Open access • Journal Article • DOI:10.1007/S10107-021-01636-Z

Sparse optimization on measures with over-parameterized gradient descent

Lénaïc Chizat

Institutions: Université Paris-Saclay

Published on: 17 Mar 2021 - Mathematical Programming (Springer Berlin Heidelberg)

Topics: Gradient descent, Convex function, Parameterized complexity and Measure (mathematics)

Sparse Optimization on Measures with Over-parameterized Gradient Descent

Lenaic Chizat

Sparse Optimization on Measures with Over-parameterized Gradient Descent

Lénaïc Chizat*

November 2, 2020

Abstract

Minimizing a convex function of a measure with a sparsity-inducing penalty is a typical problem arising, e.g., in sparse spikes deconvolution or two-layer neural networks training. We show that this problem can be solved by discretizing the measure and running non-convex gradient descent on the positions and weights of the particles. For measures on a d-dimensional manifold and under some non-degeneracy assumptions, this leads to a global optimization algorithm with a complexity scaling as $\log(1/\epsilon)$ in the desired accuracy ϵ, instead of ϵ^{-d} for convex methods. The key theoretical tools are a local convergence analysis in Wasserstein space and an analysis of a perturbed mirror descent in the space of measures. Our bounds involve quantities that are exponential in d which is unavoidable under our assumptions.

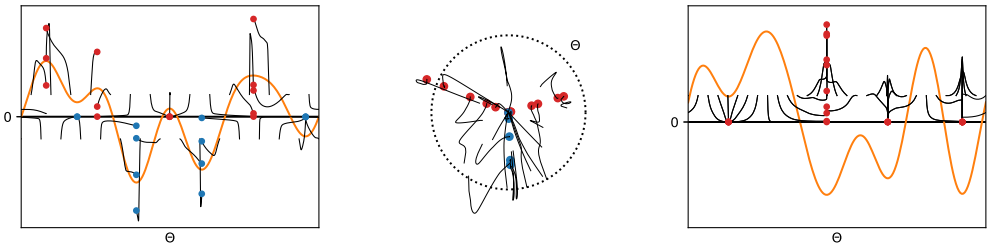
1 Introduction

Finding parsimonious descriptions of complex observations is an important problem in machine learning and signal processing. In its simplest form, this task boils down to searching for an element in a Hilbert space $\mathcal{F}$ that is close to a certain $f_0 \in \mathcal{F}$ — the observations — and that is a linear combination of a few elements from a parameterized set $\{\phi(\theta)\}_{\theta\in\Theta} \subset \mathcal{F}$ — the parsimonious description. This can be formulated as a minimization problem where the linear combination is expressed through an unknown measure ν and the distance to f_0 is quantified using a smooth convex loss function $R : \mathcal{F} \to \mathbb{R}$, such as the square loss $R(f) = \frac{1}{2}\|f - f_0\|_{\mathcal{F}}^2$. The problem to solve is then

$$J^* \coloneqq \min_{\nu\in\mathcal{M}_+(\Theta)} J(\nu), \qquad J(\nu) := R\left(\int_\Theta \phi(\theta)\mathrm{d}\nu(\theta)\right) + \lambda\nu(\Theta) \tag{1}$$

where $\mathcal{M}_+(\Theta)$ is the set of nonnegative measures ν on the parameter space Θ with finite total mass $\nu(\Theta) < \infty$ and $\lambda > 0$ is the regularization strength. This formulation also covers minimization over signed measures with total variation regularization, by replacing Θ with the disjoint union of two copies of Θ where ϕ takes opposite values, see Appendix A. A large body of research has exhibited the favorable properties of minimizers of such problems [4, 23, 43] with a statistical or variational viewpoint, showing in particular that λ favors sparser solutions and increases stability as it gets larger, at the expense of introducing a stronger bias. The present paper deals with the optimization aspect: our goal is to design algorithms that return ϵ-accurate solutions with a guaranteed computational complexity. When the set Θ is a finite set, this is a finite dimensional

*CNRS, Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, 91405, Orsay, France.



(a) Sparse deconvolution (b) Two-layer ReLU neural net (c) Generic optimization

Figure 1: Three examples of conic particle gradient descents (Algorithm 1). Trajectories of particles in black and their limits in red (positive mass) or blue (negative mass).

convex optimization problem that is well understood [9, 5]. However, convex approaches are generally inefficient when Θ is a continuous space, such as a d-dimensional manifold, where the need to discretize the space leads to a complexity scaling as ϵ^{-d} in the accuracy ϵ. We consider the following setting:

> (A1) Θ is a compact d-dimensional Riemannian manifold without boundaries. The functions $\phi : \Theta \to \mathcal{F}$ and $R : \mathcal{F} \to \mathbb{R}_+$ are twice Fréchet differentiable, with locally Lipschitz second-order derivatives, and ∇R is bounded on sublevel sets.

The algorithm that we analyze in this paper is simple to describe: *initialize with a discrete measure and run gradient descent on the positions and weights of the particles.* We will see that when the problem (1) admits sparse solutions and is non-degenerate, this over-parameterized non-convex gradient descent has a complexity scaling as $\log(1/\epsilon)$ in the accuracy ϵ. We make the following contributions:

- In Section 2, we introduce the *conic particle gradient descent* algorithm to solve optimization problems in the space of measures and discuss several of its interpretations.
- In Section 3, we show under under certain non-degeneracy assumptions that there is a sublevel of J starting from which this algorithm converges exponentially fast to minimizers.
- In Section 4, we show that for suitable choices of gradient and initialization, this algorithm converges to global minimizers. The proof combines the result of Section 3 with an analysis of a perturbed mirror descent in the space of measures. The number of iterations required to reach an accuracy ϵ is polynomial in the characteristics of the problem and logarithmic in ϵ. In contrast, the required number of particles depends exponentially on the dimension d, which is unavoidable under our assumptions.
- We report results of numerical experiments in Section 5, where the various insights brought by our analysis about local and global behaviors are investigated.

1.1 Examples of applications

As the problem of finding the simplest linear decomposition over a continuous dictionary is a very natural one, problems of the form (1) appear in a large variety of situations, see [8] for an extensive list. In this paper, our numerical illustrations are focused on two applications, chosen for their practical importance and also because they illustrate the variety of behaviors that can

be encountered. We also mention a third example to emphasize on the extreme generality — and thus the intrinsic limits — of our analysis. These three cases are illustrated on Figure 1.

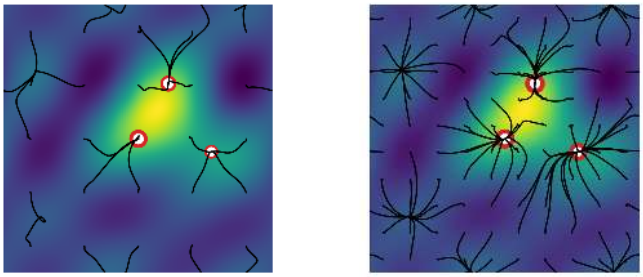
Sparse deconvolution. In this application, we want to recover a signal that consists of a mixture of spikes/impulses on Θ given a noisy and filtered observation f_0 in the space $\mathcal{F} = L^2(\Theta)$ of square-integrable real-valued functions on Θ. When one defines $\phi(\theta) : x \mapsto \psi(x - \theta)$ the translations of the filter impulse response ψ and R the squared loss, solving (1) allows to reconstruct the mixture of impulses with some guarantees, see e.g. [28, 23, 50]. In this typically low dimensional application, solving (1) to a high accuracy is crucial. Both the signed and nonnegative case have practical motivations (see Appendix A for how to handle the signed case). Figure 1-(a) illustrates the behavior of particle gradient descent for the signed case on the 1-torus, where the observed signal is shown in orange. Figure 2 illustrates the unsigned case on the 2-torus.

Two-layer neural networks. Here the goal is to select, within a specific class, a function that maps features in $\mathbb{R}^{d-1}$ to labels in $\mathbb{R}$ from the observation of a joint distribution of features and labels. This corresponds to $\mathcal{F}$ being the space of real-valued functions on $\mathbb{R}^{d-1}$ which are square-integrable under the distribution of features, R being e.g., the quadratic or the logistic loss function, and $\phi(\theta) : x \mapsto \sigma(\sum_{i=1}^{d-1} \theta_i x_i + \theta_d)$ with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Common choices are the sigmoid function or the rectified linear unit [35, 33]. In this application, d is typically large and it is not clear yet how to verify the non-degeneracy assumptions a priori, so our global convergence bounds are not useful. Still, the local analysis in Section 3 gives insights on the local behavior in the over-parameterized regularized setting and explains well the behavior observed in numerical experiments. With the ReLU activation, the method we analyze boils down to the classical gradient descent algorithm, see the remark in Section 2.2 about the 2-homogeneous case. Figure 1-(b) illustrates this case, by plotting the trajectories of $|a_i| \cdot b_i \in \mathbb{R}^2$ where $a_i \in \mathbb{R}$ is the output weight of neuron i and $b_i \in \mathbb{R}^2$ its hidden weights (the color represents the sign of a_i).

Non-convex optimization. Lastly, the minimization of any smooth function on a manifold $\phi : \Theta \to \mathbb{R}$ is covered by (1), as proved in Appendix B. For this problem, our algorithm is analogous to running independently several gradient-based minimization with diverse initializations, because the various particles simply follow the gradient field of ϕ and only interact through their masses. This case is illustrated on Figure 1-(c) where the function to minimize (here on the 1-torus) is plotted in orange. We recover the standard fact that random search as to be complemented with local search if one wants complexity that is reasonable in the precision. We stress that this is not the situation that motivates our analysis. Instead, we are interested in the case of general interactions between the particles, which is when we obtain novel insights.

1.2 Related work

Sparse optimization on measures. Problems with the structure (1) have a long history in optimization when Θ is discrete, and is typically solved with ISTA [22], mirror descent [47, 6] or variants of those algorithms. When Θ is continuous, the one dimensional case can sometimes be dealt with specific algorithms [13, 15]. In higher dimensions, the classical algorithms are conditional gradient algorithms (also known as Franck-Wolfe) [11, 25, 8], moment methods [24, 14, 27] and adaptive sampling/exchange algorithms [30, 29]. Often, these algorithms are complemented with non-convex updates on the particle positions, which considerably improves their behavior. Given an initial condition that is close to the optimum and with the same structure (i.e. without over-parameterization), the local convergence for non-convex gradient descent is studied in [56, 29].



(a) $m = 25$ particles (b) $m = 100$ particles

Figure 2: Illustration of 2D sparse deconvolution with conic particle gradient descent (Algorithm 1). Ground truth spikes in white, spatial trajectories in black and final location/mass in red.

Wasserstein gradient flows for optimization. The dynamics of two-layer neural networks optimization when the number of hidden units grows unbounded is studied in [48, 17, 45, 52, 54]. This series of work has led to various insights related to stochastic fluctuations and global convergence. The present paper can be seen as a quantitative counterpart to [17], although we consider a more restrictive setting[1]. A global rate of convergence is obtained in [60] but for a modified dynamic where particles are re-sampled at each iteration. Instead, we focus on the basic case where particles are only sampled once at the beginning of the algorithm. It should be mentioned that our analysis is different from the line of research on lazy over-parameterized models [18] initiated by [26, 36], which does not apply to the regularized case and to the unsigned case. Finally, in the parametric case where the unknown measure is assumed to belong to a finite dimensional probability model, Wasserstein natural gradient [2, 41, 16] or accelerated versions [58] have been proposed. Our analysis is however of non-parametric nature because the number of parameters is not fixed a priori in the analysis.

Related techniques. Our framework involves the theory of optimization on manifolds [1] and of Wasserstein gradient flows [3]. Some inspiration and interpretations of the algorithm under consideration come from unbalanced optimal transport theory [42, 38, 19] and in particular, from the lifting construction in [42]. Finally, our local analysis includes a functional and a gradient Łojasiewicz inequality of order 2 in Wasserstein space. Such inequalities were studied in [34, 7] for displacement convex functions, which does not cover our setting.

1.3 Notation

The set of signed (resp. nonnegative) finite Borel measures on a metric space $(\mathcal{X}, \mathrm{dist})$ is denoted by $\mathcal{M}(\mathcal{X})$ (resp. $\mathcal{M}_+(\mathcal{X})$). The relative entropy, a.k.a. Kullback-Leibler divergence, is defined for $\nu_1, \nu_2 \in \mathcal{M}_+(\mathcal{X})$ as $\mathcal{H}(\nu_1, \nu_2) = \int_{\mathcal{X}} \log(\mathrm{d}\nu_1/\mathrm{d}\nu_2)\mathrm{d}\nu_1 - \nu_1(\mathcal{X}) + \nu_2(\mathcal{X})$ if ν_1 is absolutely continuous w.r.t. ν_2, and $+\infty$ otherwise. The p-Wasserstein distance on the set $\mathcal{P}_p(\mathcal{X})$ of probability measures with finite p-th moment is defined, for $\mu_1, \mu_2 \in \mathcal{P}_p(\mathcal{X})$ as

$$W_p(\mu_1, \mu_2) = \left(\min_{\gamma \in \Pi(\mu_1, \mu_2)} \int \mathrm{dist}(x_1, x_2)^p \mathrm{d}\gamma(x_1, x_2) \right)^{1/p}$$

[1]The algorithm we study in this paper corresponds to the "2-homogeneous case" in [17]. Also, [17] allows non-smooth regularizers and does not require non-degeneracy.

where $\Pi(\mu_1, \mu_2)$ is the set of measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ_1 and μ_2. The distance W_∞ between compactly supported probabilities is defined as the limit of W_p as $p \to \infty$ and can be directly defined as $W_\infty(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \max_{(x_1, x_2) \in \operatorname{spt} \gamma} \operatorname{dist}(x_1, x_2)$ [53]. We also define the Bounded-Lipschitz norm for a continuous function $\psi : \mathcal{X} \to \mathbb{R}$ as $\|\psi\|_{\mathrm{BL}} = \|\psi\|_\infty + \mathrm{Lip}(\psi)$ where $\mathrm{Lip}(\psi)$ is the Lipschitz constant of ψ and its dual norm on $\mathcal{M}(\mathcal{X})$ as $\|\nu\|_{\mathrm{BL}}^* \coloneqq \sup_{\|\varphi\|_{\mathrm{BL}} \leq 1} \int \varphi \, \mathrm{d}\nu$. For a Riemannian manifold Θ, we denote by $T_\theta \Theta$ the tangent space of Θ at θ and by $\langle \cdot, \cdot \rangle_\theta : (T_\theta \Theta)^2 \to \mathbb{R}_+$ the metric at θ.

2 Particle gradient descent

2.1 General case

Consider a smooth increasing bijection $h : \mathbb{R}_+ \to \mathbb{R}_+$ (such as a power function $r \mapsto r^p$) and a number of particles $m \in \mathbb{N}^*$. The idea behind particle gradient-based algorithms is to parameterize the unknown measure ν as $\frac{1}{m} \sum_{i=1}^m h(r_i) \delta_{\theta_i}$ and to perform gradient-based optimization on the corresponding objective

$$F_m((r_1, \theta_1), \dots, (r_m, \theta_m)) \coloneqq R\left(\frac{1}{m} \sum_{i=1}^m h(r_i) \phi(\theta_i) \right) + \frac{\lambda}{m} \sum_{i=1}^m h(r_i), \tag{2}$$

where the parameters (r_i, θ_i) of each particle belong to $\Omega \coloneqq \mathbb{R}_+ \times \Theta$ endowed with a specific choice of metric. Clearly, if J admits a minimizer that is a mixture of $m^\star$ atoms with $m^\star \leq m$, then it is sufficient to minimize F_m from Eq. 2 for solving (1). While (2) is finite dimensional, it is typically non-convex with possibly some strict local minima. Still, when R is convex and for $h(r) = r^p$ for $p \in \{1, 2\}$, the message from [17] (see Theorem 2.2) is that solving (2) to global optimality with first-order methods is still possible by using over-parameterization, i.e. choosing m much larger than $m^\star$. Such a method involve various key hyper-parameters which role is discussed throughout the paper. They include (i) the choice of the function h (ii) the choice of the metric on Ω^m and (iii) the choice of the initialization.

Expression of the gradient. Under (A1), the objective J, seen as a function on the space $\mathcal{M}(\Theta)$ endowed with the total variation norm, is Fréchet-differentiable. Its differential at $\nu \in \mathcal{M}(\Theta)$ can be represented by the function $J'_\nu : \Theta \to \mathbb{R}$ given by

$$J'_\nu(\theta) = \left\langle \phi(\theta), \nabla R\left(\int_\Theta \phi(\theta) \mathrm{d}\nu(\theta) \right) \right\rangle_{\mathcal{F}} + \lambda, \tag{3}$$

in the sense that for any $\sigma \in \mathcal{M}(\Theta)$, it holds $\frac{d}{d\epsilon} J(\nu + \epsilon\sigma)|_{\epsilon=0} = \int_\Theta J'_\nu(\theta) \mathrm{d}\sigma(\theta)$. Now, consider a metric on $(\Omega^*)^m$ that is the average $(1/m) \sum_{i=1}^m \langle \cdot, \cdot \rangle_{(r_i, \theta_i)}$ of metrics on each factor $\Omega^* \coloneqq \mathbb{R}_+^* \times \Theta$, where $\mathbb{R}_+^*$ is the set of positive real numbers, of the form

$$\langle (\delta r_1, \delta\theta_1), (\delta r_2, \delta\theta_2) \rangle_{(r,\theta)} = \alpha(r)^{-1} \delta r_1 \delta r_2 + \beta(r)^{-1} \langle \delta\theta_1, \delta\theta_2 \rangle_\theta \tag{4}$$

where α and β are smooth functions $\mathbb{R}_+^* \to \mathbb{R}_+^*$ to be specified[2], $(r, \theta) \in \Omega^*$, $\delta r_1, \delta r_2 \in \mathbb{R}$ and $\delta\theta_1, \delta\theta_2 \in T_\theta \Theta$. Using the fact that gradients are characterized by the relation $\mathrm{d}F_m(x)(\delta x) = \langle \nabla F_m(x), \delta x \rangle$, we get that the gradient of F_m is given, in components, by

$$\begin{cases} \nabla_{r_i} F_m((r_i, \theta_i)_{i=1}^m) = \alpha(r_i) h'(r_i) J'_\nu(\theta_i) \\ \nabla_{\theta_i} F_m((r_i, \theta_i)_{i=1}^m) = \beta(r_i) h(r_i) \nabla J'_\nu(\theta_i) \end{cases} \qquad \text{where} \qquad \nu = \frac{1}{m} \sum_{i=1}^m h(r_i) \delta_{\theta_i}. \tag{5}$$

[2]Extension of the metric and gradients to the whole of Ω can be made on a case by case basis, see Section 2.2.

Lifted problem in Wasserstein space. Assume now that h has at most quadratic growth, and that the metric is defined on the whole of Ω. One can then see the discrete problem (2) as a discretization of a problem on the space $\mathcal{P}_2(\Omega)$ of probability measures on Ω with finite second moment endowed with the Wasserstein-2 metric given by

$$F^\star = \min_{\mu \in \mathcal{P}_2(\Omega)} F(\mu) \quad \text{where} \quad F(\mu) \coloneqq \left(\int_\Omega h(r)\phi(\theta) \mathrm{d}\mu(r,\theta) \right) + \lambda \int_\Omega h(r) \mathrm{d}\mu(r,\theta). \tag{6}$$

This point of view leads to insights on the properties of F_m that are independent of m, which is crucial for our theoretical analysis. For a measure $\mu \in \mathcal{P}_2(\Omega)$, we define following [42] the *homogeneous projection* operator $\mathsf{h} : \mathcal{P}_2(\Omega) \to \mathcal{M}_+(\Theta)$ where $\mathsf{h}\mu$ is characterized by

$$\int_\Theta \varphi(\theta) \mathrm{d}(\mathsf{h}\mu)(\theta) = \int_\Omega h(r)\varphi(\theta) \mathrm{d}\mu(r,\theta)$$

for any continuous function $\varphi : \Theta \to \mathbb{R}$. With this operator, we simply have $F(\mu) = J(\mathsf{h}\mu)$.

