L'École normale supérieure-PSL

Mémoire première année

Learning, optimal transport, gradient flows

Authors:
Paul Feyel
Ruben Bontorno

Supervisors:
Bertrand Maury
Quentin Mérigot

Contents

1 Introduction to the problem

1.1 Motivations and formulation of the problem

A classical problem in machine learning and signal processing is to find the minimizer of a convex function of a measure. This covers for example sparse spike deconvolution, or training a neural network with a single hidden layer. We will study a simple minimization method: we discretize the measure into a mixture of particles and run a gradient descent on their weight and position. We show that this gradient descent, although performed on a non-convex function, whith a good initialization, at the many-particle limit, converges to global minimizers.

This is based on the work L. Chizat et F. Bach did in [6]. Here, we focused on the partially 1-homogeneous case that they introduced, and expanded on the proof to hopefully make them more understandable, along with our own reproduction of their numerical experiments.

The mathematical setting is the following: we want to search for an element in an Hilbert space $\mathcal{F}$ that minimizes a smooth, convex loss function $R : \mathcal{F} \to \mathbb{R}_+$, and that is a linear combination of a few elements from a given parametrized set $\{\phi(\theta)\}_{\theta\in\Theta} \subset \mathcal{F}$.
We can encode the linear combination as a signed measure μ on the parameter space Θ and solve for:

$$J^* = \min_{\mu\in\mathcal{M}(\Theta)} J(\mu) \quad \text{where} \quad J(\mu) \overset{\text{def}}{=} R(\int \phi d\mu) + G(\mu) \tag{1}$$

where $\mathcal{M}(\Theta)$ is the set of signed measures on the parameter space Θ, and $G : \mathcal{M}(\Theta) \to \mathbb{R}$ is an optional convex regularizer, e.g. the total variation norm when sparse solutions are favored.

Our case of interest will be the infinite-dimensional case where Θ is a domain of $\mathbb{R}^d$, and ϕ is differentiable.
This setting allows for very general applications, which include:

- **Training neural networks with a single hidden layer**. 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. In this case, $\mathcal{F}$ is $L^2(\mathbb{R}^{d-1})$, with R being the quadratic or logistic loss function for example, and $\phi(\theta) : x \mapsto \sigma(\sum_{i=1}^{d-1} \theta_i x_i + \theta_d)$, with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ (e.g. the sigmoid function or the rectified linear unit, see [15] and [12]).

- **Sparse spike deconvolution**. The goal is to recover a signal which is a mixture of of impulses on Θ given a noisy and filtered observation y. Here, $\mathcal{F}$ is $L^2(\Theta)$, with $\phi(\theta) : x \mapsto \psi(x-\theta)$ (where ψ is the filter impulse response) and $R(f) = 1/2\lambda \cdot \|f-y\|_{L^2}^2$, for some λ which depends on the noise level (for more information check [9] and [10]).

- **Low rank tensor decomposition.** For background information on this subject, please refer to [14]. However, we will not delve further into this application in the report. If you are interested in recovering low-rank matrix decompositions, you can refer to [13]. For recovering mixtures of models from sketches, you can check out [18].

1.2 Study case: sparse spike deconvolution

Whilst the original paper discusses the favorable case where ϕ is homogeneous (i.e. $\phi(\lambda\theta) = \lambda\phi(\theta)$), which allows for less restrictive conditions on the initialization and less assumptions on ϕ, we will not do so. We chose to do this as it allows us to cover the most general case, and the proofs we will give are most of the time easily adaptable (and much simpler) to the homogeneous case. However they can be found on the original paper. This setting allows us to cover the problems described before when ϕ is bounded.

A particular case of interest for us is the sparse spike deconvolution, which has a lot of real-world applications, including:

Superresolution imaging. The diffraction of light imposes a physical limit on the resolution of optical images. The goal of superresolution is to remove the blur induced by diffraction as well as the effects of pixelization and noise. For images composed of a collection of point sources of light, this can be posed as a sparse inverse problem as follows. The parameters $\theta_1, \ldots, \theta_K$ denote the locations of K point sources (in $\mathbb{R}^2$ or $\mathbb{R}^3$), and w_i denotes the intensity, or brightness, of the i-th source. The image of the i-th source is given by $w_i\psi(\theta_i)$, where ψ is the pixelated point spread function of the imaging apparatus. Astronomers use this framework to deconvolve images of stars to angular resolution below the Rayleigh limit (for further details, see [19]). In biology this tool has revolutionized imaging of subcellular features (you can exlore this further in [11] and [17]).

Design of numerical quadrature rules. In many numerical computing applications we require fast procedures to approximate integration against a fixed measure. One way to do this is use a quadrature rule:

$$\int f(\theta)dp(\theta) \approx \sum_{i=1}^{k} w_i f(x_i) \, .$$

The quadrature rule, given by $w_i \in \mathbb{R}$ and $x_i \in \Theta$, is chosen so that the above approximation holds for functions f in a certain function class. The pairs (x_i, w_i) are known as quadrature nodes. In practice, we want quadrature rules with very few nodes to speed evaluation of the rule. Often we don't have an a priori description of the function class from which f is chosen, but we might have a finite number of examples of functions in the class, $f_1, \ldots, f_d$, along with their integrals against p, $y_1, \ldots, y_d$. In other words, we know that

$$\int f_i(\theta)dp(\theta) = y_i \, .$$

A reasonable quadrature rule should approximate the integrals of the known f_i well. We can phrase this task as a sparse inverse problem where each source is a single quadrature node. In our notation, $\psi(\theta) = (f_1(\theta), \ldots, f_d(\theta))$. A common choose of R for this application is simply the squared loss.

Designing radiation therapy. External radiation therapy is a common treatment for cancer in which several beams of radiation are fired at the patient to irradiate tumors. The collection of beam parameters (their intensities, positions, and angles) is called the treatment plan, and is chosen to minimize an objective function specified by an oncologist. The objective usually

rewards giving large doses of radiation to tumors, and low dosages to surrounding healthy tissue and vital organs. Plans with few beams are desired as repositioning the emitter takes time-increasing the cost of the procedure and the likelihood that the patient moves enough to invalidate the plan. A beam fired with intensity $w > 0$ and parameters θ delivers a radiation dosage $w\psi(\theta) \in \mathbb{R}^d$. Here the output is interpreted as the radiation delivered to each of d voxels in the body of a patient. The radiation dosage from beams with parameters $\theta_1, \ldots, \theta_K$ and intensities $w_1, \ldots, w_K$ add linearly, and the objective function is convex. For background see [16].

A more complete list can be found on [3], which tackles a variant of sparse spike deconvolution, where the total variation norm is bounded.

1.3 Strategy and main difficulties

At first glance, the problem might look easy: we want to minimize a smooth function that is convex, a gradient-descent based algorithm should do the trick. However, the domain is an infinite-dimensional space, and representing an arbitrary measure is impossible in a computer.

Our strategy will be to discretize the measure as a mixture of particles determined by their positions and weights. This means solving:

$$\min_{\substack{w\in\mathbb{R}^m \\ \theta\in\Theta^m}} J_m(w,\theta), \text{ where } J_m(w,\theta) \stackrel{\text{def}}{=} J(\frac{1}{m}\sum_{i=1}^{m} w_i\delta_{\theta_i}) \tag{2}$$

Which we would solve by performing the usual gradient descent-based algorithms. The idea is that, as the number of particle goes to infinity, we should recover the minimizer of problem (1).

However, one big problem is that the functions J_m are not convex and in general have multiple local minima, which is a problem for our gradient descent.

To illustrate this phenomenon, we can take the case of a sparse spike deconvolution as described in the introduction, with $\Theta = [0, 1]$, ψ a Dirichlet kernel of order 8, an ideal signal $y = \sum_{i=1}^{5} w_i \cdot \phi(\theta_i)$, for $w = (2, -5, 6, 1, 2)/4$ and $\theta = (0.1, 0.3, 0.45, 0.7, 0.95)$, no regularizer and $\lambda = 1$. Below are slices of the functions J_m for different number of particles, which clearly show that we cannot expect convexity in general.

Figure 1: J_1 on the $[-5, 5] \times [0, 1]$ box.

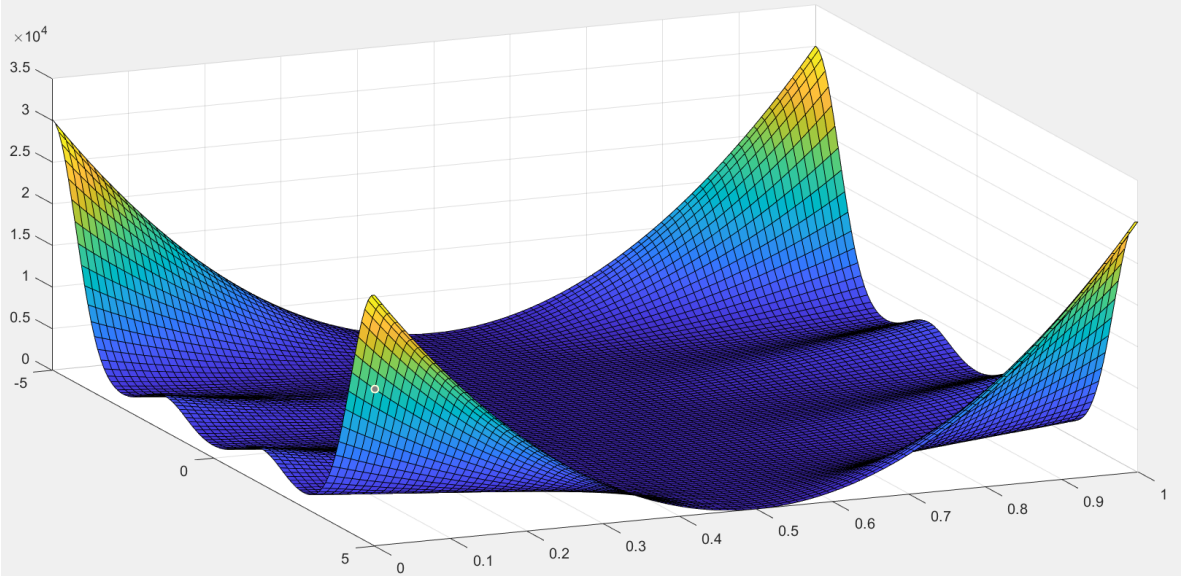


Figure 2: J_2 on the segment from (w, θ_1) to (w, θ_2) for $w = (2, 4)$, $\theta_1 = (0.2, 0.8)$ and $\theta_2 = (0.6, 0.7)$.

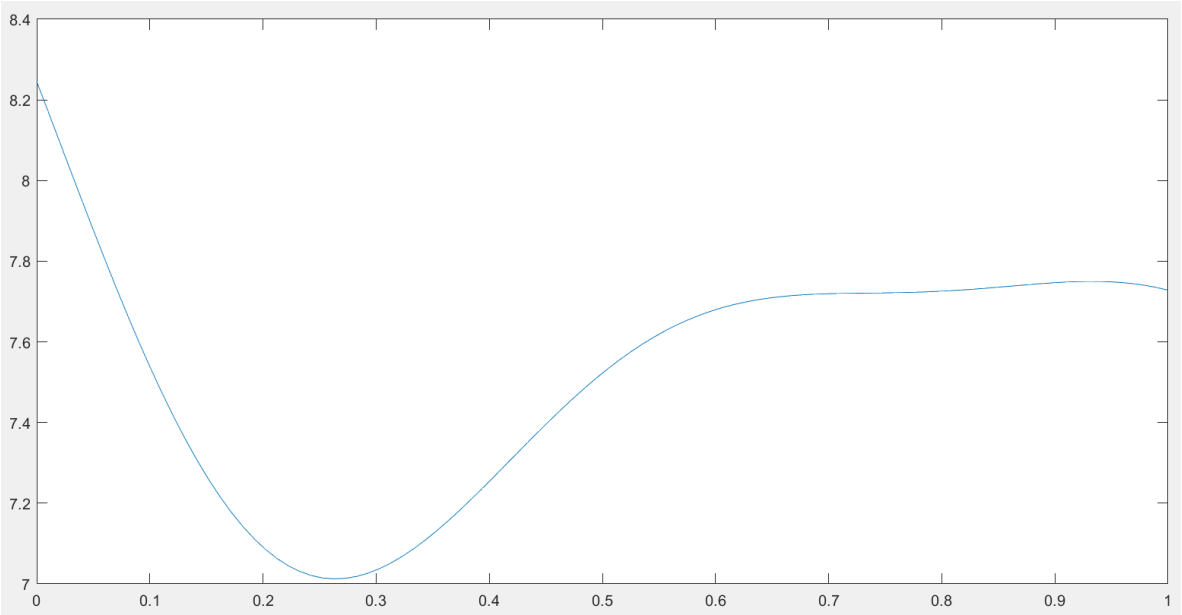


Figure 3: J_4 on the segment from (w, θ_1) à (w, θ_2) for $w = (2, 4, 2, -1)$, $\theta_1 = (0.2, 0.8, 0.1, 0.9)$ and $\theta_2 = (0.6, 0.7, 0.05, 0.6)$.

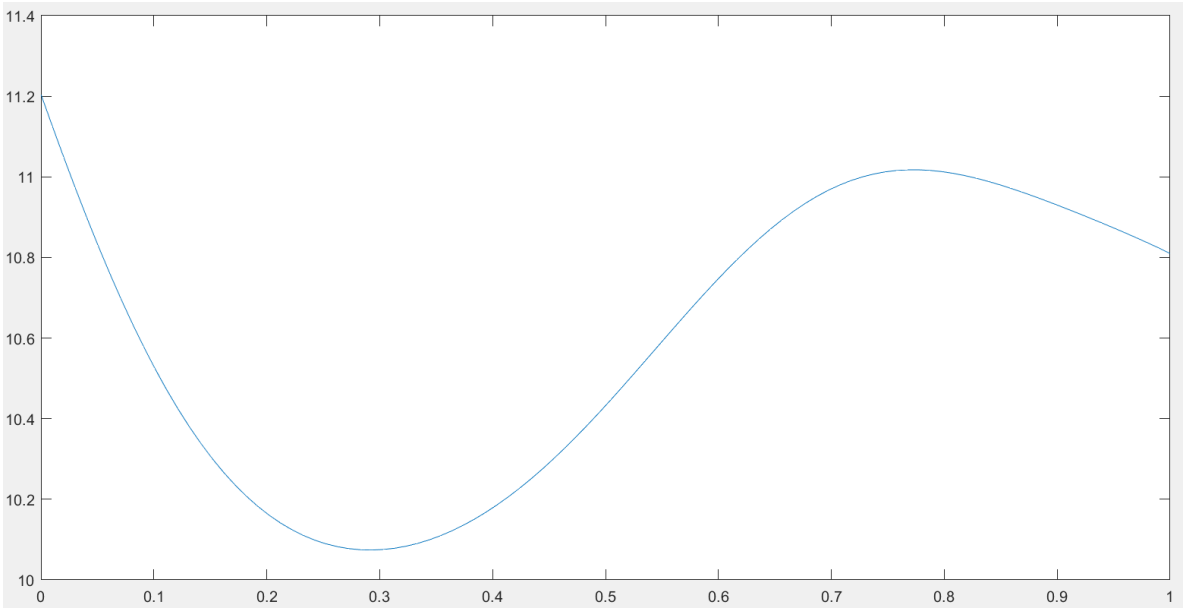
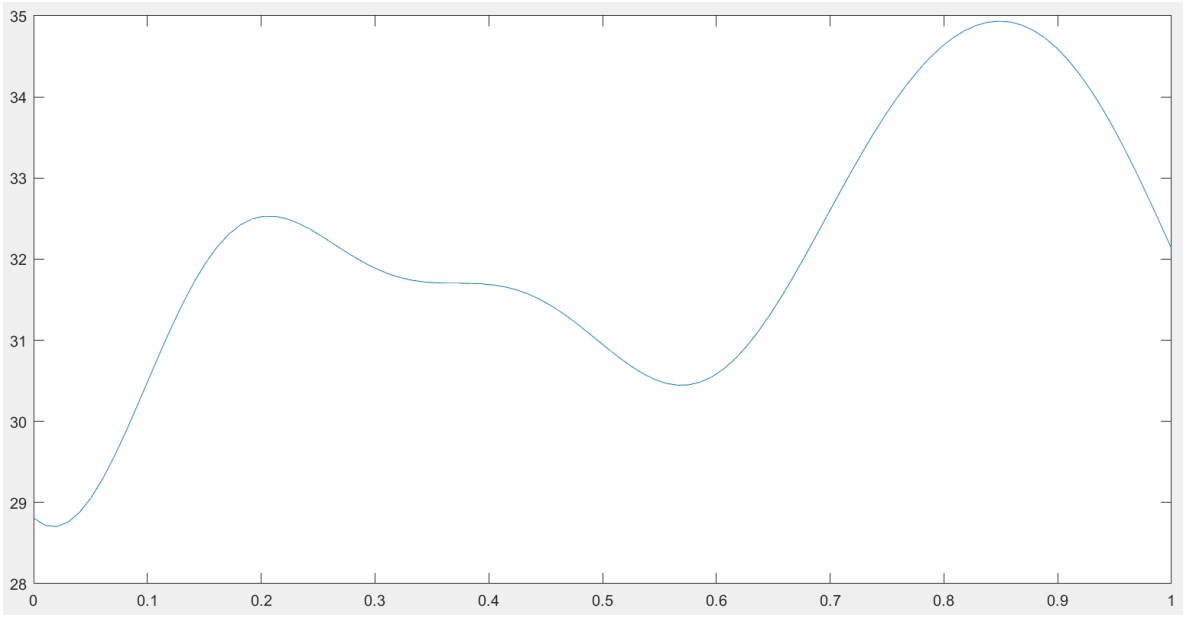


Figure 4: J_{10} on the segment from (w, θ_1) à (w, θ_2) for $w = (2, 4, 2, -1, 1, 3, 4, -4, 3, 2)$, $\theta_1 = (0.2, 0.8, 0.1, 0.9, 0.5, 0.2, 0.7, 0.6, 0.9, 0.4)$ and $\theta_2 = (0.6, 0.7, 0.05, 0.6, 0.1, 0.3, 0.9, 0.8, 0.1, 0.7)$.



