Efficient displacement convex optimization with particle gradient descent

Hadi Daneshmand [1 2 3] **Jason D. Lee** [4] **Chi Jin** [4]

Abstract

Particle gradient descent, which uses particles to represent a probability measure and performs gradient descent on particles in parallel, is widely used to optimize functions of probability measures. This paper considers particle gradient descent with a finite number of particles and establishes its theoretical guarantees to optimize functions that are *displacement convex* in measures. Concretely, for Lipschitz displacement convex functions defined on probability over $\mathbb{R}^d$, we prove that $O(1/\epsilon^2)$ particles and $O(d/\epsilon^4)$ iterations are sufficient to find the ϵ-optimal solutions. We further provide improved complexity bounds for optimizing smooth displacement convex functions. An application of our results proves the conjecture of *no optimization-barrier up to permutation invariance*, proposed by Entezari et al. (2022), for specific two-layer neural networks with two-dimensional inputs uniformly drawn from unit circle.

1. Introduction

Optimization in the space of probability measures has wide applications across various domains, including advanced generative models in machine learning (Arjovsky et al., 2017), the training of two-layer neural networks (Chizat & Bach, 2019), variational inference using Stein's method (Liu & Wang, 2016), super-resolution in signal processing (Bredies & Pikkarainen, 2013), and interacting particles in physics (McCann, 1997).

Optimization in probability spaces goes beyond the conventional optimization in Euclidean space. (Ambrosio et al., 2005) extends the notion of steepest descent in Euclidean space to the space of probability measures with the Wasserstein metric. This notion traces back to studies of the Fokker–Planck equation, a partial differential equation (PDE) describing the density evolution of Ito diffusion. The Fokker–Planck equation can be interpreted as a gradient flow in the space of probability distributions with the Wasserstein metric (Jordan et al., 1998). Gradient flows have become general tools to go beyond optimization in Euclidean space (Absil et al., 2009; Santambrogio, 2017; Chizat, 2022; Carrillo et al., 2020; Carrillo & Shu, 2022; Carrillo et al., 2021).

Gradient flows enjoy a fast global convergence on an important function class called displacement convex functions (Ambrosio et al., 2005) which is introduced to analyze equilibrium states of physical systems (McCann, 1997). Despite their fast convergence rate, gradient flows are hard to implement. Specifically, there are numerical solvers only for the limited class of linear functions with an entropy regularizer.

We study a different method to optimize functions of probability measures called particle gradient descent (Chizat & Bach, 2019; Chizat, 2022). This method restricts optimization to sparse measures with finite support (Nitanda & Suzuki, 2017; Chizat & Bach, 2019; Chizat, 2022; Li et al., 2022) as

$$\min_{w_1,\dots,w_n} F\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i}\right), \tag{1}$$

where δ_{w_i} is the Dirac measure at $w_i \in \Omega \subset \mathbb{R}^d$. Points $w_1, \dots, w_n$ are called particles. *Particle gradient descent* is the standard gradient descent optimizing the particles (Nitanda & Suzuki, 2017; Chizat & Bach, 2019; Chizat, 2022). This method is widely used to optimize neural networks (Nitanda & Suzuki, 2017; Chizat & Bach, 2019), take samples from a broad family of distributions (Li et al., 2022), and simulate gradient flows in physics (Carrillo & Shu, 2022). As will be discussed, F is not convex in particles due to its permutation-invariance to the particles. In that regard, the convergence of particle gradient descent is not guaranteed for general functions.

Gradient descent links to gradient flow as $n \to \infty$. In this asymptotic regime, (Chizat & Bach, 2019) proves that the

[1] Laboratory for Information and Decision Systems, MIT [2]Foundations of Data Science Institute (FODSI) [3]Hariri Institute for Computing and Computational Science and Engineering, Boston University [4] Department of Electrical and Computer Engineering at Princeton University. Correspondence to: Hadi Daneshmand <hdanesh@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

empirical distribution over the particles $w_1, \dots, w_n$ implements a (Wasserstein) gradient flow for F. Although the associated gradient flow globally optimizes displacement convex functions, the implication of such convergence has remained unknown for a finite number of particles.

1.1. Main contributions.

We prove that particle gradient descent efficiently optimizes displacement convex functions. Consider the sparse measure μ_n with support of size n. The error for μ_n can be decomposed as

$$F(\mu_n) - F^* := \underbrace{F(\mu_n) - \min_{\mu_n} F(\mu_n)}_{\text{optimization error}} + \underbrace{\min_{\mu_n} F(\mu_n) - F^*}_{\text{approximation error}}.$$

The optimization error in the above equation measures how much the function value of μ_n can be reduced by particle gradient descent. The approximation error is induced by the sparsity constraint. While the optimization of particles reduces the optimization error, the approximation error is independent of the optimization and depends on n.

Optimization error. For displacement convex functions, we establish the global convergence of variants of particle gradient descent. Table 1 presents the computational complexity of particle gradient descent optimizing smooth and Lipschitz displacement convex functions. To demonstrate the applications of these results, we provide examples of displacement convex functions that have emerged in machine learning, tensor decomposition, and physics.

Approximation error. Under a certain Lipschitz continuity condition, we prove the approximation error is bounded by $O(\frac{1}{\sqrt{n}})$ with a high probability. Furthermore, we prove this bound can be improved to $O(1/n)$ for convex and smooth functions in measures.

Finally, we demonstrate the application of the established results for a specific neural network with two-dimensional inputs, and zero-one activations. When the inputs are drawn uniformly from the unit circle, we prove that n-neurons achieve $O(1/n)$-function approximation in polynomial time for a specific function class.

2. Related works

There are alternatives to particle gradient descent for optimization in the space of measures. For example, conditional gradient descent optimizes smooth convex functions with a sub-linear convergence rate (Frank & Wolfe, 1956). This method constructs a sparse measure with support of size n using an iterative approach. This sparse measure is $O(\frac{1}{n})$-accurate in F (Dunn, 1979; Jaggi, 2013). However, each iteration of the conditional gradient method casts to a non-convex optimization without efficient solvers. Instead, the iterations of particle gradient descent are computationally efficient.

(Chizat & Bach, 2019) establishes the link between Wasserstein gradient flows and particle gradient descent. This study proves that particle gradient descent implements the gradient flows in the limit of infinite particles for a rich function class. The neurons in single-layer neural networks can be interpreted as the particles whose density simulates a gradient flow. The elegant connection between gradient descent and gradient flows has provided valuable insights into the optimization of neural networks (Chizat & Bach, 2019) and their statistical efficiency (Chizat & Bach, 2020). In practice, particle gradient descent is limited to a finite number of particles. Thus, it is essential to study particle gradient descent in a non-asymptotic regime. In this paper, we analyze optimization with a finite number of particles for displacement convex functions.

Displacement convexity has been used in recent studies of neural networks (Javanmard et al., 2019; Daneshmand & Bach, 2022). (Javanmard et al., 2019) establishes the global convergence of radial basis function networks using an approximate displacement convexity. (Daneshmand & Bach, 2022) proves the global convergence of gradient descent for a single-layer network with two-dimensional inputs and zero-one loss in realizable settings. Motivated by these examples, we analyze optimization for general (non-)smooth displacement convex functions.

Displacement convexity relates to the rich literature on geodesic convex optimization. Although the optimization of geodesic convex functions is extensively analyzed by (Zhang & Sra, 2016; Udriste, 2013; Absil et al., 2009) for Riemannian manifolds, less is known for the non-Riemannian manifold of probability measures with the Wasserstein-2 metric (Jordan et al., 1998).

In machine learning, various objective functions do not have any spurious local minima. This property was observed in early studies of neural networks. (Baldi & Hornik, 1989) show that the training objective of two-layer neural networks with linear activations does not have suboptimal local minima. This proof is extended to a family of matrix factorization problems, including matrix sensing, matrix completion, and robust PCA (Ge et al., 2017). Smooth displacement convex functions studied in this paper inherently do not admit spurious local minima (Javanmard et al., 2020).

For functions with no spurious minima, escaping the saddle points is crucial, which is extensively studied for smooth functions (Jin et al., 2017; Daneshmand et al., 2018). Although gradient descent may converge to suboptimal saddle points, random initialization effectively avoids the conver-

Function class	Regularity	Complexity
λ-displacement convex	ℓ-smooth	$nd\left(\frac{\ell-\lambda}{\ell+\lambda}\right)\log(\ell/\epsilon)$
star displacement convex	ℓ-smooth	$nd\ell\left(\frac{1}{\epsilon}\right)$
λ-displacement convex	L-Lipschitz	$ndL^2/(\lambda\epsilon)$
star displacement convex	L-Lipschitz	$nd\ell\left(\frac{1}{\epsilon}\right)$

Table 1. Computational complexity to reach an ϵ-optimization error. See Theorems 4.1 and 5.1 for formal statements.

gence of gradient descent to saddle points (Lee et al.). Yet, gradient descent may need a long time to escape saddles (Du et al., 2017). To speed up the escape, (Jin et al., 2017) leverages noise that allows escaping saddles in polynomial time. Building on these studies, we analyze the escaping of saddles for displacement convex functions.

