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

Yiheng Jiang* Sinho Chewi† Aram-Alexandre Pooladian‡

June 11, 2024

Abstract

We develop a theory of finite-dimensional polyhedral subsets over the Wasserstein space and optimization of functionals over them via first-order methods. Our main application is to the problem of mean-field variational inference, which seeks to approximate a distribution π over $\mathbb{R}^d$ by a product measure $\pi^\star$. 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 $\mathcal{P}_\diamond$, and (2) an algorithm for minimizing $\mathrm{KL}(\cdot \,\|\, \pi)$ over $\mathcal{P}_\diamond$ with accelerated complexity $O(\sqrt{\kappa}\log(\kappa d/\varepsilon^2))$, where κ is the condition number of π.

1 Introduction

This paper 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 (Wainwright and Jordan, 2008; Blei et al., 2017) 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 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 (Cuturi and Doucet, 2014; Zemel and Panaretos, 2019; Chewi et al., 2020; Altschuler et al., 2021; Backhoff-Veraguas et al., 2022), sampling (Jordan et al., 1998; Wibisono, 2018; Chewi, 2024), 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.

*Courant Institute of Mathematical Sciences, New York University. yj2070@nyu.edu
†School of Mathematics, Institute for Advanced Study. schewi@ias.edu
‡Center for Data Science, New York University. aram-alexandre.pooladian@nyu.edu

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 $\mathcal{C}$:

$$\pi^\star \in \arg\min_{\mu\in\mathcal{C}} \mathrm{KL}(\mu \,\|\, \pi) = \arg\min_{\mu\in\mathcal{C}} \int \log\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\Big)\,\mathrm{d}\mu\,. \tag{1}$$

For example, $\mathcal{C}$ could be taken to be the class of non-degenerate Gaussian distributions, in which case (1) is known as Gaussian VI. Recently, by leveraging the rich theory of gradient flows over the Wasserstein space, Lambert et al. (2022); Diao et al. (2023) provided algorithmic guarantees for Gaussian VI under standard tractability assumptions, i.e., strong log-concavity and log-smoothness of π.

In this paper, we instead study the problem of *mean-field* VI, in which $\mathcal{C}$ 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 Zhang and Zhou (2020); Yao and Yang (2022); Lacker (2023) have also developed algorithms based on Wasserstein gradient flows, although computational guarantees for VI are still nascent (see Section 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 (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 5.5.1, with code available here. Below, we describe our contributions in more detail.

Main contributions

Polyhedral optimization in the Wasserstein space. We study parametric sets of the following form:

$$\mathrm{cone}(\mathcal{M})_\sharp\rho := \Big\{ \big(\textstyle\sum_{T\in\mathcal{M}} \lambda_T T\big)_\sharp \rho \;\Big|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \Big\}\,,$$

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 $\mathrm{cone}(\mathcal{M})$ 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 $\mathrm{cone}(\mathcal{M})_\sharp\rho$ 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, $(\mathrm{cone}(\mathcal{M})_\sharp\rho, W_2)$ 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 (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 *accelerated* and *end-to-end* convergence rates for mean-field VI.

Organization. Our paper is outlined as follows: In Section 2 we provide relevant background information on optimal transport theory and the Wasserstein space. In Section 3, we develop known properties of compatibility which provide tools for building large compatible families, and we establish the key isometry with $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$. We prove our general continuous- and discrete-time guarantees for Wasserstein polyhedral optimization in Section 4. We apply our framework to mean-field variational inference in Section 5; in particular, the implementation is discussed in Section 5.5.1. Section 6 contains our numerical experiments. Section 7 extends our framework to mixtures of product measures. We conclude with numerous open directions.

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. (2022); 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., 2021). We provide further background information and literature on mean-field VI in Section 5.1, and omit it here to avoid redundancies.

2 Background on optimal transport

In this section, we provide background on optimal transport relevant to our work and refer to Villani (2009); Santambrogio (2015) 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

$$W_2^2(\rho,\mu) = \inf_{T:T_\sharp\rho=\mu} \int \|x - T(x)\|_2^2 \,\mathrm{d}\rho(x)\,, \tag{2}$$

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 Equation (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 $\varphi_\star^*(y) := \sup_{x\in\mathbb{R}^d}\{\langle x, y\rangle - \varphi_\star(x)\}$ is the Fenchel conjugate of $\varphi_\star$. For more information on (differentiable) convex functions and conjugacy, we suggest Rockafellar (1997); Hiriart-Urruty and Lemaréchal (2004).

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

$$\mu_t = (\nabla\varphi_t)_\sharp\mu_0 \coloneqq (\mathrm{id} + t\,(\nabla\varphi^{0\to 1} - \mathrm{id}))_\sharp\mu_0\,. \tag{3}$$

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 $\mathbb{W} \coloneqq (\mathcal{P}_2(\mathbb{R}^d), W_2)$ (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 $\mathcal{C}$ of the Wasserstein space is *geodesically convex* if it is closed under taking geodesics (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 $(\mu_t)_{t\in[0,1]}$.

3 Polyhedral sets in the Wasserstein space

In this section, we establish properties of the constraint set

$$\mathrm{cone}(\mathcal{M})_\sharp\rho \coloneqq \left\{ \left(\textstyle\sum_{T\in\mathcal{M}} \lambda_T T\right)_\sharp\rho \;\middle|\; \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \right\}, \tag{4}$$

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 (4) is valid. Otherwise, (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 $\mathrm{cone}(\mathcal{M})_\sharp\rho$ 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, $\mathrm{cone}(\mathcal{M})_\sharp\rho$ is the following set of Gaussians:

$$\mathrm{cone}(\mathcal{M})_\sharp\rho = \left\{ \mathcal{N}(0,\, (\lambda_1 I + \lambda_2 A^{1/2} + \lambda_3 B^{1/2})^2) \;\middle|\; \lambda \in \mathbb{R}_+^3 \right\}. \tag{5}$$

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 (5). This simple example illustrates that some care is required in order to define convex constraint sets in the Wasserstein space.

3.1 Compatible families of transport maps

In the Gaussian example above, geodesic convexity of $\mathrm{cone}(\mathcal{M})_\sharp\rho$ 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 Boissard et al., 2015; Bigot et al., 2017; Panaretos and Zemel, 2016; Cazelles et al., 2018; Chewi et al., 2021; Werenski et al., 2022). See Panaretos and Zemel (2020) for details.

The main motivation for compatibility is the following theorem.

Theorem 3.1 (Compatibility induces geodesic convexity)**.** *Suppose that $\mathcal{M}$ is compatible. Then,* $\mathrm{cone}(\mathcal{M})_\sharp\rho$ *is a geodesically convex set. Moreover, for any convex subset $\mathcal{K} \subseteq \mathrm{cone}(\mathcal{M})$, 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 3.2, which will show that $\mathrm{cone}(\mathcal{M})_\sharp\rho$ is isometric to a convex subset of a Hilbert space.

Motivated by this theorem, we propose the following definition.

Definition 3.2. Let $\mathcal{M}$ be a compatible and finite family of optimal transport maps. We refer to $\mathrm{cone}(\mathcal{M})_\sharp\rho$ 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 $\mathcal{K} \subseteq \mathrm{cone}(\mathcal{M})$ 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 A.

Lemma 3.3 (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 3.4 (Radial maps)**.** *Let*

$$\mathcal{M} \coloneqq \{x \mapsto g(\|x\|_2)\, x \mid g : \mathbb{R}_+ \to \mathbb{R}_+ \text{ is continuous and strictly increasing}\}.$$

Then, $\mathcal{M}$ is a compatible family.

Lemma 3.5 (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 3.6 (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 $\mathbb{R}^{d_1+d_2}$.*

[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.

Lemma 3.7 (Adding the identity). *Let $\mathcal{M}$ be a compatible family. Then, $\mathcal{M} \cup \{\mathrm{id}\}$ is a compatible family.*

Lemma 3.8 (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 3.9 (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 5.3). In particular, Lemma 3.9 is the starting point for the development of our theory of polyhedral optimization in the Wasserstein space.