Gradient flow. There are various ways to optimize (2) with first order methods. Instead of directly focusing on a specific method, we first consider the gradient flow of F_m, as it is known that (stochastic) gradient descent [32, 40] approximates this dynamics. Let us call $x = (r_i, \theta_i)_{i=1}^m \in \Omega^m$ the variable of F_m. A gradient flow of F_m is an absolutely continuous curve $(x(t))_{t \geq 0}$ in Ω^m that satisfies

$$x'(t) = -\nabla F_m(x(t))$$

for $t \geq 0$, with the gradient given in Eq. (5). Note that if $h'(r)\alpha(r)^{-1}$ does not tend to 0 as $r \to 0$, then the non-negativity constraint on r should be explicitly enforced, which requires the notion of subgradient flows, see [17] for details in our setting.

Wasserstein gradient flow. It is also possible to directly study the optimization dynamics in the space $\mathcal{P}_2(\Omega)$ for the functional F of Eq. (6). For a measure $\nu \in \mathcal{M}_+(\Theta)$, consider the vector field on Ω with expression

$$g_\nu(r,\theta) = \big(\alpha(r)h'(r)J'_\nu(\theta), \beta(r)h(r)\nabla J'_\nu(\theta)\big) \in \mathbb{R} \times T_\theta\Theta.$$

We refer to $g_{\mathsf{h}\mu}$ as the Wasserstein gradient of F at μ (this notation emphasizes that it only depends on μ through $\mathsf{h}\mu$). Gradient flows of F_m are particular cases of *Wasserstein gradient flows* of F. The latter are defined as the absolutely continuous curves $(\mu_t)_{t \geq 0}$ in $\mathcal{P}_2(\Omega)$ that satisfy

$$\partial_t \mu_t = \mathrm{div}(\mu_t g_{\mathsf{h}\mu_t}) \tag{7}$$

in the weak sense, which means that for any differentiable function $\varphi : \Omega \to \mathbb{R}$, it holds $\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \varphi \, \mathrm{d}\mu_t\right) = -\int \nabla\varphi \cdot g_{\mathsf{h}\mu_t} \mathrm{d}\mu_t$, for almost every $t \geq 0$, see [53]. This is a proper extension of the notion of gradient flow for F_m in the sense that if $x(t) = (r_i(t), \theta_i(t))_{i=1}^m$ is a gradient flow of F_m then it can be directly checked that $t \mapsto \mu_t = \frac{1}{m}\sum_{i=1}^m \delta_{(r_i(t),\theta_i(t))}$ is a Wasserstein gradient flow of F.

2.2 The conic case

As seen in Eq. (5), the choice of the homogeneity degree and of the metric on Ω determine a specific way to combine the *vertical* and the *spatial* components of the gradient (along the variable r and θ, respectively). From now on, we focus on what we refer to as the *conic case*, which corresponds to the following assumption:

(A2) The mass parameterization is $h(r) = r^2$ and the metric on Ω^* is of the form Eq. (4) with $(\alpha(r), \beta(r)) = (\alpha, \beta/r^2)$ for some $\alpha, \beta > 0$.

The corresponding geodesic distance is $\mathrm{dist}((r_1, \theta_2), (r_1, \theta_2))^2 = r_1^2 + r_2^2 - 2r_1 r_2 \cos_\pi(\mathrm{dist}(\theta_1, \theta_2))$ where $\cos_\pi(z) = \cos(\min\{\pi, z\})$. This metric can be extended as a proper metric on $\widetilde{\Omega}$, defined as the set Ω where the subset $\{0\} \times \Theta$ is identified to a single point, known as the *cone metric*, which is the canonical way to define a metric on $\tilde{\Omega}$ [12]. In our context, identifying $\{0\} \times \Theta$ to a single point is desirable because a particle located in this set is a "dead" particle carrying no mass.

Plugging the metric into Eq. (5) gives the gradient (extended by continuity to $\{0\} \times \Theta$)

$$\begin{cases} \nabla_{r_i} F_m((r_i, \theta_i)_{i=1}^m) = 2\alpha r_i J'_\nu(\theta_i) \\ \nabla_{\theta_i} F_m((r_i, \theta_i)_{i=1}^m) = \beta \nabla J'_\nu(\theta_i) \end{cases} \qquad \text{where} \qquad \nu = \frac{1}{m} \sum_{i=1}^m r_i^2 \delta_{\theta_i},$$

and the Wasserstein gradient is represented by the vector field

$$g_\nu(r, \theta) = \big(2\alpha r J'_\nu(\theta), \beta \nabla J'_\nu(\theta)\big) \in \mathbb{R} \times T_\theta \Theta, \forall (r, \theta) \in \Omega. \tag{8}$$

Existence of Wasserstein gradient flows under (A1-2), for any initialization in $\mathcal{P}_2(\Omega)$ can be proved along the same lines as in [17], see details in Appendix C.1. Abstracting away its geometric derivation, the important aspects about our choice of gradient (8) are that its leads updates in r which are multiplicative and updates in θ which are independent of r. These two properties are crucial for our local convergence analysis (Section 3). Moreover, multiplicative updates enjoy favorable convergence rates (Section 4). The resulting structure and dynamics admits several interpretations.

Transport-growth interpretation. First, the projection $\nu_t = \mathsf{h}\mu_t$ of the gradient flow solves an advection-reaction equation. Importantly, this dynamics depends on μ_t only via the initialization $\mathsf{h}\mu_0$, which is a property specific to the conic setting.

Proposition 2.1. *Under* (A1-2)*, let $(\mu_t)_{t \geq 0}$ be a Wasserstein gradient flow for F, with $\mu_0 \in \mathcal{P}_2(\Omega)$. Then $\nu_t = \mathsf{h}\mu_t$ satisfies (in the weak sense)*

$$\partial_t \nu_t = -4\alpha \nu_t J'_{\nu_t} + \beta \mathrm{div}(\nu_t \nabla J'_{\nu_t}). \tag{9}$$

Proof. For any differentiable function $\varphi : \Theta \to \mathbb{R}$, since μ_t is a Wasserstein gradient flow it holds

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\int \varphi \, \mathrm{d}\nu_t \right) = - \int \langle \nabla(\mathsf{h}^* \varphi), g_{\mathsf{h}\mu_t} \rangle_{(r,\theta)} \mathrm{d}\mu_t = - \int (4\alpha \varphi J'_{\nu_t} + \beta \nabla \varphi \cdot \nabla J'_{\nu_t}) \mathrm{d}\nu_t,$$

which is the definition of weak solutions for (9). □

When $\beta = 0$, we recover the gradient flow of J for the Fisher-Rao (or Hellinger) metric, which also corresponds to continuous time mirror descent on $\mathcal{M}_+(\Theta)$ for the entropy mirror map [39]. When $\alpha = 0$, this is the gradient flow of J for the Wasserstein metric [3]. When $\alpha, \beta > 0$, this is the gradient flow of the functional J for the Wasserstein-Fisher-Rao metric, a.k.a. Hellinger-Kantorovich metric, see e.g. [31]. Under Assumption (A2), the dynamics (7) and (9) are directly related by Proposition 2.1. In the rest of this paper, we present the statements in terms of the projected dynamics ν_t, although they also could be stated in terms of μ_t. Note that an alternative discretization of the dynamic (9) was proposed in [51] using particle birth-death.

Spherical coordinates interpretation. Consider the case when $\Theta = \mathbb{S}^d$ is the d-dimensional sphere in $\mathbb{R}^{d+1}$. Then, the space $\widetilde{\Omega}$ endowed with the cone metric and $\mathbb{R}^{d+1}$ are isometric, through the spherical to Euclidean change of coordinates $(r, \theta) \mapsto r\theta$. Identifying $\widetilde{\Omega}$ with $\mathbb{R}^{d+1}$ through this isometry, the class of functions of the form $r^p \phi(\theta)$ on $\widetilde{\Omega}$, for $p > 0$ is simply the class of p-homogeneous functions on $\mathbb{R}^{d+1}$.

It follows that the conic setting we consider boils down, when $\Theta = \mathbb{S}^d$ and $p = 2$, to objectives defined on $\mathcal{P}_2(\mathbb{R}^{d+1})$ of the form

$$F(\mu) = R\left(\int_{\mathbb{R}^{d+1}} \psi(u) \mathrm{d}\mu(u)\right) + \lambda \int_{\mathbb{R}^{d+1}} \|u\|_2^2 \mathrm{d}\mu(u) \tag{10}$$

with $\psi : \mathbb{R}^{d+1} \to \mathcal{F}$ positively 2-homogeneous. Moreover, the Wasserstein gradient on $\mathcal{P}_2(\widetilde{\Omega})$ with the cone metric can be identified with the Wasserstein gradient on $\mathcal{P}_2(\mathbb{R}^{d+1})$ with the Euclidean metric. One can thus understand our choice of conic metric and $p = 2$ as a way to emulate the structure of 2-homogeneous problems on $\mathbb{R}^{d+1}$ in more general situations.

Asymptotic global convergence. Let us recall the global convergence result of [17, Thm. 3.3], in our setting and notations. We give in Appendix C.1 a simplified proof, enabled by our stronger smoothness assumptions.

Theorem 2.2. *Under* (A1-2)*, assume that R is convex, that ϕ is d-times continuously differentiable, that $\nu_0 \in \mathcal{M}_+(\Theta)$ has full support and that the projected gradient flow $(\nu_t)_{t \geq 0}$ converges weakly to some $\nu_\infty \in \mathcal{M}_+(\Theta)$. Then ν_∞ is a global minimizer of J.*

This theorem can be understood as a consistency result for conic particle gradient descent. It also raises several questions: under which conditions does ν_∞ exist? Can we guarantee a convergence rate ? Can we relax the full support condition on the initialization? In this paper, we answer positively to these questions in the particular case of non-degenerate sparse problems.

2.3 Conic particle gradient descent algorithm

Cone compatible retractions. The definition of discrete gradient descent in a Riemannian setting requires to introduce the notion of retraction. In general, a retraction on a Riemannian manifold $\mathcal{M}$ with tangent bundle $T\mathcal{M}$ is a smooth map $\mathrm{Ret} : T\mathcal{M} \to \mathcal{M}$ such that its restriction Ret_x to $T_x\mathcal{M}$ satisfies $\mathrm{Ret}_x(0) = x$ and $\mathrm{dRet}_x(0) = \mathrm{id}_{T_x\mathcal{M}}$, see [1, Def. 4.1.1]. In our case, we need to slightly adapt the definition to deal with the cone structure.

Definition 2.3. *We say that $\mathrm{Ret} : \Omega \times (\mathbb{R} \times T\Theta) \to \Omega$ is a retraction compatible with the cone structure, if it satisfies the following:*

(i) (Retraction property) *It is a proper retraction on $\Omega^* := \mathbb{R}_+^* \times \Theta$. It is not necessarily defined everywhere but there exists $C > 0$ such that $\mathrm{Ret}_{(r,\theta)}(\delta r, \delta\theta)$ is defined as long as $\max\{|\delta r|/r, \|\delta\theta\|_\theta\} < C$.*

(ii) (Zero preserving) *It satisfies $\mathrm{Ret}_{(0,\theta)}(\delta r, \delta\theta) = (0, f(\theta, \delta\theta))$ for some arbitrary measurable f.*

(iii) (Homogeneity) *For any $r, \tilde{r} \in \mathbb{R}_+^*$, $\theta \in \Theta$, $\delta r \in \mathbb{R}$ and $\delta\theta \in T_\theta\Theta$ satisfying $\max\{|\delta r|/r, \|\delta\theta\|_\theta\} < C$, denoting $(r_1, \theta_1) = \mathrm{Ret}_{(r,\theta)}(r\delta r, \delta\theta)$ and $(r_2, \theta_2) = \mathrm{Ret}_{(\tilde{r},\theta)}(\tilde{r}\delta r, \delta\theta)$, then $\theta_1 = \theta_2$ and $\tilde{r} \cdot r_1 = r \cdot r_2$.*

These properties are satisfied in the following examples, where $\widetilde{\mathrm{Ret}}$ denotes any retraction defined on Θ (we give them names for future reference):

– the *canonical* retraction $\mathrm{Ret}_{(r,\theta)}(\delta r, \delta\theta) = (r + \delta r, \widetilde{\mathrm{Ret}}_\theta(\delta\theta))$ (here $C = 1$);

– the *mirror* retraction $\mathrm{Ret}_{(r,\theta)}(\delta r, \delta\theta) = (r\exp(\delta r/r), \widetilde{\mathrm{Ret}}_\theta(\delta\theta))$, which allows to recover a version of mirror descent when $\delta\theta = 0$ (here $C = +\infty$);

– the *induced* retraction when Θ is the d-sphere, which is the retraction induced by the isometric embedding into $\mathbb{R}^{d+1}$, see Section 2.2. It is defined as $\mathrm{Ret}_{(r,\theta)}(\delta r, \delta\theta) = (\|u\|, u/\|u\|)$ where $u = r\theta + \theta\delta r + r\delta\theta \in \mathbb{R}^{d+1}$ (here $C = 1$). With this retraction, the iterates of gradient descent on Ω with the cone metric can be identified with the iterates of (Euclidean) gradient descent in $\mathbb{R}^{d+1}$.

Gradient descent in $\mathcal{P}_2(\Omega)$. Given a retraction Ret compatible with the cone structure, we define the gradient descent as follows. Let $\mu_0 \in \mathcal{P}_2(\Omega)$ and for $k \in \mathbb{N}$ define recursively

$$\mu_{k+1} = (T_k)_\# \mu_k \tag{11}$$

where $T_k(r,\theta) = \mathrm{Ret}_{(r,\theta)}(-2\alpha r J'_{\nu_k}(\theta), -\beta\nabla J'_{\nu_k}(\theta))$ and $\nu_k = \mathsf{h}\mu_k$. The notation # stands for the pushforward operator[3]. When μ_0 is a finite discrete probability measure with uniform weights, this gives Algorithm 1, which is a gradient descent for F_m in the cone metric.

Algorithm 1 Conic Particle Gradient Descent

1. let α and β be positive step-sizes and Ret a retraction on Ω compatible with the cone structure (Definition 2.3).
2. define an initial distribution of m particles weights-locations $(r_i^{(0)}, \theta_i^{(0)})_{i=1}^m$.
3. define for $k = 0, 1, \ldots$ until a stopping criterion is satisfied

$$(r_i^{(k+1)}, \theta_i^{(k+1)}) \leftarrow \mathrm{Ret}_{(r_i^{(k)}, \theta_i^{(k)})}(-2\alpha r_i^{(k)} J'_{\nu^{(k)}}(\theta_i^{(k)}), -\beta\nabla J'_{\nu^{(k)}}(\theta_i^{(k)})) \text{ for } i \in \{1, \ldots, m\}$$

where J'_ν is given by Eq. (3) and $\nu^{(k)} \coloneqq \frac{1}{m}\sum_{j=1}^m (r_j^{(k)})^2 \delta_{\theta_j^{(k)}}$.

Transport-growth interpretation. Just like the continuous-time gradient flow, the discrete time gradient descent has a corresponding projected dynamics in $\mathcal{M}_+(\Theta)$. Here the equivalence also relies on the properties of compatible retractions.

Proposition 2.4. *Under* (A1-2)*, let* Ret *be a retraction compatible with the cone structure and let $\mu_{k+1} = (T_k)_\#\mu_k$ for some $\mu_k \in \mathcal{P}_2(\Omega)$. Let $(T_k^r(\theta), T_k^\theta(\theta)) \coloneqq T_k(1,\theta)$. Then, the projected iterates $(\nu_{k+1}, \nu_k) \coloneqq (\mathsf{h}\mu_{k+1}, \mathsf{h}\mu_k)$ satisfy*

$$\nu_{k+1} = (T_k^\theta)_\#((T_k^r)^2\nu_k). \tag{12}$$

Proof. First, remark that by Property (i) of Definition 2.3, T_k is well-defined if $\max\{\alpha,\beta\}$ is small enough and that $T_k \in L^2(\mu_k;\Omega)$ so $\mu_{k+1} \in \mathcal{P}_2(\Omega)$. For any continuous function $\psi : \Theta \to \mathbb{R}$, using Properties (ii)-(iii) of Definition 2.3, we get

$$\int \psi \mathrm{d}\nu_{k+1} = \int r^2\psi(\theta)\mathrm{d}((T_k)_\#\mu_k)(r,\theta) = \int (rT_k^r(\theta))^2\psi(T_k^\theta(\theta))\mathrm{d}\mu_k(r,\theta) = \int (T_k^r)^2(\psi\circ T_k^\theta)\mathrm{d}\nu_k$$

which proves the claim. □

[3]The pushfoward measure $T_\#\mu$ is characterized by $\int\psi\mathrm{d}(T_\#\mu) = \int(\psi\circ T)\mathrm{d}\mu$ for any continuous function ψ.

Descent property of conic particle gradient descent. The following lemma shows that, for sufficiently small step-sizes, the iterates (11) are well-defined and monotonously decrease the objective. As usual in optimization, this property is useful to convert results on gradient flows into results on gradient descent.

Lemma 2.5 (Descent property)**.** *Assume* (A1-2) *and let* Ret *be a retraction compatible with the cone structure (Definition 2.3). For any $J_{\max} \geq J^\star$, there exists $\eta_{\max} > 0$ such that if $\nu_0 \in \mathcal{M}_+(\Theta)$ satisfies $J(\nu_0) \leq J_{\max}$ then the gradient descent iteration with $\max\{\alpha, \beta\} \leq \eta_{\max}$ is well defined for all $k \geq 0$ and satisfies*

$$J(\nu_{k+1}) - J(\nu_k) \leq -\frac{1}{2}\|g_{\nu_k}\|^2_{L^2(\nu_k)} \quad \textit{where} \quad \|g_\nu\|^2_{L^2(\nu)} \coloneqq \int \left(4\alpha|J'_\nu(\theta)|^2 + \beta\|\nabla J'_\nu(\theta)\|^2_\theta\right) \mathrm{d}\nu(\theta).$$

Proof. Let us first look at one step starting from $\nu_k \in \mathcal{M}_+(\Theta)$. By Property (i) of Definition 2.3, there exists $\eta_{\max} > 0$ such that this iteration is well-defined as long as $\max\{\alpha, \beta\} \leq \eta_{\max}$. We first consider $\nu_k(\Theta)$ and $\|J'_{\nu_k}\|_{\mathcal{C}^2}$ as constants, where $\|\phi\|_{\mathcal{C}^2} = \max\{\|\phi\|_\infty, \|\nabla\phi\|_\infty, \|\nabla^2\phi\|_\infty\}$ (we will see later that they can be upper bounded independently of the iteration k). With the notations of Proposition 2.4, we have $\nu_{k+1} = (T^\theta_k)_\#((T^r_k)^2\nu_k)$ where $T^r_k(\theta) = 1 - 2\alpha J'_{\nu_k}(\theta) + O(\alpha^2 J'_{\nu_k}(\theta)^2)$ and, in normal coordinates, $T^\theta_k(\theta) = \theta - \beta\nabla J'_{\nu_k}(\theta) + O(\beta^2\|\nabla J'_{\nu_k}(\theta)\|^2)$ where the hidden constants are uniform in θ. It follows that for any twice continuously differentiable $\psi \in \mathcal{C}^2(\Theta; \mathbb{R})$, it holds

$$\begin{aligned}\int \psi \mathrm{d}(\nu_{k+1} - \nu_k) &= \int ((T^r_k(\theta))^2 - 1)\psi(T^\theta_k(\theta))\mathrm{d}\nu_k(\theta) + \int (\psi(T^\theta_k(\theta)) - \psi(\theta))\mathrm{d}\nu_k(\theta) \\ &= -\int \left(4\alpha\psi \cdot J'_{\nu_k} + \beta\nabla\psi \cdot \nabla J'_{\nu_k}\right)\mathrm{d}\nu_k + \|\psi\|_{\mathcal{C}^2}\|g_{\nu_k}\|^2_{L^2(\mu_k)} O\left(\max\{\alpha, \beta\}\right).\end{aligned}$$

In particular, using this expression with $\psi_f(\theta) = \langle\phi(\theta), f\rangle$ where $\|f\| \leq 1$ (which have uniformly bounded norms $\|\psi_f\|_{\mathcal{C}^2}$ under our assumptions), we get that

$$\left\|\int \phi \mathrm{d}(\nu_{k+1} - \nu_k)\right\|^2 = \sup_{\|f\|\leq 1} \left\|\int \psi_f \mathrm{d}(\nu_{k+1} - \nu_k)\right\|^2 = O(\max\{\alpha, \beta\}\|g_{\nu_k}\|^2_{L^2(\mu_k)}).$$

By a first order expansion of R, we have for $f, f' \in \mathcal{F}$, $R(f') - R(f) = \langle f' - f, \nabla R(f)\rangle + O(\|f' - f\|^2)$. Thus, using the expression of J'_ν from Eq. (3), it follows

$$\begin{aligned}J(\nu_{k+1}) - J(\nu_k) &= \left\langle \int \phi \mathrm{d}(\nu_{k+1} - \mathrm{d}\nu_k), \nabla R\Big(\int \phi \mathrm{d}\nu_k\Big)\right\rangle \\ &\qquad + \lambda \int (\mathrm{d}\nu_{k+1} - \mathrm{d}\nu_k) + O(\max\{\alpha, \beta\}\|g_{\nu_k}\|^2_{L^2(\mu_k)}) \\ &= \int J'_{\nu_k} \mathrm{d}(\nu_{k+1} - \nu_k) + O(\max\{\alpha, \beta\}\|g_{\nu_k}\|^2_{L^2(\mu_k)}) \\ &= (-1 + O(\max\{\alpha, \beta\}))\|g_{\nu_k}\|^2_{L^2(\mu_k)}.\end{aligned}$$

So there exists $\eta_{\max}$ such that if $\max\{\alpha, \beta\} \leq \eta_{\max}$, we have $J(\nu_{k+1}) - J(\nu_k) \leq -\frac{1}{2}\|g_{\nu_k}\|^2_{L^2(\mu_k)}$. Finally, since we have assumed that $\lambda > 0$ and ∇R is bounded on sublevel sets, the quantities $\sup_{J(\nu)\leq J(\nu_k)} \nu(\Theta)$ and $\sup_{J(\nu)\leq J(\nu_k)} \|J'_\nu\|_{\mathcal{C}^2}$ are finite. By the decrease property we just proved, these quantities decrease after one iteration if $\max\{\alpha, \beta\} \leq \eta_{\max}$. So $\eta_{\max}$, which depends on these quantities, can be chosen independently of $k \geq 0$. □

3 Exponential local convergence

We now proceed to the theoretical analysis of the projected gradient flow (9) and projected gradient descent (12) in the conic setting. In light of Propositions 2.1 and 2.4, these dynamics correspond to the gradient flow and gradient descent of F, seen through the projection operator h.

