x) \tag{B.12}$$

$$\leq C' \sum_{i \neq j} \int_{L_i} e^{-\Delta_{ij}(x)/\varepsilon} \, \mathrm{d}P(x) \leq C' \sum_{i \neq j} \frac{1}{2\|y_i - y_j\|} \int_0^\infty e^{-t/\varepsilon} h_{ij}(t) \, \mathrm{d}t \,, \tag{B.13}$$

where in the second equality, we used the definition of $s(x)$. Assumption **(S1)** ensures that the functions h_{ij}s are bounded, which implies that the right-hand side in (B.13) is of order ε. □

B.3 Stability of entropic transport plans

Proof of Proposition 3.10. Note that we may assume without loss of generality that $\nu \ll \nu'$ and that $\mathrm{KL}(\nu\|\nu') < \infty$, for otherwise the bound is vacuous. For notational convenience, we omit the dependence on ε in the subscripts. Write $\pi^{\mu,\nu} = \gamma^{\mu,\nu}(x,y)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$ for the entropic optimal plan between μ and ν, where $\gamma^{\mu,\nu} = \exp\left(\frac{1}{\varepsilon}(\langle x,y\rangle - \varphi^{\mu\to\nu}(x) - \psi^{\mu\to\nu}(y))\right)$, and analogously define $\gamma^{\mu',\nu'} = \exp\left(\frac{1}{\varepsilon}(\langle x,y\rangle - \varphi^{\mu'\to\nu'}(x) - \psi^{\mu'\to\nu'}(y))\right)$.

Consider the measure $\gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y)$. The first-order optimality condition for the pair of potentials $(\varphi^{\mu'\to\nu'}, \psi^{\mu'\to\nu'})$ implies that

$$\int \gamma^{\mu',\nu'}(y)\,\mathrm{d}\nu'(y) = 1 \quad \forall x \in \Omega\,, \tag{B.14}$$

so that $\gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\nu'(y)$ is a probability measure. Let us write $\mathrm{d}\pi^x(y) = \gamma^{\mu,\nu}(x,y)\,\mathrm{d}\nu(y)$ and $\mathrm{d}\rho^x(y) = \gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\nu'(y)$.

We make the following observations: first, $T^{\mu\to\nu}(x) = \int y\,\mathrm{d}\pi^x(y)$ and $T^{\mu'\to\nu'}(x) = \int y\,\mathrm{d}\rho^x(y)$. Second, the support of ρ^x lies inside $B(0;R)$; since any Lipschitz function f on $B(0;R)$ satisfies $\sup_x f(x) - \inf_x f(x) \le 2R$, Hoeffding's lemma (see Boucheron et al., 2013, Lemma 2.2) implies that if f is Lipschitz and $\int f\,\mathrm{d}\rho^x = 0$, then

$$\int e^{tf}\,\mathrm{d}\rho^x \le e^{2R^2t^2} \quad \forall t \in \mathbb{R}\,.$$

This implies (Bobkov and Götze, 1999, Theorem 3.1) that

$$W_1(\pi^x,\rho^x)^2 \le 8R^2\mathrm{KL}(\pi^x\|\rho^x)\,. \tag{B.15}$$

Third, Jensen's inequality implies that for any coupling γ between π^x and ρ^x,

$$\int \|y-y'\|\,\mathrm{d}\gamma(y,y') \ge \left\|\int (y-y')\,\mathrm{d}\gamma(y,y')\right\| = \|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|\,, \tag{B.16}$$

so that in particular, $\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\| \le W_1(\pi^x,\rho^x)$. Combining these facts, we obtain

$$\frac{1}{8R^2}\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|^2 \le \mathrm{KL}(\pi^x\|\rho^x) = \int \log\left(\frac{\gamma^{\mu,\nu}}{\gamma^{\mu',\nu'}}(x,y)\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\right)\gamma^{\mu,\nu}(x,y)\,\mathrm{d}\nu(y)\,. \tag{B.17}$$

Integrating both sides of this equation with respect to μ yields

$$\frac{1}{8R^2}\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|^2_{L^2(\mu)} \le \int \log\left(\frac{\gamma^{\mu,\nu}}{\gamma^{\mu',\nu'}}(x,y)\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\right)\mathrm{d}\pi^{\mu,\nu}(x,y)\,. \tag{B.18}$$

Expanding the definition of $\gamma^{\mu,\nu}$ and $\gamma^{\mu',\nu'}$ and using that

$$\int \log\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\,\mathrm{d}\pi^{\mu,\nu}(x,y) = \int \log\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\,\mathrm{d}\nu(y) = \mathrm{KL}(\nu\|\nu')$$

yields the claim. □

We now record two corollaries of this bound, which apply when either the source or the target measures of the entropic maps agree.

Corollary B.3. *For any μ, ν, ν' supported in $B(0;R)$,*

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)} \le \varepsilon^{-1}\int(\psi_\varepsilon^{\mu\to\nu'} - \psi_\varepsilon^{\mu\to\nu})\,\mathrm{d}(\nu-\nu') + \mathrm{KL}(\nu\|\nu')\,. \tag{B.19}$$

Proof. We apply Proposition 3.10 with $\mu = \mu'$, which yields (once again omitting the dependency in ε)

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)} \le \varepsilon^{-1}\left(\int(\varphi^{\mu\to\nu'} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu + \int(\psi^{\mu\to\nu'} - \psi^{\mu\to\nu})\,\mathrm{d}\nu\right) + \mathrm{KL}(\nu\|\nu')\,. \tag{B.20}$$

By definition, $(\varphi^{\mu\to\nu'}, \psi^{\mu\to\nu'})$ minimizes the expression

$$\int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu' + \varepsilon\iint e^{(\langle x,y\rangle - \varphi(x) - \psi(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y) - \varepsilon\,,$$

so, recalling that $\iint e^{(\langle x,y\rangle-\varphi^{\mu\to\nu'}(x)-\psi^{\mu\to\nu'}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y)=1$, we have in particular

$$\begin{aligned}\int\varphi^{\mu\to\nu'}\,\mathrm{d}\mu+\int\psi^{\mu\to\nu'}\,\mathrm{d}\nu' &\le \int\varphi^{\mu\to\nu}\,\mathrm{d}\mu+\int\psi^{\mu\to\nu}\,\mathrm{d}\nu'\\ &\quad+\varepsilon\iint e^{(\langle x,y\rangle-\varphi^{\mu\to\nu}(x)-\psi^{\mu\to\nu}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y)-\varepsilon\\ &=\int\varphi^{\mu\to\nu}\,\mathrm{d}\mu+\int\psi^{\mu\to\nu}\,\mathrm{d}\nu',\end{aligned}$$

where we have used that the first-order optimality condition for $(\varphi^{\mu\to\nu},\psi^{\mu\to\nu})$ implies that $\iint e^{(\langle x,y\rangle-\varphi^{\mu\to\nu}(x)-\psi^{\mu\to\nu}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y)=1$ as well (recall (1.27)). This implies

$$\int(\varphi^{\mu\to\nu'}-\varphi^{\mu\to\nu})\,\mathrm{d}\mu\le-\int(\psi^{\mu\to\nu'}-\psi^{\mu\to\nu})\,\mathrm{d}\nu'\,. \tag{B.21}$$

Applying this inequality to (B.20) yields

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu}-T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)}\le\varepsilon^{-1}\int(\psi^{\mu\to\nu'}-\psi^{\mu\to\nu})\,\mathrm{d}(\nu-\nu')+\mathrm{KL}(\nu\|\nu').$$

□

Corollary B.4. *For any μ,μ',ν supported in $B(0;R)$,*

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu}-T_\varepsilon^{\mu'\to\nu}\|^2_{L^2(\mu)}\le\varepsilon^{-1}\int(\varphi_\varepsilon^{\mu'\to\nu}-\varphi_\varepsilon^{\mu\to\nu})\,\mathrm{d}(\mu-\mu')\,. \tag{B.22}$$

Proof. We apply Proposition 3.10 with $\nu=\nu'$, yielding (dropping the dependency on ε)

$$\frac{1}{8R^2}\|T^{\mu\to\nu}-T^{\mu'\to\nu}\|^2_{L^2(\mu)}\le\varepsilon^{-1}\left(\int(\varphi^{\mu'\to\nu}-\varphi^{\mu\to\nu})\,\mathrm{d}\mu+\int(\psi^{\mu'\to\nu}-\psi^{\mu\to\nu})\,\mathrm{d}\nu\right). \tag{B.23}$$

An argument analogous to the one used in the proof of Corollary B.3 gives the inequality

$$\int\varphi^{\mu'\to\nu}\,\mathrm{d}\mu'+\int\psi^{\mu'\to\nu}\,\mathrm{d}\nu\le\int\varphi^{\mu\to\nu}\,\mathrm{d}\mu'+\int\psi^{\mu\to\nu}\,\mathrm{d}\nu\,, \tag{B.24}$$

or, equivalently,

$$\int (\psi^{\mu'\to\nu} - \psi^{\mu\to\nu})\,\mathrm{d}\nu \le -\int (\varphi^{\mu'\to\nu} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu', \tag{B.25}$$

and combining this inequality with (B.23) proves the claim. □

B.4 Strong convexity of the entropic semi-dual problem

Proposition B.5 (Strong convexity of $F_\varepsilon^{\mu\to\nu}$). *Let $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ be a measure supported on $\{y_1, \dots, y_J\} \subseteq B(0;R)$ and let μ supported on a compact convex set $\Omega \subseteq B(0;R)$ with a density p satisfying $p_{\min} \le p \le p_{\max}$ for some $p_{\max} \ge p_{\min} > 0$. For $\psi \in \mathbb{R}^J$, define $\nu_\psi = \int \pi_\varepsilon^x[\psi]\,\mathrm{d}\mu(x)$ and assume that $\nu_\psi \ge \lambda\nu$ for some $0 < \lambda \le 1$. Then, we have for $\varepsilon \in (0,1)$*

$$F_\varepsilon^{\mu\to\nu}(\psi) - \min_{\psi} F_\varepsilon^{\mu\to\nu} \ge C\lambda \cdot \mathrm{Var}_\nu(\psi - \psi_\varepsilon^{\mu\to\nu}), \tag{B.26}$$

where $C = \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1} \frac{p_{\min}}{p_{\max}}$.

Proof. As μ and ε are fixed, we will simply write ψ_ν instead of $\psi_\varepsilon^{\mu\to\nu}$, and write similarly $F_\nu = F_\varepsilon^{\mu\to\nu}$. Recall the definition (B.3) of the shifted potential $\tilde\psi_\nu(y_j) = \psi_\nu(y_j) - \varepsilon\log\nu_j$. According to Delalande (2022, Theorem 3.2), the functional F_ν is minimized at the vector $\tilde\psi_\nu$, with

$$\forall v \in \mathbb{R}^J, \quad \mathrm{Var}_\nu(v) \le \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right) v^\top \nabla^2 F_\nu(\tilde\psi_\nu) v. \tag{B.27}$$

For $t \in [0,1]$, let $\psi_t = \tilde\psi_\nu + t(\psi - \tilde\psi_\nu)$ and let $\nu_t = \int \pi_\varepsilon^x[\psi_t]\,\mathrm{d}\mu(x)$. The potential ψ_t is the (shifted) entropic Brenier potential between μ and ν_t, so that it minimizes the functional F_{ν_t} (see Section B.1). Also, note that $\nabla^2 F_\nu$ does not depend on ν, so that

$$v^\top \nabla^2 F_\nu(\psi_t) v = v^\top \nabla^2 F_{\nu_t}(\psi_t) v \ge \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1} \mathrm{Var}_{\nu_t}(v). \tag{B.28}$$

Let $v = \psi - \psi_\varepsilon^{\mu\to\nu}$. A Taylor expansion of F_ν gives

$$F_\nu(\psi) - F_\nu(\tilde{\psi}_\nu) = \int_0^1 v^\top \nabla^2 F_\nu(\psi_t) v \, \mathrm{d}t \geq \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1} \int_0^1 \mathrm{Var}_{\nu_t}(v)\, \mathrm{d}t. \tag{B.29}$$

Lemma B.6. *Write $\nu_t = \sum_{j=1}^J \nu_{t,j}\delta_{y_j}$. Then, for all $t \in [0,1]$ and $j \in [J]$, we have $\nu_{t,j} \geq \frac{p_{\min}}{p_{\max}} \nu_{0,j}^{1-t}\nu_{1,j}^t$.*

This lemma is enough to conclude the proof. Indeed, $\nu_1 = \nu_\psi \geq \lambda\nu$, so that it implies that $\mathrm{Var}_{\nu_t}(v) \geq \frac{p_{\min}}{p_{\max}}\lambda \mathrm{Var}_\nu(v)$. □

Proof of Lemma B.6. According to Delalande (2022, Proof of Proposition 4.1),

$$\Phi_\varepsilon(\psi_t)(tx + (1-t)y) \leq t\Phi_\varepsilon(\tilde{\psi}_\varepsilon^{\mu\to\nu})(x) + (1-t)\Phi_\varepsilon(\psi)(y). \tag{B.30}$$

Therefore, if we let $h_t(x) = e^{(\langle x, y_j\rangle - \psi_t(y_j) - \Phi_\varepsilon(\psi_t)(x))/\varepsilon}$, then we have $h_t(tx+(1-t)y) \geq h_0(x)^t h_1(y)^{1-t}$. By the Prékopa-Leindler inequality,

$$\nu_{t,j} = \int h_t(x)\,\mathrm{d}\mu(x) \geq p_{\min}\int_{\mathcal{X}} h_t(x)\,\mathrm{d}x \geq p_{\min}\left(\int_{\mathcal{X}} h_0(x)\,\mathrm{d}x\right)^t \left(\int_{\mathcal{X}} h_1(x)\,\mathrm{d}x\right)^{1-t} \geq \frac{p_{\min}}{p_{\max}}\nu_{0,j}^t\nu_{1,j}^{1-t}.$$

□

Proof of Proposition 3.12. As in the previous proof, we drop the ε and μ dependency in our notation. Write $\nu_k = \sum_{j=1}^J \nu_{k,j}\delta_{y_j}$ for $k = 0, 1$, and define as before the shifted potentials $\tilde{\psi}_{\nu_k}(y_j) = \psi_{\nu_1}(y_j) - \varepsilon\log\nu_{k,j}$. Let $\theta > 0$ be a parameter to fix. According to Proposition B.5, Lemma B.15, and using the inequality $F_{\nu_1}(\tilde{\psi}_{\nu_1}) \leq F_{\nu_1}(\tilde{\psi}_{\nu_0})$, we have

$$\begin{aligned} C\lambda\mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) \leq F_{\nu_0}(\tilde{\psi}_{\nu_1}) - F_{\nu_0}(\tilde{\psi}_{\nu_0}) &\leq F_{\nu_0}(\tilde{\psi}_{\nu_1}) - F_{\nu_1}(\tilde{\psi}_{\nu_1}) + F_{\nu_1}(\tilde{\psi}_{\nu_0}) - F_{\nu_0}(\tilde{\psi}_{\nu_0}) \\ &= \int (\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0})(\,\mathrm{d}\nu_0 - \,\mathrm{d}\nu_1) \\ &\leq \frac{\theta}{2}\mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) + \frac{1}{2\theta}\chi^2(\nu_1\|\nu_0). \end{aligned}$$

We pick $\theta = C\lambda$ to conclude that

$$\mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) \leq \frac{1}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0). \tag{B.31}$$

Therefore, using the inequality $|\log(a/b)| \leq |a-b|/\min\{a,b\}$ for $a, b > 0$,

$$\begin{aligned}
\mathrm{Var}_{\nu_0}(\psi_1 - \psi_0) &\leq 2\mathrm{Var}_{\nu_0}(\tilde{\psi}_1 - \tilde{\psi}_0) + 2\sum_{j=1}^{J} \nu_{0,j}\left(\log\left(\frac{\nu_{1,j}}{\nu_{0,j}}\right)\right)^2 \\
&\leq \frac{2}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0) + 2\sum_{j=1}^{J} \nu_{0,j}\left(\frac{\nu_{1,j}-\nu_{0,j}}{\min\{\nu_{0,j},\nu_{1,j}\}}\right)^2 \\
&\leq \frac{2}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0) + \frac{2}{\lambda^2}\sum_{j=1}^{J}\frac{1}{\nu_{0,j}}(\nu_{1,j}-\nu_{0,j})^2 \leq \left(\frac{2}{(C\lambda)^2} + \frac{2}{\lambda^2}\right)\chi^2(\nu_1\|\nu_0).
\end{aligned}$$

□

B.5 Control of the fluctuations in the one-sample case

Lemma B.7 (Sample complexity in the one-sample case)**.** *Assume that P satisfy **(S1)** and that Q satisfy **(S2)**. Then, it holds that $\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|^2_{L^2(P)} \lesssim \varepsilon^{-1}n^{-1}$.*

Proof. To ease notation, we write $T_{\varepsilon,n} = T_\varepsilon^{P\to Q_n}$ and $\psi_{\varepsilon,n} = \psi_\varepsilon^{P\to Q_n}$. As explained in Section 3.2, the stability result Proposition 3.10 implies that

$$\mathbb{E}\|T_{\varepsilon,n} - T_\varepsilon\|^2_{L^2(P)} \leq \frac{8R^2}{\varepsilon}\left(\frac{\mathbb{E}[\mathrm{Var}_Q(\psi_{\varepsilon,n}-\psi_\varepsilon)]}{2} + \frac{\mathbb{E}[\chi^2(Q_n\|Q)]}{2}\right) + 8R^2\mathbb{E}[\chi^2(Q_n\|Q)]\,. \tag{B.32}$$

Write $Q = \sum_{j=1}^J q_j\delta_{y_j}$ and $Q_n = \sum_{j=1}^J \hat{q}_j\delta_{y_j}$, and introduce the event $E = \{\forall j \in [J],\ \hat{q}_j \geq q_j/2\}$. If E is satisfied, we have $Q_n \geq Q/2$, so that Proposition 3.12 yields

$$\mathrm{Var}_Q(\psi_{\varepsilon,n} - \psi_\varepsilon) \leq C\chi^2(Q_n\|Q). \tag{B.33}$$

If E is not satisfied, we use the fact that the entropic potentials have a bounded amplitude (see Lemma B.1), to obtain that

$$\mathrm{Var}_Q(\psi_{\varepsilon,n} - \psi_\varepsilon) \leq C'. \tag{B.34}$$

Lemma B.8. *Let E be the event that $Q_n \geq Q/2$. Then $\mathbb{P}(E^c) \leq Je^{-cq_{\min}n}$ for some $c > 0$.*

Proof. By Vershynin (2018, Exercise 2.3.2), we have $\mathbb{P}(E^c) \leq \sum_{j=1}^J \mathbb{P}(\hat{q}_j < q_j/2) \leq Je^{-cq_{\min}n}$ for some $c > 0$. □

We obtain

$$\mathbb{E}\|\hat{T}_{\varepsilon,n} - T_\varepsilon\|^2_{L^2(P)} \lesssim \frac{R^2}{\varepsilon}\mathbb{E}[\chi^2(Q_n\|Q)] + \frac{R^2}{\varepsilon}Je^{-cq_{\min}n} \lesssim \varepsilon^{-1}n^{-1} \tag{B.35}$$

by Lemma B.16. □

B.6 Control of the fluctuations in the two-sample case

The goal of this section is to prove Theorem 3.8. We will actually prove a more general result, and show that *for any discrete measure* $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ supported on $\{y_1, \dots, y_J\}$ with $\nu_j \geq \nu_{\min} > 0$ for all $j \in [J]$, we have for $\log(1/\varepsilon) \lesssim n/\log(n)$,

$$\mathbb{E}\|T_\varepsilon^{P_n \to \nu} - T_\varepsilon^{P \to \nu}\|^2_{L_2(P)} \lesssim \varepsilon^{-1}n^{-1}. \tag{B.36}$$

Theorem 3.8 follows from (B.36) by conditioning on Q_n. Let E be the event that $Q_n \geq Q/2$. Then, by Lemma B.8,

$$\begin{aligned}\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon^{P \to Q_n}\|^2_{L_2(P)} &\leq \mathbb{E}\left[\mathbb{E}[\|\hat{T}_\varepsilon - T_\varepsilon^{P \to Q_n}\|^2_{L_2(P)}|Q_n]\mathbb{1}\{E\}\right] + R^2\mathbb{P}(E^c) \\ &\leq C\varepsilon^{-1}n^{-1} + R^2Je^{-cq_{\min}n} \lesssim \varepsilon^{-1}n^{-1}.\end{aligned}$$

We obtain Theorem 3.8 by combining this bound with Lemma B.7.

To prove (B.36), we first use Corollary B.4 which yields

$$\begin{aligned}\mathbb{E}\|T_\varepsilon^{P_n\to\nu} - T_\varepsilon^{P\to\nu}\|_{L_2(P)}^2 &\le 8R^2\varepsilon^{-1}\mathbb{E}\int (\varphi_\varepsilon^{P_n\to\nu} - \varphi_\varepsilon^{P\to\nu})\,\mathrm{d}(P_n - P)\\ &= 8R^2\varepsilon^{-1}\mathbb{E}\int (\Phi_\varepsilon(\tilde{\psi}_\varepsilon^{P_n\to\nu}) - \Phi_\varepsilon(\tilde{\psi}_\varepsilon^{P\to\nu}))\,\mathrm{d}(P_n - P),\end{aligned} \tag{B.37}$$

where we recall that for a potential ψ, the shifted potential $\tilde{\psi}$ is given by $\tilde{\psi}_j = \psi_j - \varepsilon\log\nu_j$. The remainder of the proof consists in bounding this integral by using localization arguments and standard bounds on suprema of empirical processes. Our first goal is to show that the potential $\psi_\varepsilon^{P_n\to\nu}$ is close to to the potential $\psi_\varepsilon^{P\to\nu}$ for the ∞-norm. It will be convenient to work with the "L_∞-variance"

$$\mathrm{Var}_\infty(\psi) = \inf_{c\in\mathbb{R}}\max_{j\in[J]}|\psi(y_j) - c|^2 = \left(\frac{\max\psi - \min\psi}{2}\right)^2. \tag{B.38}$$

As the measure ν is lower bounded, it holds that

$$\mathrm{Var}_\nu(\psi) \ge \nu_{\min}\mathrm{Var}_\infty(\psi). \tag{B.39}$$

Lemma B.9 (Supremum of ε-Legendre transforms)**.** *Let ψ_0 be a fixed potential and let $\tau > 0$. Then, for all $j \in [J]$,*

$$\mathbb{E}\left[\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\le\tau^2}\left|\int(\pi_\varepsilon^x(\psi)_j - \pi_\varepsilon^x(\psi_0)_j)\,\mathrm{d}(P - P_n)(x)\right|\right] \le C\sqrt{\frac{J\max\{\log(\tau/\varepsilon),1\}}{n}} \tag{B.40}$$

$$\mathbb{E}\left[\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\le\tau^2}\left|\int(\Phi_\varepsilon(\psi)(x) - \Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P - P_n)(x)\right|\right] \le C\tau\sqrt{\frac{J}{n}} \tag{B.41}$$

for some absolute constant C.