Remark 3.10. In our applications of interest, $\mathrm{cone}(\mathcal{M})$ is typically constructed as follows: let $\mathcal{M}_1, \dots, \mathcal{M}_d$ be univariate compatible families (Lemma 3.5). We then take $\mathrm{cone}(\mathcal{M})$ to be the cone generated by the direct sum of $\mathcal{M}_1, \dots, \mathcal{M}_d$ via Lemma 3.6. It is easy to see that a generating family of this cone is the set of maps $x \mapsto (0, \dots, 0, T_i(x_i), 0, \dots, 0)$, where $T_i \in \mathcal{M}_i$. This is a finite family of size $\sum_{i=1}^d |\mathcal{M}_i|$.

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 $T, \tilde{T} \in \mathcal{M}$,

$$\mathsf{d}^2_{\mathrm{LOT},\rho}(T_\sharp\rho, \tilde{T}_\sharp\rho) \coloneqq \|\tilde{T} - T\|^2_{L^2(\rho)} = \|\tilde{T} \circ T^{-1} - \mathrm{id}\|^2_{L^2(T_\sharp\rho)} = W_2^2(T_\sharp\rho, \tilde{T}_\sharp\rho)\,, \tag{6}$$

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 Wang et al. (2013); Basu et al. (2014); Kolouri and Rohde (2015); Kolouri et al. (2016); Park and Thorpe (2018); Cai et al. (2020) 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

$$T^\lambda \coloneqq \sum_{T \in \mathcal{M}} \lambda_T T\,, \qquad \mu_\lambda \coloneqq (T^\lambda)_\sharp\rho\,.$$

We can therefore consider the induced metric on $\mathbb{R}_+^{|\mathcal{M}|}$. A straightforward calculation reveals:

$$\mathsf{d}^2_{\mathrm{LOT},\rho}(\mu_\eta, \mu_\lambda) = \|\textstyle\sum_{T \in \mathcal{M}} (\eta_T - \lambda_T)\, T\|^2_{L^2(\rho)} = (\eta - \lambda)^\top Q (\eta - \lambda) = \|\eta - \lambda\|^2_Q\,,$$

where the matrix Q has entries $Q_{T,\tilde{T}} \coloneqq \langle T, \tilde{T} \rangle_{L^2(\rho)}$ 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 3.11. *Let $\mathcal{M}$ be a finite family of optimal transport maps. Then, $(\mathbb{R}_+^{|\mathcal{M}|}, \|\cdot\|_Q)$ is always isometric to $(\mathrm{cone}(\mathcal{M})_\sharp\rho, \mathsf{d}_{\mathrm{LOT},\rho})$. 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 3.11 paves the way for the application of scalable first-order Euclidean optimization algorithms for minimization problems over polyhedral subsets of the Wasserstein space.

4 Polyhedral optimization in the Wasserstein space

Let $\mathrm{cone}(\mathcal{M})_\sharp\rho$ be polyhedral and recall the Gram matrix Q from Theorem 3.11, 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 $L^2(\rho)$.

4.1 Continuous-time gradient flow

The isometry of Section 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. The latter gradient flow can be written explicitly as

$$\dot{\lambda}(t) = -Q^{-1}\,\nabla_\lambda \mathcal{F}(\mu_{\lambda(t)})\,. \tag{7}$$

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 establishes convergence rates for this continuous-time flow; see Lambert et al. (2022, Appendix D) for a proof.

Theorem 4.1. *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).*

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) $\mathcal{F}(\mu_{\lambda(t)}) - \mathcal{F}(\mu_\star) \leq \exp(-2mt)\,(\mathcal{F}(\mu_{\lambda(0)}) - \mathcal{F}(\mu_\star))$.*

4.2 Time-discretization made easy

Appealing to the isometry in Section 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

$$\min_{\lambda \in \mathbb{R}_+^{|\mathcal{M}|}} \; \mathcal{F}(\mu_\lambda)\,. \tag{8}$$

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 5.7 and Appendix C.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 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.

4.2.1 Accelerated projected gradient descent

Starting at an initial point $\lambda^{(0)} \in K$, we can solve (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 1. 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 4.2 (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 1.*

1. *If* $m = 0$*, then* $\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \lesssim M t^{-2}\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star)\,.$
2. *If* $m > 0$*, then for* $\kappa \coloneqq M/m$*,*
   (a) $W_2^2(\mu_{\lambda^{(t)}}, \mu_\star) \lesssim \kappa \exp(-t/\sqrt{\kappa})\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star).$
   (b) $\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \leq (1 - 1/\sqrt{\kappa})^t \left(\mathcal{F}(\mu_{\lambda^{(0)}}) - \mathcal{F}(\mu_\star) + \frac{m}{2}\, W_2^2(\mu_{\lambda^{(0)}}, \mu_\star)\right).$

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 $\hat{\nabla}_\lambda \mathcal{F}(\mu_\lambda)$.[2] 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.

[2] 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 1 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 $\mathcal{M}$
Set $\eta^{(0)} = \lambda^{(0)}$, $\kappa := M/m$ if $m > 0$, and $\gamma_{(0)} = 1$ if $m = 0$.
for $t = 0, 1, 2, 3, \dots$ **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 if
end for

Algorithm 2 Stochastic 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 $\mathcal{M}$, fixed step-size $h > 0$, and unbiased stochastic gradient oracle $\hat{\nabla}_\lambda \mathcal{F}(\cdot)$.
for $t = 0, 1, 2, 3, \dots$ **do**
  $\lambda^{(t+1)} \leftarrow \mathrm{proj}_{K,Q}(\lambda^{(t)} - h\, Q^{-1} \hat{\nabla}_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}))$
end for

We provide a description of stochastic projected gradient descent (SPGD) in Algorithm 2, and convergence analysis in Theorem 4.3 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

$$\mathbb{E}[\|Q^{-1} (\hat{\nabla}_\lambda \mathcal{F}(\mu_\lambda) - \nabla_\lambda \mathcal{F}(\mu_\lambda))\|_Q^2] \leq c_0 + c_1\, \mathbb{E}[W_2^2(\mu_\lambda, \mu_\star)]\,.$$

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 4.3 (Convergence results for SPGD)**.** *Let* cone$(\mathcal{M})_\sharp \rho$ *be polyhedral and* $\mathcal{K} \subseteq$ 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* $(\lambda^{(t)} : t = 0, 1, 2, 3 \dots)$ *denote the iterates of Algorithm 2. If we choose* $h \asymp \frac{m\varepsilon^2}{c_0} \leq \frac{m}{2c_1} \wedge \frac{1}{2\kappa M}$*, 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 4.3 in Appendix B.

Algorithm 3 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**
  $j^* \leftarrow \arg\min_{j \in [J]} \langle \nabla_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}), e_j - \lambda^{(t)} \rangle$
  $\lambda^{(t+1)} \leftarrow (1 - \alpha^{(t)})\, \lambda^{(t)} + \alpha^{(t)}\, e_{j^*}$ ▷ $\alpha^{(t)} = \frac{2}{t+2}$ is a standard step size choice
end for

4.2.3 Frank–Wolfe

In this section, we consider optimization over a *polytope*, i.e., a set of the form

$$\operatorname{conv}(\mathcal{M})_\sharp \rho := \left\{ \left(\textstyle\sum_{T \in \mathcal{M}} \lambda_T T \right)_\sharp \rho \;\middle|\; \lambda \in \Delta_{|\mathcal{M}|} \right\},$$

where $\mathcal{M}$ is a finite family of compatible maps and $\Delta_{|\mathcal{M}|}$ denotes the $|\mathcal{M}|$-dimensional simplex. Note that $\operatorname{conv}(\mathcal{M}) \subseteq \operatorname{cone}(\mathcal{M})$, where $\operatorname{cone}(\mathcal{M})_\sharp \rho$ is polyhedral, so that $\operatorname{conv}(\mathcal{M})$ 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 4.2 therefore apply to optimization over $\operatorname{conv}(\mathcal{M})_\sharp \rho$.