In each case, we clearly see that convexity fails. Another technical problem will be to construct gradient descent in general: R is smooth but the regularizer G might not, and for example the total variation norm is not. This means that we cannot use the classical theory of differential equation, and will instead introduce subgradients, which will allow us to cover a much larger class of regularizers.

2 Proofs and theorems for the partially 1-homogeneous case

The organization of the proofs will be as follows:

- We first introduce a more general class of problem.
- We then study the many-particle limit of the associated gradient flow for this class of problem and characterize it as a Wasserstein gradient flow.
- Finally, we show that under some assumptions on ϕ and the initialization, if this Wasserstein gradient flow converges, then the limit is a global minimizer of J.

In order to facilitate a comprehensive understanding of the subsequent discussion, we will introduce a set of key definitions that hold significant relevance.

— Any signed measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ can be decomposed as $\mu = \mu_+ - \mu_-$ where $\mu_+, \mu_- \in \mathcal{M}_+(\mathbb{R}^d)$, by the Jordan decomposition theorem. Moreover, if μ_+ and μ_- are of minimal total mass, the **variation** of μ is $|\mu| \stackrel{\text{def}}{=} \mu_+ - \mu_-$ and $|\mu|(\mathbb{R}^d)$ is **the total variation norm**.

— **The support** *spt* μ of measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ is the complement of the largest open set of measure 0. For example, for the Lebesgue measure on $\mathbb{R}$, the only open set with measure 0 is the empty set and hence it's support is $\mathbb{R}$. μ is **concentrated** on a set $S \subset \mathbb{R}^d$ if the complement of S is included in a measurable set of measure 0. Returning to our previous example, we have that the Lebesgue measure is concentrated on $\mathbb{R}\backslash\mathbb{Q}$.

— We now introduce a key concept in optimal transport: **the pushforward**. Consider two measurable sets, X and Y, in $\mathbb{R}^d$, and let $T : X \to Y$ be a measurable map, serving as our **transport map**. The pushforward is defined as $T_{\#}\mu(B) = \mu(T^{-1}(B))$ for all measurble set $B \subset Y$. Inutitively, it corresponds to the distribution of "mass" of μ after it as been displaced by T. One of its fundamental property is that it satisfies $\int_Y \phi d(T_{\#}\mu) = \int_X \phi \circ T d\mu$ whenever $\phi : Y \to \mathbb{R}$ is a measurable map such that $\phi \circ T$ is μ-integrable (see [7]). For exemple, when considering π^i the projection map, the pushforward $\pi^i_{\#}\mu$ corresponds to the marginal measure of μ on the i-th coordinates.

— **The weak convergence** of a sequence $\mu_n \in \mathcal{M}(\mathbb{R}^d)$ to μ is defined as follows: for all bounded and continuous functions $\phi : \mathbb{R}^d \to \mathbb{R}$, we have $\int \phi d\mu_n \to \int \phi d\mu$.

— **The Bounded Lipschitz norm** is defined as for any $\mu \in \mathcal{M}(\mathbb{R}^d)$: $\|\mu\|_{BL} = \sup\{\int \phi d\mu; \phi : \mathbb{R}^d \to \mathbb{R}, Lip(\phi) \leq 1, \|\phi\|_\infty \leq 1\}$ where $Lip(\phi)$ denotes the smallest Lipschitz constant of ϕ and $\|\cdot\|_\infty$ represents the supremum norm. It is important to note that if $\mu_n \in \mathcal{M}(\mathbb{R}^d)$ is a bounded sequence in total variation norm, the weak convergence and the convergence in Bounded Lipschtiz norm are equivalent.

— Given two probability measure $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we define **the p-Wasserstein distance** between μ and ν as follows: $W_p(\mu, \nu) \stackrel{\text{def}}{=} (\min \int |y - x|^p d\gamma(x, y))^{\frac{1}{p}}$, the minimization is taken over the set of probability measures $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that the marginal laws of γ (on both $\mathbb{R}^d$ factor) correspond to μ and ν. By convention, we associate μ with the first factor and ν with the second factor.

— **A positively p-homogeneous** function, where $p \geq 0$, is a function f from $\mathbb{R}^d$ to a vector space that satisfies the property for any $u \in \mathbb{R}^d$ and $\lambda > 0$: $f(\lambda u) = \lambda^p f(u)$.

— $f : \mathbb{R}^d \to \mathbb{R}$ is **semiconvex**, if there exists a $\lambda \in \mathbb{R}$ such that $f + \lambda|\cdot|^2$ is convex.

There is several key properties associated with these definitions.

— When considering only probability measures with second moments, it can be established that this set, equipped with the W_2 distance, forms a complete metric space denoted as $\mathcal{P}_2(\mathbb{R}^d)$.

— We will now state a result of [2]. In $\mathcal{P}_2(\mathbb{R}^d)$, a sequence $(\mu_m)_m$ converges if and only if, for any continuous function $\phi : \mathbb{R}^d \to \mathbb{R}$ with at most quadratic growth, we have $\int \phi d\mu_m \to \int \phi d\mu$. It is important to note that this convergence implies weak convergence.

— Finally, we can establish the following inequalities: $\|\mu - \nu\|_{BL} \leq W_1(\mu, \nu) \leq W_2(\mu, \nu)$.

— The (sub)-differential of a positively p-homogeneous function is a positively $(p-1)$-homogeneous function.

— If f is differentiable (except possibely at 0), the following identity holds: $u \cdot \nabla f(u) = pf(u)$ for $u \neq 0$.

— On a compact domain, any smooth function is semiconvex.

Part A: Rewriting in a more general setting

First, let us formally introduce the objects we are working with by specifying our general assumptions about them:

Assumptions 1: $\mathcal{F}$ is a separable Hilbert space, $\Omega \subset \mathbb{R}^d$ is the closure of a convex open set, and

(i) (smooth loss) $R : \mathcal{F} \to \mathbb{R}_+$ is differentiable, with a differential dR that is Lipschitz on bounded sets and bounded sublevel sets,

(ii) (basic regularity) $\Phi : \Omega \to \mathcal{F}$ is (Fréchet) differentiable, $V : \Omega \to \mathbb{R}_+$ is semiconvex, and

(iii) (locally Lipschitz derivatives with sublinear growth) there exists a family $(Q_r)_{r>0}$ of nested nonempty closed subsets of Ω such that:

  (a) $\{u \in \Omega;\ \mathrm{dist}(u, Q_r) \leq r'\} \subset Q_{r+r'}$ for all $r, r' > 0$,

  (b) Φ and V are bounded and $d\Phi$ is Lipschitz on each Q_r, and

  (c) there exists $C_1, C_2 > 0$ such that $\sup_{u \in Q_r}(\|d\Phi_u\| + \|\partial V(u)\|) \leq C_1 + C_2 r$ for all $r > 0$, where $\|\partial V(u)\|$ stands for the maximal norm of an element in $\partial V(u)$.

We will now consider the following minimization problem:

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \quad \text{where} \quad F(\mu) \stackrel{\text{def}}{=} R\Big(\int \Phi d\mu\Big) + \int V d\mu \tag{3}$$

However, we need to be careful about some issues regarding the well-definedness of F. By convention, we set $F(\mu) = \infty$ if μ is not concentrated on Ω. Another problematic point arises when considering $\int \Phi d\mu$, which is a Bochner integral. For background information about Bochner integral, you can refer to [7]. Its value may not be well-defined in certain cases. Specifically, it has a well-defined value if Φ is measurable and $\int |\phi| d|\mu| < \infty$. Otherwise, we also set $F(\mu) = \infty$ by convention.

We claim that (1) can be rewritten as a special case of (3). We will see that in some setting $J^* = F^*$ and that from a minimizer of J one can easily build a minimizer for F, and vice versa.

Intuitively, this new problem makes sense as we now consider weights as positions, removing the asymmetry between position and weight. Therefore, we can work with positive measures.

We will later on further restrict the framework to address the specific case of the partially 1-homogeneous case. In doing so, we will introduce a new set of assumptions that are more restrictive than Assumptions 1.

Assumptions 2: The domain $\Omega = \mathbb{R} \times \Theta$ with $\Theta \subset \mathbb{R}^{d-1}$, $\Phi(\omega, \theta) = \omega \cdot \phi(\theta)$ and $V(\omega, \theta) = |\omega| \tilde{V}(\theta)$ where ϕ and $\tilde{V}$ are bounded, differentiable with a Lipschitz differential. Moreover

(i) (smooth convex loss) The loss R is convex, differentiable with differentiable dR Lipschitz on bounded sets and bounded on sublevel sets,

(ii) (Sard-type regularity) For all $f \in \mathcal{F}$, the set of regular values of $g_f : \theta \in \Theta \mapsto \langle f, \phi(\theta) \rangle + \tilde{V}(\theta)$ is dense in its range, and

(iii) (boundary conditions) The function ϕ behaves nicely at the boundary of the domain: either

  (a) $\Theta = \mathbb{R}^{d-1}$ and for all $f \in \mathcal{F}$, $\theta \in \mathbb{S}^{d-2} \mapsto g_f(r\theta)$ converges, uniformly in $C^1(\mathbb{S}^{d-2})$ as $r \to \infty$, to a function satisfying the Sard-type regularity, or

  (b) Θ is the closure of a bounded open convex set and for all $f \in \mathcal{F}$, g_f satifies the Neumann boundary conditions (i.e., for all $\theta \in \partial\Theta$, $d(g_f)_\theta(\overrightarrow{n}_\theta) = 0$ where $\overrightarrow{n}_\theta \in \mathbb{R}^{d-1}$ is the normal to $\partial\Theta$ at θ).

First, we observe that Assumptions 2 imply Assumptions 1 by considering the family of nested sets $Q_r = [-r, r] \times \Theta$ for $r > 0$. However, it is worth noting that for parts A and B, we only need Assumptions 1 that apply to ϕ and $\tilde{V}$ (where Ω, Φ, and V are the same as in Assumptions 2, as this is the framework for the partially 1-homogeneous case). While not immediately necessary, it is beneficial to state Assumptions 2 now instead of later, as they will be necessary to obtain the final results presented in this report.

With these assumptions in place, we will soon present an equivalence lemma between equations (1) and (3). However, before we proceed, we need to introduce some notions and results.

To begin, we have a proposition that provides useful information about the function F:

Proposition A.1 For all $\mu \in \mathcal{M}_+(\Omega)$, there is $\nu \in \mathcal{P}(\Omega)$ such that $F(\mu) = F(\nu)$.

Proof. If $|\mu| = 0$, then

$$F(\mu) = R(\int \Phi d\mu) + \int V d\mu = R(0),$$

and for any $\theta_0 \in \Theta$, we have

$$F(\delta_{(0,\theta_0)}) = R(\int \Phi d\delta_{(0,\theta_0)}) + \int V d\delta_{(0,\theta_0)} = R(0 \cdot \phi(\theta_0)) + V(0,\theta_0) = R(0).$$

Otherwise, let $T : (\omega, \theta) \mapsto (|\mu|(\Omega) \cdot \omega, \theta)$, and set $\nu = T_{\#}(\mu/|\mu|(\Omega)) \in \mathcal{P}(\Omega)$.
This gives us:

$$F(\nu) = R(\int \Phi d\nu) + \int V d\nu = R(\int \Phi(T(\omega,\theta)) d\frac{\mu}{|\mu|(\Omega)}) + \int V(T(\omega,\theta)) d\frac{\mu}{|\mu|(\Omega)} =$$

$$R(\int \Phi d\mu) + \int V d\mu = F(\mu).$$

□

We now introduce an operator that establishes a connection between the original space and the lifted space. We define $h^1 : \mathcal{M}_+(\Omega) \to \mathcal{M}(\Theta)$ which satisfies $h^1(\mu)(B) = \int_{\mathbb{R}} w\mu(dw, B)$ for all $\mu \in \mathcal{P}(\Omega)$ and measurable $B \subset \Theta$. Equivalently, we can characterize h^1 using the following property: for all continuous and bounded test function $\phi : \Theta \to \mathbb{R}$: $\int_\Theta \phi(\theta) dh^1(\mu)(\theta) = \int_{\mathbb{R}\times\Theta} \omega\phi(\theta) d\mu(\omega,\theta)$. It is important to note that this operator is obviously well define only if $(\omega, \theta) \mapsto \omega$ is μ-integrable. With this definition, we can now present the equivalence theorem for these two problems, subject to certain conditions.

Lemma A.2 (Equivalence under lifting) It holds $\mathcal{M}(\Theta) \subset h^1(\mathcal{P}(\Omega)) = h^1(\mathcal{M}_+(\Omega))$. For a regularizer G on $\mathcal{M}(\Theta)$ of the form $G(\mu) = \inf_{\nu \in h^{-1}(\mu)} \int_\Omega V d\nu$, it holds $\inf_{\nu\in\mathcal{M}(\Theta)} J(\nu) = \inf_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$. If the infimum defining G is attained and if $\nu \in \mathcal{M}(\Theta)$ minimizes J, then there exists $\mu \in h^{-1}(\nu)$ that minimizes F over $\mathcal{M}_+(\Omega)$.

Proof. A measure $\nu \in \mathcal{M}(\Theta)$ can be written as $\nu = f\sigma$ where $\sigma \in \mathcal{P}(\Theta)$, and $f \in L^1(\sigma)$, for example by taking $\sigma = \frac{\nu}{|\nu|(\Omega)}$ whenever $|\nu|(\Omega) \neq 0$, and f is given by the Radon-Nykodim theorem.

Now let $\mu = (f \times id)_{\#}\sigma \in \mathcal{P}(\Omega)$. We have, for any $B \in \Theta$ measurable,

$$h^1(\mu)(B) = \int_{\mathbb{R}} w d\mu(w, B) = \int_{\mathbb{R}\times\Theta} w \mathbb{1}_B(\theta) d\mu(w,\theta) = \int_\Theta f(\theta)\mathbb{1}_B(\theta) d\sigma(\theta) = \int_B d\nu = \nu(B).$$

This means that $\nu = h^1(\mu)$, and thus h^1 is surjective.