Proof. For a metric space (A, d) and $u > 0$, we let $N(u, A, d)$ be the covering number of A at scale u, that is the smallest number of balls of radius u needed to cover A. Let B be the L_∞-ball of radius τ in $\mathbb{R}^J$, centered at ψ_0, and let $\|\cdot\|_\infty$ denote the ∞-norm. For $0 < u \le \tau$, we have $\log N(u, B, \|\cdot\|_\infty) \le J\log(\tau/u)$.

We start with the second inequality. Note that $\psi \mapsto \Phi_\varepsilon(\psi)$ is 1-Lipschitz continuous, and that the functional Φ_ε satisfies $\Phi_\varepsilon(\psi+c) = \Phi_\varepsilon(\psi)+c$ for all $c \in \mathbb{R}$. Then the set $\{\psi : \mathrm{Var}_\infty(\psi-\psi_0) \le \tau^2\}$ is equal to the set $\{\psi+c : \psi \in B,\ c \in \mathbb{R}\}$. As $\int c \,\mathrm{d}(P-P_n) = 0$, we can therefore restrict the supremum to vectors $\psi \in B$. Furthermore, an envelope function of the class $\{\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\psi_0) : \psi \in B\}$ is the constant function equal to τ. Therefore, by Lemma B.17, we obtain

$$\mathbb{E}\left[\sup_{\|\psi-\psi_0\|_\infty \le \tau} \left| \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\psi_0))(\,\mathrm{d}P - \,\mathrm{d}P_n) \right| \right] \le \frac{c_0}{\sqrt{n}} \int_0^{c_1\tau} \sqrt{J \log 2N(u, \{\Phi_\varepsilon(\psi) : \psi \in B\}, \|\cdot\|_\infty)}\,\mathrm{d}u \le \sqrt{\frac{c_3 J \tau}{n}}.$$

We repeat the same argument for the first inequality. The functional π_ε^x is invariant by translation: $\pi_\varepsilon^x(\psi + c) = \pi_\varepsilon^x(\psi)$ for all $c \in \mathbb{R}$. This implies that

$$\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\le\tau^2} \left| \int (\Phi_\varepsilon(\psi)(x) - \Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P-P_n)(x) \right| = \sup_{\|\psi-\psi_0\|_\infty\le\tau} \left| \int (\Phi_\varepsilon(\psi)(x) - \Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P-P_n)(x) \right|.$$

As the function $\psi \mapsto \pi_\varepsilon^x(\psi)_j$ is ε^{-1}-Lipschitz continuous for every $x \in \mathbb{R}^d$, we have for $0 < u \le \tau/\varepsilon$,

$$\log N(u, \{x \mapsto \pi_\varepsilon^x(\psi)_j : \psi \in B\}, \|\cdot\|_\infty) \le J \log(\tau/(u\varepsilon)).$$

Remarking furthermore that $0 \le \pi_\varepsilon^x(\psi)_j \le 1$ (so that the class of functions $\{x \mapsto \pi_\varepsilon^x(\psi)_j : \psi \in B\}$ admits the constant function 1 as an envelope function), we obtain the following control using

Lemma B.17:

$$\begin{aligned}\mathbb{E}\left[\sup_{\|\psi-\psi_0\|_\infty\le\tau}\left|\int(\pi_\varepsilon^x(\psi)_j-\pi_\varepsilon^x(\psi_0)_j)(\,\mathrm{d}P-\,\mathrm{d}P_n)(x)\right|\right]&\\ \le\frac{c_0}{\sqrt{n}}\int_0^{c_1}\sqrt{J\log 2N(u,\{x\mapsto\pi_\varepsilon^x(\psi)_j:\ \psi\in B\},\|\cdot\|_\infty)}\,\mathrm{d}u&\\ \le\sqrt{\frac{c_2J\max\{\log(\tau/\varepsilon),1\}}{n}},&\end{aligned}$$

where c_0, c_1 and c_2 are absolute constants, and the last line follows from arguing whether $c_1<\tau/\varepsilon$ or not. □

Proposition B.10. *Assume that P satisfies **(S1)** and let $\nu=\sum_{j=1}^J\nu_j\delta_{y_j}$ be a measure supported on $\{y_1,\dots,y_J\}\subset B(0;R)$, with $\nu_j\ge q_{\min}$ for all $j\in[J]$. Then, for all $0<\varepsilon\le 1$ with $\log(1/\varepsilon)\lesssim n/\log(n)$, it holds that*

$$\mathbb{E}\mathrm{Var}_\infty(\tilde\psi_\varepsilon^{P_n\to\nu}-\tilde\psi_\varepsilon^{P\to\nu})\lesssim n^{-1}. \tag{B.42}$$

Proof. To alleviate notation, we will write $\psi_n=\psi_\varepsilon^{P_n\to\nu}$ and $\psi_0=\psi_\varepsilon^{P\to\nu}$. Similarly, we write $F_n=F_\varepsilon^{P_n\to\nu}$ and $F_0=F_\varepsilon^{P\to\nu}$. Let $\nu_n=\int\pi_\varepsilon^x(\psi_\varepsilon^{P_n\to\nu})\,\mathrm{d}P(x)$. Under the event $E=\{\nu_n\ge\nu/2\}$, we have according to Proposition B.5 and the fact that $\tilde\psi_n$ minimizes F_n,

$$\begin{aligned}C\nu_{\min}\mathrm{Var}_\infty(\tilde\psi_n-\tilde\psi_0)&\le C\mathrm{Var}_\nu(\tilde\psi_n-\tilde\psi_0)\\ &\le F_0(\tilde\psi_n)-F_0(\tilde\psi_0)\\ &\le F_0(\tilde\psi_n)-F_n(\tilde\psi_n)+F_n(\tilde\psi_0)-F_0(\tilde\psi_0)\\ &=\int(\Phi_\varepsilon(\tilde\psi_n)-\Phi_\varepsilon(\tilde\psi_0))\,\mathrm{d}(P-P_n)\end{aligned} \tag{B.43}$$

Let us bound $\mathbb{P}(E^c)$. As $\tilde\psi_n$ is the minimum of F_n, we have $\nu=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}P_n(x)$ (see Section B.1). Therefore, we may write $\nu_{n,j}=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}P_n(x)+\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}(P-P_n)(x)=\nu_j+Z_j$, where

$$Z_j=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}(P-P_n)(x)=\int(\pi_\varepsilon^x(\tilde\psi_n)_j-\pi_\varepsilon^x(\tilde\psi_0)_j)\,\mathrm{d}(P-P_n)(x).$$

Note that $\mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0) \lesssim R^2$ (see Lemma B.1), so that by Lemma B.9 and Lemma B.17,

$$\mathbb{P}(E^c) \leq \sum_{j=1}^{J} \mathbb{P}(|Z_j| > v_j/2) \leq J \exp\left(-c \frac{\sqrt{n} q_{\min}}{(\sqrt{J \log(1/\varepsilon)} + \log n}\right) \lesssim n^{-1}, \tag{B.44}$$

under the condition $\log(1/\varepsilon) \lesssim n/\log(n)$.

For $k \geq 0$, let $a_k = 2^k/\sqrt{n}$ and fix some $p > 2$. Let

$$B_a = \sup_{\mathrm{Var}_\infty(\psi - \tilde{\psi}_0) \leq a^2} \left| \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) \right|$$

. Assume that E is satisfied and that $\mathrm{Var}_\infty(\tilde{\psi}_0 - \tilde{\psi}_n) \in [a^2, b^2]$. Then, according to (B.43), it holds that $B_b \geq ca^2$. Using Markov's inequality, Lemma B.9 and Lemma B.17, we bound

$$\begin{aligned}
\mathbb{E}\mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0) &\leq a_0^2 + \sum_{k \geq 0} \mathbb{P}(\mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0) \in [a_k^2, a_{k+1}^2] \text{ and } E) a_{k+1}^2 + C\mathbb{P}(E^c) \\
&\lesssim n^{-1} + \sum_{k \geq 0} \mathbb{P}\left(B_{a_{k+1}} \geq c a_k^2\right) a_{k+1}^2 \lesssim n^{-1} + \sum_{k \geq 0} \frac{\mathbb{E}[B_{a_{k+1}}^p]}{a_k^{2p}} a_{k+1}^2 \\
&\lesssim n^{-1} + \sum_{k \geq 0} \frac{(2^k/n)^p}{(4^k/n)^p} \frac{4^{k+1}}{n} \lesssim n^{-1} + \sum_{k \geq 0} \frac{2^{2k-pk}}{n} \lesssim n^{-1}.
\end{aligned}$$

□

Proposition B.11. *Under the same assumptions than Proposition B.10, it holds that*

$$\mathbb{E}\|T_\varepsilon^{P_n \to \nu} - T_\varepsilon^{P \to \nu}\|_\infty^2 \lesssim \varepsilon^{-1} n^{-1}. \tag{B.45}$$

Proof. Let $Z = \mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0)$. Let once again $a_k = 2^k/\sqrt{n}$ for $k \geq 1$, with $a_0 = 0$. Fix some $p > 2$, with $q = \frac{p}{p-1}$. For $a > 0$, let $D_a = \sup_{\mathrm{Var}_\infty(\psi - \tilde{\psi}_0) \leq a^2} \left| \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) \right|$. By Hölder

inequality and Markov inequality, we obtain,

$$\begin{aligned}
&\mathbb{E}\int (\Phi_\varepsilon(\tilde{\psi}_n) - \Phi_\varepsilon(\tilde{\psi}_0))\,\mathrm{d}(P - P_n) \\
&\quad\le \sum_{k\ge 0} \mathbb{E}\left[\mathbb{1}\{Z \in [a_k^2, a_{k+1}^2]\} \sup_{\mathrm{Var}_\infty(\psi-\tilde{\psi}_0)\le a_{k+1}^2} \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0))\,\mathrm{d}(P - P_n)\right] \\
&\quad\le \mathbb{E}[D_{a_1}] + \sum_{k\ge 1} \left(\mathbb{P}(Z \ge a_k^2)\right)^{1/q} \mathbb{E}\left[D_{a_{k+1}}^p\right]^{1/p} \\
&\quad\lesssim n^{-1} + \sum_{k\ge 0} \left(\frac{\mathbb{E}[Z]}{a_k^2}\right)^{1/q} \frac{2^k}{n} \lesssim \sum_{k\ge 0} \frac{2^{k(1-2/q)}}{n} \lesssim n^{-1},
\end{aligned}$$

where we use Proposition B.10, Lemma B.9 and Lemma B.17 at the last line. (B.37) then gives the conclusion. □

B.7 A lower bound for the performance of the 1NN estimator

In this section, we prove Proposition 3.14. We let P be the Lebesgue measure on $\Omega = [0,1]^d$, and let $y_0 = (0, 1/2, \ldots, 1/2)$ and $y_1 = (1, 1/2, \ldots, 1/2)$. We denote by P_n an empirical measure consisting of i.i.d. samples from P. As in Section B.6, we work in a general setting of a generic discrete target measure ν, which may either be fixed or may be a random measure independent of P_n. We let $\nu = \sum_{j=0,1} \nu_j \delta_{y_j}$ for $\nu_0, \nu_1 \ge \frac{1}{4}$; this latter condition will hold with overwhelming probability if ν is an empirical measure Q_n corresponding to n i.i.d. samples from $Q = \frac{1}{2}\delta_{y_0} + \frac{1}{2}\delta_{y_1}$. Following Manole et al. (2024a), we define the one-nearest neighbor estimator $\hat{T}_{1\mathrm{NN}}$ in this general context by

$$\hat{T}_{1\mathrm{NN}}(x) = \sum_{i=1}^n \sum_{j=0,1} \mathbf{1}_{V_i}(x)(n\hat{\pi}(X_i, y_j))\,,$$

where $\hat{\pi}$ is the empirical optimal coupling between P_n and ν.

We first examine the structure of the Brenier map $T_0 = \nabla\varphi_0$. The considerations in Section 3.1.3 imply that

$$T_0(x) = \begin{cases} y_0 & \langle e_1, x\rangle \le v_0 \\ y_1 & \langle e_1, x\rangle > v_0\,, \end{cases}$$

where e_1 is the first elementary basis vector. The potential φ_0 is not differentiable on the separating hyperplane $\langle e_1, x\rangle = v_0$, which has measure 0 under P, but we may arbitrarily assign points on this hyperplane to y_0.

Similar arguments imply that the empirical transport plan $\hat{\pi}$ between P_n and ν has the following property: there exists a (random) threshold $\tau \in (0, 1)$ such that

$$\hat{\pi}(x, y_0) = \begin{cases} 1 & \langle e_1, x\rangle < \tau \\ 0 & \langle e_1, x\rangle > \tau\,. \end{cases}$$

The set $\langle e_1, x\rangle = \tau$ may not have measure 0 under P_n, and $\hat{\pi}(x, y_0)$ may take values strictly between 0 and 1 on this set.

The following lemma shows that τ is close to v_0 with high probability.

Lemma B.12. *For any $t \ge 0$,*

$$(\tau \ge v_0 + t) \le e^{-2nt^2}\,.$$

Proof. If $\tau \ge v_0 + t$, this implies that $P_n(\{x : \langle e_1, x\rangle < v_0 + t\}) \le v_0$. On the other hand, $nP_n(\{x : \langle e_1, x\rangle < v_0 + t\}$ is a $\mathrm{Bin}(n, v_0 + t)$ random variable. The result then follows from Hoeffding's inequality (Boucheron et al., 2013, Theorem 2.8). □

Let us write H for the halfspace $\{x : \langle e_1, x\rangle \le v_0\}$, and $\hat{H}$ for the halfspace $\{x : \langle e_1, x\rangle \le \tau\}$. Let x be any point in Ω such that $x \in H$. We are interested in the event that there exists an element $X_i \in \{X_1, \ldots, X_n\}$ such that a) $x \in V_i$ and b) $X_i \in \hat{H}^c$. Call this event $\mathcal{E}(x)$. On this event, $\hat{T}_{1\mathrm{NN}}(x) = y_1$ and $T_0(x) = y_0$, so $\|\hat{T}_{1\mathrm{NN}}(x) - T_0(x)\|^2 = 1$.

We therefore obtain

$$\begin{aligned}\mathbb{E}\|\hat{T}_{1\mathrm{NN}} - T_0\|_{L^2(P)}^2 &= \mathbb{E}\int \|\hat{T}_{1\mathrm{NN}}(x) - T_0(x)\|^2 \,\mathrm{d}P(x) \\ &\geq \mathbb{E}\int_H \|\hat{T}_{1\mathrm{NN}}(x) - T_0(x)\|^2 \mathbb{1}\{\mathcal{E}(x)\} \,\mathrm{d}P(x) \\ &\gtrsim \mathbb{E}\int_H \mathbb{1}\{\mathcal{E}(x)\} \,\mathrm{d}P(x) \\ &= \int_H (\mathcal{E}(x)) \,\mathrm{d}P(x)\,,\end{aligned}$$

where the final equality follows from the Fubini–Tonelli theorem.

We now lower bound the probability of $\mathcal{E}(x)$. Let us write $\mathcal{A}_t$ for the event that $\tau < v_0 + t$, for $t > 0$ to be specified, and write H_t for the halfspace $\{x : \langle e_1, x\rangle \leq v_0 + t\}$. Given any $x \in H$, write $\Delta = d(x, H_t^c)$, and let B be a ball of radius 2Δ around x, intersected with Ω.

Denote by $\mathcal{F}(x)$ the event that there are no samples in $V = B \cap H_t$ but there is at least one point in $B \cap H_t^c$. Then $\mathcal{F}(x) \cap \mathcal{A}_t \subseteq \mathcal{E}(x)$, since on $\mathcal{F}(x)$ the nearest neighbor to x must be a sample in H_t^c, and on $\mathcal{A}_t$ we have $H_t^c \subseteq \hat{H}^c$.

Lemma B.13.

$$(\mathcal{F}(x) \cap \mathcal{A}_t) \geq (1 - \mathrm{vol}(V))^n - (1 - \mathrm{vol}(B))^n - e^{-2nt^2}\,.$$

Proof. We first compute $(\mathcal{F}(x))$. The probability that there are no samples in V is $(1 - \mathrm{vol}(V))^n$, and this event may be written as the disjoint union of $\mathcal{F}(x)$ and the event that all of B is empty. The latter event has probability $(1 - \mathrm{vol}(B))^n$. Therefore

$$(1 - \mathrm{vol}(V))^n = (\mathcal{F}(x)) + (1 - \mathrm{vol}(B))^n\,.$$

Since $\left(\mathcal{A}_t^c\right) \leq e^{-2nt^2}$, the claim follows. □

We need the following lemma.

Lemma B.14. *Assume that $\Delta > 0$ and that $d(x, \partial\Omega) \geq 2\Delta$. There exist positive constants $c_{d,0} < 1$ and $c_{d,1}$ such that*

$$\mathrm{vol}(V) \leq c_{d,0}\,\mathrm{vol}(B) \tag{B.46}$$

and

$$\mathrm{vol}(B) \geq c_{d,1}\Delta^d \tag{B.47}$$

Proof. This is immediate from a scaling argument: since $d(x, \partial\Omega) \geq 2\Delta$, the set B is a Euclidean ball of radius 2Δ, and the set V is a Euclidean ball of radius 2Δ minus a spherical dome cut off by a hyperplane at distance Δ from the center. When $\Delta = 1$, it is clear that the claimed inequalities hold, and the general case is obtained by dilation. □

We assume in what follows that $d(x, \partial\Omega) \geq 2\Delta$. The inequalities $(1+x)^n \geq 1 + nx$ and $e^x \leq 1 + x + x^2$, valid for all $x \in [-1, 0]$ and $n \geq 1$, imply that for any $\delta > 0$ there exists a constant $c_{d,\delta} > 0$ such that if $\Delta \leq c_{d,\delta} n^{-1/d}$, then we will have

$$(1 - \mathrm{vol}(V))^n \geq 1 - nc_{d,0}\,\mathrm{vol}(B) \tag{B.48}$$

$$(1 - \mathrm{vol}(B))^n \leq e^{-n\,\mathrm{vol}(B)} \leq 1 - (1-\delta)n\,\mathrm{vol}(B) \tag{B.49}$$

Choosing δ sufficiently small, we obtain the existence of a small $c_{d,3} > 0$ such that if $\Delta \leq c_{d,3} n^{-1/d}$, then

$$(1 - \mathrm{vol}(V))^n - (1 - \mathrm{vol}(B))^n \geq C_d n \Delta^d\,.$$

Define $\Delta_n = c_{d,4} n^{-1/d}$. Putting it all together, consider the set

$$S = \{x \in H \cap \Omega : \Delta_n/2 \leq d(x, H_t^c) \leq \Delta_n, d(x, \partial\Omega) \geq 2\Delta_n\}\,.$$

The above considerations imply that $(\mathcal{E}(x)) \geq C_d n(\Delta_n/2)^d - e^{-2nt^2} \geq C_d' - e^{-2nt^2}$ for all $x \in S$.

Choosing t to be a sufficiently large constant multiple of $n^{-1/2}$, we obtain

$$\int_H (\mathcal{E}(x))\ \mathrm{d}P(x) \geq \int_S (\mathcal{E}(x))\ \mathrm{d}P(x) \gtrsim_d \mathrm{vol}(S)\,.$$

Since $t \asymp n^{-1/2}$, we will have that $t \ll \Delta_n$ for n sufficiently large (as $d \geq 3$). Therefore, for n large enough, the set S contains the set

$$S' = \{x \in \Omega : v_0 - \Delta_n + t \leq \langle e_1, x\rangle \leq v_0 - \Delta_n/2 + t, 2\Delta_n \leq \langle e_j, x\rangle \leq 1 - 2\Delta_n \quad \forall j = 2, \ldots, d\}\,.$$

Since $\mathrm{vol}(S') \gtrsim_d \Delta_n \gtrsim n^{-1/d}$, the claim follows.

B.8 Auxiliary lemmas

Lemma B.15 (Young's inequality)**.** *Let Q_0, Q_1 be probability measures with $Q_1 \ll Q_0$ and let f be a function. Then, for $\theta > 0$,*

$$\int f(\,\mathrm{d}Q_0 - \,\mathrm{d}Q_1) \leq \frac{\theta \mathrm{Var}_{Q_0}(f)}{2} + \frac{\chi^2(Q_1\|Q_0)}{2\theta}. \tag{B.50}$$

Proof. Recall Young's inequality: for $a, b \in \mathbb{R}$, $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$. As the left-hand side is invariant by translation, we may assume without loss of generality that $\int f\,\mathrm{d}Q_0 = 0$, so that $\mathrm{Var}_{Q_0}(f) = \int f^2\,\mathrm{d}Q_0$. We write

$$\begin{aligned}\int f(\,\mathrm{d}Q_0 - \,\mathrm{d}Q_1) &= \int (\sqrt{\theta} f)\frac{\left(1 - \frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}\right)}{\sqrt{\theta}}\,\mathrm{d}Q_0 \leq \frac{\theta}{2}\int f^2\,\mathrm{d}Q_0 + \frac{1}{2\theta}\int \left(1 - \frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}\right)^2\,\mathrm{d}Q_0 \\ &= \frac{\theta \mathrm{Var}_{Q_0}(f)}{2} + \frac{\chi^2(Q_1\|Q_0)}{2\theta}\,.\end{aligned}$$

□

Lemma B.16 (Expectation of empirical χ^2-divergence)**.** *Let $Q = \sum_{j=1}^J q_j \delta_{y_j}$ be a discrete measure*

supported on J atoms, and let Q_n denote its empirical measure, consisting of n i.i.d. samples. Then,

$$\mathbb{E}[\chi^2(Q_n\|Q)] = \frac{J-1}{n}\,. \tag{B.51}$$

Proof. We can write $Q_n = \sum_{j=1}^{J} \hat{q}_j \delta_{y_j}$, where $\hat{q}_j$ is a binomial random variable with parameters n and q_j. We obtain

$$\chi^2(Q_n\|Q) = \sum_{j=1}^{J} \frac{(\hat{q}_j - q_j)^2}{q_j}\,.$$

Taking expectations, our bound reads

$$\mathbb{E}[\chi^2(Q_n\|Q)] = \sum_{j=1}^{J} \frac{\mathrm{Var}(\hat{q}_j)}{q_j} = \sum_{j=1}^{J} \frac{q_j(1-q_j)}{nq_j} = \frac{J-1}{n}.$$

□

Lemma B.17 (Control of suprema of empirical processes)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from some probability measure P on $\mathbb{R}^d$, with P_n the associated empirical measure. Consider $\mathcal{F}$ a class of functions $\mathbb{R}^d \to \mathbb{R}$ with $\|f\|_\infty \le A$ for all $f \in \mathcal{F}$. For $u > 0$, let $N(u)$ be the u-covering numbers of $\mathcal{F}$, that is the minimal number of balls of radius u for the $\|\cdot\|_\infty$-metric required to cover $\mathcal{F}$. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\int f\,\mathrm{d}(P_n - P)\right|\right] \le \frac{C_0}{\sqrt{n}}\int_0^{C_1 A}\sqrt{\log 2N(u)}\,\mathrm{d}u =: \frac{I}{\sqrt{n}} \tag{B.52}$$

for two positive absolute constants C_0 and C_1. Furthermore, for all $t > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f\,\mathrm{d}(P_n - P)\right| > t\right) \le \exp\left(-\frac{C_2\sqrt{n}t}{I + A\log n}\right), \tag{B.53}$$

for some positive absolute constant C_2. Eventually, for all $p \geq 2$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}(P_n - P) \right|^p \right]^{1/p} \leq C_p \frac{I + A}{\sqrt{n}}. \tag{B.54}$$

Proof. See Vaart and Wellner (1996, Theorem 2.14.2 and Theorem 2.14.5). □

C | Supplement to Chapter 4

C.1 Connecting the two dynamic formulations

In this section, we reconcile (at a formal level) two versions of the dynamic formulation for entropic optimal transport. We will start with (4.10) and show that this is equivalent to (4.9) by a reparameterization.

