Difference of Convex Programming in the Wasserstein Space with Applications to MMD Optimization

Clément Bonet
CMAP, CNRS, Ecole Polytechnique, IP Paris
clement.bonet.mapp@polytechnique.edu

Pierre-Cyril Aubin-Frankowski
CERMICS, CNRS, ENPC, IP Paris
pierre-cyril.aubin@enpc.fr

Youssef Mroueh
IBM Research
mroueh@us.ibm.com

Abstract

Optimizing functionals over the space of probability measures is now ubiquitous in machine learning. A widely used approach is to perform the optimization directly over the Wasserstein space, but many objective functionals of practical interest are non-convex along Wasserstein geodesics, making the analysis of standard first-order methods challenging. In this work, we study a class of objectives over the Wasserstein space that admit a difference-of-convex (DC) decomposition and we lift the classical convex-concave procedure (CCCP) to this setting. Under smoothness and strong convexity assumptions on the convex components of the decomposition, we prove almost stationarity along the iterates of the resulting algorithm. Our main focus is on the Maximum Mean Discrepancy (MMD) and the Energy Distance (ED) functionals, for which we develop explicit Wasserstein DC decompositions, and establish local convergence of the scheme under mild assumptions. Empirically, we show that well-chosen DC decompositions yield faster and more stable convergence than Wasserstein gradient descent on these MMD objectives.

1 Introduction

Optimizing over the space of probability measures is an important problem in machine learning, which has received attention to solve problems ranging from variational inference [15, 46, 65] to generative modeling [24, 29, 83], reinforcement learning [66, 92], optimization of neural networks [27, 57] or for modeling dynamics of cells [23, 64, 80]. One prominent way to optimize over the space of probability measures is to equip it with the Wasserstein distance [71, 84], and to discretize the associated Wasserstein gradient flows [43, 87]. This allows to design many optimization algorithms as counterparts of their Euclidean version, such as gradient descent [87], proximal point and gradient algorithms [43, 70], coordinate descent [88] or mirror descent [17, 73].

While these optimization methods have demonstrated good results in minimizing several objectives such as the Kullback-Leibler divergence (KL) [87], f-divergences [5, 35, 51], the Energy Distance [41] or the Sliced-Wasserstein distance [18, 30, 52], most of them are tailored for convex functionals along (generalized) geodesics. However, many popular functionals are known to be non-convex. This is the case for instance of the squared Wasserstein distance itself [4, Chapter 9], the Sliced-Wasserstein distance [20], the KL with non-log-concave target [55], the Energy distance [40] or the Maximum Mean Discrepancy (MMD) [6]. An alternative to the convexity assumption is to consider Polyak-Łojaziewicz inequalities [14, 50, 93], however these are so far only known to hold in restrictive cases such as trivially for the squared Wasserstein distance, the KL *w.r.t.* a measure satisfying the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

log-Sobolev inequality [84, Chapter 21], the Sliced-Wasserstein distance over Gaussians [81], the MMD with Coulomb kernels and smooth initializations [21, Section 2].

For the MMD specifically, to address its non-convexity in the Wasserstein space, Arbel et al. [6] proposed injecting noise to the gradient. However, tuning the amount of noise remains delicate. Gladin et al. [37] performed instead the optimization of MMD in the Wasserstein-MMD space, where MMD is convex. Their method improves performance, but it requires to change weights of the distribution, mixing two implicit steps in squared Wasserstein and MMD. More recently, Belhadji et al. [12] tackled the same problem in the Wasserstein Fisher-Rao space through a fixed-point algorithm, also changing weights. We propose instead a new particle-based algorithm on the Wasserstein space to handle non-convex functionals.

A natural candidate for such an extension is the family of difference-of-convex (DC) methods on $\mathbb{R}^d$ [49, 67], designed for objectives that can be written as the difference of two convex functions. This class is large: on $\mathbb{R}^d$ it contains in particular all twice continuously differentiable functions [42, 90]. Its most prominent algorithm is the Convex-Concave Procedure (CCCP) [90]. DC methods have found many applications in machine learning, among which kernel selection [7], clustering [79], dictionary learning [85], optimal transport [82] and neural network training [8, 9]. However, except for a few works that studied DC algorithms on Riemannian manifolds [13, 33, 74, 86], and recently on the Wasserstein space [54, 55], optimization algorithms for DC functionals have mostly been studied on Euclidean spaces.

Contributions. In this work, we focus on developing an optimization scheme on the Wasserstein space tailored for the non-convex objectives that can be written as a difference of convex objectives. To do so, we introduce the *Wasserstein Convex-Concave Procedure* (WCCCP), and analyze its theoretical convergence, proving almost stationarity along its iterates. The closest work to ours is Luu et al. [55] which used a Wasserstein Proximal Gradient algorithm [70] to solve some DC problems, but which restricted their applications to functionals whose concave part is a potential energy, *i.e.* linear. We argue that it is too restrictive to handle objectives such as MMD, which can be decomposed as a sum of two non-convex quadratic and linear terms. Thus, we deal with the more general case, where the concave part can be any differentiable functional over the Wasserstein space. Then, for several kernels, we show that with a well-chosen decomposition obtained by splitting the kernel, the scheme introduced can better optimize the MMD than Wasserstein Gradient Descent. In particular, we provide experiments on the Energy distance and on the MMD with Gaussian kernel.

Notation. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability distributions with finite second moments, and by $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ its restriction to absolutely continuous measures with respect to the Lebesgue measure. Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $\mathrm{W}_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)$ the squared Wasserstein distance, where $\Pi(\mu,\nu)$ is the set of couplings between μ and ν, and $\Pi_o(\mu,\nu)$ is its subset of optimal couplings. The metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ is called the Wasserstein space. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $L^2(\mu)$ the Hilbert space of functions $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\int \|f(x)\|_2^2 \,\mathrm{d}\mu(x) < \infty$ equipped with the norm $\|f\|_{L^2(\mu)}^2 = \int \|f(x)\|_2^2 \,\mathrm{d}\mu(x)$ and with inner product $\langle\cdot,\cdot\rangle_{L^2(\mu)}$. Given $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^d$, $\mathrm{T}_\#\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the pushforward measure of μ.

2 Background

We begin by recalling a few facts on optimization on the Wasserstein space. More precisely, we recall the notion of Wasserstein gradient, of total convexity on the Wasserstein space, and some classical optimization schemes on $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$. For more details about Wasserstein gradient flows, we refer to *e.g.* [4, 48]. Then, we provide a brief introduction to the convex-concave procedure on Euclidean spaces.

Wasserstein gradient. Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional. It admits a Wasserstein gradient $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) \in L^2(\mu)$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in \Pi_o(\mu,\nu)$, the following first order Taylor expansion is satisfied [19, 48]

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), y - x\rangle \,\mathrm{d}\gamma(x,y) + o\big(\mathrm{W}_2(\mu,\nu)\big). \tag{1}$$

When it exists, the Wasserstein gradient may not be unique in $L^2(\mu)$ in general. Nonetheless, there is only one gradient living in the tangent space $T_\mu\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$ which is a closed linear subspace

of $L^2(\mu)$. Hence, by Hilbert's decomposition theorem, any gradient can be decomposed as a part in $T_\mu \mathcal{P}_2(\mathbb{R}^d)$ and an orthogonal part $\xi(\mu)$ which satisfies $\int \langle \xi(\mu), y - x \rangle \, \mathrm{d}\gamma(x, y) = 0$, see [48, Proposition 2.11]. Thus, without loss of generality, we always work with the unique $\nabla_{\mathrm{W}_2} \mathcal{F}(\mu) \in T_\mu \mathcal{P}_2(\mathbb{R}^d)$ for a Wasserstein differentiable functional, using the shorthand W-differentiable for such functionals.

Classical functionals from $\mathcal{P}_2(\mathbb{R}^d)$ to $\mathbb{R}$ include potential energies $\mathcal{V}(\mu) = \int \mathrm{V} \mathrm{d}\mu$ and interaction energies $\mathcal{W}(\mu) = \frac{1}{2} \iint \mathrm{W}(x-y) \, \mathrm{d}\mu(x) \mathrm{d}\mu(y)$ for $\mathrm{V}, \mathrm{W} : \mathbb{R}^d \to \mathbb{R}$, with W symmetric. They are both differentiable provided V and W are differentiable and smooth enough [48], and their Wasserstein gradients read respectively as $\nabla_{\mathrm{W}_2} \mathcal{V}(\mu) = \nabla \mathrm{V}$ and as the convolution $\nabla_{\mathrm{W}_2} \mathcal{W}(\mu) = \nabla \mathrm{W} * \mu$. In this work, we will mostly focus on functionals obtained as a sum of interaction and potential energies, as they include in particular the MMD [6].

Convexity in the Wasserstein space. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$. Given $\phi : L^2(\mu) \to \mathbb{R}$ convex and Gâteaux differentiable, the Bregman divergence on $L^2(\mu)$ between T, S is defined as [34]

$$\mathrm{D}_\phi(\mathrm{T}, \mathrm{S}) = \phi(\mathrm{T}) - \phi(\mathrm{S}) - \langle \nabla \phi(\mathrm{S}), \mathrm{T} - \mathrm{S} \rangle_{L^2(\mu)}. \tag{2}$$

Given $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ W-differentiable, we can define the Bregman divergence in $L^2(\mu)$ associated to the lifted functional $\mathrm{T} \mapsto \mathcal{F}(\mathrm{T}_\# \mu)$ as

$$\mathrm{D}^\mu_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) = \mathcal{F}(\mathrm{T}_\# \mu) - \mathcal{F}(\mathrm{S}_\# \mu) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mathrm{S}_\# \mu) \circ \mathrm{S}, \mathrm{T} - \mathrm{S} \rangle_{L^2(\mu)}, \tag{3}$$

using the chain rule for the gradient of $\tilde{\mathcal{F}}_\mu : \mathrm{S} \mapsto \mathcal{F}(\mathrm{S}_\# \mu)$ which implies $\nabla \tilde{\mathcal{F}}_\mu(\mathrm{S}) = \nabla_{\mathrm{W}_2} \mathcal{F}(\mathrm{S}_\# \mu) \circ \mathrm{S}$, see [17, Proposition 1]. Note that for $\mathcal{F}(\mu) = \int \frac{1}{2} \| \cdot \|_2^2 \, \mathrm{d}\mu$, this reduces to $\mathrm{D}^\mu_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) = \frac{1}{2} \|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)}$. Let $\alpha \geq 0$, we say that $\mathcal{F}$ is α-totally convex [25, 63, 77] if for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{D}^\mu_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) \geq \frac{\alpha}{2} \|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)}$. Equivalently, it satisfies

$$\forall t \in [0,1], \ \mathcal{F}\big(\big((1-t)\mathrm{T} + t\mathrm{S}\big)_\# \mu\big) \leq (1-t)\mathcal{F}(\mathrm{T}_\# \mu) + t\mathcal{F}(\mathrm{S}_\# \mu) - \alpha \frac{t(1-t)}{2} \|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)}. \tag{4}$$

If this result holds only for $\mathrm{S} = \mathrm{Id}$ and for T the gradient of a convex function, this corresponds to the less restrictive notion of strong convexity along geodesics [4]. If V is α-strongly convex, then $\mathcal{V}$ is α-totally convex, and if W is convex, then $\mathcal{W}$ is also totally convex [4, Chapter 9]. Given $\beta \geq 0$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$ and defining the linear interpolation $\mathrm{T}_t = (1-t)\mathrm{T} + t\mathrm{S}$ for $t \in [0,1]$, $\mathcal{F}$ is said to be β-smooth along the curve $t \mapsto (\mathrm{T}_t)_\# \mu$ if for all $t, s \in [0,1]$, $\mathrm{D}^\mu_{\mathcal{F}}(\mathrm{T}_t, \mathrm{T}_s) \leq \frac{\beta}{2} \|\mathrm{T}_t - \mathrm{T}_s\|^2_{L^2(\mu)}$.

Wasserstein Gradient Descent. Optimization on $L^2(\mu)$ and on $\mathcal{P}_2(\mathbb{R}^d)$ are very much intertwined in practice, see *e.g.* [17, 31]. For instance, the Wasserstein Gradient Descent (WGD) over a W-differentiable functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ which is defined for all $k \geq 0$, $\tau > 0$, as $\mu_{k+1} = \big(\mathrm{Id} - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\big)_\# \mu_k$, can be written at each iteration in two steps: first solving an optimization problem on $L^2(\mu_k)$ to get a map $\mathrm{T}_{k+1} \in L^2(\mu_k)$, then pushing forward μ_k by T_{k+1}, *i.e.*

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \frac{1}{2\tau} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_\# \mu_k. \end{cases} \tag{5}$$

This scheme is known to converge if $\mathcal{F}$ is smooth along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_\# \mu_k$ and convex along geodesics. Other first-order algorithms have been lifted from $\mathbb{R}^d$ to $\mathcal{P}_2(\mathbb{R}^d)$. For instance, replacing the squared L^2 distance in (5) by a Bregman divergence (2) allows to lift the Mirror descent algorithm [11, 53] to $\mathcal{P}_2(\mathbb{R}^d)$ [17].

The Convex-Concave Procedure. A function $f : \mathbb{R}^d \to \mathbb{R}$ is DC if it can be written as the difference of two convex functions, *i.e.* if there exists f^+, f^- two convex functions such that $f = f^+ - f^-$. Every C^1 function with Lipschitz gradient is DC [42, Section II]. To minimize such functions, a popular algorithm is the Convex-Concave Procedure (CCCP) [90]. This algorithm amounts to linearizing the concave part around the current iterate $x_k \in \mathbb{R}^d$, which by convexity gives the lower bound,

$$\forall x \in \mathbb{R}^d, \ f^-(x) \geq f^-(x_k) + \langle \nabla f^-(x_k), x - x_k \rangle, \tag{6}$$

which entails an upper bound on $f = f^+ - f^-$, and the majorization-minimization

$$x_{k+1} = \mathrm{argmin}_x \ f^+(x) - f^-(x_k) - \langle \nabla f^-(x_k), x - x_k \rangle, \ \forall k \geq 0. \tag{7}$$

When both f^+ and f^- are differentiable, the iterates satisfy $\nabla f^+(x_{k+1}) = \nabla f^-(x_k)$ by the first order conditions. This algorithm belongs to the more general family of Difference-of-Convex algorithms (DCA) [49, 67], and is related to several optimization algorithms including Frank-Wolfe [91], the Mirror and Bregman proximal descent [32] or the Proximal gradient algorithm [68].

The first convergence analysis of CCCP focused on obtaining asymptotic convergence, showing that it converges towards a stationary point under some assumptions [47, 78]. Then, several works such as [1, 32, 91] derived non-asymptotic convergence rates. In particular, the algorithm was shown to converge in $O(1/k)$ in terms of the squared norm of the gradient. More recently, Oikonomidis et al. [62] provided an analysis of CCCP under a generalized convexity perspective, though in finite dimensions. Linear rates were also derived under Polyak-Łojaziewicz inequalities adapted to DC functions [1, 32, 60, 62]. From a complementary perspective, Maskan et al. [56] provided a convergence analysis by solving each step (7) with Frank-Wolfe and analyzed the convergence for different DC decompositions in that setting.

3 Wasserstein CCCP

In this section, we first introduce the Wasserstein Convex-Concave Procedure (WCCCP) to minimize difference-of-convex functions on the Wasserstein space as well as our assumptions. Then we provide a theoretical analysis in several settings, including the convex and non-convex ones. Finally, we discuss how we can implement these schemes in practice. All the proofs are deferred to Appendix F.

3.1 Convex-Concave Procedure in the Wasserstein Space

We focus on the problem of minimizing $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ where $\mathcal{F}$ can be decomposed as

$$\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu), \quad \forall \mu \in \mathcal{P}_2(\mathbb{R}^d), \tag{8}$$

with $\mathcal{F}^+, \mathcal{F}^- : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ both totally convex. Moreover, we assume that $\mathcal{F}^-$ is W-differentiable, and will assume $\mathcal{F}^+$ W-differentiable on a case-by-case basis. In contrast to [55], we do not restrict $\mathcal{F}^-$ to be a potential energy.

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Since $\mathcal{F}^-$ is totally convex, for any $\mathrm{T} \in L^2(\mu)$,

$$\mathrm{D}^{\mu}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}) \geq 0 \iff \mathcal{F}^-(\mathrm{T}_{\#}\mu) \geq \mathcal{F}^-(\mu) + \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)}. \tag{9}$$

Hence, we have the following upper bound on $\mathcal{F}$:

$$\mathcal{F}(\mathrm{T}_{\#}\mu) = \mathcal{F}^+(\mathrm{T}_{\#}\mu) - \mathcal{F}^-(\mathrm{T}_{\#}\mu) \leq \mathcal{F}^+(\mathrm{T}_{\#}\mu) - \mathcal{F}^-(\mu) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)}. \tag{10}$$

We define the Wasserstein Convex-Concave Procedure (WCCCP) as the majorization-minimization based on the upper bound in (10) at each iteration $k \geq 0$, *i.e.* given $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{J}(\mathrm{T}) \coloneqq \mathcal{F}^+(\mathrm{T}_{\#}\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k. \end{cases} \tag{11}$$

This extends CCCP [90] to the Wasserstein space, CCCP being recovered when $\mathcal{F}^+$ and $\mathcal{F}^-$ are potential energies. Restricting the optimization to measures of the form $\mathrm{T}_{\#}\mu_k$ for $\mathrm{T} \in L^2(\mu_k)$ is without loss of generality in two important cases: *i)* as done in our experiments, for empirical target and initial distributions with the same number of particles; *ii)* if $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, as by Brenier's theorem [22], there always exists an OT map between μ_k and any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. While for greater generality we could instead optimize over couplings with first marginal μ_k, the subproblems on $L^2(\mu_k)$ are more tractable and reflect practical implementations.

We assume existence and uniqueness in (11) for simplicity of exposition. Sufficient conditions are $\mathrm{T} \mapsto \mathcal{F}^+(\mathrm{T}_{\#}\mu_k)$ strictly convex, coercive and lower semicontinuous over $L^2(\mu_k)$, for all k. We now study the theoretical convergence of the WCCCP in several settings.

3.2 Theoretical Analysis in the Non-Convex Case

We first observe that in general, similarly to the analysis of [32] in the Euclidean case, (11) is equivalent to both a Mirror Descent and a Bregman Proximal Descent in the Wasserstein space on the objective $\mathcal{F}$, with Bregman potential respectively $\mathcal{F}^+$ and $\mathcal{F}^-$, and with step size $\tau = 1$.

Proposition 1. *Let $\mu_k \in \mathcal{P}_2(\mathbb{R}^d)$ for some $k \geq 0$. (11) is equivalent to (Bregman Proximal Descent)*

$$\mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}) + \mathcal{F}(\mathrm{T}_{\#}\mu_k), \tag{12}$$

and, if $\mathcal{F}^+$ is W-differentiable, to (Mirror Descent)

$$\mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{T}, \mathrm{Id}) + \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)}. \tag{13}$$

These equivalences allow us to provide a convergence rate in the case where $\mathcal{F}$ satisfies convexity assumptions relative to $\mathcal{F}^+$ or $\mathcal{F}^-$, leveraging the convergence analysis of Mirror descent [17] and of Bregman Proximal Descent. We refer to Appendix B for these results and we focus only on the non-convex case here.

Descent Lemma. If both $\mathcal{F}^+, \mathcal{F}^-$ are W-differentiable, then we can take the first order conditions in (11) which yield the equivalent update

$$\nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k). \tag{14}$$

Leveraging this result, we obtain the following relation between $\mathcal{F}(\mu_{k+1})$ and $\mathcal{F}(\mu_k)$ involving Bregman divergences with Bregman potential $\mathcal{F}^+$ and $\mathcal{F}^-$.

Proposition 2. *We have for all $k \geq 0$*

$$\mathcal{F}(\mu_{k+1}) = \mathcal{F}(\mu_k) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) - \mathcal{D}^k_{\mathcal{F}^+} \tag{15}$$

where $\mathcal{D}^k_{\mathcal{F}^+} \coloneqq \mathcal{F}(\mu_k) - \min_{\mathrm{T} \in L^2(\mu_k)} \ \mathcal{F}^+(\mathrm{T}_{\#}\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)}$ and

$$\mathcal{D}^k_{\mathcal{F}^+} = \mathcal{F}^+(\mu_k) - \mathcal{F}^+(\mu_{k+1}) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k), \mathrm{Id} - \mathrm{T}_{k+1} \rangle_{L^2(\mu_k)}. \tag{16}$$

Assume that $\mathcal{F}^+$ is W-differentiable. Then, for all $k \geq 0$, $\mathcal{D}^k_{\mathcal{F}^+} = \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{Id}, \mathrm{T}_{k+1})$, hence

$$\mathcal{F}(\mu_{k+1}) = \mathcal{F}(\mu_k) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) - \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{Id}, \mathrm{T}_{k+1}). \tag{17}$$

The term $\mathcal{D}^k_{\mathcal{F}^+}$, used also in the non-smooth and Euclidean case in [1, eq.(15)] or [91], is merely a proxy for $\mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{Id}, \mathrm{T}_{k+1})$ when $\mathcal{F}^+$ is possibly not W-differentiable. Note that $\mathcal{D}^k_{\mathcal{F}^+} \geq 0$, as taking $\mathrm{T} = \mathrm{Id}$ to upper bound the min, the difference vanishes. Hence, (15) implies that $\mathcal{F}$ is non-increasing along the WCCCP scheme. The condition $\mathcal{D}^k_{\mathcal{F}^+} = 0$ can thus be used as a termination criterion as it implies that μ_k is a critical point of $\mathcal{F}$, as shown in

Proposition 3. *Assume that $\mathcal{F}^+$ is W-differentiable, then $\mathcal{D}^k_{\mathcal{F}^+} = 0$ implies that $\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k) = 0$.*

Sublinear rates. We now derive sublinear rates for the non-convex case. All the next results build on the next immediate result (obtained by telescoping), which provides a bound for any arbitrary sequence and corresponds actually to the more involved [32, Theorem 5],

$$\min_{k \in \{0, \dots, K-1\}} \mathcal{F}(\mu'_k) - \mathcal{F}(\mu'_{k+1}) \leq \frac{\mathcal{F}(\mu'_0) - \mathcal{F}(\mu'_K)}{K}, \quad \forall (\mu'_k)_{k \geq 0}, \ K \geq 1. \tag{18}$$

In the next proposition, we show a sublinear convergence rate for the termination criterion $\mathcal{D}^k_{\mathcal{F}^+}$, in accordance with results over $\mathbb{R}^d$ from [1, 32, 91].

Proposition 4. *Assume that $\mathcal{F}$ is bounded from below. Then, for all $K \geq 1$,*

$$0 \leq \min_{k \in \{0, \dots, K-1\}} \mathcal{D}^k_{\mathcal{F}^+} \leq \frac{\mathcal{F}(\mu_0) - \inf \mathcal{F}}{K}. \tag{19}$$

Assuming that either $\mathcal{F}^+$ or $\mathcal{F}^-$ is strongly convex along iterates, we can also obtain a sublinear convergence result for the squared distance between iterates and the norm of the W-gradient, which guarantees that at least one iterate of WCCCP is almost a stationary point.