Now, take μ such that $\nu = h^1(\mu)$. Then,

$$\int \phi d\nu = \int \phi dh^1(\mu) = \int_\Theta \phi(\theta) dh^1(\mu)(\theta) = \int_{\mathbb{R}\times\Theta} w\phi(\theta) d\mu(w,\theta) = \int_\Omega \Phi d\mu$$

As $G(\nu) = inf_{\mu\in h^{-1}(\nu)} \int_\Omega V d\mu$, it follows that

$$F(\mu) = R(\int \Phi d\mu) + \int V d\mu = R(\int \phi d\nu) + \int V d\mu \geq R(\int \phi d\nu) + G(\nu) = J(\nu)$$

With the same reasoning, it comes that $\inf_{\mu \in h^{-1}} F(\mu) = J(\nu)$, and so $\inf_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) = \inf_{\nu \in \mathcal{M}(\Theta)} J(\nu)$.

In the case where the infimum of G is achieved, and the infimum of J is achieved, the we have a minimum on F. □

We claim that if μ is a minimizer of F, then $h^1(\mu)$ is a minimizer of J. This result is crucial as our objective is to solve the optimization problem defined by F, but we are interested in finding a solution for J. Let's provide a brief proof of this claim:

Proof. If μ is such that $F(\mu) = F^*$, let $\nu = h^1(\mu)$. We will show that $J(\nu) = J^*$.

$$J(\int_{\mathbb{R}} \omega \mu(d\omega, \cdot)) = R(\int \phi dh^1(\mu)) + G(h^1(\mu)) = R(\int_\Omega \omega \phi d\mu) + \inf_{\alpha \in h^{-1}(h^1(\mu))} \int_\Omega V d\alpha$$

This is because ϕ is continuous and bounded. The last term is equal to $R(\int_\Omega \Phi d\mu) + \inf_{\alpha \in h^{-1}(h^1(\mu))} \int_\Omega V d\alpha$. Therefore, it suffices to prove that $\inf_{\alpha \in h^{-1}(h^1(\mu))} \int_\Omega V d\alpha = \int_\Omega V d\mu$.

Since $F^* = J^* \leq J(\nu)$, we have $\inf_{\alpha \in h^{-1}(h^1(\mu))} \int_\Omega V d\alpha \geq \int_\Omega V d\mu$. On the other hand, we know that $\mu \in h^{-1}(h^1(\mu))$, which implies $\inf_{\alpha \in h^{-1}(h^1(\mu))} \int_\Omega V d\alpha \leq \int_\Omega V d\mu$. Therefore, we conclude that $J(\nu) = J^*$. □

Finally, we will demonstrate that our setting is not overly restrictive. Specifically, we will show that one of the main regularizers we intend to consider, the total variation norm, satisfies the conditions of Proposition A.2.

Proposition A.3 (Total variation) Let $V(\omega, \theta) = |\omega|$. For $\mu \in \mathcal{M}(\Omega)$, it holds $\int V d\mu \geq |h^1(\mu)|(\Theta)$ with equality if for instance, μ is a lift of $h^1(\mu)$ of the form $(f \times id)_\# \sigma$ (where $h^1(\mu) = f\sigma$ and $\sigma \in \mathcal{P}(\Theta)$).

Proof. Let $\mu \in \mathcal{P}(\Omega)$ and $\nu = h^1(\mu)$. We set $\tilde{\nu}_+ = \int_{\mathbb{R}_+} \omega \mu(d\omega, \cdot)$ and $\tilde{\nu}_- = -\int_{\mathbb{R}_-} \omega \mu(d\omega, \cdot)$. We have that $\nu = \tilde{\nu}_+ - \tilde{\nu}_-$, and by definition of total variation norm:

$$|\nu|(\Theta) = |\nu_+|(\Theta) + |\nu_-|(\Theta) \leq |\tilde{\nu}_+|(\Theta) + |\tilde{\nu}_-|(\Theta) = \int_{\mathbb{R}} |\omega| d\mu(\cdot, \Theta) = \int V d\mu$$

In addition, taking the notation from proposition A.3, if $\mu = (f \times id)_\# \sigma$, then:

$$\tilde{\nu}_+(B) = \int_{\mathbb{R} \times \Theta} \omega \mathbb{1}_{\mathbb{R}_+}(\omega) \mathbb{1}_B(\theta) d\mu(\omega, \theta) = \int_\Theta f(\theta) \mathbb{1}_{\mathbb{R}_+}(f(\theta) = \int_\Theta f(\theta) \mathbb{1}_{\mathbb{R}_+}(f(\theta)) \mathbb{1}_B(\theta) d\sigma$$

$$= \int_{B \cap f^{-1}(\mathbb{R}_+)} f(\theta) d\sigma = \nu(B \cap f^{-1}(\mathbb{R}_+))$$

In a similar way, we have $\tilde{\nu}_-(B) = \nu(B \cap f^{-1}(\mathbb{R}_-))$. This means that $spt(\tilde{\nu}_+) \cap spt(\tilde{\nu}_-)$ has $|\nu|$-measure 0. Then, we have a result that allows us to conclude that we have equality in the previous inequality (see [7]), ending the proof. □

Part B: Study of the many-particle limit, and Wasserstein gradient flow

The objective of this part is to show that particle gradient flow converges to a Wasserstein gradient flow and it is articulated around three main results:

- The first result establishes the uniqueness of gradient flows.
- The second result guarantees the uniqueness of Wasserstein gradient flows.
- The third result demonstrates the convergence of a sequence of gradient flows to a Wasserstein gradient flow.

B.1 Particle gradient flow

We start this section with some fundamental definition:

Given $f : \mathbb{R}^d \to \bar{\mathbb{R}}$, which may not be convex. **A subgradient** of f is defined as follows: For any $u_0 \in \mathbb{R}^d$, a vector $p \in \mathbb{R}^d$ is a subgradient if for all $u \in \mathbb{R}^d$, the inequality $f(u) \geq f(u_0) + p \cdot (u - u_0) + o(u - u_0)$ holds. We will denote $\partial f(u)$ the set of subgradients of f at u and we will call it **the subdifferential** of f at u.

An important property of the subdifferential (that can be found in [20]) is that $\partial f(u)$ is a closed convex set.

Let's introduce some notations. We will denote $u_m : \mathbb{R} \to \Omega^m$ to represent a function, where $m \in \mathbb{N}$. Additionally, for each evaluation of u_m, which corresponds to a point in Ω^m, we will associate a measure denoted as $\mu_{m,t} = \frac{1}{m} \sum_{i=1}^m \delta_{u_i(t)}$, and vice versa.

We define $F_m : \Omega^m \to \mathbb{R}$ as $F_m(u) = F(\frac{1}{m} \sum_{i=1}^m \delta_{u_i}) = R(\frac{1}{m} \sum_{i=1}^m \Phi(u_i)) + \frac{1}{m} \sum_{i=1}^m V(u_i)$.

Now, it is time to introduce the central object of the section:

Definition: (Particle gradient flow). A gradient flow for a functional F_m is an absolutely continious path $u : \mathbb{R}_+ \to \Omega^m$ which satisfies $u'(t) \in -m\partial F_m(u(t))$ for every $t \geq 0$.

We will use the following notation:

— We denote $R'(f) \in \mathcal{F}$ as the gradient of R at $f \in \mathcal{F}$.

— The differential $d\Phi_u$ evaluated at the j-th vector of the canonical basis of $\mathbb{R}^d$ will be denoted as $\partial_j \Phi(u) \in \mathcal{F}$.

Now, we can state the main proposition of this part:

Proposition B.1 For any initialization $u(0) \in \Omega^m$, there exists a unique gradient flow $u : \mathbb{R}_+ \to \Omega^m$ for F_m. Moreover, for almost every $t > 0$, it holds $\frac{d}{ds} F_m(u(s))|_{s=t} = -\frac{1}{m}|u'(t)|^2$ and the velocity of the i-th particlee is given by $u_i'(t) = v_t(u_i(t))$, where for $u \in \Omega$ and $\mu_{m,t} = \frac{1}{m} \sum_{i=1}^m \delta_{u_i(t)}$, $v_t(u) = \tilde{v}_t(u) - proj_{\partial V(u)}(\tilde{v}_t(u))$ with $\tilde{v}_t(u) = -[\langle R'(\int \Phi d\mu_{m,t}), \partial_j \Phi(u) \rangle]_{j=1}^d$.

Proof. F_m is a sum of a continuously differentiable function and semi-convex function, hence F_m is locally semi-convex. For such a function, the existence of a unique gradient flow on a maximal interval $[0, T[$ is classic, see [21].

A general property of gradient flows (stated in [21]) is that for almost every $t \in \mathbb{R}_+$, $u \in \Omega$, the derivative is minus the subgradient of minimal norm, so $v_t(u) = argmin\{|v|^2; \tilde{v}_t(u) - v \in \partial V(u)\} = \tilde{v}_t(u) - argmin\{|\tilde{v}_t(u) - z|^2 : z \in \partial V(u)\} = (id - proj_{V(u)})(\tilde{v}_t(u))$ where we used the characterization of the projection, which is well defined as $\partial V(u)$ is a convex closed set.
We are left to prove that the flow is global (i.e $T = \infty$).

$$F_m(u(0)) - F_m(u(t)) = -\int_0^t \frac{d}{ds} F_m(u(s))ds = \frac{1}{m}\int_0^t |u'(0)|^2 ds = \frac{t}{m}\int_0^t |u'(0)|^2 \frac{ds}{t}$$

$$\geq \frac{t}{m}(\int_0^t |u'(0)|\frac{ds}{t})^2 = \frac{1}{tm}(\int_0^t |u'(0)|ds)^2 \geq \frac{1}{tm}|u(t) - u(0)|^2$$

As F_m is lower-bounded (nonegative, as $V, R \geq 0$). We have that $tmF_m(u(0)) \geq |u(t) - u(0)|^2$. Hence if we fixed $T > 0$ then u would be bounded. If $T < \infty$, by compacity, we can extract a sequence $t_n \to T$ such that $u(t_n)$ converges (to l). We can then consider the gradient flow w for F_m with $w(0) = l$. We can extend u, which contradict the maximality of T. So $T = +\infty$. □

Notice that in the easy case when V is differentiable, we have that the velocity v_t is given by $-\nabla F$.

B.2 Wasserstein gradient flow

In this section, we build upon the concepts discussed earlier and explore the application of the general theory of Wasserstein gradient flows. This theory, developed in [2], provides a framework that extends our analysis to infinite dimensions. However, before proceeding, we need to formalize the differential framework, we have established in the finite dimensional case, to F.

Let's define the differential of F which is represent at every $\mu \in \mathcal{M}(\Omega)$ by $F'(\mu) : \Omega \to \mathbb{R}$, defined as:

$$F'(\mu)(u) \stackrel{\text{def}}{=} \langle R'(\int \Phi d\mu), \Phi(u)\rangle + V(u).$$

We can easily verify that for v_t in Proposition B.1, we have the reassuring fact that $v_t \in -\partial F'(\mu_{m,t})$. The subdifferential of F' will be referred to as the **Wasserstein subdifferential** of F, as it can be viewed as the subdifferential of F with respect to the Wasserstein distance on $\mathcal{P}_2(\Omega)$.

Definition: (Wasserstein gradient flow) A Wasserstein gradient flow for the functional F on a time interval $[0, T[$ is an absolutely continuous path $(\mu_t)_{t\in[0,T[}$ in $\mathcal{P}(\Omega)$ that satisfies, distributionally on $[0, T[\times\Omega^d$, $\partial_t\mu_t = -\text{div}(v_t\mu_t)$ where $v_t \in -\partial F'(\mu_t)$.

We will refer to the equation $\partial_t \mu_t = -\text{div}(v_t \mu_t)$ as the **continuity equation**. This equation represents the conservation of mass for a time-dependent measure under the action of a velocity field. However, there is distribution which do not have a smooth density. For these distributions, there is still a continuity equation. Specifically, for all test functions $\phi : [0, \infty[\times \mathbb{R}^d \to \mathbb{R}$ that are smooth and have compact support, we have the following equation:

$$\int_0^\infty \int_{\mathbb{R}^d} (\partial_t \phi_t(u) + \nabla_u \phi_t(u) \cdot v_t(u)) d\mu_t(u) dt = 0 \, .$$

It should also hold that for all $t_0 < T$: $\int_0^{t_0} \int_{\mathbb{R}^d} |v_t(u)| d\mu(u) dt < \infty$.

Through the following lemma, we will perform a sanity check to verify that the measure $(\mu_{m,t})_t$, which is in corresponds to a gradient flow u_m, is indeed a Wasserstein gradient flow.

Proposition B.2 If $u : \mathbb{R}_+ \to \Omega^m$ is a classical gradient flow of F_m, then $t \mapsto \mu_{m,t} = \frac{1}{m} \sum_{i=1}^m \delta_{u_i(t)}$ is a Wasserstein gradient flow of F.

Proof. For v_t the velocity field defined previously, $t \mapsto \mu_{m,t}$ is clearly absolutely continuous (for more detail see [2]).

In addition for all infinitely differentiable function $\phi :]0, \infty[\times \Omega \to \mathbb{R}$ with compact support, we have that $0 = \int_{\mathbb{R}_+} \frac{d}{dt} \phi_t(u_i(t)) dt = \int_{\mathbb{R}_+} \partial_t \phi_t(u) + \nabla_u \phi_t(u_i(t)) \cdot v_t(u_i(t)) dt$ for all $i = 1, \dots, m$. Hence $0 = \frac{1}{m} \sum_{i=1}^m \int_{\mathbb{R}_+} \partial_t \phi_t(u) + \nabla_u \phi_t(u_i(t)) \cdot v_t(u_i(t)) dt = \int_{\mathbb{R}_+} \int_\Omega \partial_t \phi_t(u) + \nabla_u \phi_t(u_i(t)) \cdot v_t(u_i(t)) d\mu_{m,t} dt$. Which allow us to conclude. □

An important property of F is that it is continuous with respect to the W_2 metric, which is atypical in the context of Wasserstein gradient flows.

Lemma B.3 (Wasserstein continuity of F) Under Assumptions 1, the function F is continuous for the Wasserstein metric W_2.

Proof. Recall that for $(\mu_m)_{m \in \mathbb{N}}$, $\mu \in \mathcal{P}_2(\Omega)$ such that $\mu_m \xrightarrow{W_2} \mu$, is equivalent to: for all ϕ continuous with at most quadratic growth, $\int \phi d\mu_m \to \int \phi d\mu$.
Assumtions 1 (iii)-(c) implies that $|V|$ and $\|\Phi\|$ have at most quadratic growth. We have that $\int V d\mu_m \to \int V d\mu$ and $\| \int \Phi d\mu_m - \int \Phi d\mu \| \leq | \int \|\Phi\| d(\mu_m - \mu)|$ (which holds by a property of Bochner integral [7]), hence $\int \Phi d\mu_m \xrightarrow{\mathcal{F}} \int \Phi d\mu$. As R is continuous, we have that $F(\mu_m) \to F(\mu)$, so F is continuous. □

Recalling the sets Q_r introduced in Assumptions 1, we now introduce:

$$F^{(r)}(\mu) = \begin{cases} F(\mu) & \text{if } \mu(Q_r) = 1 \\ \infty & \text{otherwise} \end{cases} .$$

We will use $F^{(r)}$ as a tool to obtain results about F.

Before proving our first result, it is necessary to establish certain definitions and notations. For $r > 0$:

— Given $\gamma \in \mathcal{P}(\Omega \times \Omega)$, we say it's an **admissible transport plan** if both marginals law of are concentrated on Q_r and have finite second moments.

— **The transport cost** associated to γ is for $p \geq 1$: $C_p(\gamma) = (\int |y - x|^p d\gamma(x, y))^{\frac{1}{p}}$.

— $\mathcal{F}_r \overset{\text{def}}{=} \{\int \Phi d\mu; \mu \in \mathcal{P}(\Omega), \mu(Q_r) = 1\}$ which is bounded in $\mathcal{F}$.

— $\|d\Phi\|_{\infty,r} = \sup_{u \in Q_r} \|d\Phi_u\|$.

— $\|dR\|_{\infty,r} = \sup_{f \in \mathcal{F}_r} \|dR_f\|$.

— $L_{d\Phi} = \sup_{\substack{u,\tilde{u} \in Q_r \\ u \neq \tilde{u}}} \frac{\|d\Phi_{\tilde{u}} - d\Phi_u\|}{|\tilde{u} - u|}$.

— $L_{dR} = \sup_{\substack{f,g \in \mathcal{F}_r \\ f \neq g}} \frac{\|dR_f - dR_g\|}{|f - g|}$.