We begin by recognizing that $\Delta \mathrm{p}_t = \nabla \cdot (\mathrm{p}_t \nabla \log \mathrm{p}_t)$, which allows us to write the Fokker–Planck equation as

$$\partial_t \mathrm{p}_t + \nabla \cdot ((v_t - \tfrac{\varepsilon}{2} \nabla \log \mathrm{p}_t)\mathrm{p}_t) = 0\,, \tag{C.1}$$

Inserting $b_t := v_t - \frac{\varepsilon}{2} \nabla \log \mathrm{p}_t$ into (4.10), we expand the square and arrive at

$$\inf_{(\mathrm{p}_t, b_t)} \int_0^1 \int \big(\frac{1}{2}\|b_t(x)\|^2 + \frac{\varepsilon^2}{8}\|\nabla \log \mathrm{p}_t(x)\|^2 + \frac{\varepsilon}{2} b_t^\top \nabla \log \mathrm{p}_t\big)\mathrm{p}_t(x)\,\mathrm{d}x\,\mathrm{d}t\,.$$

Up to the cross-term, this aligns with (4.9); it remains to eliminate the cross term. Using integration-by-parts and (C.1), we obtain

$$\int_0^1 \int (b_t \mathrm{p}_t)^\top \nabla \log \mathrm{p}_t \,\mathrm{d}x\,\mathrm{d}t = -\int_0^1 \int \nabla \cdot (b_t \mathrm{p}_t) \log \mathrm{p}_t \,\mathrm{d}x\,\mathrm{d}t = \int_0^1 \int (\partial_t \mathrm{p}_t) \log \mathrm{p}_t \,\mathrm{d}x\,\mathrm{d}t\,.$$

Though, we have (by product rule) the equivalence

$$\partial_t(\mathrm{p}_t \log \mathrm{p}_t) - \partial_t \mathrm{p}_t = (\partial_t \mathrm{p}_t) \log \mathrm{p}_t \,.$$

Exchanging partial derivatives under the integral, this yields the following simplification

$$\begin{aligned}
\int_0^1 \int (\partial_t \mathrm{p}_t) \log \mathrm{p}_t \,\mathrm{d}x \,\mathrm{d}t &= \int_0^1 \int \partial_t(\mathrm{p}_t \log \mathrm{p}_t) \,\mathrm{d}x \,\mathrm{d}t - \int_0^1 \int \partial_t \mathrm{p}_t \,\mathrm{d}x \,\mathrm{d}t \\
&= \int_0^1 \partial_t \int \mathrm{p}_t \log \mathrm{p}_t \,\mathrm{d}x \,\mathrm{d}t - \int_0^1 \partial_t \int \mathrm{p}_t \,\mathrm{d}x \,\mathrm{d}t \\
&= \int_0^1 \partial_t \mathcal{H}(\mathrm{p}_t) \,\mathrm{d}t + 0 \\
&= \mathcal{H}(\mathrm{p}_1) - \mathcal{H}(\mathrm{p}_0) \,,
\end{aligned}$$

where $\mathrm{p}_1 = \nu$ and $\mathrm{p}_0 = \mu$. We see that (4.10) is equivalent to

$$\frac{\varepsilon}{2}(\mathcal{H}(\nu) - \mathcal{H}(\mu)) + \inf_{(\mathrm{p}_t, b_t)} \int_0^1 \int \Big(\frac{1}{2}\|b_t(x)\|^2 + \frac{\varepsilon^2}{8}\|\nabla \log \mathrm{p}_t(x)\|^2\Big)\mathrm{p}_t(x) \,\mathrm{d}x \,\mathrm{d}t \,.$$

C.2 Connecting Markov processes and entropic Brenier maps

Here we prove Proposition 4.1. To continue, we require the following lemma.

Lemma C.1. *Fix any $t \in [0,1]$. Under M, the random variables X_0 and X_1 are conditionally independent given X_t.*

Proof. A calculation shows that the joint density of X_0, X_1, and X_t with respect to $\mu_0 \otimes \mu_1 \otimes$Lebesgue equals

$$\Lambda_\varepsilon \Lambda_{t(1-t)\varepsilon} e^{-\frac{1}{2\varepsilon t(1-t)}\|x_t - ((1-t)x_0 + t x_1)\|^2} e^{(f(x_0) + g(x_1) - \frac{1}{2}\|x_0 - x_1\|^2)/\varepsilon} = \mathsf{F}_t(x_t, x_0)\mathsf{G}_t(x_t, x_1) \,,$$

where

$$\mathsf{F}_t(x_t, x_0) = \Lambda_{\varepsilon t} e^{f(x_0)/\varepsilon} e^{-\frac{1}{2\varepsilon t}\|x_t - x_0\|^2} \quad \text{and} \quad \mathsf{G}_t(x_t, x_1) = \Lambda_{(1-t)\varepsilon} e^{g(x_1)/\varepsilon} e^{-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2} .$$

Since this density factors, the law of X_0 and X_1 given X_t is a product measure, proving the claim. □

Proof of Proposition 4.1. First, we prove that M is Markov. Let $(X_t)_{t\in[0,1]}$ be distributed according to M. It suffices to show that for any integrable $a \in \sigma(X_{[0,t]}), b \in \sigma(X_{[t,1]})$, we have the identity

$$\mathbb{E}[ab|X_t] = \mathbb{E}[a|X_t]\mathbb{E}[b|X_t] \quad \text{a.s.}$$

Using the tower property and the fact that, conditioned on X_0 and X_1, the law of the path is a Brownian bridge between X_0 and X_1, and hence is Markov, we have

$$\mathbb{E}_{\mathsf{M}}[ab|X_t] = \mathbb{E}[\mathbb{E}[ab|X_0, X_t, X_1]|X_t] = \mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t] .$$

By Lemma C.1, the sigma-algebras $\sigma(X_0, X_t)$ and $\sigma(X_t, X_1)$ are conditionally independent given X_t, hence

$$\mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t] = \mathbb{E}[\mathbb{E}[a|X_0, X_t]|X_t]\mathbb{E}[\mathbb{E}[b|X_0, X_t]|X_t] = \mathbb{E}[a|X_t]\mathbb{E}[b|X_t] ,$$

as claimed.

The proof of the second statement follows directly from the computations presented below (4.15), which hold under no additional assumptions.

We now prove the third statement. Following the approach of Föllmer (1985), the representation of M as a mixture of Brownian bridges shows that the law of $X_{[0,t]}$ for any $t < 1$ has finite entropy with respect to the law of $X_0 + \sqrt{\varepsilon}B_t$, for $X \sim \mu_0$. Hence, to verify the representation in terms of

the SDE, it suffices to compute the stochastic derivative:

$$\lim_{h\to 0} \frac{1}{h}\mathbb{E}[X_{t+h} - X_t | X_{[0,t]}]\,,$$

where the limit is taken in L^2. Using the the fact that the process is Markov and, conditioned on X_0 and X_1, the path is a Brownian bridge, we obtain

$$\lim_{h\to 0} \frac{1}{h}\mathbb{E}[X_{t+h} - X_t | X_{[0,t]}] = \lim_{h\to 0} \frac{1}{h}\mathbb{E}[\mathbb{E}[X_{t+h} - X_t | X_0, X_t, X_1] | X_t] = \frac{1}{1-t}\mathbb{E}[X_1 - X_t | X_t]\,.$$

Recalling the computations in Lemma C.1, we observe that, conditioned on $X_t = x_t$, the variable X_1 has μ_1 density proportional to $\mathsf{G}_t(x_t, x_1)$. Since π is a probability measure, in particular we have that e^g lies in $L^1(\mu_1)$. We can therefore apply dominated convergence to obtain

$$\frac{1}{1-t}\mathbb{E}[X_1 - X_t | X_t = x_t] = \frac{\int \frac{x_1 - x_t}{1-t}\mathsf{G}_t(x_t, x_1)\mu_1(\mathrm{d}x_1)}{\int \mathsf{G}_t(x_t, x_1)\mu_1(\mathrm{d}x_1)} = \varepsilon\nabla \log \mathsf{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](x_t)\,,$$

as desired.

For the fourth statement, we require the following claim.

Claim: The joint probability measure $\pi_t(z, x_1)$, defined as

$$\exp((-(1-t)f_{1-t}(z) + (1-t)g(x_1) - \tfrac{1}{2}\|z - x_1\|^2))/((1-t)\varepsilon))\mathsf{m}_t(\mathrm{d}z)\mu_1(\mathrm{d}x_1)\,,$$

is the optimal entropic coupling from m_t to ρ with regularization parameter $(1-t)\varepsilon$, where $f_{1-t}(z) \coloneqq \varepsilon \log \mathsf{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z)$. Under this claim, it is easy to verify that the definition of $\nabla\varphi_{1-t}$ is precisely this conditional expectation, which concludes the proof.

To prove the claim, we notice that π_t is already in the correct form of an optimal entropic coupling, and $\pi_t \in \Gamma(\mathsf{m}_t, ?)$ by construction. Thus, it suffices to only check the second marginal.

By the second part, above, we have that

$$\mathsf{m}_t(z) = \mathsf{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z)\mathsf{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z)\,.$$

Integrating, performing the appropriate cancellations, and applying the semigroup property, we have

$$\int \pi_t(z, \mathrm{d}x_1)\,\mathrm{d}z = e^{g(x_1)/\varepsilon}\mu_1(\mathrm{d}x_1)\mathsf{H}_{(1-t)\varepsilon}[\mathsf{H}_{t\varepsilon}[e^{f/\varepsilon}\mu_0]](x_1) = e^{g(x_1)/\varepsilon}\mu_1(\mathrm{d}x_1)\mathsf{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0](x_1)\,,$$

which proves the claim. □

C.3 Proofs for Section 4.4

C.3.1 One-sample analysis

Proof of Proposition 4.5. First, we recognize that a path with law $\tilde{\mathsf{P}}$ (resp. $\bar{\mathsf{P}}$) can be obtained by sampling a Brownian bridge between $(X_0, X_1) \sim \pi_n$ (resp. $\bar{\pi}_n$), by Proposition 4.1. Thus, by the data processing inequality,

$$\mathbb{E}[\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}\|\bar{\mathsf{P}}_{[0,\tau]})] \le \mathbb{E}[\mathrm{KL}(\tilde{\mathsf{P}}\|\bar{\mathsf{P}})] \le \mathbb{E}[\mathrm{KL}(\pi_n\|\bar{\pi}_n)] = \mathbb{E}\Big[\int \log(\pi_n/\bar{\pi}_n)\,\mathrm{d}\pi_n\Big]\,,$$

where the above manipulations are valid as both π_n and $\bar{\pi}_n$ have densities with respect to $\mu \otimes \nu_n$. Completing the expansion by explicitly writing out the densities, we obtain

$$\mathbb{E}[\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}\|\bar{\mathsf{P}}_{[0,\tau]})] \le \frac{1}{\varepsilon}\mathbb{E}\Big[\int (\hat{f} + \hat{g} - \bar{f} - g^\star)\,\mathrm{d}\pi_n\Big] = \frac{1}{\varepsilon}\mathbb{E}\Big[\mathrm{OT}_\varepsilon(\mu, \nu_n) - \int \bar{f}\,\mathrm{d}\mu - \int g^\star\,\mathrm{d}\nu_n\Big]\,.$$

We now employ the rounding trick of Stromme (2024): the rounded potential $\bar{f}$ satisfies

$$\bar{f} = \operatorname*{argmax}_{f \in L^1(\mu)} \mathcal{D}_\varepsilon^{\mu\nu_n}(f, g^\star)\,;$$

Therefore, in particular, $\mathcal{D}_\varepsilon^{\mu\nu_n}(\bar{f}, g^\star) \geq \mathcal{D}_\varepsilon^{\mu\nu_n}(f^\star, g^\star)$. Continuing from above, we obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{KL}(\tilde{\mathrm{P}}_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] &\leq \frac{1}{\varepsilon}\mathbb{E}\Big[\mathrm{OT}_\varepsilon(\mu, \nu_n) - \int f^\star \,\mathrm{d}\mu - \int g^\star \,\mathrm{d}\nu_n\Big] \\
&= \frac{1}{\varepsilon}\mathbb{E}\Big[\mathrm{OT}_\varepsilon(\mu, \nu_n) - \int f^\star \,\mathrm{d}\mu - \int g^\star \,\mathrm{d}\nu\Big] \\
&= \frac{1}{\varepsilon}\mathbb{E}\big[\mathrm{OT}_\varepsilon(\mu, \nu_n) - \mathrm{OT}_\varepsilon(\mu, \nu)\big]\,,
\end{aligned}$$

where in the penultimate equality we observed that g is independent of the data $Y_1, \ldots, Y_n$. Combined with Theorem 2.6 of Groppe and Hundrieser (2024), the proof is complete. □

Proof of Proposition 4.6. We start by applying Girsanov's theorem to obtain a difference in the drifts, which can be re-written as differences in entropic Brenier maps:

$$\mathbb{E}[\mathrm{KL}(\mathrm{P}^\star_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] \leq \int_0^\tau \mathbb{E}\|\bar{b}_t - b_t^\star\|^2_{L^2(\mathrm{p}_t)} \,\mathrm{d}t = \int_0^\tau (1-t)^{-2}\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^\star_{1-t}\|^2_{L^2(\mathrm{p}_t)} \,\mathrm{d}t\,. \quad \text{(C.2)}$$

The result then follows from Lemma C.2, where we lazily bound the resulting integral:

$$\mathbb{E}[\mathrm{KL}(\mathrm{P}^\star_{[0,\tau]} \| \bar{\mathrm{P}}_{[0,\tau]})] \leq \frac{R^2\varepsilon^{-\mathsf{k}}}{n}\int_0^\tau (1-t)^{-\mathsf{k}-2}\,\mathrm{d}t \leq \frac{R^2\varepsilon^{-\mathsf{k}}}{n}(1-\tau)^{-\mathsf{k}-2}\,.$$

□

Lemma C.2 (Point-wise drift bound). *Under the assumptions of Proposition 4.6, let $\bar{\varphi}_{1-t}$ be the entropic Brenier map between $\bar{\mathrm{p}}_t$ and $\bar{\nu}_n$ and $\varphi^\star_{1-t}$ be the entropic Brenier map between $\mathrm{p}^\star_t$ and ν, both*

with regularization parameter $(1-t)\varepsilon$. *Then*

$$\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathrm{p}_t)} \lesssim \frac{R^2}{n}((1-t)\varepsilon)^{-\mathsf{k}}\,.$$

Proof. Setting some notation, we express $\nabla\varphi^{\star}_{1-t}$ as the conditional expectation of the optimal entropic coupling $\pi^{\star}_t$ between $\mathrm{p}^{\star}_t$ and ν (recall Proposition 4.1), where we write

$$\pi^{\star}_t(z,y) = \gamma^{\star}_t(z,y)\mathrm{p}^{\star}_t(\mathrm{d}z)\nu(\mathrm{d}y)\,.$$

The rest of our proof follows a technique due to Stromme (2024): by triangle inequality, we can add and subtract the following term

$$\frac{1}{n}\sum_{j=1}^{n} Y_j\gamma^{\star}_t(z,Y_j)\,,$$

into the integrand in (C.2), resulting in

$$\begin{aligned}\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathrm{p}^{\star}_t)} \lesssim\ & \mathbb{E}\|\nabla\bar{\varphi}_{1-t} - n^{-1}\textstyle\sum_{j=1}^{n} Y_j\gamma^{\star}_t(\cdot,Y_j)\|^2_{L^2(\mathrm{p}^{\star}_t)} \\ & + \mathbb{E}\|n^{-1}\textstyle\sum_{j=1}^{n} Y_j\gamma^{\star}_t(\cdot,Y_j) - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathrm{p}^{\star}_t)}\,.\end{aligned} \tag{C.3}$$

For the second term, with the same manipulations as Stromme (2024, Lemma 20), we obtain a final bound of

$$\mathbb{E}\|n^{-1}\textstyle\sum_{j=1}^{n} Y_j\gamma^{\star}_t(\cdot,Y_j) - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathrm{p}^{\star}_t)} = \frac{R^2}{n}\|\gamma^{\star}_t\|^2_{L^2(\mathrm{p}^{\star}_t\otimes\nu)} \leq \frac{R^2}{n}((1-t)\varepsilon)^{-\mathsf{k}}\,,$$

where the final inequality is also due to Stromme (2024, Lemma 16). To control the first term in

(C.3), we also appeal to his calculations of the same theorem: observing that, from (4.25)

$$\nabla\bar{\varphi}_{1-t}(z) = \frac{1}{n}\sum_{j=1}^{n} Y_j \frac{\exp((g^\star(Y_j) - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)}{\frac{1}{n}\sum_{j=1}^{n}\exp((g^\star(Y_j) - \frac{1}{2(1-t)}\|z - Y_j\|^2)/\varepsilon)} = \frac{1}{n}\sum_{j=1}^{n} Y_j \bar{\gamma}_t(z, Y_j)\,.$$

Since the following equality is true

$$\bar{\gamma}_t(z, Y_j) = \frac{\gamma_t^\star(z, Y_j)}{\frac{1}{n}\sum_{k=1}^{n}\gamma_t^\star(z, Y_k)}\,,$$

we can verbatim apply the remaining arguments of Stromme (2024, Lemma 20). Indeed, for fixed $x \in \mathbb{R}^d$, we have

$$\|n^{-1}\textstyle\sum_{j=1}^{n} Y_j(\gamma_t^\star(x, Y_j) - \bar{\gamma}_t(x, Y_j))\|^2 \le R^2\left|\sum_{j=1}^{n}\gamma_t^\star(x, Y_j) - 1\right|^2.$$

Taking the $L^2(\mathrm{p}_t^\star)$ norm and the outer expectation, we see that the remaining term is nothing but the first component of the gradient of the dual entropic objective function (see Proposition C.6), which can be bounded via Lemma C.7, resulting in the chain of inequalities

$$\mathbb{E}\|n^{-1}\textstyle\sum_{j=1}^{n} Y_j(\gamma_t^\star(\cdot, Y_j) - \bar{\gamma}_t(\cdot, Y_j))\|^2_{L^2(\mathrm{p}_t^\star)} \lesssim \frac{R^2}{n}\|\gamma_t^\star\|^2_{L^2(\mathrm{p}_t^\star\otimes\nu)} \le \frac{R^2}{n}((1-t)\varepsilon)^{-\mathrm{k}},$$

where the last inequality again holds via Stromme (2024, Lemma 16).

□

C.3.2 Completing the results

Proof of Proposition 4.7. This proof closely follows the ideas of Chen et al. (2022b). Applying Girsanov's theorem, we obtain

$$\mathrm{TV}^2(\hat{\mathrm{P}}_{[0,\tau]}, \tilde{\mathrm{P}}_{[0,\tau]}) \lesssim \mathrm{KL}(\tilde{\mathrm{P}}_{[0,\tau]} \| \hat{\mathrm{P}}_{[0,\tau]}) = \sum_{k=0}^{N-1} \int_{k\eta}^{(k+1)\eta} \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \, \mathrm{d}t \, .$$

Recall that $\eta \in (0,1)$ is a chosen step-size based on N, the number of steps to be taken. As in prior analyses, we hope to uniformly bound the integrand above for any $t \in [k\eta, (k+1)\eta]$. Adding and subtracting the appropriate terms, we have

$$\mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \lesssim \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_{k\eta})\|^2 + \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_t(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \, . \tag{C.4}$$

By the semigroup property, we first notice that

$$\mathsf{H}_{1-k\eta}[e^{\hat{g}/\varepsilon} \nu_n] = \mathsf{H}_{t-k\eta}[\mathsf{H}_{1-t}[e^{\hat{g}/\varepsilon} \nu_n]] \, .$$

We can verbatim apply Lemma 16 of Chen et al. (2022b) with $\boldsymbol{q} \coloneqq \mathsf{H}_{1-t}[e^{\hat{g}/\varepsilon} \nu_n]$, $\boldsymbol{M}_0 = \mathrm{id}$ and $\boldsymbol{M}_1 = (t-k\eta)I$, since $\mathsf{H}_{1-k\eta}[e^{\hat{g}/\varepsilon} \nu_n] = \boldsymbol{q} * \mathcal{N}(0, (t-k\eta)I)$. This gives

$$\begin{aligned} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_{k\eta})\|^2 &= \left\| \varepsilon \nabla \log \frac{\boldsymbol{q} * \mathcal{N}(0, (t-k\eta)I)}{\boldsymbol{q}} (X_{k\eta}) \right\|^2 \\ &\lesssim L_t^2 \eta d + L_t^2 \eta^2 \|\varepsilon \nabla \log \boldsymbol{q}(X_{kh})\|^2 \, . \end{aligned}$$

Since $\varepsilon \log \boldsymbol{q}$ is L_t-smooth, we obtain the bounds

$$\begin{aligned} \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\varepsilon \nabla \log \boldsymbol{q}(X_{kh})\|^2 &\lesssim \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\varepsilon \nabla \log \boldsymbol{q}(X_t)\|^2 + L_t^2 \|X_t - X_{kh}\|^2 \\ &\leq \varepsilon L_t d + L_t^2 \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|X_t - X_{kh}\|^2 \, . \end{aligned}$$

where the final inequality is a standard smoothness inequality (see Lemma C.5). Similarly, the second term on the right-hand side of (C.4) can be bounded by

$$\mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_t(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \le L_t^2 \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|X_{k\eta} - X_t\|^2.$$

Combining the terms, we obtain

$$\mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_{k\eta}(X_{k\eta}) - \hat{b}_t(X_t)\|^2 \lesssim \varepsilon L_t^2 \eta d + L_t^2 \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|X_{k\eta} - X_t\|^2,$$

where, to simplify, we use the fact that $\eta \le 1/L_t$ (with $L_t \ge 1$), and that $\eta^2 \le \eta$ for $\eta \in [0,1]$. We now bound the remaining expectation. Under $\tilde{\mathrm{P}}_{[0,\tau]}$, we can write

$$X_t = \int_0^t \hat{b}_s(X_s)\,\mathrm{d}s + \sqrt{\varepsilon} B_t\,, X_{kh} = \int_0^{k\eta} \hat{b}_s(X_s)\,\mathrm{d}s + \sqrt{\varepsilon} B_{k\eta}\,,$$

and thus

$$X_t - X_{k\eta} = \int_{k\eta}^t \hat{b}_s(X_s)\,\mathrm{d}s + \sqrt{\varepsilon}(B_t - B_{k\eta})\,.$$

Taking squared expectations, writing $\delta \coloneqq t - k\eta \le \eta$ (recall that $t \in [k\eta, (k+1)\eta)$), we obtain (through an application of the triangle inequality and Jensen's inequality)

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|X_t - X_{k\eta}\|^2 &\lesssim \varepsilon \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|B_t - B_{k\eta}\|^2 + \delta \int_{k\eta}^t \mathbb{E}_{\tilde{\mathrm{P}}_{[0,\tau]}} \|\hat{b}_s(X_s)\|^2\,\mathrm{d}s \\
&\lesssim \varepsilon\eta d + \delta^2 L_t d \\
&\le (\varepsilon + 1)\eta d
\end{aligned}$$

where we again used Lemma C.5. Combining all like terms, we obtain the final result.