3.1 Non-degeneracy assumptions

In order to derive global optimality conditions, we assume the following.

(A3) The loss R is convex.

Commonly used losses that satisfy the smoothness and convexity conditions are the square loss and the logistic loss. Under this assumption, we have existence of minimizers and a global optimality condition.

Proposition 3.1 (Optimality condition)**.** *Under* (A1) *and* (A3)*, problem* (1) *admits minimizers. Moreover, a measure $\nu^\star \in \mathcal{M}_+(\Theta)$ is a minimizer if and only if it holds $J'_{\nu^\star}(\theta) \geq 0$ for all $\theta \in \Theta$ and $J'_{\nu^\star}(\theta) = 0$ whenever θ in the support of $\nu^\star$.*

Proof. As λ is assumed positive, the sublevel sets of J on $\mathcal{M}_+(\Theta)$ are bounded in total variation, and are thus weakly pre-compact. It follows that any minimizing sequence for J admits at least one weak limit point $\nu^\star$, which is a minimizer of (1) since J is weakly continuous. The stated optimality condition is equivalent to having $\int J'_{\nu^\star} \mathrm{d}(\nu - \nu^\star) \geq 0$ for all $\nu \in \mathcal{M}_+(\Theta)$. The latter is a sufficient optimality condition since by convexity of J, $J(\nu) - J(\nu^\star) \geq \int J'_{\nu^\star} \mathrm{d}(\nu - \nu^\star)$. It is also necessary since it holds $\frac{d}{d\epsilon} J((1-\epsilon)\nu^\star + \epsilon\nu)|_{\epsilon=0^+} = \int J'_{\nu^\star} \mathrm{d}(\nu - \nu^\star)$. □

Sparse minimizer. Our local analysis requires sparsity of the minimizers of the objective J, which can be guaranteed a priori in several settings (e.g. [28, 10]).

(A4) Problem (1) admits a unique global minimizer on $\mathcal{M}_+(\Theta)$ which is of the form $\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \delta_{\theta_i}$ with $\nu^\star(\Theta) > 0$. We denote $f^\star \coloneqq \int \phi \mathrm{d}\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \phi(\theta_i)$.

Without loss of generality, we assume $r_i > 0$ for all i and $\theta_i \neq \theta_{i'}$ whenever $i \neq i'$, so that $(r_i, \theta_i)_{i=1}^{m^\star}$ is uniquely well-defined, up to re-ordering. Let us fix from now on normal coordinates frames on the neighborhood of each θ_i. This allows to identify tensors at θ_i with their expression in coordinates and also induces a set of coordinates on the direct sum of the tangent spaces $T_{\theta_i}\Theta$, which is of dimension $m^\star \times d$.

Kernels and non-degeneracy. We define the *global kernel* $K \in \mathbb{R}^{(m^\star \times (1+d))^2}$ by

$$K_{(i,j),(i',j')} \coloneqq \langle r_i \bar{\nabla}_j \phi(\theta_i), r_{i'} \bar{\nabla}_{j'} \phi(\theta_{i'}) \rangle_{d^2 R_{f^\star}}$$

where $\bar{\nabla}\phi \coloneqq (2\alpha\phi, \beta\nabla\phi)$ can be interpreted as the gradient of $\mathsf{h}\phi$ at $(1, \theta)$. Remark that K is defined via the quadratic form associated to the Hessian of R at $f^\star$. This interaction kernel K appears naturally in the various statistical and optimization analysis of the minimization problem under consideration [28, 56]. We also use the notation for the *local kernels* for $i \in 1, \dots, m^\star$

$$H_i \coloneqq \beta^2 \nabla^2 J'_{\nu^\star}(\theta_i)$$

expressed in local coordinates. In order to simplify notations, we concatenate these matrices in a large matrix $H \in \mathbb{R}^{(m^\star \times (1+d))^2}$ of the same size as K defined as

$$H_{(i,j),(i,j')} = \begin{cases} \beta^2 \nabla^2_{j,j'} J'_{\nu^\star}(\theta_i) & \text{if } i = i' \text{ and } j, j' \geq 1, \\ 0 & \text{if } j = 0 \text{ or } j' = 0. \end{cases}$$

where here and in the proofs, we use 0 to label the r's coordinate. The local analysis will be carried under the following non-degeneracy assumptions.

(A5) The minimizer $\nu^\star$ is non-degenerate in the sense that $\nabla^2 R(f^\star)$ is positive definite and, calling $\sigma_{\min}(A)$ the smallest singular value of a linear operator A, we have global curvature $\sigma_{\min}(K) > 0$, local curvature $\sigma_{\min}(H) = \min_i \sigma_{\min}(H_i) > 0$, and strict slackness, i.e. the only points where $J'_{\nu^\star}$ vanishes are $(\theta_i)_{i=1}^{m^\star}$.

The first property is always satisfied if R is strictly convex. The second property is satisfied when the kernel associated to the feature function $\bar{\nabla}\phi$ is positive definite. The last two assumptions unfortunately depend on an *a priori* unknown object $J'_{\nu^\star}$, but are often required to perform analysis of Problem (1) [29, 28]. Yet, in some cases, they can be guaranteed to hold, see e.g. [50, 55]. In spite of this drawback, the local analysis leads to interesting qualitative insights on the dynamics in practice, see Section 5.

3.2 Convergence in $\mathcal{M}_+(\Theta)$

A first consequence of these assumptions is that convergence in value implies convergence to minimizers. The distance on $\mathcal{M}_+(\Theta)$ that naturally appears in the analysis is the Wasserstein-Fisher-Rao, a.k.a. Hellinger-Kantorovich metric $\widehat{W}_2$, which is the extension of the Wasserstein W_2 metric to unnormalized measures. It admits many equivalent definitions [42, 38, 20], the most suitable to our context being [42, Thm. 7.20]

$$\widehat{W}_2(\nu_1, \nu_2) \coloneqq \min \left\{ W_2(\mu_1, \mu_2) \; ; \; (\mu_1, \mu_2) \in \mathcal{P}_2(\Omega)^2 \text{ satisfy } (\mathsf{h}\mu_1, \mathsf{h}\mu_2) = (\nu_1, \nu_2) \right\}$$

where the Wasserstein distance on Ω is defined relative to the cone metric (in this paragraph, with $\alpha = \beta = 1$). The proof of the following result involves the construction of a transport map in the lifted space $\mathcal{P}_2(\Omega)$ and is postponed to Appendix D.4.

Proposition 3.2. *Under* (A1-5)*, for all $J_{\max} \geq J^\star$, there exists $C, C' > 0$, such that if $\nu \in \mathcal{M}_+(\Theta)$ satisfies $J(\nu) \leq J_{\max}$ then $\|\nu - \nu^\star\|^*_{\mathrm{BL}} \leq C\widehat{W}_2(\nu, \nu^\star) \leq C'(J(\nu) - J^\star)^{\frac{1}{2}}$.*

3.3 Sharpness of the objective

Our first main result is a lower bound on the squared norm of the gradient in terms of the sub-optimality gap, an inequality known as sharpness, or Polyak-Łojasiewicz inequality [49, 37], which is a special case of Łojasiewicz gradient inequality. It involves the $L^2(\nu)$ norm of the gradient, which we denote for $\nu = \mathsf{h}\mu$ by

$$\|g_\nu\|^2_{L^2(\nu)} \coloneqq \int_\Omega \left(\frac{1}{\alpha} |2\alpha r J'_\nu(\theta)|^2 + \frac{r^2}{\beta} \|\beta \nabla J'_\nu\|^2_\theta \right) \mathrm{d}\mu(r, \theta) = \int_\Theta \left(4\alpha |J'_\nu(\theta)|^2 + \beta \|\nabla J'_\nu(\theta)\|^2_\theta \right) \mathrm{d}\nu(\theta).$$

Theorem 3.3 (Sharpness)**.** *Under* (A1-5)*, there exists $J_0 > J^\star$ and $\kappa_0 > 0$, such that for all $\nu \in \mathcal{M}_+(\Theta)$ satisfying $J(\nu) \leq J_0$ and $\alpha, \beta > 0$, one has*

$$\frac{1}{2} \|g_\nu\|^2_{L^2(\nu)} \geq \kappa_0 \min\{\alpha, \beta\} \, (J(\nu) - J^\star).$$

While the objective is non-convex in the Wasserstein geometry and has typically an infinity of bad stationary points, this inequality guarantees exponential convergence to global minimizers of various gradient-based dynamics as long as their initialization ν_0 has a small enough objective value. Crucially, the specific structure of ν does not matter, beyond the fact that is is close enough to optimality: it applies indifferently to discrete and absolutely continuous measures. Once Theorem 3.3 is established, it is straightforward to prove exponential convergence of gradient flow and gradient descent.

Corollary 3.4 (Local convergence of gradient flow)**.** *Under* (A1-5)*, let J_0 and κ_0 be given by Theorem 3.3. Consider $(\nu_t)_{t\geq 0}$ a projected gradient flow for J as in Eq.* (9)*. If $J(\nu_0) \leq J_0$ then*

$$J(\nu_t) - J^* \leq (J(\nu_0) - J^*) \exp(-2\kappa_0 \min\{\alpha, \beta\} t).$$

Proof. By Theorem 3.3 and direct computations, one has

$$\frac{d}{dt}(J(\nu_t) - J^\star) = \int_\Omega J'_{\nu_t} \mathrm{d}(\partial_t \nu_t) = -\|g_{\nu_t}\|^2_{L^2(\nu_t)} \leq -2\kappa_0 \min\{\alpha, \beta\}(J(\nu_t) - J^\star)$$

and the result follows by Grönwall's lemma. □

Corollary 3.5 (Local convergence of gradient descent)**.** *Assume* (A1-5)*, let J_0 and κ_0 be given by Theorem 3.3, and let* Ret *be a retraction compatible with the cone structure (Definition 2.3). There exists $\eta_{\max} > 0$ such that for any projected gradient descent $(\nu_k)_{k\geq 0}$ for J following recursion* (11)*, if $J(\nu_0) \leq J_0$ and $\max\{\alpha, \beta\} \leq \eta_{\max}$, then*

$$J(\mu_k) - J^\star \leq (J(\nu_0) - J^\star) \left(1 - \kappa_0 \min\{\alpha, \beta\}\right)^k.$$

Proof. By Lemma 2.5, there exists $\eta_{\max}$ such that if $\max\{\alpha, \beta\} \leq \eta_{\max}$, then $J(\nu_{k+1}) - J(\nu_k) \leq -\frac{1}{2}\|g_{\nu_k}\|^2_{L^2(\nu_k)}$. Combining this inequality with Theorem 3.3, one has $J(\nu_{k+1}) - J(\nu_k) \leq -\kappa_0 \min\{\alpha, \beta\}(J(\nu_k) - J^\star)$. Rearranging the terms, we get $J(\nu_{k+1}) - J^\star \leq (1 - \kappa_0 \min\{\alpha, \beta\})(J(\nu_k) - J^\star)$ and the result follows by recursion. □

3.4 Proof strategy for the sharpness theorem

The proof of Theorem 3.3, in Appendix D, is based on a local expansion of $J(\nu)$ in terms of some local moments of ν. For a radius $\tau > 0$ (that shall be fixed at some small enough value in the course of the proof), we define the sets for $i \in \{1, \dots, m^\star\}$,

$$\Theta_i := \{\theta \in \Theta \,;\, \mathrm{dist}(\theta, \theta_i) < \tau\}.$$

We assume that τ is smaller than 1 and small enough so that these sets together with $\Theta_0 := \Theta \setminus \cup_{i=1}^{m^\star} \Theta_i$ form a partition of Θ and that the exponential map at θ_i has injectivity radius larger than τ, for $i \in \{1, \dots, m^\star\}$. We then say that τ is an *admissible* radius.

Definition 3.6 (Local moments)**.** *Given an admissible radius $\tau > 0$ and a measure $\nu \in \mathcal{M}_+(\Theta)$, we define for $i \in \{0, \dots, m^\star\}$ the local masses $\bar{r}_i^2 = \nu(\Theta_i)$ and the local means $\bar{\theta}_i := \frac{1}{\bar{r}_i^2} \int_{\Theta_i} \theta \mathrm{d}\nu(\theta)$ if $\nu(\Theta_i) > 0$ and $\bar{\theta}_i = \theta_i$ otherwise. Finally, we define for $i \in \{1, \dots, m^\star\}$ the* weighted biases

$$b_i^r := \frac{\bar{r}_i^2 - r_i^2}{2\alpha r_i}, \qquad b_i^\theta := \frac{\bar{r}_i^2}{\beta r_i}(\bar{\theta}_i - \theta_i), \qquad b_i = (b_i^r, b_i^\theta),$$

and the weighted covariances $\Sigma_i := \frac{1}{\bar{r}_i^2 \beta^2} \int_{\Theta_i} (\theta - \bar{\theta}_i)(\theta - \bar{\theta}_i)^\intercal \mathrm{d}\nu(\theta)$.

If ν has only 1 atom in each Θ_i then its spatial coordinate is $\bar{\theta}_i$ and $\Sigma_i = 0$. When moreover $\nu(\Theta_0) = 0$, the optimization reduces to a more classical gradient flow in $\Omega^{m^\star}$ which local behavior has already been studied [56, 29], but obtaining measures of this form is typically almost as hard as solving the original problem. This decomposition can be reminiscent of proof techniques used to study log-Sobolev inequalities (another type of sharpness inequality in Wasserstein space [7]) in the small temperature regime [46].

It turns out that the local moments of Definition 3.6 are sufficient to characterize the behavior of J near optimality. In particular, we have the following approximations for J and its gradient around optimality. These formulas are obtained as an intermediate step in the proof of Theorem 3.3 and follow by combining the bounds of Proposition D.4 and Proposition D.5 with Lemma D.3.

Proposition 3.7 (Local expansion)**.** *Assuming* (A1-5)*, for any $\nu \in \mathcal{M}_+(\Theta)$ it holds*

$$J(\nu) - J^\star = \frac{1}{2} b^\intercal (K + H) b + \frac{1}{2} \sum_{i=1}^{m^\star} r_i^2 \operatorname{tr}(\Sigma_i H_i) + \int_{\Theta_0} J'_{\nu^\star} \mathrm{d}\nu + \operatorname{err}(\tau, J(\nu) - J^\star) \tag{13}$$

and

$$\frac{1}{2} \|g_\nu\|^2_{L^2(\nu)} = \frac{1}{2} b^\intercal (K + H)^2 b + \frac{1}{2} \sum_{i=1}^{m^\star} r_i^2 \operatorname{tr}(\Sigma_i H_i^2) + \frac{1}{2} \|g_\nu\|^2_{L^2(\nu|_{\Theta_0})} + \operatorname{err}(\tau, J(\nu) - J^\star) \tag{14}$$

where $\|g_\nu\|^2_{L^2(\nu|_{\Theta_0})} = \int_{\Theta_0} (\alpha |J'_\nu|^2 + \beta \|\nabla J'_\nu\|^2_\theta) \mathrm{d}\nu$ *and* $\operatorname{err}(\tau, \Delta) = O(\Delta\tau + \Delta^{\frac{3}{2}} \tau^{-6})$.

3.5 Discussion on the local behavior

Let us now explore what the expansion from Proposition 3.7 teaches us about the local behavior of the dynamics. In order to simplify the discussion, let us fix a small admissible radius τ_0 and ignore the error terms in Proposition 3.7.

Effect of over-parameterization. When there is no over-parameterization ($m = m^\star$) and we have a single particle in the neighborhood Θ_i of each optimal particle, then there is no local variance: $\Sigma_i = 0$ for $i = 1, \dots, m^\star$. In this case, we recover the Taylor expansion of $F_{m^\star}$ around its minimizer

$$J(\nu) - J^\star \approx \frac{1}{2} b^\intercal (K + H) b,$$

and the local convergence rate is dictated by the conditioning of $(K + H)$. Now, for an arbitrary over-parameterization i.e. $\nu \in \mathcal{M}_+(\Theta)$ but with the support of the solution approximately identified, i.e. $\nu(\Theta_0) = 0$, the objective is still entirely characterized locally by the local moments of ν, since

$$J(\nu) - J^\star \approx \frac{1}{2} b^\intercal K b + \frac{1}{2} \sum_{i=1}^{m^\star} b_i^\intercal H_i b_i + \frac{1}{2} \sum_{i=1}^{m^\star} r_i^2 \operatorname{tr}(\Sigma_i H_i).$$

This expression gives a clear picture of the energy landscape, so let us comment on it. If we think of the particles in Θ_i as a cluster, then the first term consists in a global interaction between the clusters, which only depends on the biases of each cluster relatively to their respective ground truth particles. The two other terms are local interactions within each cluster, which are due to the local curvature of $J'_{\nu^\star}$ at each θ_i. Note in particular that the only term in this expansion that penalizes the variance of each cluster Σ_i consists of local interactions.