Proposition 5. *Let $\alpha^+, \alpha^- \geq 0$ such that $\alpha^+ + \alpha^- > 0$. Assume $\mathcal{F}^+$ and $\mathcal{F}^-$ are respectively α^+ and α^- totally-convex. Then for all $K \geq 1$,*

$$\min_{0\leq k\leq K-1} \mathrm{W}_2^2(\mu_k, \mu_{k+1}) \leq \min_{0\leq k\leq K-1} \|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \leq \frac{2}{\alpha^+ + \alpha^-} \frac{\big(\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)\big)}{K}. \quad (20)$$

Furthermore, if $\mathcal{F}^+$ is W-differentiable and $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ satisfies the Lipschitz condition:

$$\|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k)\|_{L^2(\mu_k)} \leq L\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)} \, \forall k \geq 0, \quad (21)$$

then

$$\min_{0\leq k\leq K-1} \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|^2_{L^2(\mu_k)} \leq \frac{2L^2}{\alpha^+ + \alpha^-} \frac{\big(\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)\big)}{K}. \quad (22)$$

3.3 Computing WCCCP

To solve the WCCCP scheme (11), we need to minimize at each iteration J on $L^2(\mu_k)$. First, we observe that if $\mathcal{F}^+$ is convex along curves of the form $t \mapsto \big((1-t)\mathrm{T} + t\mathrm{S}\big)_{\#}\mu_k$ for any $\mathrm{S}, \mathrm{T} \in L^2(\mu_k)$, then J is convex on $L^2(\mu_k)$ as $\mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{T}, \mathrm{S}) \geq 0$, see Section 2.

If $\mathcal{F}^+$ is W-differentiable, then $\mathrm{T} \mapsto \mathcal{F}^+(\mathrm{T}_{\#}\mu_k)$ is Fréchet differentiable and taking the first order conditions in (11), we get the equivalent update (14). If $\mathcal{F}^+(\mu) = \int \mathrm{V} \mathrm{d}\mu$ with V strictly convex, then $\mathrm{T}_{k+1} = \nabla \mathrm{V}^* \circ \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)$ with $\mathrm{V}^*(x) = \sup_y \langle x, y\rangle - \mathrm{V}(y)$ the Legendre transform of V. For more general functionals, the problem is implicit and, to the best of our knowledge, cannot be solved in closed-form.

Nonetheless, since J is convex on $L^2(\mu_k)$, under the Lipschitz condition (21) (corresponding to L-smoothness of $\mathrm{T} \mapsto \mathcal{F}^+(\mathrm{T}_{\#}\mu_k)$ for all $k \geq 0$), we can perform a gradient descent on $L^2(\mu_k)$, which is of the form, for $0 < \tau \leq 1/L$ and $\tilde{\mathrm{T}}_k^0 = \mathrm{Id}$,

$$\forall \ell \geq 0, \; \tilde{\mathrm{T}}_k^{\ell+1} = \tilde{\mathrm{T}}_k^{\ell} - \tau\big(\nabla_{\mathrm{W}_2}\mathcal{F}^+\big((\tilde{\mathrm{T}}_k^{\ell})_{\#}\mu_k\big) \circ \tilde{\mathrm{T}}_k^{\ell} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)\big). \quad (23)$$

Then, we approximate the solution of (11) by $\tilde{\mathrm{T}}_{k+1} = \tilde{\mathrm{T}}_k^M$ and $\mu_{k+1} = (\tilde{\mathrm{T}}_{k+1})_{\#}\mu_k$. If $\mu_k^\ell = (\tilde{\mathrm{T}}_k^\ell)_{\#}\mu_k = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{k,\ell}}$, then the update translates to the particles as

$$\forall k \geq 0, \ell \geq 0, \; x_i^{k,\ell+1} = x_i^{k,\ell} - \tau\big(\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k^\ell)(x_i^{k,\ell}) - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)(x_i^{k,0})\big). \quad (24)$$

In practice, one can follow this procedure for any τ small enough, and we run the algorithm for $0 < k \leq K$ outer steps (11), and, at each k, $0 < \ell \leq M$ inner steps of (24) to minimize J. Alternatively one could rely on root-finding algorithms instead, such as Newton's method [94].

3.4 Connection with Wasserstein Proximal Gradient Algorithm

Luu et al. [55] proposed to solve the DC problem on $\mathcal{P}_2(\mathbb{R}^d)$ by using a Wasserstein Proximal Gradient Descent scheme [70], alternating between a gradient descent step on $-\mathcal{F}^-$ and a proximal (JKO) step on $\mathcal{F}^+$, *i.e.* using

$$\begin{cases} \nu_{k+1} = \big(\mathrm{Id} + \tau\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)\big)_{\#}\mu_k \\ \mu_{k+1} = \mathrm{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \; \frac{1}{2\tau}\mathrm{W}_2^2(\mu, \nu_{k+1}) + \mathcal{F}^+(\mu). \end{cases} \quad (25)$$

In Proposition 6, we show that this is equivalent to minimizing an upper bound similar to that of WCCCP (10), with an additional quadratic cost.

Proposition 6. *Assume $\mu_k, \nu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ for all $k \geq 0$. Then (25) is equivalent to*

$$\begin{cases} \tilde{\mathrm{T}}_{k+1} = \underset{\mathrm{T}\in L^2(\mu_k)}{\mathrm{argmin}} \; \mathcal{F}^+(\mathrm{T}_{\#}\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} + \frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} \\ \mu_{k+1} = (\tilde{\mathrm{T}}_{k+1})_{\#}\mu_k. \end{cases} \quad (26)$$

Hence, this scheme can be seen as a lifting of CCCP algorithms with additional quadratic terms [76] which have been in particular used to deal with the stochastic setting [26, 59, 89]. Note that it would also be possible to use instead a Bregman divergence on $L^2(\mu_k)$, and that it would be equivalent with

a Bregman proximal gradient scheme, see *e.g.* [17, Appendix F]. Furthermore, the WCCCP algorithm can be seen as minimizing in (12) a regularization $\mathcal{F}(\mathrm{T}_{\#}\mu_k) + \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}, \mathrm{Id})$ of $\mathcal{F}(\mathrm{T}_{\#}\mu_k)$, and (26) hence corresponds to replacing $\mathcal{F}(\mathrm{T}_{\#}\mu_k)$ by $\mathcal{F}(\mathrm{T}_{\#}\mu_k) + \frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)}$ in our analysis.

Luu et al. [55] restrict their attention to functionals whose concave part is a potential energy, $\mathcal{F}^-(\mu) = \int \mathrm{V}^- \,\mathrm{d}\mu$ with $\mathrm{V}^- : \mathbb{R}^d \to \mathbb{R}$ convex. This already covers a large class of objectives, including the KL divergence with respect to a possibly non-log-concave target whose potential admits a DC decomposition $\mathrm{V} = \mathrm{V}^+ - \mathrm{V}^-$ with $\mathrm{V}^+, \mathrm{V}^-$ respectively α^+- and α^--strongly convex. Our analysis of Section 3 also covers these objectives. In this case, $\mathcal{F}^-(\mu) = \int \mathrm{V}^- \,\mathrm{d}\mu$ and $\mathcal{F}^+(\mu) = \mathcal{H}(\mu) + \int \mathrm{V}^+ \,\mathrm{d}\mu + \mathrm{cst}$, where $\mathcal{H}(\mu) = \int \log \rho \,\mathrm{d}\mu$ denotes the negative entropy of $\mathrm{d}\mu = \rho \,\mathrm{dLeb}$ (and $\mathcal{H}(\mu) = +\infty$ otherwise). While $\mathcal{F}^-$ is α^--totally convex, $\mathcal{F}^+$ is only α^+-convex along generalized geodesics, *i.e.* along curves $t \mapsto \big((1-t)\mathrm{T} + t\mathrm{S}\big)_{\#}\mu$ where T and S are gradients of convex functions, hence OT maps [4, Chapter 9]. This weaker notion is nevertheless sufficient: we show in Proposition 7 that (20) still holds for this functional.

Proposition 7. *Let $\alpha^+ \geq 0$ and $\alpha^- > 0$, $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, $\mathrm{V}^- : \mathbb{R}^d \to \mathbb{R}$ and for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathcal{F}^-(\mu) = \int \mathrm{V}^- \,\mathrm{d}\mu$ and $\mathcal{F}^+(\mu) = \mathcal{H}(\mu) + \mathcal{G}(\mu)$. Assume V^- is α^--strongly convex, and $\mathcal{F}^+$ is α^+-convex along generalized geodesics. Then, we also have almost stationarity of WCCCP as given in (20).*

In the next section, we focus instead on the MMD, which can be decomposed as a DC function, whose concave part is a sum of an interaction and of a potential energy, and for which $\mathcal{F}^+$ and $\mathcal{F}^-$ are both totally convex.

4 DC Decomposition for the Maximum Mean Discrepancy

We now focus on finding a DC decomposition to the MMD [38]. Given $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a kernel, the squared MMD is defined as

$$\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \mathrm{MMD}_k^2(\mu, \nu) = \iint k(x, y)\ \mathrm{d}(\mu - \nu)(x)\mathrm{d}(\mu - \nu)(y). \tag{27}$$

It is well known that the squared MMD can be decomposed as a sum of an interaction energy and a potential energy [6], *i.e.*

$$\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu) = \frac{1}{2}\iint k(x, y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V}\ \mathrm{d}\mu + c(\nu), \tag{28}$$

with $\mathrm{V}(x) = -\int k(x, y)\ \mathrm{d}\nu(y)$ and $c(\nu) = \frac{1}{2}\iint k(x, y)\ \mathrm{d}\nu(x)\mathrm{d}\nu(y)$. The first term in (28) is an interaction term and the second a potential. This objective is in general not (geodesically) convex, but only semi-convex [6, Proposition 5], *i.e.* λ-totally convex with $\lambda \in \mathbb{R}$. Moreover, the performance of WGD to minimize it depends heavily on the kernel as observed in *e.g.* [6, 41, 44].

For an L-smooth kernel k, *i.e.* satisfying $\|\nabla k(x, y) - \nabla k(x', y')\|_2^2 \leq L^2\big(\|x - x'\|_2^2 + \|y - y'\|_2^2\big)$ for all $x, x', y, y' \in \mathbb{R}^d$, Luu et al. [55, Appendix A.2] proposed to use, for $\alpha \geq L$,

$$\mathcal{F}^+(\mu) = \alpha \int \|x\|_2^2 \,\mathrm{d}\mu(x) + \frac{1}{2}\iint k(x, y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + c(\nu),\ \mathcal{F}^-(\mu) = \int \big(\alpha\|x\|_2^2 - \mathrm{V}(x)\big)\ \mathrm{d}\mu(x) \tag{29}$$

as a DC decomposition of (28). Instead, we propose to obtain a Difference-of-Convex function for this objective by decomposing the kernel itself. For this, we focus on translation-invariant kernels.

MMD with translation-invariant kernel. A large class of useful kernels are the translation-invariant ones, *i.e.* those of the form $k(x, y) = \psi(x - y)$ for some symmetric function $\psi : \mathbb{R}^d \to \mathbb{R}$ [58, 75]. This class includes, among others, the Riesz (a.k.a. negative distance) kernel with $\psi(z) = -\|z\|_2$, the Gaussian kernel with $\psi(z) = e^{-\|z\|_2^2/(2h)}$ or the inverse multiquadric kernel with $\psi(z) = (c^2 + \|z\|_2^2)^{-\alpha}$ and $\alpha > 1$ [58]. Assuming that ψ admits a DC decomposition $\psi = \psi^+ - \psi^-$, we obtain the following DC decomposition of the MMD.

Table 1: DC decompositions satisfying Prop.9, i.e., $k(x,y) = q_+(s) - q_-(s)$, where $s = \|x-y\|_2^2$.

kernel	q_+	q_-	$\underline{\lambda}[q_+]$	$\underline{\lambda}[q_-]$	$\overline{\Lambda}[q_+]$	$\overline{\Lambda}[q_-]$	Ω	Ref
Smooth Riesz	0	$\sqrt{\varepsilon+s}$	0	$\frac{\varepsilon}{(\varepsilon+S_*)^{3/2}}$	0	$\varepsilon^{-1/2}$	compact	Lemma D.6
Gauss-Jordan	$e^{-s/(2h)} + s/2h$	$s/2h$	0	$1/h$	$\frac{1+2e^{-3/2}}{h}$	$1/h$	$\mathbb{R}^d$	Proposition D.13
Gauss-cosh/sinh	$\cosh(s/(2h))$	$\sinh(s/(2h))$	0	$1/h$	(49)	(50)	compact	Lemma D.4

Proposition 8. *Let k be a translation-invariant kernel of the form $k(x,y) = \psi(x-y)$ for all $x, y \in \mathbb{R}^d$, with a ψ admitting a DC decomposition $\psi = \psi^+ - \psi^-$, ψ^+, ψ^- being α^+, α^--strongly convex. Then (28) admits the DC decomposition $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ where for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\begin{cases} \mathcal{F}^+(\mu) = \frac{1}{2}\iint \psi^+(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V}^-\mathrm{d}\mu + c(\nu), & \mathrm{V}^-(\cdot) = \int \psi^-(\cdot - y)\,\mathrm{d}\nu(y), \\ \mathcal{F}^-(\mu) = \frac{1}{2}\iint \psi^-(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V}^+\mathrm{d}\mu, & \mathrm{V}^+(\cdot) = \int \psi^+(\cdot - y)\,\mathrm{d}\nu(y), \end{cases} \tag{30}$$

and $\mathcal{F}^+, \mathcal{F}^-$ are respectively α^- and α^+ totally convex. Note that ψ^+, ψ^- can always be chosen symmetric as, by symmetry of ψ, $2\psi(x) = \psi(x) + \psi(-x) = \psi^+(x) + \psi^+(-x) - \psi^-(x) - \psi^-(-x)$. They are also locally Lipschitz since they are convex and have full domain.

We now focus on the subclass of radial kernels, *i.e.* those for which there exists $q : \mathbb{R}_+ \to \mathbb{R}$ such that $\psi(z) = q(\|z\|_2^2)$. For such kernels, DC decompositions can be obtained via a Jordan decomposition

$$\begin{cases} q_+(x) = q(0) + (q'(0) + A)x + \int_0^x (x-t)\max\big(0, q''(t)\big)\,\mathrm{d}t, \\ q_-(x) = Ax - \int_0^x (x-t)\min\big(0, q''(t)\big)\,\mathrm{d}t, \end{cases} \tag{31}$$

whenever these integrals of the second derivative are computable, and taking $A \geq \max(0, -q'(0))$ to ensure that q_+ and q_- are nondecreasing (hence ψ_+ and ψ_- are convex). For the Gaussian kernel, $q(t) = e^{-t/(2h)}$ for $t \geq 0$, and (31) gives a simple DC decomposition.

Alternatively, if q is analytic, *i.e.* $q(t) = \sum_{i\in\mathbb{N}} a_i t^i$ for all $t \geq 0$, then we can set $I^+ \coloneqq \{i \,|\, a_i \geq 0\}$ and $I^- \coloneqq \{i \,|\, a_i \leq 0\}$, $q_+(t) = \sum_{i\in I^+} a_i t^i$, $q_-(t) = \sum_{i\in I^-} |a_i| t^i$. For the Gaussian kernel, this gives a DC decomposition based on cosh and sinh.

Lemma 1. *Let $q : t \mapsto e^{-\alpha t}$ for $\alpha > 0$. Then the DC decomposition obtained by (31) is $q_+(t) = e^{-\alpha t} + \alpha t$ and $q_-(t) = \alpha t$ for all $t \geq 0$. Similarly the algebraic DC decomposition is $q_+(t) = \cosh(\alpha t)$ and $q_-(t) = \sinh(\alpha t)$ for all $t \geq 0$.*

These two choices are explored for the Gaussian kernel (Gauss Jordan vs cosh/sinh) in Section 5, and their forms and properties are detailed in Table 1.

Let $\Omega = \mathbb{R}^d$ or a compact convex subset and $S_* = \sup_{x,y\in\Omega} \|x-y\|_2$. Define for $\psi(z) = q(\|z\|_2^2)$,

$$\underline{\lambda}[q] \coloneqq \inf_{0\leq s\leq S_*} \min\big\{2q'(s),\, 2q'(s)+4sq''(s)\big\}, \quad \overline{\Lambda}[q] \coloneqq \sup_{0\leq s\leq S_*} \max\big\{2q'(s), 2q'(s)+4sq''(s)\big\}.$$

Having $\underline{\lambda}[q_\pm] \geq 0$ and finite $\overline{\Lambda}[q_\pm]$, which are related to bounds on the Hessian of $\mathcal{F}^\pm$, allows for a sufficient condition on a DC decomposition of q and therefore of the MMD. Owing to this decomposition we can apply Proposition 8 and all the convergence results of Section 3.

Proposition 9. *Let $q \in C^2([0, S_*])$. Assume there exists $q_\pm \in C^2([0, S_*])$ such that $q = q_+ - q_-$. If $\underline{\lambda}[q_+], \underline{\lambda}[q_-] \geq 0$, then the translation-invariant kernel $k(x,y) = \psi(x-y) = q(\|x-y\|_2^2)$ satisfies Proposition 8 with $\psi_\pm(z) = q_\pm(\|z\|_2^2)$, $\alpha^\pm = \underline{\lambda}[q_\pm]$. Moreover, if $\overline{\Lambda}[q_+], \overline{\Lambda}[q_-] < \infty$, for all $k \geq 0$, the Lipschitz condition in (21) holds for $L = \sqrt{2}\cdot\overline{\Lambda}[q_+] + \overline{\Lambda}[q_-] < \infty$. Hence, if $\underline{\lambda}[q_+] + \underline{\lambda}[q_-] > 0$,*

1. *for $S_* = \infty$,* i.e. $\Omega = \mathbb{R}^d$, *WCCCP leads to an almost stationary measure,* i.e. (22) *holds;*
2. *for $S_* < \infty$,* i.e. Ω *compact and convex subset of $\mathbb{R}^d$, stationarity (22) of WCCCP provided the additional condition that the iterates $(\mu_k)_k$ remain in Ω.*

We refer to Table 1 for decompositions of the Gaussian and (smoothed) Riesz kernels satisfying Proposition 9. For more discussion about kernels satisfying Proposition 9, we refer to Appendix D. In Appendix D.3, we also provide sufficient conditions under which the WCCCP scheme converges locally towards a critical point of the MMD in the compact (Proposition D.15) and non-compact cases (Proposition D.16).

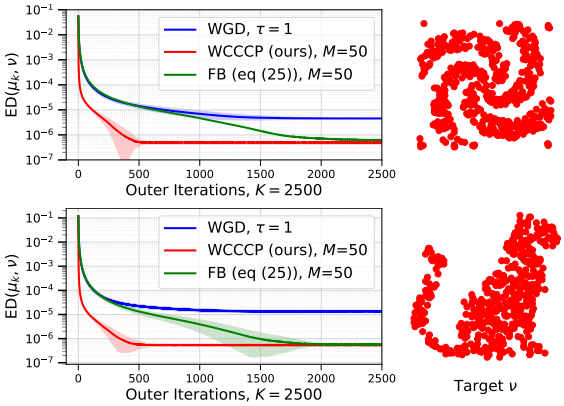


Figure 1: Convergence of WGD, WCCCP and FB to minimize $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{ED}(\mu, \nu)$ with ν as uniform distribution over the spiral and cat shapes.

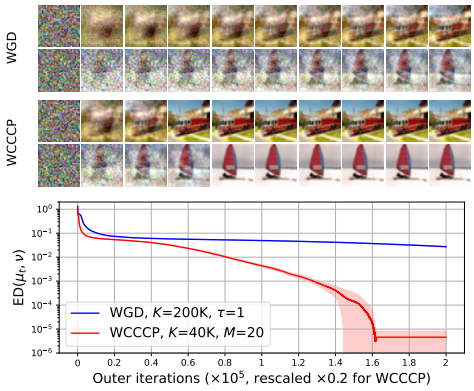


Figure 2: Convergence of WGD and WCCCP to minimize $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{ED}(\mu, \nu)$ with ν sampled from CIFAR10.

5 Numerical Experiments

We now apply the WCCCP algorithm on the Energy Distance and on the MMD with Gaussian kernel. We compare its performance with the Wasserstein Gradient Descent (WGD) and the Wasserstein Proximal Gradient (25) proposed in [55] and denoted Forward-Backward (FB). For the Gaussian kernel, we compare WCCCP with the three different DC decompositions introduced in Section 4, namely the Jordan decomposition, the cosh/sinh decomposition (Lemma 1) and the one of [55] recalled in (29). For more numerical experiments and implementation details, we refer to Appendix E[1].

Energy distance. The Energy distance (ED) [72] corresponds to (28) induced by $\psi(z) = -\|z\|_2$, *i.e.* $k(x, y) = -\|x - y\|_2$ and, for $c(\nu) = -\frac{1}{2}\iint \|x - y\|_2 \, \mathrm{d}\nu(x)\mathrm{d}\nu(y)$,

$$\frac{1}{2}\mathrm{ED}(\mu, \nu) = -\frac{1}{2}\iint \|x - y\|_2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V} \, \mathrm{d}\mu + c(\nu), \ \mathrm{V}(\cdot) = \int \|\cdot - y\|_2 \, \mathrm{d}\nu(y). \quad (32)$$

While not convex, its Wasserstein gradient flow has a good behavior [28] and has demonstrated good results for different machine learning applications [3, 36, 39, 40, 41]. Nonetheless, it can be naturally decomposed as a DC functional using Proposition 8 with $\psi^+ = 0$ and $\psi^-(z) = \|z\|_2$. Thus, we propose to apply the WCCCP algorithm to minimize it.

On Figures 1 and 2, we minimize the Energy distance with respect to ν_n, an empirical distribution of $n = 500$ samples drawn uniformly respectively over spiral or cat shapes, and from CIFAR10 [45]. In both cases, we start from an empirical distribution of n samples of $\mu_0 = \mathcal{N}(0, I_d)$, and use the stepsize $\tau = 1$ for WGD and FB, as WCCCP can be seen as Mirror Descent with $\tau = 1$. This allows comparing the three schemes in the same setting, assuming access to an oracle for WCCCP and FB. We add in Figure E.7 a comparison with different step sizes for WGD and the same computational budget for WCCCP, *i.e.* with the same number of gradient evaluations. Results were averaged over 100 different source and target samples on Figure 1 and 5 on Figure 2. For both FB and WCCCP, the inner optimization of J is performed via gradient descent, possibly with momentum. We display the number of outer iterations, focusing on the behavior in objective values on Figure 1 (FB and WCCCP having $M = 50$ extra inner iterations). On Figure 2 instead, to be fair on the computational time (1h30 for WGD, 1h15 for WCCCP on a Nvidia V100 GPU) for $x \in \mathbb{R}^d$ and $d \approx 3$K, in total 200K iterations were performed for WGD against 40K outer iterations for WCCCP, and we show snapshots every 20K iterations for WGD, and every 4K iterations for WCCCP. Even with this rescaling, the convergence of WCCCP remains much faster.