The estimates for the Lipschitz constant follow from Lemma C.4. □

C.4 Proofs for Section 4.4.3

C.4.1 Computing Equation 4.26

The Föllmer drift is a special case of the Schrödinger bridge, where $\mu = \delta_a$ for any $a \in \mathbb{R}^d$. Let $(f^{\mathsf{F}}, g^{\mathsf{F}})$ denote the optimal entropic potentials in this setting. Note that they these potentials are defined up to translation (i.e., the solution is the same if we take $f^{\mathsf{F}} + c$ and $g^{\mathsf{F}} - c$ for any $c \in \mathbb{R}$). So, we further impose the condition that $f^{\mathsf{F}}(a) = 0 = c$. Then the optimality conditions yield

$$g^{\mathsf{F}}(y) = \frac{1}{2\varepsilon}\|y\|^2 \,. \tag{C.5}$$

Plugging this into the expression for the Schrödinger bridge drift, we obtain

$$b_t^{\mathsf{F}}(z) = \varepsilon \nabla \log \mathrm{H}_{(1-t)\varepsilon}[e^{\frac{1}{2\varepsilon}\|\cdot\|^2}\nu](z) = (1-t)^{-1}\left(-z + \frac{\int y e^{\frac{1}{2\varepsilon}\|y\|^2 - \frac{1}{2(1-t)\varepsilon}\|z-y\|^2}\nu(\mathrm{d}y)}{\int e^{\frac{1}{2\varepsilon}\|y\|^2 - \frac{1}{2(1-t)\varepsilon}\|z-y\|^2}\nu(\mathrm{d}y)}\right).$$

Replacing the integrals with respect to ν with their empirical counterparts yields the estimator.

C.4.2 Proof of Proposition 4.9

Our goal is to prove the following lemma.

Lemma C.3. *Let p_τ be the Föllmer bridge at time $\tau \in [0,1)$ between $\mu = \delta_0$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\varepsilon = 1$ and suppose the squared second moment of ν is bounded above by d. Then*

$$W_2^2(\mathrm{p}_\tau, \nu) \le d(1-\tau) \,.$$

Proof. Note that $\mathrm{p}_\tau = \mathrm{P}_{1-\tau}$, where $\mathrm{P}_{1-\tau}$ is the *reverse* bridge, which starts at ν and ends at $\mu = \delta_0$. This reverse bridge is well known to satisfy a simple SDE (Föllmer, 1985): the measure $\mathrm{P}_{1-\tau}$ is the

law of $Y_{1-\tau}$, where Y_s solves

$$\mathrm{d}Y_s = -\frac{Y_s}{1-s}\mathrm{d}s + \mathrm{d}B_s, \qquad Y_0 \sim \nu,$$

which has the explicit solution

$$Y_s = (1-s)Y_0 + (1-s)\int_0^s \frac{1}{1-r}\mathrm{d}B_r\,.$$

In particular, we obtain

$$\begin{aligned} W_2^2(\mathrm{P}_s, \nu) &\le \mathbb{E}\|Y_s - Y_0\|^2 \\ &= \mathbb{E}\left\|-sY_0 + (1-s)\int_0^s \frac{1}{1-r}\mathrm{d}B_r\right\|^2 \\ &= s^2\mathbb{E}\|Y_0\|^2 + ds(1-s) \\ &\le ds\,, \end{aligned}$$

which proves the claim. □

C.5 Technical lemmas

Lemma C.4 (Hessian calculation and bounds)**.** *Let (p_t, b_t) be the optimal density-drift pair satisfying the Fokker–Planck equation (4.10) between μ_0 and μ_1. For $t \in [0,1)$, b_t is Lipschitz with constant L_t given by*

$$L_t := \sup_x \|\nabla b_t(x)\|_{\mathrm{op}} \le \frac{1}{(1-t)}\Big(1 \vee \|\nabla^2\varphi_{1-t}(x)\|_{\mathrm{op}}\Big),$$

where $\nabla\varphi_{1-t}$ is the entropic Brenier map between p_t and μ_1 with regularization parameter $(1-t)\varepsilon$. Moreover, if the support of μ_1 is contained in $B(0,R)$, then

$$L_t \le (1-t)^{-1}(1 \vee R^2((1-t)\varepsilon)^{-1})\,. \tag{C.6}$$

Proof. Taking the Jacobian of b_t, we arrive at

$$\nabla b_t(x) = (1-t)^{-1}(\nabla^2\varphi_{1-t}(x) - I)\,,$$

As entropic Brenier potentials are convex (recall that their Hessians are covariance matrices; see (4.7)), we have the bounds

$$-(1-t)^{-1}I \preceq \nabla b_t(x) \preceq (1-t)^{-1}\nabla^2\varphi_{1-t}(x)\,.$$

The first claim follows by considering the larger of the two operator norms of both sides.

The second claim follows from the fact that since φ_{1-t} is an optimal entropic Brenier potential, its Hessian is the conditional covariance of an optimal entropic coupling $\pi_t \in \Gamma(\mathrm{p}_t, \mu_1)$, so

$$\|\nabla^2\varphi_{1-t}(z)\|_{\mathrm{op}} = \frac{1}{(1-t)\varepsilon}\|\mathrm{Cov}_{\pi_t}[Y|X_t = z]\|_{\mathrm{op}} \le \frac{R^2}{(1-t)\varepsilon}\,,$$

since $\mathrm{supp}(\mu_1) \subseteq B(0,R)$. □

Lemma C.5. *Let (p_t, b_t) be the optimal density-drift pair satisfying the Fokker–Planck equation (4.10) between μ_0 and μ_1. Then for any $t \in [0,1)$*

$$\mathbb{E}_{\mathrm{p}_t}\|b_t\|^2 \le \frac{\varepsilon}{2}L_t d\,.$$

Proof. This proof follows the ideas of Vempala and Wibisono (2019, Lemma 9). We note that the

generator given by the forward Schrödinger bridge with volatility ε is

$$\mathcal{L}_t f = \frac{\varepsilon}{2}\Delta f - \langle b_t, \nabla f\rangle\,,$$

for a smooth function f. Writing $b_t = \nabla(\varepsilon \log \mathsf{H}_{1-t}[e^{g/\varepsilon}\mu_1])$, we obtain

$$0 = \mathbb{E}_{\mathrm{p}_t}\mathcal{L}_t(\varepsilon \log \mathsf{H}_{1-t}[e^{g/\varepsilon}\mu_1]) \implies \mathbb{E}_{\mathrm{p}_t}\|b_t(X_t)\|^2 = \frac{\varepsilon}{2}\mathbb{E}_{\mathrm{p}_t}[\nabla \cdot b_t] \le \frac{\varepsilon}{2}L_t d\,.$$

□

Lemma C.6. *(Stromme, 2024, Proposition 3.1) Let P, Q be probability measures on $\mathbb{R}^d$, and fix $\varepsilon > 0$. For every pair $h_1 = (f_1, g_1) \in L^\infty(P) \times L^\infty(Q)$, there exists an element of $L^\infty(P) \times L^\infty(Q)$ which we denote by $\nabla \mathcal{D}_\varepsilon^{PQ}(f_1, g_1)$ such that for all $h_0 = (f_0, g_0) \in L^\infty(P) \times L^\infty(Q)$,*

$$\begin{aligned}\langle \nabla \mathcal{D}_\varepsilon^{PQ}(h_1), h_0\rangle_{L^2(P)\times L^2(Q)} = &\int f_0(x)\Big(1 - \int e^{-\varepsilon^{-1}(c(x,y)-f_1(x)-g_1(y))}\,\mathrm{d}Q(y)\Big)\,\mathrm{d}P(x)\\ &+ \int g_0(y)\Big(1 - \int e^{-\varepsilon^{-1}(c(x,y)-f_1(x)-g_1(y))}\,\mathrm{d}P(x)\Big)\,\mathrm{d}Q(x).\end{aligned}$$

In other words, the gradient of $\mathcal{D}_\varepsilon^{PQ}$ at (f_1, g_1) is the marginal error corresponding to (f_1, g_1).

Lemma C.7. *Following Proposition C.6, suppose $P = \mu$ and $Q = \nu_n$, where ν_n is the empirical measure of some measure ν on the basis of n i.i.d. samples. Let $(f^\star, g^\star)$ be the optimal entropic potentials between μ and ν, which induce an optimal entropic coupling $\pi^\star$ (recall (1.26)). Then*

$$\mathbb{E}\|\nabla \mathcal{D}_\varepsilon^{\mu\nu_n}(f^\star, g^\star)\|^2_{L^2(\mu)\times L^2(\nu_n)} \lesssim \frac{\|\gamma^\star\|^2_{L^2(\mu\otimes\nu)}}{n}\,,$$

where the expectation is with respect to the data, and $\gamma^\star = \frac{\mathrm{d}\pi^\star}{\mathrm{d}(\mu\otimes\nu)}$.

Proof. Writing out the squared-norm of the gradient explicitly in the norm $L^2(\mu) \times L^2(\nu_n)$, we

obtain

$$\mathbb{E}\|\nabla \mathcal{D}_\varepsilon^{\mu\nu_n}(f^\star, g^\star)\|^2_{L^2(\mu)\times L^2(\nu_n)} = \mathbb{E}\int \Big(\frac{1}{n}\sum_{j=1}^n \gamma^\star(x, Y_j) - 1\Big)^2 \mu(\mathrm{d}x) + \mathbb{E}\frac{1}{n}\sum_{j=1}^n \Big(\int \gamma^\star(x, Y_j)\mu(\mathrm{d}x) - 1\Big)^2 .$$

Note that by the optimality conditions, $\int \gamma^\star(x, Y_j)\mu(\mathrm{d}x) = 1$ for all Y_j. Thus, writing $Z_j \coloneqq \gamma^\star(x, Y_j)$ which are i.i.d., we see that

$$\mathbb{E}\int \Big(\frac{1}{n}\sum_{j=1}^n \gamma^\star(x, Y_j) - 1\Big)^2 \mu(\mathrm{d}x) = \int \mathbb{E}\Big(\frac{1}{n}\sum_{j=1}^n (Z_j - \mathbb{E}[Z_j])\Big)^2 = \mathrm{Var}_{\mu\otimes\nu}\Big(\frac{1}{n}\sum_{j=1}^n Z_j\Big) = \frac{1}{n}\mathrm{Var}_{\mu\otimes\nu}(Z_1) .$$

The remaining component of the squared gradient vanishes, and we obtain

$$\mathbb{E}\|\nabla \mathcal{D}_\varepsilon^{\mu\nu_n}(f^\star, g^\star)\|^2_{L^2(\mu)\times L^2(\nu_n)} = \frac{1}{n}\mathrm{Var}_{\mu\otimes\nu}(\gamma^\star) \le \frac{\|\gamma^\star\|^2_{L^2(\mu\otimes\nu)}}{n} .$$

□

D | Supplement to Chapter 5

D.1 Proof of the Cramér–Rao lower bound

For any smooth and compactly supported test function $h : \mathbb{R}^d \to \mathbb{R}$, integration by parts yields

$$\mathbb{E}_P \nabla h = \int \nabla h \, \mathrm{d}P = -\int (h \, \nabla \ln P) \, \mathrm{d}P = \int (h - \mathbb{E}_P h) \, \nabla V \, \mathrm{d}P$$

where we used the fact that $\mathbb{E}_P \nabla \ln P = 0$. Therefore,

$$\langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} \, \mathbb{E}_P \nabla h \rangle = \int (h - \mathbb{E}_P h) \, \langle \nabla V, (\mathbb{E}_P \nabla^2 V)^{-1} \, \mathbb{E}_P \nabla h \rangle \, \mathrm{d}P \,. \tag{D.1}$$

Applying the Cauchy–Schwarz inequality,

$$(\text{D.1}) \leqslant \sqrt{(\mathrm{Var}_P \, h) \int \langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} \, (\nabla V)^{\otimes 2} \, (\mathbb{E}_P \nabla^2 V)^{-1} \, \mathbb{E}_P \nabla h \rangle \, \mathrm{d}P} \,.$$

Integration by parts shows that $\int \nabla V^{\otimes 2} \, \mathrm{d}P = \int \nabla^2 V \, \mathrm{d}P$, and upon rearranging we deduce that

$$\mathrm{Var}_P \, h \geqslant \langle \mathbb{E}_P \nabla h, (\mathbb{E}_P \nabla^2 V)^{-1} \, \mathbb{E}_P \nabla h \rangle \,. \tag{D.2}$$

By approximation, this continues to hold for any locally Lipschitz $h : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_P \, \|\nabla h\| < \infty$.

Specializing the inequality (D.2) to $h := \langle e, \cdot \rangle$ for a unit vector $e \in \mathbb{R}^d$ then recovers the

Cramér–Rao inequality of Lemma 5.5.

D.2 Gaussian case

Suppose $P = \mathcal{N}(0, A)$ and $Q = \mathcal{N}(0, B)$ are Gaussians. Then, it is known that the Hessian of the Brenier potential is given by (see Gelbrich, 1990)

$$\nabla^2 \varphi_0(x) = A^{-1/2} \left(A^{1/2} B A^{1/2}\right)^{1/2} A^{-1/2} .$$

If we have $A^{-1} \preceq \beta I$ and $B^{-1} \succeq \alpha I \succ 0$, then Caffarelli's contraction theorem (Theorem 5.1) implies

$$\|\nabla^2 \varphi_0\|_{\mathrm{op}} \leq \sqrt{\beta/\alpha} .$$

This matches the bound of Altschuler et al. (2021, Lemma 2).

For $\varepsilon > 0$, the upper bound from Theorem 5.6 implies

$$\left\|\nabla^2 \varphi_\varepsilon\right\|_{\mathrm{op}} \leq \frac{1}{2} \left(\sqrt{4\beta/\alpha + \varepsilon^2 \beta^2} - \varepsilon\beta\right) . \tag{D.3}$$

On the other hand, from Janati et al. (2020); Mallasto et al. (2022), it is known that

$$\nabla^2 \varphi_\varepsilon(x) = A^{-1/2} \left(A^{1/2} B A^{1/2} + \frac{\varepsilon^2}{4} I\right)^{1/2} A^{-1/2} - \frac{\varepsilon}{2} A^{-1} .$$

In particular, if we take $A = \beta^{-1} I$ and $B = \alpha^{-1} I$, then (D.3) is an equality. Hence, Theorem 5.6 is sharp for every $\varepsilon > 0$.

E | Supplement to Chapter 6

E.1 Proof of Lemma 6.1

We temporarily omit the superscript in ν for ease of reading. Note that

$$\mathbb{E}_{Z\sim\pi_\varepsilon(\cdot|x)}[e^{\langle h,Z\rangle}] = e^{-\varphi_\varepsilon(x)/\varepsilon}\int e^{(\langle x+\varepsilon h,z\rangle-\psi_\varepsilon(z))/\varepsilon}\,\mathrm{d}\nu(z) = e^{(\varphi_\varepsilon(x+\varepsilon h)-\varphi_\varepsilon(x))/\varepsilon}\,.$$

From this we can conclude, since for all $z\in\mathbb{R}^d$

$$\begin{aligned}\frac{\mathrm{d}\mathcal{T}_h\pi_\varepsilon(\cdot|x)}{\mathrm{d}\nu}(z) &= e^{(\langle x+\varepsilon h,z\rangle-\varphi_\varepsilon(x)-\psi_\varepsilon(z))/\varepsilon}e^{(\varphi_\varepsilon(x)-\varphi_\varepsilon(x+\varepsilon h))/\varepsilon}\\ &= e^{(\langle x+\varepsilon h,z\rangle-\varphi_\varepsilon(x+\varepsilon h)-\psi_\varepsilon(z))/\varepsilon}\\ &= \frac{\mathrm{d}\pi_\varepsilon(\cdot|x+\varepsilon h)}{\mathrm{d}\nu}(z)\,.\end{aligned}$$

E.2 Proof of Corollary 6.2

If the domain of φ_ε^ν is not $\mathbb{R}^d$, then $H_{\max}(\varphi_\varepsilon^\nu)=+\infty$, and the proposition is vacuous. Otherwise, fix $x\in\mathbb{R}^d$. By definition of tilt stability, it suffices to compute an upper bound on the covariance of $\mathcal{T}_h\pi_\varepsilon^\nu(\cdot|x)$ which holds uniformly over all tilts $h\in\mathbb{R}^d$. This follows by direct computation, as

Lemma 6.1 and (6.10) imply that

$$\mathrm{Cov}(\mathcal{T}_h \pi_\varepsilon^\nu(\cdot|x)) = \mathrm{Cov}(\pi_\varepsilon^\nu(\cdot|x+\varepsilon h)) \preceq H_{\max}(\varphi_\varepsilon^\nu) I\,,$$

where the last inequality holds by taking the supremum over both x and h arguments. Note that the argument is symmetric for the other conditional entropic coupling.

E.3 Proof of Proposition 6.8

We assume that φ_ε^ν is finite everywhere, for otherwise there is nothing to prove. We have that

$$\begin{aligned}
T_\varepsilon^\mu(x) - T_\varepsilon^\nu(x) &= \int y \,\mathrm{d}\pi_\varepsilon^\mu(y|x) - \int z \,\mathrm{d}\pi_\varepsilon^\nu(z|x) \\
&= \iint y \gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\tau(y,z) - \int z \,\mathrm{d}\pi_\varepsilon^\nu(z|x) \\
&= \iint (y-z)\gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\tau(y,z) + \iint z(\gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\tau(y,z) - \,\mathrm{d}\pi_\varepsilon^\nu(z|x)) \\
&= \iint (y-z)\gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\tau(y,z) + \int z \,\mathrm{d}(Q(z|x) - \pi_\varepsilon^\nu(z|x))\,.
\end{aligned}$$

Taking the $L^2(\rho)$-norm of both sides and applying Minkowski's and Jensen's inequalities yields

$$\|T_\varepsilon^\mu - T_\varepsilon^\nu\|_{L^2(\rho)} \le \left(\iint \|y-z\|^2 \gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\tau(y,z) \,\mathrm{d}\rho(x)\right)^{1/2} + \left\| \int z \,\mathrm{d}(Q(z|\cdot) - \pi_\varepsilon^\nu(z|\cdot)) \right\|_{L^2(\rho)}.$$

Since τ is an optimal coupling between μ and ν and $\int \gamma_\varepsilon^\mu(x,y) \,\mathrm{d}\rho(x) = 1$, the first term is $W_2(\mu,\nu)$. For the second term, Corollary 6.2 implies for all $x \in \mathbb{R}^d$

$$\left\| \int z \,\mathrm{d}(Q(z|x) - \pi_\varepsilon^\nu(z|x)) \right\| \le \sqrt{2H_{\max}(\varphi_\varepsilon^\nu)\mathrm{KL}(Q(\cdot|x)\|\pi_\varepsilon^\nu(\cdot|x))}\,.$$

Therefore

$$\left\| \int z \, \mathrm{d}(Q(z|\cdot) - \pi_\varepsilon^\nu(z|\cdot)) \right\|_{L^2(\rho)} \leq \sqrt{2H_{\max}(\varphi_\varepsilon^\nu) \int \mathrm{KL}(Q(\cdot|x) \| \pi_\varepsilon^\nu(\cdot|x)) \, \mathrm{d}\rho(x)}$$
$$= (2H_{\max}(\varphi_\varepsilon^\nu) I)^{1/2} ,$$

which completes the proof.

E.4 Proof of Proposition 6.10

We assume that ψ_ε^ν is finite everywhere, for otherwise there is nothing to prove. Recall that $S_\varepsilon^\nu(z) = \int x \, \mathrm{d}\pi_\varepsilon^\nu(x|z)$ and similarly for $S_\varepsilon^\mu(y)$. By Corollary 6.2, we have the following bound

$$\begin{aligned} \|S_\varepsilon^\nu(z) - S_\varepsilon^\mu(y)\|^2 &= \left\| \int x \, \mathrm{d}(\pi_\varepsilon^\nu(x|z) - \pi_\varepsilon^\mu(x|y)) \right\|^2 \\ &\leq 2H_{\max}(\psi_\varepsilon^\nu) \mathrm{KL}(\pi_\varepsilon^\mu(\cdot|y) \| \pi_\varepsilon^\nu(\cdot|z)) \\ &= 2H_{\max}(\psi_\varepsilon^\nu) \int \log\Big(\frac{\gamma_\varepsilon^\mu(x,y)}{\gamma_\varepsilon^\nu(x,z)}\Big) \gamma_\varepsilon^\mu(x,y) \, \mathrm{d}\rho(x) . \end{aligned}$$

Integrating with respect to τ concludes the proof.

E.5 Proof of the bias term

Recall that our target measures are discrete measures of the form

$$\mu = \sum_{j=1}^{J} \mu_j \delta_{y_j} .$$

and that we write the Laguerre cells as L_i for $i \in \{1, \ldots, J\}$.

We require the following definitions, which we borrow from Altschuler et al. (2022). For $x \in L_i$

and any other $j \in \{1, \dots, J\}$, we write

$$\Delta_{ij}(x) := 2(\langle x, y_i - y_j \rangle - \psi_0^\mu(y_i) + \psi_0^\mu(y_j)),$$

and $H_{ij}(t) = \{x \in L_i : \Delta_{ij}(x) = t\}$, which represents the trace on L_i of the hyperplane spanned by the boundary between L_i and L_j, shifted by t (should the two cells have non-empty intersection). Moreover, we have the following co-area formula: for every nonnegative measurable function $f : \mathbb{R} \to \mathbb{R}_+$,

$$\int_{L_i} f(\Delta_{ij}(x))\rho(x)\,\mathrm{d}x = \frac{1}{2\|y_i - y_j\|}\int_0^\infty f(t)h_{ij}(t)\,\mathrm{d}t,$$

where

$$h_{ij}(t) = \int_{H_{ij}(t)} \rho(x)\,\mathrm{d}\mathcal{H}_{d-1}(x), \tag{E.1}$$

and $\mathcal{H}_{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

Proof of Proposition 6.11. Let $x \in L_i$. For $j \in \{1, \dots, J\}$ other than i, we have the upper bound

$$\begin{aligned}
\pi_\varepsilon^\mu(y_j|x) &= e^{(\langle x, y_j\rangle - \varphi_\varepsilon^\mu(x) - \tilde{\psi}_\varepsilon^\mu(y_j))/\varepsilon} \\
&= \frac{e^{(\langle x, y_j\rangle - \tilde{\psi}_\varepsilon^\mu(y_j))/\varepsilon}}{\sum_{k=1}^J e^{(\langle x, y_k\rangle - \tilde{\psi}_\varepsilon^\mu(y_k))/\varepsilon}} \\
&\le \frac{e^{(\langle x, y_j\rangle - \tilde{\psi}_\varepsilon^\mu(y_j))/\varepsilon}}{e^{(\langle x, y_i\rangle - \tilde{\psi}_\varepsilon^\mu(y_i))/\varepsilon} + e^{(\langle x, y_j\rangle - \tilde{\psi}_\varepsilon^\mu(y_j))/\varepsilon}}.
\end{aligned}$$

Adding and subtracting appropriate factors of $\psi_0^\mu(y_i)$ and $\psi_0^\mu(y_j)$, we obtain

$$\pi_\varepsilon^\mu(y_j|x) \le e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon}\frac{e^{(\langle x, y_j\rangle - \psi_0^\mu(y_j))/\varepsilon}}{e^{(\langle x, y_i\rangle - \psi_0^\mu(y_i))/\varepsilon} + e^{(\langle x, y_j\rangle - \psi_0^\mu(y_j))/\varepsilon}} = e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon}\left(1 + e^{\Delta_{ij}(x)/2\varepsilon}\right)^{-1},$$

By an application of Jensen's inequality, we have

$$\|T_\varepsilon^\mu(x) - y_i\|^2 \leq \sum_{j=1}^J \pi_\varepsilon^\mu(y_j|x)\|y_i - y_j\|^2 \leq e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon} \sum_{j=1}^J \|y_i - y_j\|^2 \left(1 + e^{\Delta_{ij}(x)/2\varepsilon}\right)^{-1},$$

so integrating against ρ (partitioned into the J Laguerre cells) yields

$$\begin{aligned}
\|T_\varepsilon^\mu - T_0^\mu\|_{L^2(\rho)}^2 &\leq e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon} \sum_{i,j} \|y_i - y_j\|^2 \int_{L_i} \left(1 + e^{\Delta_{ij}(x)/2\varepsilon}\right)^{-1} \mathrm{d}\rho(x) \\
&= e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon} \sum_{i,j} \|y_i - y_j\|/2 \int_0^\infty h_{ij}(t)\left(1 + e^{t/2\varepsilon}\right)^{-1} \mathrm{d}t \\
&= e^{2\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty/\varepsilon} \varepsilon \sum_{i,j} \|y_i - y_j\|/2 \int_0^\infty h_{ij}(u\varepsilon)\left(1 + e^{u/2}\right)^{-1} \mathrm{d}u\,,
\end{aligned}$$

where the second line follows from the definition of the co-area formula, and the last line is a change of variables $u = t/\varepsilon$. This gives (6.16).