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:

$$\eta^{(t)} = \operatorname*{arg\,min}_{\eta \in \Delta_{|\mathcal{M}|}} \langle \nabla_\lambda \mathcal{F}(\mu_{\lambda^{(t)}}), \eta - \lambda^{(t)} \rangle . \tag{9}$$

Finding this direction can be substantially cheaper than the projection step in Algorithm 1. Indeed, the sub-problem (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 3. 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 4.4 (Convergence results for FW)**.** *Suppose that $\mathcal{F}$ is geodesically convex and M-smooth over $\operatorname{conv}(\mathcal{M})_\sharp \rho$, 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 3, with $\alpha^{(t)} = 2/(t+2)$. Then,*

$$\mathcal{F}(\mu_{\lambda^{(t)}}) - \mathcal{F}(\mu_\star) \lesssim M t^{-1} \operatorname{diam}\big(\operatorname{conv}(\mathcal{M})_\sharp \rho\big)^2 . \tag{10}$$

Via the isometry in Section 3.2, $\operatorname{diam}(\operatorname{conv}(\mathcal{M})_\sharp \rho)$ equals the diameter of $\operatorname{conv}(\mathcal{M})$ 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 4.5. 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 $\delta > 0$ 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.

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 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 3.9 and Lemma 3.7, $\mathrm{cone}(\mathcal{M} \cup \{\mathrm{id}\})$ is a compatible family, and then we can restrict to the convex subset $\mathcal{K} \subseteq \mathrm{cone}(\mathcal{M} \cup \{\mathrm{id}\})$ corresponding to λ for which the coefficient λ_{id} in front of id is α. The guarantees of Section 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 3.8. 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 Proposition 5.10) 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.

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 $\mathcal{C}$, the VI optimization problem reads

$$\pi_{\mathcal{C}}^{\star} \in \arg\min_{\mu \in \mathcal{C}} \mathrm{KL}(\mu \,\|\, \pi) \coloneqq \arg\min_{\mu \in \mathcal{C}} \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z \,, \tag{11}$$

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 5.3 and Section 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 5.9).

Lemma 5.1. *Suppose $\mathcal{C}$ 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 $\mathcal{C}$.*

Despite the widespread use of variational inference in numerous settings (see for example Wainwright and Jordan, 2008; Blei et al., 2017), explicit guarantees have only recently been established for a few constraint families. Recently, Lambert et al. (2022); Diao et al. (2023) 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 5.2 in Section 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 5.4 in Section 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 5.6, Theorem 5.7, and Theorem 5.8 in Section 5.3).

4. Provide convergence guarantees for optimizing the KL divergence over these polyhedral sets (see Theorem 5.11 in Section 5.4).

5. Describe implementation details for our final algorithm (see Section 5.5.1).

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

$$\pi^\star \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{R})^{\otimes d}} \mathrm{KL}(\mu \,\|\, \pi)\,, \tag{12}$$

where, by design, the constraint set forces the minimizers to be of the form

$$\pi^\star(x_1, \ldots, x_d) = \bigotimes_{i=1}^{d} \pi_i^\star(x_i)\,. \tag{13}$$

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 (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.

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.[3]

Mean-field equations. Via calculus of variations, one can readily derive the following system of *mean-field equations* from (12): for $i \in [d]$,

$$\pi_i^\star(x_i) \propto \exp\Big(-\int_{\mathbb{R}^{d-1}} V(x_1, \ldots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\Big)\,. \tag{14}$$

These are also sometimes called *self-consistency equations*; we give a derivation in Appendix C.1. From the structure of $\pi^\star$, we can prove the following result.

Proposition 5.2. *Suppose that π is well-conditioned (WC). Then, (12) admits a* unique *minimizer of the form (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 5.1 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 (14).

[3] Tran et al. (2023) also study discrete-time rates for mean-field VI, though their draft has numerous errors, and incorrectly applies several prior results.

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 (12) to

$$\pi^\star_\diamond := (T^\star_\diamond)_\sharp \rho \in \arg\min_{\mu \in \mathcal{P}_\diamond} \mathrm{KL}(\mu \,\|\, \pi)\,, \tag{15}$$

where $\mathcal{P}_\diamond$ is a polyhedral subset of the Wasserstein space (Definition 3.2). Recall that polyhedral subsets of the Wasserstein space are geodesically convex (see Theorem 3.1). Combined with Lemma 5.1, the following corollary is immediate.

Corollary 5.3. *Suppose that $\mathcal{P}_\diamond$ is a polyhedral subset of the Wasserstein space, and that π is well-conditioned (WC). Then the minimizer to (15) is* unique*, denoted by $\pi^\star_\diamond$.*

Borrowing inspiration from the existing literature combining normalizing flows and optimal transport (Chen et al., 2018; Finlay et al., 2020a,b; Huang et al., 2021), 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. In this work, 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 5.3.

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 5.4. *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 (12), written $T^\star(x) = (T^\star_1(x_1), \dots, T^\star_d(x_d))$. Each map $T^\star_i$ satisfies*

$$\sqrt{1/L_V} \le (T^\star_i)' \le \sqrt{1/\ell_V}\,.$$

Moreover, we have the higher-order regularity bounds

$$|(T^\star_i)''(x)| \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,(1 + |x|)\,, \qquad \text{and} \qquad |(T^\star_i)'''(x)| \lesssim \frac{\kappa^2}{\sqrt{\ell_V}}\,(1 + |x|^2)\,. \tag{16}$$

The bounds on $(T^\star_i)'$ in Theorem 5.4 in fact follow immediately from two landmark results in optimal transport, and Proposition 5.2. 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 (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 C.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 5.8 in Section 5.3.

Remark 5.5. In prior works that statistically estimate optimal transport maps on the basis of samples (such as Deb et al. (2021); Hütter and Rigollet (2021); Manole et al. (2021); Pooladian and Niles-Weed (2021); Divol et al. (2022)), 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)).

5.3 Approximating the mean-field solution with compatible maps

So, Theorem 5.4 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 $\mathcal{P}_\diamond := \underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})_\sharp \rho$,

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

is close to $\pi^\star$ in the Wasserstein distance. Then, Section 5.4 will provide guarantees for computing $\pi^\star_\diamond$ 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 $\int \|D(\bar{T} - \hat{T})\|^2_{\mathrm{F}}\,\mathrm{d}\rho$.

Theorem 5.6. *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 5.4, there exists $\hat{T} \in \underline{\mathrm{cone}}(\mathcal{M}; \alpha\,\mathrm{id})$, with $\alpha = 1/\sqrt{L_V}$, such that $W_2(\bar{T}_\sharp\rho, \hat{T}_\sharp\rho) = \|\bar{T} - \hat{T}\|_{L^2(\rho)} \le \varepsilon/\ell_V^{1/2}$ 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 C.3.

This piecewise linear construction exploits the smoothness of $\bar{T}$ up to order two, but no further. On the other hand, from Theorem 5.4 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 C.3.

Theorem 5.7. *There exists a polyhedral set $\mathcal{P}_\diamond$ with an explicit generating family (see Appendix C.3) of size $\widetilde{O}(\kappa^{2/3} d^{7/6} / \varepsilon^{1/3})$ with the following property. In the setting of Theorem 5.6, assume also that each component $\bar{T}_i$ of $\bar{T}$ obeys the third derivative bound in Theorem 5.4. Then, there exists $\hat{T} \in \mathcal{P}_\diamond$ such that $W_2(\bar{T}_\sharp \rho, \hat{T}_\sharp \rho) = \|\bar{T} - \hat{T}\|_{L^2(\rho)} \le \varepsilon / \ell_V^{1/2}$ and $\|D(\bar{T} - \hat{T})\|_{L^2(\rho)} \lesssim \kappa^{2/3} d^{1/6} \varepsilon^{2/3} / \ell_V^{1/2}$.*

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* $\pi_\diamond^\star$ of the KL divergence over $\mathcal{P}_\diamond$, and hence can be computed using the algorithms in Section 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 5.6 and Theorem 5.7). The proof of Theorem 5.8 is also found in Appendix C.3.

Theorem 5.8. *The mean-field solution $\pi^\star$ is close to the* minimizer *$\pi_\diamond^\star$ of the KL divergence over $\mathcal{P}_\diamond$ with corresponding generating family $\mathcal{M}$, in the sense that $\sqrt{\ell_V}\, W_2(\pi_\diamond^\star, \pi^\star) \le \varepsilon$, in the following two cases.*

1. *For the piecewise linear construction of Theorem 5.6, the size of the family is bounded by $|\mathcal{M}| \le \widetilde{O}(\kappa^2 d^{3/2} / \varepsilon)$.*
2. *For the higher-order approximation scheme of Theorem 5.7, the size of the family is bounded by $|\mathcal{M}| \le \widetilde{O}(\kappa^{3/2} d^{5/4} / \varepsilon^{1/2})$.*

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

$$\mathrm{KL}(\mu \,\|\, \pi) = \mathcal{V}(\mu) + \mathcal{H}(\mu) \coloneqq \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z\,, \tag{17}$$

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 5.6, although the discussion below can be adapted to more general polyhedral sets. As in Section 4.2, we apply Euclidean optimization algorithms over the parameterization of $\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})$; see Section 5.5.1 for a discussion of implementation.