The quantities above are finite by Assumptions 1.

We are now ready to prove the first lemma which will gives us interesting technical properties about $F^{(r)}$. Note that for sake of simplicity in this lemma, we assume $V = 0$. This is because it is already well studied in [2].

Lemme B.4 (Properties of $F^{(r)}$ in Wasserstein geometry) Under Assumptions 1, suppose that $V = 0$. For all $r > 0$, $F^{(r)}$ is proper and continuous for W_2 on its closed domain. Moreover,

(i) there exists $\lambda_r > 0$ such that for all admissible transport plan γ, considering the transport interpolation $\mu_t^\gamma = ((1-t)\pi^1 + t\pi^2)_\# \gamma$, the function $t \mapsto F(\mu_t^\gamma)$ is differentiable with a $\lambda_r C_2^2(\gamma)$-Lipschitz derivative;

(ii) for μ concentrated on Q_r, a velocity field $v \in L^2(\mu, \mathbb{R}^d)$ satisfies, for any admissible transport plan γ with first marginal μ,

$$F(\pi^2_\# \gamma) \geq F(\mu) + \int v(u) \cdot (\tilde{u} - u) d\gamma(u, \tilde{u}) + o(C_2(\gamma))$$

if and only if $v(u) \in \partial(F'(\mu) + \iota_{Q_r})(u)$ for μ-almost every $u \in \Omega$ where ι_{Q_r} is the convex function on Ω that is worth 0 on Q_r and ∞ outside.

Proof. We will prove the two point separately. We first recall that by Holder: $C_1(\gamma) \leq C_2^2(\gamma)$.

Proof of (i). Let us denote $h(t) = F^{(r)}(\mu_t^\gamma)$. We will first show that $t \mapsto \int \Phi d\mu_t^\gamma = \int \Phi((1 - t)x + ty) d\gamma(x, y)$ is differentiable. The marginal laws of γ are concentrated on Q_r, so the $(\mu_t^\gamma)_t$ are also concentrated on Q_r.

On Q_r, $d\Phi$ is uniformly bounded (so integrable for γ as it is a probability measure) so with the dominated convergence theorem for Bochner integrals (see [7]), the function is differentiable. As R is also differentiable, we have that h is differentiable and $h'(t) = \langle R'(\int \Phi d\mu_t^\gamma), \int d\Phi_{(1-t)x+ty} (y - x) d\gamma \rangle$.

Now, we have that for $0 \leq t_1 < t_2 < 1$:

$$|h'(t_1) - h'(t_2)| \leq (I) + (II)$$

where

$$\begin{aligned}(I) &= |\langle R'(\int \Phi d\mu_{t_2}^{\gamma}) - R'(\int \Phi d\mu_{t_1}^{\gamma}), \int d\Phi_{(1-t_2)x+t_2y}(x-y)d\gamma(x,y)|\\ &\leq L_{dR}\|d\Phi\|_{\infty,r}|t_2-t_1|C_1(\gamma)C_1(\gamma)\|d\Phi\|_{\infty,r}\\ &\leq L_{dR}\|d\Phi\|_{\infty,r}^2|t_2-t_1|C_2^2(\gamma)\rangle\end{aligned}$$

and

$$\begin{aligned}(II) &= |\langle R'(\int \Phi d\mu_{t_1}), \int (d\Phi_{(1-t_2)x+t_2y} - d\Phi_{(1-t_1)x+t_1y})(y-x)d\gamma(x,y)|\\ &\leq L_{d\Phi}\|dR\|_{\infty,r}C_2^2(\gamma)|t_2-t_1|.\end{aligned}$$

So for $\lambda_r = L_{dR}\|d\Phi\|_{\infty,r}^2 + L_{d\Phi}\|dR\|_{\infty,r}$, h' is $\lambda_r C_2^2(\gamma)$-Lipschitz.

Proof of (ii). For all $u, \tilde{u} \in Q_r$ and $f, g \in \mathcal{F}_r$, $\Phi(\tilde{u}) = \Phi(u) + d\Phi(\tilde{u}-u) + M(u,\tilde{u})$ and $R(g) = R(f) + \langle R'(f), g-f\rangle + N(f,g)$ where $\|M(u,\tilde{u})\| \leq \frac{1}{2}L_{d\Phi}|\tilde{u}-u|^2$ and $\|N(f,g)\| \leq \frac{1}{2}L_{dR}\|f-g\|^2$ because $d\Phi$ and dR are Lipschitz on Q_r and $\mathcal{F}_r$.

Using the notation $\mu = \pi_{1\#}\gamma$ and $\nu = \pi_{2\#}\gamma$, both concentrated on Q_r, we have, by composition:

$$\begin{aligned}\int \Phi d\nu &= \int \Phi(\tilde{u})d\nu = \int \Phi(\tilde{u})d\gamma(u,\tilde{u}) = \int (\Phi(u) + d\Phi_u(\tilde{u}-u) + M(u,\tilde{u}))d\gamma(u,\tilde{u})\\ &= \int \Phi(u)d\mu + \int d\Phi_u(\tilde{u}-u)d\gamma(u,\tilde{u}) + \int M(u,\tilde{u})d\gamma(u,\tilde{u}).\end{aligned}$$

So,

$$\begin{aligned}F^{(r)}(\nu) &= R(\int \Phi d\mu + \int \Phi d(\nu-\mu))\\ &= F^{(r)}(\mu) + \langle R'(\int \Phi d\mu), \int \Phi d(\nu-\mu)\rangle + N(\int \Phi d\mu, \int \Phi d\nu)\\ &= F^{(r)}(\mu) + \langle R'(\int \Phi d\mu), \int d\Phi_u(\tilde{u}-u)d\gamma(u,\tilde{u})\rangle + (I) + (II).\end{aligned}$$

Where:

$$(I) = \langle R'(\int \Phi d\mu), \int M(u,\tilde{u})d\gamma(u,\tilde{u})\rangle \leq \|dR\|_{\infty,r}\frac{1}{2}L_{d\Phi}C_2^2(\gamma) = o(C_2(\gamma)),$$

and

$$\begin{aligned}(II) &= N(\int \Phi d\mu, \int \Phi d\nu)\\ &\leq \frac{1}{2}L_{dR}\|\int \Phi d(\mu-\nu)\|^2\\ &\leq \frac{1}{2}L_{dR}\|\int d\Phi_u(\tilde{u}-u)d\gamma(u,\tilde{u}) + \int M(u,\tilde{u})d\gamma(u,\tilde{u})\|^2\\ &\leq \frac{1}{2}L_{dR}(\|d\Phi\|_{\infty,r}C_1(\gamma) + \frac{1}{2}L_{d\Phi}C_2^2(\gamma))^2\\ &\leq \frac{1}{2}L_{dR}(\|d\Phi\|_{\infty,r} + \frac{1}{2}L_{d\Phi})^2C_2^4(\gamma) = o(C_2(\gamma)).\end{aligned}$$

Hence $F^{(r)}(\nu) = F^{(r)}(\mu) + \int \langle R'(\int \Phi d\mu), d\Phi_u(\tilde{u} - u)\rangle d\gamma(u, \tilde{u}) + o(C_2(\gamma))$.

As a result, in the inside of Q_r, we have a well defined velocity field $v(u)$. Recalling that $(\nabla F'(\mu)(u))_i = \langle R'(\int \Phi d\mu), d\Phi_u(e_i)\rangle$, we have that $v(u) \in \partial F'(\mu)$. On the boundary of Q_r, we have more choice as the constraint is that $\pi_{2\#}\gamma$ is concentrated on Q_r. So $v(u) - \nabla F'(\mu)(u)$ can be the normal cone of Q_r, i.e $\partial \iota_{Q_r}$. The condition is hence relaxed into $v(u) \in \partial(F'(\mu) + \iota_{Q_r})(u)$. □

We claim that this guarantees that Wasserstein gradient flows for the functionals $F^{(r)}$ are well defined.

Lemme B.5 Under Assumptions 1, there exists a unique Wasserstein gradient flow for $F^{(r)}$ starting from any $\mu_0 \in \mathcal{P}_2(\Omega)$ concentrated on Q_r, i.e a curve $(\mu_t^{(r)})_{t\geq 0}$ continuous in $\mathcal{P}_2(\Omega)$, that solves $\partial_t \mu_t^{(r)} + div(v_t^{(r)} \mu_t^{(r)}) = 0$ where, for all $t \geq 0$, $v_t^{(r)} \in \partial(F'(\mu_t^{(r)})(u) + \iota_{Q_r}(u)$, for $\mu_t^{(r)}$-a.e $u \in \Omega$.

Proof. If V is λ_V-semiconvex, then the function $\mu \mapsto \int V d\mu$ is λ_V-semiconvex along generalized geodesics (see [2]). With lemma B.4 (i), we have that $F^{(r)}$ is $(\lambda_V + \lambda_r)$-semiconvex along generalized geodesics.
In addition to that lemma B.4 (ii) guarantee that $F^{(r)}$ admits strong Wasserstein sub-diffential (in the sense of [2]). And again, it is an easy adaptation to show that (ii) still holds with a potential term. So the existence of a unique Wasserstein gradient flow characterized as above is guaranteed (see [2]). □

With this lemma, we are now prepared to prove the main result for F. Notice that, according to the characterization in Lemma B.5, the Wasserstein gradient flows for the functions $F^{(r)}$ all coincide for $r > r_0 > 0$ on $[0, T]$ if $\mu_t^{(2r_0)}$ is concentrated on Q_{r_0} for all $t \in [0, T]$. In the following proposition, our strategy is therefore to ensure that for all times T, such an r_0 exists. In other words, we aim to ensure that the support of the gradient flows does not grow too fast.

Proposition B.6 (Existence and uniqueness) Under Assumptions 1, if $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on a set $Q_{r_0} \subset \Omega$, then there exists a unique Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ for F starting from μ_0. It satisfies the continuity equation with the velocity field defined in Proposition B.1 (with μ_t in place of $\mu_{m,t}$).

Proof. Let r_0 such that μ_0 is concentrated on Q_{r_0}. With lemma B.5, for all $r > r_0$, there exists a unique, globally defined, Wasserstein gradient flow $(\mu_t^{(r)})_{t\geq 0}$ for $F^{(r)}$.

Set $t_r = \inf\{t > 0 : \mu_t^{(2r)}(Q_r) < 1\}$. Hence, if $t_r > 0$, we have existence of a Wasserstein gradient flow on $[0, t_r]$. As a result, we are left with showing $t_r \to +\infty$ as $r \to +\infty$.

With the result we have on $v^{(r)}$ in the lemma B.4 (ii), for all $0 \leq t \leq t_r$, we have that $v_t^{(r)} \in \partial F'(\mu_t^{(r)})$ in the sens of $L^2(\mu_t^{(r)}, \mathbb{R}^d)$. Hence with Assumption 1 (iii)-(c) and as dR is bounded on sublevels of R, we have that for $0 \leq t \leq t_r$:

$$|v_t^{(r)}| \leq C_1 + C_2 r$$

where C_1 and C_2 are independent of u, r and t.

Indeed, $F'(\mu_t^{(r)}) : u \mapsto \langle R'(\int \Phi d\mu_t^{(r)}), \Phi(u)\rangle + V(u)$ which leads to $v_t^{(r)} \in \partial V(u) + \langle R'(\int \Phi d\mu_t^{(r)}), d\Phi_u\rangle$. Hence $|v_t^{(r)}| \leq \|\partial V(u)\| + \|dR((\int \Phi d\mu_t^{(r)})\|\|d\Phi_u\|$.

But dR is bounded on sub-levels of R, and as $\mu_t^{(r)}$ is a gradient flow for F, $t \mapsto F(\mu_t^{(r)})$ is decreasing, so we stay on the same sub-level (independently of r) along the dynamics which allow us to conclude.

With Grönwall lemma, we deduce that $\mu_t^{(r)}$ is concentrated on $\{u \in \Omega : dist(u, Q_{r_0}) \leq (r_0 + \frac{C_1}{C_2})e^{tC_2}\}$.

As a result, for all $T > 0$, there exists $r > 0$ such that $t_r > T$. □

With this proposition in hand, we can take advantage of the opportunity to present a result that will be useful in the subsequent steps of this report.

Lemma B.7 (Representation of the flow) Under the assumptions of Proposition B.6, let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow of F and (v_t) the associated velocity fields. Consider the flow $X : \mathbb{R} \times \Omega \to \Omega$ which for all $u \in \Omega$, is an absolutely continuous solution to

$$X(0, u) = u \text{ and } \partial_t X(t, u) = v_t(X(t, u)) \text{ for a.e. } t \geq 0.$$

Then X is uniquely well-defined, continuous, $X(t, \cdot)$ is Lipschtiz on Q_r, uniformly on compact time intervals for all $r > 0$, and it holds $\mu_t = (X_t)_{\#}\mu_0$.

Proof. The claims concerning X are classical and follow from the fact that v_t satisfies a one-sided Lipschitz property on Q_r, uniformly on compact time intervals (by [2]). The expression as a pushforward is also a general property of the continuity equation (see [2]). □

B.3 Many-particle limit

We will now combine the two aforementioned results into a more general one. We will prove the main result of part B, which demonstrates the convergence of a sequence of gradient flows to a Wasserstein gradient flow. It is worth noting that this proof provides an alternative argument for the existence of a Wasserstein gradient flow (Proposition B.6 is still useful for establishing uniqueness).

Theorem B.8 (Many-particle limit). Consider $(t \mapsto u_m(t))_{m\in\mathbb{N}}$ a sequence of classical gradient flows for F_m initialized in a set $Q_{r_0} \subset \Omega$. If $\mu_{m,0}$ converges to some $\mu_0 \in \mathcal{P}_2(\Omega)$ for the Wasserstein distance W_2, then $\mu_{m,t}$ converges weakly, as $m \to \infty$, to the unique Wasserstein gradient flow of F starting from μ_0.

Proof. We will prove it in four steps.

Step (i): Let $t_r = \inf\{t > 0 : \exists m \in \mathbb{N}, \mu_{m,t}(Q_r) < 1\}$. We would like to show that $t_r > 0$. With

Assumptions 1 (iii)-(a), in order to leave Q_r from Q_{r_0}, the minimal distance is $r - r_0$. So if we remind the velocity of each particle:

$$v_t(u) = (id - proj_{\partial V(u)}(\tilde{v}_t(u))$$

We get that for $u \in Q_r$, $|v_t(u)| \leq L_{V,r} + \|dR\|_{\infty,r}\|d\Phi\|_{\infty,r}$. It then follows that $t_r \geq \frac{r-r_0}{L_{V,r}+\|dR\|_{\infty,r}\|d\Phi\|_{\infty,r}} > 0$.

Step (ii): We know work in $[0, t_r]$. Let show the existence of a limit curve: $t \mapsto \mu_t$ in $\mathcal{P}_2(\Omega)$.
We have

$$\begin{aligned} W_2(\mu_{m,t_1}, \mu_{m,t_2})^2 &\leq \frac{1}{m}\sum_{i=1}^{m} |u_{m,i}(t_2) - u_{m,i}(t_1)|^2 \\ &\leq \frac{1}{m}\sum_{i=1}^{m} (\int_{t_1}^{t_2} |u'_{m,i}(s)|ds)^2 \\ &\leq \frac{(t_2-t_1)^2}{m}\sum_{i=1}^{m} (\int_{t_1}^{t_2} |u'_{m,i}(s)|\frac{ds}{t_2-t_1})^2 \\ &\leq \frac{(t_2-t_1)}{m}\sum_{i=1}^{m} \int_{t_1}^{t_2} |u'_{m,i}(s)|^2 ds. \end{aligned}$$

But as $\frac{d}{ds}F(\mu_{m,t}) = -\frac{1}{m}|u'(t)|^2 = -\frac{1}{m}\sum_{i=1}^{m} |u'_{m,i}(t)|^2$, we then have

$$\begin{aligned} W_2(\mu_{m,t_1}, \mu_{m,t_2})^2 &\leq (t_2 - t_1)\int_{t_1}^{t_2} -\frac{d}{ds}F(\mu_{m,t})ds \\ &\leq (t_2 - t_1)(F(\mu_{m,t_1}) - F(\mu_{m,t_2}) \\ &\leq (t_2 - t_1)(\sup_{m\in\mathbb{N}} F(\mu_{m,0}) - \inf_{\mu\in\mathcal{P}_{\Omega)}} F(\mu)). \end{aligned}$$

Hence the curve family $(t \mapsto \mu_{m,t})_m$ is equicontinuous on $[0, t_r]$ for W_2, uniformly in m. In addition, for all $t \in [0, t_r]$, the family $(\mu_{m,t})$ is in a W_2 ball:

$$W_2(\mu_0, \mu_{m,t}) \leq W_2(\mu_0, \mu_{m,0}) + W_2(\mu_{m,0}, \mu_{m,t}) \leq W_2(\mu_0, \mu_{m,0}) + t(\sup_{m\in\mathbb{N}} F(\mu_{m,0}) - \inf_{\mu\in\mathcal{P}_2\Omega} F(\mu))$$

Which is bounded independently of m as $W_2(\mu_0, \mu_{m,0}) \to 0$ as $m \to \infty$. As $\|\cdot\|_{BL} \leq W_2$, the ball is weakly precompact.