With the additional assumptions **(T1)** and **(T2)**, we can use Corollary 2.2 by Delalande (2022), which tells us that

$$\varepsilon^{-1}\|\psi_0^\mu - \tilde{\psi}_\varepsilon^\mu\|_\infty \leq C_1\varepsilon^\alpha\,, \tag{E.2}$$

where the underlying constant depends on $d, R, J, \mu_{\min}, \min_{i\neq j}\|y_i - y_j\|, \rho_{\min}, \rho_{\max}$, and on the maximum angle formed by three non aligned points among the atoms $\{y_j\}_{j=1}^J$. This gives and upper bound of

$$\|T_\varepsilon^\mu - T_0^\mu\|_{L^2(\rho)}^2 \leq e^{C_1\varepsilon^\alpha}\left(\sum_{i,j} \|y_i - y_j\|/2 \int_0^\infty h_{ij}(u\varepsilon)\left(1 + e^{u/2}\right)^{-1} \mathrm{d}u\right)\varepsilon\,.$$

Since $\|y_i - y_j\| \leq 2R$, $h_{ij}(u\varepsilon)$ is bounded under our assumptions on ρ, the proof is concluded. □

F | Supplement to Chapter 7

F.1 Proofs for Section 7.3

Proof of Lemma 7.3. Take $T_1(x) = A_1 x$ and $T_2(x) = A_2 x$ for A_1, A_2 positive definite, and mutually diagonalizable: there exists an orthogonal matrix U such that $A_i = U\Lambda_i U^{-1}$ with Λ_i diagonal with positive entries. Then

$$(T_1 \circ (T_2)^{-1})(x) = U\Lambda_1 U^{-1} \left(U\Lambda_2 U^{-1}\right)^{-1} x = U\Lambda_1\Lambda_2^{-1}U^{-1}x = \tilde{A}x\,,$$

with $\tilde{A} \succ 0$; this completes the claim. □

Proof of Lemma 7.4. See Panaretos and Zemel (2020, Section 2.3.2). □

Proof of Lemma 7.5. Let $S, T \in \mathcal{M}$, and for simplicity assume they are strictly increasing. Note that T^{-1} is also strictly increasing, so $S \circ T^{-1}$ is strictly increasing. □

Proof of Lemma 7.6. Take $S_1, T_1 \in \mathcal{M}_1$ and $S_2, T_2 \in \mathcal{M}_2$. Take $(x, y) \in \mathbb{R}^{d_1 \times d_2}$, and write $S(x, y) = (S_1(x), S_2(y))$, and similarly for T. Since each of $S_1 \circ T_1^{-1}$ and $S_2 \circ T_2^{-1}$ are gradients of convex functions, then $S \circ T^{-1} = (S_1 \circ T_1^{-1}, S_2 \circ T_2^{-1})$ is also the gradient of a convex (and separable) function. □

Proof of Lemma 7.7. For any $T \in \mathcal{M}$, T and T^{-1} are both gradients of convex functions, so the claim is immediate. □

Proof of Lemma 7.8. Suppose $T_1, T_2 \in \mathcal{M}$ are compatible i.e., $T_1 \circ (T_2)^{-1}$ is the gradient of a convex function. Write $\tilde{T}_1 = \nabla\tilde{\varphi}_1 = \nabla(\varphi_1 + \langle u, \cdot\rangle)$ and $\tilde{T}_2 = \nabla\tilde{\varphi}_2 = \nabla(\varphi_2 + \langle v, \cdot\rangle)$. One can check that $\tilde{\varphi}_2^*(y) = \varphi_2^*(y - v)$, and then by convex duality $(\tilde{T}_2)^{-1} = \nabla\varphi_2^*(\cdot - v)$ is the gradient of a convex function. So,

$$\tilde{T}_1(\tilde{T}_2^{-1}(y)) = \nabla\varphi_1(\nabla\varphi_2^*(y - v)) + u\,,$$

which is the gradient of a sum of convex functions. □

Proof of Lemma 7.9. For $\eta, \lambda \in \mathbb{R}_+^{|\mathcal{M}|}$, write $S^\eta = \sum_{S\in\mathcal{M}} \eta_S S$ and $T^\lambda = \sum_{T\in\mathcal{M}} \lambda_T T$ in $\mathrm{cone}(\mathcal{M})$. Assume $\eta, \lambda \neq 0$ or otherwise the statement is trivial. The composition reads

$$T^\lambda \circ (S^\eta)^{-1} = \textstyle\sum_{T\in\mathcal{M}} \lambda_T T \circ \left(\sum_{S\in\mathcal{M}} \eta_S S\right)^{-1}\,,$$

so it suffices to show that $\tilde{T} \coloneqq T \circ \left(\sum_{S\in\mathcal{M}} \eta_S S\right)^{-1}$ is the gradient of a convex function. Since each $S \in \mathcal{M}$ is the gradient of a convex function, we have that

$$\tilde{T}^{-1} = \Big(\sum_{S\in\mathcal{M}} \eta_S S\Big) \circ T^{-1} = \sum_{S\in\mathcal{M}} \eta_S \,(S \circ T^{-1})\,.$$

Since $\tilde{T}^{-1}$ is the gradient of a convex function, by conjugacy, it holds that $\tilde{T}$ is the gradient of a convex function. □

F.2 Proofs for Section 7.4.2

Proof of Theorem 7.14. For an iteration number $t \in \mathbb{N}$, we use the shorthand $\hat{\nabla}_\lambda \mathcal{F}_t \coloneqq \hat{\nabla}_\lambda \mathcal{F}(\mu_{\lambda^{(t)}})$, and similarly for the true gradient.

Since projections are contractive, a first manipulation gives

$$\|\lambda^{(t+1)} - \lambda^\star\|_Q^2 \le \|\lambda^{(t)} - \lambda^\star\|_Q^2 + h^2 \, \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2 + 2h \, \langle \hat{\nabla}_\lambda \mathcal{F}_t, \lambda^\star - \lambda^{(t)} \rangle \, .$$

Taking expectations conditioned on $\lambda^{(t)}$ yields, by linearity,

$$\mathbb{E}_t \|\lambda^{(t+1)} - \lambda^\star\|_Q^2 \le \|\lambda^{(t)} - \lambda^\star\|_Q^2 + h^2 \, \mathbb{E}_t \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2 + 2h \, \langle \nabla_\lambda \mathcal{F}_t, \lambda^\star - \lambda^{(t)} \rangle \, .$$

By m-strong convexity of $\mathcal{F}$, we obtain

$$\begin{aligned} \mathbb{E}_t \|\lambda^{(t+1)} - \lambda^\star\|_Q^2 &\le (1 - mh) \, \|\lambda^{(t)} - \lambda^\star\|_Q^2 + h^2 \, \mathbb{E}_t \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2 + 2h \, (\mathcal{F}(\mu_\star) - \mathcal{F}(\mu_{\lambda^{(t)}})) \\ &\le (1 - 2mh) \|\lambda^{(t)} - \lambda^\star\|_Q^2 + h^2 \, \mathbb{E}_t \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2 \, . \end{aligned}$$

Taking expectations again,

$$\mathbb{E} \|\lambda^{(t+1)} - \lambda^\star\|_Q^2 \le (1 - 2mh) \, \mathbb{E} \|\lambda^{(t)} - \lambda^\star\|_Q^2 + h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2] \, .$$

Adding and subtracting the true gradient at iterate $\lambda^{(t)}$, written $Q^{-1} \, \nabla_\lambda \mathcal{F}_t$, the second term can be bounded via smoothness of $\mathcal{F}$:

$$\begin{aligned} h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, \hat{\nabla}_\lambda \mathcal{F}_t\|_Q^2] &\le 2h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, (\hat{\nabla}_\lambda \mathcal{F}_t - \nabla_\lambda \mathcal{F}_t)\|_Q^2] + 2h^2 \, \mathbb{E} \|Q^{-1} \, \nabla_\lambda \mathcal{F}_t\|_Q^2 \\ &\le 2h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, (\hat{\nabla}_\lambda \mathcal{F}_t - \nabla_\lambda \mathcal{F}_t)\|_Q^2] + 2M^2 h^2 \, \mathbb{E} \|\lambda^{(t)} - \lambda^\star\|_Q^2 \, . \end{aligned}$$

Combining this with our previous bound results in

$$\begin{aligned} \mathbb{E} \|\lambda^{(t+1)} - \lambda^\star\|_Q^2 &\le (1 - 2mh + 2M^2 h^2) \, \mathbb{E} \|\lambda^{(t)} - \lambda^\star\|_Q^2 + 2h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, (\hat{\nabla}_\lambda \mathcal{F}_t - \nabla_\lambda \mathcal{F}_t)\|_Q^2] \\ &\le (1 - mh) \, \mathbb{E} \|\lambda^{(t)} - \lambda^\star\|_Q^2 + 2h^2 \, \mathbb{E} [\mathbb{E}_t \|Q^{-1} \, (\hat{\nabla}_\lambda \mathcal{F}_t - \nabla_\lambda \mathcal{F}_t)\|_Q^2] \, , \end{aligned}$$

where in the last step we took $h \le \frac{1}{2\kappa M}$.

By (VB), we obtain

$$\mathbb{E}\|\lambda^{(t+1)} - \lambda^\star\|_Q^2 \le (1 - mh + c_1 h^2)\, \mathbb{E}\|\lambda^{(t)} - \lambda^\star\|_Q^2 + c_0 h^2 \,.$$

If $c_1 h^2 \le mh/2$, i.e., $h \le m/(2c_1)$, then

$$\mathbb{E}\|\lambda^{(t+1)} - \lambda^\star\|_Q^2 \le (1 - mh/2)\, \mathbb{E}\|\lambda^{(t)} - \lambda^\star\|_Q^2 + c_0 h^2 \,.$$

Iterating this bound gives

$$\mathbb{E}\|\lambda^{(t)} - \lambda^\star\|_Q^2 \le (1 - \tfrac{mh}{2})^t \, \|\lambda^{(0)} - \lambda^\star\|_Q^2 + \frac{2c_0 h}{m} \le \exp(-mht/2)\, \|\lambda^{(0)} - \lambda^\star\|_Q^2 + \frac{2c_0 h}{m} \,.$$

Choosing $h \asymp m\varepsilon^2/c_0$ concludes the proof. □

F.3 Proofs for Section 7.5

F.3.1 Proofs for Section 7.5.1

To derive the mean-field equations, we recall that the KL divergence is

$$\mathrm{KL}(\mu\|\pi) = \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z \,.$$

Over the space of product measures, we obtain the functional

$$(\mu_1, \dots, \mu_d) \mapsto \mathrm{KL}\Big(\bigotimes_{i=1}^d \mu_i \,\Big\|\, \pi\Big) = \int V \,\mathrm{d} \bigotimes_{i=1}^d \mu_i + \sum_{i=1}^d \int \log \mu_i \,\mathrm{d}\mu_i + \log Z \,.$$

If we take the first variation of this functional (c.f. Santambrogio, 2015, Section 7.2) w.r.t. μ_i, we obtain the equation

$$\left[\delta_{\mu_i} \mathrm{KL}\Big(\bigotimes_{j=1}^{d} \mu_j \;\Big\|\; \pi\Big)\right](x_i) = \int V(x_1, \ldots, x_d) \bigotimes_{j \neq i} \mu_j(\mathrm{d}x_j) + \log \mu_i(x_i) + \text{const}\,.$$

At optimality, the first variation must equal a constant, which leads to

$$\pi_i^\star(x_i) \propto \exp\Big(-\int V(x_1, \ldots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\Big)\,.$$

F.3.2 Proofs for Section 7.5.2

In this section, we prove the regularity bounds on the optimal transport maps given as Theorem 7.21. Recall that $\pi^\star$ denotes the mean-field VI solution and $T^\star$ is the optimal transport map from ρ to $\pi^\star$. Let $\pi_i^\star$ and $T_i^\star$ denote the i-th components respectively, and recall also from (7.14) that $\pi_i^\star \propto \exp(-V_i)$, where

$$V_i(x_i) \coloneqq \int V(x_1, \ldots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\,.$$

We begin with a few simple lemmas which show that $T_i^\star(0)$, the mean of $\pi_i^\star$, and the mode of $\pi_i^\star$ are all close to each other.

Lemma F.1. *Let T denote the optimal transport map from $\rho = \mathcal{N}(0,1)$ to μ, and let m denote the mean of μ. If $T' \leq \beta$, then $|T(0) - m| \leq \sqrt{2/\pi}\,\beta$.*

Proof. Let $Z \sim \mathcal{N}(0,1)$, so that $T(Z) \sim \mu$ and $m \coloneqq \mathbb{E}T(Z)$. Since $T' \leq \beta$,

$$|T(0) - m| = |\mathbb{E}(T(0) - T(Z))| \leq \beta\,\mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}\,\beta\,.$$

□

Lemma F.2. *Let m and $\tilde{m}$ denote the mean and the mode of μ, respectively, where μ is ℓ_V-strongly log concave and univariate. Then, $|m - \tilde{m}| \leqslant 1/\sqrt{\ell_V}$.*

Proof. This is a standard consequence of strong log-concavity, see, e.g., Dalalyan et al. (2022, Proposition 4). □

We are now ready to prove Theorem 7.21.

Proof of Theorem 7.21. As the main text contains the proof of the bounds on the first derivative of T, we continue with the second and third derivative bounds.

We, obviously, start with the second derivative bounds. Recall the Monge–Ampère equation (or the change of variables formula) yields

$$\log \pi_i^\star \circ T_i^\star(x) = -\frac{x^2}{2} - \log(T_i^\star)'(x) - \tfrac{1}{2}\log(2\pi)\,. \tag{F.1}$$

Differentiating once yields

$$(\log \pi_i^\star \circ T_i^\star)'(x) = -V_i'(T_i^\star(x))\,(T_i^\star)'(x) = -x - \frac{(T_i^\star)''(x)}{(T_i^\star)'(x)}\,. \tag{F.2}$$

Rearranging to isolate $(T_i^\star)''$ gives

$$(T_i^\star)''(x) = -(T_i^\star)'(x)\left(x - V_i'(T_i^\star(x))\,(T_i^\star)'(x)\right). \tag{F.3}$$

Let $m_i^\star$ and $\tilde{m}_i^\star$ denote the mean and mode of $\pi_i^\star$ respectively. Recall also that $0 < 1/\sqrt{L_V} \leq$

$(T_i^\star)' \le 1/\sqrt{\ell_V}$. By Lemma F.1 and Lemma F.2,

$$\begin{aligned}
|V_i'(T_i^\star(x))| &\leqslant \underbrace{|V_i'(\tilde{m}_i^\star)|}_{=0} + L_V\,|T_i^\star(x) - \tilde{m}_i^\star| \\
&\leqslant L_V\,(|T_i^\star(x) - T_i^\star(0)| + |T_i^\star(0) - m_i^\star| + |m_i^\star - \tilde{m}_i^\star|) \\
&\leqslant L_V\,\Big(\frac{1}{\sqrt{\ell_V}}\,|x| + \sqrt{\frac{2}{\pi}}\,\frac{1}{\sqrt{\ell_V}} + \frac{1}{\sqrt{\ell_V}}\Big) \lesssim \frac{L_V}{\sqrt{\ell_V}}\,(1+|x|)\,.
\end{aligned}$$

Substituting this into (F.3), we obtain

$$|(T_i^\star)''(x)| \lesssim \frac{1}{\sqrt{\ell_V}}\,\Big(|x| + \frac{L_V}{\ell_V}\,(1+|x|)\Big) \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,(1+|x|)\,.$$

For the third derivative control, we differentiate (F.2) again to yield

$$\begin{aligned}
(\log \pi_i^\star \circ T_i^\star)''(x) &= -\big(V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2 + V_i'(T_i^\star(x))\,(T_i^\star)''(x)\big) \\
&= -1 - \frac{(T_i^\star)'''(x)\,(T_i^\star)'(x) - (T_i^\star)''(x)^2}{(T_i^\star)'(x)^2}\,.
\end{aligned}$$

Again, we rearrange and isolate, giving

$$(T_i^\star)'''(x) = \frac{(T_i^\star)''(x)^2}{(T_i^\star)'(x)} - (T_i^\star)'(x)\,\big(1 - V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2 - V_i'(T_i^\star(x))\,(T_i^\star)''(x)\big)\,.$$

Taking absolute values, we can collect the terms one by one:

$$\begin{aligned}
|(T_i^\star)''(x)^2/(T_i^\star)'(x)| &\lesssim \frac{\kappa^2}{\sqrt{\ell_V}}\,(1+|x|^2)\,, \\
|V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2| &\le \kappa\,, \\
|V_i'(T_i^\star(x))\,(T_i^\star)''(x)| &\lesssim \frac{L_V}{\sqrt{\ell_V}}\,(1+|x|)\cdot\frac{\kappa}{\sqrt{\ell_V}}\,(1+|x|) \lesssim \kappa^2\,(1+|x|^2)\,.
\end{aligned}$$

To obtain the first bound, note that by (F.3) and the subsequent calculations, we have

$$|(T_i^\star)''(x)/(T_i^\star)'(x)| \lesssim \kappa\,(1+|x|)\,.$$

Square and use $(T_i^\star)'(x) \leqslant 1/\sqrt{\ell_V}$. Hence, the final bound scales as

$$|(T_i^\star)'''(x)| \lesssim \frac{\kappa^2}{\sqrt{\ell_V}}\,(1+|x|^2)\,.$$

□

F.3.3 Proofs for Section 7.5.3

For our approximation results, we begin with a simple construction via piecewise linear maps. Let $R > 0$ denote a truncation parameter, and partition the interval $[-R, +R]$ into sub-intervals of length $\delta > 0$. Let ψ be the elementary step function

$$\psi : \mathbb{R} \to \mathbb{R}, \qquad \psi(x) \coloneqq \begin{cases} 0\,, & x \leqslant 0\,, \\ x\,, & x \in [0,1]\,, \\ 1\,, & x \geqslant 1\,. \end{cases}$$

We then define the following family of compatible maps:

$$\mathcal{M} \coloneqq \left\{ x \mapsto \psi(\delta^{-1}\,(x_i - a))\,e_i \;\middle|\; i \in [d],\ I = [a, a+\delta] \text{ is a sub-interval} \right\}.$$

We suppress the dependence on the parameters R, δ in the notation.

Proof of Theorem 7.23. Owing to the isometry, we wish to show that we can find a map $\hat{T} \in$

$\underline{\text{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})$, with $\alpha = 1/\sqrt{L_V}$, such that

$$\|\bar{T} - \hat{T}\|^2_{L^2(\rho)} \leqslant \varepsilon^2/\ell_V \qquad \text{and} \qquad \|D(\bar{T} - \hat{T})\|^2_{L^2(\rho)} \leqslant \varepsilon_1^2/\ell_V\,. \tag{F.4}$$

Here, $\|D(\bar{T} - \hat{T})\|^2_{L^2(\rho)} \coloneqq \int \|D(\bar{T} - \hat{T})\|^2_{\mathrm{F}}\,\mathrm{d}\rho$.

We first make a series of reductions. By assumption, $D\bar{T} \succeq \alpha I$, and by definition, $\hat{T}$ is of the form $\alpha\,\mathrm{id} + \sum_{T\in\mathcal{M}} \lambda_T T + v$. By replacing $\bar{T}$ with $\bar{T} - \alpha\,\mathrm{id}$, it suffices to prove the following statement: assuming that $0 \preceq D\bar{T} \preceq \ell_V^{-1/2}\, I$ together with the second derivative bound on $\bar{T}$, there exists $\hat{T}$ of the form $\sum_{T\in\mathcal{M}} \lambda_T T + v$ such that (F.4) holds. However, from the structure of $\mathcal{M}$, the problem now separates across the coordinates and it suffices to prove this statement with $d = 1$ and ε replaced with $\varepsilon/\sqrt{d}$.

Truncation procedure. We will construct $\hat{T}$ so that $\bar{T}(-R) = \hat{T}(-R)$ and $\bar{T}(+R) = \hat{T}(+R)$. Assuming that this holds, the bound on $\bar{T}'$ and the fact that $\hat{T}$ is constant on $(-\infty, -R]$ and on $[+R, +\infty)$ readily imply

$$|\bar{T}(x) - \hat{T}(x)| \leqslant \frac{1}{\sqrt{\ell_V}}\,(|x| - R)\,, \qquad \text{for } |x| \geqslant R\,.$$

The error contributed by the tails is therefore bounded by

$$\int_{\mathbb{R}\setminus(-R,+R)} |\bar{T} - \hat{T}|^2\,\mathrm{d}\rho \leqslant \frac{1}{\sqrt{2\pi}\,\ell_V} \int_{\mathbb{R}\setminus(-R,+R)} (|x| - R)^2 \exp(-x^2/2)\,\mathrm{d}x\,.$$

Similarly,

$$|\bar{T}'(x) - \hat{T}'(x)| \leqslant 1/\sqrt{\ell_V}\,, \qquad \text{for } |x| \geqslant R\,,$$

which gives

$$\int_{\mathbb{R}\setminus(-R,+R)} |\bar{T}' - \hat{T}'|^2 \,\mathrm{d}\rho \leqslant \frac{1}{\sqrt{2\pi}\,\ell_V} \int_{\mathbb{R}\setminus(-R,+R)} \exp(-x^2/2)\,\mathrm{d}x\,.$$

Standard Gaussian tail bounds and the Cauchy–Schwarz inequality imply that with the choice $R \asymp \sqrt{\log(1/(\ell_V \varepsilon^2))}$, we obtain $\|\bar{T} - \hat{T}\|^2_{L^2(\tilde{\rho})} \vee \|\bar{T}' - \hat{T}'\|^2_{L^2(\tilde{\rho})} \lesssim \varepsilon^2$, where $\tilde{\rho}$ is the Gaussian measure restricted to the set $\mathbb{R} \setminus [-R, R]$.