MMD with Gaussian kernel. Another natural translation-invariant kernel is the Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/(2h)}$ for which $q(s) = e^{-s/(2h)}$. However, minimizing the MMD with this kernel has proven challenging, as its convergence depends a lot on the value of its bandwidth h [6]. On the one hand, we can use the decomposition (29) proposed by Luu et al. [55] with choice $\alpha = \frac{2}{h}$ (see Example D.1). On the other hand, we can use DC decompositions based on decomposing

[1] Code available at `https://github.com/clbonet/Wasserstein_Convex_Concave_Procedure`

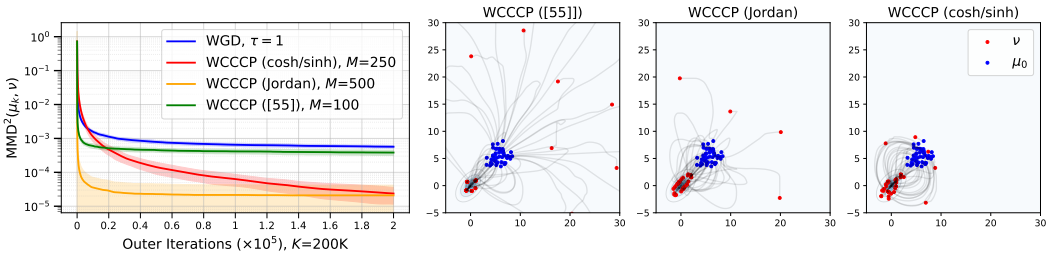


Figure 3: Optimization of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu)$ for ν a Gaussian target and k the Gaussian kernel. (**Left**) Evolution of the squared MMD along the flow. (**Right**) Trajectories of the particles over time. The initial particles are in blue and the final particles in red.

the kernel as introduced in Proposition 8. We do so by choosing q_+, q_- either with the Jordan decomposition (31) or the algebraic one which involves $\cosh$ and $\sinh$ (see Lemma 1). Their properties are detailed in Table 1 and derived in Appendix D.

We compare on Figure 3 the minimization of the squared MMD with Gaussian kernel and bandwidth $h = 10$ between WGD and WCCCP with these three decompositions. We use $n = 500$ particles initially sampled from $\mu_0 = \mathcal{N}(5\mathbb{1}_2, I_2)$ and set the target as an empirical distribution $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ with y_i sampled independently from $\nu = \mathcal{N}(0, \Sigma)$ with $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. The results are averaged over 25 runs with different source and target samples. The inner problems of the WCCCP scheme with $\cosh$ / $\sinh$ are solved using a gradient descent with momentum, with $m = 0.9$ and step size $\tau = 5 \cdot 10^{-4}$ for $M = 250$ iterations, while for the WCCCP scheme with Jordan decomposition, we use a gradient descent with $\tau = 0.1$ and $M = 500$ iterations. In particular, this decomposition is smoother and less prone to numerical instabilities than the $\cosh$ / $\sinh$ one. We observe that WGD and WCCCP with the decomposition (29) get stuck in a local minimum where some samples drift away of the target distribution, while WCCCP with $\cosh$ / $\sinh$ and Jordan decompositions converge much better. We note that the Jordan decomposition can still sometimes be stuck in local minima, which are nonetheless better than WGD. We hypothesize that this is due to the inexact solver used at each iteration of (11). On the other hand, all the samples of WCCCP with $\cosh$ / $\sinh$ appear to eventually converge, instead of being sent away, but at a slower rate. We also compare the results with the FB algorithm in Appendix E, which also showcases results that depend heavily on the choice of DC decomposition.

This experiment exhibits the importance of the choice of the DC decomposition of the target functional, as different DC decompositions induce very different behaviors. We leave for future work providing guidelines on how to choose the DC decomposition to have better convergence guarantees for WCCCP, and include a preliminary experiment in Appendix E.2.

6 Conclusion

In this work, we lifted the convex-concave procedure to the Wasserstein space, and analyzed its convergence in the convex and non-convex settings. Then, we used it to minimize the Maximum Mean Discrepancy, and showed improved performance over the Wasserstein Gradient Descent for the negative distance and Gaussian kernel. Nonetheless, the improved convergence of WCCCP strongly depends on the choice of the DC decomposition of the kernel. Future work will therefore focus on better understanding the impact of different DC decompositions on the performance, with the goal of designing more effective, automatic, and adaptive decomposition strategies [2]. On the theoretical side, taking into account the inexact solvers used for the inner optimization problems would also be a natural avenue for future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable comments. This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011015891R1 made by GENCI. CB's work was supported by the Ecole Polytechnique Foundation as part of its campaign "Servir la Science", and by the French National Research Agency (ANR) through the France 2030 program under the MacLeOD project (ANR-25-PEIA-0005) and the program "Hi! PARIS", grant number ANR-23-IACL-0005.

References

[1] Hadi Abbaszadehpeivasti, Etienne de Klerk, and Moslem Zamani. On the rate of convergence of the difference-of-convex algorithm (DCA). *Journal of Optimization Theory and Applications*, 202(1):475–496, 2024. (Cited on p. 4, 5, 18)

[2] Amir Ali Ahmadi and Georgina Hall. DC Decomposition of Nonconvex Polynomials with Algebraic Techniques. *Mathematical Programming*, 169(1):69–94, 2018. (Cited on p. 10, 18)

[3] Fabian Altekrüger, Johannes Hertrich, and Gabriele Steidl. Neural Wasserstein Gradient Flows for Discrepancies with Riesz Kernels. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 664–690. PMLR, 23–29 Jul 2023. (Cited on p. 9)

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer, 2008. (Cited on p. 1, 2, 3, 7, 18, 20, 35)

[5] Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining Deep Generative Models via Discriminator Gradient Flow. In *International Conference on Learning Representations*, 2021. (Cited on p. 1)

[6] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on p. 1, 2, 3, 7, 9, 28)

[7] Andreas Argyriou, Raphael Hauser, Charles A Micchelli, and Massimiliano Pontil. A DC-Programming Algorithm for Kernel Selection. In *Proceedings of the 23rd international conference on Machine learning*, pages 41–48, 2006. (Cited on p. 2)

[8] Mohammad Askarizadeh, Alireza Morsali, Sadegh Tofigh, and Kim Khoa Nguyen. Convex-Concave Programming: An Effective Alternative for Optimizing Shallow Neural Networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(4):2894–2907, 2024. (Cited on p. 2)

[9] Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. DC-programming for neural network optimizations. *Journal of Global Optimization*, pages 1–17, 2024. (Cited on p. 2)

[10] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. (Cited on p. 18)

[11] Amir Beck and Marc Teboulle. Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters*, 31(3):167–175, 2003. (Cited on p. 3)

[12] Ayoub Belhadji, Daniel Sharp, and Youssef Marzouk. Weighted Quantization Using MMD: From Mean Field to Mean Shift via Gradient Flows. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. (Cited on p. 2)

[13] Ronny Bergmann, Orizon P Ferreira, Elianderson M Santos, and João Carlos O Souza. The Difference of Convex Algorithm on Hadamard Manifolds. *Journal of Optimization Theory and Applications*, 201(1):221–251, 2024. (Cited on p. 2)

[14] Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275(7):1650–1673, 2018. (Cited on p. 1, 18)

[15] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. (Cited on p. 1)

[16] Immanuel M. Bomze and Marco Locatelli. Undominated d.c. Decompositions of Quadratic Functions and Applications to Branch-and-Bound Approaches. *Computational Optimization and Applications*, 28(2):227–245, 2004. (Cited on p. 18)

[17] Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and Preconditioned Gradient Descent in Wasserstein Space. In *Thirty-eight Conference on Neural Information Processing Systems*, 2024. (Cited on p. 1, 3, 5, 7, 18, 19, 33, 36)

[18] Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025. (Cited on p. 1)

[19] Benoît Bonnet. A Pontryagin Maximum Principle in Wasserstein Spaces for Constrained Optimal Control Problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:52, 2019. (Cited on p. 2)

[20] Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013. (Cited on p. 1)

[21] Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4): 4556–4587, 2025. (Cited on p. 2)

[22] Yann Brenier. Polar Factorization and Monotone Rearrangement of Vector-Valued Functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991. (Cited on p. 4)

[23] Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022. (Cited on p. 1)

[24] Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient Flow Drifting: Generative Modeling via Wasserstein Gradient Flows of KDE-Approximated Divergences. *arXiv preprint arXiv:2603.10592*, 2026. (Cited on p. 1)

[25] Giulia Cavagnari, Giuseppe Savaré, and Giacomo Enrico Sodini. A Lagrangian approach to totally dissipative evolutions in Wasserstein spaces. *Journal of Differential Equations*, 470: 114395, 2026. (Cited on p. 3, 18)

[26] El Mahdi Chayti and Martin Jaggi. Stochastic Difference-of-Convex Optimization with Momentum. *arXiv preprint arXiv:2510.17503*, 2025. (Cited on p. 6)

[27] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-Parameterized Models using Optimal Transport. *Advances in neural information processing systems*, 31, 2018. (Cited on p. 1)

[28] Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative Convergence of Wasserstein Gradient Flows of Kernel Mean Discrepancies. *arXiv preprint arXiv:2603.01977*, 2026. (Cited on p. 9)

[29] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative Modeling via Drifting. *arXiv preprint arXiv:2602.04770*, 2026. (Cited on p. 1)

[30] Chao Du, Tianbo Li, Tianyu Pang, Shuicheng Yan, and Min Lin. Nonparametric Generative Modeling with Conditional Sliced-Wasserstein Flows. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8565–8584. PMLR, 23–29 Jul 2023. (Cited on p. 1)

[31] Théo Dumont, Théo Lacombe, and François-Xavier Vialard. Learning Monge maps by lifting and constraining Wasserstein gradient flows. *arXiv preprint arXiv:2603.25182*, 2026. (Cited on p. 3)

[32] Oisin Faust, Hamza Fawzi, and James Saunderson. A Bregman Divergence View on the Difference-of-Convex Algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 3427–3439. PMLR, 2023. (Cited on p. 4, 5, 18, 19, 20)

[33] OP Ferreira, DS Gonçalves, MS Louzeiro, SZ Németh, and J Zhu. A subdifferential characterization via Busemann functions and applications to DC optimization on Hadamard manifolds. *arXiv preprint arXiv:2602.20931*, 2026. (Cited on p. 2)

[34] Bela A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional Bregman Divergence. In *2008 IEEE International Symposium on Information Theory*, pages 1681–1685. IEEE, 2008. (Cited on p. 3)

[35] Yuan Gao, Yuling Jiao, Yang Wang, Yao Wang, Can Yang, and Shunkang Zhang. Deep Generative Learning via Variational Gradient Flow. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2093–2101. PMLR, 09–15 Jun 2019. (Cited on p. 1)

[36] Jonathan Geuter, Clément Bonet, Anna Korba, and David Alvarez-Melis. DDEQs: Distributional Deep Equilibrium Models through Wasserstein Gradient Flows. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. (Cited on p. 9)

[37] Egor Gladin, Pavel Dvurechensky, Alexander Mielke, and Jia-Jie Zhu. Interaction-Force Transport Gradient Flows. *Advances in Neural Information Processing Systems*, 37:14484–14508, 2024. (Cited on p. 2, 28)

[38] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-sample Test. *The journal of machine learning research*, 13(1):723–773, 2012. (Cited on p. 7)

[39] Paul Hagemann, Johannes Hertrich, Fabian Altekrüger, Robert Beinert, Jannis Chemseddine, and Gabriele Steidl. Posterior Sampling Based on Gradient Flows of the MMD with Negative Distance Kernel. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on p. 9)

[40] Johannes Hertrich, Manuel Gräf, Robert Beinert, and Gabriele Steidl. Wasserstein Steepest Descent Flows of Discrepancies with Riesz Kernels. *Journal of Mathematical Analysis and Applications*, 531(1):127829, 2024. (Cited on p. 1, 9)

[41] Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative Sliced MMD Flows with Riesz Kernels. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on p. 1, 7, 9, 25, 28)

[42] Jean-Baptiste Hiriart-Urruty. Generalized Differentiability, Duality and Optimization for Problems Dealing with Differences of Convex Functions. In J. Ponstein, editor, *Convexity and Duality in Optimization*, volume 256 of *Lecture Notes in Economics and Mathematical Systems*, pages 37–70. Springer, Berlin, 1985. (Cited on p. 2, 3)

[43] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998. (Cited on p. 1)

[44] Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein Discrepancy Descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR, 2021. (Cited on p. 7)

[45] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. 2009. (Cited on p. 9)

[46] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational Inference via Wasserstein Gradient Flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022. (Cited on p. 1)

[47] Gert Lanckriet and Bharath K Sriperumbudur. On the Convergence of the Concave-Convex Procedure. *Advances in neural information processing systems*, 22, 2009. (Cited on p. 4)

[48] Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-Order Conditions for Optimization in the Wasserstein Space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025. (Cited on p. 2, 3)

[49] Hoai An Le Thi and Tao Pham Dinh. DC programming and DCA: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018. (Cited on p. 2, 4)

[50] Linshan Liu, Mateusz B Majka, and Łukasz Szpruch. Polyak–Łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics & Optimization*, 87(3):48, 2023. (Cited on p. 1, 18)

[51] Song Liu, Jiahao Yu, Jack Simons, Mingxuan Yi, and Mark Beaumont. Minimizing f-Divergences by Interpolating Velocity Fields. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32308–32331. PMLR, 21–27 Jul 2024. (Cited on p. 1)

[52] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions. In *International Conference on machine learning*, pages 4104–4113. PMLR, 2019. (Cited on p. 1)

[53] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. (Cited on p. 3, 18)

[54] Hoang Phuc Hau Luu and Zhongjian Wang. DC-LA: Difference-of-Convex Langevin Algorithm. In *Forty-third International Conference on Machine Learning*, 2026. (Cited on p. 2)

[55] Hoang Phuc Hau Luu, Hanlin Yu, Bernardo Williams, Petrus Mikkola, Marcelo Hartmann, Kai Puolamäki, and Arto Klami. Non-geodesically-convex optimization in the Wasserstein space. *Advances in Neural Information Processing Systems*, 37:16772–16809, 2024. (Cited on p. 1, 2, 4, 6, 7, 9, 20, 21, 25, 29)

[56] Hoomaan Maskan, Yikun Hou, Suvrit Sra, and Alp Yurtsever. Revisiting Frank-Wolfe for Structured Nonconvex Optimization. *Advances in Neural Information Processing Systems*, 38: 43651–43678, 2025. (Cited on p. 4)

[57] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-Layer Neural Networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. (Cited on p. 1)

[58] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and TrendsÂ® in Machine Learning*, 10(1-2):1–141, 2017. (Cited on p. 7)

[59] Atsushi Nitanda and Taiji Suzuki. Stochastic Difference of Convex Algorithm and its Application to Training Deep Boltzmann Machines. In *Artificial intelligence and statistics*, pages 470–478. PMLR, 2017. (Cited on p. 6)

[60] Yi-Shuai Niu. Continuous-Time Dynamics of the Difference-of-Convex Algorithm. *arXiv preprint arXiv:2604.06926*, 2026. (Cited on p. 4, 18)

[61] Yi-Shuai Niu, Hoai An Le Thi, and Dinh Tao Pham. On Difference-of-SOS and Difference-of-Convex-SOS Decompositions for Polynomials. *SIAM Journal on Optimization*, 34(2): 1852–1878, 2024. (Cited on p. 18)

[62] Konstantinos Oikonomidis, Emanuel Laude, and Panagiotis Patrinos. Forward-backward splitting under the light of generalized convexity. *arXiv preprint arXiv:2503.18098*, 2025. (Cited on p. 4, 18, 19, 20)

[63] Guy Parker. Some convexity criteria for differentiable functions on the 2-Wasserstein space. *Bulletin of the London Mathematical Society*, 56(5):1839–1858, 2024. (Cited on p. 3, 18)

[64] Mikhail Persiianov, Jiawei Chen, Petr Mokrov, Alexander Tyurin, Evgeny Burnaev, and Alexander Korotin. Learning of Population Dynamics: Inverse Optimization Meets JKO Scheme. In *The Fourteenth International Conference on Learning Representations*, 2026. (Cited on p. 1)

[65] Marguerite Petit-Talamon, Marc Lambert, and Anna Korba. Variational Inference with Mixtures of Isotropic Gaussians. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (Cited on p. 1)

[66] David Pfau, Ian Davies, Diana L Borsa, João Guilherme Madeira Araújo, Brendan Daniel Tracey, and Hado van Hasselt. Wasserstein Policy Optimization. In *Forty-second International Conference on Machine Learning*, 2025. (Cited on p. 1)

[67] Tao Pham Dinh and Hoai An Le Thi. Recent advances in DC programming and DCA. *Transactions on computational intelligence XIII*, pages 1–37, 2014. (Cited on p. 2, 4)

[68] Teodor Rotaru, Panagiotis Patrinos, and François Glineur. Tight Analysis of Difference-of-Convex Algorithm (DCA) Improves Convergence Rates for Proximal Gradient Descent. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. (Cited on p. 4)

[69] Nicolaj Rux, Michael Quellmalz, and Gabriele Steidl. Smoothed distance kernels for MMDs and applications in Wasserstein gradient flows. *Advances in Computational Mathematics*, 52 (2):24, 2026. (Cited on p. 25)

[70] Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein Proximal Gradient Algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020. (Cited on p. 1, 2, 6)

[71] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015. (Cited on p. 1)

[72] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013. (Cited on p. 9, 25)

[73] Louis Sharrock, Lester Mackey, and Christopher Nemeth. Learning Rate Free Sampling in Constrained Domains. *Advances in Neural Information Processing Systems*, 36:65380–65415, 2023. (Cited on p. 1)

[74] JC d O Souza and Paulo Roberto Oliveira. A proximal point algorithm for DC fuctions on Hadamard manifolds. *Journal of Global Optimization*, 63(4):797–810, 2015. (Cited on p. 2)

[75] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7), 2011. (Cited on p. 7)

[76] Wen-yu Sun, Raimundo JB Sampaio, and MAB Candido. Proximal point algorithm for minimization of DC function. *Journal of computational Mathematics*, pages 451–462, 2003. (Cited on p. 6)

[77] Ken'ichiro Tanaka. Accelerated gradient descent method for functionals of probability measures by new convexity and smoothness based on transport maps. *arXiv preprint arXiv:2305.05127*, 2023. (Cited on p. 3, 18)

[78] Pham Dinh Tao and Hoai An Le Thi. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997. (Cited on p. 4)

[79] Pham Dinh Tao et al. New and efficient DCA based algorithms for minimum sum-of-squares clustering. *Pattern Recognition*, 47(1):388–401, 2014. (Cited on p. 2)

[80] Antonio Terpin, Nicolas Lanzetti, Martín Gadea, and Florian Dorfler. Learning diffusion at lightspeed. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (Cited on p. 1)

[81] Gauthier Thurin, Claire Boyer, and Kimia Nadjahi. Convergence Rates for Distribution Matching with Sliced Optimal Transport. In Steve Hanneke and Tor Lattimore, editors, *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 6156–6196. PMLR, 2026. (Cited on p. 2)

[82] Quang Huy Tran, Hicham Janati, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Factored couplings in multi-marginal optimal transport via difference of convex programming. *arXiv preprint arXiv:2110.00629*, 2021. (Cited on p. 2)

[83] Erkan Turan and Maks Ovsjanikov. Generative Drifting is Secretly Score Matching: a Spectral and Variational Perspective. *arXiv preprint arXiv:2603.09936*, 2026. (Cited on p. 1)

[84] Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009. (Cited on p. 1, 2)

[85] Xuan Thanh Vo, Hoai An Le Thi, Tao Pham Dinh, and Thi Bich Thuy Nguyen. DC Programming and DCA for Dictionary Learning. In *Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part I*, pages 295–304. Springer, 2015. (Cited on p. 2)

[86] Melanie Weber and Suvrit Sra. Global optimality for Euclidean CCCP under Riemannian convexity. In *International Conference on Machine Learning*, pages 36790–36803. PMLR, 2023. (Cited on p. 2)

[87] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on learning theory*, pages 2093–3027. PMLR, 2018. (Cited on p. 1)

[88] Yewei Xu and Qin Li. Random Coordinate Descent on the Wasserstein Space of Probability Measures. *arXiv preprint arXiv:2604.01606*, 2026. (Cited on p. 1)

[89] Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers with Non-Asymptotic Convergence. In *International conference on machine learning*, pages 6942–6951. PMLR, 2019. (Cited on p. 6)

[90] Alan L. Yuille and Anand Rangarajan. The Concave-Convex Procedure (CCCP). *Advances in neural information processing systems*, 14, 2001. (Cited on p. 2, 3, 4)

[91] Alp Yurtsever and Suvrit Sra. CCCP is Frank-Wolfe in disguise. *Advances in Neural Information Processing Systems*, 35:35352–35364, 2022. (Cited on p. 4, 5)

[92] Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy Optimization as Wasserstein Gradient Flows. In *International Conference on machine learning*, pages 5737–5746. PMLR, 2018. (Cited on p. 1)

[93] Shuailong Zhu and Xiaohui Chen. Convergence Analysis of the Wasserstein Proximal Algorithm beyond Geodesic Convexity. *arXiv preprint arXiv:2501.14993*, 2025. (Cited on p. 1)

[94] Zico Zolter, David Duvenaud, and Matt Johnson. Deep Implicit Layers - Neural ODEs, Deep Equilibirum Models, and Beyond. *Neurips 2020 Tutorial*, 2020. (Cited on p. 6)

Appendix

The appendix is organized as follows. In Appendix A, we discuss some limitations of the paper. In Appendix B, we detail the theoretical analysis of the WCCCP algorithm in the convex case, leveraging the mirror and Bregman proximal descent formulations. In Appendix C, we discuss the derivations of Polyak-Łojaziewicz inequalities for DC functionals. In Appendix D, we provide the full theoretical analysis of DC decompositions of the MMD. In Appendix E, we include more details about the numerical experiments. Finally, in Appendix F, we state all the proofs.

Appendix Contents

A Limitations **18**

B Theoretical Analysis of WCCCP in the Convex Case **18**

C Enquiry on Polyak-Łojaziewicz Inequality for DC Functionals **19**

D DC Theory for MMD **20**

D.1 Strong Convexity of $\mathcal{F}^-$ and $\mathcal{F}^+$ for Radial Kernels 21

D.2 Lipschitz Continuity of $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ and Stationarity 23

D.3 Critical points and Local Convergence 24

E Numerical Applications on MMD **25**

E.1 Energy Distance 25

E.2 MMD with Gaussian Kernel 27

F Proofs **30**

F.1 Proof of Proposition 1 30

F.2 Proof of Proposition 2 31

F.3 Proof of Proposition 3 31

F.4 Proof of Proposition 4 31

F.5 Proof of Proposition 5 31

F.6 Proof of Proposition 6 32

F.7 Proof of Proposition 7 33

F.8 Proof of Proposition 8 34

F.9 Proof of Lemma 1 35

F.10 Proof of Proposition 9 35

F.11 Proof of Proposition B.11 35

F.12 Proof of Lemma D.2 36

F.13 Proof of Lemma D.4 37

F.14 Proof of Proposition D.13 37

F.15 Proof of Lemma D.6 38

F.16 Proof of Lemma D.7 38

F.17 Proof of Proposition D.14 39

F.18 Proof of Theorem D.1 40
F.19 Proof of Proposition D.15 41
F.20 Proof of Proposition D.16 41

A Limitations

Our empirical findings show that the improved convergence of WCCCP on MMD strongly depends on the choice of the DC decomposition of the kernel k. Our current strategy is restricted to a few simple decompositions of the kernels, without taking into account the geometry of the optimization problem. Hence, future work could focus on developing automatic and adaptive ways to find DC decompositions of the kernels better suited to the problem of minimizing the MMD, *e.g.* based on algebraic decompositions and polynomial DC decompositions minimizing a suitable objective [2, 16, 61].