Particle-based algorithms are also widely used in the context of Stein's variational inference (Liu & Wang, 2016; Yang et al., 2020; Li et al., 2022; Korba et al., 2021). We comment that Stein's variational gradient descent studied in that line of works is different from particle gradient descent considered in this paper in the following two important perspectives: (1) the update equations are different—their algorithms can be viewed as a discretization an infinite dimensional gradient descent on the relative entropy (Korba et al., 2020), and their update can not be written as a gradient descent of an objective function of discrete measures; (2) their results only apply to objective functions of specific forms that are either derived from Stein's identity or more generally admit a variational form (Yang et al., 2020). Our analysis does not rely on a variational form for the objective function. To elaborate on our problem settings and assumptions, we provide examples for which our assumptions hold.

3. Displacement convex functions

Note that the objective function F is invariant to the permutation of the particles. This permutation invariance concludes that F is not convex as the next Proposition states.

Proposition 3.1. *Suppose that $w_1^*, \dots, w_n^*$ is the unique minimizer of an arbitrary function $F(\frac{1}{n}\sum_{i=1}^n \delta_{w_i})$ such that $w_1^* \neq w_2^*$. If F is invariant to the permutation of $w_1, \dots, w_n$, then it is non-convex.*

The condition of having distinct optimal particles, required in the last Proposition, ensures the minimizer is not a trivial minimizer for which all the particles are equal. Since there is no global optimization method for non-convex functions, we study the optimization of the specific family of displacement convex functions.

3.1. Optimal transport

To introduce displacement convexity, we need to review the basics of optimal transport theory. Consider two probability measures μ and ν over $\mathbb{R}^d$. A transport map from μ to ν is a function $T : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\int_A \nu(x)dx = \int_{T^{-1}(A)} \mu(x)dx \tag{2}$$

holds for any Borel subset A of $\mathbb{R}^d$ (Santambrogio, 2017). The optimal transport T^* has the minimum transportation cost:

$$T^* = \arg\min_T \int \text{cost}(T(x), x)d\mu(x).$$

We use the standard squared Euclidean distance function for the transportation cost (Santambrogio, 2017). Remarkably, the transport map between distributions may not exist. For example, one can not transport a Dirac measure to a continuous measure.

In this paper, we frequently use the optimal transport map for two n-sparse measures in the following form

$$\mu = \frac{1}{n}\sum_{i=1}^n \delta_{w_i}, \quad \nu = \frac{1}{n}\sum_{i=1}^n \delta_{v_i}. \tag{3}$$

For the sparse measures, a permutation of $[1, \dots, n]$, denoted by σ, transports μ to ν. Consider the set Λ, containing all permutations of $[1, \dots, n]$ and define

$$\sigma^* = \arg\min_{\sigma\in\Lambda} \sum_{i=1}^n \|w_i - v_{\sigma(i)}\|^2. \tag{4}$$

The optimal permutation in the above equation yields the optimal transport map from μ to ν as $T^*(w_i) = v_{\sigma_i^*}$, and the Wasserstein-2 distance between μ and ν:

$$W_2^2(\mu, \nu) = \sum_{i=1}^n \|w_i - v_{\sigma^*(i)}\|_2^2. \tag{5}$$

Note that we omit the factor $1/n$ in W_2^2 for ease of notation.

3.2. Displacement convex functions

The displacement interpolation between μ and ν is defined by the optimal transport map as (McCann, 1997)

$$\mu_t = ((1-t)\text{Identity} + tT^*)_{\#}\,\mu, \tag{6}$$

where $G_{\#}\mu$ denotes the measure obtained by pushing μ with G. Note that the above interpolation is different from the convex combination of measures, i.e., $(1-t)\mu + t\nu$. For sparse measure, the displacement interpolation is $(1-t)w_i - tw_{\sigma*(i)}$ for the optimal permutation σ^* defined in Eq. (4).

λ-displacement convexity asserts Jensen's inequality along the displacement interpolation (McCann, 1997) as

$$F(\mu_t) \leq (1-t)F(\mu)+tF(\nu)-\frac{\lambda}{2}(1-t)(t)W_2^2(\mu,\nu).$$

A standard example of a displacement convex function is a convex quadratic function of measures.

Example 3.1. *Consider*

$$Q(\mu) = \int K(x-y)d\mu(x)d\mu(y)$$

where μ is a measure over $\mathbb{R}^d$ and $K(\Delta)$ is convex in $\Delta \in \mathbb{R}^d$; then, Q is 0-displacement convex (McCann, 1997).

The optimization of Q over a sparse measure is convex [1]. However, this is a very specific example of displacement convex functions. Generally, displacement convex functions are not necessarily convex.

Recall the sparse measures defined in Eq. (3). While convexity asserts Jensen's inequality for the interpolation of $\{w_i\}$ with all $n!$ permutations of $\{v_j\}$, displacement convexity only relies on a specific permutation. In that regard, displacement convexity is weaker than convexity. In the following example, we elaborate on this difference.

Example 3.2. *The energy distance between measures over $\mathbb{R}$ is defined as*

$$E(\mu,\nu) = 2\int |x-y|d\mu(x)d\nu(y) - \int |x-y|d\mu(x)d\mu(x) - \int |x-y|d\nu(x)d\nu(y). \quad (7)$$

$E(\mu,\nu)$ is 0-displacement convex in μ (Carrillo et al., 2020).

According to Proposition 3.1, E does not obey Jensen's inequality for interpolations with an arbitrary transport map. In contrast, E obeys Jensen's inequality for the optimal transport map, since it is monotone in $\mathbb{R}$ (Carrillo et al., 2020). This key property concludes E is displacement convex.

Remarkably, the optimization of the energy distance has applications in machine learning and physics. Daneshmand & Bach (2022) show that the training of two-layer neural networks with two-dimensional inputs (uniformly drawn from the unit sphere) casts to minimizing $E(\mu,\nu)$ in a sparse measure μ. The optimization of the energy distance has been also used in clustering (Székely & Rizzo, 2017). In physics, the gradient flow on the energy distance describes interacting particles from two different species (Carrillo et al., 2020).

[1] Q does not satisfy the condition of Proposition 3.1.

3.3. Star displacement convex functions

Our convergence analysis extends to a broader family of functions. Let $\widehat{\mu}$ denote the optimal n-sparse solution for the optimization in Eq. (1), and μ_t is obtained by the displacement interpolation between μ and $\widehat{\mu}$. Star displacement convex function F obeys

$$\sum_i \langle w_i - T(w_i), \partial_{w_i} F(\mu)\rangle \geq F(\mu) - F(\widehat{\mu}),$$

where T is the optimal transport map from μ to $\widehat{\mu}$. The above definition is inspired by the notion of star-convexity (Nesterov & Polyak, 2006). It is easy to check that 0-displacement convex functions are star displacement convex.

Star displacement convex optimization is used for generative models in machine learning. An important family of generative models optimizes the Wasserstein-2 metric (Arjovsky et al., 2017). Although Wasserstein 2 is not displacement convex (Santambrogio, 2017), it is star displacement convex.

Example 3.3. *$W_p(\mu,\nu)$ is star displacement convex in μ as long as μ and ν has sparse supports of the same size.*

Star displacement convexity holds for complete orthogonal tensor decomposition. Specifically, we consider the following example of tensor decomposition.

Example 3.4. *Consider the orthogonal complete tensor decomposition of order 3, namely*

$$\min_{w_1,\ldots,w_d\in\mathbb{R}^d} \left(G\left(\frac{1}{n}\sum_{i=1}^d \delta_{w_i}\right) = -\sum_{i=1}^d\sum_{j=1}^d \left\langle \frac{w_j}{\|w_j\|}, v_i\right\rangle^3 \right),$$

where $v_1,\ldots,v_d$ are orthogonal vectors over the unit sphere denoted by $\mathcal{S}_{d-1}$.

Although orthogonal tensor decomposition is not convex (Anandkumar et al., 2014), the next lemma proves that it is star displacement convex.

Lemma 3.2. *G is star displacement convex for $w_1,\ldots,w_n \in \mathcal{S}_{d-1}$.*

To prove the above lemma, we leverage the properties of the optimal transport map used for displacement interpolation.

There are more examples of displacement convex functions in machine learning (Javanmard et al., 2020) and physics (Carrillo & Slepčev, 2009). Motivated by these examples, we analyze displacement convex optimization.

4. Optimization of smooth functions

Gradient descent is a powerful method to optimize smooth functions (see Appendix A for the definition) that enjoy a dimension-free convergence rate to a critical point (Nesterov, 2003). More interestingly, a variant of gradient descent converges to local optimum (Daneshmand et al., 2018; Jin et al., 2017; Ge et al., 2015; Xu et al., 2018; Zhang et al., 2017). Here, we prove gradient descent globally optimizes the class of (star) displacement convex functions. Our results are established for the standard gradient descent, namely the following iterates

$$w_i^{(k+1)} = w_i^{(k)} - \gamma \partial_{w_i} F(\mu_k), \quad \mu_k := \frac{1}{n}\sum_{i=1}^{n} \delta_{w_i^{(k)}}, \quad (8)$$

where $\partial_{w_i} F$ denotes the gradient of F with respect to w_i. The next Theorem establishes the convergence of gradient descent.