Uniform approximation over a compact domain. We now show that $\hat{T}$ can be chosen to uniformly approximate $\bar{T}$ on $[-R, +R]$. Indeed, we take

$$\hat{T}(x) = \bar{T}(-R) + \sum_{m=0}^{2R/\delta - 1} \lambda_m \psi\Big(\frac{x - (-R + m\delta)}{\delta}\Big),$$

where the λ_m are chosen so that $\bar{T}$ and $\hat{T}$ agree at each of the endpoints of the sub-intervals of size δ. Consider such a sub-interval $I = [a, a + \delta]$. Then, for $x \in I$,

$$|\bar{T}(x) - \hat{T}(x)| = \Big|\bar{T}(x) - \bar{T}(a) - \frac{\bar{T}(a+\delta) - \bar{T}(a)}{\delta}\,(x - a)\Big|.$$

By the mean value theorem, $\bar{T}(x) = \bar{T}(a) + \bar{T}'(c_1)\,(x - a)$ and $\bar{T}(a + \delta) = \bar{T}(a) + \bar{T}'(c_2)\,\delta$ for some $c_1, c_2 \in I$. Together with the second derivative bound on $\bar{T}$, it yields

$$|\bar{T}(x) - \hat{T}(x)| = |(\bar{T}'(c_1) - \bar{T}'(c_2))\,(x - a)| \lesssim \frac{\kappa R}{\sqrt{\ell_V}}\,\delta^2\,.$$

Similarly, for the derivative,

$$|\bar{T}'(x) - \hat{T}'(x)| = \Big|\bar{T}'(x) - \frac{\bar{T}(a+\delta) - \bar{T}(a)}{\delta}\Big| = |\bar{T}'(x) - \bar{T}'(c_2)| \lesssim \frac{\kappa R}{\sqrt{\ell_V}}\,\delta\,.$$

To obtain our desired error bounds, we take $\delta = \widetilde{\Theta}(\sqrt{\varepsilon/\kappa})$. Finally, to obtain the stated bounds in

the theorem in dimension d, replace ε with $\varepsilon/\sqrt{d}$.

With this choice of δ, we then obtain $\varepsilon_1 = \widetilde{O}(\sqrt{d}\,\kappa\delta) = \widetilde{O}(\kappa^{1/2} d^{1/4} \varepsilon^{1/2})$.

Size of the generating family. Finally, the size of $\mathcal{M}$ is $O(dR/\delta) = \widetilde{O}(\kappa^{1/2} d^{5/4}/\varepsilon^{1/2})$, which completes the proof. □

In the proof above, we have used the bounds on the first and second derivatives of $\bar{T}$. However, from Theorem 7.21, we actually have control on the third derivative as well, so we can expect to exploit this added degree of smoothness to obtain better approximation rates.

As above, we fix a truncation parameter $R > 0$ and a mesh size $\delta > 0$. Our family of maps will be constructed from the following basic building blocks.

- **Linear function.** We let $\psi^{\mathrm{lin}}(x) := x$ for $x \in \mathbb{R}$.
- **Piecewise quadratics.** Define the piecewise quadratic

$$\psi^{\mathrm{quad},\pm}(x) := \pm \begin{cases} 0\,, & x \leqslant 0\,, \\ x^2\,, & x \in [0,1]\,, \\ 2x - 1\,, & x \geqslant 1\,. \end{cases}$$

- **Piecewise cubics.** Define the piecewise cubic,

$$\psi^{\mathrm{cub},\pm}(x) := \pm \begin{cases} 0\,, & x \leqslant 0\,, \\ x^2\,(3 - 2x)\,, & x \in [0,1]\,, \\ 1\,, & x \geqslant 1\,. \end{cases}$$

Given a univariate function ψ and $i \in [d]$, we extend it to a map $\psi_i : \mathbb{R}^d \to \mathbb{R}^d$ by setting $\psi_i(x) := \psi(x_i)$. Also, given a sub-interval $I = [a, a+\delta]$, we define the map $\psi_{I,i} : \mathbb{R}^d \to \mathbb{R}^d$ via $\psi_{I,i}(x) := \psi(\delta^{-1}\,(x_i - a))$.

Let $\mathcal{I}$ denote the set of sub-intervals. Our generating family will consist of

$$\mathcal{M} \coloneqq \{\psi_i^{\mathrm{lin}} \mid i \in [d]\} \cup \bigcup_{I \in \mathcal{I}} \bigcup_{i=1}^{d} \{\psi_{I,i}^{\mathrm{quad},-}, \psi_{I,i}^{\mathrm{quad},+}, \psi_{I,i}^{\mathrm{cub},-}, \psi_{I,i}^{\mathrm{cub},+}\},$$

which consists of $(4\,|\mathcal{I}| + 1)\,d$ elements. However, we will not consider the full cone generated by $\mathcal{M}$—indeed, if we did, then the presence of the *negative* piecewise quadratics and cubics would mean that we obtain non-monotone maps.

Elements of our polyhedral set will be of the form $x \mapsto \alpha\,\mathrm{id} + \sum_{T \in \mathcal{M}} \lambda_T T + v$, where $v \in \mathbb{R}^d$ and we may decorate components of λ according to the elements of $\mathcal{M}$ to which they correspond, e.g., $\lambda_{I,i}^{\mathrm{quad},-}$ is the coefficient in front of $\psi_{I,i}^{\mathrm{quad},-}$.

To provide some intuition, we will use the linear function and the piecewise quadratics to approximate the *derivative* of $\bar{T}$. Indeed, suppose for the moment that $\bar{T}$ is univariate and note that the derivatives of the linear and piecewise quadratic functions give rise to piecewise linear interpolations of $\bar{T}'$. The interpolation of $\bar{T}'$, once integrated, does not necessarily interpolate $\bar{T}$, and the piecewise cubics will be used to remedy this issue.

Toward this end, note that since $\bar{T}$ is monotonically increasing, $\bar{T}'$ is non-negative. We will want our approximating $\hat{T}$ to have the same property, which will be ensured by imposing *linear constraints* on λ. We consider the following polyhedral subset of $\mathbb{R}_+^{|\mathcal{M}|}$:

$$\begin{aligned} K \coloneqq \Big\{ \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \;\Big|\; & \forall i \in [d],\ \frac{2}{\delta} \sum_{I \in \mathcal{I}} (\lambda_{I,i}^{\mathrm{quad},+} - \lambda_{I,i}^{\mathrm{quad},-}) + \lambda_i^{\mathrm{lin}} \geqslant 0\,, \\ & \text{and} \qquad \forall I \in \mathcal{I},\ \forall i \in [d],\ \frac{6\lambda^{\mathrm{cub},-}}{\delta} \leqslant \frac{\alpha}{2} \Big\}. \end{aligned} \tag{F.5}$$

We then take $\mathcal{K} \coloneqq \{x \mapsto \alpha x + \sum_{T \in T} \lambda_T T + v \mid \lambda \in K,\ v \in \mathbb{R}^d\}$ and $\mathcal{P}_\diamond \coloneqq \mathcal{K}_\sharp \rho$. The first constraint ensures that the sum of the linear and piecewise quadratic functions has non-negative slope. As for the second constraint, it ensures that the sum of the negative piecewise cubic functions has slope at least $-\alpha/2$. Since we always add α id, each of our maps will have slope at least $\alpha/2$ and therefore

be increasing. With this choice, our family consists of gradients of strongly convex functions with convexity parameter *less* than that of the true map $T^\star$, which does affect some of the other results of this paper (e.g., the geodesic smoothness of the KL divergence in Proposition 7.28), but only by at most a constant factor, and henceforth we ignore this technical issue.

We are now ready to prove our improved approximation result.

Proof of Theorem 7.24. We start with the same reductions as in the proof of Theorem 7.23, reducing to the univariate case.

Truncation procedure. The truncation procedure is similar to the one before, except that $\hat{T}$ is no longer constant on $(-\infty, -R]$ and on $[+R, +\infty)$. Instead, on these intervals, $\hat{T}$ will be linear, with the additional conditions $\bar{T}'(-R) = \hat{T}'(-R)$ and $\bar{T}'(+R) = \hat{T}'(+R)$. However, the arguments still go through, and we can take $R \asymp \sqrt{\log(1/(\ell_V \varepsilon^2))}$ as before.

Uniform approximation over a compact domain. We will first construct a preliminary version of $\hat{T}$ without using the piecewise cubics. Recall from the discussion above that using the linear and piecewise quadratic functions, we can ensure that $\hat{T}'$ is a linear interpolation of $\bar{T}'$. Namely, we set

$$\hat{T}' = \bar{T}'(-R) + \sum_{I \in \mathcal{I}} \left[\lambda_I^{\mathrm{quad},-} \, (\psi^{\mathrm{quad},-})' + \lambda_I^{\mathrm{quad},+} \, (\psi^{\mathrm{quad},+})' \right],$$

where the coefficients are chosen such that $\bar{T}'$ and $\hat{T}'$ agree at each of the endpoints of the sub-intervals. Following the argument as before, for a sub-interval $I = [a, a + \delta]$ and $x \in I$,

$$|\bar{T}'(x) - \hat{T}'(x)| = \left| \bar{T}'(x) - \bar{T}'(a) - \frac{\bar{T}'(a+\delta) - \bar{T}'(a)}{\delta} \, (x - a) \right|.$$

By the mean value theorem, $\bar{T}'(x) = \bar{T}'(a) + \bar{T}''(c_1) \, (x - a)$ and $\bar{T}'(a + \delta) = \bar{T}'(a) + \bar{T}''(c_2) \, \delta$ for

some $c_1, c_2 \in I$. Using the bounds on the derivatives of $\bar{T}$,

$$|\bar{T}'(x) - \hat{T}'(x)| = |(\bar{T}''(c_1) - \bar{T}''(c_2))\,(x-a)| \lesssim \frac{\kappa^2 R^2}{\sqrt{\ell_V}}\,\delta^2\,. \tag{F.6}$$

Next, we wish to control $|\bar{T}(x) - \hat{T}(x)|$. Here, $\hat{T}$ is defined by integrating $\hat{T}'$, and choosing the shift v so that $\bar{T}(-R) = \hat{T}(-R)$. First, *suppose* that $\bar{T}(a) = \hat{T}(a)$. We can then use the fundamental theorem of calculus to obtain

$$|\bar{T}(x) - \hat{T}(x)| = \left|\int_a^x (\bar{T}'(y) - \hat{T}'(y))\,\mathrm{d}y\right| \lesssim \frac{\kappa^2 R^2}{\sqrt{\ell_V}}\,\delta^3\,. \tag{F.7}$$

In particular, $|\bar{T}(a+\delta) - \hat{T}(a+\delta)|$ is of order δ^3.

To ensure that $\bar{T}$ and $\hat{T}$ agree at each of these endpoints, we scan the set of sub-intervals left to right, and we iteratively add non-negative multiples of the piecewise cubics in order to achieve this interpolating condition. Since the original endpoint error is bounded in (F.7), it follows that the coefficients of the piecewise cubics that we add are small: $0 \leqslant \lambda_I^{\mathrm{cub},\pm} \lesssim \kappa^2 R^2 \delta^3/\sqrt{\ell_V}$. In particular, the constraint on $\lambda_I^{\mathrm{cub},-}$ in (F.5) is met for small δ.

The key property of the piecewise cubics is that $(\psi^{\mathrm{cub},\pm})'(0) = (\psi^{\mathrm{cub},\pm})'(1) = 0$. This means that even after adding the piecewise cubics, $\bar{T}'$ and $\hat{T}'$ agree at all of the endpoints of the sub-intervals. However, we must check that adding these piecewise cubics does not destroy the approximation rates (F.6) and (F.7). Since $|(\psi_I^{\mathrm{cub},\pm})'| \lesssim 1/\delta$, the bound on the coefficients for the piecewise cubics shows that the derivative of the piecewise cubic part of $\hat{T}$ is bounded in magnitude by $O(\kappa^2 R^2 \delta^2/\sqrt{\ell_V})$, so that (F.6) is intact. Similarly, (F.7) is also intact, either by integrating (F.6) or by using the bound on the coefficients of the piecewise cubics. Thus, $|\bar{T}(x) - \hat{T}(x)| \lesssim \kappa^2 R^2 \delta^3/\sqrt{\ell_V}$, and setting this to be at most $\varepsilon/\sqrt{d\ell_V}$ yields the choice $\delta = \widetilde{\Theta}(\varepsilon^{1/3}/(\kappa^{2/3} d^{1/6}))$.

Size of the generating family. The size of the generating family is then $O(dR/\delta) = \widetilde{O}(\kappa^{2/3} d^{7/6}/\varepsilon^{1/3})$, which completes the proof. □

Using the bounds on the Jacobian $D\hat{T}_\diamond$ of the approximating map, we can bound the change in the KL divergence on the path from $\hat{\pi}_\diamond$ to $\pi^\star$. This shows that $\hat{\pi}_\diamond$ has a small *suboptimality gap* for KL minimization over $\mathcal{P}_\diamond$. The following calculation is similar to the one for Proposition 7.28, which establishes smoothness of the KL divergence over $\mathcal{P}_\diamond$. However, since $\pi^\star$ does not lie in $\mathcal{P}_\diamond$, it does not apply here.

Corollary F.3. *Assume that π is well-conditioned (WC). Let $\hat{\pi}_\diamond = (\hat{T}_\diamond)_\sharp \rho$ denote the approximation to $\pi^\star$ given by the piecewise linear construction (Theorem 7.23). Then,*

$$\mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi_\diamond^\star \| \pi) \leqslant \mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi^\star \| \pi) \lesssim \kappa^3 d^{1/2} \varepsilon \,.$$

If, on the other hand, $\hat{\pi}_\diamond = (\hat{T}_\diamond)_\sharp \rho$ is given by the construction of Theorem 7.24,

$$\mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi_\diamond^\star \| \pi) \leqslant \mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi^\star \| \pi) \lesssim \kappa^{10/3} d^{1/3} \varepsilon^{4/3} \,.$$

Proof. Let $(\mu_t)_{t \in [0,1]}$ denote the geodesic joining $\pi^\star$ to $\hat{\pi}_\diamond$. Then, by differentiating the KL divergence along this geodesic twice, we obtain the following expressions; see Chewi (2024) and Diao et al. (2023, Appendix B.2) for derivations. We write $T = \hat{T}_\diamond \circ (T^\star)^{-1}$ for the optimal transport map from $\pi^\star$ to $\hat{\pi}_\diamond$, and $T_t = (1-t)\,\mathrm{id} + t\,T$.

For the potential energy term,

$$\partial_t^2 \mathcal{V}(\mu_t) = \mathbb{E}_{\pi^\star} \langle T - \mathrm{id}, (\nabla^2 V \circ T_t)\,(T - \mathrm{id}) \rangle \leqslant L_V\, \mathbb{E}_{\pi^\star} \|T - \mathrm{id}\|^2 = L_V\, \mathbb{E}_\rho \|\hat{T}_\diamond - T^\star\|^2 \,.$$

Next, for the entropy term,

$$\partial_t^2 \mathcal{H}(\mu_t) = \mathbb{E}_{\pi^\star} \|(DT_t)^{-1}\,(DT - I)\|_{\mathrm{F}}^2 \,,$$

By Theorem 7.21, $DT = D((T^\star)^{-1})\, D\hat{T}_\diamond \geq 1/\sqrt{\kappa}$, so $DT_t \geq 1/\sqrt{\kappa}$. Also, $DT^\star \geq 1/\sqrt{L_V}$. Therefore,

we obtain

$$\begin{aligned}\partial_t^2 \mathcal{H}(\mu_t) &\leqslant \kappa\, \mathbb{E}_{\pi^\star} \| D((T^\star)^{-1})\, D\hat{T}_\diamond \circ (T^\star)^{-1} - I \|_{\mathrm{F}}^2 \\ &\leqslant \kappa L_V\, \mathbb{E}_{\pi^\star} \| D\hat{T}_\diamond \circ (T^\star)^{-1} - D((T^\star)^{-1}) \|_{\mathrm{F}}^2 = \kappa L_V\, \mathbb{E}_\rho \| D\hat{T}_\diamond - DT^\star \|_{\mathrm{F}}^2 \,.\end{aligned}$$

Therefore, adding the two terms together,

$$\partial_t^2 \mathrm{KL}(\mu_t \| \pi) \leqslant L_V\, \| \hat{T}_\diamond - T^\star \|_{L^2(\rho)}^2 + \kappa L_V\, \| D(\hat{T}_\diamond - T^\star) \|_{L^2(\rho)}^2 \,.$$

Integrating this expression from $t = 0$ to $t = 1$,

$$\begin{aligned}\mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi^\star \| \pi) \leqslant\ & \mathbb{E}_{\pi^\star} \langle [\nabla_{\mathbb{W}} \mathrm{KL}(\cdot \| \pi)](\pi^\star), T - \mathrm{id} \rangle \\ & + \frac{L_V}{2} \left(\| \hat{T}_\diamond - T^\star \|_{L^2(\rho)}^2 + \kappa\, \| D\hat{T}_\diamond - DT^\star \|_{L^2(\rho)}^2 \right) .\end{aligned}$$

However, since $\hat{\pi}_\diamond$, $\pi^\star$ both belong to the geodesically convex set of product measures, and $\pi^\star$ minimizes the KL divergence over this set, we must have $\mathbb{E}_{\pi^\star} \langle [\nabla_{\mathbb{W}} \mathrm{KL}(\cdot \| \pi)](\pi^\star), T - \mathrm{id} \rangle = 0$.

We are now in a position to apply the approximation guarantees. Applying the result of Theorem 7.23, we obtain

$$\mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi^\star \| \pi) \lesssim \kappa \varepsilon^2 + \kappa^3 d^{1/2} \varepsilon \,.$$

If we instead use the improved guarantee of Theorem 7.24, we obtain

$$\mathrm{KL}(\hat{\pi}_\diamond \| \pi) - \mathrm{KL}(\pi^\star \| \pi) \lesssim \kappa \varepsilon^2 + \kappa^{10/3} d^{1/3} \varepsilon^{4/3} \,.$$

□

Finally, from the small suboptimality gap of $\hat{\pi}_\diamond$ and the strong geodesic convexity of the

KL divergence, we are able to prove that $\pi^\star$ is close, not just to our constructed $\hat{\pi}_\diamond$, but to the minimizer $\pi_\diamond^\star$ of the KL divergence over $\mathcal{P}_\diamond$, which in turn can be computed via the algorithms in Section 7.5.4.

Proof of Theorem 7.26. By triangle inequality, we have

$$W_2(\pi_\diamond^\star, \pi^\star) \le W_2(\pi_\diamond^\star, \hat{\pi}_\diamond) + W_2(\hat{\pi}_\diamond, \pi^\star)\,,$$

and since we can control the second term (recall Theorem 7.23), it suffices to control the first. Since $\mathrm{KL}(\cdot\|\pi)$ is ℓ_V-strongly geodesically convex, the first term can be bounded above by

$$\ell_V W_2^2(\pi_\diamond^\star, \hat{\pi}_\diamond)/2 \le \mathrm{KL}(\hat{\pi}_\diamond\|\pi) - \mathrm{KL}(\pi_\diamond^\star\|\pi) \lesssim \kappa^3 d^{1/2}\tilde{\varepsilon}\,,$$

where the final bound is obtained from Corollary F.3 (we only take the worst-case scaling term), and $\tilde{\varepsilon}$ is the approximation accuracy guaranteed by Theorem 7.23. Setting this equal to ε^2, we apply Theorem 7.23 with $\frac{\varepsilon^2}{\kappa^3 d^{1/2}}$ replacing ε and we see that $|\mathcal{M}| = \widetilde{O}(\kappa^2 d^{3/2}/\varepsilon)$.

Similarly, for the higher-order approximation scheme, we use Corollary F.3 and apply Theorem 7.24 with $\frac{\varepsilon^{3/2}}{\kappa^{5/2} d^{1/4}}$ replacing ε, obtaining $|\mathcal{M}| = \widetilde{O}(\kappa^{3/2} d^{5/4}/\varepsilon^{1/2})$. □

F.3.4 Proofs for Section 7.5.4

Proof of Proposition 7.28. We write

$$\mathrm{KL}(\mu\|\pi) = \mathcal{V}(\mu) + \mathcal{H}(\mu) := \int V\,\mathrm{d}\mu + \int \log\mu\,\mathrm{d}\mu + \log Z\,.$$

To prove smoothness, it suffices to show that the Wasserstein Hessians for both $\mathcal{V}$ and $\mathcal{H}$ are bounded. Since we work with the augmented cone, we let

$$T^{\lambda,v} := \alpha\,\mathrm{id} + \sum_{T\in\mathcal{M}} \lambda_T T + v\,, \qquad \mu_{\lambda,v} := (T^{\lambda,v})_\sharp \rho\,.$$

Our goal is to upper bound the following quadratic forms

$$\begin{aligned}
\nabla^2_{\mathbb{W}}\mathcal{V}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] &= \mathbb{E}_{\mu_{\lambda,v}}[(T^{\eta,u}_{\lambda,v} - \mathrm{id})^\top \nabla^2 V\,(T^{\eta,u}_{\lambda,v} - \mathrm{id})]\,,\\
\nabla^2_{\mathbb{W}}\mathcal{H}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] &= \mathbb{E}_{\mu_{\lambda,v}}\|DT^{\eta,u}_{\lambda,v} - I\|^2_{\mathrm{F}}\,,
\end{aligned}$$

in terms of the squared Wasserstein distance between $\mu_{\lambda,v}$ and $\mu_{\eta,u}$, and $T^{\eta,u}_{\lambda,v}$ is the optimal transport map from $\mu_{\lambda,v}$ to $\mu_{\eta,u}$. See Chewi (2024) and Diao et al. (2023, Appendix B.2) for derivations of these expressions. We bound the two terms separately.

An upper bound on the potential term is straightforward. By (WC), $\nabla^2 V \preceq L_V I$, and so

$$\begin{aligned}
\nabla^2_{\mathbb{W}}\mathcal{V}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] &= \mathbb{E}_{\mu_{\lambda,v}}[(T^{\eta,u}_{\lambda,v} - \mathrm{id})^\top \nabla^2 V\,(T^{\eta,u}_{\lambda,v} - \mathrm{id})]\\
&\le L_V\,\mathbb{E}_{\mu_{\lambda,v}}\|T^{\eta,u}_{\lambda,v} - \mathrm{id}\|^2 = L_V\,W_2^2(\mu_{\lambda,v}, \mu_{\eta,u})\,.
\end{aligned}$$

The entropy term needs a bit more work. To start, we note that by compatibility,

$$T^{\eta,u}_{\lambda,v} = T^{\eta,u} \circ (T^{\lambda,v})^{-1} = T^\eta \circ (T^\lambda)^{-1}(\cdot - v) + u\,, \tag{F.8}$$

where we write $T^{\lambda,v} = T^\lambda + v$ and similarly $T^{\eta,u} = T^\eta + u$. By the chain rule,

$$DT^{\eta,u}_{\lambda,v}(\cdot) = [DT^\eta \circ (T^\lambda)^{-1}(\cdot - v)]\,D[(T^\lambda)^{-1}](\cdot - v)\,.$$

(For simplicity, the reader may wish to first read the following calculations setting $u = v = 0$.)