In order to apply the algorithmic guarantees from Section 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 3.1 and Section 4.3).

Proposition 5.9 (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 $\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})$, since $\underline{\mathrm{cone}}(\mathcal{M};\, \alpha\,\mathrm{id})$ is augmented by translations. Assuming now that $\mathcal{M}$ is centered, we recall the Gram matrix Q with entries $Q_{T,\tilde{T}} \coloneqq \langle T, \tilde{T} \rangle_{L^2(\rho)}$. We also form the Gram matrix of the Jacobians, $Q^{(1)}$, with entries $Q^{(1)}_{T,\tilde{T}} \coloneqq \langle DT, D\tilde{T} \rangle_{L^2(\rho)} \coloneqq \int \langle DT, D\tilde{T} \rangle \,\mathrm{d}\rho$. 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 3.6 (c.f. Remark 3.10), as in our approximation results (Section 5.3), the matrices Q and $\tilde{Q}$ have a $d \times d$ block diagonal structure; see Section 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".[4]

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 5.10 (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*

$$M \leq L_V + \Upsilon/\alpha^2\,.$$

From Theorem 5.4, 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 4.2.

Theorem 5.11 (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 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 5.11 with our approximation result in Theorem 5.8, 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. To the best of our knowledge, this constitutes the first *accelerated* and *end-to-end* convergence result for mean-field VI. See Section 5.1 for comparisons with the literature.

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.

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 $\mathbb{R}^d$.

[4] However, if one wishes to maintain the same quality of approximation in high dimension, our approximation results in Section 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. We leave the detailed quantitative analysis of Υ for future work.

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 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 $\mathcal{M}_1 = \{T_1, \dots, T_J\}$.

Parameterization of the cone. As discussed in Section 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 $Q_{(i,j);(i',j')} \coloneqq \int \langle T_j(x_i)\, e_i, T_{j'}(x_{i'})\, e_{i'} \rangle\, \rho(\mathrm{d}x)$. 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 $Q^{\mathcal{M}_1}_{j,j'} \coloneqq \int T_j T_{j'}\, \mathrm{d}\rho_1$ (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 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,

$$\mathrm{KL}((T^{\lambda,v})_\sharp \rho \,\|\, \pi) = \int [V(T^{\lambda,v}(x)) - \log\det DT^\lambda(x)]\, \rho(\mathrm{d}x) + \int \log\rho \,\mathrm{d}\rho + \log Z\,.$$

The partial derivatives are therefore computed to be

$$\begin{aligned} \partial_{\lambda_{i,j}} \mathrm{KL}((T^{\lambda,v})_\sharp \rho \,\|\, \pi) &= \int \left[\partial_i V(T^{\lambda,v}(x))\, T_j(x_i) - \langle e_i, (DT^\lambda)^{-1}(x)\, e_i \rangle\, T_j'(x_i)\right] \rho(\mathrm{d}x)\,, \\ \nabla_v \mathrm{KL}((T^{\lambda,v})_\sharp \rho \,\|\, \pi) &= \int \nabla V(T^{\lambda,v}(x))\, \rho(\mathrm{d}x)\,. \end{aligned} \tag{18}$$

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:

$$DT^\lambda(x) = \alpha I_d + \mathrm{diag}\left(\sum_{j=1}^J \lambda_{i,j} T_j'(x_i)\right)_{i=1}^d.$$

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 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).

5.5.2 Convergence for stochastic mean-field VI

In Section 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 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 5.6, 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 $\bar{Q}_{T,\tilde{T}}(x) = \langle T(x), \tilde{T}(x)\rangle$ for $T, \tilde{T} \in \mathcal{M}_1$. Here, $J := |\mathcal{M}_1|$.

The following lemma established a variance bound of the type (VB) which, when combined with Theorem 4.3, proves Theorem 5.13.

Lemma 5.12 (Variance bound for stochastic mean-field VI)**.** *Assume that 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 C.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:*

$$\mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(Q^{-1/2}\,\hat{\nabla}_\lambda \mathrm{KL}(\mu_\lambda\,\|\,\pi))] \le 2L_V^2 \Xi J\, W_2^2(\mu_\lambda, \pi^\star_\diamond) + 4L_V \Xi J\,(L_V\, W_2^2(\pi^\star_\diamond, \pi^\star) + \kappa d)\,.$$

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 5.3. The next theorem follows immediately from Theorem 4.3 and the previous lemma.

Theorem 5.13 (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, 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 $h \asymp \varepsilon^2/(L_V \Xi \kappa J d)$.

6 Numerical experiments

We showcase our proposed MFVI algorithm on numerical experiments. Experimental details are deferred to Appendix D, 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 5.6) 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.

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.

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 Appendix D.2 for a calculation of this fact).

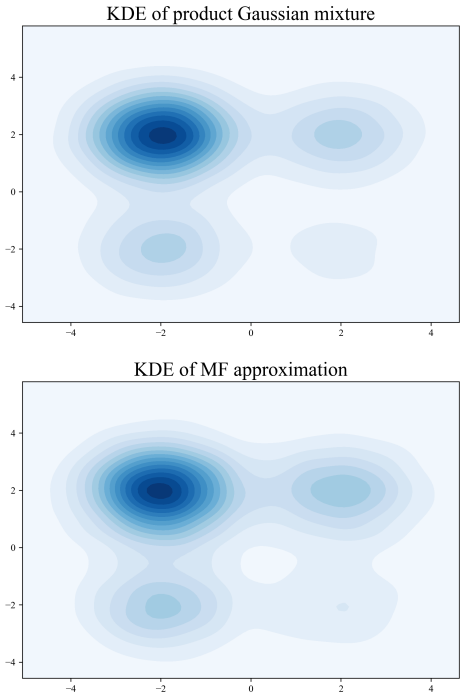

Figure 1: KDEs for the optimal product Gaussian mixture and our algorithm.

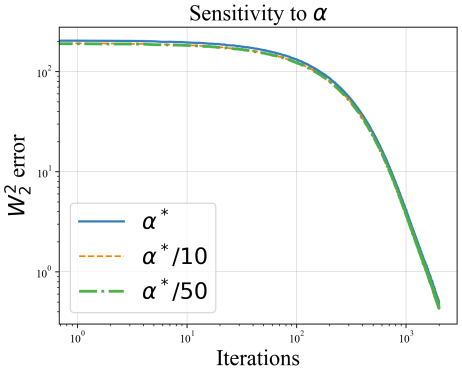

Figure 2: Our algorithm is **robust to the choice of** α.

In Figure 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.

6.3 Synthetic Bayesian logistic regression

As a final example, we turn to Bayesian logistic regression on synthetic data; precise details are deferred to Appendix D.3. In summary, we are given i.i.d. data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$, from which we want to recover a parameter θ. When assuming an *improper* (Lebesgue) prior on θ, the posterior is given by

$$V(\theta) = \sum_{i=1}^{n} \left[\log(1 + \exp(\theta^\top X_i)) - Y_i \theta^\top X_i \right].$$

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 3 contains the 20 marginals for both our approach and LMC, which are closely aligned.