This work also focused on DC decompositions of MMD, which can be written as a sum of simple functionals, which are potential and interaction energies. For MMD with the ubiquitous radial kernels, we were thus able to derive DC decompositions based on decompositions of functions on $\mathbb{R} \to \mathbb{R}$. Future work could focus on the problem of deriving DC decompositions of more complex functionals.

On the theoretical part, as highlighted in Section 1, there are versions of Polyak-Łojaziewicz (PL) inequalities for both the Wasserstein space [14, 50] and DC functions [1, 32, 60, 62]. So far our framework does not include them despite preliminary research. We discuss reasons for this in Section C.

We also did not take into account that we actually solve each inner scheme approximately, which leads to a gap between theory and practice.

B Theoretical Analysis of WCCCP in the Convex Case

We focus in this section on the case where $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ is also convex along a curve of interest which we will detail. For this, we will leverage Proposition 1, where we showed that (11) is equivalent to both a mirror descent and a Bregman proximal descent in the Wasserstein space. Consequently, on the one hand, we will inherit the results from [17] for mirror descent, and on the other hand we will derive novel convergence results for Bregman proximal descent.

Relative convexity and smoothness. First, we need to introduce the notions of relative convexity and smoothness. Let $\alpha, \beta \geq 0$. Following [17], we say that $\mathcal{F}$ is α-convex relative to $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ along $t \mapsto \big((1-t)\mathrm{T} + t\mathrm{S}\big)_{\#}\mu$ if $\mathrm{D}^{\mu}_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) \geq \alpha \mathrm{D}^{\mu}_{\mathcal{G}}(\mathrm{T}, \mathrm{S})$. Equivalently, we have that $\mathcal{F} - \alpha\mathcal{G}$ is convex along this curve, and

$$\forall t \in [0,1],\ \mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mathrm{T}_{\#}\mu) + t\mathcal{F}(\mathrm{S}_{\#}\mu) - \alpha t(1-t)\mathrm{D}^{\mu}_{\mathcal{G}}(\mathrm{T}, \mathrm{S}). \tag{33}$$

Likewise, $\mathcal{F}$ is β-smooth relative to $\mathcal{G}$ along this curve if $\mathrm{D}^{\mu}_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) \leq \beta \mathrm{D}^{\mu}_{\mathcal{G}}(\mathrm{T}, \mathrm{S})$. These notions enable lifting the notion of relative convexity and smoothness [10, 53] to $\mathcal{P}_2(\mathbb{R}^d)$ for differentiable functionals.

If the convexity holds for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathcal{G}(\mu) = \int \frac{1}{2}\|\cdot\|_2^2 \mathrm{d}\mu$, then we say that $\mathcal{F}$ is α-totally convex [25, 63, 77]. Following [4], if convexity only holds for $\mathrm{T} = \mathrm{Id}$ and S gradient of convex functions, it coincides with strong convexity along geodesics [4]. Convexity along generalized geodesics corresponds to both T and S being the gradients of convex functions. Examples of totally convex functionals include potential and interaction energies, provided V, W are convex, lower semi-continuous and have a negative part with quadratic growth [4, Section 9.3]. Note also that all three notions of convexity are equivalent for continuous functionals (for $d \geq 2$) [25, 63].

Bregman Wasserstein distance. Similarly to [17], let us introduce the Bregman Optimal Transport problem W_{ϕ} associated with a Wasserstein differentiable functional $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, defined for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\mathrm{W}_{\phi}(\nu, \mu) \coloneqq \inf_{\gamma \in \Pi(\nu,\mu)} \phi(\nu) - \phi(\mu) - \int \langle \nabla_{\mathrm{W}_2}\phi(\mu)(y), x - y \rangle \, \mathrm{d}\gamma(x, y). \tag{34}$$

By [17, Proposition 15], if $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, then there exists $\mathrm{T} \in L^2(\mu)$ such that $\mathrm{W}_\phi(\nu,\mu) = \mathrm{D}_\phi^\mu(\mathrm{T},\mathrm{Id})$, *i.e.* this problem admits an optimal transport map.

Convergence results. We now provide a first convergence result for $\mathcal{F}$ smooth and strongly convex relative to $\mathcal{F}^+$, relying on the mirror descent formulation (13) and the results of [17]. More precisely, $\mathcal{F}$ needs to be smooth along iterates and convex along what would be the analog of geodesics on the space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_{\mathcal{F}^+})$.

Proposition B.10. *Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional admitting a DC decomposition $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ with both $\mathcal{F}^+, \mathcal{F}^-$ Wasserstein differentiable on $\mathcal{P}_2(\mathbb{R}^d)$. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $0 \leq \alpha \leq \beta \leq 1$ and $(\mathrm{T}_k)_{k\geq 1}$, $(\mu_k)_{k\geq 0}$ iterates of (11). Assume that for all $k \geq 0$, $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, and denote $\mathrm{T}^{\mu_k,\nu} = \mathrm{argmin}_{\mathrm{T}_\# \mu_k = \nu} \mathrm{D}_{\mathcal{F}^+}^{\mu_k}(\mathrm{T},\mathrm{Id})$, which exists as $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Furthermore, assume that $\mathcal{F}$ is β-smooth along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_\# \mu_k$ and $\mathcal{F}$ is α-convex relative to $\mathcal{F}^+$ along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}^{\mu_k,\nu}\big)_\# \mu_k$. Then, for all $k \geq 1$,*

$$\mathcal{F}(\mu_k) - \mathcal{F}(\nu) \leq \frac{\alpha}{(1-\alpha)^{-k} - 1} \mathrm{W}_{\mathcal{F}^+}(\nu,\mu_0) \leq \frac{1-\alpha}{k} \mathrm{W}_{\mathcal{F}^+}(\nu,\mu_0). \tag{35}$$

Moreover, if $\alpha > 0$, taking $\nu = \mu^*$ the minimizer of $\mathcal{F}$, we obtain a linear rate, i.e. *for all $k \geq 0$,*

$$\mathrm{W}_{\mathcal{F}^+}(\mu^*,\mu_k) \leq (1-\alpha)^k \mathrm{W}_{\mathcal{F}^+}(\mu^*,\mu_0). \tag{36}$$

Proof. The assumptions imply that we can use [17, Proposition 4] for the mirror descent formulation (13). □

Assuming analog conditions as for the convergence of the JKO scheme adapted to the Bregman Proximal Gradient Descent (12), we obtain the following linear rate convergence result.

Proposition B.11. *Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ a functional admitting a DC decomposition $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ with both $\mathcal{F}^+, \mathcal{F}^-$ Wasserstein differentiable on $\mathcal{P}_2(\mathbb{R}^d)$. Let $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ be the minimizer of $\mathcal{F}$, $\alpha \geq 0$ and $(\mathrm{T}_k)_{k\geq 1}$, $(\mu_k)_{k\geq 0}$ iterates of (11). Assume that for all $k \geq 0$, $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, and denote $\mathrm{T}_k^* = \mathrm{argmin}_{\mathrm{T}_\# \mu_k = \mu^*} \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T},\mathrm{Id})$, which exists as $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Furthermore, assume that $\mathcal{F}$ is α-convex relative to $\mathcal{F}^-$ along $t \mapsto \big((1-t)\mathrm{T}_k^* + t\mathrm{T}_{k+1}\big)_\# \mu_k$. Then, for all $k \geq 0$,*

$$\mathrm{W}_{\mathcal{F}^-}(\mu^*,\mu_k) \leq \left(\frac{1}{1+\alpha}\right)^k \mathrm{W}_{\mathcal{F}^-}(\mu^*,\mu_0), \quad \textit{and} \quad \mathcal{F}(\mu_{k+1}) - \mathcal{F}(\mu^*) \leq \left(\frac{1}{1+\alpha}\right)^k \mathrm{W}_{\mathcal{F}^-}(\mu^*,\mu_0). \tag{37}$$

Proof. See Section F.11. □

The convergence results are given in the geometry induced by the Bregman divergence with potential given by $\mathcal{F}^-$.

C Enquiry on Polyak-Łojaziewicz Inequality for DC Functionals

An analog of the DC PL inequality of [62, Definition 6.1], writing it only using the Bregman divergences of $\mathcal{F}^+$ and $\mathcal{F}^-$, reads as follows

Definition C.1. *We say that $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ satisfies the Wasserstein DC PL inequality analog to [62, Definition 6.1], if there exists $\eta_1 \geq 0$, $\eta_2 \geq 0$ such that $\eta_1 + \eta_2 > 0$ and for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\eta_1\big(\mathcal{F}(\mu) - \inf \mathcal{F}\big) + \eta_2\big(\mathcal{F}(\bar{\mathrm{T}}_\# \mu) - \inf \mathcal{F}\big) \leq \mathrm{D}_{\mathcal{F}^-}^\mu(\bar{\mathrm{T}},\mathrm{Id}) + \mathrm{D}_{\mathcal{F}^+}^\mu(\mathrm{Id},\bar{\mathrm{T}}) \tag{38}$$

where $\bar{\mathrm{T}}$ solves WCCCP, e.g. in its version (12), giving $\bar{\mathrm{T}} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu)} \mathrm{D}_{\mathcal{F}^-}^\mu(\mathrm{T},\mathrm{Id}) + \mathcal{F}(\mathrm{T}_\# \mu)$.

Under this ideal condition, we show that we have the following linear convergence rate, analogously to [32, Lemma 1] and [62, Theorem 6.2].

Proposition C.12. *Assume $\mathcal{F}$ satisfies the Wasserstein DC-PL inequalities (38). Then, for all $k \geq 0$,*

$$\mathcal{F}(\mu_k) - \inf \mathcal{F} \leq \left(\frac{1-\eta_1}{1+\eta_2}\right)^k \big(\mathcal{F}(\mu_0) - \inf \mathcal{F}\big). \tag{39}$$

If $\eta_1 \geq 1$, then we can set $\eta_1 = 1$ and we have convergence in one step.

Proof. Since $\mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k$, (38) gives

$$\begin{aligned}\eta_1\big(\mathcal{F}(\mu_k) - \inf \mathcal{F}\big) + \eta_2\big(\mathcal{F}(\mu_{k+1}) - \inf \mathcal{F}\big) &\leq \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_{k+1}, \mathrm{Id}) + \mathrm{D}_{\mathcal{F}^+}^{\mu_k}(\mathrm{Id}, \mathrm{T}_{k+1}) \\ &\overset{(17)}{=} \mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}).\end{aligned}$$

Rearranging, we obtain

$$\mathcal{F}(\mu_{k+1}) - \inf \mathcal{F} \leq \frac{1-\eta_1}{1+\eta_2}\big(\mathcal{F}(\mu_k) - \inf \mathcal{F}\big) \leq \left(\frac{1-\eta_1}{1+\eta_2}\right)^{k+1} \big(\mathcal{F}(\mu_0) - \inf \mathcal{F}\big). \tag{40}$$

□

Although [62, Lemma 6.3] in the Euclidean case shows that [32, Definition 1] implies their [62, Definition 6.1], the latter is difficult to check directly for a given functional and is more of a target inequality to establish. We thus want to argue that [32, Definition 1] is more promising to stand as DC PL inequality and has received a more developed discussion of when it holds, *e.g.* under strong convexity. However this definition rests upon Fenchel duality on $\mathbb{R}^d$, for which there is no publicly available counterpart on the Wasserstein space at the time of this submission. Consequently it is unclear for now whether Definition C.1 holds for functionals with strongly convex DC decompositions.

While [55, Theorems 4 and 5] do rest upon a Łojaziewicz inequality, the latter is unrelated to Bregman divergences. Luu et al. [55] achieve instead their linear rates using the extra Wasserstein regularization appearing in (26). While we discussed the similarity of the two approaches in Section 3.4, one cannot use their theory without the Wasserstein regularization, as otherwise their constant becomes vacuous.

D DC Theory for MMD

Let k be a translation-invariant kernel of the form $k(x,y) = \psi(x-y)$ for all $x, y \in \mathbb{R}^d$, with ψ admitting a DC decomposition $\psi = \psi^+ - \psi^-$. Then by Proposition 8, the squared MMD (28) admits the DC decomposition $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ where for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\begin{cases} \mathcal{F}^+(\mu) = \frac{1}{2}\iint \psi^+(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V}^-\mathrm{d}\mu + c(\nu), & \mathrm{V}^-(\cdot) = \int \psi^-(\cdot - y)\,\mathrm{d}\nu(y), \\ \mathcal{F}^-(\mu) = \frac{1}{2}\iint \psi^-(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int \mathrm{V}^+\mathrm{d}\mu, & \mathrm{V}^+(\cdot) = \int \psi^+(\cdot - y)\,\mathrm{d}\nu(y), \end{cases} \tag{41}$$

and $\mathcal{F}^+, \mathcal{F}^-$ are both totally convex. Moreover, if ψ^+ and ψ^- are respectively $\alpha^+ \geq 0$ and $\alpha^- \geq 0$ strongly convex, then $\mathcal{F}^+$ is α^--totally convex and $\mathcal{F}^-$ is α^+-totally convex as they inherit the strong convexity of the potential, while the interaction terms are only convex, see [4, Section 9.3].

We will now focus on radial kernels. Let $\Omega \subset \mathbb{R}^d$ be a nonempty compact convex set. Consider the geodesically convex set $\mathcal{P}_2(\Omega)$. Let

$$S_* \coloneqq \sup_{x,y\in\Omega} \|x-y\|_2. \tag{42}$$

We consider in what follows:

$$\psi^{\pm}(z) \coloneqq q_{\pm}(\|z\|_2^2), \quad z \in \Omega - \Omega \coloneqq \{x - y \,|\, x, y \in \Omega\}, \tag{43}$$

where $q_{\pm} \in C^2([0, S_*])$. The MMD functional $\mathcal{F}$ is therefore defined as a function of $q_{\pm}$.

D.1 Strong Convexity of $\mathcal{F}^-$ and $\mathcal{F}^+$ for Radial Kernels

We first study the Hessian of radial functions, in order to derive conditions for $\psi_\pm$ to be (strongly) convex and to be able to apply Proposition 8.

Let $q \in C^2([0, S_*])$ and define:

$$\underline{\lambda}[q] := \inf_{0\le s\le S_*} \min\big\{2q'(s),\, 2q'(s)+4sq''(s)\big\}, \quad \overline{\Lambda}[q] := \sup_{0\le s\le S_*} \max\big\{2q'(s),\, 2q'(s)+4sq''(s)\big\}. \tag{44}$$

We use the shorthands $\lambda_\pm := \underline{\lambda}[q_\pm]$ and $\Lambda_\pm := \overline{\Lambda}[q_\pm]$. First, we compute the Hessian of $\psi(z) = q(\|z\|_2^2)$ and bound its eigenvalues using $\underline{\lambda}[q]$ and $\overline{\Lambda}[q]$.

Lemma D.2 (Radial Hessian Bounds)**.** *Let $\psi(z) = q(\|z\|_2^2)$ on $z \in \Omega - \Omega$, we have*

$$\nabla^2\psi(z) = 2q'(\|z\|_2^2)I_d + 4q''(\|z\|_2^2)zz^\top. \tag{45}$$

Hence, we have the following bounds on $\nabla^2\psi$:

$$\underline{\lambda}[q]I_d \preceq \nabla^2\psi(z) \preceq \overline{\Lambda}[q]I_d, \quad z \in \Omega - \Omega. \tag{46}$$

If $\underline{\lambda}(q) \ge 0$, then ψ is convex on $\Omega - \Omega$.

Proof. See Section F.12. □

Now, building on the previous Lemma, we deduce sufficient conditions under which ψ is convex.

Lemma D.3 (Sufficient Condition for Convexity)**.** *Assume $q \in C^2([0, S_*])$ such that $q'(s) \ge 0$ and $q''(s) \ge 0$ for all $s \in [0, S_*]$, then $z \mapsto \psi(z) = q(\|z\|_2^2)$ is convex in $\Omega - \Omega$, and $\overline{\Lambda}(q) \ge \underline{\lambda}(q) \ge 0$.*

Proof of Lemma D.3. We always have $\overline{\Lambda}(q) \ge \underline{\lambda}(q)$ and under these conditions it is easy to see that $\underline{\lambda}(q) \ge 0$. □

This lemma provides sufficient conditions to get a DC decomposition of the form (41) by Proposition 8.

Corollary D.1 (Strong Convexity of DC decomposition in MMD)**.** *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a radial kernel,* i.e. *such that $k(x, y) = q(\|x - y\|_2^2)$, where q admits a decomposition $q = q_+ - q_-$, with q_+ and q_- satisfying the conditions of Lemma D.3. Then, considering $\mathcal{F}^+, \mathcal{F}^-$ as defined in (41), we have on $\mathcal{P}_2(\Omega)$, that*

1. *$\mathcal{F}^+$ is α^+-strongly totally convex with $\alpha^+ = \lambda_-$;*
2. *$\mathcal{F}^-$ is α^--strongly totally convex with $\alpha^- = \lambda_+$.*

Proof. We apply Proposition 8. □

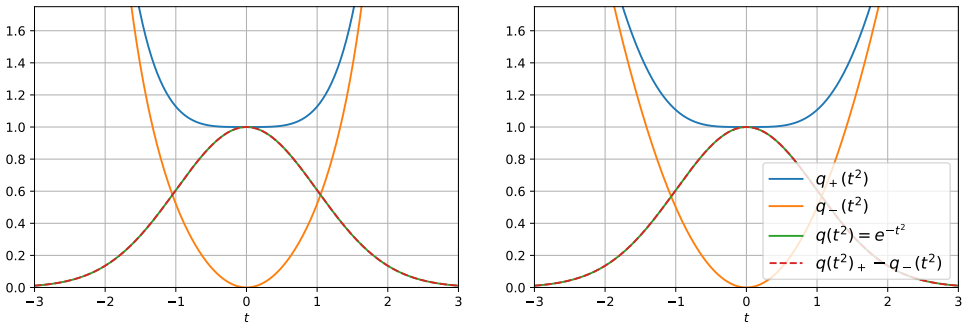
Building on these two lemmas, we show that some usual kernels admit such decompositions. We begin by discussing several DC decompositions for the Gaussian kernel that we can use for MMD. The first one is based on a remark from [55], and the other ones are based on observing that it is a radial kernel with $q(t) = e^{-t/(2h)}$. Hence, we simply need to find a DC decomposition of $q = q_+ - q_-$ which satisfies the assumptions in Lemma D.3.

Gaussian kernel. Recall that the Gaussian kernel is a radial kernel with $q(t) = e^{-\alpha t}$ with $\alpha > 0$ (often taken as $\alpha = 1/(2h)$, h being the bandwidth).

Luu et al. [55] observed in their Appendix A.2 that for k differentiable and L-smooth, *i.e.* satisfying $\|\nabla k(x, y) - \nabla k(x', y')\|_2^2 \le L^2\big(\|x - x'\|_2^2 + \|y - y'\|_2^2\big)$ for all $x, x', y, y' \in \mathbb{R}^d$, a DC decomposition of (28) is given by

$$\mathcal{F}^+(\mu) = \alpha \int \|\cdot\|_2^2 \,\mathrm{d}\mu + \frac{1}{2}\iint k(x, y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + c(\nu), \quad \mathcal{F}^-(\mu) = \int \big(\alpha\|x\|_2^2 - \mathrm{V}(x)\big)\,\mathrm{d}\mu(x) \tag{47}$$

for any $\alpha \ge L$. Actually we can specify in the next example the best choice of α for the Gaussian kernel.



(a) $q_+(t) = \cosh(t),\ q_-(t) = \sinh(t)$ (b) $q_+(t) = e^{-\alpha t} + \alpha t,\ q_-(t) = \alpha t$

Figure D.4: Plot of two possible DC decompositions of $t \mapsto q(t^2) = e^{-t^2}$ that we use to get the DC decomposition of MMD with Gaussian kernel. On the left, we show the $\cosh / \sinh$ decomposition, and on the right, we show the decomposition based on the sign of the Hessian.

Example D.1. *For the Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/(2h)}$, let us find the best constant L. We have $\nabla_x k(x, y) = -\frac{1}{h} e^{-\|x-y\|_2^2/(2h)}(x - y)$ and $\nabla_x^2 k(x, y) = \frac{1}{h^2} e^{-\|x-y\|_2^2/(2h)}(x - y)(x - y)^\top - \frac{1}{h} e^{-\|x-y\|_2^2/(2h)} I_d = \frac{1}{h} e^{-\|x-y\|_2^2/(2h)} \big(\frac{1}{h}(x - y)(x - y)^\top - I_d\big)$. Its eigenvalues are $\lambda_0 = -\frac{1}{h} e^{-\|x-y\|_2^2/(2h)}$ and $\lambda_1 = (\frac{1}{h}\|x - y\|_2^2 - 1)\frac{1}{h} e^{-\|x-y\|_2^2/(2h)}$. Let $t = \frac{1}{h}\|x - y\|_2^2$, then the operator norm of the Hessian is $\|\nabla_x^2 k(x, y)\|_{\text{op}} = \frac{1}{h} e^{-t/2} \max(1, |t - 1|)$. The maximum in t is obtained for $t = 0$, thus $\|\nabla^2 k(x, y)\|_{\text{op}} \leq \frac{2}{h}$. Thus, we can use $\alpha = \frac{2}{h}$ in (47).*

Another natural DC decomposition is based on the observation that $e^{-t} = \cosh(t) - \sinh(t)$. We show in the next lemma that this gives a valid decomposition for Proposition 8. We plot the decomposition on Figure D.4a.

Lemma D.4 (DC decomposition of the Gaussian Kernel based on $\cosh / \sinh$)**.** *For $x, y \in \Omega$ and $\alpha > 0$, a Gaussian kernel $k(x, y) = \exp(-\alpha\|x - y\|_2^2)$ admits a DC decomposition as follows: $k(x, y) = \psi^+(x - y) - \psi^-(x - y)$, where $\psi^\pm(z) = q_\pm(\|z\|_2^2)$ with*

$$q_+(s) = \cosh(\alpha s), \quad q_-(s) = \sinh(\alpha s). \tag{48}$$

For $s \in [0, S_*]$ the maximum and minimum eigenvalues of the Hessian satisfy

$$\lambda_+ = 0, \quad \Lambda_+ = 2\alpha \sinh(\alpha S_*) + 4\alpha^2 S_* \cosh(\alpha S_*), \tag{49}$$

and

$$\lambda_- = 2\alpha, \quad \Lambda_- = 2\alpha \cosh(\alpha S_*) + 4\alpha^2 S_* \sinh(\alpha S_*). \tag{50}$$

Proof. See Section F.13. □

Applying Proposition 8, we can deduce that $\mathcal{F}^+$ is 2α-totally convex, with $\alpha > 0$, while $\mathcal{F}^-$ is only totally convex. Hence, we can apply Proposition 5 and obtain convergence of WCCCP towards a stationary point at a sublinear rate, provided the iterates remain in Ω.