Theorem 4.1. *Assume F is ℓ-smooth, and particle gradient descent starts from distinct particles $w_1^{(0)} \neq \ldots \neq w_n^{(0)}$. Let $\widehat{\mu}$ denote the optimal solution of (1).*

(a) For $(\lambda > 0)$-displacement functions,

$$F(\mu_{k+1}) - F(\widehat{\mu}) \leq \ell \left(1 - \left(\frac{2\lambda\ell\gamma}{\ell+\lambda}\right)\right)^k W_2^2(\mu_0, \widehat{\mu})$$

holds as long as $\gamma \leq 2/(\lambda + \ell)$.

(b) Under 0-displacement convexity,

$$F(\mu_{k+1}) - F(\widehat{\mu}) \leq \frac{2(F(\mu_0) - F(\widehat{\mu}))W_2^2(\mu_0, \widehat{\mu})}{2W_2^2(\mu_0, \widehat{\mu}) + (F(\mu_0) - F(\widehat{\mu}))\gamma k}$$

holds for $\gamma \leq 1/\ell$.

(c) Suppose that F is star displacement convex and $\max_{m\in\{1,\ldots,k\}} W_2^2(\mu_m, \widehat{\mu}) \leq r^2$; then

$$F(\mu_{k+1}) - F(\widehat{\mu}) \leq \frac{2(F(\mu_0) - F(\widehat{\mu}))r^2}{2r^2 + (F(\mu_0) - F(\widehat{\mu}))\gamma k}$$

holds for $\gamma \leq 1/\ell$.

Table 2 compares convergence rates for convex and displacement convex functions. We observe an analogy between the rates. The main difference is the replacement of Euclidean distance by the Wasserstein distance in the rates for displacement convex functions. This replacement is due to the permutation invariance of F. The Euclidean distance between $(w_1^*, \ldots, w_n^*)$ and permuted particles $(w_{\sigma(1)}^*, \ldots, w_{\sigma(n)}^*)$ can be arbitrary large, while F is invariant to the permutation of particles $w_1, \ldots, w_n$. Proven by the last Theorem 4.1, Wasserstein distance effectively replaces the Euclidean distance for permutation invariant displacement convex functions.

Function class	Convergence rate
λ-disp. convex	$\ell \left(\frac{\ell-\lambda}{\ell+\lambda}\right)^k W_2^2(\mu_0, \widehat{\mu})$
λ-strongly convex	$\frac{\ell}{2} \left(\frac{\ell-\lambda}{\ell+\lambda}\right)^k \sum_i \|w_i - w_i^*\|_2^2$
0-disp. convex	$2LW_2^2(\mu_0, \widehat{\mu})k^{-1}$
convex	$2L\left(\sum_i \|w_i - w_i^*\|_2^2\right)(k+4)^{-1}$

Table 2. Convergence rates for the optimization of ℓ-smooth functions. We use the optimal choice for the stepsize γ to achieve the best possible rate. Recall $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{w_i^*}$ denotes the optimal solution for Eq. (1). Rates for convex functions: (Nesterov, 2003). Rates for displacement convex functions: Theorem 4.1.

Smooth displacement convex functions are non-convex, hence have saddle points. However, displacement convex functions do not have suboptimal local minima (Javanmard et al., 2020). Such property has been frequently observed for various objective functions in machine learning. To optimize functions without suboptimal local minima, escaping saddle points is crucial since saddle points may avoid the convergence of gradient descent (Ge et al., 2015). (Lee et al.) proves that random initialization effectively avoids the convergence of gradient descent to saddle points. Similarly, the established global convergence results rely on a weak condition on initialization: The particles have to be distinct. A regular random initialization satisfies this condition.

Escaping saddles with random initialization may require considerable time for general functions. (Du et al., 2017) propose a smooth function on which escaping saddles may take an exponential time with the dimension. Notably, the result of the last theorem holds specifically for displacement convex functions. For this function class, random initialization not only enables escaping saddles but also leads to global convergence.

Remarkably, the convergence bound for star displacement convex functions requires $W_2^2(\mu_m, \widehat{\mu})$ to be bounded by a constant for all $m = 1, \ldots, k$. We postulate this technical assumption is not necessary and can be proven in future research.

5. Optimization of Lipschitz functions

The smoothness is a strong restriction. The training loss of neural networks with the standard ReLU activation is not smooth. In physics, energy functions often are not smooth (McCann, 1997; Carrillo & Shu, 2022). Furthermore, recent sampling methods are developed based on non-smooth optimization with particle gradient descent (Li et al., 2022). Thus, it is important to study the optimization of non-smooth. Here, we focus on non-smooth Lipschitz functions (see Appendix A for more details) which obey displacement convexity.

To optimize non-smooth functions, we add noise to gradient iterations as

$$w_i^{(k+1)} = w_i^{(k)} - \gamma_k \left(\partial_{w_i} F(\mu_k) + \frac{1}{\sqrt{n}} \xi_i^{(k)} \right) \quad \text{(PGD)}$$

where $\xi_1^{(k)}, \dots \xi_n^{(k)} \in \mathbb{R}^d$ are random vectors uniformly drawn from the unit ball. $\partial_{w_i} F$ denotes the sub-gradient of F with respect to w_i. In the appendix D.1, we prove the set of sub-gradients is not empty for displacement convex functions.

The above perturbed gradient descent (PGD) is widely used in smooth optimization to escape saddle points (Ge et al., 2015). The next Theorem proves this random perturbation can be leveraged for optimization of non-smooth functions, which are (star) displacement convex.

Theorem 5.1. *Consider the optimization of a L-Lipschitz function with PGD starting from $w_1^{(0)} \neq \dots \neq w_n^{(0)}$.*

a. If F is λ-displacement convex, then

$$\min_{k \in \{1,\dots,m\}} \{\mathbb{E}\left[F(\mu_k) - F(\widehat{\mu})\right]\} \leq \frac{2(L^2+1)}{\lambda(m+1)}$$

holds for $\gamma_k = 2/(\lambda(k+1))$.

b. If F is star displacement convex, then

$$\min_{k \in \{1,\dots,m\}} \{\mathbb{E}\left[F(\mu_k) - F(\widehat{\mu})\right]\} \leq \frac{1}{\sqrt{m}} \left(W_2^2(\mu_0, \widehat{\mu}) + L + 1 \right)$$

holds for $\gamma_1 = \dots = \gamma_m = 1/\sqrt{m}$.

Notably, the above expectations are taken over random vectors $\xi_1^{(k)}, \dots \xi_n^{(k)}$.

Thus, PGD yields an ϵ-optimization error with $O(1/\epsilon^2)$ iterations to reach ϵ-suboptimal solution for Lipschitz displacement convex functions. This rate holds for the optimization of the energy distance since it is 2-Lipschitz and 0-displacement convex (Carrillo et al., 2020). Daneshmand & Bach (2022) also establishes the convergence of gradient descent on the specific example of the energy distance. The last Theorem extends this convergence to the general function class of non-smooth Lipschitz displacement convex functions.

6. Approximation error

Now, we turn our focus to the approximation error. We provide bounds on the approximation error for two important function classes:

(i) Lipschitz functions in measures.

(ii) Convex and smooth functions in measures.

For (i), we provide the probabilistic bound $O\left(\frac{1}{\sqrt{n}}\right)$ on the approximation error; then, we improve the bound to $O(\frac{1}{n})$ for (ii).

6.1. Lipschitz functions in measures

We introduce a specific notion of Lipschitz continuity for functions of probability measures. This notion relies on Maximum Mean Discrepancy (MMD) between probability measures. Given a positive definite kernel K, MMD_K is defined as

$$\left(\text{MMD}_K(\mu,\nu)\right)^2 = \int K(w,v) d\mu(w) d\mu(v) - 2\int K(w,v) d\mu(w) d\nu(v) + \int K(w,v) d\nu(w) d\nu(v)$$

MMD is widely used for the two-sample test in machine learning (Gretton et al., 2012). Leveraging MMD, we define the following Lipschitz property.

Definition 6.1 (L-MMD_K Lipschitz)**.** F is L-MMD_K Lipschitz, if there exists a positive definite Kernel K such that

$$|F(\mu) - F(\nu)| \leq L \times \text{MMD}_K(\mu,\nu)$$

holds for all probability measures μ and ν.