By Ascoli theorem, we can then extract a subsequence which converges weakly to the curve $(\mu_t)_{t\geq 0}$, continuous for the weak topology, which is so also concentrated on Q_r. We also have uniform convergence in the bounded Lipschitz metric.
For the following parts, we will denote $(\mu_m)_m$ for the sub-sequence.

Step (iii): We now need to check that the limit (μ_t) satisfies the continuity equation.

Let's consider $v_{m,t}$ the velocity fields of the particle at time t, and let's set $v_t(u) = (id -$

$proj_{\partial V(u)}(\tilde{v}_t(u))$ where $\tilde{v}_t(u) = -[\langle R'(\int \Phi d\mu_t), \partial_i \Phi(u)\rangle]_{i=1}^d$. We now set $E_m = v_{m,t}\mu_{m,t}dt$ and $E = v_t\mu_t dt$.

Then for all bounded and continuous functions $\phi : [0, t_r] \times \mathbb{R}^d \to \mathbb{R}^d$, we have

$$\begin{aligned}|\int \phi d(E_m - E)| &\leq |\int \phi \cdot (v_{m,t} - v_t) d\mu_{m,t}dt + \int \phi \cdot v_t d(\mu_{m,t} - \mu_t)dt| \\ &\leq \|\phi\|_\infty \int |v_{m,t} - v_t| d\mu_{m,t}(u)dt + |\int \phi \cdot v_t d(\mu_{m,t} - \mu_t)dt|.\end{aligned}$$

We will now show that both those terms go to 0 at the limit.
For the first one, as the $\mu_{m,t}$ are concentrated on Q_r, we only need to show that the sequences of $(t, u) \mapsto v_{m,t}(u)$ converges uniformly to $(t, u) \mapsto v_t(u)$ on $[0, t_r] \times Q_r$. Using the fact that the projection over a convex closed set is 1-Lipshitz, we have

$$\begin{aligned}|v_{m,t} - v_t(u)| &\leq 2|v_{m,t} - \tilde{v}_t(u)| \\ &\leq \|d\Phi\|_{\infty,r} \|R'(\int \Phi d\mu_{m,t}) + R'(\int \Phi d\mu_t)\| \\ &\leq 2\|d\Phi\|_{\infty,r}\|dR\|_{\infty,r}\| \int \Phi d\mu_{m,t} - \int \Phi d\mu_t\| \\ &\leq 2\|d\Phi\|_{\infty,r}\|dR\|_{\infty,r} \sup_{f\in\mathcal{F}, \|f\|\leq 1} \int \langle f, \Phi(u)\rangle d(\mu_{m,t} - \mu_t)(u).\end{aligned}$$

But $\Phi / \max\{\|d\Phi\|_{\infty,1}, \|\Phi\|_{\infty,1}\}$ is bounded in norm by 1 and its Lipschitz constant is at most 1. So,

$$\begin{aligned}&\sup_{f\in\mathcal{F}, \|f\|\leq 1} \int \langle f, \Phi(u)\rangle d(\mu_{m,t} - \mu_t)(u) \\ &= \max\{\|d\Phi\|_{\infty,1}, \|\Phi\|_{\infty,1}\} \cdot \sup_{f\in\mathcal{F}, \|f\|\leq 1} \langle f, \int \frac{\Phi}{\max\{\|d\Phi\|_{\infty,1}, \|\Phi\|_{\infty,1}\}} d(\mu_{m,t} - \mu_t)\rangle \\ &\leq \max\{\|d\Phi\|_{\infty,1}, \|\Phi\|_{\infty,1}\} \cdot \|\mu_{m,t} - \mu_t\|_{BL}.\end{aligned}$$

As the convergence of $(t \mapsto \mu_{m,t})$ is uniform in the BL norm we can conclude.

For the second term, $(t, u) \mapsto v_t(u)$ is bounded over Q_r, and ϕ is bounded an d continuous, so the weak convergence tells us that the second term also goes to 0.

As a result, $E_m \to E$ weakly. In particular, the continuity equation goes to the limit.
In addition, $(v_t)_t$ is uniformly bounded in time over Q_r, so $\int_0^{t_r} \int_\Omega |v_t(u)|^2 d\mu_t dt < +\infty$, so (μ_t) is an absolutely continuous path for W_2.

Step (iv): Now let's show that $t_r \to +\infty$.
As $F(\mu_{m,0}) \to F(\mu_0)$ and that all path $(\mu_{m,t})_t$ decrease monotonously along F, everything is on a sub-level of R, where dR is bounded.
We get: $v_{m,t}| \leq \|\partial V(u)\| + \|dR\| \cdot \|d\Phi\|$ and similarly for v.

With Assumptions 1 (iii)-(c), we have a growth at most linear in r. By Grönwall's lemma, we deduce that $t_r \to +\infty$ as $n \to +\infty$.

Combining these result with the unicity previously obtained, this means that for every subsequence, we can extract a subsequence that converges to the Wasserstein gradient flow of F starting from μ_0. This implies convergence of the sequence, finishing the proof.

□

Part C: Convergence to the global minimizers

The objective of this section is to prove that if the Wasserstein gradient flow converges, then it converges to the global minima.

This is organized as follows:

- First, we give global optimality conditions.
- Then, we give a criteria for the gradient flow to escape neighbourhoods of non-optimal stationary points.
- After that, using topological degree theory, we show that some separation properties of the support of the measure are preserved throughout the dynamic.
- Finally, we prove the main theorem.

C.1 Optimality Conditions

Let us first better understand what characterizes global minimizers in our setting.

Proposition C.1 (Minimizers) Assume that R is convex. A measure $\mu \in \mathcal{M}_+(\Omega)$ such that $F(\mu) < \infty$ minimizes F on $M_+(\Omega)$ iff $F'(\mu) \geq 0$ and $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$.

Proof. For $\mu, \sigma \in \mathcal{M}(\Omega)$ such that $F(\mu), F(\sigma) < \infty$, we have that

$$\begin{aligned} F(\mu + t\sigma) &= R(\int \Phi d\mu + t \int \Phi d\sigma) + \int V d\mu + t \int V d\sigma \\ &= F(\mu) + t(\langle R'(\int \Phi d\mu), \int \Phi d\sigma \rangle + \int V d\sigma) + o(t \int V d\sigma). \end{aligned}$$

So $\frac{d}{dt} F(\mu + t\sigma)|_{t=0} = \int_w F'(\mu) d\sigma$, for $F'(\mu) : u \mapsto \langle R(\int \Phi d\mu, \Phi(u) \rangle + V(u)$.
It is well defined as $\int |F'(\mu)| d\sigma \leq \|R(\int \Phi d\mu)\|, \int \|\Phi\| d\sigma + \int \|V\| d\sigma < +\infty$ as $F(\mu), F(\sigma) < +\infty$.

Let now $\mu, \nu \in \mathcal{M}_+(\Omega)$ with $\nu - \mu, F(\mu), F(\nu) < +\infty$, and consider $\sigma = \nu - \mu$. Suppose that μ is a minimizer of F. For ν as previously, $\mu + t\nu \in \mathcal{M}_+(\Omega)$ for all $t > 0$. So the necessary condition of optimality is $\int_\Omega F'(\mu) d\nu \geq 0$. Testing it on $\mu = \delta_u$, $u \in \Omega$, we have that $F'(\mu) \geq 0$

on Ω. For $\nu = \mu$, $t \in]-1, 1[$ is allowed so we have that $\int_w F'(\mu)d\mu = 0$ and so $F'(\mu) = 0$ μ-a.e with the previous point.

Those conditions are also sufficient. Indeed, we have that

$$\frac{d}{dt}F(\mu + t\sigma)|_{t=0} = \int F'(\mu)d\sigma = \int F'(\mu)d\nu - \int F'(\mu)d\mu \geq 0$$

Hence by convexity, $0 \leq \frac{d}{dt}F(\mu + t\sigma) \leq \frac{d}{dt}(F(\mu)(1-t) + F(\nu)t) \leq F(\nu) - F(\mu)$.

□

C.2 A criteria to escape from non-optimal stationary points

We will establish a criteria for Wasserstein gradient flow to avoid non-optimal stationary points.
We will denote by $\|\cdot\|_{C^1}$ the maximum of the supremum norm of a function and the supremum norm of its gradient.

Proposition C.2 (Criteria to escape local minima) Under Assumptions 2, let $\mu \in \mathcal{M}(\Omega)$ be such that $F'(\mu)$ is not nonnegative. Then there exists $\epsilon > 0$ and a set $A \subset \Omega$ such that if $(\mu_t)_t$ is a Wasserstein gradient flow of F satisfying $\|h^1(\mu) - h^1(\mu_{t_0})\|_{BL} < \epsilon$ for some $t_0 \geq 0$ and $\mu_{t_0} > 0$ then there exists $t_1 > t_0$ such that $\|h^1(\mu) - h^1(\mu_{t_1})\|_{BL} \geq \epsilon$.

Such a set is given by $A = (\mathbb{R}_+ \times K^+) \cup (\mathbb{R}_- \times K^-)$ where K^+ (respectively K^-) is the $(-\eta)$-sublevel set of $\theta \mapsto F'(\mu)(1, \theta)$ (respectively of $\theta \mapsto F'(\mu)(-1, \theta)$) for some $\eta > 0$ that can be chosen arbitrarily close to 0.

Proof. Suppose that $F'(\mu)$ only takes negative values on $\mathbb{R}_+ \times \Theta$ (the case $\mathbb{R}_- \times \Theta$ is similar).

We set $g_\mu : \Theta \to \mathbb{R}$ the restriction on $F'(\mu)$ to $\{1\} \times \Theta$, i.e $g_\mu(\theta) = \langle R'(\int \Phi d\mu), \phi(\theta)\rangle + \tilde{V}(\theta)$.

Let $-\eta < 0$ a negative regular value of g (whose existence is guaranteed by the hypothesis) and denote $K^+ \subset \Theta$ the associated sub level. Using the regular value theorem, we have $\partial K^+ = g_\mu^{-1}(-\eta)$ is a differentiable manifold of dimension $(d-2)$ and is orthogonal to the gradient field of g_μ.

If Θ is bounded, ∂K^+ is compact, so there exists $\beta > 0 : \inf_{\theta \in \partial K^+} |dg_\mu(\theta)| \geq \beta$.

If $\Theta = \mathbb{R}^{d-1}$ and K^+ is not bounded, we can choose η such that it is a regular value of the function from the sphere $\mathbb{S}^{d-2}$ to which g_μ uniformly converges at infinity. If so, the same type of bound exists. It follows that on K^+, $g_\mu \leq -\eta$ and $\nabla g_\mu(\theta) \cdot \vec{n_\theta} < -\beta$, for $\theta \in \partial K^+$, with $\vec{n_\theta}$ is the unit normal vector to ∂K^+ in θ.

Let $C_0 > 0$ fixed big enough, and consider the measures ν such that $\|h^1(\nu)\|_{BL} < C_0$. For $\epsilon = \frac{\min\{\eta, \beta\}}{4\alpha\|\phi\|^2_{C_1}}$, for $\alpha > 0$ given by lemma C.3 (see below), if $\|h^1(\nu) - h^1(\mu)\|_{BL} < \epsilon$ then $g_\nu \leq \frac{-\eta}{2}$ over K and $\nabla g_\nu(\theta) \cdot \vec{n_\theta} < \frac{-\beta}{2}$ (because $\|g_\nu\| \leq \|g_\nu - g_\mu\| + \|g_\mu\|$ etc ...)

Let's now consider the Wasserstein gradient flow $(\mu_t)_t$ of F such that μ_0 in concentrated on $[-r_0, r_0] \times \Theta$ for $r_0 > 0$, and $\|h^1(\mu_0) - h^1(\mu)\|_{BL} < \epsilon$. As long as this stays true for μ_t, we also have that $\|h^1(\mu_t)\|_{BL} < C_0$, because $\|h^1(\mu_t)\|_{BL} \leq \|h^1(\mu)\|_{BL} + \epsilon$.

Let's denote $t_1 > 0$ the infinimum of time such that this conditions holds. Consider the flow X of the lemma B.7. By construction of K^+, every path from $t \mapsto (w_t, \theta_t) = X(t, (w_0, \theta_0))$ with $(w_0, \theta_0) \in \mathbb{R}_+ \times K^+$ stays in $\mathbb{R}_+ \times K_+$ for $t \leq t_1$.

Moreover, $F'(\mu_t)$ being homogeneous in the w variable, the condition $v_t \in \partial F'(\mu_t)$ implies that the component with respect to w of v_t is bounded by below by $-g_\mu$, and so by $\frac{\eta}{2}$. It follows (by integration) that $w_t \geq w_0 + t \cdot \frac{\eta}{2}$. This guarantee the path never enters in $\mathbb{R}_- \times K^+$.

The paths in $\mathbb{R}_- \times K^+$ verify $w_t \geq w_0 + t \cdot \frac{\eta}{2}$ as $F'(-1, \cdot) = -F'(1, \cdot)$. We then have:

$$\begin{aligned} h^1(\mu_t)(K^+) &= \int_{\mathbb{R}} w d\mu_t(w, K^+) \\ &= \int_{\mathbb{R}} w dX_{t\#}\mu_0(w, K^+) \\ &= \int w_t \mathbf{1}_{w_t > 0} d\mu_0(w, K^+) + \int w_t \mathbf{1}_{w_t < 0} d\mu_0(w, K^+) \, . \end{aligned}$$

Using that μ_0 is concentrated on $[-r_0, r_0] \times \Theta$, it follows:

$$\begin{aligned} &\geq t \cdot \frac{\eta}{2} \mu_0(\mathbb{R}_+ \times K^+) + \int w_t \mathbf{1}_{0 > w_t > -r_0} d\mu_0(w, K^+) \\ &\geq t \cdot \frac{\eta}{2} \mu_0(\mathbb{R}_+ \times K^+) + \min\{0, t \cdot \frac{\eta}{2} - r_0\} \mu_0(\mathbb{R}_- \times K^+) \, . \end{aligned}$$

Hence, as long as $\mu_0(\mathbb{R}_+ \times K^+) > 0$, $h^1(\mu_t)(K^+)$ grows at least linearly.

If $\Theta = K^+$, then we have that (using the constant to 1 function in the definition of the BL norm) $\|h^1(\mu_t)\|_{BL} \geq h^1(\mu_t)(K^+) \to +\infty$ (as $t \to \infty$), so $t_1 < +\infty$.

Otherwise, we consider another sub-level $\tilde{K}^+$ of g_μ for $\tilde{\eta} \in]-\eta, 0[$, and such that $\tilde{K}^+$ does not cover all Θ. As g_μ is Lipschitz, there exists $\Delta \in]0, 1[$ such that $d(K^+, \Theta \backslash \tilde{K}^+) \geq \Delta$. We can choose ϵ smaller if needed, in order to repeat the previous argument for $\tilde{\eta}$ and $\tilde{K}^+$.

By taking again the above inequality, we see that either $t_1 \leq \frac{2r_0}{\tilde{\eta}}$ (in which case we can conclude), or there exists $\tilde{t} > 0$ such that $h^1(\mu_t)(\tilde{K}^+) > 0$ over $[\tilde{t}, t_1]$, and more precisely, $h^1(\mu_t)(\tilde{K}^+)$ grows at most linearly.