Next, we consider $q(t) = e^{-\alpha t}$ directly. Its derivative is $q'(t) = -\alpha e^{-\alpha t}$ and second derivative $q''(t) = \alpha^2 e^{-\alpha t}$. We note that it is convex but it does not satisfy Lemma D.3 as $q' < 0$. However, adding a linear term large enough, we can obtain a DC decomposition of q satisfying the assumptions of Lemma D.3. For this, we apply (31) with $A = \alpha$.

Lemma D.5. *Let $q : t \mapsto e^{-\alpha t}$ for $\alpha > 0$. Then the DC decomposition written in (31) is $q_+(t) = e^{-\alpha t} + \alpha t$ and $q_-(t) = \alpha t$ for all $t \in \mathbb{R}$.*

Proof. See Section F.9. □

Note that for this decomposition the definition of the bounds λ do not need to be restricted to a compact set. The decomposition is plotted on Figure D.4b.

Proposition D.13 (DC decomposition of the Gaussian Kernel based on (31))**.** *For $x, y \in \mathbb{R}^d$, $\alpha > 0$, a Gaussian kernel $k(x,y) = e^{-\alpha\|x-y\|_2^2}$ has a DC decomposition $k(x,y) = q_+(\|x-y\|_2^2) - q_-(\|x-y\|_2^2)$ with, for all $s \in [0, +\infty)$,*

$$q_+(s) = e^{-\alpha s} + \alpha s, \quad q_-(s) = \alpha s. \tag{51}$$

The maximum and minimum eigenvalues of the Hessian satisfy

$$\lambda_+ = 0, \quad \Lambda_+ = 2\alpha(1 + 2e^{-\frac{3}{2}}), \tag{52}$$

and

$$\lambda_- = 2\alpha, \quad \Lambda_- = 2\alpha. \tag{53}$$

Proof. See Section F.14. □

Smoothed Riesz Kernel. We now deal with the negative distance (Riesz) kernel $k(x,y) = -\|x-y\|_2$. Since it is not differentiable in $x = y$, we instead focus on a smoothed version of it, defined as $k_\varepsilon(x,y) = -\sqrt{\varepsilon + \|x-y\|_2^2}$. In this case, we have $q(t) = -\sqrt{t+\varepsilon}$ for all $t \geq 0$.

Lemma D.6 (DC decomposition of the smoothed Riesz Kernel)**.** *For $\varepsilon > 0$, define $k_\varepsilon(x,y) = -\sqrt{\varepsilon + \|x-y\|_2^2}$. The smoothed Riesz kernel has a DC decomposition as follows: $k_\varepsilon(x-y) = \psi^+(x-y) - \psi^-(x-y)$, where $\psi^\pm(z) = q_\pm(\|z\|_2^2)$ with*

$$q_+(s) = 0, \quad q_-(s) = \sqrt{\varepsilon + s}. \tag{54}$$

For $s \in [0, S_*]$ the maximum and minimum eigenvalues satisfy

$$\lambda_+ = 0, \quad \Lambda_+ = 0, \quad \lambda_- = \frac{\varepsilon}{(\varepsilon + S_*)^{3/2}}, \quad \Lambda_- = \frac{1}{\sqrt{\varepsilon}}. \tag{55}$$

Proof. See Section F.15. □

Rational Quadratic Kernel. We now turn to the rational quadratic kernel $k(x,y) = \psi(x-y)$, where $\psi(z) = \frac{1}{(c^2+\|z\|_2^2)^\alpha}, \alpha \geq 1, c > 0$.

Lemma D.7 (Rational Quadratic Kernel DC decomposition based on (31))**.** *For $x, y \in \mathbb{R}^d$, $\alpha \geq 1$, a rational quadratic kernel has a DC decomposition $k(x,y) = q_+(\|x-y\|_2^2) - q_-(\|x-y\|_2^2)$ with, for all $s \in [0, +\infty)$,*

$$q_+(s) = \frac{1}{(c^2+s)^\alpha} + \alpha c^{-2(\alpha+1)} s, \qquad q_-(s) = \alpha c^{-2(\alpha+1)} s. \tag{56}$$

The maximum and minimum eigenvalues of the Hessian satisfy:

$$\lambda_+ = 0, \quad \Lambda_+ = f(s^*), \quad \lambda_- = 2\alpha c^{-2(\alpha+1)}, \quad \Lambda_- = 2\alpha c^{-2(\alpha+1)}, \tag{57}$$

where $f(s) = 2q_+'(s) + 4sq_+''(s)$ and $s^* = \frac{6c^2}{4(\alpha+2)-6}$.

Proof. See Section F.16. □

D.2 Lipschitz Continuity of $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ and Stationarity

We now show that for a DC decomposition based on radial kernels as in Lemma D.3, the Wasserstein gradient of $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ satisfies a Lipschitz condition, which corresponds to the assumption in Proposition 5.

Proposition D.14 (Lipschitz Continuity of $\nabla_{\mathrm{W}_2}\mathcal{F}^+$)**.** *Assume $\lambda_\pm \geq 0$ and $\Lambda_\pm < \infty$. Let $\sigma, \mu \in \mathcal{P}_2(\Omega)$ be two a.c measures, consider T such that $\sigma = \mathrm{T}_\#\mu$, then we have*

$$\|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\sigma) \circ \mathrm{T} - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu)\|_{L^2(\mu)} \leq L\|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)}, \tag{58}$$

with $L = \sqrt{2}\Lambda_+ + \Lambda_-$.

Proof. See Section F.17. □

Using this Proposition, we can apply Proposition 5 to obtain a sublinear rate over the minimum of the Wasserstein gradient over the scheme.

Theorem D.1 (WCCCP with MMD leads to a Stationary Measure)**.** *Let $\nu \in \mathcal{P}_2(\Omega)$, where Ω is a compact and convex nonempty set in $\mathbb{R}^d$. Consider $(\mu_k)_{k\geq 0}$ the WCCCP iterates (11) for the MMD functional $\mathcal{F}$ with a radial translation-invariant kernel that admits a DC decomposition: $k(x,y) = \psi^+(x-y) - \psi^-(x-y)$. Assume $\lambda_+, \lambda_- \geq 0$, $\lambda_+ + \lambda_- > 0$ and $0 \leq \Lambda_\pm < \infty$, and WCCCP iterates belong to $\mathcal{P}_2(\Omega)$, then these iterates satisfy, for all $K \geq 1$,*

$$\min_{0\leq k\leq K-1} \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|^2_{L^2(\mu_k)} \leq \frac{2(\sqrt{2}\Lambda_+ + \Lambda_-)^2}{\lambda_+ + \lambda_-} \frac{\big(\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)\big)}{K}. \tag{59}$$

Proof. See Section F.18. □

Remark 1. *We can get quantitative bounds for the Gaussian kernel with* cosh / sinh *DC split using Lemma D.4, and with the Jordan decomposition using Proposition D.13. Similarly, we also have these bounds for the smoothed Riesz kernel using Lemma D.6 and for the rational quadratic kernel with Lemma D.7.*

D.3 Critical points and Local Convergence

In the following we analyze the conditions under which we obtain local convergence of WCCCP for the MMD functional with base space Ω. All the statements are in the weak topology and assume a continuous kernel k.

First, we consider a condition on the initialization *w.r.t.* the critical value gap as follows:

Assumption 1. *The set Ω is convex and compact and the following holds true:*

1. **Strong Convexity & Lipschitz Continuity.** *$\mathcal{F}^+$ and $\mathcal{F}^-$ are totally strongly convex, $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ is Lipschitz continuous, and $\mu \mapsto \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)\|_{L^2(\mu)}$ is lower semicontinuous;*
2. **Critical-value gap.** *Setting*

$$c_* := \inf\{\mathcal{F}(\mu) : \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)\|_{L^2(\mu)} = 0, \mu \neq \nu\},$$

the initialization μ_0 satisfies:

$$\mathcal{F}(\mu_0) < c_*.$$

Under Assumption 1, we have the following local convergence result, interpreted as: if one starts with a near optimal initialization, by descent, there must be convergence. The condition $\mathcal{F}(\mu_0) < c_*$ entails that ν is an isolated critical point.

Proposition D.15 (Local Convergence of WCCCP for MMD under Critical Value Gap Assumption)**.** *Under Assumption 1, the sequence produced by WCCCP converges to ν, i.e., $\mu_k \to \nu$.*

Proof. See Section F.19.

□

The following proposition gives local linear convergence without a compactness assumption on the base space, which may be $\mathbb{R}^d$, under local quadratic growth, a local Wasserstein PL inequality, and a suitable initialization; compactness of the orbit closure follows from convergence.

Proposition D.16 (Local Linear Convergence of WCCCP under Quadratic Growth and a PL Inequality)**.** *Let $(\mu_k)_{k\geq 0}$ be the exact WCCCP iterates (11) for the MMD functional $\mathcal{F}$ with target ν. Assume $\mathcal{F}^+$ and $\mathcal{F}^-$ are W-differentiable and respectively λ_+- and λ_--totally convex, with $\lambda_\pm \geq 0$ and $\lambda := \lambda_+ + \lambda_- > 0$. Assume also that $\nabla_{\mathrm{W}_2}\mathcal{F}^+$ satisfies the Lipschitz condition (21) with constant $L > 0$. Suppose there exist $r, c > 0$ such that, writing*

$$B_r(\nu) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathrm{W}_2(\mu,\nu) \leq r\},$$

(i) $\mathcal{F}$ *has local quadratic growth at* ν*:*

$$\mathcal{F}(\mu) \geq c\mathrm{W}_2^2(\mu,\nu), \qquad \mu \in B_r(\nu);$$

(ii) the initialization satisfies

$$\mu_0 \in B_r(\nu), \qquad E_0 := \mathcal{F}(\mu_0) < cr^2;$$

(iii) $\mathcal{F}$ *satisfies a local Wasserstein PL inequality: there exists* $m > 0$ *such that*

$$\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)\|^2_{L^2(\mu)} \geq 2m\mathcal{F}(\mu), \qquad \mu \in B_r(\nu), \quad \mathcal{F}(\mu) \leq E_0. \tag{60}$$

Note that necessarily $m \leq L^2/\lambda$*; set* $q := 1 - \lambda m/L^2 \in [0,1)$*. Then all iterates remain in* $B_r(\nu)$ *and, for every* $k \geq 0$*,*

$$\mathcal{F}(\mu_k) \leq q^k E_0, \qquad \mathrm{W}_2^2(\mu_k,\nu) \leq \frac{E_0}{c}\, q^k.$$

In particular, $\mu_k \to \nu$ *in* W_2*.*

Proof. See Section F.20.

□

Remark 2 (Scope of the local PL assumption)**.** *Assumptions (i) and (iii) do not follow from the total convexity of* $\mathcal{F}^{\pm}$ *or from the Lipschitz condition* (21)*, and whether they hold depends on the target. We leave establishing if they hold for the MMD kernels considered in the paper for future works. The argument above obtains the DC-PL estimate only along the iterates, where the Lipschitz condition* (21) *is assumed.*

E Numerical Applications on MMD

We now detail the experiments of Section 5, as well as extra ones, starting with the experiments on the Energy Distance, and then focus on MMD with the Gaussian kernel. All the numerical applications are done on a Nvidia V100 GPU.

E.1 Energy Distance

We recall that the Energy distance [72] is of the form, for $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{ED}(\mu,\nu) = -\iint \|x-y\|_2 \,\mathrm{d}(\mu-\nu)(x)\mathrm{d}(\mu-\nu)(y), \tag{61}$$

which by Proposition 8, for a fixed target ν, can be decomposed as $\frac{1}{2}\mathrm{ED}(\mu,\nu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu)$ where

$$\mathcal{F}^+(\mu) = \iint \|x-y\|_2 \,\mathrm{d}\nu(y)\mathrm{d}\mu(x) + c(\nu), \quad \mathcal{F}^-(\mu) = \frac{1}{2}\iint \|x-y\|_2 \,\mathrm{d}\mu(x)\mathrm{d}\mu(y), \tag{62}$$

with $c(\nu) = -\frac{1}{2}\iint \|x-y\|_2 \,\mathrm{d}\nu(x)\mathrm{d}\nu(y)$. The functional $\mathcal{F}^+$ is a potential energy with $V(x) = \int \|x-y\|_2 \,\mathrm{d}\nu(y)$ and $\mathcal{F}^-$ an interaction energy, and both are totally convex. If we use the smoothed Riesz kernel $k(x,y) = -\sqrt{\varepsilon + \|x-y\|_2^2}$ and assume that $\mu,\nu \in \mathcal{P}_2(\Omega)$ for Ω a compact convex set, then by Lemma D.6 and Corollary D.1, $\mathcal{F}^+$ is also α-totally convex with $\alpha = \varepsilon/(\varepsilon + S_*)^{\frac{3}{2}}$ with $S_* = \sup_{x,y\in\Omega} \|x-y\|_2$. In practice, we mostly use the non-smooth Riesz kernel $k(x,y) = -\|x-y\|_2$ as it works well [41]. Nonetheless, some smoothed versions based on convolutions have been also shown to have more favorable theoretical properties [69].

Experiments on shapes. For the experiment of Figure 1, we compare the Wasserstein gradient descent (WGD) (5), the Forward-Backward (FB) scheme (25) from [55], and our scheme WCCCP (11). We use a spatial discretization as described in Section 3.3, *i.e.* we first sample $n = 500$ particles from $\mu_0 = \mathcal{N}(0, I_2)$, and n uniform independent particles from the target shape to obtain the target distribution ν_n. Then, at each iteration, we compute the map T_{k+1} and apply it to move each of the

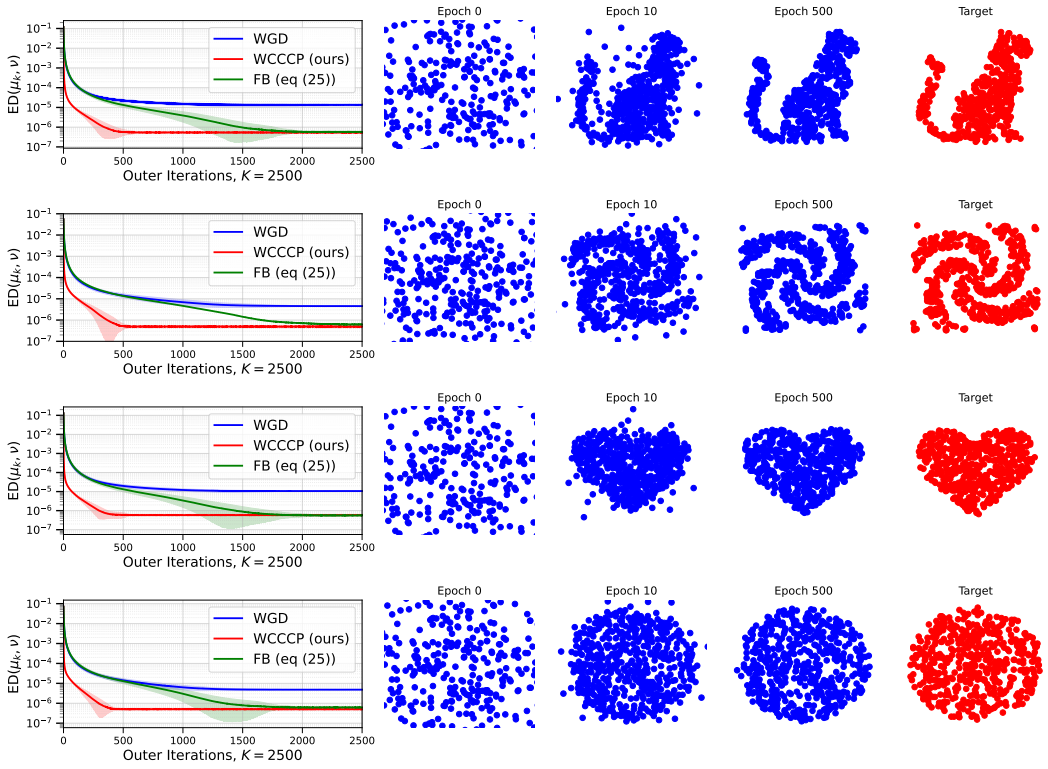


Figure E.5: Convergence of the Wasserstein Gradient Descent (WGD), Forward-Backward (FB) and Wasserstein Convex-Concave Procedure (WCCCP) on the squared MMD with kernel $k(x, y) = -\|x - y\|_2$ (**Left**) and particles of WCCCP along the scheme (**Right**).

particles. For WGD and FB, we use as step size $\tau = 1$. For WCCCP and FB, we solve each inner optimization problem with a gradient descent with momentum $m = 0.9$ for $M = 50$ iterations and step size 0.1. We report on Figure 1 the values of the objective $\mathrm{ED}(\mu_k, \nu_n)$ for the cat and spiral shapes depending on the number of outer iterations (hence WGD is much faster but still converges to a local minimum and does not improve further). We performed the experiment 100 times, and report the average values with standard deviation to get confidence intervals. We add on Figure E.5 the results for the heart and disk shapes, as well as particles for the WCCCP at iterations 0, 10 and 500.

On Figure E.6, we performed the same experiment with target samples from a Gaussian $\nu = \mathcal{N}(0, \Sigma)$ for $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$ (top), and with target samples from a mixture of 3 Gaussian with equal weights, means $m_1 = (0, 0)$, $m_2 = (3, -1)$ and $m_3 = (1, 4)$, and covariances $\Sigma_1 = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $\Sigma_2 = I_2$ and $\Sigma_3 = \left(\begin{smallmatrix} 3 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. For both, we use the same initial distribution $\mu_0 = \mathcal{N}(5, I_2)$ and the same hyperparameters as the shape experiment.

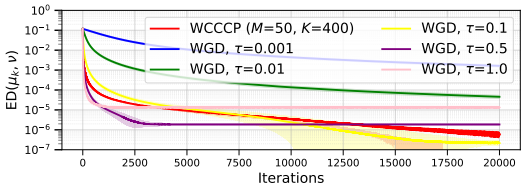


Figure E.7: Convergence of WGD (with different step size and K=20K) and WCCCP with the same computational budget, *i.e.* $M = 50$ and $K = 400$ (each iteration thus corresponding to one inner step). The target is the cat shape.

We also report on Figure E.7 the convergence for the cat target uniform distribution, with the exact same number of iterations between WCCCP and WGD with different step sizes, *i.e.* using $K = 20$K iterations of WGD and $K = 400$ outer iterations of WCCCP and $M = 50$ inner iterations to solve each sub-problem. We observe that for $\tau = 0.1$, WGD converges better but slower than for $\tau = 1$. With the same computational budget for WCCCP with $M = 50$, $K = 400$, we observe that the algorithm converges in similar computational time.

Experiments on images. On Figure 2, we performed the same experiment with target samples from the CIFAR10 dataset, whose images are of size $3 \times 32 \times 32$. More precisely, we sampled 50 points

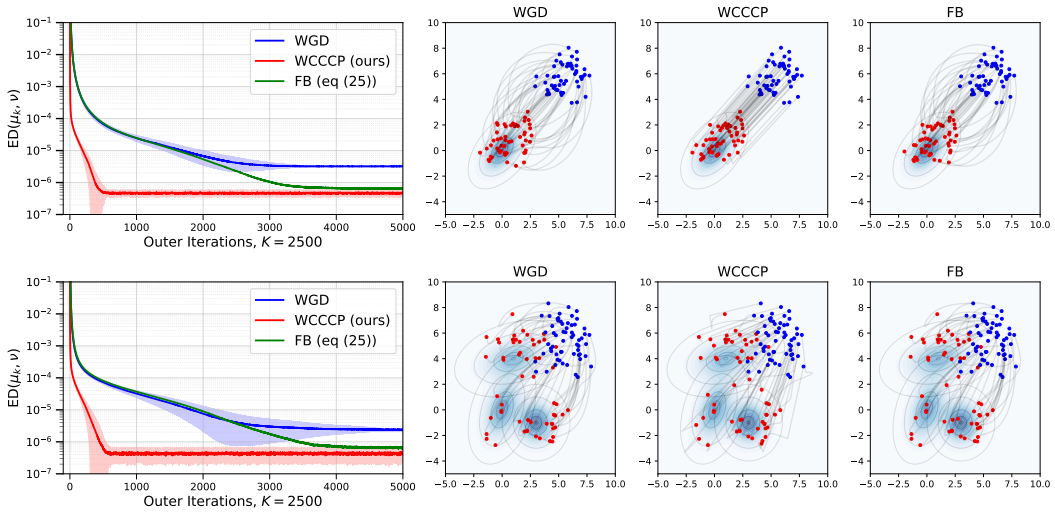


Figure E.6: Optimization of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{ED}(\mu, \nu)$ for ν a Gaussian target (**Top**) and a Gaussian mixture (**Bottom**). (**Left**) Evolution of the squared Energy distance along the flow. (**Right**) Trajectories of the particles over time. The initial particles are in blue and the final particles in red.

by class, and hence also worked with $n = 500$ particles. We started the flows from $\mu_0 = \mathcal{N}(0, I_d)$. We compared WGD with stepsize $\tau = 1$ and WCCCP, using a gradient descent to solve the inner optimization scheme with $\tau = 1$. We ran WGD for K =200K iterations and WCCCP for K =40K iterations, with $M = 20$ iterations to solve each of the subproblems (11). We choose to use a different number of iterations to be fair in comparing the two methods, as WGD solves each inner problem in closed-form, and is thus less computationally expensive. With a Nvidia v100 GPU, WCCCP took about 1h15 while WGD took 1h30. On Figure 2, we plot samples along the scheme of WGD every 20K iterations, and samples along the scheme of WCCCP every 4K iterations. Below, we plot the value of the loss across iterations, where we rescaled the abscissa for WCCCP to match the abscissa of WGD as done on the image samples. We observe overall a much faster convergence of WCCCP compared to WGD on this high dimensional challenging dataset. On Figure E.8, we add more samples along the schemes of WGD and WCCCP, every 2K iterations for WCCCP and 10K iterations for WGD.

We also performed the experiment on MNIST and report the results on Figure E.8. Here, the samples are reported every 1K iterations for WCCCP and 5K iterations for WGD. The schemes took about 18 minutes to run for both WGD and WCCCP.

We show on Figure E.9 the evolution of the Energy distance along the scheme, averaged over 5 runs with different samples of the source and of the target. We observe that on the MNIST experiment, both schemes converge, but WCCCP converges faster (even though the iterations are rescaled). On the CIFAR10 experiment, even after 200K iterations, WGD is very far from converging, whereas WCCCP has already converged. Hence, WCCCP can be promising to accelerate convergence in high dimensions (here $d = 3 \times 32 \times 32 = 3072$).

E.2 MMD with Gaussian Kernel

We now focus on MMD with Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/(2h)}$ for a bandwidth $h > 0$.