Effect of the regularization parameter. In this paper, the assumption that λ is non-zero is not crucial as such. Instead the crucial assumption for the local analysis is (A5). Still, this assumption is intimately connected to the regularization: in the signed case (detailed in Appendix A), it is necessary that $\lambda > 0$ to have (A5), because with $\lambda = 0$, the minimizer $\nu^\star$ is a global minimizer in the space of signed measures and thus the global optimality condition $J'_{\nu^\star} = 0$ holds. In fact, a finer analysis of the behavior as $\lambda \to 0$ is possible in the signed case: it can be shown that one has (emphasizing the dependency in λ in the notation):

$$K_\lambda = K_0 + o(\lambda), \qquad H_\lambda = \lambda H_0 + o(\lambda), \qquad J'_{\nu*,\lambda} = \lambda J'_0 + o(\lambda),$$

for some K_0, H_0 and J'_0 [28, Prop. 1 and Thm. 2] (where the result is proved for R being the square loss but can be directly generalized to R smooth and strongly convex around the minimizer). Under the assumption that J'_0 is non-degenerate in the sense of (A5), as soon as $\nu(\Theta_0) > 0$ or $\Sigma_i \neq 0$ for some $i \in \{1, \dots, m^\star\}$, the local rate κ_0 is thus of order λ and for $\lambda = 0$, the exponential convergence rate is lost. This shows that regularization is necessary for fast local convergence in the signed case, and in particular – remembering the previous paragraph – for the variance of each cluster of particles to vanish quickly. Note that it is an open question to even show local convergence when (A5) does not hold.

Choice of the metric and conditioning. While our statements, in particular Corollary 3.5, seem to imply that it is best to choose $\alpha = \beta$, this is in fact just an artefact of the way the upper bounds are presented, with some hidden constants. Instead, these parameters should be chosen, as usual, so as to make the local expression of J above well-conditioned. Without additional information, a possible heuristic is to make the block diagonal matrix $\mathrm{diag}(K + H, H)$ well-conditioned by choosing (α, β) satisfying $2\alpha\|\phi\|_\infty \approx \beta \mathrm{Lip}(\phi)$.

Polynomial dependency. It can be seen from the proof of Theorem 3.3 that $(J_0 - J^\star)^{-1}$ and κ_0^{-1} depend polynomially on the characteristics of the problem, which are the regularization λ, the regularity parameters of ϕ and R, the ratio $\max_i r_i / \min_i r_i$, the inverses of the $\sigma_{\min}(\nabla^2 R(f^\star))$, $\sigma_{\min}(H)$, $\sigma_{\min}(K)$ and finally the quantity $v^\star$ that quantifies the strict slackness assumption, in the following sense: $v^* > 0$ is such that for any local minimum θ of $J'_{\nu^\star}$, either $\theta = \theta_i$ for some $i \in \{1, \dots, m^\star\}$ or $J'_{\nu^\star}(\theta) \geq v^*$.

4 Quantitative global convergence

There are several convex optimization-based algorithms that are known to return approximate minimizers of J which are mixtures of atoms (with typically $m > m^\star$) with a guaranteed complexity, see Section 1.2. Starting from any such approximate minimizer, the results of the previous section imply that conic particle gradient descent converges exponentially fast to minimizers of J. However, such a "two-algorithms" approach comes with a drawback: one has to decide when to switch from one algorithm to another. In this section, we show that it is possible to reach global optimality by only performing non-convex gradient descent. This is true under two main conditions: (i) the initialization samples Θ densely enough, and (ii) the ratio β/α is small, at least in the early stages of the algorithm.

4.1 Statement of the main results

In order to state the condition on the initialization, we first choose a reference measure $\rho \in \mathcal{M}_+(\Theta)$ with a smooth positive density, also denoted by ρ, which represents our prior knowledge about

the solution $\nu^\star$. We introduce the quantity (analogous to a log-likelihood)

$$\bar{\mathcal{H}}(\nu^\star, \rho) \coloneqq \sum_{i=1}^{m^\star} r_i^2 \log\left(\frac{r_i^2}{\rho(\theta_i)}\right) - \nu^\star(\Theta) + \rho(\Theta).$$

It quantifies how good is ρ as a prior for the unknown minimizer $\nu^\star$ and we will see that our convergence bounds are better when $\bar{\mathcal{H}}(\nu^\star, \rho)$ is smaller. If nothing is known about the optimal positions θ_i, we should choose ρ as a uniform density $\alpha\mathrm{vol}$ over Θ for some $\alpha > 0$. Minimizing $\bar{\mathcal{H}}(\nu^\star, \alpha\,\mathrm{vol})$ in α suggests to choose $\alpha = \frac{\nu^\star(\Theta)}{\mathrm{vol}(\Theta)}$.

To obtain an implementable algorithm, we then discretize ρ and consider an initialization $\nu_0 \in \mathcal{M}_+(\Theta)$ which is close to ρ in the W_∞ distance (our statements do not require ν_0 to be discrete but this is necessary to obtain an implementable algorithm). We now state our main theorem.

Theorem 4.1 (Global convergence of gradient flow)**.** *Under* (A1-5)*, let J_0 and κ_0 given by Theorem 3.3, let $\rho = \rho\mathrm{vol} \in \mathcal{M}_+(\Theta)$ an absolutely continuous reference measure with $\log\rho$ L-Lipschitz and let $B_{\nu_0} = \sup_{J(\nu)\leq J(\nu_0)} \|J'_\nu\|_{\mathrm{BL}}$, where $\nu_0 \in \mathcal{M}_+(\Theta)$ is the initialization. For any $0 < \epsilon \leq 1/2$, there exists $C_\epsilon > 0$ that only depend on ϵ and bounds on the curvature of Θ such that if it holds $\beta/\alpha \leq (4B_{\nu_0}/\max\{1, L\})^2$,*

$$W_\infty(\nu_0, \rho) \leq \frac{J_0 - J^\star}{2B_{\nu_0}\nu^\star(\Theta)} \qquad \textit{and} \qquad \frac{\beta}{\alpha} \leq \left(\frac{J_0 - J^\star}{2(4B_{\nu_0})^\epsilon\left(\bar{\mathcal{H}}(\nu^\star, \rho) + B_{\nu_0}^2 + C_\epsilon\nu^\star(\Theta)d\right)}\right)^{\frac{2}{1-\epsilon}}$$

then the projected gradient flow $(\nu_t)_{t\geq 0}$ initialized with ν_0 converges to the global minimizer $\nu^\star$. Denoting $t_0 = 1/\sqrt{\alpha\beta}$ it satisfies, for $t \geq t_0$,

$$J(\nu_t) - J^\star \leq (J_0 - J^*)\exp(-2\kappa_0\min\{\alpha, \beta\}(t - t_0)).$$

We also state a similar result for gradient descent, but without tracking the constants. The proof follows the same lines as that of Theorem 4.1 and is given in Appendix F.

Theorem 4.2 (Global convergence of gradient descent)**.** *Under* (A1-5)*, let J_0 and κ_0 be given by Theorem 3.3 and $\rho = \rho\mathrm{vol} \in \mathcal{M}_+(\Theta)$ an absolutely continuous reference measure with $\log\rho$ Lipschitz. For any $0 < \epsilon \leq 1/2$ and $\nu_0 \in \mathcal{M}_+(\Theta)$, there exists $C, C' > 0$ that depends on the characteristics of the problem and increasingly on $\bar{\mathcal{H}}(\nu^\star, \nu_0)$ and $1/\epsilon$, such that if*

$$W_\infty(\nu_0, \rho) \leq (J_0 - J^\star)/C, \qquad \alpha \leq (J_0 - J^*)^{1+\epsilon/2}/C, \qquad \textit{and} \qquad \beta \leq (J_0 - J^*)\alpha^2/C'$$

then the projected gradient descent $(\nu_k)_{k\in\mathbb{N}}$ initialized with ν_0 converges to the global optimum $\nu^\star$. Denoting $k_0 = C/(J_0 - J^*)^{2+\epsilon}$ it satisfies, for $k \geq k_0$,

$$J(\nu_k) - J^\star \leq (J(\nu_0) - J^\star)\left(1 - \kappa_0 \cdot \min\{\alpha, \beta\}\right)^{k-k_0}.$$

We can make the following comments:

– The non-asymptotic convergence rate does not appear explicitly in Theorem 4.1, because the result is obtained by trading-off various error terms. In an the idealized setting where $\nu_0 = \rho$ and $\beta = 0$, a direct consequence of Lemma 4.3 and Lemma E.1 is that $J(\nu_t) - J^\star$ decreases as $O(\log(t)/t)$ for the gradient flow and in $O(\log(k)/\sqrt{k})$ for the gradient descent in general. For the specific case of the mirror retraction, we show in Appendix G that a faster rate in $O(\log(k)/k)$ holds.

– The condition on the initialization can be achieved by taking ν_0 a weighted empirical distribution of m samples from ρ (typically the normalized volume measure), and it is known that the rate of convergence in W_∞ of such approximation is in $\tilde{O}(m^{-1/d})$, see [57]. Unfortunately, this exponential dependence in the dimension is unavoidable when approximating densities in Wasserstein distances [59]. This corresponds to a quantitative version of the condition on the initialization in Theorem 2.2. Also, note that $J_0 - J^\star$ gets smaller as the problem becomes more difficult, in which case the overparameterization m must increase, and the convergence speed slows down. In particular, the necessary condition $m \geq m^\star$ is implicitly implied by our assumptions.

– The fact that the sublevel J_0 from Theorem 3.3 does not depend on the metric parameters (α, β) is crucial to prove these theorems. However, the local exponential rate of convergence in Theorem 4.2 may be deceptively bad if β/α is extremely small. An natural fix is to start with a small ratio β/α as required by Theorem 4.2, and to increase this ratio at each iteration so as to improve the conditioning of J near optimality. The interest of Theorem 4.2 lies mostly in the qualitative insights it brings. In practice, we would advise to choose $W_\infty(\nu_0, \rho)$, α and β via heuristics or parameter search rather than trying to derive the constants of Theorem 4.2, which could be deceptively conservative.

4.2 Proof of global convergence for gradient flows

The proof of Theorem 4.1 mostly relies on the the following general lemma which applies to any type of initialization or any structure of minimizers. It gives an upper bound on the optimality gap during along gradient flows in terms of a *mirror rate* function $\mathfrak{Q}_{\nu_0,\nu^\star} : \mathbb{R}_+^* \to \mathbb{R}_+$ defined for $\nu^\star, \nu_0 \in \mathcal{M}_+(\Theta)$ and $\tau > 0$ as

$$\mathfrak{Q}_{\nu^\star,\nu_0}(\tau) = \inf_{\nu \in \mathcal{M}_+(\Theta)} \|\nu^\star - \nu\|_{\mathrm{BL}}^* + \frac{1}{\tau}\mathcal{H}(\nu, \nu_0). \tag{15}$$

This is a continuous and decreasing function of τ that satisfies

$$\lim_{\tau \to \infty} \mathfrak{Q}_{\nu^\star,\nu_0}(\tau) = \inf_{\operatorname{spt}\nu \subset \operatorname{spt}\nu_0} \|\nu^\star - \nu\|_{\mathrm{BL}}^*$$

which is 0 if and only if $\operatorname{spt}(\nu^\star) \subset \operatorname{spt}(\nu_0)$. When $\beta = 0$, this function directly controls the rate of convergence of this mirror descent dynamics hence the name *mirror rate* function.

Lemma 4.3. *Assume* $(\mathsf{A1}-\mathsf{3})$ *and that J admits a minimizer $\nu^\star \in \mathcal{M}_+(\Theta)$. Then for all $\nu_0 \in \mathcal{M}_+(\Theta)$, denoting $B_{\nu_0} \coloneqq \sup_{J(\nu) \leq J(\nu_0)} \|J'_\nu\|_{\mathrm{BL}}$, it holds for $t \geq 0$,*

$$J(\nu_t) - J^\star \leq \inf_{s \in [0,t]} \left(B_{\nu_0} \cdot \mathfrak{Q}_{\nu^\star,\nu_0}(4\alpha B_{\nu_0} s) + \beta B_{\nu_0}^2 s \right).$$

A direct consequence of this lemma is that $\lim_{t\to\infty} J(\nu_t) - J^\star$ is guaranteed to be small as β gets smaller and as $\operatorname{spt}\nu_0$ gets closer to $\operatorname{spt}\nu^\star$. In Appendix E we give an upper bound on $\mathfrak{Q}$ for the situation of interest here, leading to explicit convergence rates when combined with Lemma 4.3.

Proof. Let $\nu_0^\epsilon \in \mathcal{M}_+(\Theta)$ be a measure to be specified later that satisfies $\mathcal{H}(\nu_0^\epsilon, \nu_0) < +\infty$, and let ν_t^ϵ satisfy $\partial_t \nu_t^\epsilon = \operatorname{div}(\beta \nu_t^\epsilon \nabla J'_{\nu_t})$ for $t \geq 0$ weakly (this is a continuity equation with a smooth velocity field which admits a unique weak solution). Differentiating the relative entropy with respect to

its second argument and using the invariance of the relative entropy under diffeomorphisms, it holds, for $t \geq 0$,

$$\begin{aligned}\frac{1}{4\alpha}\frac{d}{dt}\mathcal{H}(\nu_t^\epsilon, \nu_t) &= \int_\Theta J'_{\nu_t} \mathrm{d}(\nu_t^\epsilon - \nu_t) \\ &= \int_\Theta J'_{\nu_t} \mathrm{d}(\nu^\star - \nu_t) + \int_\Theta J'_{\nu_t} \mathrm{d}(\nu_t^\epsilon - \nu^\star) \\ &\leq -(J(\nu_t) - J^\star) + \|J'_{\nu_t}\|_{\mathrm{BL}} \cdot \|\nu^\star - \nu_t^\epsilon\|_{\mathrm{BL}}^*\end{aligned}$$

where the first term comes from the convexity of J and the second from the definition of $\|\cdot\|_{\mathrm{BL}}^*$. After integrating in time and rearranging the terms we get

$$\left(\frac{1}{t}\int_0^t J(\nu_s)\mathrm{d}s\right) - J^\star \leq \frac{1}{4\alpha t}\left(\mathcal{H}(\nu_0^\epsilon, \nu_0) - \mathcal{H}(\nu_t^\epsilon, \nu_t)\right) + \frac{B_{\nu_0}}{t}\int_0^t \|\nu^\star - \nu_s^\epsilon\|_{\mathrm{BL}}^* \mathrm{d}s.$$

For the last integral term, we use the triangular inequality

$$\frac{B_{\nu_0}}{t}\int_0^t \|\nu^\star - \nu_s^\epsilon\|_{\mathrm{BL}}^* \mathrm{d}s \leq B_{\nu_0}\|\nu^\star - \nu_0^\epsilon\|_{\mathrm{BL}}^* + \frac{B_{\nu_0}}{t}\int_0^t \|\nu_0^\epsilon - \nu_s^\epsilon\|_{\mathrm{BL}}^* \mathrm{d}s \leq B_{\nu_0}\|\nu^\star - \nu_0^\epsilon\|_{\mathrm{BL}}^* + B_{\nu_0}^2 \beta t$$

where the last term is obtained by bounding the integrated flow of the velocity field $(\nabla J'_{\nu_t})_{t\geq 0}$. Since $\mathcal{H}(\nu_\epsilon, \nu_t) \geq 0$ and $J(\nu_s)$ is decreasing, it follows

$$J(\nu_t) - J^\star \leq \inf_{\nu^\epsilon \in \mathcal{M}_+(\Theta)} \left(\frac{1}{4\alpha t}\mathcal{H}(\nu^\epsilon, \nu_0) + B_{\nu_0}\|\nu^\star - \nu^\epsilon\|_{\mathrm{BL}}^*\right) + B_{\nu_0}^2 \beta t. \qquad \square$$

Proof of Theorem 4.1 (gradient flow). By Lemma E.1, we have for $\tau \geq L = \mathrm{Lip}(\log \rho)$, by writing $\bar{\mathcal{H}} \coloneqq \bar{\mathcal{H}}(\nu^\star, \rho \,\mathrm{vol})$,

$$\mathcal{Q}_{\nu^\star, \hat{\nu}_0}(\tau) \leq \frac{\bar{\mathcal{H}} + d \cdot \nu^\star(\Theta) \cdot (\log(\tau) + C_\Theta)}{\tau} + \nu^\star(\Theta) \cdot W_\infty(\nu_0, \rho \,\mathrm{vol}).$$

Combining this bound with Lemma 4.3, we get that for $t \geq L/(4\alpha B_{\nu_0})$,

$$J(\nu_t) - J^\star \leq \frac{\bar{\mathcal{H}} + d \cdot \nu^\star(\Theta) \cdot (\log(4\alpha B_{\nu_0} t) + C_\Theta)}{4\alpha B_{\nu_0} t} + B_{\nu_0} \cdot \nu^\star(\Theta) \cdot W_\infty(\nu_0, \rho \,\mathrm{vol}) + \beta B_{\nu_0}^2 t.$$

In particular, for $t = (\alpha\beta)^{-\frac{1}{2}}$, we get

$$J(\nu_t) - J^\star \leq \sqrt{\frac{\beta}{\alpha}}\left(\bar{\mathcal{H}} + d \cdot \nu^\star(\Theta)(\log(4B_{\nu_0}\sqrt{\alpha/\beta}) + C_\Theta) + B_{\nu_0}^2\right) + B_{\nu_0}\nu^\star(\Theta) \cdot W_\infty(\nu_0, \rho \,\mathrm{vol}). \tag{16}$$

Since this is valid only when $t \geq L/(4\alpha B_{\nu_0})$, we require $(\alpha\beta)^{-\frac{1}{2}} \geq L/(4\alpha B_{\nu_0})$ which leads to the first condition on β/α. Now, we want the right-hand side of (16) to be smaller than $\Delta_0 \coloneqq J_0 - J^\star$ so that we can conclude with Corollary 3.4. To this end, we require, on the one hand $W_\infty(\nu_0, \rho \,\mathrm{vol}) \leq \Delta_0/(2B_{\nu_0}\nu^\star(\Theta))$. On the other hand, we use the bound $\log(u) \leq C_\epsilon u^\epsilon$ for $\epsilon \in]0, 1/2]$, require $4B_{\nu_0}\sqrt{\alpha/\beta} \geq 1$ and obtain the condition

$$(4B_{\nu_0})^\epsilon (\beta/\alpha)^{(1-\epsilon)/2}\left(\bar{\mathcal{H}} + d \cdot \nu^\star(\Theta)(C_\Theta + C_\epsilon) + B_{\nu_0}^2\right) \leq \frac{1}{2}\Delta_0.$$

This leads to the second condition on β/α is the theorem. $\square$

4.3 Fully non-convex gradient descent

The results in the previous section require to set β/α at a small initial value. This might appear undesirable because the asymptotic convergence result of Theorem 2.2 holds irrespective of the choice of β/α. Also, in practice, this condition does not seem required, at least in the examples that we have considered (see Section 5). While the proof technique from Section 4.2 fails without controlling β/α, the question of wether it is possible to obtain convergence rates for any ratio β/α is a natural one.

For such a result, the key challenge is to obtain a convergence rate for the gradient flow dynamics (9) when initialized with a positive density, without conditions on (α, β). While we were not able to prove such a result, in order to point out at the theoretical difficulty, we show in Appendix H with a proof technique inspired by [60], that a convergence rate in objective value in $O(1/\sqrt{\eta t})$ holds as long as the density ν_t is lower bounded by some $\eta > 0$ (at least on a certain subset of Θ).

Proposition 4.4. *Under* (A1-3)*, for any $J_{\max} \geq J^\star$, there exists $C > 0$ such that for any $\eta, t > 0$ and $\nu_0 \in \mathcal{M}_+(\Theta)$ satisfying $J(\nu_0) \leq J_{\max}$, if the projected gradient flow* (9) *satisfies for $0 \leq s \leq t$,*

$$\nu_s|_{S_t} \geq \eta \operatorname{vol}|_{S_t}$$

where $S_t = \{\theta \in \Theta \,;\, J'_{\nu_s}(\theta) \leq 0$ for some $s \in [0, t]\}$, then $J(\nu_t) - J^\star \leq \frac{C}{\sqrt{\alpha \eta t}}$.