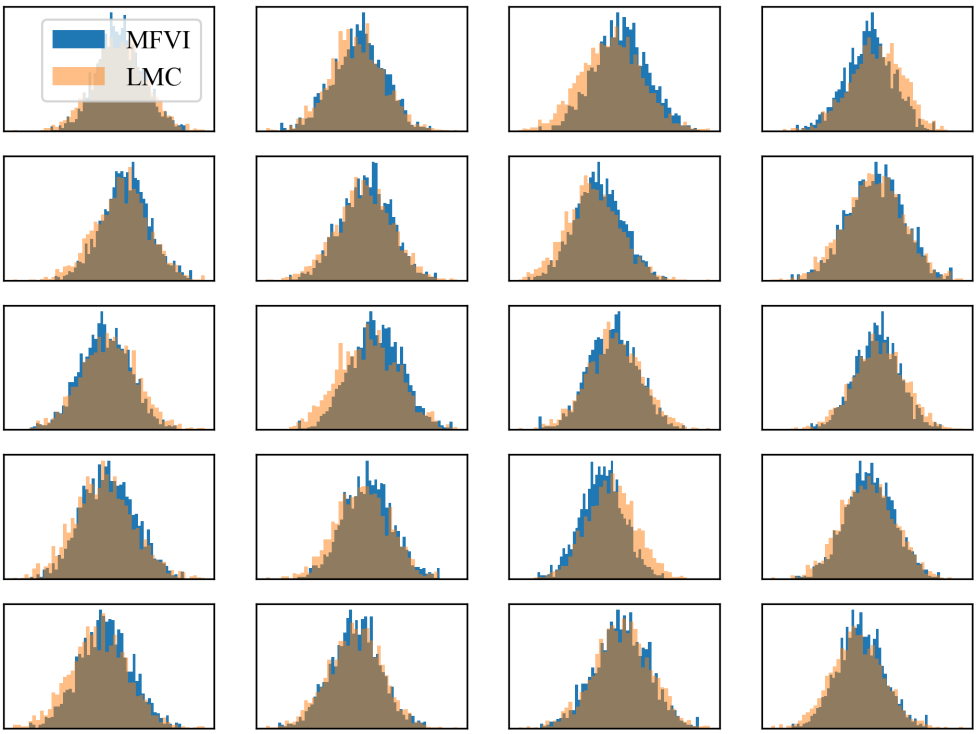

Figure 3: Histograms of the first ten marginals computed via our mean-field VI algorithm vs. Langevin Monte Carlo for a 20-dimensional Bayesian logistic regression example.

7 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 (Eldan, 2018; Eldan and Gross, 2018; Austin, 2019; Jain et al., 2019).

In Section 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 E.

Theorem 7.1. *The Wasserstein gradient flow of* $P \mapsto \mathrm{KL}(\mu_P \,\|\, \pi)$ *is the flow* $(P^{(t)})_{t \geq 0}$

specified as follows. For each $t \geq 0$, $P^{(t)}$ 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

$$P = \frac{1}{K}\sum_{k=1}^{K} \delta_{(\lambda[k], v[k])}\,, \qquad \mu_P = \frac{1}{K}\sum_{k=1}^{K} (T^{\lambda[k], v[k]})_\sharp \rho\,. \tag{19}$$

The system of ODEs above then becomes an interacting particle system:

$$\dot{\lambda}_T^{(t)}[k] = -\int \left\langle \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}, T \right\rangle \mathrm{d}\rho\,, \qquad \text{for } T \in \mathcal{M}\,,$$
$$\dot{v}^{(t)}[k] = -\int \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho\,.$$

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.

The dynamics (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 geometry, which captures unbalanced optimal transport (Liero et al., 2016; Chizat et al., 2018; Liero et al., 2018). The use of this geometry for sampling was pioneered in Lu et al. (2019).

Theorem 7.2. *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 $t \geq 0$, 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,*

$$\dot{\lambda}_T^{(t)}[k] = -\int \left\langle \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}, T \right\rangle \mathrm{d}\rho\,, \qquad \textit{for } T \in \mathcal{M}\,,$$
$$\dot{v}^{(t)}[k] = -\int \nabla \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho\,,$$
$$\dot{r}^{(t)}[k] = -\Big(\int \log \frac{\mu_{P^{(t)}}}{\pi} \circ T^{\lambda^{(t)}[k], v^{(t)}[k]}\, \mathrm{d}\rho - \int \log \frac{\mu_{P^{(t)}}}{\pi}\, \mathrm{d}\mu_{P^{(t)}} \Big)\, r^{(t)}[k]\,.$$

We leave it as an open question to obtain convergence rates for this flow.

8 Conclusion

In this paper, we have put forth a definition of polyhedral subsets of the Wasserstein space, resting upon the notion of compatibility. Compatibility induces an isometry with Euclidean geometry, allowing for the application of scalable first-order algorithms for convex optimization. We then applied our framework to yield end-to-end guarantees for mean-field VI, showing first that the mean-field solution lies arbitrarily close to the KL minimizer over a polyhedral set, and then providing complexity guarantees for computing that KL minimizer.

Looking ahead, our work leaves open several questions which could be fruitful for future study. Regarding our notion of Wasserstein polyhedra, one could investigate statistical convergence guarantees as in Gunsilius et al. (2022), or develop further interesting applications of polyhedral optimization.

As for our application to mean-field VI, one direction to explore would be to explicitly quantify the dependence of the parameter Υ on $\mathcal{M}$, thereby initiating the search for optimal choices of $\mathcal{M}$. We also ask if our analysis of SGD can be improved, e.g., by removing the boundedness assumption in (Ξ). Finally, another important question is to move beyond the well-conditioned setting.

Acknowledgements

The authors are grateful to Jonathan Niles-Weed for helpful discussions at all stages of this work. YJ acknowledges support from NYU through the SURE program, where this project originated under the supervision of AAP. SC acknowledges the support of the Eric and Wendy Schmidt Fund at the Institute for Advanced Study. AAP acknowledges the support of Meta AI research, as well as the National Science Foundation through NSF Award 1922658.

References

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*.

Altschuler, J. M., Chewi, S., Gerber, P. R., and Stromme, A. J. (2021). Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P. S., Vaughan, J. W., and Dauphin, Y., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22132–22145. Curran Associates, Inc.

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.

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.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

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*.

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.

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.

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.

Bubeck, S. (2015). Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

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.

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. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2019). Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278.

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. *Reports. Mathematical*, 361:1471–1482.

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.

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.

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.

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.

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*.

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.

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.

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.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

Gunsilius, F., Hsieh, M. H., and Lee, M. J. (2022). Tangential Wasserstein projections. *arXiv preprint arXiv:2207.14727*.

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. (2021). Convex potential flows: universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.*, 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.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17.

Kent, C., Blanchet, J., and Glynn, P. (2021). Frank–Wolfe methods in probability space. *arXiv preprint arXiv:2105.05352*.

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.

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.

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.

Lu, Y., Lu, J., and Nolen, J. (2019). Accelerating Langevin sampling with birth-death. *arXiv preprint 1905.09863*.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*.

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. 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.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174.

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.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.

Rockafellar, R. T. (1997). *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Tran, M.-N., Tseng, P., and Kohn, R. (2023). Particle mean field variational Bayes. *arXiv preprint arXiv:2303.13930*.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

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., 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.

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.

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*.

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.

A Proofs for Section 3

Proof of Lemma 3.3. 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 3.4. See Panaretos and Zemel (2020, Section 2.3.2). □

Proof of Lemma 3.5. 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 3.6. 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 3.7. For any $T \in \mathcal{M}$, T and T^{-1} are both gradients of convex functions, so the claim is immediate. □

Proof of Lemma 3.8. 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 3.9. 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

$$\tilde{T}^{-1} = \Big(\sum_{S\in\mathcal{M}} \eta_S S\Big) \circ T^{-1} = \sum_{S\in\mathcal{M}} \eta_S \left(S \circ T^{-1}\right).$$

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. □

B Proofs for Section 4.2

Proof of Theorem 4.3. 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

$$\begin{aligned} \mathbb{E} \|\lambda^{(t+1)} - \lambda^\star\|_Q^2 &\le (1 - \tfrac{mh}{2})^t \, \|\lambda^{(0)} - \lambda^\star\|_Q^2 + \frac{2c_0 h}{m} \\ &\le \exp(-mht/2) \, \|\lambda^{(0)} - \lambda^\star\|_Q^2 + \frac{2c_0 h}{m} \, . \end{aligned}$$

Choosing $h \asymp m\varepsilon^2/c_0$ concludes the proof. □

C Proofs for Section 5

C.1 Proofs for Section 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

$$\Big[\delta_{\mu_i} \mathrm{KL}\Big(\bigotimes_{j=1}^{d} \mu_j \,\Big\|\, \pi\Big)\Big](x_i) = \int V(x_1, \dots, x_d) \bigotimes_{j \neq i} \mu_j(\mathrm{d}x_j) + \log \mu_i(x_i) + \text{const} \,.$$

At optimality, the first variation must equal a constant, which leads to

$$\pi_i^\star(x_i) \propto \exp\Big(- \int V(x_1, \dots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j)\Big) \,.$$

C.2 Proofs for Section 5.2

In this section, we prove the regularity bounds on the optimal transport maps given as Theorem 5.4. 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 (14) that $\pi_i^\star \propto \exp(-V_i)$, where

$$V_i(x_i) \coloneqq \int V(x_1, \dots, x_d) \bigotimes_{j \neq i} \pi_j^\star(\mathrm{d}x_j) \,.$$

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 C.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 $T' \leq \beta$,

$$|T(0) - m| = |\mathbb{E}(T(0) - T(Z))| \leq \beta\, \mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}\,\beta \,. \qquad \square$$

Lemma C.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, $|m - \tilde{m}| \leq 1/\sqrt{\ell_V}$.*

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 5.4.