Decomposition based on the radial kernel. By Corollary D.1, we can have a DC decomposition of the squared MMD if we find q_+, q_- satisfying $q'_+, q'_-, q''_+, q''_- \geq 0$ and $q = q_+ - q_-$. We discuss here two natural decompositions of $q : t \mapsto e^{-t/(2h)}$.

The first one is based on the remark that $e^{-z} = \cosh(z) - \sinh(z)$ since $\cosh(z) = \left(e^z + e^{-z}\right)/2$ and $\sinh(z) = \left(e^z - e^{-z}\right)/2$, and hence $q_+(t) = \cosh\left(t/(2h)\right)$, $q_-(t) = \sinh\left(t/(2h)\right)$. This is based on the algebraic decomposition described in Section 4 as $\cosh(t) = \sum_{k \text{ even}} \frac{t^k}{k!}$ and

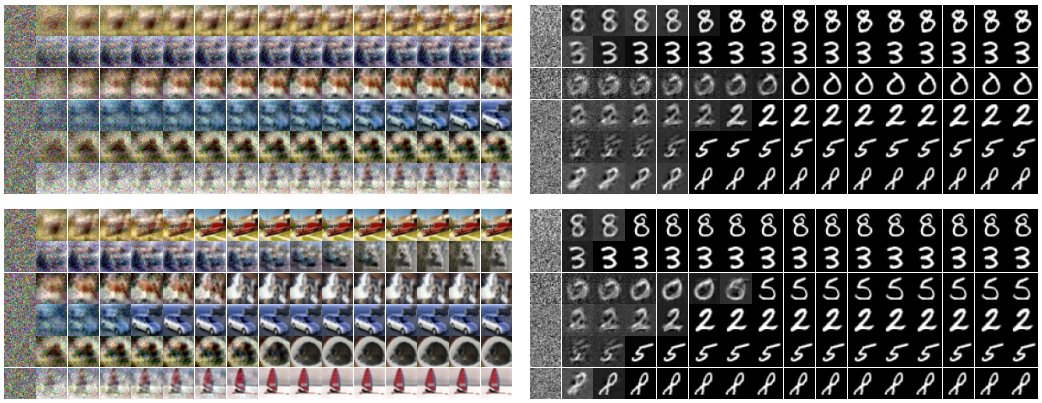


Figure E.8: Samples along the scheme of WGD (**Top**) and WCCCP (**Bottom**) with $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{ED}(\mu, \nu)$ as objective. On CIFAR10, we plot samples every 2K iterations for WCCCP and every 10K iterations for WGD. On MNIST, we plot samples every 1K iterations for WCCCP and every 5K iterations for WGD.

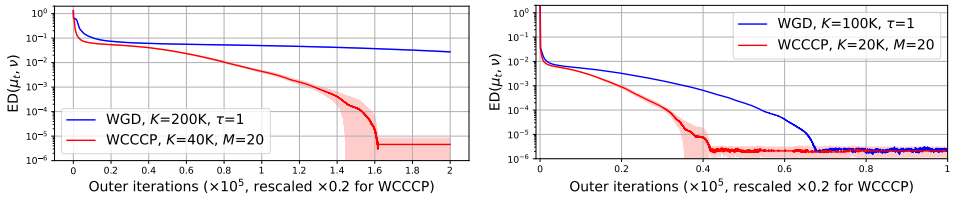


Figure E.9: Evolution of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{ED}(\mu, \nu)$ along the WCCCP and WGD schemes for ν composed of samples of CIFAR10 (**Left**) and of MNIST (**Right**). The results are averaged over 5 different runs with different samples of the source and target distribution.

$\sinh(t) = \sum_{k \text{ odd}} \frac{t^k}{k!}$. This decomposition is valid to apply Proposition 8 as shown in Lemma D.4. We show this decomposition on Figure D.4a. We note that both q_+ and q_- tend to take very large values, which might be prone to numerical instabilities, in particular for small bandwidths.

The second one is based on decomposing the second derivative, and taking the convex part as the function having as second derivative the non-negative part, and q_- as the function whose second derivative is the negative part of q''. This can be done using the Jordan decomposition (31). The exact formula gives $q_+(s) = e^{-\alpha s} + \alpha s$ and $q_-(s) = \alpha s$. This also satisfies Proposition 8 as shown in Proposition D.13. We plot on Figure D.4b the resulting DC decomposition. We observe here that this decomposition does not blow up, and should be numerically more stable.

Experiments. The Wasserstein gradient descent on $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu)$ with $k(x, y) = e^{-\|x-y\|_2^2/(2h)}$ is known to heavily depend on the choice of the bandwidth [6, 41], and might not converge well in practice. In particular, on simple examples such as a single Gaussian distribution, Arbel et al. [6], Gladin et al. [37] observed that some particles get stuck further from the mode, and never converge. In our experiments, we compare several DC decompositions in the WCCCP algorithm and test for which decompositions we obtain a better convergence or not.

For this, we choose the same setting as [37]. We first take $n = 500$ samples from $\mu_0 = \mathcal{N}(5, I_2)$ to get the initial distribution. Then, we take $n = 500$ samples from the target distribution ν. We experiment with 2 choices for ν. The first one is a Gaussian $\nu = \mathcal{N}(0, \Sigma)$ with $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. The second is a mixture of three Gaussians with uniform weights, means $m_1 = (0, 0)$, $m_2 = (3, -1)$, $m_3 = (1, 4)$ and covariances $\Sigma_1 = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $\Sigma_2 = I_2$, $\Sigma_3 = \left(\begin{smallmatrix} 3 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. Moreover, the bandwidth is fixed to $h = 10$.

In each experiment, we use for WGD and FB a stepsize $\tau = 1$. For FB and WCCCP, we optimize the inner problem with a gradient descent with momentum $m = 0.9$, step size $\tau = 5 \cdot 10^{-4}$ and for 250 iterations.

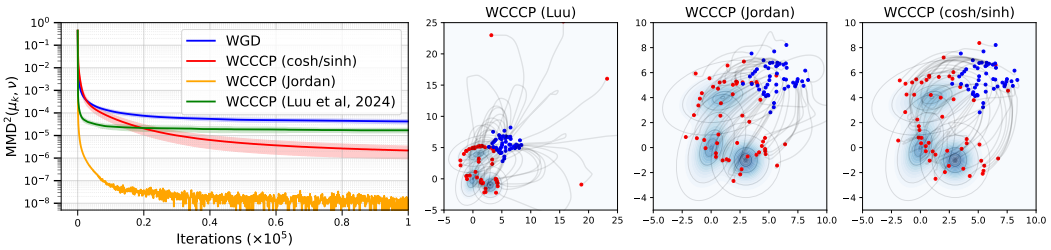


Figure E.10: Optimization of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu)$ for ν a Gaussian mixture target and k the Gaussian kernel. (**Left**) Evolution of the squared MMD along the flow. (**Right**) Trajectories of the particles over times. The initial particles are in blue and the final particles in red.

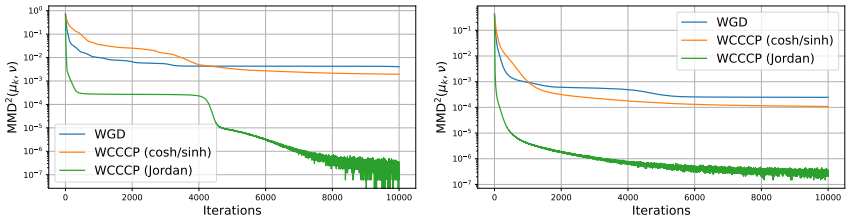


Figure E.11: Loss for one run for MMD with Gaussian kernel, and Gaussian target (**Left**) and Gaussian mixture target (**Right**).

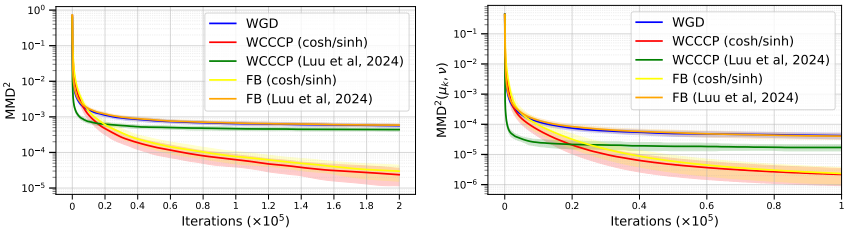


Figure E.12: Evolution of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}_k^2(\mu, \nu)$ for ν a Gaussian target (**Left**) and a Gaussian mixture target (**Right**) over iterations, comparing WGD, WCCCP and FB with the $\cosh / \sinh$ decomposition and the DC decomposition (47). We observe that both FB and WCCCP work well with the $\cosh / \sinh$ decomposition.

On Figure 3, we show the evolution of the loss over 200K iterations of the algorithms, comparing WGD with WCCCP for the DC decomposition (47), and the DC decomposition of the radial kernel based on $\cosh$ and $\sinh$ as presented in Lemma D.4 as well as the decomposition based on the Jordan decomposition presented in Proposition D.13. We show the evolution of the objective over the iterations and average the results over 100 different set of initial and target samples.

We observe that the two radial decompositions perform better than WGD and the DC decomposition (47). The $\cosh / \sinh$ decomposition seems to converge more consistently than the one based on the Jordan decomposition, even though it takes longer to converge. Indeed, the Jordan decomposition converges much faster, but also shows a higher variance, and sometimes, particles get stuck away from the target mode, which does not seem to happen with the $\cosh / \sinh$ decomposition.

We also show the result on Figure E.10 of the same experiment on the Gaussian mixture target. On this target, the Jordan decomposition outperforms all the others. On Figure E.11, we add the loss for one run and $n = 50$. We observe that WCCCP with Jordan decomposition can plateau during the training, and seems to have several convergence regimes for the Gaussian target. Moreover, it is much faster in its first iterations than the $\cosh / \sinh$ decomposition.

On Figure E.12, we plot the evolution of the loss for the Gaussian and Gaussian mixture targets, comparing WGD with WCCCP and the Forward-Backward (FB) algorithm studied in [55]. We compare WCCCP and FB with two DC decompositions: the one based on the DC decomposition

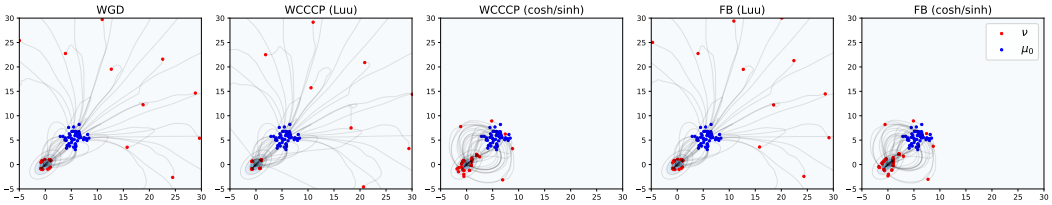


Figure E.13: Evolution of the particles over the scheme for the Gaussian target.

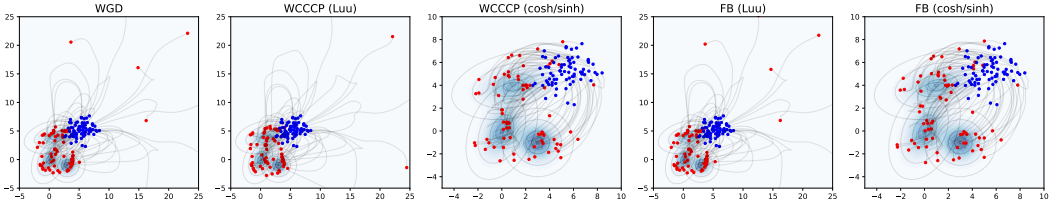


Figure E.14: Evolution of the particles over the scheme for the Gaussian mixture target.

of the radial kernel with $\cosh / \sinh$ as presented in Lemma D.4 and the one based on the DC decomposition (47). We observe that both WCCCP and FB performance heavily depend on the choice of the DC decomposition. In particular, the $\cosh / \sinh$ decomposition seems to work well for both algorithms, while the iterates of the decomposition (47) get stuck in a local minimum, like WGD. We add on Figure E.13 and Figure E.14 the evolutions of particles over the flows of WGD, and WCCCP and FB for both decompositions.

Ablation on the DC Decomposition The experiments exhibit the importance of the choice of the DC decomposition of the target functional, as different DC decompositions induce very different behaviors. We investigate how the performance of the WCCCP method evolves when we change the Jordan decomposition. To do so, we define $q_+^a(s) = e^{-s/(2h)} + a \cdot s/(2h)$ and $q_-^a(s) = a \cdot s/(2h)$ for $a > 0$. For $a = 1$, this recovers the Jordan decomposition. We show in Figure E.15 the value of the squared MMD objective after running WCCCP for $a \in]0, 4]$ with the decomposition $q = q_+^a - q_-^a$, for 10K iterations, and averaged for each a over 5 different initializations. For $a < 1$, note that the decomposition is not valid as q_+ is not non-decreasing. We can run nonetheless WCCCP. We observe that the loss at convergence is pretty stable for $a \in [0.5, 1]$ as there is no clear minimum, but the results seem to get worse as a increases. In the regime where q_+^a is far from convex, *i.e.* $a < 0.5$, we observe very bad convergence.

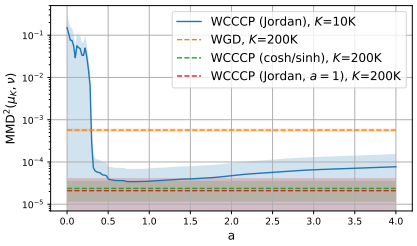


Figure E.15: Evolution of the loss at convergence for $q_+^a(s) = e^{-s/2h} + as/(2h)$, $q_-^a(s) = as/(2h)$.

F Proofs

F.1 Proof of Proposition 1

Let $k \geq 0$, $\mu_k \in \mathcal{P}_2(\mathbb{R}^d)$. On one hand, we have for any $\mathrm{T} \in L^2(\mu_k)$, setting J as the r.h.s. of (10),

$$
\begin{aligned}
\mathrm{J}(\mathrm{T}) &:= \mathcal{F}^+(\mathrm{T}_\# \mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\
&= \mathcal{F}^+(\mathrm{T}_\# \mu_k) - \mathcal{F}^+(\mu_k) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\
&\qquad + \mathcal{F}^+(\mu_k) + \langle \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\
&\quad - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\
&= \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{T}, \mathrm{Id}) + \mathcal{F}^+(\mu_k) - \mathcal{F}^-(\mu_k) + \langle \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\
&= \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{T}, \mathrm{Id}) + \mathcal{F}(\mu_k) + \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)}.
\end{aligned}
\tag{63}
$$

Hence, (11) is equivalent to a mirror descent on $\mathcal{F}$ with geometry induced by the Bregman divergence with Bregman potential $\mathcal{F}^+$.

On the other hand, we also have, for any $\mathrm{T} \in L^2(\mu_k)$,

$$\begin{aligned}\mathrm{J}(\mathrm{T}) &= \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\ &= \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mathrm{T}_\#\mu_k) + \mathcal{F}^-(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\ &= \mathcal{F}(\mathrm{T}_\#\mu_k) + \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}).\end{aligned} \tag{64}$$

Hence, (11) is also equivalent with a Bregman proximal descent on $\mathcal{F}$ with geometry induced by the Bregman divergence with Bregman potential $\mathcal{F}^-$.

F.2 Proof of Proposition 2

$$\begin{aligned}\mathcal{F}(\mu_{k+1}) &= \mathcal{F}(\mu_k) + \mathcal{F}^+(\mu_{k+1}) - \mathcal{F}^+(\mu_k) - \mathcal{F}^-(\mu_{k+1}) + \mathcal{F}^-(\mu_k) \\ &= \mathcal{F}(\mu_k) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) + \mathcal{F}^+(\mu_{k+1}) - \mathcal{F}^+(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T}_{k+1} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\ &= \mathcal{F}(\mu_k) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) - \mathcal{D}^k_{\mathcal{F}^+}.\end{aligned} \tag{65}$$

If $\mathcal{F}^+$ is W-differentiable, observe that by the first order conditions in (11), $\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)$, and thus $\mathcal{D}^k_{\mathcal{F}^+} = \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{Id}, \mathrm{T}_{k+1})$.

F.3 Proof of Proposition 3

If $\mathcal{D}^k_{\mathcal{F}^+} = 0$, then $\mathrm{Id} \in \operatorname{argmin}_{\mathrm{T} \in L^2(\mu_k)} \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}$. Hence the first order condition gives $0 = \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mathrm{Id}_\#\mu_k) \circ \mathrm{Id} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k) = \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)$.

F.4 Proof of Proposition 4

We just use a telescopic sum on (15), discarding $\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id})$ (nonnegative by convexity of $\mathcal{F}^-$), and then use the observation (18).

F.5 Proof of Proposition 5

By (15), for all $k \in \{0, \dots, K-1\}$,

$$\mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) = \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) + \mathcal{D}^k_{\mathcal{F}^+}. \tag{66}$$

$\mathcal{F}^+$ and $\mathcal{F}^-$ are respectively α^+ and α^--convex along iterates with respect to $\mu \mapsto \int \frac{1}{2}\|\cdot\|_2^2 \,\mathrm{d}\mu$. As $\mathcal{F}^-$ is W-differentiable, we have $\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) \geq \frac{\alpha^-}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)}$. For all $t \in [0,1]$, define $\rho_t = \big((1-t)\mathrm{T}_{k+1} + t\mathrm{Id}\big)_\#\mu_k$ and

$$\begin{aligned}g(t) &:= \mathrm{J}\big((1-t)\mathrm{T}_{k+1} + t\mathrm{Id}\big) - \mathcal{F}^-(\mu_k) \\ &= \mathcal{F}^+(\rho_t) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), (1-t)(\mathrm{T}_{k+1} - \mathrm{Id})\rangle_{L^2(\mu_k)}.\end{aligned} \tag{67}$$

By α^+-total convexity of $\mathcal{F}^+$, for all $t \in [0,1]$,

$$\begin{aligned}g(t) &= \mathcal{F}^+(\rho_t) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), (1-t)(\mathrm{T}_{k+1} - \mathrm{Id})\rangle_{L^2(\mu_k)} \\ &\leq (1-t)\mathcal{F}^+(\rho_0) + t\mathcal{F}^+(\rho_1) - \alpha^+\frac{t(1-t)}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \\ &\quad - (1-t)\mathcal{F}^-(\mu_k) - t\mathcal{F}^-(\mu_k) - (1-t)\langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T}_{k+1} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\ &= (1-t)g(0) + tg(1) - \alpha^+\frac{t(1-t)}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)}.\end{aligned} \tag{68}$$

By definition of T_{k+1} in (11), $g(0) = \mathrm{J}(\mathrm{T}_{k+1}) - \mathcal{F}^-(\mu_k) \leq g(t)$ for all $t \in [0,1]$. Hence,

$$\begin{aligned}&g(0) \leq (1-t)g(0) + tg(1) - \alpha^+\frac{t(1-t)}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \\ &\iff \alpha^+\frac{(1-t)}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \leq g(1) - g(0) = \mathrm{J}(\mathrm{Id}) - \mathrm{J}(\mathrm{T}_{k+1}) = \mathcal{D}^k_{\mathcal{F}^+}.\end{aligned} \tag{69}$$

Taking the limit $t \to 0$, we get $\mathcal{D}^k_{\mathcal{F}^+} \geq \frac{\alpha^+}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)}$.

Hence, by (66),

$$\mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) \geq \frac{\alpha^+ + \alpha^-}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)}, \tag{70}$$

and by (18),

$$\begin{aligned}\min_{k\in\{0,\dots,K-1\}} \|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} &\leq \frac{2}{\alpha^+ + \alpha^-} \min_{k\in\{0,\dots,K-1\}} \mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) \\ &\leq \frac{2}{\alpha^+ + \alpha^-} \cdot \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)}{K}.\end{aligned} \tag{71}$$

Under the additional assumption that $\mathcal{F}^+$ is W-differentiable and $\|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k)\|_{L^2(\mu_k)} \leq L\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}$, then by the first order condition (14),

$$\begin{aligned}\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|_{L^2(\mu_k)} &= \|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)\|_{L^2(\mu_k)} \\ &= \|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1}\|_{L^2(\mu_k)} \\ &\leq L\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}.\end{aligned} \tag{72}$$

Hence, combining this with the previous result,

$$\min_{k\in\{0,\dots,K-1\}} \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|^2_{L^2(\mu_k)} \leq \frac{2L^2}{\alpha^+ + \alpha^-} \cdot \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)}{K}. \tag{73}$$

F.6 Proof of Proposition 6

Let $\mathrm{S}_{k+1} \coloneqq \mathrm{Id} + \tau\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)$. First, notice that since $\nu_{k+1} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ by assumption, then (25) is equivalent to

$$\begin{cases} \mathrm{S}_{k+1} = \mathrm{argmin}_{\mathrm{S}\in L^2(\mu_k)} \ \frac{1}{2\tau}\|\mathrm{S} - \mathrm{Id}\|^2_{L^2(\mu_k)} - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{S} - \mathrm{Id}\rangle_{L^2(\mu_k)}, & \nu_{k+1} = (\mathrm{S}_{k+1})_\#\mu_k, \\ \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T}\in L^2(\nu_{k+1})} \ \frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\nu_{k+1})} + \mathcal{F}^+(\mathrm{T}_\#\nu_{k+1}), & \mu_{k+1} = (\mathrm{T}_{k+1})_\#\nu_{k+1}. \end{cases} \tag{74}$$

The first equation is obtained by the first order condition, and the second by observing that T_{k+1} is necessarily an OT map between ν_{k+1} and μ_{k+1}. Indeed, as $\nu_{k+1} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, by Brenier's theorem, there exists a unique OT map between ν_{k+1} and μ_{k+1}, and if T_{k+1} is not such that $\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\nu_{k+1})} = \mathrm{W}_2^2(\mu_{k+1}, \nu_{k+1})$, then we can replace T_{k+1} by the OT map $\mathrm{T}^{\mu_{k+1}}_{\nu_{k+1}}$, which will not change $\mathcal{F}^+\big((\mathrm{T}^{\mu_{k+1}}_{\nu_{k+1}})_\#\nu_{k+1}\big) = \mathcal{F}^+(\mu_{k+1})$ but will give a better transport cost.