Indeed, the above Lipschitz continuity is an extension of the standard Lipschitz continuity to functions of probability measures. A wide range of objective functions obeys the above Lipschitz continuity. Particularly, (Chizat & Bach, 2019) introduces a unified formulation for training two-layer neural networks, sparse deconvolution, and tensor decomposition as

$$R\left(\int \Phi(w) d\mu(w)\right) \quad (9)$$

where $\Phi : \mathbb{R}^d \to \mathcal{H}$ is a map whose range lies in the Hilbert space $\mathcal{H}$ and $R : \mathcal{H} \to \mathbb{R}_+$. Under a weak assumption, R is L-MMD$_K$ Lipschitz.

Proposition 6.2. *If R is L-Lipschitz in its input, then it is L-MMD$_K$-Lipschitz for $K(w, v) = \langle \Phi(w), \Phi(v) \rangle$.*

Thus, the class of Lipschitz functions is rich. For this function class, $O(\frac{1}{\sqrt{n}})$-approximation error is achievable.

Proposition 6.3. *Suppose that there exists a uniformly bounded kernel $\|K\|_\infty \leq B$ such that F is L-MMD$_K$ Lipschitz; then,*

$$\min_{\mu_n} F(\mu_n) - F^* \leq \frac{3\sqrt{B}}{\sqrt{n}}$$

holds with probability at least $1 - \exp(-1/n)$.

The last Proposition is a straightforward application of Theorem 7 in (Gretton et al., 2012). Combining the above result with Theorem 5.1 concludes the total complexity of $O(d/\epsilon^4)$ to find an ϵ-optimal solution for Lipschitz displacement functions. The complexity can be improved to $O(d/\epsilon^2)$ for smooth functions according to Theorem 4.1.

The established bound $O(1/\sqrt{n})$ can be improved under assumptions on the kernel K associated with the Lipschitz continuity. For d-differentiable shift-invariant kernels, (Xu et al., 2022) establishes a considerably tighter bound $O(\frac{\log(n)^d}{n})$ when the support of the optimal measure is a subset of the unit hypercube.

6.2. Convex functions in measures

If F is convex and smooth in μ, we can get a tighter bound on the approximation error.

Lemma 6.4. *Suppose F is convex and smooth in μ. If the probability measure μ is defined over a compact set, then*

$$\min_{\mu_n} F(\mu_n) - F^* = O\left(\frac{1}{n}\right)$$

holds for all n.

The proof of the last Lemma is based on the convergence rate of the Frank-Wolfe algorithm (Jaggi, 2013). This algorithm optimizes a smooth convex function by adding particles one by one. After n iterates, the algorithm obtains an n-sparse measure which is $O(1/n)$-suboptimal. Bach (2017) leverages this proof technique to bound the approximation error for neural networks. The last lemma extends this result to a broader function class.

Remarkably, the energy distance is convex and smooth in μ, hence enjoys $O(1/n)$-approximation error as stated in the next lemma.

Lemma 6.5. *$E(\mu, \nu)$ is convex and smooth in μ when μ and ν have a bounded support.*

6.3. Applications for neural networks

The training loss of neural networks is non-convex due to the invariance to the permutation of neurons. Based on extensive experimental observations, Entezari et al. (2022) postulate the optimization of neural networks has no barrier up to this permutation invariance. An application of our analysis proves this conjecture for a toy neural network with two-dimensional inputs.

Consider the class of functions in the following form

$$f(x) = \int \varphi(x^\top w) d\nu(w) \tag{10}$$

where $x, w \in \mathbb{R}^2$ lies on the unit circle and ν is a measure with support contained in the upper-half unit circle. φ is the standard zero-one ridge function:

$$\varphi(a) = \begin{cases} 1 & a > 0 \\ 0 & a \leq 0 \end{cases}. \tag{11}$$

The above function is used in the original MacCulloch-Pitts model for neural networks (McCulloch & Pitts, 1943). To approximate function f, one may use a neural network with a finite number of neurons implementing the following output function:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \varphi(x^\top w_i), \tag{12}$$

where $w_1, \ldots, w_n$ are points over the unit circle representing the parameters of the neurons. To optimize the location of $w_1, \ldots, w_n$, one may minimize the standard mean-squares loss as

$$\min_{w_1, \ldots, w_n} \left(L(w) := \mathop{\mathbb{E}}_x \left(f_n(x) - f(x) \right)^2 \right). \tag{13}$$

As is stated in the next corollary, PGD optimizes L up to the approximation error when the input x is distributed uniformly over the unit circle.

Corollary 6.6. *Suppose that the input x is drawn uniformly over the unit circle. After a specific transformation of the coordinates for $w_1, \ldots, w_n$, PGD with n particles with stepsize $\gamma_k = 1/\sqrt{k}$ obtains $w^{(k)} := [w_1^{(k)}, \ldots, w_n^{(k)}]$ after k iteration such that*

$$\min_{i \in \{1, \ldots, k\}} \mathbb{E}\left[L(w^{(i)}) \right] = O(\frac{n}{\sqrt{k}} + \frac{1}{n}) \tag{14}$$

holds where the expectation is taken over the algorithmic randomness of PGD.

The last corollary is the consequence of part b of Theorem 5.1, and the approximation error established in

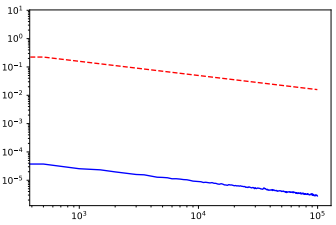


Figure 1. **The convergence of PGD for the energy distance.** Horizontal: $\log(k)$; vertical: $\log(E(\mu_k,\nu)-E(\widehat{\mu},\nu))$. The red dashed line is the theoretical convergence rate. The blue line is the convergence observed in practice for the average of 10 independent runs.

Lemma 6.4. For the proof, we use the connection between L and the energy distance derived by (Daneshmand & Bach, 2022). While (Daneshmand & Bach, 2022) focuses on realizable settings, the last corollary holds for non-realizable settings when the measure ν is not an n-sparse measure.

A line of research investigates hidden convexity structure in optimization of neural networks (Bartan & Pilanci, 2023; Mishkin et al., 2022; Ergen & Pilanci, 2021). In particular, these studies cast the training of two-layer neural networks to a convex optimization. While these interesting results rely on *overparameterization*, the result of the last lemma holds for the population loss, hence in an under-parameterized regime.

6.4. Applications for one-dimensional clustering

The energy distance is used for clustering (Székely & Rizzo, 2017). Clustering can be formulated as an optimization over sparse measure measures as (Peyré et al., 2019)

$$\min_{w_1,\dots,w_n} \text{dist}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i},\mu\right)$$

where dist denotes a proper distance metric for probability distributions. For example, k-means relies on Wasserstein distance (Peyré et al., 2019). An alternative to Wasserstein distance is the energy distance (Székely & Rizzo, 2017). The last theorem ensures $O(1/\sqrt{k})$ global-convergence-rate for PGD optimizing the energy distance for one-dimensional clustering. While k-means algorithms rely on expectation maximization, we leverage particle gradient descent to achieve this global convergence. Thus, we call for future research on clustering with particle gradient descent.

7. Experiments

We experimentally validate established bounds on the approximation and optimization error. Specifically, we validate the results for the example of the energy distance, which obeys the required conditions for our theoretical results[2].

7.1. Optimization of the energy distance

As noted in Example 3.2, the energy distance is displacement convex. Furthermore, it is easy to check that this function is 2-Lipschitz. For the sparse measures in Eq. (3), the energy distance has the following form

$$n^2E(\mu,\nu) = 2\sum_{i,j=1}^{n}|w_i-v_j| - \sum_{i,j=1}^{n}|v_i-v_j| - \sum_{i,j=1}^{n}|w_i-w_j|,$$

where $n=100$ for this experiment. We draw $v_1,\dots,v_n$ at random from uniform$[0,1]$. Since E is not a smooth function, we use PGD to optimize $w_1,\dots,w_n\in\mathbb{R}$. In particular, we use $\xi_1^{(k)}$ i.i.d. from uniform$[-0.05,0.05]$. For the stepsize, we use $\gamma_k=1/\sqrt{k}$ required for the convergence result in Theorem 5.1 (part b). In Figure 1, we observe a match between the theoretical and experimental convergence rate for PGD.

7.2. Approximation error for the energy distance

Lemma 6.4 establish $O(1/n)$ approximation error for convex functions of measures. Although the energy distance $E(\mu,\nu)$ is not convex in the support of μ, it is convex and smooth in μ as stated in Lemma 6.5. Thus, $O(1/n)$-approximation error holds for the energy distance. We experimentally validate this result. Consider the recover of $\nu=$uniform$[-1,1]$ by minimizing the energy distance as

$$E(\mu,\nu) = \frac{2}{n}\sum_{i=1}^{n}|w_i-v|d\nu(v) - \frac{1}{n^2}\sum_{i,j=1}^{n}|w_i-w_j| - \int|v-v'|d\nu(v)d\nu(v').$$

The above integrals can be computed in closed forms using $\int_{-1}^{1}|w-v|d\nu(v)=w^2+1$. Hence, we can compute the derivative of E with respect to w_i. We run PGD with stepsize determined in the part b of Theorem 5.1 for $k=3\times 10^5$ iterations and various $n\in\{2^2,\dots,2^8\}$. Figure 2 shows how the error decreases with n in the log-log scale. In this plot, we observe that E enjoys a mildly better approximation error compared to the established bound $O(1/n)$.