Proof of Theorem 5.4. 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

$$\log \pi_i^\star \circ T_i^\star(x) = -\frac{x^2}{2} - \log(T_i^\star)'(x) - \tfrac{1}{2}\log(2\pi)\,. \tag{20}$$

Differentiating once yields

$$(\log \pi_i^\star \circ T_i^\star)'(x) = V_i'(T_i^\star(x))\,(T_i^\star)'(x) = -x - \frac{(T_i^\star)''(x)}{(T_i^\star)'(x)}\,. \tag{21}$$

Rearranging to isolate $(T_i^\star)''$ gives

$$(T_i^\star)''(x) = -(T_i^\star)'(x)\left(x + V_i'(T_i^\star(x))\,(T_i^\star)'(x)\right). \tag{22}$$

Let $m_i^\star$ and $\tilde{m}_i^\star$ denote the mean and mode of $\pi_i^\star$ respectively. Recall also that $0 < 1/\sqrt{L_V} \le (T_i^\star)' \le 1/\sqrt{\ell_V}$. By Lemma C.1 and Lemma C.2,

$$\begin{aligned}
|V_i'(T_i^\star(x))| &\le \underbrace{|V_i'(\tilde{m}_i^\star)|}_{=0} + L_V\,|T_i^\star(x) - \tilde{m}_i^\star| \\
&\le L_V\,(|T_i^\star(x) - T_i^\star(0)| + |T_i^\star(0) - m_i^\star| + |m_i^\star - \tilde{m}_i^\star|) \\
&\le L_V\left(\frac{1}{\sqrt{\ell_V}}\,|x| + \sqrt{\frac{2}{\pi}}\,\frac{1}{\sqrt{\ell_V}} + \frac{1}{\sqrt{\ell_V}}\right) \lesssim \frac{L_V}{\sqrt{\ell_V}}\,(1+|x|)\,.
\end{aligned}$$

Substituting this into (22), we obtain

$$|(T_i^\star)''(x)| \lesssim \frac{1}{\sqrt{\ell_V}}\left(|x| + \frac{L_V}{\ell_V}\,(1+|x|)\right) \lesssim \frac{\kappa}{\sqrt{\ell_V}}\,(1+|x|)\,.$$

For the third derivative control, we differentiate (21) again to yield

$$\begin{aligned}
(\log \pi_i^\star \circ T_i^\star)''(x) &= V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2 + V_i'(T_i^\star(x))\,(T_i^\star)''(x) \\
&= -1 - \frac{(T_i^\star)'''(x)\,(T_i^\star)'(x) - (T_i^\star)''(x)^2}{(T_i^\star)'(x)^2}\,.
\end{aligned}$$

Again, we rearrange and isolate, giving

$$(T_i^\star)'''(x) = \frac{(T_i^\star)''(x)^2}{(T_i^\star)'(x)} - (T_i^\star)'(x)\left(1 + V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2 + V_i'(T_i^\star(x))\,(T_i^\star)''(x)\right).$$

Taking absolute values, we can collect the terms one by one:

$$
\begin{aligned}
|(T_i^\star)''(x)^2/(T_i^\star)'(x)| &\lesssim \frac{\kappa^{3/2}}{\sqrt{\ell_V}}\,(1+|x|^2)\,,\\
|V_i''(T_i^\star(x))\,(T_i^\star)'(x)^2| &\le \kappa\,,\\
|V_i'(T_i^\star(x))\,(T_i^\star)''(x)| &\lesssim \frac{L_V}{\sqrt{\ell_V}}\,(1+|x|)\cdot\frac{\kappa}{\sqrt{\ell_V}}\,(1+|x|) \lesssim \kappa^2\,(1+|x|^2)\,.
\end{aligned}
$$

Hence, the final bound scales as

$$
|(T_i^\star)'''(x)| \lesssim \frac{\kappa^2}{\sqrt{\ell_V}}\,(1+|x|^2)\,. \qquad \square
$$

C.3 Proofs for Section 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

$$
\psi:\mathbb{R}\to\mathbb{R}\,,\qquad \psi(x)\coloneqq\begin{cases}0\,, & x\le 0\,,\\ x\,, & x\in[0,1]\,,\\ 1\,, & x\ge 1\,.\end{cases}
$$

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 5.6. Owing to the isometry, we wish to show that we can find a map $\hat{T}\in\underline{\mathrm{cone}}(\mathcal{M};\,\alpha\,\mathrm{id})$, with $\alpha=1/\sqrt{L_V}$, such that

$$
\|\bar{T}-\hat{T}\|_{L^2(\rho)}^2\le d\varepsilon_0^2 \qquad\text{and}\qquad \|D(\bar{T}-\hat{T})\|_{L^2(\rho)}^2\le d\varepsilon_1^2\,. \tag{23}
$$

Here, $\|D(\bar{T}-\hat{T})\|_{L^2(\rho)}^2\coloneqq\int\|D(\bar{T}-\hat{T})\|_{\mathrm{F}}^2\,\mathrm{d}\rho$.

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 (23) holds. However, from the structure of $\mathcal{M}$, the problem now separates across the coordinates and it suffices to prove this statement with $d=1$.

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 implies

$$
|\bar{T}(x)-\hat{T}(x)|\le\frac{1}{2\sqrt{\ell_V}}\,(|x|-R)\,,\qquad\text{for }|x|\ge R\,.
$$

The error contributed by the tails is therefore bounded by

$$\int_{\mathbb{R}\setminus(-R,+R)} |\bar{T} - \hat{T}|^2 \,\mathrm{d}\rho \le \frac{1}{4\sqrt{2\pi}\,\ell_V} \int_{\mathbb{R}\setminus(-R,+R)} (|x| - R)^2 \exp(-x^2/2)\,\mathrm{d}x\,.$$

Similarly,

$$|\bar{T}'(x) - \hat{T}'(x)| \le 1/\sqrt{\ell_V}\,, \qquad \text{for } |x| \ge R\,,$$

which gives

$$\int_{\mathbb{R}\setminus(-R,+R)} |\bar{T}' - \hat{T}'|^2 \,\mathrm{d}\rho \le \frac{1}{\sqrt{2\pi}\,\ell_V} \int_{\mathbb{R}\setminus(-R,+R)} \exp(-x^2/2)\,\mathrm{d}x\,.$$

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 $\|\bar{T} - \hat{T}\|^2_{L^2(\rho)} \vee \|\bar{T}' - \hat{T}'\|^2_{L^2(\rho)} \lesssim \varepsilon^2$.

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 $\varepsilon/\sqrt{d}$.

Size of the generating family. Finally, the size of $\mathcal{M}$ is $O(R/\delta) = \widetilde{O}(\kappa^{1/2} d^{1/4}/\varepsilon^{1/2})$, 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 5.4, 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 $\psi^{\mathrm{lin}}(x) \coloneqq x$ for $x \in \mathbb{R}$.
- **Piecewise quadratics.** Define the piecewise quadratic

$$\psi^{\mathrm{quad},\pm}(x) \coloneqq \pm \begin{cases} 0\,, & x \le 0\,, \\ x^2\,, & x \in [0,1]\,, \\ 2x-1\,, & x \ge 1\,. \end{cases}$$

- **Piecewise cubics.** Define the piecewise cubic,

$$\psi^{\mathrm{cub},\pm}(x) \coloneqq \pm \begin{cases} 0\,, & x \le 0\,, \\ x^2\,(3-2x)\,, & x \in [0,1]\,, \\ 1\,, & x \ge 1\,. \end{cases}$$

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 $\psi_i(x) \coloneqq \psi(x_i)$. Also, given a sub-interval $I = [a, a+\delta]$, we define the map $\psi_{I,i} : \mathbb{R}^d \to \mathbb{R}^d$ via $\psi_{I,i}(x) \coloneqq \psi(\delta^{-1}\,(x_i - a))$.