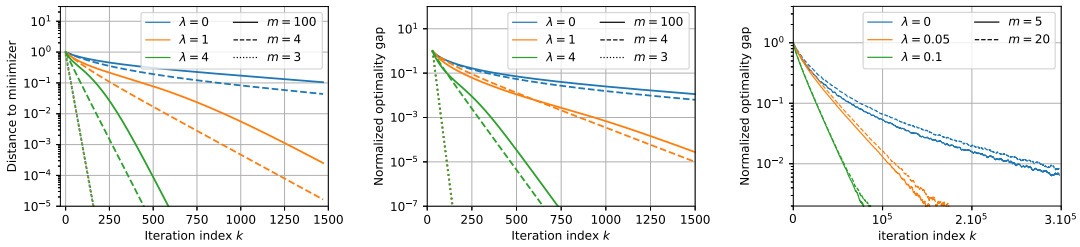
Unfortunately, this result is not sufficient to obtain a convergence rate because the lower bound on the density may decrease too fast. When this happens, the gradient flow may stagnate an *a priori* unbounded time in neighborhoods of saddle points, although it is guaranteed to eventually escape by Lemma C.1. Note that the result above does not requires $\mathcal{F}$ to be finite dimensional nor $\lambda = 0$ while this would be needed for a proof based on the positive definiteness of the tangent kernel [36].

5 Numerical experiments

All experiments can be reproduced with the Julia code available online[4]. Our goal here is not to demonstrate the superiority of Algorithm 1 over other algorithms, but rather to illustrate the insights obtained by the analysis. We consider the following problems introduced in Section 1.1 :

- (Sparse deconvolution) We consider the Dirichlet low-pass filter of order $n_f \in \mathbb{N}_*$ on the d-torus with values in $L^2(\mathbb{T}^d)$ i.e. $\phi(\theta) : x \mapsto \sum_{k=-n_f}^{n_f} \exp(k\sqrt{-1}(x - \theta))$ when $d = 1$. We use the square-loss and solve problem (1) with conic particle gradient descent (Algorithm 1) with the "mirror retraction" from Section 2.3.
- (Two-layer neural net) We consider the function $\phi(w) : x \mapsto \max\left\{\sum_{j=1}^d x_j w_j \cdot |w_j|, 0\right\}$ which is 2-homogeneous on $\mathbb{R}^{d+1}$ with $d + 1 = 20$. We use the square loss and solve problem (1) with stochastic gradient descent with a small fixed step-size for an input data distribution uniform on the sphere $\mathbb{S}^d$. This corresponds to a stochastic version of Algorithm 1 with the "induced retraction" from Section 2.3. For our purposes, the advantage of this architecture over classical ReLU neural networks (as presented in Section 1.1) is that here ϕ is differentiable on $\mathbb{S}^d$ (see, e.g. [17, Lem. D.5]).

[4] `https://github.com/lchizat/2019-sparse-optim-measures`



(a) Deconvolution 2D (distance) (b) Deconvolution 2D (loss) (c) Two-layer neural net (loss)

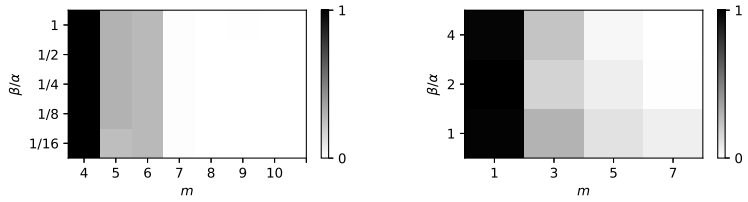
Figure 3: Convergence plots (normalized to 1 after a burning period of 30 iterations). The initialization is close to the minimizer for deconvolution, and is random for neural nets.

We focus in both cases on the "teacher-student" setting without noise with the square loss, because it guarantees that even the unregularized problem ($\lambda = 0$) has sparse solutions, in spite of $\mathcal{F}$ being infinite dimensional. We thus have $R(f) = \frac{1}{2}\|f - f^\star\|_{\mathcal{F}}^2$ where $f^\star = \sum_{i=1}^{m_0} r_i^2 \phi(\theta_i)$ and $m_0 \in \mathbb{N}^*$ is the number of atoms for the teacher.

Local convergence rate. We observe on Figure 3 the effect of the regularization parameter λ and of the over-parameterization parameter m on the local convergence rates (in $\widehat{W}_2$ distance – approximated by mapping each particle to its final position/mass – or in optimality gap). In accordance with the expansion of Proposition 3.7, we observe exponential convergence whenever $\lambda > 0$, with a rate that improves as λ increases. For sparse deconvolution, we observe fast exponential convergence when $m = m_0 = 3$ which is explained by only the first term in the local expansion (13) being non-zero. By adding just a single particle, the second term comes into play and the behavior is qualitatively similar than with 20 particles. For Figure 3-(c), the initialization is random and $m_0 = 5$. Here the behavior for $m = m_0$ follows that of $m > m_0$ which suggests that the first term in the local expansion of Eq. (13) dominates.

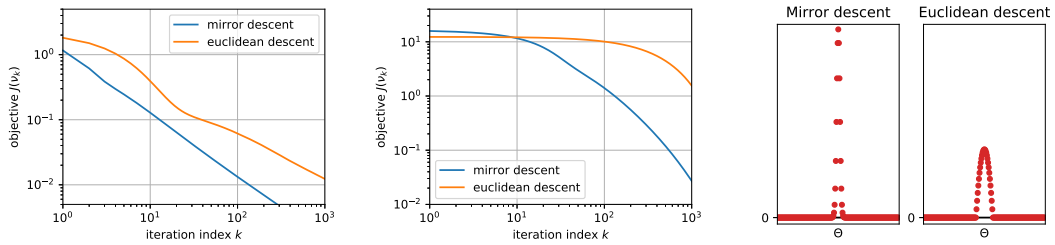
Global convergence. We observe on Figure 4 the effect on the success/failure of optimization of the two main parameters that appear in Theorem 4.1: the over-parameterization parameter m (used to decrease the W_∞ criterion) and the ratio of the vertical/spatial step-sizes β/α. In both (a) and (b) we have $m_0 = 5$ and $\lambda > 0$, and the final loss is averaged over 5 random experiments. Without surprise, minimizers cannot be reached when m is too small. It is also observed that increasing m increases the chances of success even when $m \geq m_0$. In contrast, these experiments do not reveal a clear role for β/α, beyond a change in the convergence speed (see Section 4.3).

Comparison of vertical geometries. Finally, we compare on Figure 5 the behavior of mirror descent against that of Euclidean descent (here integrated with ISTA algorithm [22]). This corresponds respectively to $h(r) = r^2$ and $h(r) = r$ in Eq. 2 and $\beta = 0$. We consider the problem of recovering a single spike ($m_0 = 1$) for 1D and 2D sparse deconvolution, starting from the uniform measure on Θ densely sampled on a grid ($m = 100$). We report the behavior in early stages of optimization, before the effect of the discretization comes into play. We observe that mirror descent outperforms Euclidean descent and enjoys a convergence rate of order $\sim 1/k$ around iteration number $k = 100$. This is in accordance with the result of Appendix G, where we show a convergence rate for mirror descent with continuous densities in $O(\log(k)/k)$, independent



(a) Deconvolution 1D (b) Two-layer neural net

Figure 4: Effect of m and β/α on the excess loss after a fixed large number of iterations (λ fixed). The shades go from white (lowest objective achieved) to black (highest).



(a) Deconvolution 1D (b) Deconvolution 2D (c) Some iterate ν_k

Figure 5: Comparison of mirror ($h(r) = r^2$) and Euclidean ($h(r) = r$) dynamics when $\beta = 0$.

of the dimension. The difference in behavior is illustrated on Figure 5-(c) where we plot ν_{1000} (in the setting of panel (a)).

6 Conclusion

In this paper, we have studied particle gradient descent for sparse convex optimization on measures and obtained complexity guarantees under non-degeneracy assumptions. One central idea underlying our analysis is to directly study the iterates in Wasserstein space. We believe that this approach, at the crossroads between analysis and optimization, may lead to other insights for over-parameterized and non-convex gradient descent.

An avenue for future research is to study the unregularized case. This may require to exploit finer properties of the problem than mere smoothness and could improve our understanding of the implicit bias of over-parameterized gradient descent. Another important question is to find theoretical explanations for the favorable behavior observed in high dimensions for two layer neural networks optimization.

Acknowledgments

The author thanks Francis Bach for fruitful discussions related to this work and the anonymous referees for their thorough reading and suggestions.

References

[1] P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

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

[5] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[6] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] 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.

[8] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[10] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.

[11] Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

[12] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

[13] Emmanuel J. Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[14] Paul Catala, Vincent Duval, and Gabriel Peyré. A low-rank approach to off-the-grid sparse deconvolution. In *Journal of Physics: Conference Series*, volume 904, page 012015. IOP Publishing, 2017.

[15] Frédéric Champagnat and Cedric Herzet. Atom selection in continuous dictionaries: reconciling polar and SVD approximations. In *ICASSP 2019-IEEE 44th International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5. IEEE, 2019.

[16] Yifan Chen and Wuchen Li. Wasserstein natural gradient in statistical manifolds with continuous sample space. *arXiv preprint arXiv:1805.08380*, 2018.

[17] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3040–3050, 2018.

[18] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[19] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[20] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[21] Donald L. Cohn. *Measure theory*, volume 165. Springer, 1980.

[22] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[23] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[24] Yohann De Castro, Fabrice Gamboa, Didier Henrion, and J.-B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2017.

[25] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.

[26] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

[27] Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*. Springer, 2007.

[28] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[29] Axel Flinth, Frédéric de Gournay, and Pierre Weiss. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, pages 1–37, 2020.

[30] Axel Flinth and Pierre Weiss. Exact solutions of infinite dimensional total-variation regularized problems. *Information and Inference: A Journal of the IMA*, 8(3):407–443, 2019.

[31] Thomas Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

[32] Walter Gautschi. *Numerical analysis*. Springer Science & Business Media, 1997.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[34] Daniel Hauer and José Mazón. Kurdyka–Łojasiewicz–Simon inequality for gradient flows in metric spaces. *Transactions of the American Mathematical Society*, 2019.

[35] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1994.

[36] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[37] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[38] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

[39] Walid Krichene, Alexandre Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

[40] Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[41] Wuchen Li and Guido Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, 2018.

[42] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[43] Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[44] Stefania Maniglia. Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *Journal de mathématiques pures et appliquées*, 87(6):601–626, 2007.

[45] 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.

[46] Georg Menz and André Schlichting. Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014.

[47] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. *Wiley Interscience*, 1983.

[48] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[49] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[50] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*, 2018.

[51] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2019.

[52] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in neural information processing systems*, pages 7146–7155, 2018.

[53] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

[54] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[55] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[56] Yann Traonmilin and Jean-François Aujol. The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem. *Inverse Problems*, 36(4):045003, 2020.

[57] Nicolás Garcia Trillos and Dejan Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

[58] Yifei Wang and Wuchen Li. Accelerated information gradient flow. *arXiv preprint arXiv:1909.02102*, 2019.

[59] Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[60] Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

A Dealing with signed measures

Let us show that problems over signed measures with total variation regularization are covered by problem (1), after a suitable reformulation. Consider a function $\tilde{\phi} : \tilde{\Theta} \to \mathcal{F}$ and the functional on signed measures $\tilde{J} : \mathcal{M}(\tilde{\Theta}) \to \mathbb{R}$ defined as

$$\tilde{J}(\mu) = R\left(\int \tilde{\phi} \mathrm{d}\mu\right) + \lambda |\mu|(\tilde{\Theta}). \tag{17}$$

where $|\mu|(\tilde{\Theta})$ is the total variation of μ. This is a continuous version of the LASSO problem, known as BLASSO [23]. Define Θ as the disjoint union of two copies $\tilde{\Theta}_+$ and $\tilde{\Theta}_-$ of $\tilde{\Theta}$ and define the symmetrized function $\phi : \Theta \to \mathcal{F}$ as

$$\phi(\theta) = \begin{cases} +\tilde{\phi}(\theta) & \text{if } \theta \in \tilde{\Theta}_+ \\ -\tilde{\phi}(\theta) & \text{if } \theta \in \tilde{\Theta}_- \end{cases}.$$

With this choice of ϕ, minimizing (17) or minimizing (1) are equivalent, in a sense made precise in Proposition A.1. This symmetrization procedure, also suggested in [17], is simple to implement in practice: in Algorithm 1, we fix at initialization the sign attributed to each particle — depending on whether it belongs to $\tilde{\Theta}_+$ or $\tilde{\Theta}_-$ — and do not change it throughout the iterations.

Proposition A.1. *The infima of* (17) *and* (1) *are the same and:*

(i) if $\tilde{\mu}$ is a minimizer of $\tilde{J}$ and $\tilde{\mu} = \tilde{\mu}_+ - \tilde{\mu}_-$ is its Jordan decomposition, then the measure which restriction to $\tilde{\Theta}_+$ (resp. $\tilde{\Theta}_-$) coincides with $\tilde{\mu}_+$ (resp. μ_-) is a minimizer of J;

(ii) reciprocally, if μ is a minimizer of J then $\mu_+ - \mu_-$ where μ_+ (resp. μ_-) is the restriction of μ to $\tilde{\Theta}_+$ (resp. Θ_-) is a minimizer of $\tilde{J}$.

Proof. We recall that for any decomposition of a signed measure as the difference of nonnegative measures $\tilde{\mu} = \tilde{\mu}_+ - \tilde{\mu}_-$, it holds $|\tilde{\mu}|(\Theta) \leq \tilde{\mu}_+(\Theta) + \tilde{\mu}_-(\Theta)$, with equality if and only if $(\tilde{\mu}_+, \tilde{\mu}_-)$ is the Jordan decomposition of $\tilde{\mu}$ [21, Sec. 4.1]. It follows that starting from any $\tilde{\mu} \in \mathcal{M}(\tilde{\Theta})$, the construction in (i) yields a measure $\mu \in \mathcal{M}_+(\Theta)$ satisfying $\tilde{J}(\tilde{\mu}) = J(\mu)$. Also, starting from any $\mu \in \mathcal{M}_+(\Theta)$, the construction in (ii) yields a measure $\tilde{\mu} \in \mathcal{M}(\tilde{\Theta})$ satisfying $\tilde{J}(\tilde{\mu}) \leq J(\mu)$, with equality if and only if (μ_+, μ_-) is a Jordan decomposition. The conclusion follows. □

B Generic non-convex minimization

In this section, we show that any smooth optimization problem on a manifold is equivalent to solving a problem of the form (1). This corresponds to the case of a scalar-valued ϕ.

Proposition B.1. *Let $\phi : \Theta \to \mathbb{R}$ be a smooth function with minimum $\phi^\star < 0$ that admits a global minimizer, and let*

$$\nu^\star \in \arg\min_{\nu \in \mathcal{M}_+(\Theta)} J(\nu) \qquad \textit{where} \qquad J(\nu) \coloneqq \frac{1}{2}\left(2 + \int_\Theta \phi(\theta) \mathrm{d}\nu(\theta)\right)^2 + \lambda \nu(\Theta) \tag{18}$$

where $0 < \lambda < -2\phi^\star$. Then $\emptyset \neq \operatorname{spt} \nu^\star \subset \arg\min \phi$ so minimizers of ϕ can be built from $\nu^\star$. Reciprocally, from a minimizer of ϕ, one can build a minimizer for (18).

Proof. For a measure $\nu \in \mathcal{M}_+(\Theta)$, we define $f_\nu \coloneqq \int_\Theta \phi(\theta)\mathrm{d}\nu(\theta) \in \mathbb{R}$. It holds

$$\int_\Theta J'_\nu(\theta)\mathrm{d}\nu(\theta) = \int_\Theta \left(\phi(\theta)(2+f_\nu)+\lambda\right)\mathrm{d}\nu(\theta) = f_\nu^2 + 2f_\nu + \lambda\nu(\Theta).$$

Now suppose that ν is a global minimizer of J. Then the optimality condition in Proposition 3.1 implies that

$$f_\nu^2 + 2f_\nu + \lambda\nu(\Theta) = 0. \tag{19}$$

Solving for f_ν is possible if $\lambda\nu(\Theta) < 1$ and leads to $f_\nu = \sqrt{1-\lambda\nu(\Theta)} - 1$. We also deduce from the fact that $f_\nu > -1$ that $\arg\min J'_\nu = \arg\min \phi$, and so $\operatorname{spt}\nu \subset \arg\min\phi$. It remains to find under which condition $\nu(\Theta) > 0$. We use the fact that $f_\nu = \phi^\star\nu(\Theta)$ in Equation (19), and get

$$\nu(\Theta) = \max\left\{0, \frac{-2\phi^\star - \lambda}{(\phi^\star)^2}\right\}$$

which in particular satisfies $\lambda\nu(\Theta) < 1$. Thus, as long as $-2\phi^\star > \lambda$, we have $\nu(\Theta) > 0$. Finally, we verify that global minimizers exist, so that the above reasoning makes sense. If $-2\phi^\star - \lambda \leq 0$, then $\nu = 0$ satisfies the global optimality conditions. Otherwise, choose $\theta^\star$ a minimizer for $\phi^\star$ and define $\nu = \nu(\Theta)\delta_{\theta^\star}$ with the value above for $\nu(\Theta)$, which also satisfies the global optimality conditions. □

C Wasserstein gradient flow

In this section, we recall and adapt some results and proofs from [17], for the sake of completeness.

C.1 Existence

For this result, we assume (A1-2). For a compactly supported initial condition $\mu_0 \in \mathcal{P}_2(\Omega)$, the proof of existence for Wasserstein gradient flows (Eq. (7)) in [17] goes through, as it is simply based on a compactness arguments which can be directly translated to this Riemannian setting (more precisely, we apply here Arzelà-Ascoli compactness criterion for curves in the Wasserstein space on the cone of Θ, which is a complete metric space [42]). Note that these arguments do not require convexity of R, but in order to guarantee global existence in time, we need to assume that ∇R is bounded in sub-level sets of F.

For the existence of solutions for projected dynamics on Θ for any $\nu_0 \in \mathcal{M}_+(\Theta)$, consider a measure $\mu_0 \in \mathcal{M}_+(\Omega)$ such that $\mathsf{h}\mu_0 = \nu_0$ (see [42] for such a construction) and the corresponding Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ for F. Then $\mathsf{h}\mu_t$ is a solution to (9).

For the existence of Wasserstein gradient flows (Eq. (7)) for F when μ_0 is not compactly supported, proceed as follows: there exists a Wasserstein-Fisher-Rao gradient flow ν_t satisfying $\nu_0 = \mathsf{h}\mu_0$. Now we can simply define μ_t as the solution to $\partial_t\mu_t = \operatorname{div}(\mu_t J'_{\nu_t})$. It can be directly checked that $\mathsf{h}\mu_t = \nu_t$ for $t \geq 0$ and thus μ_t is a solution to Eq. (7).

We do not attempt to show uniqueness in the present work. Note that it is proved in [17] for the case where Θ is a sphere, by applying the theory developed in [3].

C.2 Asymptotic global convergence

In this section, we give a short proof of Theorem 2.2, adapted from [17]. The next lemma is the crux of the global convergence proof. It gives a criterion to espace from the neighborhood of measures which are not minimizers.