By taking as test function $\min\{d(\Theta \backslash \mathring{\tilde{K}}, \cdot), 1\}$, we have that $\|h^1(\mu_t)\|_{BL} \geq \Delta \cdot h^1(\mu_t)$, which grows linearly. So once again, $t_1 < +\infty$. □

If we take the example of atomic measures for the Wasserstein gradient flow, the criteria corresponds roughly to: when the flow approaches a local minima, at least one particle needs to be in a 0-sublevel of the current potential $F'(\mu)$. This means that if we think of the negativity of the potential as a "measure" for non-optimality (recall the criterion for optimality), if the flow "detects" this defect by giving mass to it, we escape from the neighbourhood.

Lemma C.3 For all $C_0 > 0$, there exists $\alpha > 0$ such that for all $\mu, \nu \in \mathcal{M}_+(\Omega)$ that satisfy $\|h^1(\mu)\|_{BL}, \|h^1(\nu)\|_{BL} < C_0$, it holds

$$\|g_\mu - g_\nu\|_{C^1} \leq \alpha \|\phi\|_{C_1}^2 \cdot \|h^1(\mu) - h^1(\nu)\|_{BL}.$$

Proof. We take $\alpha > 0$ the Lipschitz constant of dR over $\{\int \Phi d\mu : \mu \in \mathcal{P}(\mathbb{R}^d), \|h^1(\mu)\|_{BL} \leq C_0\}$ which is bounded over $\mathcal{F}$.

We have:

$$\begin{aligned}
\|g_\mu - g_\nu\|_{C^1} &\leq \|\langle R'(\int \Phi d\mu) - R'(\int \Phi d\nu)\phi(\cdot)\rangle\|_{C^1} \\
&\leq \alpha \|\phi\|_{C^1} \| \int \Phi d\mu - \int \Phi d\nu \| \\
&\leq \alpha \|\phi\|_{C^1} \| \int \Phi dh^1(\mu) - \int \Phi dh^1(\nu) \| \\
&\leq \alpha \|\phi\|_{C^1} \cdot \sup_{f \in \mathcal{F}, \|f\| \leq 1} \int \langle f, \phi \rangle d(h^1(\mu) - h^1(\nu)) \\
&\leq \alpha \|\phi\|_{C^1}^2 \cdot \|h^1(\mu) - h^1(\nu)\|_{BL}.
\end{aligned}$$

Where we used the fact that $\langle f, \phi \rangle$ is $\|\phi\|_{C^1}$-Lipschitz and bounded by $\|\phi\|_{C^1}$ when $\|f\| \leq 1$.

□

To finish this section, we give a general property of stationary points.

Lemma C.4 Under Assumptions 2, let $(\mu_t)_t$ be a Wasserstein gradient flow of F. If $h^1(\mu_t)$ converges weakly to $\nu \in \mathcal{M}_+(\Omega)$, then $F'(\nu)$ vanishes ν-a.e.

Proof. At the point $(1, \theta) \in \Omega$, the velocity field $(v_t)_t$ associated to the gradient flow is given by applying the function $(\mathrm{id} - \mathrm{proj}_{\partial V(1,\theta)})$ (which is 2-Lipschitz) to the vector which has $g_{\mu_t}(\theta)$ as first components and $\nabla g_{\mu_t}(\theta)$ for the other ones.

Using Lemma C.3, we get that v_{μ_t} converges uniformly to v_ν over $\{1\} \times \Theta$.

Suppose now that there exists $\theta_0 \in \Theta$ such that $g_\nu(\theta_0) > 0$, then we can repeat the proof of the Proposition C.2 for a regular value $0 < \eta < g_\nu(\theta_0)$.

We write K the sub-level associated to $g_\nu(\theta_0)$ ($\theta_0 \in \mathring{K}$). The boundary is the same as in the proof of C.2, we then have the same conclusion.

As $h^1(\mu_t) \to \nu$, we find the same conclusion as in the C.2, but with g_{μ_t} bounded below

by $\frac{\eta}{2}$.

Repeating the arguments, we arrive to the inequality for all $t \geq t_0$ (where t_0 is the entry time in the ϵ-ball):

$$h^1(\mu_t)(K) \leq -t\frac{g_\nu(\theta_0)}{2}\mu_{t_0}(\mathbb{R}_+ \times K) + \min\{0, -t\frac{g_\nu(\theta_0)}{2} - r_0\}\mu_{t_0}(\mathbb{R}_- \times K) + h^1(\mu_{t_0})(K).$$

However here, as $h^1(\mu_t) \to \nu$, we need to have $\mu_{t_0}(\mathbb{R}_+ \times K) = 0$, and coming back to the above inequality, we have in particular that $\nu(K) = 0$. Hence, we deduce that $F'(\nu) \leq 0$ ν-a.e.

In addition, as the conclusion of Proposition C.2 is not achieved, we have that $F'(\nu) \geq 0$ ν-a.e. Combined with the other part, we get the result. □

C.3 Stability of separation properties

In this section, our goal will be to prove that a certain separation property of the support of the initialization measure is preserved by Wasserstein gradient flows. It will help us use the criteria to escape non-optimal stationary points.

The proof is based on topological degree theory, which allows us to cover the case of discontinuous velocity fields, which appear when V is non-differentiable. Indeed, if we recall the expression of the velocity field given in proposition B.1, we notice that when V is non-differentiable, $\partial V(u)$ can go from a singleton to a set 'discontinuously', which in turn means that the projection onto $\partial V(u)$ is discontinuous.

If you want to avoid this, and simply understand the idea, a more regular setting (V differentiable etc...) allows μ_t to be the pushforward measure of μ_0 by a homeomorphism. This would make the following results much easier to prove.

Definition (Topological degree) Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous map, $A \subset \mathbb{R}^d$ a bounded open set and $y \notin f(\partial A)$. The topological degree $deg(f, A, y)$ is a signed integer that satisfies:

1. If $deg(f, A, y) \neq 0$, then there exists $x \in A$ such that $f(x) = y$. If $y \in A$, then $deg(id, A, y) = 1$.
2. If A_1, A_2 are disjoint open subsets of A and $y \notin f(\overline{A}\backslash(A_1 \cup A_2))$ then $deg(f, A, y) = deg(f, A_1, y) + deg(f, A_2, y)$.
3. If $X : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ is continuous and $y : [0, 1] \to \mathbb{R}^d$ is a continuous curve such that $y(t) \notin X_t(\partial A)$ for all $t \in [0, 1]$, then $deg(X_t, A, y_t)$ is constant on $[0, 1]$.

These properties characterize a uniquely well-defined map deg from the set of triplets (f, A, y) as above to the set of signed integers (see [4]). Intuitively, it gives an algebraic count of the number of solutions to the equation $f(x) = y$. Algebraic meaning a solution x counts as $+1$ if f preserves orientation around x and -1 otherwise.

Let us first begin by a little lemma about the operations "taking the support of a measure"

and "taking the pushforward of a measure by a continuous map".

Lemma C.5 If $f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous map and $\mu \in \mathcal{M}_+(\mathbb{R}^d)$, then $spt(f_\# \mu) = \overline{f(spt\ \mu)}$.

Proof. Let $y \in f(spt\ \mu)$ and $\mathcal{V}$ a neighbourhood of y. By continuity, $f^{-1}(V)$ is the neigbourhood of a point in $spt(\mu)$, so $0 < \mu(f^{-1}(\mathcal{V})) = f_\#(\mathcal{V})$, meaning $y \in spt(f_\#(\mu))$ and $f(spt\ \mu) \subset spt(f_\#(\mu))$.

Conversely, let $y \in \overline{f(spt\ \mu)}^c$ and let $\mathcal{V}$ a neigbouroud of y that does not intersect $\overline{f(spt\ \mu)}$. This neighborhood satisfies $f^{-1}(\mathcal{V}) \subset (spt\ \mu)^c$, so it holds that $f_\#\mu(\mathcal{V}) = \mu(f^{-1}(\mathcal{V})) \leq \mu((spt\ \mu)^c) = 0$. Hence, $y \in (spt f_\#(\mu))^c$ so $\overline{f(spt\ \mu)}^c \subset (spt f_\# \mu)^c$ which implies $spt f_\# \mu \subset f\overline{spt\ \mu}$. □

Notice that in this lemma, if the map f is closed, then the two operation commute.
Now, let us introduce the property that interests us:

Property A (Separation) K a closed set contained in a box $Q_r \overset{\text{def}}{=} [-r, r] \times \Theta$ and separates $\{-r\} \times \Theta$ from $\{r\} \times \Theta$, for some $r > 0$, in the ambient space $\Omega = \mathbb{R} \times \Theta$. Here, separates means that any continuous path in the ambient space with endpoints in $\{-r\} \times \Theta$ and $\{r\} \times \Theta$ intersects K.

First, let us prove an abstract topological result:

Proposition C.6 (Set separation, boxes) Let $\Theta \subset \mathbb{R}^d$ be the closure of a bounded, connected, open set and, for some $T > 0$, let $X : [0, T] \times (\mathbb{R} \times \Theta) \to \mathbb{R} \times \Theta$ be a continuous map such that $X(0, \cdot) = id$, and $X_t(\mathbb{R} \times \Theta) \subset \mathbb{R} \times \partial\Theta$ for all $t \in [0, T]$. If K satisfies Property A, then $X_t(K)$ satifies it too for all $t \in [0, T]$.

Proof. Let $0 < \epsilon < \alpha < \beta$ be such that $X_t(K) \subset]-\alpha-\epsilon, \alpha+\epsilon[\times\Theta$, and $[-\alpha, \alpha] \times \Theta \subset X_t(]-\beta-\epsilon, \beta+\epsilon[\times\Theta)$ for all $t \in [0, T]$, and let A be the intersection of $]-\beta, \beta[$ with the unique connected component of $(\mathbb{R}\times\Theta)\backslash K$ that contains $\{\alpha\}\times\Theta$. The set A is thus bounded and open in $\mathbb{R}\times\mathbb{R}^{d-1}$. Consider the function $Y : (t, x) \mapsto (t, X_t(x))$ and the set $S = Y([0, T]\times\partial A)$ which a compact subset of $[0, T] \times (\mathbb{R} \times \Theta)$. Since connected component of S^c (the complement of S in $[0, T] \times (\mathbb{R} \times \Theta)$ are path connected, recalling the definition of the topological degree (point 3), it follows that

$$(t, (w, \theta)) \mapsto deg(X_t, A, (w, \theta))$$

is constant on each connected component of S^c.
Moreover, this degree is 1 on $[0, T] \times (\{\alpha \times \Theta)$ and is 0 on $[0, T] \times (\{-\alpha \times \Theta)$, using point 1 of the definition and $X(0, \cdot) = id$.

This means that for a fixed $t \in [0, T]$, any path joining $\{-\alpha\} \times \Theta$ to $\{\alpha\} \times \Theta$ must intersect $X_t(\partial A)$. It is in particular true for paths entirely contained in $[-\alpha, \alpha] \times \text{int}\, A$. It remains to notice that $\partial A \subset K \cup (\mathbb{R} \times \partial\Theta) \cup (\{\beta\} \times \Theta)$ and so, thanks to our assumption on X,

$$X_t(\partial A) \cap ([-\alpha, \alpha] \times \text{int}\, \Theta) \subset X_t(K).$$

This shows that $X_t(K)$ separates $\{-\alpha\} \times \Theta$ from $\{\alpha\} \times \Theta$ in the ambient space $\mathbb{R} \times \text{int}\, \Theta$.

Moreover, notice that K is compact, so $X_t(K)$ is closed, which means that the separation is still true in the ambient space $\mathbb{R} \times \Theta$. □

We can now prove the desired result:

Lemma C.7 (Stability of the separation property) Under Assumptions 2, let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow of F. If the support of μ_0 satisfies Property A, then so does the support of μ_t, for all $t > 0$.

Proof. Let X be the velocity field introduced in lemma B.7. It is continuous and satisfies $\mu_t = X_{t\#}\mu_0$. Moreover:

$$\begin{aligned}
|X_t(u)| &\geq |u| - |X_t(u) - u| \\
&\geq |u| - \int_0^t |\partial_t(X_s(u))ds \\
&\geq |u| - \int_0^t |v_s(X_s(u))|ds \\
&\geq |u| - t \cdot C
\end{aligned}$$

As v is bounded on finite time (which follows from the linear growth of $\|V(u)\|$, and the fact that in finite time, we stay on a Q_r).

So X_t is coercive, implying closed. By lemma C.5, we get $spt(\mu_t) = X_t(spt(\mu_0))$.

To conclude, we just need to verify the assumptions of Proposition C.6.

When Θ is bounded, Assumptions 2 (iii)-(a) imply that for all $(t, u) \in \mathbb{R} \times \partial\Theta$, $\partial_t X_t(u) = 0$, so $X_t = id$, and $X_t(\mathbb{R} \times \partial\theta) \subset \mathbb{R} \times \partial\Theta$.
Using the abstract result from above, the proof is finished.

In the case where $\Theta = \mathbb{R}^{d-1}$, let us consider $\Psi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \times \text{int}\,(B(0,1))$, $(\omega, \theta) \mapsto (\omega, \frac{\theta}{|\theta|} \cdot \tanh(|\theta|)$ if $\theta \neq 0$ if not $(w, 0)$. Notice that Ψ is a diffeomorphism.

Let $Y_t = \Psi \circ X_t \circ \Psi^{-1}$. By the chain rule, we have that $\partial_t Y_t(y) = d\Psi_{v_t \circ \Psi^{-1}(y)}(v_t \circ \Psi^{-1}(y)) = \tilde{v}_t(y)$.

So Y_t is the flow for the associated velocity field $\tilde{v}_t$, defined a priori on $\mathbb{R} \times \text{int}\,(B(0,1))$.
We can extend it by continuity on $\mathbb{R} \times \mathbb{S}^{d-2}$ by $(g_\infty(\theta) \cdot sign(\omega), 0)$ where g_∞ is the limit in Assumptions 2 (iii)-(a), with $sign(0) = 0$.
This means that for (ω, θ) stays the same throughout the dynamic, as there is a null velocity on this compact. This means that $Y_t(\mathbb{R} \times \mathbb{S}^{(d-2)}) \subset \mathbb{R} \times \mathbb{S}^{(d-2)}$.

So by C.6, $Y_t(spt(\mu_0))$ satisfies Property A.

If we notice that $spt\mu_0$ satisfies Property A implies $spt(\Psi_{\#}\mu_0)$ satisfies Property A for $\Theta = B(0,1)$, then we get that:

$$Y_t(spt(\Psi_{\#}\mu_0)) = \Psi \circ X_t(spt\mu_0) = \Psi(spt\mu_t)$$

satisfies Property A. Ψ being a homeomorphism, we get that $spt(\mu_t)$ satisfies Property A, proving the result.

□

C.4 Main theorem

First let us prove a lemma about the convergence of the Wasserstein gradient flows, that will be useful later.

Lemma C.8 Under Assumtions 1, let (μ_t) be a Wasserstein gradient flow whose initialization is concentrated on a set Q_{r_0} and such that $F(\mu_T) \to F^*$. If $(\mu_{0,m})_m$ is a sequence of measures concentrated on a set Q_{r_0} that converges to μ_0 in W_2, then

$$F^* = \lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = \lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t}).$$

Proof. As the atomic measures are dense in $\mathcal{P}_2(\Omega)$ for W_2, the results of theorem B.6 can be extended for a general initialization sequence $(\mu_{m,0})$.

Using continuity of F for W_2, we get that $\lim_{t\to\infty} F(\mu_{m,t}) = F(\mu_t)$ giving us $\lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = F^*$.

Moreover, let $\epsilon > 0$, as $t \mapsto F(\mu_t)$ is decreasing, take $t_0 > 0$: $F(\mu_{t_0}) < F^* + \frac{\epsilon}{2}$.
Now, let $n_0 \in \mathbb{N}$: for all $n \geq n_0$, $F(\mu_{m,t_0}) < F(\mu_{t_0}) + \frac{\epsilon}{2}$.

We get that $F(\mu_{m,t_0}) \leq F^* + \epsilon$. Which implies, $t \mapsto F(\mu_{m,t})$ is decreasing, $\lim_{t\to\infty} F(\mu_{m,t}) \leq F^* + \epsilon$, finally implying, $\lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t}) \leq F^* + \epsilon$.