7.3. Clustering in $\mathbb{R}$

In Section 6.4, we discussed clustering using particle gradient descent. We illustrate the outputs of such clustering for

[2]The implementation is available on the GitHub repository https://github.com/hadidaneshmand/icml23_pgd

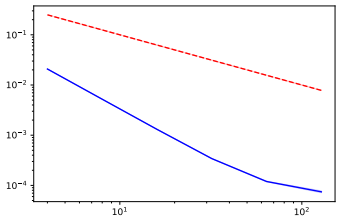
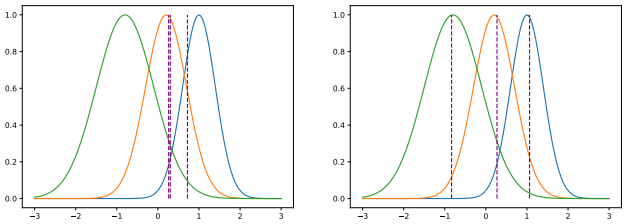


Figure 2. **Approximation error for the energy distance.** Horizontal: n; vertical: $E(\widehat{\mu}_n, \nu)$ where $\widehat{\mu}_n$ is obtained by 3×10^5 iterations of PGD with n particles. The red dashed line is the theoretical $O(1/n)$-bound for the approximation error. The plot is in the log-scale for both axes. The (blue) plot shows the average of 10 independent runs.

a mixture of Gaussian over $\mathbb{R}$. Specifically, we run particle gradient descent with three particles to cluster a mixture of three Gaussian distributions. Figure 3 shows particles before and after optimization, demonstrating the convergence of particles to the means of mixtures.



Initialization After optimization

Figure 3. **Clustering in $\mathbb{R}$ with PGD.** Solid curves show the mixture distribution. Vertical dashed lines marks the locations of $\{w_1, w_2, w_3\}$ at initialization (left) and after 20000 iterations of PGD (right). In this experiment, we use 100 samples from each distribution.

8. Discussions

We establish a non-asymptotic convergence rate for particle gradient descent when optimizing displacement convex functions of measures. Leveraging this convergence rate, we prove the optimization of displacement convex functions of (infinite-dimensional) measures can be solved in polynomial time with input dimension, and the desired accuracy rate. This finding will be of interest to various communities, including the communities of non-convex optimization, optimal transport theory, particle-based sampling, and theoretical physics.

The established convergence rates are limited to particle gradient descent. Yet, there may be other algorithms that converge faster than this algorithm. There is ample research on lower-bound complexities required to optimize convex function (Nesterov, 2003). Given that displacement convex functions do not obey the conventional notion of convexity, it is not clear whether these lower bounds extend to this specific class of non-convex functions. More research is needed to establish (Oracle-based) lower-computational-complexities for displacement convex optimization.

Nesterov's accelerated gradient descent enjoys a considerably faster convergence compared to gradient descent in convex optimization. Indeed, this method attains the optimal convergence rate using only first-order derivatives of smooth convex functions (Nesterov, 2003). This motivates future research to analyze the convergence of accelerated gradient descent on displacement convex functions.

We provided examples of displacement convex functions, including the energy distance. Displacement convex functions are not limited to these examples. A progression of this work is to assess the displacement convexity of various non-convex functions. In particular, non-convex functions invariant to permutation of the coordinates, including latent variable models and matrix factorization (Anandkumar et al., 2014), may obey displacement convexity under weak assumptions.

A major limitation of our result is excluding displacement convex functions with entropy regularizers that have emerged frequently in physics (McCann, 1997). The entropy is displacement convex. It is challenging to estimate entropy using sparse measures. Thus, particle gradient descent is not practical for the optimization of functions with an entropy regularizer. To optimize such functions, the existing literature uses a system of interacting particles solving a stochastic differential equation (Philipowski, 2007). In asymptotic regimes, this algorithm implements a gradient flow converging to the global optimal measure (Philipowski, 2007). To assess the complexity of these particle-based algorithms, we need non-asymptotic analyses for a finite number of particles.

Acknowledgments and Disclosure of Funding

We thank Ashkan Soleymani, Mert Pilanci, Zhuoran Yang, Xiang Cheng, Francis Bach, Lenaic Chizat, and Philippe Rigollet for their helpful discussions on the related literature on particle-based sampling, the energy distance minimization and Riemannian optimization. We appreciate very helpful comments of *anonymous reviewers of ICML23*. This project was mainly funded by the Swiss National Science Foundation (grant P2BSP3_195698). We also acknowledge support from the NSF TRIPODS program (award DMS-2022448).

References

Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 2014.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 2017.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Bartan, B. and Pilanci, M. Convex optimization of deep polynomial and relu activation neural networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Bredies, K. and Pikkarainen, H. K. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 2013.

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

Carrillo, J., Mateu, J., Mora, M., Rondi, L., Scardia, L., and Verdera, J. The equilibrium measure for an anisotropic nonlocal energy. *Calculus of Variations and Partial Differential Equations*, 60(3):1–28, 2021.

Carrillo, J. A. and Shu, R. Global minimizers of a large class of anisotropic attractive-repulsive interaction energies in 2d. *arXiv preprint arXiv:2202.09237*, 2022.

Carrillo, J. A. and Slepčev, D. Example of a displacement convex functional of first order. *Calculus of Variations and Partial Differential Equations*, 2009.

Carrillo, J. A., Francesco, M. D., Esposito, A., Fagioli, S., and Schmidtchen, M. Measure solutions to a system of continuity equations driven by newtonian nonlocal interactions. *Discrete and Continuous Dynamical Systems*, 40 (2):1191–1231, 2020.

Chizat, L. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 2022.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Proceedings of Conference on Neural Information Processing Systems*, 2019.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Daneshmand, H. and Bach, F. Polynomial-time sparse measure recovery: From mean field theory to algorithm design, 2022.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, 2018.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 2017.

Dunn, J. C. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 1979.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. *ICLR*, 2022.

Ergen, T. and Pilanci, M. Convex geometry and duality of over-parameterized neural networks. *The Journal of Machine Learning Research*, 22(1):9646–9708, 2021.

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

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, 2015.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242, 2017.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435, 2013.

Javanmard, A., Mondelli, M., and Montanari, A. Analysis of a two-layer neural network via displacement convexity. 2019.

Javanmard, A., Mondelli, M., and Montanari, A. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732. PMLR, 2017.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 2020.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on Learning Theory*.

Li, L., Liu, Q., Korba, A., Yurochkin, M., and Solomon, J. Sampling with mollified interaction energy descent. *arXiv preprint arXiv:2210.13400*, 2022.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 2016.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 1997.

McCulloch, W. S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 1943.

Mishkin, A., Sahiner, A., and Pilanci, M. Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions. In *International Conference on Machine Learning*, pp. 15770–15816. PMLR, 2022.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003.

Nesterov, Y. and Polyak, B. T. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 2006.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019.

Philipowski, R. Interacting diffusions approximating the porous medium equation and propagation of chaos. *Stochastic processes and their applications*, 2007.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 2017.

Székely, G. J. and Rizzo, M. L. The energy of data. *Annual Review of Statistics and Its Application*, 2017.

Udriste, C. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

Xu, L., Korba, A., and Slepcev, D. Accurate quantization of measures via interacting particle-based optimization. In *International Conference on Machine Learning*, pp. 24576–24595. PMLR, 2022.

Xu, Y., Jin, R., and Yang, T. First-order stochastic algorithms for escaping from saddle points in almost linear time. *Advances in neural information processing systems*, 31, 2018.

Yang, Z., Zhang, Y., Chen, Y., and Wang, Z. Variational transport: A convergent particle-based algorithm for distributional optimization. *arXiv preprint arXiv:2012.11554*, 2020.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638. PMLR, 2016.

Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pp. 1980–2022, 2017.

A. Function classes

Our focus is on four main function classes:

- F is ℓ-smooth when it is continuously differentiable w.r.t $w_1, \ldots, w_n$ and its gradient is ℓ-Lipschitz.
- F is L-Lipschitz if its subgradients are bounded by L.
- F is smooth in measure μ when its functional differentiable and its differential in μ is Lipschitz.
- The Lipschitzness in μ is defined in Definition 6.1.

B. Displacement convexity

B.1. Proof of Proposition 3.1

Suppose that $w_1^*, \ldots, w_n^*$ is the unique minimizer of F such that $w_1^* \neq w_2^*$. Let σ is a non monotone permutation of $\{1, \ldots, n\}$. Interpolating the parameters $\{w_i^*\}$ with the permuted $\{w_{\sigma(i)}^*\}$ concludes F is not convex since

$$F\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{(1-t)w_i^*+tw_{\sigma(i)}^*}\right) > tF\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i^*}\right) + (1-t)F\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_{\sigma(i)^*}}\right)$$

holds for the unique minimizer $w_1^*, \ldots, w_n^*$. Thus, F is not convex.