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 $\mathbb{R}_+^{|\mathcal{M}|}$:

$$\begin{aligned} K \coloneqq \Big\{ \lambda \in \mathbb{R}_+^{|\mathcal{M}|} \;\Big|\; \forall i \in [d],\ \frac{2}{\delta} \sum_{I \in \mathcal{I}} (\lambda_{I,i}^{\mathrm{quad},+} - \lambda_{I,i}^{\mathrm{quad},-}) + \lambda_i^{\mathrm{lin}} \ge 0\,, \\ \text{and} \qquad \forall I \in \mathcal{I},\ \forall i \in [d],\ \frac{6\lambda^{\mathrm{cub},-}}{\delta} \le \frac{\alpha}{2} \Big\}\,. \end{aligned} \tag{24}$$

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 $\alpha\,\mathrm{id}$, 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 5.10), 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 5.7. We start with the same reductions as in the proof of Theorem 5.6, 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 $\bar{T}$,

$$|\bar{T}'(x) - \hat{T}'(x)| = |(\bar{T}''(c_1) - \bar{T}''(c_2)) \, (x-a)| \lesssim \frac{\kappa^2 R^2}{\sqrt{\ell_V}} \, \delta^2 . \tag{25}$$

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

$$|\bar{T}(x) - \hat{T}(x)| = \left| \int_a^x (\bar{T}'(y) - \hat{T}'(y)) \, \mathrm{d}y \right| \lesssim \frac{\kappa^2 R^2}{\sqrt{\ell_V}} \, \delta^3 . \tag{26}$$

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 (26), it follows that the coefficients of the piecewise cubics that we add are small: $0 \le \lambda_I^{\mathrm{cub},\pm} \lesssim \kappa^2 R^2 \delta^3 / \sqrt{\ell_V}$. In particular, the constraint on $\lambda_I^{\mathrm{cub},-}$ in (24) 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 (25) and (26). 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 (25) is intact. Similarly, (26) is also intact, either by integrating (25) or by using the bound on the coefficients of the piecewise cubics. The proof is concluded via bookkeeping. □

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 5.10, 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 C.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 5.6). Then,*

$$\mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi_\diamond^\star \,\|\, \pi) \le \mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi^\star \,\|\, \pi) \lesssim \kappa^3 d^{1/2} \varepsilon \,.$$

If, on the other hand, $\hat{\pi}_\diamond = (\hat{T}_\diamond)_\sharp \rho$ is given by the construction of Theorem 5.7,

$$\mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi_\diamond^\star \,\|\, \pi) \le \mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi^\star \,\|\, \pi) \lesssim \kappa^{10/3} d^{1/3} \varepsilon^{4/3} \,.$$

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,

$$\partial_t^2 \mathcal{V}(\mu_t) = \mathbb{E}_{\pi^\star} \langle T - \mathrm{id}, (\nabla^2 V \circ T_t)\,(T - \mathrm{id}) \rangle \le L_V\, \mathbb{E}_{\pi^\star} \|T - \mathrm{id}\|^2 = L_V\, \mathbb{E}_\rho \|\hat{T}_\diamond - T^\star\|^2 \,.$$

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 5.4, $DT = D((T^\star)^{-1})\, D\hat{T}_\diamond \succeq 1/\sqrt{\kappa}$, so $DT_t \succeq 1/\sqrt{\kappa}$. Also, $DT^\star \succeq 1/\sqrt{L_V}$. Therefore, we obtain

$$\begin{aligned}
\partial_t^2 \mathcal{H}(\mu_t) &\le \kappa\, \mathbb{E}_{\pi^\star} \|D((T^\star)^{-1})\, D\hat{T}_\diamond \circ (T^\star)^{-1} - I\|_{\mathrm{F}}^2 \\
&\le \kappa L_V\, \mathbb{E}_{\pi^\star} \|D\hat{T}_\diamond \circ (T^\star)^{-1} - D((T^\star)^{-1})\|_{\mathrm{F}}^2 = \kappa L_V\, \mathbb{E}_\rho \|D\hat{T}_\diamond - DT^\star\|_{\mathrm{F}}^2 \,.
\end{aligned}$$

Therefore, adding the two terms together,

$$\partial_t^2\, \mathrm{KL}(\mu_t \,\|\, \pi) \le L_V\, \|\hat{T}_\diamond - T^\star\|_{L^2(\rho)}^2 + \kappa L_V\, \|D(\hat{T}_\diamond - T^\star)\|_{L^2(\rho)}^2 \,.$$

Integrating this expression from $t = 0$ to $t = 1$,

$$\mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi^\star \,\|\, \pi) \le \mathbb{E}_{\pi^\star}\langle [\nabla_{\mathbb{W}} \mathrm{KL}(\cdot \,\|\, \pi)](\pi^\star), T - \mathrm{id}\rangle + \frac{L_V}{2}\left(\|\hat{T}_\diamond - T^\star\|^2_{L^2(\rho)} + \kappa \,\|D\hat{T}_\diamond - DT^\star\|^2_{L^2(\rho)}\right).$$

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 5.6, 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 5.7, we obtain

$$\mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi^\star \,\|\, \pi) \lesssim \kappa\varepsilon^2 + \kappa^{10/3} d^{1/3}\varepsilon^{4/3}\,. \qquad \square$$

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 $\pi^\star_\diamond$ of the KL divergence over $\mathcal{P}_\diamond$, which in turn can be computed via the algorithms in Section 5.4.

Proof of Theorem 5.8. By triangle inequality, we have

$$W_2(\pi^\star_\diamond, \pi^\star) \le W_2(\pi^\star_\diamond, \hat{\pi}_\diamond) + W_2(\hat{\pi}_\diamond, \pi^\star)\,,$$

and since we can control the second term (recall Theorem 5.6), 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

$$\ell_V W_2^2(\pi^\star_\diamond, \hat{\pi}_\diamond)/2 \le \mathrm{KL}(\hat{\pi}_\diamond \,\|\, \pi) - \mathrm{KL}(\pi^\star_\diamond \,\|\, \pi) \lesssim \kappa^3 d^{1/2}\tilde{\varepsilon}\,,$$

where the final bound is obtained from Corollary C.3 (we only take the worst-case scaling term), and $\tilde{\varepsilon}$ is the approximation accuracy guaranteed by Theorem 5.6. Setting this equal to ε^2, we apply Theorem 5.6 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 C.3 and apply Theorem 5.7 with $\frac{\varepsilon^{3/2}}{\kappa^{5/2} d^{1/4}}$ replacing ε, obtaining $|\mathcal{M}| = \widetilde{O}(\kappa^{3/2} d^{5/4}/\varepsilon^{1/2})$. $\square$

C.4 Proofs for Section 5.4

Proof of Proposition 5.10. We write

$$\mathrm{KL}(\mu \,\|\, \pi) = \mathcal{V}(\mu) + \mathcal{H}(\mu) \coloneqq \int V \,\mathrm{d}\mu + \int \log \mu \,\mathrm{d}\mu + \log Z\,.$$

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

$$T^{\lambda,v} \coloneqq \alpha\,\mathrm{id} + \sum_{T\in\mathcal{M}} \lambda_T T + v\,, \qquad \mu_{\lambda,v} \coloneqq (T^{\lambda,v})_\sharp \rho\,.$$

Our goal is to upper bound the following quadratic forms

$$\nabla^2_{\mathbb{W}} \mathcal{V}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] = \mathbb{E}_{\mu_{\lambda,v}}[(T^{\eta,u}_{\lambda,v} - \mathrm{id})^\top \nabla^2 V \, (T^{\eta,u}_{\lambda,v} - \mathrm{id})] \,,$$
$$\nabla^2_{\mathbb{W}} \mathcal{H}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] = \mathbb{E}_{\mu_{\lambda,v}} \|DT^{\eta,u}_{\lambda,v} - I\|^2_{\mathrm{F}} \,,$$

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,

$$T^{\eta,u}_{\lambda,v} = T^{\eta,u} \circ (T^{\lambda,v})^{-1} = T^\eta \circ (T^\lambda)^{-1}(\cdot - v) + u \,, \tag{27}$$