As ϵ is arbitrary, this gives us the desired result. □

We are now ready to prove the main theorem:

Theorem C.9 Under Assumptions 2, let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow of F such that for some $r_0 > 0$, the support of μ_0 is contained in $[-r_0, r_0] \times \Theta$ and separates $\{-r_0\} \times \Theta$ from $\{r_0\} \times \Theta$. If $h^1(\mu_t)$ converges weakly, then its limit is a global minimizer of F over $\mathcal{M}_+(\Omega)$ and $\lim_{t\to\infty} F(\mu_t) = F^*$.

Proof. Let ν be a weak limit of $h^1(\mu_t)$, we see it as a measure on $\{1\} \times \Theta$. By lemma C.4, $F'(\nu)$ vanish ν-a.e.

By contradiction, let us assume that ν is not a minimizer of F over $\mathcal{M}_+(\Omega)$, meaning $F'(\nu)$ is not nonnegative. Let $A \subset \Omega$ and ϵ be the radius of the $\|\cdot\|_{BL}$-ball provided by proposition C.1.
As $h^1(\mu_t) \to \nu$, let $t_0 > 0$ such that $t > t_0$, $\|h^1(\mu_t) - \nu\|_{BL} < \epsilon$.

Taking notations from C.2, let us first consider the case where Θ is bounded. Let $\theta_0 \in K^+$ be a local minimum of g_ν in the interior of K^+ relatively to Θ (the case where K^+ is empty andK^- is not is similar).

The Neumann boundary conditions guarantee that $\nabla g_\nu(\theta_0) = 0$ even if $\theta_0 \in \partial\Theta$.
By lemma C.7, the line $\mathbb{R} \times \{\theta_0\}$ intersect the support of μ_{t_0}.
If the intersection happens in $\mathbb{R} \times K^+$, we can immediately conclude by C.2. Otherwise, let $M > 0$ be such that μ_{t_0} is concentrated on $[-M, M] \times \Theta$.

Let $r_0 > 0$ be such that $B(\theta_0, r_0) \cap \Theta \subset K^+$. By lemma C.10 (see below), there exists $t_1 > t$ such that if $spt(\mu_{t_1}) \cap [-M, 0] \times \{\theta_0\} \neq \emptyset$, then the support intersect $\mathbb{R} \times K^+$ at a later time, and we conclude by C.4.

So it remains to check that $spt(\mu_{t_1}) \cap [-M, 0] \times \{\theta_0\} \neq \emptyset$. By C.7, $spt(\mu_{t_1})$ intersects $\mathbb{R}_- \times \{\theta_0\}$ at the point (ω_0, θ_0). If we recall the properties of K^+ developped in C.2, we have that the ω component of the velocity field is lower bounded by $\frac{\eta}{2}$ for $t > t_0$. This means that $\omega_0^2 > -M$, which allow us to conclude.

For the case $\Theta = \mathbb{R}^{d-1}$, we use the same technique as in C.7, where we map the flow to the unit ball, and use the previous reasoning. □

We state the technical result used in the proof:

Lemma C.10 Consider, for a measure $\nu \in \mathcal{M}(\Theta)$, a point $\theta_0 \in \Theta$ such that $|\nabla g_\nu(\theta)| = 0$ and $g_\nu(\theta) \leq -\eta$ for some $\eta > 0$. For any $M > 0$ and $r_0 > 0$, there exists $T, \epsilon > 0$ such that if $(\mu_t)_t$ is a Wasserstein gradient flow of F that satisfies for all $t \in [0, T]$, $\|g_{\mu_t} - g_\nu\|_{C^1} \leq \epsilon$ and denoting $(w(t), \theta(t))$ the solution of the flow introduced in section B starting from (w_0, θ_0) with $w_0 \in [-M, 0]$, it holds $w(T) = 0$ and $|\theta(t) - \theta_0| < r_0$.

Proof. The Lipschitz regularity of g_ν and its derivative implies that there exists $L > 0$ such that $\max\{|g_\nu(\theta) - g_\nu(\theta_0)|, |\nabla g_\nu(\theta) - \nabla g_\nu(\theta_0)|\} \leq L|\theta - \theta_0|$ for all $\theta \in \Theta$. Without loss of generality (take L big enough), we assume that $r_0 < \eta/(4L)$. Consider $\epsilon \in]0, \eta/4[$ and assume that there exists $\overline{T} > 0$ such that $\|g_{\mu_t} - g_\nu\|_{C^1} \leq \epsilon$ for $t \in [0, \overline{T}]$. Writing $q(t) = |\theta(t) - \theta_0|$, it holds for $t \in [0, \overline{T}]$,

$$\begin{cases} \frac{dq}{dt} \leq -w(\epsilon + Lq) \\ \frac{dw}{dt} \geq \eta - \epsilon - Lq \end{cases}.$$

In particular, if we can make sure that $|q(t)| < r_0$ for $t \in [0, \overline{T}]$ and if $\overline{T} > 2/\eta$ then, as $(dw/dt) \geq \eta/2$ on this interval, there exists $T < 2/\eta$ such that $w(T) = 0$.

It remains to make sure that $|q(t)| < r_0$ for $t \in [0, \overline{T}]$, by adjusting if necessary the value of ϵ. Parametrizing in w instead of t (it is an admissible reparametrization thanks to the positive lower bound on its derivative), we get

$$(dq/dw) = (dq/dt) \cdot (dt/dw) \leq -w(\epsilon + Lq) \cdot 2/\eta.$$

We can then apply Grönwall's lemma to $\tilde{q}(w) \overset{\text{def}}{=} \epsilon + Lq(w)$ which satisfies $(d/dw)\tilde{q}(w) \leq (-2L/\eta) \cdot w \cdot \tilde{q}$ and obtain

$$\tilde{q}(w) \leq \tilde{q}(w_0) \exp(-(2L/\eta) \int_{w_0}^{0} s ds) = \epsilon \exp(L w_0^2/\eta).$$

Thus, choosing $\epsilon < Lr_0/(\exp(Lw_0^2/\eta) - 1)$, it is guaranteed that $q(t) < r_0$ for $t \in [0, \overline{T}]$.

□

In theorem C.9, the convergence of the Wasserstein gradient flow comes as an assumption. In order to prove convergence of gradient flows, we generally need: (i) compacteness of the trajectories and (ii) a *Łojasiewiscz inequality* which roughly controls how much a function flattens around its critical points. As compactness in W_2 is a strong requirement, we relaxed the topology where convergence is required for more reasonable assumptions. However, even when a gradient flow is in a compact set, there are some cases where it does not converges.

3 Consequences for sparse spike deconvolution and neural networks with a single hidden layer with a sigmoid activation

After the quite theoretical part 2, it is time to see how the theory that we developed can be used to state results in more concrete cases, namely sparse spike deconvolution (for setting, see the introduction) and neural networks with a single hidden layer with a sigmoid activation.

3.1 Loss functions

In this section, we will give sufficient conditions to satisfy the assumptions on the loss R, when the Hilbert space is $\mathcal{F} = L^2(\rho)$ for a probability measure ρ on a space $\mathcal{X}$, which is either a domain of $\mathbb{R}^d$ or the torus. In this setting, typical losses are of the form $R(f) = \int r(x, f(x))d\rho(x)$ for a function $r : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$.

Lemma (Properties of the loss) If r is convex in the second variable, then R is convex. If r is differentiable in the second variable with $\partial_2 r$ Lipschitz, uniformly in the first variable, then R is differentiable with differential dR Lipschitz. If moreover $|\partial_2 r|^2 \leq C_1 r + C_2$ for some constants $C_1, C_2 > 0$, then dR is bounded on sublevel sets.

Proof. The convexity property follows immediately from the linearity and monotonicity of the integral.
If $\partial_2 r$ is Lipschitz, uniformly in the first variable, then denoting $dR_f : h \mapsto \int r'(x, f(x))h(x)d\rho(x)$, $\exists L > 0$ such that for all $f, h \in \mathcal{F}$,

$$|R(f + h) - R(f) - dR_f(h)| \leq \frac{L}{2} \int |h(x)|^2 d\rho(x) = \frac{L}{2}||h||^2 = o(||h||),$$

which means that dR_f is the differential of R at the point f.
Using that, it follows that dR is Lipschitz in the operator norm.

Finally, if $|\partial_2 r|^2 \leq C_1 r + C_2$, then

$$||dR_f||^2 = \int |\partial_2 r(x, f(x))|^2 d\rho(x) \leq C_1 R(f) + C_2,$$

so dR is bounded on sublevel sets. □

3.2 Sparse deconvolution

For sparse deconvolution, it is typical to consider a signal $y \in \mathcal{F} \overset{\text{def}}{=} L^2(\Theta)$ on the d-torus $\Theta = \mathbb{R}^d/\mathbb{Z}^d$. The loss function is $R(f) = (1/2\lambda) = \|y - f\|^2_{L^2}$, for some $\lambda > 0$, a parameter that increases with the noise level and the regularization is $V(w, \theta) = |w|$. Consider now a filter impulse response $\psi : \Theta \to \mathbb{R}$ and let $\Phi(w, \theta) : x \to w \cdot \psi(x - \theta)$. The object sought after is a signed measure on Θ, which is obtained from a probability measure on $\mathbb{R} \times \Theta$ by applying the operator h^1 defined previously. This corresponds to going back from the lifting introduced before.

Proposition (Spare deconvolution) Assume that the filter impulse response ψ is $\min\{d, 2\}$ times continuously differentiable, and that the support of μ_0 contains $\{0\} \times \Theta$. If the projection $h^1(\mu_t))_t$ of the Wasserstein gradient flow of F weakly converges to $\nu \in \mathcal{M}(\Theta)$, then ν is a global minimizer of

$$\min_{\mu \in \mathcal{M}(\Theta)} \frac{1}{2\lambda} \|y - \int \psi d\mu\|^2_{L^2} + |\mu|(\Theta).$$

Proof. Let us show that Assumptions 2 holds in this case.

The choice of $\Theta = \mathbb{R}^d/\mathbb{Z}^d$ means we are in the bounded case without the difficulties related to the boundary. On the separable Hilbert space $\mathcal{F} = L^2(\Theta)$ for the normalized Lebesgue measure on the d-torus, the loss R verifies the assumptions for the lemma of part 3.1, with $r(x, f) = (f(y) - y(x))^2$, with the regularisation term $\tilde{V} = 1$.
The norm of the function $\phi(\theta) : x \mapsto \psi(x - \theta)$ does not depend on θ, so it is bounded. The differentiability of ψ means that ϕ is continuously differentiable with $d\phi_\theta(\bar{\theta}) : x \mapsto \nabla\psi(x - \theta) \cdot \bar{\theta}$ which is bounded (again the norm does not depend on θ), and we can show it is Lipschitz the same way as the proof of the Lemma from 3.1.

It remains to check the Morse-type regularity assumption i.e., to check that for all $f \in \mathcal{F}$, the function $\theta \mapsto \langle f, \phi(\theta)\rangle = \int f(x)\psi(x - \theta)dx$ has a set of regular values which is dense in its range. If this function is constantly 0 then this is trivially true, otherwise, its range is an interval of $\mathbb{R}$. By Morse-Sard's lemma (see [1]), if this function is $d - 1$-times continuously differentiable, then the set of critical values has zero Lebesgue measure and our assumption holds. By differentiating under the integral sign, this assumption is thus satisfied if ϕ is $d - 1$-times continuously differentiable.

As everything checks out, we can use theorem C.9 to conclude. □

3.3 Neural network with a single hidden layer: sigmoid activation

Let's first build the mathematical framework. Consider a joint distribution of features and labels, $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$, and $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$ as the marginal distribution of features. We define the loss as the expected risk, i.e., $R(f) = \int l(f(x), y)d\rho(x, y)$. In this context, we take $\mathcal{F}$ to be $L^2(\rho_x)$ and $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}+$ to be either the squared loss or the logistic loss. Finally, we define the function $\Phi(\omega, \theta) : x \mapsto \omega\sigma\left(\sum i = 1^{d-2}\theta_i x_i + \theta_{d-1}\right)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a sigmoid function, such as $\sigma(s) = (1 + e^{-s})^{-1}$. The domain Θ is chosen to be $\mathbb{R}^{d-1}$. Additionally, note that the natural option (though not obligatory) for the regularization term is $V(\omega, \theta) = |\omega|$,

penalizing the l^1 norm of the weights.

Before utilizing theorem C.9, we first state a key result that needs to be proven. This result will guarantee that ϕ satisfies Assumptions 2.

Lemma If ρ_x has finite moments up to order 4, then the function $\phi : \mathbb{R}^{d-1} \to \mathcal{F}$ defined as $\phi(x) = \sigma(z \cdot \theta)$ is differentiable, with a Lipschitz and bounded differential $d\phi_\theta(h)(x) = (h \cdot z)\sigma'(z \cdot \theta)$, where $z = (x, 1)$.

Proof. Before proceeding, it is worth mentioning that in this proof and throughout the rest of this report, we will denote by ρ_z the distribution of z when x is distributed according to ρ_x. Let's now check that the function $d\phi$ defined above is indeed the differential of ϕ. For $\theta, h \in \mathbb{R}^{d-1}$, we have

$$\Delta(h)^2 \stackrel{\text{def}}{=} \|\phi(\theta + h) - \phi(\theta) - d\phi_\theta(h)\|^2$$

$$= \int_{\mathcal{X}} |\sigma(\theta \cdot z + h \cdot z) - \sigma(\theta \cdot z) - (h \cdot z)\sigma'(z \cdot \theta)|^2 d\rho_z(z) \leq \frac{L^2}{4} \int_{\mathcal{X}} |h \cdot z|^4 d\rho_z(z)$$

where L denotes the Lipschitz constant of σ'. So if ρ_z has finite 4-th order moment $M_4(\rho_z)$ then $\delta(h) \leq \frac{L\sqrt{M_4(\rho_z}}{2}|h|^2$ and $d\phi$ is indeed the differential of ϕ. This differential is bounded and Lipschitz since $\|d\phi_\theta\| \leq \|\sigma'\|_\infty \sqrt{M_2(\rho_z)}$ and $\|d\phi_\theta - d\phi_{\tilde{\theta}}\| \leq L\sqrt{M_4(\rho_z)}|\theta - \tilde{\theta}|$ for all $\theta, \tilde{\theta} \in \mathbb{R}^{d-1}$. Finally, it is clear that if ρ_x has finite 4-th moment then so does ρ_z. □

Let's demonstrate that this result applies to the sigmoid activation framework.

Propositon (Sigmoid activation) Assume that ρ_x has finite moment up to order $\max\{4, 2d-2\}$, that the support of μ_0 is $\{0\} \times \Theta$ and that boundary conditions Assumption 2 (iii)-(a) holds. If the Wasserstein gradient flow of F converges in W_2 to μ_∞, then μ_∞ is a global minimizer of F.

Proof. We need to verify if Assumptions 2 are satisfied in order to apply theorem C.9. Since the boundary condition is assumed, we only need to check the following:

(i) Both ϕ and $\tilde{V}$ are bounded and differentiable with Lipschitz differentials.
(ii) Smooth convex loss.
(iii) Sard-type regularity.

Proof of (i): It is trivial to see that the regularization term $\tilde{V} = 1$ satisfies the required assumptions. Moreover, by the previous lemma, we know that ϕ also satisfies these assumptions.

Proof of (ii): We write the disintegration of ρ with respect to the variable x as $\rho(dx \bigotimes dy) = \rho(dy|x) \bigotimes \rho_x(dx)$ where ρ_x is the marginal law of ρ on $\mathcal{X}$ and $(\rho(\cdot|x))_{x \in \mathcal{X}}$ a family of conditional probabilities on $\mathbb{R}$ (see [2]). On the separable Hilbert space $L^2(\rho_x)$, the loss R is as in Lemma $D.1$ with $r(x, p) = \int_{\mathbb{R}} l(p, y)p(dy|x)$. Thus, it satisfies the necessary assumptions.

Proof of (iii): We need to check that for all $f \in \mathcal{F}, \theta \mapsto \langle f, \phi(\theta)\rangle) \int_{\mathcal{X}} f(x)((x, 1) \cdot \theta)d\rho_x(x)$ has a set of regular values which is dense in it's range. If the function in constantly 1 then this is trivially true, otherwise, its range is an interval of $\mathbb{R}$. If ρ_x has finite moments up to order $2d - 2$ then the function above is $d - 1$ continuously differentiable and the conclusion follows from Morse-Sard's lemma.