B.2. Proof of Proposition 6.2

The range of Φ lies in a Hilbert space with a norm induced by an inner product. Thus,

$$\begin{aligned}\left\|\int \Phi d\mu - \int \Phi d\nu\right\|^2 &= \left\langle \int \Phi d\mu - \int \Phi d\nu, \int \Phi d\mu - \int \Phi d\nu \right\rangle \\ &= \left\langle \int \Phi d\mu, \int \Phi d\mu \right\rangle - 2\left\langle \int \Phi d\mu, \int \Phi d\nu \right\rangle + \left\langle \int \Phi d\nu, \int \Phi d\nu \right\rangle \\ &= \left(\mathrm{MMD}_K(\mu,\nu)\right)^2.\end{aligned}$$

The above result together with the Lipschitz property of R concludes the proof.

B.3. Example 3.3

$$\text{Recall:} \qquad \mu_t = \frac{1}{n}\sum_{i=1}^{n}\delta_{(1-t)w_i+tv_{\sigma^*(i)}}.$$

According to the definition, we get

$$W_p(\mu_t,\nu) = (\frac{1}{n}\sum_i \|(1-t)w_i + tT^*(w_i) - T^*(w_i)\|^p)^{1/p} = (\frac{1}{n}\sum_i (1-t)^p\|w_i - T^*(w_i)\|^p)^{1/p} = (1-t)W_p(\mu,\nu) \tag{15}$$

B.4. Example 3.4

$$\text{Define:} \qquad \sigma = \arg\min_{\sigma'}\sum_i \left\|w_i - v_{\sigma'(i)}\right\|^2.$$

$$\text{Define:} \qquad g(t) = G\left(\underbrace{(1-t)w_1 + tv_{\sigma(1)}}_{w_1(t)}, \ldots, (1-t)w_n + tv_{\sigma(n)}\right).$$

To prove G is star displacement convex, we need to prove that $-g'(0) \geq (G - G^*)$ holds. To validate this inequality, we take the derivative of g with respect to t:

$$\frac{1}{3}g'(t) = -\sum_j \sum_i \left(\frac{\langle v_i, w_j(t)\rangle^2 \langle v_i, (v_{\sigma(j)} - w_j(t))\rangle}{\|w_j(t)\|^3} + \frac{\langle v_i, w_j(t)\rangle^3 \langle w_j(t), w_j - v_{\sigma(j)}\rangle}{\|w_j(t)\|^3} \right).$$

For $w_1, \ldots, w_n \in \mathcal{S}_{d-1}$, we get

$$\frac{1}{3}g'(0) = \sum_j \langle w_j, v_{\sigma(j)}\rangle \left(\sum_i \langle v_i, w_j\rangle^3 - \langle w_j, v_{\sigma(j)}\rangle \right), \tag{16}$$

where we used the orthogonality of $v_1, \ldots, v_n$. Since σ is an optimal permutation, we have

$$n \sum_j \langle w_j, v_{\sigma(j)}\rangle^2 \geq \sum_i \sum_j \langle w_i, v_j\rangle^2 = n, \tag{17}$$

where we use the fact that $v_1, \ldots, v_n$ are orthogonal. Thus,

$$-\frac{1}{3}g'(0) \geq 1 - \sum_{i,j=1}^{n} \langle v_i, w_j\rangle^3 = G - G^*. \tag{18}$$

B.5. A characterization for λ-convex functions

Let T denote the optimal transport map from μ to ν which achieves

$$W_2^2(\mu, \nu) = \sum_{i=1}^{n} \|w_i - T(w_i)\|_2^2 \tag{19}$$

Let $g(t) = F(\mu_t)$ where μ_t is obtained by the displacement interpolation of μ and ν. Taking the derivative of g with respect to t leads to the following important inequality (Santambrogio, 2017):

$$\underbrace{\sum_i \langle w_i - T(w_i), \partial_{w_i} F(\mu)\rangle}_{g'(0)} \geq F(\mu) - F(\nu) + \frac{\lambda}{2} W_2^2(\mu, \nu) \tag{20}$$

B.6. Smooth and displacement convex functions

The next Theorem establishes properties of smooth and displacement convex functions, which will be repeatedly used in our future analysis.

Theorem B.1. *An ℓ-smooth F and $(\lambda \geq 0)$-displacement convex function obeys*

$$\sum_i \|\partial_{w_i} F(\mu) - \partial_{T(w_i)} F(\nu)\| \leq \ell^2 W_2^2(\mu, \nu) \tag{i}$$

$$\|\partial_{w_i} F(\mu) - \partial_{w_j} F(\mu)\| \leq \ell \|w_i - w_j\| \tag{ii}$$

$$F(\nu) \geq F(\mu) + \sum_i \langle \partial_{w_i} F(\mu), T(w_i) - w_i\rangle + \frac{1}{2\ell} \sum_i \|\partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\mu)\|^2 \tag{iii}$$

$$\sum_i \langle \partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\nu), T(w_i) - w_i\rangle \geq \frac{1}{\ell} \sum_i \|\partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\mu)\|^2 \tag{iv}$$

$$\sum_i \langle \partial_{w_i} F(\mu) - \partial_{T(w_i)} F(\nu), w_i - T(w_i)\rangle \geq \frac{\lambda\ell}{\ell + \lambda} \sum_i \|w_i - T(w_i)\|^2 + \frac{1}{\lambda + \ell} \|\partial_{w_i} F(\mu) - \partial_{w_i} F(\nu)\|^2 \tag{v}$$

Proof. **(i)** Since the smoothness holds for all permutation of particles, we get

$$\sum_{i=1}^{n} \|\partial_{w_i} F(\mu) - \partial_{T(w_i)} F(\nu)\|^2 = \|\partial F(w) - \partial F(v)\|^2 \tag{21}$$

$$\leq \ell^2 \|w - v\|^2 = \ell^2 W_2^2(\mu, \nu), \tag{22}$$

where use a permutation of particles to get the the last equality.
(ii) Suppose that v is obtained by swapping w_i and w_j for $i \neq j$ in $w = (w_1, \ldots, w_n)$. Then, the smoothness implies

$$\|\partial_{w_i} F(w) - \partial_{w_i} F(v)\| \leq \ell \|w_i - w_j\| \tag{23}$$

(iii) Akin to the proof of Theorem 2.1.5 in (Nesterov, 2003), we define

$$Q(\nu) = F(\nu) - \sum_i \langle T(w_i), \partial_{w_i} F(\mu) \rangle$$

Displacement convexity, more precisely Eq. (20), ensures that μ is the minimizer of the above functional since

$$F(\nu) - F(\mu) - \sum_i \langle T(w_i) - w_i, \partial_{w_i} F(\mu) \rangle \geq 0 \tag{24}$$

ℓ-smoothness ensures (Nesterov, 2003)

$$Q(\mu) \leq Q\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{v_i - \frac{1}{\ell} \partial_{v_i} Q(\nu)} \right) \tag{25}$$

$$\leq Q(\nu) - \frac{1}{2\ell} \underbrace{\sum_{i=1}^{n} \|\partial_{v_i} Q(\nu)\|^2}_{= \sum_i \|\partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\mu)\|^2} \tag{26}$$

(iv) Inequality (iii) ensures

$$F(\nu) \geq F(\mu) + \sum_i \langle \partial_{w_i} F(\mu), T(w_i) - w_i \rangle + \frac{1}{2\ell} \sum_i \|\partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\mu)\|^2 \tag{27}$$

$$F(\mu) \geq F(\nu) + \sum_i \langle \partial_{T(w_i)} F(\nu), w_i - T(w_i) \rangle + \frac{1}{2\ell} \sum_i \|\partial_{T(w_i)} F(\nu) - \partial_{w_i} F(\mu)\|^2 \tag{28}$$

Summing up the above two inequalities concludes (iv).