Let $J(\mathrm{T}) = \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} + \frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)}$. The gradient of J gives

$$\begin{aligned}\nabla J(\mathrm{T}) &= \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mathrm{T}_\#\mu_k) \circ \mathrm{T} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k) + \frac{1}{\tau}(\mathrm{T} - \mathrm{Id}) \\ &= \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mathrm{T}_\#\mu_k) \circ \mathrm{T} + \frac{1}{\tau}\big(\mathrm{T} - (\mathrm{Id} + \tau\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k))\big) \\ &= \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mathrm{T}_\#\mu_k) \circ \mathrm{T} + \frac{1}{\tau}\big(\mathrm{T} - \mathrm{S}_{k+1}\big).\end{aligned} \tag{75}$$

Taking the first order conditions,

$$\nabla J(\tilde{\mathrm{T}}_{k+1}) = 0 \iff \tilde{\mathrm{T}}_{k+1} = \mathrm{argmin}_{\tilde{\mathrm{T}}\in L^2(\mu_k)} \ \mathcal{F}^+(\tilde{\mathrm{T}}_\#\mu_k) + \frac{1}{2\tau}\|\tilde{\mathrm{T}} - \mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}. \tag{76}$$

Now let us show that $\tilde{\mathrm{T}}_{k+1} = \mathrm{T}_{k+1} \circ \mathrm{S}_{k+1}$, *i.e.*

$$\min_{\tilde{\mathrm{T}}\in L^2(\mu_k)} \mathcal{F}^+(\tilde{\mathrm{T}}_\#\mu_k) + \frac{1}{2\tau}\|\tilde{\mathrm{T}} - \mathrm{S}_{k+1}\|^2_{L^2(\mu_k)} = \min_{\mathrm{T}\in L^2(\nu_{k+1})} \mathcal{F}^+(\mathrm{T}_\#\nu_{k+1}) + \frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\nu_{k+1})}. \tag{77}$$

On one hand, notice by a change of variables, since $\nu_{k+1} = (\mathrm{S}_{k+1})_\#\mu_k$, that for all $\mathrm{T} \in L^2(\nu_{k+1})$,

$$\frac{1}{2\tau}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\nu_{k+1})} + \mathcal{F}^+(\mathrm{T}_\#\nu_{k+1}) = \frac{1}{2\tau}\|\mathrm{T}\circ\mathrm{S}_{k+1} - \mathrm{S}_{k+1}\|^2_{L^2(\mu_k)} + \mathcal{F}^+\big((\mathrm{T}\circ\mathrm{S}_{k+1})_\#\mu_k\big). \tag{78}$$

Since $\{\mathrm{T}\circ\mathrm{S}_{k+1},\ \mathrm{T}\in L^2(\nu_{k+1})\}\subset L^2(\mu_k)$,

$$\begin{aligned}
&\min_{\mathrm{T}\in L^2(\nu_{k+1})}\ \frac{1}{2\tau}\|\mathrm{T}-\mathrm{Id}\|^2_{L^2(\nu_{k+1})}+\mathcal{F}^+(\mathrm{T}_\#\nu_{k+1})\\
&=\min_{\mathrm{T}\in L^2(\nu_{k+1})}\ \frac{1}{2\tau}\|\mathrm{T}\circ\mathrm{S}_{k+1}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}+\mathcal{F}^+\big((\mathrm{T}\circ\mathrm{S}_{k+1})_\#\mu_k\big)\\
&\geq\min_{\mathrm{T}\in L^2(\mu_k)}\ \frac{1}{2\tau}\|\mathrm{T}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}+\mathcal{F}^+(\mathrm{T}_\#\mu_k)\\
&\overset{(76)}{=}\frac{1}{2\tau}\|\tilde{\mathrm{T}}_{k+1}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}+\mathcal{F}^+\big((\tilde{\mathrm{T}}_{k+1})_\#\mu_k\big).
\end{aligned}\tag{79}$$

Now, let us suppose by contradiction that it is a strict inequality, *i.e.*

$$\min_{\mathrm{T}\in L^2(\nu_{k+1})}\ \frac{1}{2\tau}\|\mathrm{T}-\mathrm{Id}\|^2_{L^2(\nu_{k+1})}+\mathcal{F}^+(\mathrm{T}_\#\nu_{k+1})>\frac{1}{2\tau}\|\tilde{\mathrm{T}}_{k+1}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}+\mathcal{F}^+\big((\tilde{\mathrm{T}}_{k+1})_\#\mu_k\big).\tag{80}$$

Define $\eta=(\tilde{\mathrm{T}}_{k+1})_\#\mu_k$. Since $\nu_{k+1}\in\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, there exists an OT map $\mathrm{T}^\eta\in L^2(\nu_{k+1})$ such that $\mathrm{W}_2^2(\eta,\nu_{k+1})=\|\mathrm{T}^\eta-\mathrm{Id}\|^2_{L^2(\nu_{k+1})}$. Notice that $(\tilde{\mathrm{T}}_{k+1},\mathrm{S}_{k+1})_\#\mu_k\in\Pi(\eta,\nu_{k+1})$, hence $\|\tilde{\mathrm{T}}_{k+1}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}\geq\mathrm{W}_2^2(\eta,\nu_{k+1})$ and

$$\begin{aligned}
\min_{\mathrm{T}\in L^2(\nu_{k+1})}\ \frac{1}{2\tau}\|\mathrm{T}-\mathrm{Id}\|^2_{L^2(\nu_{k+1})}+\mathcal{F}^+(\mathrm{T}_\#\nu_{k+1})&>\frac{1}{2\tau}\|\tilde{\mathrm{T}}_{k+1}-\mathrm{S}_{k+1}\|^2_{L^2(\mu_k)}+\mathcal{F}^+\big((\tilde{\mathrm{T}}_{k+1})_\#\mu_k\big)\\
&\geq\frac{1}{2\tau}\|\mathrm{T}^\eta-\mathrm{Id}\|^2_{L^2(\nu_{k+1}}+\mathcal{F}^+((\mathrm{T}^\eta)_\#\nu_{k+1}).
\end{aligned}\tag{81}$$

However, $\mathrm{T}^\eta\in L^2(\nu_{k+1})$ and thus this is a contradiction. Therefore, we necessarily have an equality in (77) and $\tilde{\mathrm{T}}_{k+1}=\mathrm{T}_{k+1}\circ\mathrm{S}_{k+1}$.

F.7 Proof of Proposition 7

Similarly as in Section F.5, since V^- is α^--strongly convex, $\mathcal{F}^-$ is α^--totally convex and we have $\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1},\mathrm{Id})\geq\frac{\alpha^-}{2}\|\mathrm{T}_{k+1}-\mathrm{Id}\|^2_{L^2(\mu_k)}$. Let us show that $\mathcal{D}^k_{\mathcal{F}^+}\geq\frac{\alpha^+}{2}\|\mathrm{T}_{k+1}-\mathrm{Id}\|^2_{L^2(\mu_k)}$.

First, notice by (12) that T_{k+1} is an OT map for the Bregman OT problem $\mathrm{W}_{\mathcal{F}^-}$, *i.e.* $\mathrm{W}_{\mathcal{F}^-}(\mu_{k+1},\mu_k)=\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1},\mathrm{Id})$ (see Section B for the definition of $\mathrm{W}_{\mathcal{F}^-}$). Indeed, $\mu_k\in\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ as there is $\mathcal{H}$ in the optimized functional. Hence, there exist transport maps from μ_k to μ_{k+1}. If T_{k+1} is not the optimal one, we can find another one T such that $\mathrm{T}_\#\mu_k=\mu_{k+1}$ which leaves $\mathcal{F}(\mathrm{T}_\#\mu_k)=\mathcal{F}((\mathrm{T}_{k+1})_\#\mu_k)$ unchanged and will give a better cost $\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T},\mathrm{Id})<\mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1},\mathrm{Id})$, which contradicts that T_{k+1} is optimal.

As T_{k+1} is an OT map for $\mathrm{W}_{\mathcal{F}^-}$, by [17, Proposition 15], there exists a convex function $u_k:\mathbb{R}^d\to\mathbb{R}$ such that $\mathrm{T}_{k+1}=\nabla u_k\circ\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)=\nabla u_k\circ\nabla\mathrm{V}^-$ μ_k-a.e. Define for all $t\in[0,1]$, $\rho_t=\big((1-t)\mathrm{T}_{k+1}+t\mathrm{Id}\big)_\#\mu_k$ an interpolating curve between $\rho_0=\mu_{k+1}$ and $\rho_1=\mu_k$. Since V^- is $\alpha^->0$-strongly convex, $\nabla\mathrm{V}^-$ is invertible and $(\nabla\mathrm{V}^-)^{-1}=\nabla(\mathrm{V}^-)^*$. Hence, the inverse is the gradient of a convex function. Thus, $\rho_t=\big((1-t)\nabla u_k+t\nabla(\mathrm{V}^-)^*\big)_\#(\nabla\mathrm{V}^-)_\#\mu_k$ is a generalized geodesic with base measure $(\nabla\mathrm{V}^-)_\#\mu_k$. Define now for all $t\in[0,1]$

$$\begin{aligned}
g(t)&:=\mathrm{J}\big((1-t)\mathrm{T}_{k+1}+t\mathrm{Id}\big)-\mathcal{F}^-(\mu_k)\\
&=\mathcal{F}^+(\rho_t)-\mathcal{F}^-(\mu_k)-\langle\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k),(1-t)(\mathrm{T}_{k+1}-\mathrm{Id})\rangle_{L^2(\mu_k)}.
\end{aligned}\tag{82}$$

Then, by convexity of $\mathcal{F}^+$ along generalized geodesics, we have for all $t\in[0,1]$,

$$\begin{aligned}
g(t)&=\mathcal{F}^+(\rho_t)-\mathcal{F}^-(\mu_k)-\langle\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k),(1-t)(\mathrm{T}_{k+1}-\mathrm{Id})\rangle_{L^2(\mu_k)}\\
&\leq(1-t)g(0)+tg(1)-\alpha^+\frac{t(1-t)}{2}\|\mathrm{T}_{k+1}-\mathrm{Id}\|^2_{L^2(\mu_k)}.
\end{aligned}\tag{83}$$

We recover (68) and can apply the same proof as in Section F.5 starting from there to find (20).

F.8 Proof of Proposition 8

We start with two preliminary lemmas.

Lemma F.8. *Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\alpha \geq 0$. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a α-strongly convex function that is ν-integrable. Then $\mathrm{V} : x \mapsto \int \psi(x-y) \,\mathrm{d}\nu(y)$ is α-strongly convex.*

Proof. Let $x_0, x_1 \in \mathbb{R}^d$, $t \in [0,1]$, then

$$\begin{aligned}
\mathrm{V}\big((1-t)x_0 + tx_1\big) &= \int \psi\big((1-t)x_0 + tx_1 - y\big) \,\mathrm{d}\nu(y) \\
&= \int \psi\big((1-t)(x_0-y) + t(x_1-y)\big) \,\mathrm{d}\nu(y) \\
&\leq (1-t)\int \psi(x_0-y) \,\mathrm{d}\nu(y) + t\int \psi(x_1-y) \,\mathrm{d}\nu(y) - \frac{\alpha t(1-t)}{2}\|x_0-x_1\|_2^2 \\
&= (1-t)\mathrm{V}(x_0) + t\mathrm{V}(x_1) - \frac{\alpha t(1-t)}{2}\|x_0-x_1\|_2^2.
\end{aligned} \tag{84}$$

□

Lemma F.9. *Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be ν-integrable and define $\mathrm{V} : x \mapsto \int \psi(x-y) \,\mathrm{d}\nu(y)$. Assume there exists $a, b \in \mathbb{R}$ such that $\psi(x) \geq -a - b\|x\|_2^2$ for all $x \in \mathbb{R}^d$. Then, there exists $a', b' \in \mathbb{R}$ such that $\mathrm{V}(x) \geq -a' - b'\|x\|_2^2$.*

Proof. Let $x \in \mathbb{R}^d$, then

$$\mathrm{V}(x) = \int \psi(x-y) \,\mathrm{d}\nu(y) \geq -a - b\int \|x-y\|_2^2 \,\mathrm{d}\nu(y). \tag{85}$$

If $b \geq 0$, we can use that $\int \|x-y\|_2^2 \,\mathrm{d}\nu(y) \leq 2\|x\|_2^2 + 2\int \|y\|_2^2 \,\mathrm{d}\nu(y)$, and thus we have the result for $a' = a + 2b\int \|y\|_2^2 \,\mathrm{d}\nu(y)$ and $b' = 2b$. If $b < 0$, then $\mathrm{V}(x) \geq -a$ and we can use $a' = a$, $b' = 0$. □

We now move to proving Proposition 8. First, let us focus on the potential term of MMD, *i.e.*

$$\mathcal{V}(\mu) = \int \mathrm{V} \,\mathrm{d}\mu, \quad \mathrm{V}(\cdot) = -\int k(\cdot, y) \,\mathrm{d}\nu(y). \tag{86}$$

Since for all $x, y \in \mathbb{R}^d$, $k(x,y) = \psi(x-y) = \psi^+(x-y) - \psi^-(x-y)$, we can rewrite V as, for all $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathrm{V}(x) &= -\int \psi(x-y) \,\mathrm{d}\nu(y) \\
&= -\int \big(\psi^+(x-y) - \psi^-(x-y)\big) \,\mathrm{d}\nu(y) \\
&= \int \psi^-(x-y) \,\mathrm{d}\nu(y) - \int \psi^+(x-y) \,\mathrm{d}\nu(y) \\
&= \mathrm{V}^-(x) - \mathrm{V}^+(x).
\end{aligned} \tag{87}$$

Moreover, as ψ^- and ψ^+ are respectively α^- and α^+-strongly convex and locally Lipschitz, V^- and V^+ are respectively α^- and α^+-strongly convex by Lemma F.8 and continuous.

We now show ψ^- and ψ^+ have a negative with at most quadratic growth. Since they have full domain, they have nonempty subdifferentials everywhere, in particular in 0. Take for instance $p \in \partial\psi^-(0)$, then $\psi^-(x) \geq \psi^-(0) + \langle p, x\rangle$. If $p = 0$, then ψ^- is lower bounded. Assume $p \neq 0$, then

$$\psi^-(x) \geq \psi^-(0) + \langle p, x\rangle \geq \psi^-(0) - \|p\|_2\|x\|_2 \geq \psi^-(0) - \|p\|_2(1 + \|x\|_2^2). \tag{88}$$

The same reasoning applies to ψ^+. Consequently there exists $a^+, a^-, b^+, b^- \in \mathbb{R}$ such that $\psi^+(\cdot) \geq -a^+ - b^+\|\cdot\|_2^2$ and $\psi^-(\cdot) \geq -a^- - b^-\|\cdot\|_2^2$. Consequently, V^- and V^+ have a negative part with a quadratic growth using Lemma F.9.

Hence, by [4, Proposition 9.3.2], $\mathcal{V}$ can be decomposed as a difference of two strongly totally convex potential energies $\mathcal{V} = \mathcal{V}^- - \mathcal{V}^+$ with $\mathcal{V}^- = \int \mathrm{V}^- \mathrm{d}\mu$ α^--totally convex, $\mathcal{V}^+ = \int \mathrm{V}^+ \mathrm{d}\mu$ α^+-totally convex.

Similarly, the interaction energy term $\mathcal{W}(\mu) = \frac{1}{2}\iint k(x,y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y)$ can be decomposed as a difference of totally convex interaction energies $\mathcal{W} = \mathcal{W}^+ - \mathcal{W}^-$ (by [4, Proposition 9.3.5]) with

$$\mathcal{W}^+(\mu) = \frac{1}{2}\iint \psi^+(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y), \quad \mathcal{W}^-(\mu) = \frac{1}{2}\iint \psi^-(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y). \tag{89}$$

Defining $\mathcal{F}^+(\mu) = \mathcal{W}^+ + \mathcal{V}^- + c$ and $\mathcal{F}^-(\mu) = \mathcal{W}^- + \mathcal{V}^+$, we have for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu), \tag{90}$$

where $\mathcal{F}^+$ is α^--totally convex and $\mathcal{F}^-$ is α^+-totally convex, as the sum of a convex term (the interactions) and of a strongly-convex term (the potentials).

F.9 Proof of Lemma 1

Concerning the Jordan decomposition, we have for all $t \geq 0$, $q(t) = e^{-\alpha t}$, hence $q'(t) = -\alpha e^{-\alpha t}$ and $q''(t) = \alpha^2 e^{-\alpha t} \geq 0$. Setting $A = \max(0, -q'(0)) = \alpha$,

$$q_-(t) = \alpha t - \int_0^t (t-s)\min\big(0, q''(s)\big)\,\mathrm{d}s = \alpha t, \tag{91}$$

and

$$q_+(t) = q(t) + q_-(t) = e^{-\alpha t} + \alpha t. \tag{92}$$

Concerning the algebraic decomposition, for all $t \geq 0$,

$$\begin{aligned} q(t) = e^{-\alpha t} = \sum_{k\geq 0} \frac{(-1)^k(\alpha t)^k}{k!} &= \sum_{k\geq 0,\text{even}} \frac{(\alpha t)^k}{k!} - \sum_{k\geq 0,\text{odd}} \frac{(\alpha t)^k}{k!} \\ &= \sum_{\ell\geq 0} \frac{(\alpha t)^{2\ell}}{(2\ell)!} - \sum_{\ell\geq 0} \frac{(\alpha t)^{2\ell+1}}{(2\ell+1)!} \\ &= \cosh(\alpha t) - \sinh(\alpha t) \\ &= q_+(t) - q_-(t). \end{aligned} \tag{93}$$

F.10 Proof of Proposition 9

The assumption $\underline{\lambda}[q_+], \underline{\lambda}[q_-] \geq 0$ allows to deduce that $\psi^+ : z \mapsto q_+(\|z\|_2^2)$ and $\psi^- : z \mapsto q_-(\|z\|_2^2)$ are $\underline{\lambda}[q_+] \geq 0$ and $\underline{\lambda}[q_-] \geq 0$ strongly convex. Hence, we can apply Proposition 8 and obtain that $\mathcal{F}^+$ is $\underline{\lambda}[q_-]$-totally convex, and $\mathcal{F}^-$ is $\underline{\lambda}[q_+]$-totally convex.

Then applying Proposition D.14, we obtain the second result.

F.11 Proof of Proposition B.11

We recall that for all $k \geq 0$,

$$\mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T}\in L^2(\mu_k)}\ \mathcal{F}(\mathrm{T}_{\#}\mu_k) + \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}). \tag{94}$$

Taking the first order conditions, we obtain

$$\begin{aligned} &\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1}) \circ \mathrm{T}_{k+1} + \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k) = 0 \\ &\qquad \iff \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1}) \circ \mathrm{T}_{k+1} = -\big(\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)\big). \end{aligned} \tag{95}$$

Notice that $\mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T}_{\#}\mu_k = \mu_{k+1}} \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id})$, and define $\mathrm{T}^*_k = \mathrm{argmin}_{\mathrm{T}_{\#}\mu_k = \mu^*} \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id})$. Both exist as by assumption, $\mu_k \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ for all $k \geq 0$.

By hypothesis, $\mathcal{F}$ is α-convex relative to $\mathcal{F}^-$ along $t \mapsto \big((1-t)\mathrm{T}_k^* + t\mathrm{T}_{k+1}\big)_{\#}\mu_k$, thus we have

$$\begin{aligned}&\mathrm{D}_{\mathcal{F}}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}) \geq \alpha \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}) \\ &\iff \mathcal{F}(\mu^*) - \mathcal{F}(\mu_{k+1}) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1}) \circ \mathrm{T}_{k+1}, \mathrm{T}_k^* - \mathrm{T}_{k+1}\rangle_{L^2(\mu_k)} \geq \alpha \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}).\end{aligned} \tag{96}$$

By definition of μ^*, $\mathcal{F}(\mu^*) - \mathcal{F}(\mu_{k+1}) \leq 0$. Using the first order conditions, we get the inequality

$$\langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T}_k^* - \mathrm{T}_{k+1}\rangle_{L^2(\mu_k)} \geq \alpha \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}). \tag{97}$$

By the 3-point equality (see [17, Lemma 28]) applied with $\mathrm{T} := \mathrm{T}_{k+1}$, $\mathrm{S} := \mathrm{T}_k^*$ and $\mathrm{U} := \mathrm{Id}$,

$$\begin{aligned}&\langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T}_k^* - \mathrm{T}_{k+1}\rangle_{L^2(\mu_k)} \\ &\qquad = \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{Id}) - \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}) - \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_{k+1}, \mathrm{Id}).\end{aligned} \tag{98}$$

Plugging (98) into (97), we get

$$\mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{Id}) - \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_{k+1}, \mathrm{Id}) \geq (\alpha+1)\mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}). \tag{99}$$

Using that $\mathcal{F}^-$ is convex along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_{\#}\mu_k$, we get that $\mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_{k+1}, \mathrm{Id}) \geq 0$. Using the definition of T_k^*, we have $\mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{Id}) = \mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_k)$ and since $\gamma = (\mathrm{T}_k^*, \mathrm{T}_{k+1})_{\#}\mu_k \in \Pi(\mu^*, \mu_{k+1})$, we also have $\mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_{k+1}) \leq \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1})$. Thus, we obtain by induction

$$\mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_{k+1}) \leq \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{T}_{k+1}) \leq \frac{1}{1+\alpha}\mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_k) \leq \left(\frac{1}{1+\alpha}\right)^{k+1} \mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_0). \tag{100}$$

Moreover, by definition of T_{k+1},

$$\mathcal{F}(\mu_{k+1}) + \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_{k+1}, \mathrm{Id}) \leq \mathcal{F}(\mu^*) + \mathrm{D}_{\mathcal{F}^-}^{\mu_k}(\mathrm{T}_k^*, \mathrm{Id}). \tag{101}$$

Hence,

$$\mathcal{F}(\mu_{k+1}) - \mathcal{F}(\mu^*) \leq \mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_k) \leq \left(\frac{1}{1+\alpha}\right)^{k} \mathrm{W}_{\mathcal{F}^-}(\mu^*, \mu_0). \tag{102}$$

F.12 Proof of Lemma D.2

Let $s = \|z\|_2^2$ and recall that $\psi(z) = q(\|z\|_2^2)$. Differentiating, we have $\nabla\psi(z) = 2q'(s)z$, hence differentiating a second time we obtain

$$\nabla^2\psi(z) = 2q'(s)I_d + 4q''(s)zz^\top. \tag{103}$$

Fix z and let $e \perp z$ and $\|e\|_2 = 1$ then $zz^\top e = 0$, and

$$\nabla^2\psi(z)e = 2q'(s)e, \tag{104}$$

then e is a tangential eigenvector, with eigenvalue $2q'(s)$. We can find $d-1$ vectors that are linearly independent and $\perp z$, hence $2q'(s)$ is an eigenvalue with multiplicity $d-1$. Now consider in the radial direction $v = \frac{z}{\|z\|_2}$ (when $z \neq 0$), we have:

$$\begin{aligned}\nabla^2\psi(z)v &= \big(2q'(s)I_d + 4q''(s)zz^\top\big)\, v \\ &= 2q'(s)v + 4q''(s)zz^\top \frac{z}{\|z\|_2} \\ &= 2q'(s)v + 4q''(s)\|z\|_2 z \\ &= 2q'(s)v + 4q''(s)\|z\|_2^2 v \\ &= \big(2q'(s) + 4q''(s)s\big)v.\end{aligned} \tag{105}$$

Hence v is an eigenvector of $\nabla^2\psi(z)$ with eigenvalue $(2q'(s) + 4q''(s)s)$ of multiplicity 1. Hence we have identified the full spectrum of $\nabla^2\psi(z)$, and the smallest eigenvalue is

$$\min\big\{2q'(s),\, 2q'(s) + 4sq''(s)\big\}, \tag{106}$$

and the largest eigenvalue is

$$\max\big\{2q'(s),\, 2q'(s) + 4sq''(s)\big\}. \tag{107}$$

F.13 Proof of Lemma D.4

Recall that $k(x,y) = \psi(x-y)$, where $\psi(x-y) = \exp(-\alpha\|x-y\|_2^2)$. Let $q(s) = \exp(-\alpha s)$, we have $q(s) = q_+(s) - q_-(s)$ where

$$q_+(s) = \cosh(\alpha s), \quad q_-(s) = \sinh(\alpha s). \tag{108}$$

Taking the derivatives, we have

$$q'_+(s) = \alpha \sinh(\alpha s), \quad q''_+(s) = \alpha^2 \cosh(\alpha s), \tag{109}$$

and

$$q'_-(s) = \alpha \cosh(\alpha s), \quad q''_-(s) = \alpha^2 \sinh(\alpha s). \tag{110}$$

It is easy to see that $q'_\pm(s) \geq 0$ and $q''_\pm(s) \geq 0$ for $s \geq 0$. Therefore $z \mapsto \psi_\pm(z) = q_\pm(\|z\|_2^2)$ are convex on $\Omega - \Omega$ by Lemma D.3, and we have a DC decomposition of the Gaussian kernel.