We can conclude by applying theorem C.9, as the framework satisfies Assumptions 2, as demonstrated above. □

It might seem unusual that we assumed the boundary condition. However, the reality is that verifying it can be challenging. For example, consider the setting where $R(f) = \frac{1}{2}\|f - f^*\|^2_{\mathcal{F}}$ where f^* is the optimal Bayes regressor that we may assume smooth. As required in the boundary assumptions, consider a function $f \in \mathcal{F}$ of the form $f = R'(\int \Phi d\mu) = \int \Phi d\mu - f^*$ for some μ in the domain of the functional F. In the limit $r \to \infty$, the function $g_f(r\theta) \overset{\text{def}}{=} \langle f, \phi(r\theta)\rangle = \int f(x)(r\theta \cdot (x,1))d\rho_x(x)$ converges to the function $\overline{g}_f(\theta) = \int_{\theta\cdot(x,1)\geq 0} f(x)d\rho_x(x)$. This function is continuously differentiable on the sphere if the density of ρ_x is in $C_0(\mathbb{R}^{d-2})$ and f is bounded and continuous (this is the case here) and the convergence of $g_f(r\cdot) \to \overline{g}_f$ is indeed in C^1. However, we cannot guarantee a very high regularity for f in general: differentiable under the integral sign $d-1$ times requires to have moments of order $(d-1)$ bounded for μ, which cannot be assumed a priori (μ is just know to be in the domain of F). This prevents us from applying Morse-Sard's lemma.

4 Numerical illustration

In this section, we will illustrate the previous results with concrete examples. More precisely, we will look at sparse deconvolution, in the setting of section 3.2.
Thanks to the Proposition from section 3.2, we know that with big enough amount of particle, by running a gradient descent, we can find the global minimum.

In our experiment, we consider the signal $y = \sum_{i=1}^{3} w_i \cdot \phi(\theta_i)$, for $w = (3.5, -4, 4)$ and $\theta = (0.1, 0.3, 0.8)$, where the filter impulse response is a Dirichlet kernel of order 7 and $\lambda = 0.3$.
The gradient flow is integrated using the forward-backward algorithm described in [8], and the particles are initialized on $\{0\} \times \Theta$ uniformly. For implementation details, see the code appendix below.
In red, the trajectories of the particles, in black the final position, the x-axis is Θ and the y-axis is the weights.

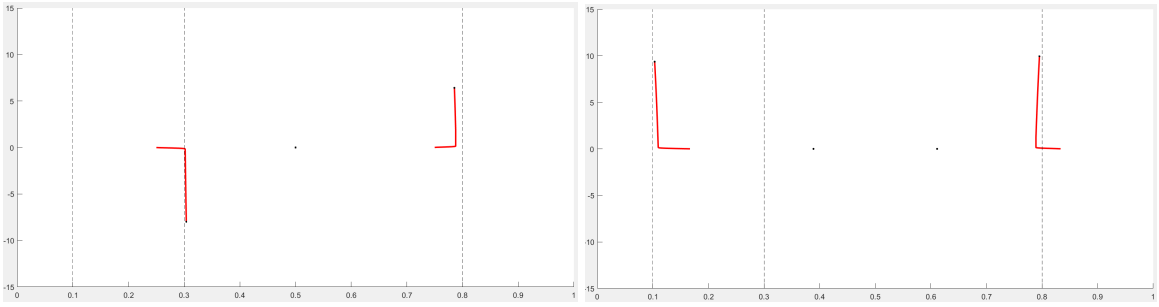


Figure 5: Failure to reach minimum with 3 and 4 particles.

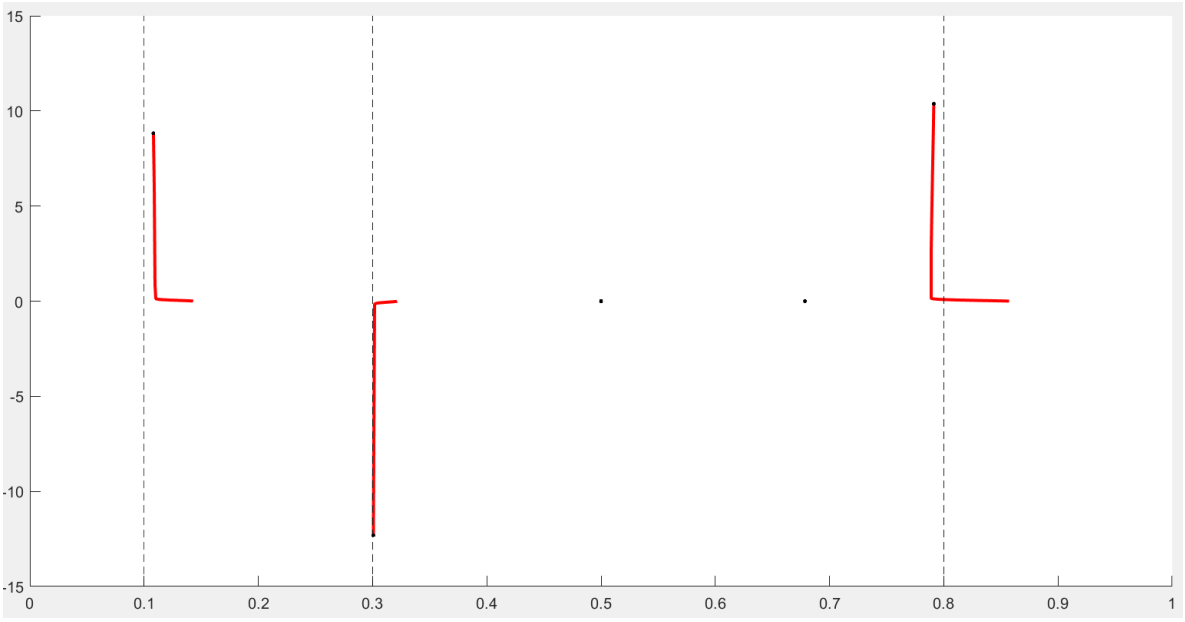


Figure 6: Success to reach minimum with 5 particles.

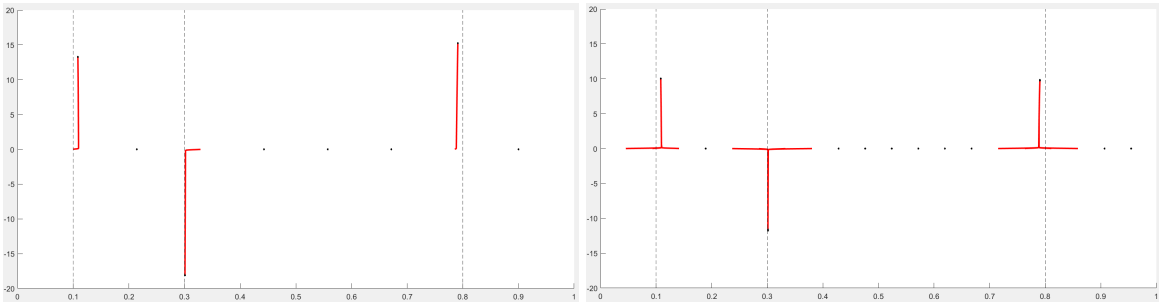


Figure 7: Success with 8 and 20 particles.

Another strategy that is much easier to implement both in theory and practice is to discretize the parameter space Θ by picking a number of fixed positions, and then optimize on the weights. The resulting functional is convex, meaning gradient descent-based algorithm work, and for a large enough number of positions, we can get approximate minimizers. However, our approach has some advantages over this method:

- Whilst the original idea behind searching the minimizer of the original functional was to find a sparse description of the signal, implementing the 'convex' approach in code for the same signal as before gives us the following results:

  We can see that the distribution of weights does not gives us much sparsness, but are spread around the optimal positions. On the other hand, the previous strategy, even when there was not enough particle to find a minimizer, gave us very sparse solutions (always less or equal number of spikes than the ground truth).
  This means this approach fails at giving us a parsimonious description despite the regularizer.

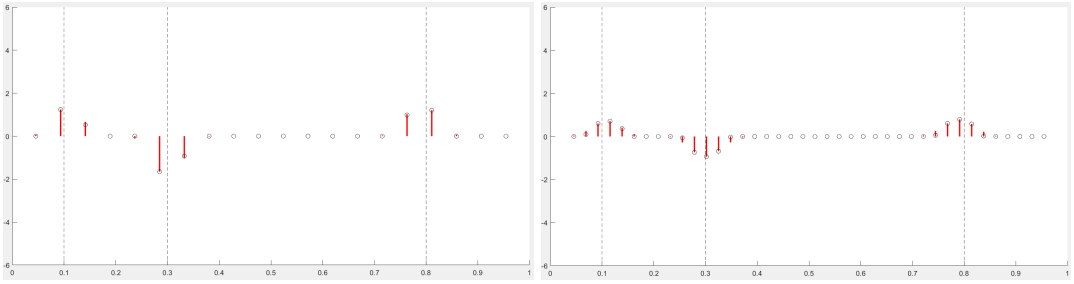


Figure 8: Gradient descent on the weights with fixed positions for 20 and 40 uniformly spaced particles

- Another point is the number of particles that we need to use in order to reach the minimum. While our approach only needs slight over-parametrization in some cases, the 'convex' approach behaves worse with the same amount of particles, and it needs much more to reach the same performance.
- Under more restrictive assumptions (mainly regularity-type assumptions), and with a modified gradient descent, the complexity scales as $\log(1/\epsilon)$ in the desired accuracy ϵ instead of ϵ^{-d} for the convex method (d being the dimension of the parameter space Θ). This means that for big dimensions, our approach is considerably better. For more details, see [5], which is a follow-up to the article studied here and offers quantitative results.

5 Conclusion

We have established asymptotic global optimality properties for a family of non-convex gradient flows. These results were enabled by the study of a Wasserstein gradient flow: this object simplifies the handling of many-particle regimes, analogously to a mean-field limit. The particle-complexity to reach global optimality turns out very favorable on synthetic numerical problems. This confirms the relevance of our qualitative results and calls for quantitative ones that would further exploit the properties of such particle gradient flows (see [5] as explained before). Multiple layer neural networks are also an interesting avenue for future research.

References

[1] Ralph Abraham and Joel Robbin. *Transversal mappings and flows*. WA Benjamin New York, 1967.

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

[3] 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.

[4] Felix E. Browder. Fixed point theory and nonlinear problems. *Proc. Sym. Pure. Math*, 39:49–88, 1983.

[5] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent, 2020.

[6] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport, 2018.

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

[8] Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

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

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

[11] D. Evanko. Primer: fluorescence imaging under the diffraction limit. *Nature Methods*, 6:19–20, 2009.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[13] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems 30*, 2017.

[14] Benjamin D. Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017.

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

[16] Haitham A. Hindi. A tutorial on optimization methods for cancer radiation treatment planning. *2013 American Control Conference*, pages 6804–6816, 2013.

[17] M. Bates M. Rust and X. Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature Methods*, 3:793–796, 2006.

[18] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. A dual certificates analysis of compressive off-the-grid recovery. *arXiv preprint arXiv:1802.08464*, 2018.

[19] K. G. Puschmann and F. Kneer. On super-resolution in astronomical imaging. *Astronomy and Astrophysics*, 436:373–378, 2005.

[20] Ralph T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[21] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

Appendix Code

In this appendix, we give the matlab code that we used to produce the results presented in section 4.

First, the function which computes the gradient descent:

```
function [ws, thetas, loss] = GD_1(w_init, theta_init, w_obs,
    theta_obs, lambda, alpha, niter)

    m = length(w_init);
    ws = zeros(m,niter);
    thetas = zeros(m,niter);
    gradw = zeros(1, m);
    gradtheta = zeros(1, m);
    loss = zeros(1, niter);

    %observed signal
    f_obs = @(x) sum(w_obs.*phi(x-theta_obs));

    function s = A(w, theta)
            f = @(x) sum(w.*phi(x - theta))./m;
            f1 = @(x) (f(x)-f_obs(x)).^2;
            s = integral(f1, 0, 1);
    end

    %proximity operator for our non-differentiable part
    function p = prox(x, P)
            p=zeros(m, 1);
            for l=1:m
                if x(l, 1)>P/m
                    p(l, 1) = x(l, 1)-P/m;
                elseif x(l, 1)<-P/m
                    p(l, 1) = x(l, 1)+P/m;
                else
                    p(l, 1) = 0;
                end
            end
    end

    ws(:, 1) = w_init;
    thetas(:, 1) = theta_init;
    loss(1,1) = 1./(2.*lambda).*A(w_init, theta_init) + sum(
       abs(w_init))./m;

    for iter=2:niter

        w= ws(:, iter-1);
```

```
        theta= thetas(:, iter-1);

        %compute the error
        loss(1, iter) = 1./(2.*lambda).*A(w, theta) + sum(abs
           (w))./m;

        for i = 1:m %compute the gradient

            g = @(x) (f_obs(x) - sum(w.*phi(x-theta))./m ).*
               phi(x-theta(i,1));
            h = @(x) (f_obs(x) - sum(w.*phi(x-theta))./m).*w(
               i,1).*phi_der(x-theta(i,1));

            gradw(1, i) = -integral(g, 0, 1)./(lambda.*m);
            gradtheta(1, i) = integral(h, 0, 1)./(lambda.*m);
        end

        %update of the step
        beta=alpha./(1+max(abs(w)./25));

        %update of the weights and positions with forward-
           backward method
        ws(:, iter) = w + 0.5.*(prox(w -beta.*gradw', beta) -
            w);
        thetas(:, iter) = theta + 0.5.*(- beta.*gradtheta');

        %theta must stay in the interval
        for k = 1:m
            if (thetas(k, iter)< 0)
                thetas(k, iter) = 0;
                ws(k, iter) = 0;
            end
            if (thetas(k, iter)>1)
                thetas(k, iter)=1;
                ws(k, iter) = 0;
            end

        end

    end

end
```

Then, the function that computes the gradient descent where we fix a priori the positions and optimize the weights:

```
function [ws, loss] = CO(w_obs, theta_obs, w_init, theta_init
   , lambda, alpha, niter )
```

```
2
3       m = length(w_init);
4       ws = zeros(m,niter);
5       gradw = zeros(1, m);
6       loss = zeros(1, niter);
7
8       %observed signal
9       f_obs = @(x) sum(w_obs.*phi(x-theta_obs));
10
11      function int = A(w)
12              f = @(x) sum(w.*phi(x - theta_init))./m;
13              f1 = @(x) (f(x)-f_obs(x)).^2;
14              int = integral(f1, 0, 1);
15      end
16
17      %proximity operator
18      function p = prox(x, P)
19           p=zeros(m, 1);
20           for l=1:m
21               if x(l, 1)>P/m
22                       p(l, 1) = x(l, 1)-P/m;
23               elseif x(l, 1)<-P/m
24                       p(l, 1) = x(l, 1)+P/m;
25               else
26                       p(l, 1) = 0;
27               end
28           end
29       end
30
31      ws(:, 1) = w_init;
32      loss(1,1) = 1./(2.*lambda).*A(w_init) + sum(abs(w_init))
           ./m;
33
34      for iter=2:niter
35
36          w= ws(:, iter-1);
37
38          %compute the error
39          loss(1, iter) = 1./(2.*lambda).*A(w) + sum(abs(w))./m
               ;
40
41          for i = 1:m %compute the gradient
42
43              g = @(x) (f_obs(x) - sum(w.*phi(x-theta_init))./m
                   ).*phi(x-theta_init(i,1));
44
45              gradw(1, i) = -integral(g, 0, 1)./(lambda.*m);
```

```
        end
        %update the weights
        ws(:, iter) = prox(w -alpha.*gradw', alpha);

    end

end
```

Finally, some auxiliary functions:

```
function t = phi_der(x)
    nf = 7;
    p= zeros(size(x));
    for i = -nf:nf
        p =  p + 2.*pi.*1i.*i.*exp(2.*pi.*1i.*i.*x);
    end
    t = real(p)./nf;
end
```

```
function phi = phi(x)
    nf = 7;
    phi = diric(2.*pi.*x, nf);
end
```