(v) The proof is similar to the proof of Theorem 2.1.11 in (Nesterov, 2003). Let T denote the optimal transport map from μ to ν and μ_t is the displacement interpolation between μ and ν. First, we define function $\phi(\mu) = F(\mu) - \frac{\lambda}{2} \sum_{i=1}^{n} \|w_i\|^2$. We prove ϕ is 0-displacement convex.

$$\phi(\mu_t) \leq (1-t) F(\mu) + t F(\nu) - \left(\frac{(1-t) t \lambda}{2} \right) W_2^2(\mu, \nu) \tag{29}$$

$$\phi(\mu) = F(\mu) - \frac{\lambda}{2} \sum_{i=1}^{n} \|w_i\|^2 \tag{30}$$

$$\phi(\nu) = F(\nu) - \frac{\lambda}{2} \sum_{i=1}^{n} \|T(w_i)\|^2 \tag{31}$$

Putting the above three equations together yields

$$(1-t)\phi(\mu)+t\phi(\nu)-\phi(\mu_t) \geq \left(\frac{\lambda t(1-t)}{2}\right)\sum_i \|w_i - T(w_i)\|^2 - \frac{\lambda(1-t)}{2}\sum_i \|w_i\|^2 - \frac{\lambda}{2}\sum_i \|T(w_i)\|^2 + \frac{\lambda}{2}\sum_i \|(1-t)w_i + tT(w_i)\|^2 \quad (32)$$

Expanding the last term concludes ϕ is 0-displacement convex. Thus, we can use (iv) to get

$$\sum_i \langle \partial_{w_i}\phi(\mu) - \partial_{T(w_i)}\phi(\nu), w_i - T(w_i)\rangle \geq \frac{1}{\ell-\lambda}\sum_i \|\partial_{w_i}\phi(\mu) - \partial_{T(w_i)}\phi(\nu)\|^2 \quad (33)$$

A rearrangement of the terms concludes (v). □

C. Smooth displacement convex optimization

C.1. Proof of Theorem 4.1.a

Contraction

We first prove that the particles contract in W_2.

Lemma C.1 (Contraction)**.** *Suppose that μ_k is obtained by particle gradient descent starting from n-sparse measures with distinct particles. If F is λ-displacement convex and ℓ-smooth, then*

$$W_2^2(\mu_{(k+1)}, \widehat{\mu}) \leq \left(1 - \left(\frac{2\lambda\ell\gamma}{\ell+\lambda}\right)\right) W_2^2(\mu_k, \widehat{\mu}) \quad (34)$$

holds for $\gamma \leq 2/(\ell+L)$.

Proof. In order to leverage displacement convexity, we need to ensure the optimal transport map exists from $\mu_{(k)}$ to $\widehat{\mu}$ for all k. The next lemma proves that μ_k has distinct particles assuming μ_0 has distinct particles, hence the optimal transport map exists under weak conditions on μ_0.

Lemma C.2. *For ℓ-smooth F and $\gamma < 1/\ell$, $w_i^{(k+1)} = w_j^{(k+1)}$ hold only if $w_i^{(k)} = w_j^{(k)}$.*

Therefore, the optimal transport from $\mu := \mu_k$ to $\widehat{\mu}$ is well defined and denoted by T. We leverage this transports to couple $\mu_+ := \mu_{k+1}$ with $\widehat{\mu}$. The optimality of the transport implies that

$$\sum_i \|w_i - T(w_i)\|^2 d\mu(w) = W_2^2(\mu, \widehat{\mu}) \quad (35)$$

Using the recurrence of gradient descent, we get

$$W_2^2(\mu_+, \widehat{\mu}) \leq \sum_i \left(\|w_i - T(w_i)\|^2 + 2\gamma\langle T(w_i) - w_i, \partial_{w_i}F(\mu)\rangle\right) + \gamma^2 \sum_i \|\partial_{w_i}F(\mu) - \underbrace{\partial_{T(w_i)}F(\widehat{\mu})}_{=0}\|^2. \quad (36)$$

According to Theorem B.1.v. we have

$$\sum_i \left(\langle w_i - T(w_i), \partial_{w_i}F(\mu)\rangle\right) \geq \frac{\lambda\ell}{\ell+\lambda}W_2^2(\mu, \widehat{\mu}) + \frac{1}{\ell+\lambda}\sum_i \|\partial_{w_i}F(\mu) - \partial_{T(w_i)}F(\widehat{\mu})\|^2$$

Combining the last two inequalities, we get

$$W_2^2(\mu_+, \widehat{\mu}) \leq W_2^2(\mu, \widehat{\mu}) - \left(\frac{2\lambda\ell\gamma}{\ell+\lambda}\right) W_2^2(\mu, \widehat{\mu}) + \left(\gamma^2 - \frac{2\gamma}{\ell+\lambda}\right)\sum_i \|\partial_{w_i}F(\mu) - \underbrace{\partial_{T(w_i)}F(\widehat{\mu})}_{=0}\|^2$$

For $\gamma \leq 2/(\ell + \lambda)$, we get

$$W_2^2(\mu_+, \widehat{\mu}) \leq \left(1 - \left(\frac{2\lambda\ell\gamma}{\ell + \lambda}\right)\right) W_2^2(\mu, \widehat{\mu}). \tag{37}$$

□

Convergence proof

A straight forward applications of Cauchy-Schwarz for Eq.20 yields

$$W_2(\mu, \widehat{\mu}) \sqrt{\sum_i \|\partial_{w_i} F(\mu)\|^2} \geq F(\mu) - F(\widehat{\mu}) \tag{38}$$

holds for all μ with n particles. Invoking Theorem B.1.i, we get

$$\sum_i \|\partial_{w_i} F(\mu)\|^2 \leq \ell^2 W_2^2(\mu, \widehat{\mu}) \tag{39}$$

Combining the last two inequalities, we get

$$F(\mu) - F(\widehat{\mu}) \leq \ell W_2^2(\mu, \widehat{\mu}) \tag{40}$$

Combining the above inequality with the contraction established in the last lemma, we get

$$|F(\mu_{k+1}) - F(\widehat{\mu})| \leq \ell \left(1 - \frac{2\lambda\ell\gamma}{\ell + \lambda}\right)^k W_2^2(\mu_0, \widehat{\mu}) \tag{41}$$

C.2. Proof of Theorem 4.1.c.

It is known that gradient descent decreases smooth functions in each iteration as long as $\gamma \leq 1/\ell$ (see Theorem 2.1.13 of (Nesterov, 2003)):

$$F(\mu_{k+1}) \leq F(\mu_k) - \gamma\,(1 - \ell\gamma/2) \sum_i \|\partial_{w_i} F(\mu_k)\|^2 \tag{42}$$

star displacement convexity ensures

$$\sum_i \langle w_i - T(w_i), \partial_{w_i} F(\mu)\rangle \geq F(\mu_k) - F(\widehat{\mu}). \tag{43}$$

A straightforward application of Cauchy-Schwarz yields

$$\sum_i \langle w_i - T(w_i), \partial_{w_i} F(\mu)\rangle \leq \sum_i \|w_i - T(w_i)\| \|\partial_{w_i} F(\mu)\| \tag{44}$$

$$\leq W_2(\mu_k, \widehat{\mu}) \sqrt{\sum_i \|\partial_{w_i} F(\mu)\|^2} \tag{45}$$

where the last inequality holds since $W_2^2(\mu_k, \widehat{\mu})$ is decreasing for $\gamma < 1/\ell$. Combining the last two inequalities, we get

$$\sum_i \|\partial_{w_i} F(\mu)\|^2 \geq \frac{(F(\mu_k) - F(\widehat{\mu}))^2}{W_2^2(\mu_k, \widehat{\mu})} \tag{46}$$

We introduce the compact notion $\Delta_k = F(\mu_k) - F(\widehat{\mu})$. Plugging the last inequality into Eq. 42 obtains

$$\Delta_{k+1} \leq \Delta_k - \underbrace{\left(\frac{\gamma}{W_2^2(\mu_k, \widehat{\mu})}\right)}_{\geq \gamma/r^2} (1 - \ell\gamma/2)\Delta_k^2 \tag{47}$$

According to Theorem 2.1.13 of (Nesterov, 2003), the above inequality concludes the proof.

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k(1 - c\Delta_k)} \tag{48}$$

C.3. Proof of Theorem 4.1.b

Suppose that T is the optimal transport map from μ_k to $\widehat{\mu}$. Using ℓ-smoothness, we have

$$W_2^2(\mu_{k+1}, \widehat{\mu}) \leq \sum_i \|T(w_i) - w_i\|^2 + 2\gamma\langle T(w_i) - w_i, \partial_{w_i} F(\mu_k)\rangle + \gamma^2 \|\partial_{w_i} F(\mu_k)\|^2 \tag{49}$$

Replacing $\nu = \widehat{\mu}$ and $\mu = \mu_k$ in Theorem B.1.iii yields

$$\sum_i \langle w_i - T(w_i), \partial_{w_i} F(\mu)\rangle \geq \frac{1}{2\ell} \sum_i \|\partial_{w_i} F(\mu)\|^2 \tag{50}$$

Incorporating the above inequality into the established bound for W_2 leads to

$$W_2^2(\mu_{k+1}, \widehat{\mu}) \leq W_2^2(\mu_k, \widehat{\mu}) + (\gamma^2 - \gamma/\ell) \sum_i \|\partial_{w_i} F(\mu_k)\|^2. \tag{51}$$

Thus, $W_2(\mu_k, \widehat{\mu})$ is non-increasing in k for $\gamma \leq 1/\ell$. Since a 0-displacement convex function is also weak displacement convex, invoking part c. of Theorem 4.1 with $r^2 = W_2^2(\mu_0, \widehat{\mu})$ concludes the proof.