Lemma C.1 (Criteria to espace local minima)**.** *Under* (A1-3)*, let* $\nu \in \mathcal{M}_+(\Theta)$ *be such that* $v^\star := \min_{\theta\in\Theta} J'_\nu(\theta) < 0$*. Then there exists* $v \in [2v^\star/3, v^\star/3]$ *and* $\epsilon > 0$ *such that if* $(\nu_t)_{t\geq 0}$ *is a projected gradient flow of* J *satisfying* $\|\nu - \nu_{t_0}\|^*_{\mathrm{BL}} < \epsilon$ *for some* $t_0 \geq 0$ *and* $\nu_{t_0}((J'_\nu)^{-1}(]-\infty, v])) > 0$ *then there exists* $t_1 > t_0$ *such that* $\|\nu - \nu_{t_1}\|^*_{\mathrm{BL}} \geq \epsilon$*.*

Proof. We first assume that J'_ν takes nonnegative values and let $v \in [2v^\star/3, v^\star/3]$ be a regular value of g_ν, i.e. be such that $\|\nabla J'_\nu\|$ does not vanish on the v level-set of J'_ν. Such a v is guaranteed to exist thanks to Morse-Sard's lemma and our assumption that ϕ is d-times continuously differentiable, which implies that J'_ν is the same. Let $K_v = (J'_\nu)^{-1}(]-\infty, v]) \subset \Theta$ be the corresponding sublevel set. By the regular value theorem, the boundary ∂K_v of K_v is a differentiable orientable compact submanifold of Θ and is orthogonal to $\nabla J'_\nu$. By construction, it holds for all $\theta \in K_v$, $J'_\nu(\theta) \leq v^\star/3$ and, for some $u > 0$, by the regular value property, $\nabla J'_\nu(\theta) \cdot \vec{n}_\theta > u$ for all $\theta \in \partial K_v$ where $\vec{n}_\theta$ is the unit normal vector to ∂K_v pointing outwards. Since the map $\nu \mapsto J'_\nu$ is locally Lipschitz as a map $(\mathcal{M}_+(\Theta), \|\cdot\|^*_{\mathrm{BL}}) \to (\mathcal{C}^1(\Theta), \|\cdot\|_{\mathrm{BL}})$, there exists $\epsilon > 0$ such that if $\nu_t \in \mathcal{M}_+(\Theta)$ satisfies $\|\nu_t - \nu\|^*_{\mathrm{BL}} < \epsilon$, then

$$\forall \theta \in K_v, \quad J'_{\nu_t}(\theta) \leq v^\star/4 \qquad \text{and} \qquad \forall \theta \in \partial K_v, \quad \nabla J'_{\nu_t}(\theta) \cdot \vec{n}_\theta > u/2.$$

Now let us consider a projected gradient flow $(\nu_t)_{t\geq 0}$ such that $\|\nu_0 - \nu\|^*_{\mathrm{BL}} < \epsilon$ and let $t_1 > 0$ be the first time such that $\|\nu_{t_1} - \nu\|^*_{\mathrm{BL}} \geq \epsilon$, which might a priori be infinite. For $t \in [t_0, t_1[$, it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\nu_t(K_\nu) \geq -4\alpha \int_{K_v} J'_{\nu_t} \mathrm{d}\nu_t \geq v^\star \alpha \nu_t(K_v)$$

where the first inequality can be seen by using the "characteristic" representation of solutions to (9), see [44]. It follows by Grönwall's lemma that $\nu_t(K_v) \geq \exp(\alpha v^\star t)\nu_0(K_v)$ which implies that t_1 is finite. Finally, if we had not assumed that 0 is in the range of J'_ν in the first place, then we could simply take $K = \Theta$ and conclude by similar arguments. □

Proof of Theorem 2.2. Let $\nu_\infty \in \mathcal{M}(\Theta)$ be the weak limit of $(\nu_t)_t$. It satisfies the stationary point condition $\int |J'_{\nu_\infty}|^2 \mathrm{d}\nu_\infty = 0$. Then by the optimality conditions in Proposition 3.1, either ν_∞ is a minimizer of J, or J'_{ν_∞} is not nonnegative. For the sake of contradiction, assume the latter. Let ϵ be given by Lemma C.1 and let $t_0 = \sup\{t \geq 0 \ ; \ \|\nu_t - \nu_\infty\|^*_{\mathrm{BL}} \geq \epsilon\}$ which is finite since we have assumed that ν_t weakly converges to ν_∞. But ν_{t_0} has full support since it can be written as the pushforward of a rescaled version of ν_0 by a diffeomorphism, see [44, Eq. (1.3)] (note that this step is considerably simplified here by the fact that we do not have a potentially non-smooth regularizer, unlike in [17] where topological degree theory comes into play). Then the conclusion of Lemma C.1 contradicts the definition of t_0. □

D Proof of the gradient inequality

In this whole section, we consider without loss of generality $\alpha = \beta = 1$ (we explain in Section D.7 how to adapt the results to arbitrary α, β). For simplicity, we only track the dependencies in ν and τ. Any quantity that is independent of ν and τ is treated as a constant and represented by $C, C', C'' > 0$, and the quantity these symbols refer to can change from line to line.

D.1 Bound on the transport distance to minimizers

Given a measure $\nu \in \mathcal{M}_+(\Theta)$, we consider the local centered moments introduced in Definition 3.6 and in addition, for $i \in \{1, \dots, m^\star\}$,

$$\delta\theta_i = \bar{\theta}_i - \theta_i \qquad \tilde{b}^\theta_i = \bar{r}_i \delta\theta_i, \qquad s_i := \bar{r}_i (\operatorname{tr} \Sigma_i)^{\frac{1}{2}}.$$

Finally, we will quantify errors with the following quantity

$$W_\tau(\nu)^2 \coloneqq \bar{r}_0^2 + \|b\|^2 + \|s\|^2 = \bar{r}_0^2 + \sum_{i=1}^{m^\star} \left(|b_i^r|^2 + \|b_i^\theta\|^2 + s_i^2\right) \tag{20}$$

which also controls the $\widehat{W}_2$ distance (introduced in Section 3.1) to the minimizer $\nu^\star$ of J, as shown in the next proposition.

Lemma D.1. *It holds* $\widehat{W}_2(\nu, \nu^\star) \leq W_\tau(\nu)(1 + O(\tau^2) + O(W_\tau(\nu)^2))$.

Proof. Note that for $W_\tau(\nu)$ small enough, it holds $\nu(\Theta_i) > 0$ for $i \in \{1, \dots, m^\star\}$. Let $\mu \in \mathcal{P}_2(\Omega)$ be such that $\mathsf{h}\mu = \nu$ and consider the transport map $T : \Omega \to \Omega$ defined as

$$T(r, \theta) = \begin{cases} (r\frac{\bar{r}_i}{\bar{r}_i}, \theta_i) & \text{if } \theta \in \Theta_i \text{ and } \bar{r}_i > 0,\ i \in \{1, \dots, m^\star\}, \\ (0, \theta) & \text{otherwise.} \end{cases}$$

By construction, it holds $\mathsf{h}(T_\# \mu) = \nu^\star$. Let us estimate the transport cost associated to this map

$$\mathcal{T} = \int_\Omega \operatorname{dist}((r, \theta), T(r, \theta))^2 \mathrm{d}\mu(r, \theta).$$

The geodesic distance associated to the cone metric is

$$\begin{aligned} \operatorname{dist}((r_1, \theta_1), (r_2, \theta_2))^2 &= r_1^2 + r_2^2 - 2r_1 r_2 \cos(\min\{\operatorname{dist}(\theta_1, \theta_2), \pi\}) \\ &= (r_1 - r_2)^2 + r_1 r_2 \operatorname{dist}(\theta_1, \theta_2)^2 + O(r_1 r_2 \operatorname{dist}(\theta_1, \theta_2)^4) \end{aligned}$$

Now, if we only consider points $\theta \in \Theta_i$ with $\tilde{\theta}$ their coordinates in a normal frame centered at θ_i (note that in all other proofs, we do not need to distinguish between θ and $\tilde{\theta}$), we have the approximation

$$\operatorname{dist}((r_1, \theta_1), (r_2, \theta_2))^2 = (r_1 - r_2)^2 + r_1 r_2 \|\tilde{\theta}_1 - \tilde{\theta}_2\|^2 (1 + O(\tau^2)).$$

Let us decompose $T(r, \theta)$ as $(rT^r(\theta), T^\theta(\theta))$ and estimate the two contributions forming $\mathcal{T}$ separately. On the one hand, we have

$$\int (rT^r(\theta) - r)^2 \mathrm{d}\mu(r, \theta) = \bar{r}_0^2 + \sum_{i=1}^{m^\star} (\bar{r}_i - r_i)^2 = \bar{r}_0^2 + \|b^r\|^2 (1 + O(W_\tau(\nu)^2)).$$

On the other hand, we have

$$\int r^2 T^r(\theta) \|T^\theta(\theta) - \theta\|^2 \mathrm{d}\mu(r, \theta) = \sum_{i=1}^{m^\star} \bar{r}_i r_i (\operatorname{tr} \Sigma_i + \|\delta\theta_i\|^2) = (\|s\|^2 + \|b^\theta\|^2)(1 + O(W_\tau(\nu)^2)).$$

As a consequence, we have $\mathcal{T} = W_\tau(\nu)(1 + O(W_\tau(\nu)^2) + O(\tau^2))$. Remark that this estimate does not depend on the chosen lifting μ satisfying $\mathsf{h}\mu = \nu$. We then conclude by using the characterization in [42, Thm. 7.20] for the distance $\widehat{W}_2$:

$$\widehat{W}_2(\nu_1, \nu_2) = \min\left\{W_2(\mu_1, \mu_2)\ ;\ (\mathsf{h}\mu_1, \mathsf{h}\mu_2) = (\nu_1, \nu_2)\right\}.$$

Thus $\widehat{W}_2(\nu, \nu^\star)^2 \leq W_2(\mu, T_\#(\mu))^2 \leq \mathcal{T}$, and the result follows. □

D.2 Local expansion lemma

Lemma D.2 (Expansion around $\nu^\star$)**.** *Let ψ be any (vector or real-valued) smooth function on Θ and $\nu \in \mathcal{M}_+(\Theta)$. If $\tau > 0$ is an admissible radius, then the following first and second-order expansions hold*

$$\begin{aligned}\int \psi \mathrm{d}(\nu - \nu^\star) &= \sum_{i=1}^m r_i \bar{\nabla}\psi(\theta_i)^\intercal b_i + \int_{\Theta_0} \psi \mathrm{d}\nu + \sum_{i=1}^m \int_{\Theta_i} M_{2,\psi}(\theta_i, \theta)\mathrm{d}\nu(\theta) \\ &= \sum_{i=1}^m r_i \bar{\nabla}\psi(\theta_i)^\intercal b_i + \frac{1}{2}\sum_{i=1}^m \bar{r}_i^2 \left(\mathrm{tr}\left(\nabla^2\psi(\theta_i)\Sigma_i\right) + \delta\theta_i^\intercal \nabla^2\psi(\theta_i)\delta\theta_i\right) \\ &\quad + \int_{\Theta_0} \psi \mathrm{d}\nu + \sum_{i=1}^m \int_{\Theta_i} M_{3,\psi}(\theta_i, \theta)\mathrm{d}\nu(\theta)\end{aligned}$$

where $M_{k,\psi}(\theta_i, \theta)$ is the remainder in the $k-1$-th order Taylor expansion of ψ around θ_i in local coordinates (and we recall that $\bar{\nabla}\psi := (2\psi, \nabla\psi)$).

Proof. By a Taylor expansion of ψ around θ_i for $i \in \{1, \dots, m^\star\}$, it holds

$$\int_{\Theta_i} \psi \mathrm{d}\nu = \int_{\Theta_i} \left(\psi(\theta_i) + \nabla\psi(\theta_i)^\intercal(\theta - \theta_i) + (\theta - \theta_i)^\intercal \nabla^2\psi(\theta_i)(\theta - \theta_i) + M_{3,\psi}(\theta_i, \theta)\right) \mathrm{d}\nu(\theta)$$

and substracting $\int_{\Theta_i} \psi \mathrm{d}\nu^\star = r_i^2 \phi(\theta_i)$ yields

$$\begin{aligned}\int_{\Theta_i} \psi \mathrm{d}(\nu - \nu^\star) &= (\bar{r}_i^2 - r_i^2)\psi(\theta_i) + \bar{r}_i^2 \nabla\psi(\theta_i)^\intercal \delta\theta_i \\ &\quad + \frac{1}{2}\sum_{i=1}^m \bar{r}_i^2 \left(\mathrm{tr}\left(\nabla^2\psi(\theta_i)\Sigma_i\right) + \delta\theta_i^\intercal \nabla^2\psi(\theta_i)\delta\theta_i\right) + \sum_{i=1}^m \int_{\Theta_i} M_{3,\psi}(\theta_i, \theta)\mathrm{d}\nu(\theta)\end{aligned}$$

where we have used a bias-variance decomposition for the quadratic term. The result follows by summing the integrals over each Θ_i and using the expression of b. □

D.3 Bound on the distance to minimizers

In the next lemma, we globally bound the quantity $W_\tau(\nu)$ from Eq. (20) in terms of the function values. It involves the quantity $v^\star > 0$ which is such that for any local minimum θ of $J'_{\nu^\star}$, either $\theta = \theta_i$ for some $i \in \{1, \dots, m^\star\}$ or $J'_{\nu^\star}(\theta) \geq v^*$ (which is non-zero under (A5)). We also recall that $\tilde{b}_i^\theta = \bar{r}_i \delta\theta_i$, as defined in Section D.1.

Lemma D.3 (Global distance bound)**.** *Under (A1-5), let τ_{adm} be an admissible radius τ as in Definition 3.6, fix some $J_{\max} > 0$ and let*

$$\tau_0 = \min\left\{\tau_{\mathrm{adm}}, 2\sqrt{\frac{v^\star}{\sigma_{\min}(H)}}, \frac{3\sigma_{\min}(H)}{2\mathrm{Lip}(\nabla^2 J'_{\nu^\star})}\right\}.$$

Then there exists $C, C' > 0$ such that for all $\tau \leq \tau_0$ and $\nu \in \mathcal{M}_+(\Theta)$ such that $J(\nu) \leq J_{\max}$, it holds

$$W_\tau(\nu) \leq \frac{C}{\tau^2}(J(\nu) - J^\star)^{\frac{1}{2}} \qquad \text{and} \qquad \|\tilde{b}^\theta\|^2 + \|s\|^2 \leq C'(J(\nu) - J^\star).$$

Proof. Let us write $f_\nu \coloneqq \int \phi \mathrm{d}\nu$ and $f^\star = \int \phi \mathrm{d}\nu^\star$. By strong convexity of R at $f^\star$, and optimality of $\mu^\star$, there exists $C > 0$ such that for all $\nu \in \mathcal{M}_+(\Theta)$ it holds

$$J(\nu) - J^\star \geq \int_\Theta J'_{\nu^\star} \mathrm{d}\nu + C \min\{\|f_\nu - f^\star\|^2, \|f_\nu - f^\star\|\}. \tag{21}$$

To prove the first claim, we thus have to bound $W_\tau(\nu)$ using the terms in the right-hand side of (21).

Step 1. By a Taylor expansion, one has for $\theta \in \Theta_i$ for $i \in \{1, \dots, m^\star\}$,

$$|J'_{\nu^\star}(\theta) - \tfrac{1}{2}(\theta - \theta_i)^\intercal H_i(\theta - \theta_i)| \leq \tfrac{1}{6}\mathrm{Lip}(\nabla^2 J'_{\nu^\star})\|\theta - \theta_i\|^3.$$

Thus, if $\|\theta - \theta_i\| \leq 3\sigma_{\min}(H)/(2\mathrm{Lip}(\nabla^2 J'_{\nu^\star}))$, then $J'_{\nu^\star}(\theta) \geq \frac{1}{4}(\theta - \theta_i)^\intercal H_i(\theta - \theta_i)$ for $\theta \in \Theta_i$. Decomposing the integral of this quadratic term into bias and variance, we get

$$\int_{\Theta_i} (\theta - \theta_i)^\intercal H_i(\theta - \theta_i) \mathrm{d}\nu(\theta) = \bar{r}_i^2 \left(\delta\theta_i^\intercal H_i \delta\theta_i + \mathrm{tr}(\Sigma_i H_i)\right)$$

and we deduce a first bound by summing the terms for $i \in \{1, \dots, m^\star\}$,

$$\int_{\Theta \setminus \Theta_0} J'_{\nu^\star} \mathrm{d}\nu \geq \frac{\sigma_{\min}(H)}{4}(\|\tilde{b}^\theta\|^2 + \|s\|^2).$$

Step 2. In order to lower bound the integral over Θ_0, we first derive a lower bound for $J'_{\nu^\star}$ on Θ_0. This is a continuously differentiable and nonnegative function on a closed domain Θ_0 so its minimum is attained either at a local minima in the interior of Θ_0 or on its boundary. Using the quadratic lower bound from the previous paragraph, it follows that for $\theta \in \Theta_0$,

$$J'_{\nu^\star}(\theta) \geq \min\{v^\star, \tau^2 \sigma_{\min}(H)/4\}.$$

Thus, if we also assume that $\tau \leq 2\sqrt{v^\star/\sigma_{\min}(H)}$ then $J'_{\nu^\star}(\theta) \geq \tau^2\sigma_{\min}(H)/4$ for $\theta \in \Theta_0$ and it follows that

$$\int_{\Theta_0} J'_{\nu^\star} \mathrm{d}\nu \geq \frac{\sigma_{\min}(H)}{4} \tau^2 \bar{r}_0^2.$$

Using inequality (21) we have shown so far that

$$\tilde{W}_\tau(\nu)^2 \coloneqq \bar{r}_0^2 + \|\tilde{b}^\theta\|^2 + \|s\|^2 \leq \frac{C}{\tau^2}(J(\mu) - J^\star). \tag{22}$$

Notice that $\tilde{W}_\tau(\nu)$ is similar to $W_\tau(\nu)$ but it does not contain the terms controlling the deviations of mass $|\bar{r}_i - r_i|$. These quantities can be controlled by using the coercivity of R, i.e. the last term in (21), as we do now.

Step 3. Using the first order expansion of Lemma D.2 then squaring gives

$$\left| \|f_\nu - f^\star\|^2 - \frac{1}{2} b^\intercal K b \right| \leq C \left(\|b\| \tilde{W}_\tau(\nu)^2 + \tilde{W}_\tau(\nu)^4 \right).$$

Since we have assumed that K is positive definite, it follows

$$\|f_\nu - f^\star\|^2 \geq C\|b\|^2 - C\tilde{W}_\tau(\nu)^2\|b\| - C\tilde{W}_\tau(\nu)^4$$

and thus, after rearranging the terms

$$(\|b\| - C\tilde{W}_\tau(\nu)^2)^2 \leq C\|f_\nu - f^\star\|^2 + C\tilde{W}_\tau(\nu)^4.$$

It follows that $\|b\| \leq C\|f_\nu - f^\star\| + C\tilde{W}_\tau(\nu)^2$. Also, by inequality (21), if $J(\nu) \leq J_{\max}$, then $\|f_\nu - f^\star\|^2 \leq C(J(\nu) - J^\star)$. Moreover, by inequality (22), we get

$$\|b\| \leq \frac{C}{\tau^2}(J(\nu) - J^\star) + C(J(\nu) - J^\star)^{\frac{1}{2}} \leq \frac{C}{\tau^2}(J(\nu) - J^\star)^{\frac{1}{2}}.$$

We finally combine with the bound on $\tilde{W}_\tau(\nu)$ to conclude since $W_\tau(\nu)^2 \leq \tilde{W}_\tau(\nu)^2 + \|b\|^2$ □

D.4 Proof of the distance inequality (Proposition 3.2)

By Lemma D.1, it holds

$$\widehat{W}_2(\nu, \nu^\star) \leq W_\tau(\nu)(1 + O(\tau^2) + O(W_\tau(\nu)^2)).$$

Moreover, by Lemma D.3, there exists $\tau_0 > 0$ and $C > 0$ such that

$$W_\tau(\nu) \leq \frac{C}{\tau_0^2}(J(\nu) - J^\star)^{\frac{1}{2}}.$$

Combining these two lemmas, it follows that for some $C' > 0$, we have

$$\widehat{W}_2(\nu, \nu^\star)^2 \leq C'(J(\nu) - J^\star)^{\frac{1}{2}}.$$

This also implies a control on the Bounded-Lipschitz distance since it holds $(\|\nu - \nu^\star\|_{\mathrm{BL}}^*)^2 \leq (2 + \pi^2/2)(\nu(\Theta) + \nu^\star(\Theta))\widehat{W}_2(\nu, \nu^\star)^2$, see [42, Prop. 7.18].

D.5 Local estimate of the objective

We now prove a local expansion formula for J.