Performing the appropriate change of variables, the Wasserstein Hessian of $\mathcal{H}$ reads

$$\begin{aligned}
\mathbb{E}_{\mu_{\lambda,v}} \|DT^{\eta,u}_{\lambda,v} - I\|_{\mathrm{F}}^2 &= \mathbb{E}_{\mu_{\lambda,v}} \|\left[DT^{\eta} \circ (T^{\lambda})^{-1}(\cdot - v)\right] D[(T^{\lambda})^{-1}](\cdot - v) - I\|_{\mathrm{F}}^2 \\
&= \mathbb{E}_{\rho} \|DT^{\eta}\, D[(T^{\lambda})^{-1}] \circ (T^{\lambda,v} - v) - I\|_{\mathrm{F}}^2 \\
&= \mathbb{E}_{\rho} \|DT^{\eta}\, D[(T^{\lambda})^{-1}] \circ T^{\lambda} - I\|_{\mathrm{F}}^2 \\
&= \mathbb{E}_{\rho} \|DT^{\eta}\, (DT^{\lambda})^{-1} - I\|_{\mathrm{F}}^2 ,
\end{aligned}$$

where we invoked the inverse function theorem in the last step. Given our set of maps, we know that for any $\lambda \in \mathbb{R}_+^{|\mathcal{M}|}$, $DT^{\lambda} \succeq \alpha I$, and since $DT^{\lambda}\,(DT^{\lambda})^{-1} = I$, we obtain

$$\mathbb{E}_{\mu_{\lambda,v}} \|DT^{\eta,u}_{\lambda,v} - I\|_{\mathrm{F}}^2 \leq \frac{1}{\alpha^2}\, \mathbb{E}_{\rho} \|DT^{\eta} - DT^{\lambda}\|_{\mathrm{F}}^2 .$$

Since our maps are regular (i.e., (ϒ) holds), there exists $\Upsilon > 0$ such that

$$\mathbb{E}_{\rho} \|DT^{\eta} - DT^{\lambda}\|_{\mathrm{F}}^2 = \mathbb{E}_{\rho} \left\| \sum_{T \in \mathcal{M}} (\lambda_T - \eta_T)\, DT \right\|_{\mathrm{F}}^2 = \langle \eta - \lambda, Q^{(1)}\,(\eta - \lambda) \rangle \leqslant \Upsilon \,\langle \eta - \lambda, Q\,(\eta - \lambda) \rangle .$$

Finally, note that

$$\begin{aligned}
W_2^2(\mu_{\lambda,v}, \mu_{\eta,u}) = \mathbb{E}_{\rho} \left\| \sum_{T \in \mathcal{M}} (\eta_T - \lambda_T)\, T + u - v \right\|^2 &= \mathbb{E}_{\rho} \left\| \sum_{T \in \mathcal{M}} (\eta_T - \lambda_T)\, T \right\|^2 + \|u - v\|^2 \\
&= \langle \eta - \lambda, Q\,(\eta - \lambda) \rangle + \|u - v\|^2 ,
\end{aligned}$$

where we used the fact that the maps in $\mathcal{M}$ are *centered*. This shows that

$$\nabla^2_{\mathbb{W}} \mathcal{H}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] \leqslant \frac{\Upsilon}{\alpha^2}\, W_2^2(\mu_{\lambda,v}, \mu_{\eta,u}) .$$

Combining all of the terms completes the proof. □

Proof of Lemma 7.30. We restrict our attention to the piecewise linear family denoted $\mathcal{M}$ in dimension one with $|\mathcal{M}| = J$. This suffices due to the tensorization property of Υ, see the remark after the definition of Υ. It suffices to prove, for all $\lambda \in \mathbb{R}^J$,

$$\Big\| \sum_{T \in \mathcal{M}} \lambda_T T' \Big\|_{L^2(\rho)}^2 \le \Upsilon \Big\| \sum_{T \in \mathcal{M}} \lambda_T T \Big\|_{L^2(\rho)}^2 ,$$

where $\rho = \mathcal{N}(0,1)$. We truncate the domain of ρ to $[-R, R]$, where $R \asymp \sqrt{\log(1/(\ell_V \varepsilon^2))}$. On some interval $[a, a+\delta]$, note that

$$T^\lambda(x) = T^\lambda(a) + \lambda_T \left((x-a)/\delta \right)_+ , \quad DT^\lambda(x) = \lambda_T/\delta .$$

It thus suffices to prove the statement on such an interval. This is equivalent to proving that

$$\int_a^{a+\delta} \Big(\frac{\lambda_T}{\delta} \Big)^2 \rho(\mathrm{d}x) \le \Upsilon \int_a^{a+\delta} \Big(T^\lambda(a) + \lambda_T \, \frac{x-a}{\delta} \Big)^2 \rho(\mathrm{d}x) .$$

Rearranging, it suffices to show that

$$\delta^{-2} \Upsilon^{-1} \le \frac{\inf_{m \in \mathbb{R}} \int_a^{a+\delta} \big(\frac{x-a}{\delta} - m \big)^2 \rho(\mathrm{d}x)}{\int_a^{a+\delta} \rho(\mathrm{d}x)} = \frac{\mathrm{var} X}{\delta^2} ,$$

or $\Upsilon^{-1} \le \mathrm{var} X$, with $X \sim \rho|_{[a,a+\delta]}$.

Letting $m_a \coloneqq \mathbb{E}X$, suppose WLOG $m_a \le a + \delta/2$. We compute

$$\mathbb{E}[(X - m_a)^2] \ge \mathbb{E}[(\delta/4)^2 \, \mathbf{1}_{X \ge a+3\delta/4}] \gtrsim \delta^2 \, \mathbb{P}(X \ge a + 3\delta/4) = \delta^2 \, \frac{\int_{a+3\delta/4}^{a+\delta} \rho(\mathrm{d}x)}{\int_a^{a+\delta} \rho(\mathrm{d}x)} \gtrsim \delta^2 ,$$

provided $\delta \lesssim 1/R$; indeed, for this choice of δ, $|\log \rho(x) - \log \rho(y)| \lesssim 1$ for all $x, y \in [a, a+\delta]$. Stringing together the inequalities, we obtain the desired claim. □

F.3.5 Proofs for Section 7.5.5.2

In this section, we prove our variance bounds for SPGD for mean-field VI. We start with a gradient bound under $\pi^\star$.

Lemma F.4. *Let π be a (WC) measure, and let $\pi^\star$ be the mean-field approximation. Then*

$$\mathbb{E}_{\pi^\star} \nabla V = 0\,, \qquad \mathbb{E}_{\pi^\star} \|\nabla V\|^2 \leqslant L_V \kappa d\,. \tag{F.9}$$

Proof. Recall our definition of $\pi^\star$ with components $\pi_i \propto \exp(-V_i)$ with

$$V_i(x_i) \int_{\mathbb{R}^{d-1}} V(x_1, \dots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)$$

Assuming the first claim, we can prove the second by applying the Brascamp–Lieb inequality (Brascamp and Lieb, 1976):

$$\mathbb{E}_{\pi^\star} \|\nabla V - \mathbb{E}_{\pi^\star} \nabla V\|^2 \leq \mathbb{E}_{\pi^\star} \operatorname{tr}((\nabla^2 V)^2 \operatorname{diag}(\vec{V}'')^{-1})\,,$$

where $\vec{V}'' \coloneqq (V_1'', \dots, V_d'')$. By Proposition 7.19, each component satisfies the bound $(V_i'')^{-1} \leq 1/\ell_V$, and we also have by assumption $\nabla^2 V \preceq L_V I$. Together, the bound is clear:

$$\mathbb{E}_{\pi^\star} \|\nabla V\|^2 \leq \operatorname{tr}((L_V I)^2)/\ell_V = L_V \kappa d\,.$$

It remains to prove the first equality. Recall that for $i \in [d]$,

$$V_i(x_i) = \int V(x) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\,.$$

Consider a test vector $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^d$. Appropriately interchanging the order of integration, one can check that we obtain

$$\mathbb{E}_{\pi^\star} \nabla V^\top e_1 = \int V_1'(x_1)\, \pi_1^\star(\mathrm{d}x_1) = \int V_1'(x) \, \frac{\exp(-V_1(x_1))}{\int \exp(-V_1(x_1'))\, \mathrm{d}x_1'} \, \mathrm{d}x_1 = 0\,,$$

by an application of integration by parts. The same is true for the other coordinates. □

Proof of Lemma 7.33. We want to bound the quantity

$$\mathbb{E}[\|Q^{-1}\,(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda))\|_Q^2] = \mathbb{E}[\|Q^{-1/2}\,(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda))\|^2]\,.$$

Using convenient notation choices, we first recall the expressions of the stochastic and non-stochastic gradients of the potential energy:

$$\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) = \boldsymbol{T}(\hat{X})\, \nabla V(T^\lambda(\hat{X}))\,, \qquad \nabla_\lambda \mathcal{V}(\mu_\lambda) = \mathbb{E}_\rho[\boldsymbol{T}\, \nabla V \circ T^\lambda]\,,$$

where $\hat{X} \sim \rho$ is a random draw, and $\boldsymbol{T}(\hat{X}) = (T_1(\hat{X}), \dots, T_{|\mathcal{M}|}(\hat{X})) \in \mathbb{R}^{|\mathcal{M}|} \times \mathbb{R}^d$ is the evaluation of the *whole* dictionary at the random draw.

We begin by exploiting symmetry in the problem, reducing it to one dimension. First, note that $\boldsymbol{T}$ can be equivalently expressed as d repetitions of the following vectors,

$$\boldsymbol{T} = (T_{1:J}, \dots, T_{1:J})\,,$$

where $T_{1:J}$ denotes the first J maps in our dictionary (the same maps exist in all dimensions) (This is a slight abuse of notation because the i-th occurrence of $T_{1:J}$ above acts only on the i-th coordinate of the input.) Thus, the matrix $Q^{-1/2}$ is block-diagonal, written

$$Q^{-1/2} = I_d \otimes Q_{1:J}^{-1/2}\,,$$

where $Q_{1:J}$ is the first $J \times J$ block of the full Q matrix, which is $Jd \times Jd$. We can similarly express the gradients with respect to λ in this way (i.e., only differentiating the first J components), which results in controlling the following quantity

$$\mathbb{E}[\|Q^{-1/2}\left(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda)\right)\|^2] = \sum_{i=1}^{d} \mathbb{E}[\|Q_{1:J}^{-1/2}\left(\hat{\nabla}_{1:J} \mathcal{V}(\mu_\lambda) - \nabla_{1:J} \mathcal{V}(\mu_\lambda)\right)\|^2] \,.$$

Combining these reductions, we are left with bounding the following term in each dimension:

$$\begin{aligned}\operatorname{tr} \operatorname{Cov}\left(Q_{1:J}^{-1/2}\, T_{1:J}(\hat{X}_i)\, \partial_i V(T^\lambda(\hat{X})\right) &= \mathbb{E}\left[\langle T_{1:J}(\hat{X}_i)\, T_{1:J}(\hat{X}_i)^\top, Q_{1:J}^{-1}\rangle\, \partial_i V(T^\lambda(\hat{X}))^2\right] \\ &\leqslant \Xi J\, \mathbb{E}[\partial_i V(T^\lambda(\hat{X}))^2] \,,\end{aligned}$$

where we invoked (Ξ) in the last inequality. Summing over the coordinates,

$$\mathbb{E}[\|Q^{-1/2}\left(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda)\right)\|^2] \leqslant \Xi J\, \mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2 \,.$$

We bound the remaining expectation by repeatedly invoking smoothness of V. First,

$$\begin{aligned}\mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2 &\leq 2\, \mathbb{E}_\rho \|\nabla V \circ T^\lambda - \nabla V \circ T_\diamond^\star\|^2 + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2 \\ &\leq 2L_V^2\, \|T^\lambda - T_\diamond^\star\|_{L^2(\rho)}^2 + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2 \\ &= 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2 \,.\end{aligned}$$

For the next term, we apply the same trick, but we compare against $\pi^\star$, the true mean-field approximation:

$$\begin{aligned}\mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2 &\leq 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 4\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star - \nabla V \circ T^\star\|^2 + 4\, \mathbb{E}_\rho \|\nabla V \circ T^\star\|^2 \\ &\leq 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 4L_V^2\, W_2^2(\pi_\diamond^\star, \pi^\star) + 4L_V \kappa d \,,\end{aligned}$$

where we used Lemma F.4 in the last step.

Our full variance bound reads

$$\mathbb{E}[\|Q^{-1}\left(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda)\right)\|_Q^2] \leq 2L_V^2 \Xi J\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 4L_V \Xi J\left(L_V\, W_2^2(\pi_\diamond^\star, \pi^\star) + \kappa d\right).$$

□

Finally, we also prove the bound on Ξ for the piecewise linear dictionary.

Proof of Lemma 7.32. If we can show that $Q \geq \gamma I$ for some $\gamma > 0$, then

$$\langle Q^{-1}, \bar{Q}(x)\rangle \leqslant \gamma^{-1} \operatorname{tr} \bar{Q}(x) = \gamma^{-1} \sum_{T \in \mathcal{M}} T(x)^2 \leqslant \gamma^{-1} J,$$

where we use the fact that the elements of the piecewise linear dictionary are uniformly bounded by 1.

To prove the lower bound on Q, we note that for any $\lambda \in \mathbb{R}^J$,

$$\langle \lambda, Q\,\lambda\rangle = \Big\| \sum_{T \in \mathcal{M}} \lambda_T T \Big\|_{L^2(\rho)}^2.$$

On an interval $[a, a+\delta]$, since $T^\lambda(x) = T^\lambda(a) + \lambda_T\,((x-a)/\delta)_+$,

$$\begin{aligned}\int_a^{a+\delta} T^\lambda(x)^2\,\rho(\mathrm{d}x) = \int_a^{a+\delta} \Big(T^\lambda(a) + \lambda_T\,\frac{x-a}{\delta}\Big)^2 \rho(\mathrm{d}x) &\geqslant \lambda_T^2 \inf_{m \in \mathbb{R}} \int_a^b \big(\frac{x-a}{\delta} - m\big)^2 \rho(\mathrm{d}x) \\ &\gtrsim \lambda_T^2 \delta^2 \int_a^{a+\delta} \rho(\mathrm{d}x),\end{aligned}$$

where we used the variance bound from the proof of Lemma 7.30. Summing across the intervals, we find that $\langle \lambda, Q\,\lambda\rangle \gtrsim \delta^2 \sum_{T \in \mathcal{M}} \lambda_T^2$, so we can take $\gamma \asymp \delta^2$. This leads to an upper bound on Ξ of order $\delta^{-2} \asymp J^2$. □

F.4 Remaining implementation details

F.4.1 Product Gaussian mixture

Let V_1 (resp. V_2) be the potential for a univariate Gaussian mixture with weights $w_{1,1}$ and $w_{1,2}$ (resp. $w_{2,1}$ and $w_{2,2}$) that sum to unity, and centers $m_{1,1}$ and $m_{1,2}$ (resp. $m_{2,1}$ and $m_{2,2}$), where all the mixture components have unit variance. Then, $V : \mathbb{R}^2 \to \mathbb{R}$ defined by $V(x, y) = V_1(x) + V_2(y)$ is the potential for the Gaussian mixture with mean-weight pairs given by

$$\{([m_{1,1}, m_{2,1}], w_{1,1}w_{2,1}), ([m_{1,1}, m_{2,2}], w_{1,1}w_{2,2}), ([m_{1,2}, m_{2,1}], w_{1,2}w_{2,1}), ([m_{1,2}, m_{2,2}], w_{1,2}w_{2,2})\}\,.$$

We take $m_{1,1} = m_{2,1} = 2$, $m_{1,2} = m_{2,2} = -2$, with $w_{1,1} = w_{2,2} = 0.25$ and $w_{1,2} = w_{2,1} = 0.75$. As for the hyperparameters of our model, we chose $J = 28$, $\alpha = 0.1$, a step-size $h = 10^{-3}$ (for both λ and v), ran for 3000 iterations, and initialized at $\lambda^{(0)} = 0_{2\times J} \in \mathbb{R}^{2\times J}$, and $v^{(0)} = 0_2 \in \mathbb{R}^2$. The KDE plots were generated via `sklearn`, after we generated 50,000 samples from the ground truth density and from our algorithm.

F.4.2 Non-isotropic Gaussian

We generated $A \in \mathbb{R}^{d\times d}$ with entries $A_{i,j} \sim \mathcal{N}(0, 1)$, and defined $\Sigma = AA^\top$ for $d = 5$, which is fixed once and for all. We computed the optimal $\alpha^* = 1/\sqrt{L_V}$, since the potential is a Gaussian. For the remaining hyper-parameters of our model, we chose $J = 28$, a step-size $h = 10^{-4}$ (for both λ and v), ran for 2000 iterations, and initialized at $\lambda^{(0)} = \mathbf{1}_{d\times J} \in \mathbb{R}^{d\times J}$, the all-ones matrix, and $v^{(0)} = 0_d \in \mathbb{R}^d$. At each step, we computed $\hat{\Sigma}_{\mathrm{MF}}$ by pushing forward 10,000 samples, computing the empirical covariance, and computing the Bures–Wasserstein distance to Σ_{MF}.

We now compute the fact that Σ_{MF} is diagonal with components $1/(\Sigma^{-1})_{i,i}$ for $i \in [d]$. Recall

the KL divergence between two Gaussians with mean zero is given by

$$\mathrm{KL}(\mathcal{N}(0,A)\|\mathcal{N}(0,\Sigma)) = \frac{1}{2}\left[\mathrm{tr}(\Sigma^{-1}A) - d + \log\det(\Sigma) - \log\det(A)\right].$$

Now, we impose that A is a diagonal matrix with entries $A_{i,i} = a_i$ for some $a_i \geq 0$. In this case, up to constants denoted by C, the above reads

$$\mathrm{KL}(\mathcal{N}(0,A)\|\mathcal{N}(0,\Sigma)) = \frac{1}{2}\sum_{i=1}^{d}\left[(\Sigma^{-1})_{i,i}a_i - \log(a_i)\right] + C\,.$$

Taking the derivative in a_i, we see that the optimality conditions yield

$$1/(\Sigma^{-1})_{i,i} = a_i^\star$$

for every $i \in [d]$, which completes the calculation.

F.4.3 Bayesian logistic regression

We first randomly drew $\theta^\star \sim \mathcal{N}(0, I_d)$ in $d = 20$ as the ground truth parameter. Further, we let $n = 100$ and randomly generated $X \in \mathbb{R}^{n\times d}$ as in the non-isotropic Gaussian experiment (here, X takes the role of A), but we divided the matrix by $\lambda_{\max}(X^\top X)$ for normalization purposes. Subsequently, Y_i was generated for each i independently according to

$$Y_i \mid X_i \sim \mathrm{Bern}(\exp(\theta^\top X_i))\,,$$

where X_i is a row of X. Using this data, and assuming an improper (Lebesgue) prior on θ, the potential of the posterior is given by

$$V(\theta) = \sum_{i=1}^{n} \left[\log(1 + \exp(\theta^\top X_i)) - Y_i \, \theta^\top X_i \right].$$

With access to V and ∇V, we ran standard Langevin Monte Carlo (LMC) for 5000 iterations with a step size of $h = 10^{-2}$, where we generated 2000 samples.

For the hyperparameters of our model, we chose $J = 28$, $\alpha = 0.1$, a step size $h = 10^{-2}$ for the λ iterates, and $h_v = 10^{-1}$ for updating v, and ran for 2000 iterations. We initialized at $\lambda^{(0)} = \mathbf{1}_{d\times J}/(Jd) \in \mathbb{R}^{d\times J}$, and $v^{(0)} = 0_d \in \mathbb{R}^d$. The final histograms were generated using 2000 samples from both the mean-field VI algorithm and LMC.

F.5 Proofs for Section 7.7

In this section, we derive the gradient flows in Section 7.7.

Proof of Theorem 7.36. We refer to Lambert et al. (2022, Appendix F) for the relevant background. The first variation of the functional $\mathcal{F}(P) \coloneqq \mathrm{KL}(\mu_P \| \pi)$ is given by

$$\delta\mathcal{F}(P) : (\lambda, v) \mapsto \int (V + \log \mu_P + 1) \, \mathrm{d}\mu_{\lambda,v} = \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} + 1 \,. \tag{F.10}$$

Therefore, the Wasserstein gradient is given by

$$\nabla_{\mathbb{W}} \mathcal{F}(P)(\lambda, v) = \left(Q^{-1} \nabla_\lambda \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v}, \, \nabla_v \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} \right). \tag{F.11}$$

These terms are further computed as follows. First,

$$\partial_{\lambda_T} \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} = \partial_{\lambda_T} \int \log \frac{\mu_P}{\pi} \circ T^{\lambda,v} \, \mathrm{d}\rho = \int \left\langle \nabla \log \frac{\mu_P}{\pi} \circ T^{\lambda,v}, T \right\rangle \mathrm{d}\rho \,.$$

Similarly, we have

$$\nabla_v \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} = \int \nabla \log \frac{\mu_P}{\pi} \circ T^{\lambda,v} \, \mathrm{d}\rho \, .$$

This concludes the proof. □

Proof of Theorem 7.37. This theorem follows from the expression of the first variation computed in (F.10), see Lambert et al. (2022, Appendix H). □

Bibliography

Ahidar-Coutrix, A., Le Gouic, T., and Paris, Q. (2020). Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probab. Theory Related Fields*, 177(1-2):323–368.

Albergo, M. S., Boffi, N. M., Lindsey, M., and Vanden-Eijnden, E. (2024). Multimarginal generative modeling with stochastic interpolants. In *The Twelfth International Conference on Learning Representations*.

Albergo, M. S. and Vanden-Eijnden, E. (2022). Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*.

Altschuler, J., Bach, F., Rudi, A., and Niles-Weed, J. (2019). Massively scalable sinkhorn distances via the nyström method. *Advances in neural information processing systems*, 32.

Altschuler, J., Chewi, S., Gerber, P., and Stromme, A. (2021). Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30*.

Altschuler, J. M., Niles-Weed, J., and Stromme, A. J. (2022). Asymptotics for semidiscrete entropic optimal transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition.

Anari, N., Jain, V., Koehler, F., Pham, H. T., and Vuong, T.-D. (2021a). Entropic independence i: Modified log-Sobolev inequalities for fractionally log-concave distributions and high-temperature Ising models. *arXiv preprint arXiv:2106.04105*.

Anari, N., Liu, K., and Gharan, S. O. (2021b). Spectral independence in high-dimensional expanders and applications to the hardcore model. *SIAM Journal on Computing*, (0):FOCS20–1.

Andoni, A., Naor, A., and Neiman, O. (2015). Snowflake universality of Wasserstein spaces. *arXiv preprint arXiv:1509.08677*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223.

Arnese, M. and Lacker, D. (2024). Convergence of coordinate ascent variational inference for log-concave measures via optimal transport. *arXiv preprint arXiv:2404.08792*.