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

$$\begin{aligned}\mathbb{E}_{\mu_{\lambda,v}} \|DT^{\eta,u}_{\lambda,v} - I\|^2_{\mathrm{F}} &= \mathbb{E}_{\mu_{\lambda,v}} \|[DT^\eta \circ (T^\lambda)^{-1}(\cdot - v)] \, D[(T^\lambda)^{-1}](\cdot - v) - I\|^2_{\mathrm{F}} \\ &= \mathbb{E}_\rho \|DT^\eta \, D[(T^\lambda)^{-1}] \circ (T^{\lambda,v} - v) - I\|^2_{\mathrm{F}} \\ &= \mathbb{E}_\rho \|DT^\eta \, D[(T^\lambda)^{-1}] \circ T^\lambda - I\|^2_{\mathrm{F}} \\ &= \mathbb{E}_\rho \|DT^\eta \, (DT^\lambda)^{-1} - I\|^2_{\mathrm{F}} \,,\end{aligned}$$

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

$$\mathbb{E}_{\mu_{\lambda,v}} \|DT^{\eta,u}_{\lambda,v} - I\|^2_{\mathrm{F}} \le \frac{1}{\alpha^2} \, \mathbb{E}_\rho \|DT^\eta - DT^\lambda\|^2_{\mathrm{F}} \,.$$

Since our maps are regular (i.e., (Υ) holds), there exists $\Upsilon > 0$ such that

$$\begin{aligned}\mathbb{E}_\rho \|DT^\eta - DT^\lambda\|^2_{\mathrm{F}} = \mathbb{E}_\rho \Big\| \sum_{T \in \mathcal{M}} (\lambda_T - \eta_T) \, DT \Big\|^2_{\mathrm{F}} &= \langle \eta - \lambda, Q^{(1)} \, (\eta - \lambda) \rangle \\ &\le \Upsilon \, \langle \eta - \lambda, Q \, (\eta - \lambda) \rangle \,.\end{aligned}$$

Finally, note that

$$\begin{aligned}W_2^2(\mu_{\lambda,v}, \mu_{\eta,u}) = \mathbb{E}_\rho \Big\| \sum_{T \in \mathcal{M}} (\eta_T - \lambda_T) \, T + u - v \Big\|^2 &= \mathbb{E}_\rho \Big\| \sum_{T \in \mathcal{M}} (\eta_T - \lambda_T) \, T \Big\|^2 + \|u - v\|^2 \\ &= \langle \eta - \lambda, Q \, (\eta - \lambda) \rangle + \|u - v\|^2 \,,\end{aligned}$$

where we used the fact that the maps in $\mathcal{M}$ are *centered*. This shows that

$$\nabla^2_{\mathbb{W}} \mathcal{H}(\mu_{\lambda,v})[T^{\eta,u}_{\lambda,v} - \mathrm{id}, T^{\eta,u}_{\lambda,v} - \mathrm{id}] \le \frac{\Upsilon}{\alpha^2} \, W_2^2(\mu_{\lambda,v}, \mu_{\eta,u}) \,.$$

Combining all of the terms completes the proof. □

C.5 Proofs for Section 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 C.4. *Let π be a (WC) measure, and let $\pi^\star$ be the mean-field approximation. Then*

$$\mathbb{E}_{\pi^\star} \nabla V = 0\,, \qquad \mathbb{E}_{\pi^\star} \|\nabla V\|^2 \le L_V \kappa d\,. \tag{28}$$

Proof. Assuming the first, we can prove the second by applying the Brascamp–Lieb inequality (Brascamp and Lieb, 1976):

$$\mathbb{E}_{\pi^\star} \|\nabla V - \mathbb{E}_{\pi^\star} \nabla V\|^2 \le \mathbb{E}_{\pi^\star} \operatorname{tr}((\nabla^2 V)^2 \operatorname{diag}(\vec{V}'')^{-1})\,,$$

where $\vec{V}'' \coloneqq (V_1'', \dots, V_d'')$. By Proposition 5.2, each component satisfies the bound $(V_i'')^{-1} \le 1/\ell_V$, and we also have by assumption $\nabla^2 V \preceq L_V I$. Together, the bound is clear:

$$\mathbb{E}_{\pi^\star} \|\nabla V\|^2 \le \operatorname{tr}((L_V I)^2)/\ell_V = L_V \kappa d\,.$$

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 5.12. 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:

$$\begin{aligned}\operatorname{tr}\operatorname{Cov}\left(Q_{1:J}^{-1/2}\, T_{1:J}(\hat{X}_i)\, \partial_i V(T^\lambda(\hat{X}))\right) &= \mathbb{E}\left[\langle T_{1:J}(\hat{X}_i)\, T_{1:J}(\hat{X}_i)^\top, Q_{1:J}^{-1}\rangle\, \partial_i V(T^\lambda(\hat{X}))^2\right] \\ &\le \Xi J\, \mathbb{E}[\partial_i V(T^\lambda(\hat{X}))^2],\end{aligned}$$

where we invoked (Ξ) in the last inequality. Summing over the coordinates,

$$\mathbb{E}[\|Q^{-1/2}\left(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda)\right)\|^2] \le \Xi J\, \mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2.$$

We bound the remaining expectation by repeatedly invoking smoothness of V. First,

$$\begin{aligned}\mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2 &\le 2\, \mathbb{E}_\rho \|\nabla V \circ T^\lambda - \nabla V \circ T_\diamond^\star\|^2 + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2 \\ &\le 2L_V^2\, \|T^\lambda - T_\diamond^\star\|_{L^2(\rho)}^2 + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2 \\ &= 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 2\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star\|^2.\end{aligned}$$

For the next term, we apply the same trick, but we compare against $\pi^\star$, the true mean-field approximation:

$$\begin{aligned}\mathbb{E}_\rho \|\nabla V \circ T^\lambda\|^2 &\le 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 4\, \mathbb{E}_\rho \|\nabla V \circ T_\diamond^\star - \nabla V \circ T^\star\|^2 + 4\, \mathbb{E}_\rho \|\nabla V \circ T^\star\|^2 \\ &\le 2L_V^2\, W_2^2(\mu_\lambda, \pi_\diamond^\star) + 4L_V^2\, W_2^2(\pi_\diamond^\star, \pi^\star) + 4L_V \kappa d,\end{aligned}$$

where we used Lemma C.4 in the last step.

Our full variance bound reads

$$\begin{aligned}\mathbb{E}[\|Q^{-1}\left(\hat{\nabla}_\lambda \mathcal{V}(\mu_\lambda) - \nabla_\lambda \mathcal{V}(\mu_\lambda)\right)\|_Q^2] \le\ & 2L_V^2 \Xi J\, W_2^2(\mu_\lambda, \pi_\diamond^\star) \\ & + 4L_V \Xi J \left(L_V\, W_2^2(\pi_\diamond^\star, \pi^\star) + \kappa d\right).\end{aligned}$$

□

D Remaining implementation details

D.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.

D.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

$$\mathrm{KL}(\mathcal{N}(0, A) \,\|\, \mathcal{N}(0, \Sigma)) = \frac{1}{2}\big[\mathrm{tr}(\Sigma^{-1}A) - d + \log\det(\Sigma^{-1}) - \log\det(A)\big] .$$

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

$$\mathrm{KL}(\mathcal{N}(0, A) \,\|\, \mathcal{N}(0, \Sigma)) = \frac{1}{2}\sum_{i=1}^{d}\big[(\Sigma^{-1})_{i,i}a_i - \log(a_i)\big] + C .$$

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.

D.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.

E Proofs for Section 7

In this section, we derive the gradient flows in Section 7.

Proof of Theorem 7.1. 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

$$\delta \mathcal{F}(P) : (\lambda, v) \mapsto \int (V + \log \mu_P + 1) \, \mathrm{d}\mu_{\lambda,v} = \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} + 1 \, . \tag{29}$$

Therefore, the Wasserstein gradient is given by

$$\nabla_{\mathbb{W}} \mathcal{F}(P)(\lambda, v) = \left(Q^{-1} \, \nabla_\lambda \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v}, \, \nabla_v \int \log \frac{\mu_P}{\pi} \, \mathrm{d}\mu_{\lambda,v} \right) . \tag{30}$$

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 7.2. This theorem follows from the expression of the first variation computed in (29), see Lambert et al. (2022, Appendix H). □