Proposition D.4 (Local expansion)**.** *It holds*

$$J(\nu) - J^\star = \frac{1}{2}b^\intercal K b + \frac{1}{2}\sum_{i=1}^m \bar{r}_i^2 \left(\mathrm{tr}(\Sigma_i H_i) + \delta\theta^\intercal H_i \delta\theta_i\right) + \int_{\Theta_0} J'_{\nu^\star} \mathrm{d}\nu + \mathrm{err}(\tau, \nu)$$

where $\mathrm{err}(\tau, \nu) = O(\tau(\|\tilde{b}^\theta\|^2 + \|s\|^2) + W_\tau(\nu)^3)$*. In particular, if* τ *is fixed small enough,*

$$J(\nu) - J^\star \leq \sigma_{\max}(K)\|b\|^2 + \sigma_{\max}(H)(\|b^\theta\|^2 + \|s\|^2) + \|J'_{\nu^\star}\|_\infty \bar{r}_0^2 + O(W_\tau(\nu)^3).$$

Proof. Let us write $f_\nu \coloneqq \int \phi \mathrm{d}\nu$ and $f^\star = \int \phi \mathrm{d}\nu^\star$. By a second order Taylor expansion of R around $f^\star$, we have

$$J(\nu) - J^\star = \int_\Theta J'_{\nu^\star} \mathrm{d}\nu + \frac{1}{2}\|f_\nu - f^\star\|_\star^2 + O(\|f_\nu - f^\star\|^3).$$

Using the first order expansion of Lemma D.2 for ϕ, we get $\|f_\nu - f^\star\|_\star^2 = b^\intercal K b + O(W_\tau(\nu)^3)$. Also, using the second order expansion of Lemma D.2 for $J'_{\nu^\star}$ and using the fact that $J'_{\nu^\star}$ and its gradient vanish for all θ_i, we get

$$\int_\Theta J'_{\nu^\star} \mathrm{d}\nu = \frac{1}{2}\sum_{i=1}^m \bar{r}_i^2 \left(\mathrm{tr}(\Sigma_i H_i) + \delta\theta^\intercal H_i \delta\theta_i\right) + \int_{\Theta_0} J'_{\nu^\star} \mathrm{d}\nu + O(\tau(\|s\|^2 + \|b^\theta\|^2))$$

and the expansion follows. Notice also that in the expression of $J(\nu)$, $\bar{r}_i$ and r_i are interchangeable up to introducing higher order error, since $|r_i - \bar{r}_i| = O(|b_i^r|)$ (and also $\|\tilde{b}^\theta\| = \|b^\theta\|(1 + O(W_\tau(\nu)))$). □

D.6 Local estimate of the gradient norm

Proposition D.5 (Gradient estimate)**.** *For $\nu \in \mathcal{P}_2(\Omega)$, it holds*

$$\|g_\nu\|^2_{L^2(\nu)} = b^\intercal (K+H)^2 b + \sum_{i=1}^m \bar{r}_i^2 \operatorname{tr}(\Sigma_i H_i^2) + \|g_\nu\|^2_{L^2(\nu|_{\Theta_0})} + \operatorname{err}(\tau, \nu)$$

where $\operatorname{err}(\tau,\nu) \lesssim \tau(\|\tilde{b}^\theta\|^2 + \|s\|^2) + W_\tau(\nu)^3$*. In particular, if τ is fixed small enough*

$$\|g_\nu\|^2_{L^2(\nu)} \geq \frac{1}{2}(\sigma_{\min}(K) + \sigma_{\min}(H))^2 \|b\|^2 + \frac{1}{2}\sigma_{\min}(H)^2\|s\|^2 + \frac{1}{4}\bar{r}_0^2 \sigma_{\min}(H)^2 \tau^4 + O(W_\tau(\nu)^3).$$

Proof. For this proof, we write $f_\nu - f^\star = \delta f_0 + \delta f_b + \delta f_{\text{err}}$ where

$$\delta f_0 \coloneqq \int_{\Theta_0} \phi(\theta) \mathrm{d}\nu(\theta), \qquad \delta f_b \coloneqq \sum_{i=1}^m r_i \bar{\nabla}\phi(\theta_i) b_i, \qquad \delta f_{\text{err}} \coloneqq \sum_{i=1}^m \int_{\Theta_i} M_{\phi,2}(\theta_i, \theta) \mathrm{d}\nu(\theta).$$

where the decomposition follows from Lemma D.2. The expression for the norm of the gradient is as follows:

$$\|g_\nu\|^2_{L^2(\nu)} = \int_\Theta \|\bar{\nabla} J'_\nu(\theta)\|^2 \mathrm{d}\nu(\theta)$$

where $\bar{\nabla} J = (2J, \nabla J)$. We start with the following decomposition for $\theta \in \Theta_i$ (recall that $J'_\nu(\theta) = \langle \phi(\theta), \nabla R(\int \phi \mathrm{d}\nu)\rangle + \lambda$):

$$\begin{aligned} J'_\nu(\theta) = \lambda + \left\langle \phi(\theta_i) + (\theta - \theta_i)^\intercal \nabla\phi(\theta_i) + \frac{1}{2}(\theta-\theta_i)^\intercal \nabla^2 \phi(\theta_i)(\theta-\theta_i) + M_{\phi,3}(\theta_i,\theta), \nabla R(f^\star) \right\rangle \\ + \langle \phi(\theta_i) + (\theta-\theta_i)^\intercal \nabla\phi(\theta_i) + M_{\phi,2}(\theta_i,\theta), f_\mu - f^\star\rangle_\star + \langle \phi(\theta), M_{\nabla R,2}(f^\star, f)\rangle \end{aligned}$$

Here we use the notation $\langle \cdot, \cdot \rangle_\star$ to denote the quadratic form associated to $\nabla^2 R(f^\star)$. Thanks to the optimality conditions $\bar{\nabla} J'_{\nu^\star}(\theta_i) = 0$ for $i \in \{1, \dots, m\}$, we get

$$\begin{aligned} \bar{\nabla}_j J'_\nu(\theta) &= [H_i(\theta - \theta_i)]_j + \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0 + \delta f_b\rangle_\star + [N(\theta_i,\theta)]_j \\ &= [H_i(\theta - \bar{\theta}_i)]_j + \big([H_i(\bar{\theta}_i - \theta_i)]_j + \langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star\big) + \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0\rangle_\star + [N(\theta_i,\theta)]_j \end{aligned}$$

where N collects the higher order terms and is defined as

$$[N(\theta_i,\theta)]_j = \langle \bar{\nabla}_j M_{\phi,3}(\theta_i,\theta), \nabla R(f^\star)\rangle + \langle \bar{\nabla}_j M_{\phi,2}(\theta_i,\theta), f - f^\star\rangle_\star + \langle \bar{\nabla}_j \phi(\theta), M_{\nabla R,2}(f^\star, f)\rangle$$

where $\|\bar{\nabla}_j M_{\phi,3}(\theta_i,\theta)\| = O(\|\theta - \theta_i\|^2)$ if $j > 0$ and $O(\|\theta-\theta_i\|^3)$ if $j = 0$. Expanding the square

gives the following ten terms:

$$\int_{\Omega\setminus\Omega_0} \|\bar{\nabla} J'_\nu\|^2 \mathrm{d}\nu(\theta) = \sum_{i=1}^m \int_{\Theta_i} \sum_{j=0}^d [H_i(\theta-\bar{\theta}_i)]_j^2 \mathrm{d}\nu \tag{I}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} \sum_{j=0}^d \left(\langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star + [H_i(\bar{\theta}_i-\theta_i)]_j\right)^2 \mathrm{d}\nu(\theta) \tag{II}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} \sum_{j=0}^d \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0\rangle_\star^2 \mathrm{d}\nu(\theta) \tag{III}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d [H_i(\theta-\bar{\theta}_i)]_j \cdot (\langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star + [H_i(\bar{\theta}_i-\theta_i)]_j) \mathrm{d}\nu(\theta) \tag{IV}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d [H_i(\theta-\bar{\theta}_i)]_j \cdot \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0\rangle_\star \mathrm{d}\nu(\theta) \tag{V}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d (\langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star + [H_i(\bar{\theta}_i-\theta_i)]_j) \cdot \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0\rangle_\star \mathrm{d}\nu(\theta) \tag{VI}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d [N(\theta_i,\theta)]_j \cdot [H_i(\theta-\bar{\theta}_i)]_j \mathrm{d}\nu(\theta) \tag{VII}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d [N(\theta_i,\theta)]_j \cdot (\langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star + [H_i(\bar{\theta}_i-\theta_i)]_j) \mathrm{d}\nu(\theta) \tag{VIII}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} 2\sum_{j=0}^d [N(\theta_i,\theta)]_j \cdot \langle \bar{\nabla}_j \phi(\theta_i), \delta f_0\rangle_\star \mathrm{d}\nu(\theta) \tag{IX}$$

$$+ \sum_{i=1}^m \int_{\Theta_i} \sum_{j=0}^d [N(\theta_i,\theta)]_j^2 \mathrm{d}\nu(\theta) \tag{X}$$

Terms (I) to (II) are the main terms in the expansion, while the other terms are higher order. The term (I) is a local curvature term and can be expressed as (I) $= \sum_{i=1}^m \bar{r}_i^2 \operatorname{tr} \Sigma_i H_i^2$. The term (II) is a global interaction term that writes

$$\begin{aligned}
\text{(II)} &= \sum_{i=1}^m \bar{r}_i^2 \sum_{j=0}^d |\langle \bar{\nabla}_j \phi(\theta_i), \delta f_b\rangle_\star + [H_i(\bar{\theta}_i-\theta_i)]_j|^2 \\
&= \sum_{i=1}^m \sum_{j=0}^d \left| \sum_{i'=1}^m \sum_{j'=0}^d (\langle \bar{r}_i \bar{\nabla}_j \phi(\theta_i), r_{i'} \bar{\nabla}_{j'} \phi(\theta_{i'})\rangle_\star + \bar{H}_{(i,j),(i,j')})(b_{i',j'}) \right|^2 \\
&= \|(\bar{K}+H)(b)\|^2
\end{aligned}$$

where the entries of $\bar{K}$ and $\bar{H}$ differ from those of K and H by a factor $\bar{r}_i/r_i$. More precisely,

$$[\bar{K}-K]_{(i,j),(i',j')} = (\bar{r}_i/r_i - 1) K_{(i,j),(i',j')}$$

and similarly for $\bar{H}-H$. Since $|\bar{r}_i/r_i - 1| = O(|b_i^r|)$ we have $\sigma_{\max}(\bar{K}-K) = O(W_\tau(\nu))$. It

follows, by expanding the square, that

$$\|(\bar{K}+\bar{H})(b)\|^2 = \|(K+H)(b)\|^2 + O(W_\tau(\nu)^3).$$

The remaining terms are error terms, that we estimate directly in terms of $W_\tau(\nu)$ and τ. We use in particular the fact that by Hölder's inequality, $\int_{\Theta_i} \|\theta - \bar{\theta}_i\| \mathrm{d}\nu(\theta) = O(\bar{r}_i^2 \operatorname{tr} \Sigma_i^{\frac{1}{2}})$. One has

- (III) $= O\left(\sum_{i=1}^m \bar{r}_i^2 \bar{r}_0^4\right) = O(W_\tau^4(\nu))$;
- (IV) $=$ (V) $= 0$ because the integral of the terms $H_i(\theta - \bar{\theta}_i)$ vanishes;
- (VI) $= O\left((\sum_{i=1}^m \bar{r}_i^2 (\|b\| + \|\delta\theta_i\|) \cdot \bar{r}_0^2\right) = O(W_\tau^3(\nu))$;
- (VII) $= O(\tau(\|\tilde{b}^\theta\|^2 + \|s\|^2)) + O(W_\tau(\nu)^3)$;
- (VIII) $= O(W_\tau^3(\nu))$;
- (IX) $= O(W_\tau^4(\nu))$;
- (X) $= O(\tau^2(\|\tilde{b}^\theta\|^2 + \|s\|^2)) + O(W_\tau(\nu)^4)$.

It follows that overall, the error term is in $O(\tau(\|\tilde{b}^\theta\|^2 + \|s\|^2) + W_\tau(\nu)^3)$. There remains to lower bound the norm of the gradient over Θ_0, which can be done as follows. As seen in the proof of Lemma D.3, if τ is small enough then $J'_{\nu^\star}(\theta) \geq \tau^2 \sigma_{\min}(H)/4$ for $\theta \in \Theta_0$. Considering only the first component of the gradient, it holds

$$\int_{\Theta_0} \|\bar{\nabla} J'_\nu(\theta)\|^2 \mathrm{d}\nu(\theta) \geq \int_{\Theta_0} 4|J'_\nu(\theta)|^2 \mathrm{d}\nu(\theta).$$

Using the expansion $J'_\nu(\theta) = J'_{\nu^\star}(\theta) + \langle \phi(\theta), M_{\nabla R,1}(f^\star, f_\nu)\rangle$, we get

$$\int_{\Theta_0} \|\bar{\nabla} J'_\nu(\theta)\|^2 \mathrm{d}\nu(\theta) \geq C\bar{r}_0^2 \tau^4 + O(W_\tau(\nu)^3).$$

The result follows by collecting all the estimates above. □

D.7 Proof of the sharpness inequality (Theorem 3.3)

By Proposition D.4 we have that for $\tau > 0$ small enough

$$J(\nu) - J^\star \leq C W_\tau(\nu)^2 + O(W_\tau(\nu)^3)$$

where $C = \sigma_{\max}(K+H) + \|J'_{\nu^\star}\|_\infty$.

Similarly, by Proposition D.5, for τ small enough, it holds

$$\|g_\nu\|^2_{L^2(\nu)} \geq C' W_\tau(\nu)^2 + O(W_\tau(\nu)^3)$$

where $C' = \frac{1}{8}\sigma_{\min}(H)^2\tau^4$. Now fix $\tau > 0$ satisfying the hypothesis of Lemma D.3 and the two previous inequalities. By Lemma D.3, $W_\tau(\nu) = O((J(\nu) - J^\star)^{\frac{1}{2}})$. We deduce that there exists $J_0 > J^\star$ and $\kappa_0 > 0$, such that whenever $\nu \in \mathcal{M}_+(\Theta)$ satisfies $J(\nu) < J_0$, one has

$$\|g_\nu\|^2_{L^2(\nu)} \geq \kappa_0 (J(\nu) - J^\star).$$

Finally, notice that if different metric factors $(\alpha, \beta) \neq (1,1)$ are introduced, one can always lower bound the new gradient squared norm as

$$\int_\Theta \left(4\alpha |J'_\nu(\theta)|^2 + \beta \|\nabla J'_\nu(\theta)\|_\theta^2\right) \mathrm{d}\nu(\theta) \geq \min\{\alpha,\beta\} \int_\Theta \left(4|J'_\nu(\theta)|^2 + \|\nabla J'_\nu(\theta)\|_\theta^2\right)$$

which proves the statement for any (α, β). Note however that if one wants to make a more quantitative bound, then there are values (α_0, β_0) that would lead to a better conditioning and potentially higher values for J_0. In this case, the factor appearing in the sharpness inequality should rather be $\min\{\alpha/\alpha_0, \beta/\beta_0\}$.

E Estimation of the mirror rate function

We provide an upper bound for the *mirror rate* function $\mathcal{Q}$ in the situation that is of interest to us, with $\nu^\star$ sparse. Note that this approach could be generalized to arbitrary $\nu^\star$.

Lemma E.1. *Under* (A1)*, there exists $C_\Theta > 0$ that only depends on the curvature of Θ, such that for all $\nu^\star, \nu_0 \in \mathcal{M}_+(\Theta)$ where $\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \delta_{\theta_i}$ and $\nu_0 = \rho \operatorname{vol}$ where $\log \rho$ is L-Lipschitz, then*

$$\mathcal{Q}_{\nu^\star,\nu_0}(\tau) \leq \frac{1}{\tau}\left(\nu^\star(\Theta)\cdot d \cdot \left(C_\Theta + \log(\tau) + L/\tau\right) + \nu_0(\Theta) - \nu^\star(\Theta) + \sum_{i=1}^{m^\star} r_i^2 \log\left(\frac{r_i^2}{\rho(\theta_i)}\right)\right).$$

Moreover, for any other $\hat{\nu}_0 \in \mathcal{M}_+(\Theta)$, it holds $\mathcal{Q}_{\nu^\star,\hat{\nu}_0}(\tau) \leq \mathcal{Q}_{\nu^\star,\nu_0}(\tau) + \nu^\star(\Theta)\cdot W_\infty(\nu_0,\hat{\nu}_0)$.

In the context of Lemma E.1, we introduce the quantity,

$$\bar{\mathcal{H}}(\nu^\star,\rho) \coloneqq \sum_{i=1}^{m^\star} r_i^2 \log\left(\frac{r_i^2}{\rho(\theta_i)}\right) - \nu^\star(\Theta) + \nu_0(\Theta).$$

which measures how much ρ is a good prior for the (a priori unknown) minimizer $\nu^\star$. With this quantity, the conclusion of Lemma E.1 reads, for $\tau \geq L$,

$$\mathcal{Q}_{\nu^\star,\nu_0}(\tau) \leq \frac{\bar{\mathcal{H}}(\nu^\star,\rho) + \nu^\star(\Theta)\cdot d\cdot(\log(\tau) + C_\Theta)}{\tau}.$$

Proof. Let us build ν_ϵ in such a way that the quantity defining $\mathcal{Q}_{\nu^\star,\nu_0}(\tau)$ in Eq. (15) is small. For this, consider a radius $\epsilon > 0$ and consider the measure ν_ϵ defined as the normalized volume measure on each geodesic ball of radius τ around each θ_i, with mass r_i^2 on this ball, and vanishing everywhere else. Using the transport map that maps these balls to their centers θ_i, we get if Θ is flat,

$$\|\nu_\epsilon - \nu^\star\|_{\mathrm{BL}}^* \leq W_1(\nu_\epsilon,\nu^\star) \leq \sum_{i=1}^{m^\star} \frac{r_i^2}{V^{(d)}(\epsilon)} \int_0^\epsilon s \frac{d}{ds} V^{(d)}(s) \mathrm{d}s$$

where $V^{(d)}(\epsilon)$ is the volume of a ball of radius ϵ in $\mathbb{R}^d$, that scales as ϵ^d. Using an integration by parts, it follows

$$\frac{1}{V^{(d)}(\epsilon)} \int_0^\epsilon s \frac{d}{ds} V^{(d)}(s) \mathrm{d}s = \epsilon - \int_0^\epsilon \frac{V^{(d)}(s)}{V^{(d)}(\epsilon)} \mathrm{d}s = \epsilon - \int_0^\epsilon \left(\frac{s}{\epsilon}\right)^d \mathrm{d}s = \frac{\epsilon d}{d+1},$$

thus $W_1(\nu_\epsilon,\nu^\star) \leq \nu^\star(\Theta)\epsilon$. In the general case where Θ is a potentially curved manifold, this upper bound also depends on the curvature of Θ around each θ_i, a dependency that we hide

in the multiplicative constant so $W_1(\nu_\epsilon, \nu^\star) \le C\nu^\star(\Theta)\epsilon$. Let us now control the entropy term. Writing $\rho_\epsilon = \mathrm{d}\nu_\epsilon/\mathrm{dvol}$ and Θ_i for the geodesic ball of radius ϵ around θ_i, it holds

$$\mathcal{H}(\nu_\epsilon, \nu_0) = \nu_0(\Theta) - \nu^\star(\Theta) + \sum_{i=1}^{m^\star} \int_{\Theta_i} \rho_\epsilon(\theta) \log(\rho_\epsilon(\theta)/\rho(\theta)) \mathrm{dvol}(\theta).$$

The integral term can be estimated as follows,

$$\begin{aligned}\int_{\Theta_i} \rho_\epsilon(\theta) \log(\rho_\epsilon(\theta)/\rho(\theta)) \mathrm{dvol}(\theta) &= \int_{\Theta_i} \frac{r_i^2}{V^{(d)}(\epsilon)} \Big(\log(r_i^2) - \log V^{(d)}(\epsilon) - \log(\rho(\theta)) \Big) \mathrm{dvol}(\theta) \\ &\le r_i^2 \Big(\log(r_i^2) - \log V^{(d)}(\epsilon) - \log \rho(\theta_i) + \mathrm{Lip}(\log \rho) \cdot \epsilon \Big).\end{aligned}$$

Recalling that $-\log V^{(d)}(\epsilon) \le -d \log(\epsilon) + C$ for some C that only depends on the curvature of Θ, we get that the right-hand side of (15) is bounded by

$$\frac{1}{\tau} \left(C\nu^\star(\Theta) + \nu_0(\Theta) - \nu^\star(\Theta) d \log(\epsilon) + \sum_{i=1}^{m^\star} r_i^2 \log\left(\frac{r_i^2}{\rho(\theta_i)} \right) + \epsilon L \nu^\star(\Theta) \right) + C\nu^\star(\Theta)\epsilon.$$

Let us fix $\epsilon > 0$ by minimizing $C\nu^\star(\Theta)\epsilon - \nu^\star(\Theta) d \log(\epsilon)/\tau$, which gives $\epsilon = d/(C\tau)$. The first claim follows by plugging this value for ϵ in the expression above.