On $[0, S_*]$ we have:

$$\lambda_+ = \inf_{s\in[0,S_*]} \min\left(2\alpha\sinh(\alpha s), 2\alpha\sinh(\alpha s) + 4\alpha^2 s\cosh(\alpha s)\right) = \inf_{s\in[0,S_*]} 2\alpha\sinh(\alpha s) = 0, \tag{111}$$

and

$$\begin{aligned}\Lambda_+ &= \sup_{s\in[0,S_*]} \max\left(2\alpha\sinh(\alpha s), 2\alpha\sinh(\alpha s) + 4\alpha^2 s\cosh(\alpha s)\right) \\ &= \sup_{s\in[0,S_*]} 2\alpha\sinh(\alpha s) + 4\alpha^2 s\cosh(\alpha s) \\ &= 2\alpha\sinh(\alpha S_*) + 4\alpha^2\cosh(\alpha S_*).\end{aligned} \tag{112}$$

Similarly we have:

$$\lambda_- = 2\alpha, \quad \Lambda_- = 2\alpha\cosh(\alpha S_*) + 4\alpha^2\sinh(\alpha S_*). \tag{113}$$

F.14 Proof of Proposition D.13

Recall that $k(x,y) = \psi(x-y)$, where $\psi(x-y) = e^{-\alpha\|x-y\|_2^2}$. Let $q(s) = e^{-\alpha s}$, then $q(s) = q_+(s) - q_-(s)$ where for all $s \geq 0$,

$$q_+(s) = e^{-\alpha s} + \alpha s, \quad q_-(s) = \alpha s. \tag{114}$$

Taking the derivatives, we get

$$q'_+(s) = \alpha(1 - e^{-\alpha s}), \quad q''_+(s) = \alpha^2 e^{-\alpha s}, \tag{115}$$

and

$$q'_-(s) = \alpha, \quad q''_-(s) = 0. \tag{116}$$

Hence, for $s \geq 0$, $q'_\pm(s) \geq 0$ and thus $z \mapsto \psi_\pm(\|z\|_2^2)$ are convex by Lemma D.3.

Moreover, we get the following minimal and maximal eigenvalues,

$$\lambda_+ = \inf_{s\in[0,+\infty)} \min\left\{2\alpha(1-e^{-\alpha s}), 2\alpha(1-e^{-\alpha s}) + 4s\alpha^2 e^{-\alpha s}\right\} = \inf_{s\in[0,+\infty)} 2\alpha(1-e^{-\alpha s}) = 0, \tag{117}$$

and

$$\begin{aligned}\Lambda_+ &= \sup_{s\in[0,+\infty)} \max\left\{2\alpha(1-e^{-\alpha s}), 2\alpha(1-e^{-\alpha s}) + 4s\alpha^2 e^{-\alpha s}\right\} \\ &= \sup_{s\in[0,+\infty)} 2\alpha(1-e^{-\alpha s}) + 4s\alpha^2 e^{-\alpha s} = 2\alpha + 2\alpha(2s\alpha - 1)e^{-\alpha s}.\end{aligned} \tag{118}$$

Let $f(s) = 2\alpha + 2\alpha(2s\alpha - 1)e^{-\alpha s}$, its derivative gives $f'(s) = 4\alpha^2 e^{-\alpha s} - 2\alpha^2(2s\alpha - 1)e^{-\alpha s} = 2\alpha^2 e^{-\alpha s}(3 - 2s\alpha) = 0 \iff s = \frac{3}{2\alpha}$. Moreover, $f''(s) = -2\alpha^3 e^{-\alpha s}(3 - 2s\alpha) - 4\alpha^3 e^{-s\alpha} = -2\alpha^3 e^{-s\alpha}(5 - 2s\alpha) \leq 0 \iff s \leq \frac{5}{2\alpha}$. Hence $s = \frac{3}{2\alpha}$ is the maximizer, and

$$\Lambda_+ = 2\alpha(1 + 2e^{-\frac{3}{2}}). \tag{119}$$

Similarly, $\lambda_- = \Lambda_- = 2\alpha$.

F.15 Proof of Lemma D.6

The derivatives of q_- give, for all $s \in \mathbb{R}$,

$$q'_-(s) = \frac{1}{2\sqrt{\varepsilon + s}}, \quad q''_-(s) = -\frac{1}{4(\varepsilon+s)^{3/2}}. \tag{120}$$

For the minimum eigenvalue we have

$$\begin{aligned} \min\{2q'_-(s),\, 2q'_-(s) + 4sq''_-(s)\} &= 2q'_-(s) + 4sq''_-(s) \\ &= \frac{1}{\sqrt{\varepsilon+s}} - \frac{s}{(s+\varepsilon)^{\frac{3}{2}}} \\ &= \frac{\varepsilon}{(s+\varepsilon)^{3/2}}. \end{aligned} \tag{121}$$

Therefore for $s \in [0, S_*]$, we have

$$\lambda_- = \frac{\varepsilon}{(S_*+\varepsilon)^{3/2}}. \tag{122}$$

For the maximum eigenvalue we have

$$\max\{2q'_-(s),\, 2q'_-(s) + 4sq''_-(s)\} = 2q'_-(s) = \frac{1}{\sqrt{\varepsilon+s}}. \tag{123}$$

Therefore for $s \in [0, S_*]$, we have

$$\Lambda_- = \frac{1}{\sqrt{\varepsilon}}. \tag{124}$$

F.16 Proof of Lemma D.7

$\psi(z) = \frac{1}{(c^2+\|z\|_2^2)^\alpha}$, this corresponds to $q(s) = \frac{1}{(c^2+s)^\alpha}$, $\alpha \geq 1$. We have for $s \geq 0$,

$$q'(s) = -\alpha\left(c^2+s\right)^{-\alpha-1}, \quad q''(s) = \alpha(\alpha+1)\left(c^2+s\right)^{-\alpha-2}. \tag{125}$$

Hence we have $q'(s) < 0$ and $q''(s) > 0$, and thus $q(s)$ is convex on $s \geq 0$, but ψ is not convex since $q'(s) < 0$. Set $A = \max\left(0, -q'(0)\right) = \alpha c^{-2(\alpha+1)}$, hence

$$q_-(s) = As - \int_0^s (s-t)\min\left(0, q''(t)\right)\,\mathrm{d}t = \alpha c^{-2(\alpha+1)} s, \tag{126}$$

and

$$q_+(s) = q(s) + q_-(s) = \frac{1}{(c^2+s)^\alpha} + \alpha c^{-2(\alpha+1)} s. \tag{127}$$

Now turning to minimum of the Hessian we have

$$\underline{\lambda}[q_+] \coloneqq \inf_{s\geq 0} \min\{2q'_+(s),\, 2q'_+(s) + 4sq''_+(s)\}, \quad \underline{\lambda}[q_-] \coloneqq \inf_{s\geq 0} \min\{2q'_-(s),\, 2q'_-(s) + 4sq''_-(s)\}. \tag{128}$$

By construction $q'_+(s) \geq 0$ and $q''_+(s) \geq 0$, and hence q'_+ is non decreasing and $\underline{\lambda}[q_+] = \inf_{s\geq 0} 2q'_+(s) = 2q'_+(0) = -A + A = 0$. On the other hand $q'_-(s) = \alpha c^{-2(\alpha+1)}$ and $q''_-(s) = 0$ and hence $\underline{\lambda}[q_-] = 2\alpha c^{-2(\alpha+1)}$. For the maximum of the Hessian we have as well,

$$\overline{\Lambda}[q_+] = \sup_{s\geq 0} \max\{2q'_+(s),\, 2q'_+(s) + 4sq''_+(s)\} = \sup_{s\geq 0}\; 2q'_+(s) + 4sq''_+(s). \tag{129}$$

Let us note for $s \geq 0$,

$$f(s) = 2q'_+(s) + 4sq''_+(s), \quad f'(s) = 6q''_+(s) + 4sq'''_+(s). \tag{130}$$

We have

$$\begin{aligned} f'(s) &= 6\alpha(\alpha+1)(c^2+s)^{-\alpha-2} - 4\alpha(\alpha+1)(\alpha+2)s(c^2+s)^{-\alpha-3} \\ &= \alpha(\alpha+1)(c^2+s)^{-\alpha-2}\left(6 - 4(\alpha+2)\frac{s}{c^2+s}\right). \end{aligned} \tag{131}$$

Let s^* such that

$$6 - 4(\alpha+2)\frac{s^*}{c^2+s^*} = 0, \tag{132}$$

equivalently

$$s^* = \frac{6c^2}{4(\alpha+2)-6} \geq 0. \tag{133}$$

We have $f'(s) \geq 0$ for $s \in [0, s^*]$ and $f'(s) \leq 0$ for $s \in [s^*, +\infty[$, hence $f(s^*)$ is the global sup. The result for $\overline{\Lambda}[q_-]$ is immediate.

F.17 Proof of Proposition D.14

Recall that based on the DC decomposition of Proposition 8, $\mathcal{F}^+ = \mathcal{W}^+ + \mathcal{V}^- + c$ where

$$\mathcal{W}^+(\mu) = \frac{1}{2}\iint \psi^+(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y), \quad \mathcal{V}^-(\mu) = \int \mathrm{V}^-(x)\mathrm{d}\mu(x), \tag{134}$$

and

$$\mathrm{V}^-(x) = \int \psi^-(x-y)\,\mathrm{d}\nu(y). \tag{135}$$

As $\mathcal{W}^+$ is an interaction energy, and $\mathcal{V}^-$ a potential energy, their Wasserstein gradients at $\mu \in \mathcal{P}_2(\Omega)$ read as, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} \nabla_{\mathrm{W}_2}\mathcal{W}^+(\mu)(x) &= \nabla(\psi^+ * \mu)(x) \\ \nabla_{\mathrm{W}_2}\mathcal{V}^-(\mu)(x) &= \nabla \mathrm{V}^-(x) = \nabla(\psi^- * \nu)(x). \end{aligned} \tag{136}$$

Hence, the Wasserstein gradient of $\mathcal{F}^+$ at μ is, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu)(x) &= \nabla_{\mathrm{W}_2}\mathcal{W}^+(\mu)(x) + \nabla_{\mathrm{W}_2}\mathcal{V}^-(\mu)(x) \\ &= \nabla(\psi^+ * \mu)(x) + \nabla(\psi^- * \nu)(x). \end{aligned} \tag{137}$$

Let $\mathrm{T} \in L^2(\mu)$ such that $\mathrm{T}_{\#}\mu = \sigma$. Define $a, b : \mathbb{R}^d \to \mathbb{R}^d$ as

$$a(x) \coloneqq \nabla(\psi^+ * \mu)(x) - \nabla(\psi^+ * \sigma)\big(\mathrm{T}(x)\big), \quad b(x) \coloneqq \nabla(\psi^- * \nu)(x) - \nabla(\psi^- * \nu)\big(\mathrm{T}(x)\big). \tag{138}$$

Then, we have the following relation between the Wasserstein gradient of $\mathcal{F}^+$ at μ and σ:

$$\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu)(x) - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\sigma)(\mathrm{T}(x)) = a(x) + b(x). \tag{139}$$

Bounding the term $\|a\|_{L^2(\mu)}$. Looking at the term a we have:

$$\begin{aligned} a(x) &= \int \nabla\psi^+(x-y)\,\mathrm{d}\mu(y) - \int \nabla\psi^+\big(\mathrm{T}(x)-z\big)\,\mathrm{d}\sigma(z) \\ &= \int \big(\nabla\psi^+(x-y) - \nabla\psi^+\big(\mathrm{T}(x)-\mathrm{T}(y)\big)\big)\,\mathrm{d}\mu(y), \end{aligned} \tag{140}$$

since $\mathrm{T}_{\#}\mu = \sigma$. $\nabla\psi_+$ is Λ_+-Lipschitz by assumption on $\Omega - \Omega$ and hence we have:

$$\|a(x)\|_2 \leq \Lambda_+ \int \|(x-y) - (\mathrm{T}(x) - \mathrm{T}(y))\|_2\,\mathrm{d}\mu(y). \tag{141}$$

Now, Jensen inequality and rearranging terms we obtain:

$$\|a(x)\|_2^2 \leq \Lambda_+^2 \int \|(x - \mathrm{T}(x)) - (y - \mathrm{T}(y))\|_2^2\,\mathrm{d}\mu(y) \tag{142}$$

Integrating on x *w.r.t* μ,

$$\int \|a(x)\|_2^2\,\mathrm{d}\mu(x) \leq \Lambda_+^2 \int \|(x - \mathrm{T}(x)) - (y - \mathrm{T}(y))\|_2^2\,\mathrm{d}\mu(x)\mathrm{d}\mu(y). \tag{143}$$

Let us now provide an intermediary lemma before bounding a.

Lemma F.10. *Let $H \in L^2(\mu)$, we have:*

$$\int \|H(x) - H(y)\|_2^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) = 2 \int \|H(x)\|_2^2 \, \mathrm{d}\mu(x) - 2 \left\| \int H(x) \, \mathrm{d}\mu(x) \right\|_2^2 \tag{144}$$

Proof of Lemma F.10. Note that:

$$\begin{aligned} \int \|H(x) - H(y)\|_2^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) &= \int (\|H(x)\|^2 + \|H(y)\|^2 - 2\langle H(x), H(y)\rangle) \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\ &= 2 \int \|H(x)\|_2^2 \, \mathrm{d}\mu(x) - 2 \left\| \int H(x) \, \mathrm{d}\mu(x) \right\|_2^2 . \end{aligned} \tag{145}$$

□

Using Lemma F.10 in Equation (143) for $H(x) = x - \mathrm{T}(x)$ we obtain:

$$\begin{aligned} \int \|a(x)\|_2^2 \, \mathrm{d}\mu(x) &\leq 2\Lambda_+^2 \int \|x - \mathrm{T}(x)\|_2^2 \, \mathrm{d}\mu(x) - 2\Lambda_+^2 \left\| \int (x - \mathrm{T}(x)) \, \mathrm{d}\mu(x) \right\|_2^2 \\ &\leq 2\Lambda_+^2 \|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)}^2. \end{aligned} \tag{146}$$

Therefore we have:

$$\|a\|_{L^2(\mu)} \leq \sqrt{2}\Lambda_+ \|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)}. \tag{147}$$

Bounding the term $\|b\|_{L^2(\mu)}$. Note that $\nabla \mathrm{V}^-$ is Λ_--Lipschitz since $\nabla^2 \psi^- \preceq \Lambda_- I_d$ and hence for all $x \in \Omega$,

$$\nabla^2 \mathrm{V}^-(x) = \int \nabla^2 \psi^-(x - y) \, \mathrm{d}\nu(y) \preceq \Lambda_- I_d. \tag{148}$$

Hence

$$\|b(x)\|_2 = \|\nabla \mathrm{V}^-(x) - \nabla \mathrm{V}^-(\mathrm{T}(x))\|_2 \leq \Lambda_- \|x - \mathrm{T}(x)\|_2. \tag{149}$$

Integrating on x and by Jensen inequality, we have finally

$$\|b\|_{L^2(\mu)} \leq \Lambda_- \|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)}. \tag{150}$$

Final Bound. Combining equations (143) and (150) we obtain:

$$\begin{aligned} \|\nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu) - \nabla_{\mathrm{W}_2} \mathcal{F}^+(\sigma) \circ \mathrm{T}\|_{L^2(\mu)} &= \|a(x) + b(x)\|_{L^2(\mu)} \\ &\leq \|a\|_{L^2(\mu)} + \|b\|_{L^2(\mu)} \\ &\leq \left(\sqrt{2}\Lambda_+ + \Lambda_-\right) \|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)}. \end{aligned} \tag{151}$$

F.18 Proof of Theorem D.1

In Wasserstein CCCP we have $\nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k)$ and hence

$$\begin{aligned} \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k) &= \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2} \mathcal{F}^-(\mu_k) \\ &= \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1}. \end{aligned} \tag{152}$$

Hence we have

$$\begin{aligned} \|\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\|_{L^2(\mu_k)} &= \|\nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k) - \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1}\|_{L^2(\mu_k)} \\ &\leq L \|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}, \end{aligned} \tag{153}$$

where we used Proposition D.14 and $L = \sqrt{2}\Lambda_+ + \Lambda_-$. Since $\lambda_\pm \geq 0$ and $\lambda_+ + \lambda_- > 0$, we can apply Proposition 5 with $\alpha_\pm = \lambda_\pm$. Taking minimum on both sides in inequality (153), and applying Proposition 5 we have finally

$$\begin{aligned} \min_{0 \leq k \leq K-1} \|\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\|_{L^2(\mu_k)}^2 &\leq L^2 \min_{0 \leq k \leq K-1} \|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}^2 \\ &\leq \frac{2L^2}{\alpha^+ + \alpha^-} \frac{\left(\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)\right)}{K}. \end{aligned} \tag{154}$$

In particular, $\min_{0 \leq k \leq K-1} \|\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\|_{L^2(\mu_k)} \to 0$ as $K \to \infty$.

F.19 Proof of Proposition D.15

Let $(\mu_k)_k$ be the sequence produced by WCCCP. The compactness of Ω implies the compactness of $\mathcal{P}_2(\Omega)$, hence we can extract a subsequence $\mu_{k_j} \to \mu^*$ in W_2.

Using Theorem D.1 (satisfied under the first point in Assumption 1), $\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k_j})\|_{L^2(\mu_{k_j})} \to 0$. Indeed, by (70) and (72) which are obtained while proving Theorem D.1, we have the following inequalities

$$\begin{cases} \frac{\lambda_+ + \lambda_-}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \leq \mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}), \\ \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|_{L^2(\mu_k)} \leq L\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}. \end{cases} \tag{155}$$

Then, since $\mathcal{F}$ is bounded below, telescoping gives

$$\frac{\lambda_+ + \lambda_-}{2}\sum_{k=0}^{K-1}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \leq \mathcal{F}(\mu_0) - \mathcal{F}(\mu_K) \leq \mathcal{F}(\mu_0) - \inf\mathcal{F}. \tag{156}$$

Letting $K \to \infty$ yields

$$\sum_{k=0}^{\infty}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \leq \frac{2(\mathcal{F}(\mu_0) - \inf\mathcal{F})}{\lambda_+ + \lambda_-} < \infty. \tag{157}$$

Therefore, $\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)} \to 0$ and $\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|_{L^2(\mu_k)} \to 0$ by (72). Thus, the norm of the gradient vanishes along the full sequence, and in particular along the subsequence.

By lower semicontinuity of $\mu \mapsto \|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)\|_{L^2(\mu)}$, we have $\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu^*)\|_{L^2(\mu^*)} = 0$ and μ^* is then a critical point. Since under our assumptions WCCCP guarantees descent of values and under Assumption 1 we have $\mathcal{F}(\mu_0) < c_*$, we obtain:

$$\mathcal{F}(\mu^*) \leq \mathcal{F}(\mu_0) < c_*.$$

By the definition of c_*, the limit μ^* must be equal to ν ($\mu^* = \nu$). We finally conclude that every subsequence μ_{k_j} converges to ν, and hence $\mu_k \to \nu$.

F.20 Proof of Proposition D.16

Local quadratic growth and the initialization assumption give $\mathrm{W}_2(\mu_0, \nu) < r$. To prove that the iterates remain in the ball, define the majorizing surrogate

$$Q_k(\mathrm{T}) := \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mu_k) - \langle\nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}.$$

By (10), total convexity of $\mathcal{F}^+$, and optimality of T_{k+1}, the interpolations $\mathrm{T}_{k,t} := (1-t)\mathrm{Id} + t\mathrm{T}_{k+1}$ and $\mu_{k,t} := (\mathrm{T}_{k,t})_\#\mu_k$ satisfy

$$\mathcal{F}(\mu_{k,t}) \leq Q_k(\mathrm{T}_{k,t}) \leq (1-t)Q_k(\mathrm{Id}) + tQ_k(\mathrm{T}_{k+1}) \leq \mathcal{F}(\mu_k), \qquad t \in [0,1].$$

The path is continuous in W_2, since

$$\mathrm{W}_2(\mu_{k,t}, \mu_{k,u}) \leq |t-u|\,\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}.$$

If $\mathrm{W}_2(\mu_k, \nu) < r$ and $\mathcal{F}(\mu_k) \leq E_0$, this path cannot reach the boundary of $B_r(\nu)$, where quadratic growth would give $\mathcal{F} \geq cr^2 > E_0$. Induction therefore yields $\mathrm{W}_2(\mu_k, \nu) < r$ and $\mathcal{F}(\mu_k) \leq E_0$ for all k, so the local PL inequality applies at every iterate.

The descent and gradient estimates established in the proof of Proposition 5, together with (60), give

$$\begin{aligned} \mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) &\geq \frac{\lambda}{2}\|\mathrm{T}_{k+1} - \mathrm{Id}\|^2_{L^2(\mu_k)} \\ &\geq \frac{\lambda}{2L^2}\|\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\|^2_{L^2(\mu_k)} \\ &\geq \frac{\lambda m}{L^2}\mathcal{F}(\mu_k). \end{aligned}$$

By (17) and $\inf\mathcal{F} = \mathcal{F}(\nu) = 0$, this is the DC-PL estimate (38) along the iterates with $\eta_1 = \lambda m/L^2$ and $\eta_2 = 0$. The iteration argument in Proposition C.12 thus yields $\mathcal{F}(\mu_k) \leq q^k E_0$. Applying local quadratic growth gives the stated bound on $\mathrm{W}_2^2(\mu_k, \nu)$ and convergence to ν.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are to extend the CCCP algorithm to the Wasserstein space, proving almost stationarity along of its iterates and applying it to the MMD. The Wasserstein CCCP is introduced in Section 3 along with its theoretical analysis, and the application to MMD is described in Section 4 and Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of the current analysis and method in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the assumptions are provided in each theoretical results. And all the proofs are in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiments are described in Section 5 and the full details to reproduce them are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We attached to the submission supplementary materials with the code to reproduce experiments, with script to reproduce them in the same setting.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All the details on the datasets and optimizers are provided in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All plot showing convergence results are run several times, and are plotted with standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All the experiments were done on a Nvidia V100 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper is conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The research conducted in this paper is conform with the NeurIPS Code of Ethics.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We did not use LLMs for core method development.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.