Aurenhammer, F., Hoffmann, F., and Aronov, B. (1998). Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76.

Austin, T. (2019). The structure of low-complexity Gibbs measures on product spaces. *Ann. Probab.*, 47(6):4002–4023.

Backhoff-Veraguas, J., Fontbona, J., Rios, G., and Tobar, F. (2022). Bayesian learning with Wasserstein barycenters. *ESAIM Probab. Stat.*, 26:436–472.

Bakry, D., Gentil, I., and Ledoux, M. (2014). *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham.

Bansil, M. and Kitagawa, J. (2022). Quantitative stability in the geometry of semi-discrete optimal transport. *International Mathematics Research Notices*, 2022(10):7354–7389.

Baptista, R., Pooladian, A.-A., Brennan, M., Marzouk, Y., and Niles-Weed, J. (2024). Conditional simulation via entropic optimal transport: Toward non-parametric estimation of conditional brenier maps. *arXiv preprint arXiv:2411.07154*.

Basu, S., Kolouri, S., and Rohde, G. K. (2014). Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453.

Bauerschmidt, R., Bodineau, T., and Dagallier, B. (2023). Stochastic dynamics and the Polchinski equation: an introduction. *arXiv preprint arXiv:2307.07619*.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

Berman, R. J. (2021). Convergence rates for discretized Monge–Ampère equations and quantitative stability of optimal transport. *Foundations of Computational Mathematics*, 21(4):1099–1140.

Bernton, E., Ghosal, P., and Nutz, M. (2022). Entropic optimal transport: Geometry and large deviations. *Duke Mathematical Journal*, 171(16):3363–3400.

Bernton, E., Heng, J., Doucet, A., and Jacob, P. E. (2019). Schrödinger bridge samplers. *arXiv preprint arXiv:1912.13170*.

Bhattacharya, A., Pati, D., and Yang, Y. (2023). On the convergence of coordinate ascent variational inference. *arXiv preprint arXiv:2306.01122*.

Bigot, J., Gouet, R., Klein, T., and López, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. Henri Poincaré Probab. Stat.*, 53(1):1–26.

Bing, X., Bunea, F., and Niles-Weed, J. (2023). Estimation and inference for the Wasserstein distance between mixing measures in topic models. *arXiv preprint 2206.12768*.

Birgé, L. (2001). An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, pages 113–133.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bobkov, S. G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.*, 163(1):1–28.

Bobkov, S. G. and Ledoux, M. (2000). From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.*, 10(5):1028–1052.

Bogachev, V. I. (2007). *Measure Theory*, volume 1. Springer Science & Business Media.

Boissard, E., Le Gouic, T., and Loubes, J.-M. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759.

Bonneel, N., Peyré, G., and Cuturi, M. (2016). Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Brascamp, H. J. and Lieb, E. H. (1976). On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Functional Analysis*, 22(4):366–389.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. (2022). The union of manifolds hypothesis and its implications for deep generative modelling. *arXiv preprint arXiv:2207.02862*.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Bunne, C., Hsieh, Y.-P., Cuturi, M., and Krause, A. (2023a). The Schrödinger bridge between Gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR.

Bunne, C., Papaxanthos, L., Krause, A., and Cuturi, M. (2022). Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. (2023b). Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768.

Bures, D. (1969). An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212.

Caffarelli, L. A. (1992). Boundary regularity of maps with convex potentials. *Communications on pure and applied mathematics*, 45(9):1141–1151.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496.

Caffarelli, L. A. (2000). Monotonicity properties of optimal transportation and the FKG and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563.

Cai, T., Cheng, J., Craig, N., and Craig, K. (2020). Linearized optimal transport for collider events. *Physical Review D*, 102(11):116019.

Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192.

Carlier, G., Chizat, L., and Laborde, M. (2024). Displacement smoothness of entropic optimal transport.

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418.

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., and Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM J. Sci. Comput.*, 40(2):B429–B456.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Chen, R. T. Q., Amos, B., and Nickel, M. (2022a). Semi-discrete normalizing flows through differentiable tessellation. In *Advances in Neural Information Processing Systems*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022b). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Chen, T., Liu, G.-H., and Theodorou, E. A. (2021a). Likelihood training of Schrödinger bridge using forward-backward SDEs theory. *arXiv preprint arXiv:2110.11291*.

Chen, Y. and Eldan, R. (2022). Localization schemes: a framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*, pages 110–122. IEEE Computer Soc., Los Alamitos, CA.

Chen, Y., Georgiou, T. T., and Pavon, M. (2016). On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169:671–691.

Chen, Y., Georgiou, T. T., and Pavon, M. (2021b). Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *Siam Review*, 63(2):249–313.

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2019). Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278.

Chen, Y., Goldstein, M., Hua, M., Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. (2024). Probabilistic forecasting with stochastic interpolants and Föllmer processes. *arXiv preprint arXiv:2403.13724*.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.

Chewi, S. (2024). Log-concave sampling. Book draft available at https://chewisinho.github.io.

Chewi, S., Clancy, J., Le Gouic, T., Rigollet, P., Stepaniants, G., and Stromme, A. J. (2021). Fast and smooth interpolation on Wasserstein space. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3061–3069. PMLR.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. (2020). Gradient descent algorithms for Bures–Wasserstein barycenters. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1276–1304. PMLR.

Chewi, S. and Pooladian, A.-A. (2023). An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *Comptes Rendus. Mathématique*, 361(G9):1471–1482.

Chiarini, A., Conforti, G., Greco, G., and Tamanini, L. (2022). Gradient estimates for the Schrödinger potentials: Convergence to the Brenier map and quantitative stability. *arXiv preprint arXiv:2207.14262*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). An interpolating distance between optimal transport and Fisher–Rao metrics. *Found. Comput. Math.*, 18(1):1–44.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269.

Chizat, L., Zhang, S., Heitz, M., and Schiebinger, G. (2022). Trajectory inference via mean-field Langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742.

Colombo, M., Figalli, A., and Jhaveri, Y. (2017). Lipschitz changes of variables between perturbations of log-concave measures. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 17(4):1491–1519.

Conforti, G. (2024). Weak semiconvexity estimates for schrödinger potentials and logarithmic sobolev inequality for schrödinger bridges. *Probability Theory and Related Fields*, 189(3):1045–1071.

Conforti, G., Durmus, A., and Greco, G. (2023). Quantitative contraction rates for Sinkhorn algorithm: Beyond bounded costs and compact marginals. *arXiv preprint arXiv:2304.04451*.

Conforti, G. and Tamanini, L. (2021). A formula for the time derivative of the entropic cost and applications. *Journal of Functional Analysis*, 280(11):108964.

Cordero-Erausquin, D. (2017). Transport inequalities for log-concave measures, quantitative forms, and applications. *Canad. J. Math.*, 69(3):481–501.

Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *ECML PKDD*, pages 274–289.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158.

Cuesta-Albertos, J. A., Matrán-Bea, C., and Tuero-Diaz, A. (1996). On lower bounds for the L^2-Wasserstein metric in a Hilbert space. *J. Theoret. Probab.*, 9(2):263–283.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693. PMLR.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*.

Dalalyan, A. S., Karagulyan, A., and Riou-Durand, L. (2022). Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *J. Mach. Learn. Res.*, 23:Paper No. 235, 38.

Daniels, M., Maunu, T., and Hand, P. (2021). Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965.

Dantzig, G. B. (1951). Application of the simplex method to a transportation problem. *Activity analysis and production and allocation*.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753.

del Barrio, E., González-Sanz, A., and Loubes, J.-M. (2022a). Central limit theorems for semidiscrete Wasserstein distances. *arXiv preprint arXiv:2202.06380*.

del Barrio, E., Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022b). An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.*, 47(2):926–951.

Delalande, A. (2022). Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1619–1642. PMLR.

Delalande, A. and Mérigot, Q. (2023). Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*.

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM J. Imaging Sci.*, 13(2):936–970.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. (2023). Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2022). Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2025). Tight stability bounds for entropic brenier maps. *International Mathematics Research Notices*, 2025(7):rnaf078.

Domke, J. (2020). Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning*, pages 2587–2596. PMLR.

Domke, J., Gower, R., and Garrigos, G. (2023). Provable convergence guarantees for black-box variational inference. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 66289–66327. Curran Associates, Inc.

Dudley, R. M. (1969). The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*.

Eckstein, S. and Nutz, M. (2023). Convergence rates for regularized optimal transport via quantization. *Mathematics of Operations Research*.

Eldan, R. (2018). Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.*, 28(6):1548–1596.

Eldan, R. and Gross, R. (2018). Decomposition of mean-field Gibbs distributions into product measures. *Electron. J. Probab.*, 23:Paper No. 35, 24.

Eldan, R., Lehec, J., and Shenfeld, Y. (2020). Stability of the logarithmic Sobolev inequality via the Föllmer process.

Fathi, M., Gozlan, N., and Prod'homme, M. (2020). A proof of the Caffarelli contraction theorem via entropic regularization. *Calculus of Variations and Partial Differential Equations*, 59(3):1–18.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. (2017). Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer.

Feydy, J., Glaunès, A., Charlier, B., and Bronstein, M. (2020). Fast geometric learning with symbolic matrices. *Advances in Neural Information Processing Systems*, 33:14448–14462.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR.

Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. (2020a). Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033*.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. (2020b). How to train your neural ODE: the world of Jacobian and kinetic regularization. In *International Conference on Machine Learning*, pages 3154–3164. PMLR.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Föllmer, H. (1985). An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pages 156–163. Springer, Berlin.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. (2019). Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR.

Fortet, R. (1940). Résolution d'un système d'équations de m. Schrödinger. *Journal de Mathématiques Pures et Appliquées*, 19(1-4):83–105.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

Gelbrich, M. (1990). On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203.

Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Gentil, I., Léonard, C., Ripani, L., and Tamanini, L. (2020). An entropic interpolation proof of the HWI inequality. *Stochastic Processes and their Applications*, 130(2):907–923.

Ghosal, P., Nutz, M., and Bernton, E. (2022). Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622.

Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *The Annals of Statistics*, 50:1012–1037.

Gigli, N. (2011). On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409.

Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2024a). Limit theorems for entropic optimal transport maps and sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2024b). Statistical inference with regularized optimal transport. *Information and Inference: A Journal of the IMA*, 13(1):iaad056.

Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. *arXiv preprint arXiv:2207.07427*.

Gozlan, N. and Juillet, N. (2020). On a mixture of Brenier and Strassen theorems. *Proc. Lond. Math. Soc. (3)*, 120(3):434–463.

Gozlan, N. and Sylvestre, M. (2025). Global regularity estimates for optimal transport via entropic regularisation. *arXiv preprint arXiv:2501.11382*.

Graf, S. and Luschgy, H. (2007). *Foundations of quantization for probability distributions*. Springer.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2018). Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*.

Greco, G., Noble, M., Conforti, G., and Durmus, A. (2023). Non-asymptotic convergence bounds for Sinkhorn iterates and their gradients: a coupling approach. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 716–746. PMLR.

Groppe, M. and Hundrieser, S. (2024). Lower complexity adaptation for empirical entropic optimal transport. *Journal of Machine Learning Research*, 25(344):1–55.

Gunsilius, F., Hsieh, M. H., and Lee, M. J. (2024). Tangential wasserstein projections. *Journal of Machine Learning Research*, 25(69):1–41.

Gunsilius, F. and Xu, Y. (2021). Matching for causal effects via multimarginal optimal transport. *arXiv preprint arXiv:2112.04398*.

Gushchin, N., Kolesov, A., Mokrov, P., Karpikova, P., Spiridonov, A., Burnaev, E., and Korotin, A. (2023). Building the bridge of Schrödinger: A continuous entropic optimal transport benchmark. *Advances in Neural Information Processing Systems*, 36:18932–18963.

Hallin, M., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach.

Haviv, D., Pooladian, A.-A., Pe'er, D., and Amos, B. (2024). Wasserstein flow matching: Generative modeling over families of distributions. *arXiv preprint arXiv:2411.00698*.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2004). *Fundamentals of convex analysis*. Springer Science & Business Media.

Huang, C.-W., Chen, R. T. Q., Tsirigotis, C., and Courville, A. (2021a). Convex potential flows: universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*.

Huang, H. (2024). One-step data-driven generative model via Schrödinger bridge. *arXiv preprint arXiv:2405.12453*.

Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J., and Liu, Y. (2021b). Schrödinger–Föllmer sampler: Sampling without ergodicity. *arXiv preprint arXiv:2106.10880*.

Hundrieser, S., Klatt, M., and Munk, A. (2024a). Limit distributions and sensitivity analysis for empirical entropic optimal transport on countable spaces. *The Annals of Applied Probability*, 34(1B):1403–1468.

Hundrieser, S., Staudt, T., and Munk, A. (2024b). Empirical optimal transport between different measures adapts to lower complexity. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 60, pages 824–846. Institut Henri Poincaré.

Hütter, J.-C. and Mao, C. (2017). Notes on adaptive estimation with Lepski's method.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.

Jaggi, M. (2013). Revisiting Frank–Wolfe: projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR.

Jain, V., Risteski, A., and Koehler, F. (2019). Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236. ACM, New York.

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17.

Kantorovitch, L. (1942). On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201.

Kassraie, P., Pooladian, A.-A., Klein, M., Thornton, J., Niles-Weed, J., and Cuturi, M. (2024). Progressive entropic optimal transport solvers. *Advances in Neural Information Processing Systems*, 37:19561–19590.

Kato, K. (2024). Large deviations for dynamical Schrödinger problems. *arXiv preprint arXiv:2402.05100*.

Kawakita, G., Kamiya, S., Sasai, S., Kitazono, J., and Oizumi, M. (2022). Quantifying brain state transition cost via Schrödinger bridge. *Network Neuroscience*, 6(1):118–134.

Kent, C., Blanchet, J., and Glynn, P. (2021). Frank–Wolfe methods in probability space. *arXiv preprint arXiv:2105.05352*.

Khurana, V., Kannan, H., Cloninger, A., and Moosmüller, C. (2023). Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1):1.

Kim, K., Oh, J., Wu, K., Ma, Y., and Gardner, J. (2023). On the convergence of black-box variational inference. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44615–44657. Curran Associates, Inc.

Kim, Y.-H. and Milman, E. (2012). A generalization of Caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862.

Klatt, M., Tameling, C., and Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443.

Klein, M., Pooladian, A.-A., Ablin, P., Ndiaye, E., Niles-Weed, J., and Cuturi, M. (2024). Learning elastic costs to shape monge displacements. *Advances in Neural Information Processing Systems*, 37:108542–108565.

Kolesnikov, A. V. (2011). Mass transportation and contractions. *arXiv preprint arXiv:1103.1479*.

Kolouri, S. and Rohde, G. K. (2015). Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884.

Kolouri, S., Tosun, A. B., Ozolek, J. A., and Rohde, G. K. (2016). A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. Informs.

Lacker, D. (2023). Independent projections of diffusions: gradient flows for variational inference and optimal mean field approximations. *arXiv preprint arXiv:2309.13332*.

Lacker, D., Mukherjee, S., and Yeung, L. C. (2024). Mean field approximations via log-concavity. *International Mathematics Research Notices*, 2024(7):6008–6042.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447.

Lavenant, H. and Zanella, G. (2024). Convergence rate of random scan coordinate ascent variational inference under log-concavity. *SIAM Journal on Optimization*, 34(4):3750–3761.

Lavenant, H., Zhang, S., Kim, Y.-H., Schiebinger, G., et al. (2024). Toward a mathematical theory of trajectory inference. *The Annals of Applied Probability*, 34(1A):428–500.

Le Gouic, T., Paris, Q., Rigollet, P., and Stromme, A. J. (2022). Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 25(6):2229–2250.

Ledoux, M. (2018). Remarks on some transportation cost inequalities.

Lee, D., Lee, D., Bang, D., and Kim, S. (2024). Disco: Diffusion Schrödinger bridge for molecular conformer optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13365–13373.

Lee, H., Lu, J., and Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR.

Léonard, C. (2012). From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920.

Léonard, C. (2014). A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34(4):1533–1574.

Letrouit, C. and Mérigot, Q. (2024). Gluing methods for quantitative stability of optimal transport maps. *arXiv preprint arXiv:2411.04908*.

Liero, M., Mielke, A., and Savaré, G. (2016). Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM J. Math. Anal.*, 48(4):2869–2911.

Liero, M., Mielke, A., and Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, G.-H., Chen, T., So, O., and Theodorou, E. (2022a). Deep generalized Schrödinger bridge. *Advances in Neural Information Processing Systems*, 35:9374–9388.

Liu, X., Gong, C., and Liu, Q. (2022b). Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lu, Y., Lu, J., and Nolen, J. (2019). Accelerating Langevin sampling with birth-death. *arXiv preprint 1905.09863*.

Lunardi, A. (2009). *Interpolation theory*, volume 9. Edizioni della normale Pisa.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR.

Mallasto, A., Gerolin, A., and Minh, H. Q. (2022). Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Information Geometry*, 5(1):289–323.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2024a). Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998.

Manole, T., Bryant, P., Alison, J., Kuusela, M., and Wasserman, L. (2024b). Background modeling for double higgs boson production: Density ratios and optimal transport. *The Annals of Applied Statistics*, 18(4):2950–2978.

Manole, T. and Niles-Weed, J. (2024). Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135.

Marino, S. D. and Gerolin, A. (2020). An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):1–28.

Masud, S. B., Werenski, M., Murphy, J. M., and Aeron, S. (2023). Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32.

Mérigot, Q., Delalande, A., and Chazal, F. (2020). Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR.

Mérigot, Q., Santambrogio, F., and Sarrazin, C. (2021). Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12810–12821. Curran Associates, Inc.

Mikulincer, D. and Shenfeld, Y. (2023). On the lipschitz properties of transportation along heat flows. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2020-2022*, pages 269–290. Springer.

Mikulincer, D. and Shenfeld, Y. (2024). The Brownian transport map. *Probability Theory and Related Fields*, pages 1–66.

Mokrov, P., Korotin, A., Kolesov, A., Gushchin, N., and Burnaev, E. (2023). Energy-guided entropic neural optimal transport. *arXiv preprint arXiv:2304.06094*.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704.

Moriel, N., Senel, E., Friedman, N., Rajewsky, N., Karaiskos, N., and Nitzan, M. (2021). Novosparc:

flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200.

Neeman, J. (2022). Lipschitz changes of variables via heat flow. *arXiv preprint arXiv:2201.03403*.

Nemirovski, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547.

Nusken, N., Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., and Lawrence, N. (2022). Bayesian learning via neural Schrödinger–Föllmer flows. *STATISTICS AND COMPUTING*, 33.

Nutz, M. (2021). Introduction to entropic optimal transport. *Lecture notes, Columbia University*.

Nutz, M. and Wiesel, J. (2021). Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, pages 1–24.

Nutz, M. and Wiesel, J. (2023). Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *The Annals of Probability*, 51(2):699–722.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. (2021). Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation.

Pal, S. (2024). On the difference between entropic cost and the optimal transport cost. *The Annals of Applied Probability*, 34(1B):1003–1028.

Panaretos, V. M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.*, 44(2):771–812.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. Springer Nature.

Park, S. and Thorpe, M. (2018). Representing and learning high dimensional data with the optimal transport map from a probabilistic viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7864–7872.

Pavon, M., Trigila, G., and Tabak, E. G. (2021). The data-driven Schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pooladian, A.-A., Cuturi, M., and Niles-Weed, J. (2022). Debiaser beware: Pitfalls of centering regularized transport maps. In *International Conference on Machine Learning*, pages 17830–17847. PMLR.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning*, pages 28128–28150. PMLR.

Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.

Rigollet, P. and Stromme, A. J. (2022). On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*.

Ripani, L. (2019). Convexity and regularity properties for entropic interpolations. *Journal of Functional Analysis*, 277(2):368–391.

Rockafellar, R. T. (1997). *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Sadhu, R., Goldfeld, Z., and Kato, K. (2024). Stability and statistical inference for semidiscrete optimal transport maps. *The Annals of Applied Probability*, 34(6):5694–5736.

Sadhu, R., Goldfeld, Z., and Kato, K. (2025). Approximation rates of entropic maps in semidiscrete optimal transport. *Electronic Communications in Probability*, 30:1–13.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. (2018). Improving GANs using optimal transport. In *International Conference on Learning Representations*.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.

Schrödinger, E. (1932). Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2018). Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations*.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. (2024). Diffusion Schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36.

Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2022). Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pages 1792–1802. PMLR.

Sinkhorn, R. (1967). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stromme, A. (2023). Sampling from a Schrödinger bridge. In *International Conference on Artificial Intelligence and Statistics*, pages 4058–4067. PMLR.

Stromme, A. J. (2024). Minimum intrinsic dimension scaling for entropic optimal transport. In *International Conference on Soft Methods in Probability and Statistics*, pages 491–499. Springer.

Thornton, J., Hutchinson, M., Mathieu, E., De Bortoli, V., Teh, Y. W., and Doucet, A. (2022). Riemannian diffusion Schrödinger bridge. *arXiv preprint arXiv:2207.03024*.

Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., and Bengio, Y. (2023). Simulation-free Schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*.

Torous, W., Gunsilius, F., and Rigollet, P. (2024). An optimal transport approach to estimating causal effects via nonlinear difference-in-differences. *Journal of Causal Inference*, 12(1):20230004.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes with applications to statistics. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Vacher, A., Muzellec, B., Bach, F., Vialard, F.-X., and Rudi, A. (2024). Optimal estimation of smooth transport maps with kernel sos. *SIAM Journal on Mathematics of Data Science*, 6(2):311–342.

Valdimarsson, S. I. (2007). On the Hessian of the optimal transport potential. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 6(3):441–456.

Vargas, F., Ovsianas, A., Fernandes, D., Girolami, M., Lawrence, N. D., and Nüsken, N. (2023). Bayesian learning via neural Schrödinger–Föllmer flows. *Statistics and Computing*, 33(1):3.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. (2021). Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Villani, C. (2009). *Optimal transport: Old and new*, volume 338. Springer.

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

von Luxburg, U. and Bousquet, O. (2003). Distance-based classification with Lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., and Rohde, G. K. (2010). An optimal transportation approach for nuclear structure-based pathology. *IEEE transactions on medical imaging*, 30(3):621–631.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101:254–269.

Werenski, M., Jiang, R., Tasissa, A., Aeron, S., and Murphy, J. M. (2022). Measure estimation in the barycentric coding model. In *International Conference on Machine Learning*, pages 23781–23803. PMLR.

Werenski, M., Murphy, J. M., and Aeron, S. (2023). Estimation of entropy-regularized optimal transport maps between non-compactly supported measures. *arXiv preprint arXiv:2311.11934*.

Wibisono, A. (2018). Sampling as optimization in the space of measures: the Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR.

Wibisono, A., Wu, Y., and Yang, K. Y. (2024). Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4958–4991. PMLR.

Yang, K. D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. (2020). Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828.

Yao, R. and Yang, Y. (2022). Mean field variational inference via Wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*.

Yi, M. and Liu, S. (2023). Bridging the gap between variational inference and Wasserstein gradient flows. *arXiv preprint arXiv:2310.20090*.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. (2023). Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*.

Yue, M.-C., Kuhn, D., and Wiesemann, W. (2022). On linear optimization over Wasserstein balls. *Mathematical Programming*, 195(1-2):1107–1122.

Zemel, Y. and Panaretos, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976.

Zhang, A. Y. and Zhou, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, 48(5):2575–2598.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.