For the second claim of the statement, let us build a suitable candidate $\hat{\nu}_\epsilon$ in order to upper bound the infimum that defines $\mathcal{Q}_{\nu^\star, \hat{\nu}_0}(\tau)$. Let T be an optimal transport map from ν_0 to $\hat{\nu}_0$ for W_∞, i.e. a measurable map $T : \Theta \to \Theta$ satisfying $T_\# \nu_0 = \hat{\nu}_0$ and $\max\{\mathrm{dist}(\theta, T(\theta)) \ ; \ \theta \in \mathrm{spt}\, \nu_0 (= \Theta)\} = W_\infty(\nu_0, \hat{\nu}_0)$ (see [53, Sec. 3.2], the absolute continuity of ν_0 is sufficient for such a map to exist). Now we define $\hat{\nu}_\epsilon = T_\# \nu_\epsilon$ where ν_ϵ is such that $\mathcal{H}(\nu_\epsilon, \nu_0) < \infty$. Since the relative entropy is non-increasing under pushforwards, it holds $\mathcal{H}(\hat{\nu}_\epsilon, \hat{\nu}_0) \le \mathcal{H}(\nu_\epsilon, \nu_0)$. Moreover, it holds $\|\nu_\epsilon - \hat{\nu}_\epsilon\|_{\mathrm{BL}}^* \le W_1(\nu_\epsilon, \hat{\nu}_\epsilon) \le \nu^\star(\Theta) W_\infty(\nu_\epsilon, \hat{\nu}_\epsilon)$. Thus we have

$$\begin{aligned}\mathcal{Q}_{\nu^\star, \hat{\nu}_0}(\tau) &\le \|\nu^\star - \hat{\nu}_\epsilon\|_{\mathrm{BL}}^* + \frac{1}{\tau} \mathcal{H}(\hat{\nu}_\epsilon, \hat{\nu}_0) \\ &\le \|\nu_\epsilon - \hat{\nu}_\epsilon\|_{\mathrm{BL}} + \|\nu^\star - \nu_\epsilon\|_{\mathrm{BL}} + \frac{1}{\tau} \mathcal{H}(\nu_\epsilon, \nu_0) \\ &\le \nu^\star(\Theta) W_\infty(\nu_\epsilon, \hat{\nu}_\epsilon) + \|\nu^\star - \nu_\epsilon\|_{\mathrm{BL}}^* + \frac{1}{\tau} \mathcal{H}(\nu_\epsilon, \nu_0).\end{aligned}$$

The claim follows by noticing that, by construction, $W_\infty(\nu_\epsilon, \hat{\nu}_\epsilon) \le W_\infty(\nu_0, \hat{\nu}_0)$ and then by taking the infimum in ν_ϵ. □

F Global convergence for gradient descent

In the following, result, we study the non-convex gradient descent updates $\mu_{k+1} = (T_k)_\# \mu_k$ and $\nu_k = \mathsf{h}\mu_k$ where

$$T_k(r, \theta) = \mathrm{Ret}_{(r,\theta)}(-2\alpha_k J'_{\nu_k}(\theta), -\beta_k \nabla J'_{\nu_k}(\theta))$$

with step-sizes $\alpha, \beta > 0$. When $\beta = 0$, we recover mirror descent updates in $\mathcal{M}_+(\Theta)$ with the entropy mirror map (more specifically, this is true when Ret is the "mirror" retraction defined in Section 2.3).

Lemma F.1. *Assume* $(\mathsf{A1}-3)$ *and that* J *admits a minimizer* $\nu^\star \in \mathcal{M}_+(\Theta)$*. Then there exists* $C, \eta_{\max} > 0$ *such that for all* $\nu_0 \in \mathcal{M}_+(\Theta)$*, denoting* $B = \sup_{J(\nu)\le J(\nu_0)} \|J'_\nu\|_{\mathrm{BL}}$*, if* $\max\{\alpha,\beta\} < \beta_{\max}$*, it holds*

$$J(\nu_k) - J^\star \le \inf_{k'\in[0,k]} \left(B \cdot \mathcal{Q}_{\nu^\star,\nu_0}(4B\alpha k) + C\alpha + \beta B^2 k\right).$$

Proof. As in the proof of Lemma 2.5, we define $(T_k^r(\theta), T_k^\theta(\theta)) \coloneqq T_k(1,\theta)$ and we define recursively $\nu^\epsilon_{k+1} = (T_k^\theta)_\# \nu_k^\epsilon$ where ν_0^ϵ is such that $\mathcal{H}(\nu_0^\epsilon, \nu_0) < \infty$. Using the invariance of the relative entropy under diffeomorphisms (indeed, T_k^θ is a diffeomorphism of Θ for β small enough), and doing a first order expansion of $T_k^r = 1 - 2\alpha J'_{\nu_k} + O(\alpha^2)$ it holds for β small enough

$$\begin{aligned}
\frac{1}{4\alpha}\left(\mathcal{H}(\nu^\epsilon_{k+1},\nu_{k+1}) - \mathcal{H}(\nu_k^\epsilon,\nu_k)\right) &= \frac{1}{4\alpha}(\mathcal{H}(\nu_k^\epsilon, (T_k^r)^2\nu_k) - \mathcal{H}(\nu_k^\epsilon,\nu_k)) \\
&= \frac{1}{4\alpha}\Big(\int \log((T_k^r)^{-2})\mathrm{d}\nu_k^\epsilon + \int ((T_k^r)^2 - 1)\mathrm{d}\nu_k\Big) \\
&= \int J'_{\nu_k}\cdot \mathrm{d}(\nu_k^\epsilon - \nu_k) + O(\alpha) \\
&= \int J'_{\nu_k}\mathrm{d}(\nu^\star - \nu_k) + \int J'_{\nu_k}\mathrm{d}(\nu_k^\epsilon - \nu^\star) + O(\alpha) \\
&\le -(J(\nu_k) - J^\star) + \|J'_{\nu_k}\|_{\mathrm{BL}} \cdot \|\nu^\star - \nu_k^\epsilon\|^*_{\mathrm{BL}} + C\alpha
\end{aligned}$$

where the term in $O(\alpha)$ originates from a first order approximation of the retraction. Now, taking $\max\{\alpha,\beta\}$ small enough to ensure decrease of $(J(\nu_k))_k$ (by Lemma 2.5) so that C above can be chosen independently of k, it follows

$$\begin{aligned}
\left(\frac{1}{k}\sum_{k'=0}^{k-1} J(\nu_{k'})\right) - J^\star &\le \frac{1}{4\alpha k}\mathcal{H}(\nu_0^\epsilon,\nu_0) + B\|\nu^\star - \nu_0^\epsilon\|^*_{\mathrm{BL}} + \frac{B}{k}\left(\sum_{k'=0}^{k-1}\|\nu_{k'}^\epsilon - \nu_0^\epsilon\|\right) + C\alpha \\
&\le B\mathcal{Q}_{\nu^\star,\nu_0}(4B\alpha k) + C\alpha + \frac{1}{2}\beta B^2(k-1)
\end{aligned}$$

by bounding each term $\|\nu_{k'}^\epsilon - \nu_0^\epsilon\|$ by $B\beta k'$. □

Proof of Theorem 4.2 (gradient descent). The proof follows closely that of Theorem 4.1 but we do not track the "constants" (this would be more tedious). By Lemma E.1, there exists $C > 0$ (that depends on $\bar{\mathcal{H}}$, the curvature of Θ and $\nu^\star(\Theta)$) such that $\mathcal{Q}_{\nu^\star,\hat\nu_0}(\tau) \le C(\log\tau)/\tau + \nu^\star(\Theta)W_\infty(\nu_0,\hat\nu_0)$. Combining this with Lemma F.1, we get that when $\max\{\alpha,\beta\} \le \eta_{\max}$,

$$J(\nu_k) - J^\star \le C\frac{\log(B\alpha k)}{\alpha k} + \beta B^2 k + C'\alpha + B\nu^\star(\Theta)\cdot W_\infty(\nu_0,\hat\nu_0).$$

Our goal is to choose k_0, α, β and $W_\infty(\nu_0,\hat\nu_0)$ so that this is quantity smaller than $\Delta_0 \coloneqq J_0 - J^\star$. With $\alpha = 1/\sqrt{k}$ and $\beta = \beta_0/k$ we get

$$J(\nu_k) - J^\star \le \frac{C'\log(Bk)}{\sqrt{k}} + B^2\beta_0 + B\nu^\star(\Theta)\cdot W_\infty(\nu_0,\hat\nu_0).$$

Then, using a bound $\log(u) \le C_\epsilon u^\epsilon$, we may choose $k \gtrsim \Delta_0^{-2-\epsilon}$, $\beta_0 \le \frac{1}{3}\Delta_0/B^2$ and $W_\infty(\nu_0,\hat\nu_0) \le \frac{1}{3}\Delta_0/(B\nu^\star(\Theta))$ in order to have $J(\nu_k) - J^\star \le \Delta_0$. This gives $\alpha \lesssim \Delta_0^{1+\epsilon/2}$, $\beta \lesssim \Delta_0^{3+\epsilon}$ and the regime of exponential convergence kicks off after $k = \Delta_0^{-2-\epsilon}$ iterations. □

G Faster rate for mirror descent

In this section, we show that for a specific choice of retraction, the convergence rate of $O(\log(t)/t)$ for the gradient flow is preserved for the gradient descent.

Proposition G.1 (Mirror flow, fast rate)**.** *Assume* (A1-4) *and consider the infinite dimensional mirror descent update*

$$\nu_{k+1} = \exp(-4\alpha J'_{\nu_k})\nu_k$$

which corresponds to the so-called mirror retraction in Section 2.3 and $\beta = 0$. For any $\nu_0 \in \mathcal{M}_+(\Theta)$, there exists $\alpha_{\max} > 0$ such that for $\alpha \leq \alpha_{\max}$ it holds, denoting $B_{\nu_0} = \sup_{J(\nu)\leq J(\nu_0)} \|J'_\nu\|_{\mathrm{BL}}$,

$$J(\nu_k) - J^\star \leq B_{\nu_0}\Omega_{\nu^\star,\nu_0}(2\alpha B_{\nu_0}k).$$

In particular, combining with Lemma E.1, if $\nu_0 = \rho\mathrm{vol}$ has a smooth positive density, then $J(\nu_k) - J^\star = O(\log(k)/k)$.

Proof. Consider $\nu_\epsilon \in \mathcal{M}_+(\Theta)$ such that $\mathcal{H}(\nu_\epsilon, \nu_0) < \infty$. It holds

$$\begin{aligned}\frac{1}{4\alpha}\left(\mathcal{H}(\nu_\epsilon,\nu_k) - \mathcal{H}(\nu_\epsilon,\nu_{t+1})\right) &= -\frac{1}{4\alpha}\int \log\left(\frac{\nu_{k+1}}{\nu_k}\right)\mathrm{d}(\nu_{k+1}-\nu_\epsilon) + \frac{1}{4\alpha}\mathcal{H}(\nu_{k+1},\nu_k)\\ &= \int J'_{\nu_k}\mathrm{d}(\nu_{k+1}-\nu_\epsilon) + \frac{1}{4\alpha}\mathcal{H}(\nu_{k+1},\nu_k)\end{aligned}$$

where the first equality is obtained by rearranging terms in the definition of $\mathcal{H}$, and the second one is specific to the mirror retraction. Let us estimate the two terms in the right-hand side. Using convexity inequalities, we get

$$\begin{aligned}\int J'_{\nu_k}\mathrm{d}(\nu_{k+1}-\nu_\epsilon) &= \int J'_{\nu_k}\mathrm{d}(\nu_k-\nu^\star) + \int J'_{\nu_k}\mathrm{d}(\nu_{k+1}-\nu_k) + \int J'_{\nu_k}\mathrm{d}(\nu^\star-\nu_\epsilon)\\ &\geq J(\nu_k) - J(\nu^\star) + J(\nu_{k+1}) - J(\nu_k) + O(\alpha\|g_{\nu_k}\|^2_{L^2(\nu_k)}) + \int J'_{\nu_k}\mathrm{d}(\nu^\star-\nu_\epsilon)\\ &\geq J(\nu_{k+1}) - J^\star + O(\alpha\|g_{\nu_k}\|^2_{L^2(\nu_k)}) - \|J'_{\nu_k}\|_{\mathrm{BL}}\cdot\|\nu^\star-\nu_\epsilon\|^*_{\mathrm{BL}}.\end{aligned}$$

Here the term in $O(\alpha\|g_{\nu_k}\|^2_{L^2(\nu_k)})$ comes from the proof of Lemma 2.5 (note that the iterates remain in a sublevel of J for α small enough). As for the relative entropy term, we have, using the convexity inequality $\exp(u) \geq 1 + u$,

$$\begin{aligned}\frac{1}{\alpha}\mathcal{H}(\nu_{k+1},\nu_k) &= \frac{1}{\alpha}\int\left(\exp(-2\alpha J'_{\nu_k})(-2\alpha J'_{\nu_k}-1)+1\right)\mathrm{d}\nu_k\\ &\geq \frac{1}{\alpha}\int\left(4\alpha^2|J'_{\nu_k}\|^2 - 1 + 1\right)\mathrm{d}\nu_k = \int 4\alpha|J'_{\nu_k}|^2\mathrm{d}\nu_k = \|g_{\nu_k}\|^2_{L^2(\nu_k)}.\end{aligned}$$

We use this inequality in place of the strong convexity of the mirror function used in the usual proof of mirror descent (because there is no Pinsker inequality on $\mathcal{M}_+(\Theta)$). Coming back to the first equality we have derived, it holds,

$$B_{\nu_0}\|\nu^\star-\nu_\epsilon\|^*_{\mathrm{BL}} + \frac{1}{4\alpha}\left(\mathcal{H}(\nu_\epsilon,\nu_k) - \mathcal{H}(\nu_\epsilon,\nu_{t+1})\right) \geq J(\nu_{k+1}) - J^\star + \frac{1}{4}\|g_{\nu_k}\|^2_{L^2(\nu_k)} + O(\alpha\|g_{\nu_k}\|^2_{L^2(\nu_k)})$$

Thus for α small enough, it holds

$$B_{\nu_0}\|\nu^\star-\nu_\epsilon\|^*_{\mathrm{BL}} + \frac{1}{4\alpha}\left(\mathcal{H}(\nu_\epsilon,\nu_k) - \mathcal{H}(\nu_\epsilon,\nu_{t+1})\right) \geq J(\nu_{k+1}) - J^\star.$$

Summing over K iterations and dividing by K, we get

$$\left(\frac{1}{K}\sum_{k=1}^{K} J(\nu_k)\right) - J^\star \leq \frac{1}{4\alpha K}\mathcal{H}(\nu_\epsilon, \nu_K) + B_{\nu_0}\|\nu^\star - \nu_\epsilon\|^*_{\mathrm{BL}}.$$

Since for α small enough $(J(\nu_k))_{k\geq 1}$ is decreasing (by Lemma 2.5), the result follows. □

H Convergence rate for lower bounded densities

In this section, we justify the claim made in Section 4.3 about the convergence without condition on β/α. Let us recall the result that we want to prove.

Proposition H.1. *Under* (A1-3)*, for any $J_{\max} > J^\star$, there exists $C > 0$ such that for any $\eta, t > 0$ and $\nu_0 \in \mathcal{M}_+(\Theta)$ satisfying $J(\nu_0) \leq J_{\max}$, if the projected gradient flow* (9) *satisfies for $0 \leq s \leq t$,*

$$\nu_s|_{S_t} \geq \eta \,\mathrm{vol}\,|_{S_t}$$

where $S_t = \{\theta \in \Theta \,;\, J'_{\nu_s}(\theta) \leq 0$ for some $s \in [0,t]\}$, then $J(\nu_t) - J^\star \leq \frac{C}{\sqrt{\alpha\eta t}}$.

Proof. Following [60], we start with the convexity inequality

$$J(\nu_t) - J^\star \leq \int J'_{\nu_t}\mathrm{d}\nu_t - \int J'_{\nu_t}\mathrm{d}\nu^\star.$$

Let us control these two terms separately. On the one hand, one has by Jensen's inequality

$$\left(\int J'_{\nu_t}\mathrm{d}\nu_t\right)^2 = (\nu_t(\Theta))^2\left(\frac{1}{\nu_t(\Theta)}\int J'_{\nu_t}\mathrm{d}\nu_t\right)^2 \leq \nu_t(\Theta)\int |J'_{\nu_t}|^2\mathrm{d}\nu_t \leq -\frac{\nu_t(\Theta)}{4\alpha}\frac{\mathrm{d}}{\mathrm{d}t}J(\nu_t).$$

Using the fact that on sublevels of J, $\nu(\Theta)$ and $\|g_\nu\|^2_{L^2(\nu)}$ are bounded, we have, for some $C > 0$,

$$\int J'_{\nu_t}\mathrm{d}\nu_t \leq C\left(-\frac{1}{\alpha}\frac{\mathrm{d}}{\mathrm{d}t}J(\nu_t)\right)^{1/3}.$$

On the other hand, we have

$$\int J'_{\nu_t}\mathrm{d}\nu^\star \geq \nu^\star(\Theta)\min\left\{0, \min_{\theta\in\Theta} J'_{\nu_t}(\theta)\right\} =: \nu^\star(\Theta)\cdot v_t.$$

where the last equality defines $v_t \leq 0$. Using the gradient flow structure, let us show that a non-zero v_t and a lower bound η on the density of ν_t (at least on the set $\{J'_{\nu_t} \leq 0\}$) guarantees a decrease of the objective. Indeed, letting $\Theta_t = \{\theta \in \Theta \,;\, J'_{\nu_t}(\theta) \leq v_t/2\}$ (which could be empty), we get

$$-\frac{\mathrm{d}}{\mathrm{d}t}J(\nu_t) \geq 4\alpha\int_{\Theta_t} |J'_{\nu_t}|^2\mathrm{d}\nu_t \geq 4\alpha(v_t/2)^2\nu_t(\Theta_t) = \alpha\cdot v_t^2\cdot\eta\cdot\mathrm{vol}(\Theta_t).$$

Moreover, the Lipschitz regularity of J'_ν is bounded on sublevels of J, and thus along gradient flow trajectories, so there exists $C' > 0$ such that $\mathrm{vol}(\Theta_t) \geq C'\cdot|v_t|$. It follows

$$|v_t|^3 \leq -\frac{1}{C'\alpha\eta}\frac{\mathrm{d}}{\mathrm{d}t}J(\nu_t) \quad\Rightarrow\quad v_t \geq -\left(-\frac{1}{C'\alpha\eta}\frac{\mathrm{d}}{\mathrm{d}t}J(\nu_t)\right)^{1/3}.$$

Coming back to our first inequality, we have

$$J(\nu_t) - J^\star \leq C \left(-\frac{1}{\alpha} \frac{\mathrm{d}}{\mathrm{d}t} J(\nu_t) \right)^{1/3} + \left(-\frac{1}{C' \alpha \eta} \frac{\mathrm{d}}{\mathrm{d}t} J(\nu_t) \right)^{1/3} \leq \frac{C''}{(\alpha\eta)^{1/3}} \left(-\frac{\mathrm{d}}{\mathrm{d}t} J(\nu_t) \right)^{1/3}$$

for some $C'' > 0$ that, given $J(\nu_0)$, is independent of α, η and ν_t. It remains to remark that a continuously differentiable and positive function h that satisfies $h(t) \leq C^{-1/3} \cdot (-h'(t))^{1/3}$ satisfies $C \leq -h'(t)/h(t)^3 = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t}(h(t)^{-2})$ and, after integrating between 0 and t, $h(t) \leq \left(2Ct + h(0)^{-2}\right)^{-1/2} \leq \frac{1}{\sqrt{2Ct}}$. We conclude by taking $h(t) = J(\nu_t) - J^\star$ and $C \propto \alpha\eta$. □