C.4. Proof of Lemma C.2

The proof is similar to the analysis of smooth programs in (Lee et al.). According to the definition, $w_i^{(k+1)} = w_j^{(k+1)}$ if

$$w_i^{(k)} - \gamma \partial_{w_i} F(\mu_k) = w_j^{(k)} - \gamma \partial_{w_i} F(\mu_k) \tag{52}$$

A rearrangement of terms together with Theorem B.1.II concludes the proof:

$$\|w_i^{(k)} - w_j^{(k)}\| = \gamma \|\partial_{w_i} F(\mu_k) - \partial_{w_j} F(\mu_k)\| \leq \gamma\ell \|w_i^{(k)} - w_j^{(k)}\| < \|w_i^{(k)} - w_j^{(k)}\|. \tag{53}$$

D. Lipschitz displacement convex optimization

D.1. The existence of sub-gradients

Although the subgradient may not exist for general non-convex non-smooth functions, the next Lemma proves it does exists for displacement convex functions under a weak assumption.

Lemma D.1. *The sub-gradient at $w_1 \neq \ldots \neq w_n$ does exist for displacement convex functions of n-sparse measures with a support whose elements are bounded.*

Proof. As a warm-up, we prove the statement for Fréchet differentiable functions using a straightforward application of displacement convexity. Recall the definition of displacement convexity as

$$F\left(\frac{1}{n}\sum_{i=1}^n \delta_{(1-t)w_i + tT^*(w_i)}\right) \leq (1-t)F(\mu) + tF(\nu) \tag{54}$$

Note that the above inequality turns into equality at $t = 0$ and $t = 1$. Thus, the derivative of left-side and the right side are equal at $t = 0$ and $t = 1$, namely

$$\frac{d}{dt}|_{t=0} F\left(\frac{1}{n}\sum_{i=1}^n \delta_{(1-t)w_i + tT^*(w_i)}\right) \leq F(\nu) - F(\mu) \quad (*) \tag{55}$$

When F is Fréchet differentiable, the above inequality coincides with

$$\sum_i \langle \nabla_{w_i} F, T^*(w_i) - w_i\rangle \leq F(\nu) - F(\mu). \tag{56}$$

Thus, the gradients are also sub-gradient for differentiable functions. To extend the proof to non-differentiable functions, we need to establish a bound on

$$\liminf_{t \to 0_+} t^{-1} \left(F\left(\frac{1}{n} \sum_{i=1}^n \delta_{w_i + t(T^*(w_i) - w_i)} \right) - F(\mu) \right) \tag{57}$$

where $t \to 0_+$ denotes any decreasing sequence of positive real numbers. Since the optimal transport between μ and μ_t remains constant for a small t, F is convex in a small neighborhood of μ. This local convexity structure ensures the existence of subgradients $g_1(\mu), \dots, g_n(\mu) \in \mathbb{R}^d$ such that

$$\liminf_{t \to 0_+} t^{-1} \left(F\left(\frac{1}{n} \sum_{i=1}^n \delta_{w_i + t(T^*(w_i) - w_i)} \right) - F(\mu) \right) \geq \sum_i \langle g_i(\mu), T^*(w_i) - w_i \rangle$$

Plugging the above inequality into (*) concludes the proof. □

D.2. Proof of Theorem 5.1.a

The proof is inspired by the convergence analysis of gradient descent for non-smooth convex functions (Theorem 3.9 of (Bubeck et al., 2015)). The injection with noise ensures that the particles remain distinct with probability one, hence the optimal transport from μ_k to $\widehat{\mu}$ denoted by T exists with probability one. Leveraging the optimal transport T and inequality (20) obtained by λ-diplacement convexity, we get

$$\mathbb{E}\left[W_2^2\left(\mu_{k+1}, \widehat{\mu}\right)\right] \leq \mathbb{E}\left[\sum_i \|T(w_i) - w_i\|^2 + \gamma_k \underbrace{2 \sum_i \langle T(w_i) - w_i, \partial_{w_i} F'(\mu_k) \rangle}_{\leq -\lambda W_2^2(\mu_k, \widehat{\mu}) + 2F(\widehat{\mu}) - 2F(\mu_k))} \right] + \gamma_k^2 \underbrace{\sum_i \mathbb{E}\left\|\partial_{w_i} F(\mu_k)\right\|^2}_{\leq L^2} + \frac{\gamma_k^2}{n} \sum_{i=1}^n \mathbb{E}\left[\|\xi_i^{(k)}\|^2\right] \tag{58}$$

A rearrangement of terms yields

$$k\,\mathbb{E}\left[F(\mu_k) - F(\widehat{\mu})\right] \leq \lambda k(k-1)\,\mathbb{E}\left[W_2^2(\mu_k, \widehat{\mu})\right] - \lambda k(k+1)\,\mathbb{E}\left[W_2^2(\mu_{k+1}, \widehat{\mu})\right] + \frac{(L^2+1)}{\lambda} \tag{59}$$

Summing over $k = 1, \dots, m$ concludes the proof as

$$\left(\frac{m(m+1)}{2}\right) \min_{k \leq m} \left(\mathbb{E}\left[F(\mu_k) - F(\widehat{\mu})\right]\right) \leq \sum_{k=1}^m k \left(\mathbb{E}\left[F(\mu_k) - F(\widehat{\mu})\right]\right) \leq \frac{m(L^2+1)}{\lambda}. \tag{60}$$

D.3. Proof of Theorem 5.1.b

The proof is inspired by Theorem 3.2.2 of (Nesterov, 2003). Suppose $\widehat{\mu}$ the minimizer and let T denotes the optimal mapping from μ_k to $\widehat{\mu}$. The injection with noise ensure that particles of μ_k are distinct, hence T does exists.

$$\mathbb{E}\left[W_2^2\left(\mu_{k+1}, \mu_*\right)\right] \leq \mathbb{E}\left[\sum_i \left(\|T(w_i) - w_i\|^2 + 2\gamma \langle T(w_i) - w_i, \partial_{w_i} F(\mu_k) \rangle \right) \right] + \gamma^2 \underbrace{\sum_i \mathbb{E}\left\|\partial_{w_i} F(\mu)\right\|^2}_{\leq L^2} + \frac{\gamma^2}{n} \sum_{i=1}^n \mathbb{E}\,\|\xi_i^{(k)}\|^2 \tag{61}$$

Using Eq. (20), we conclude that

$$-\int \langle T(w)-w, \partial_w F(\mu)\rangle d\mu_k(w) \geq \underbrace{F(\mu_k)-F(\widehat{\mu})}_{\Delta_k} \geq 0 \tag{62}$$

Therefore,

$$\mathbb{E}\left[W_2^2\left(\mu_{k+1},\widehat{\mu}\right)\right] \leq \mathbb{E}\left[W_2^2\left(\mu_k,\widehat{\mu}\right)\right] - 2\gamma\,\mathbb{E}\left[\Delta_k\right] + \gamma^2(L^2+1) \tag{63}$$

Summing over k concludes the proof

$$(m\gamma)\min_{k\in\{1,\dots,m\}} \mathbb{E}\left[\Delta_k\right] \leq \sum_{k=1}^{m} \gamma\,\mathbb{E}\left[\Delta_k\right] \leq W_2^2(\mu_0,\widehat{\mu}) + m\gamma^2(L^2+1) \tag{64}$$

E. Approximation error

E.1. Proof of Lemma 6.4

The proof is based on the convergence of Frank-Wolf algorithm. This proof technique is previously used for specific functions in neural networks (Bach, 2017). Since this algorithm uses a (infinite-dimensional) non-convex optimization in each step, it is not implementable. Yet, we can use its convergence properties to bound the approximation error. To introduce the algorithm, we first need to formulate the optimization over a compact domain in Banach space. We optimize F over L^2 Hilbert spaces of functions. Let D is the set of probability measures over a compact set, which is a subset of L^2. Frank-Wolfe method optimizes F through the following iterations (Jaggi, 2013)

$$\mu^{(k+1)} = (1-\gamma)\mu^{(k)} + \gamma s, \quad s = \arg\max_{\nu\in D} \int F'(\mu^{(k)})(x)d\nu(x), \tag{65}$$

where F' is the functional derivative of F with respect to μ (Santambrogio, 2017). It is easy to check that s is always a Dirac measure at $\max_x F'(\mu^{(k)})(x)$. Hence, $\mu^{(n-1)}$ is a sparse measure over n particles as long as $\mu^{(0)} = \delta_{w_0}$. The compactness of D, convexity and smoothness of F ensures the rate $O(1/n)$ for the convergence of Frank-Wolfe method (Jaggi, 2013).

E.2. Proof of Lemma 6.5

The proof is a straightforward application of the key observation in (Daneshmand & Bach, 2022). Daneshmand & Bach (2022) show the energy distance can be written alternatively as an MMD_K distance with a positive definite kernel K, which is quadratic convex function in μ. Since the kernel K is Lipschitz, L is smooth in the measure.

E.3. Proof of Corollary 6.6

Daneshmand & Bach (2022) prove that L is equivalent to the E in polar coordinates. Invoking part b of Theorem 5.1 concludes the rate.