École polytechnique fédérale de Lausanne

Spring 2025

Department of Mathematics

Gradient Flows in Wasserstein Spaces

Variational Inference and Sampling

Author:
Luca Raffo

Supervisor:
Dr. Leonardo V. Santoro

Contents

Introduction

In many areas of science, there is a fundamental need to model randomness and uncertainty. To formalize these concepts, mathematicians introduced the notion of **probability measures**, which provide a rigorous and systematic framework to describe the likelihood of different outcomes. Over time, these objects have become ubiquitous in probability theory and statistics.

The question that arises suddenly is how to compare different probability measures. A natural way is to ask: *given two probability measures, how much effort is required to morph one into the other?* This question is at the heart of optimal transport theory, a field that studies the most efficient way to transfer mass between two distributions while minimizing a given transportation cost. Surprisingly, the **optimal map**, i.e. the map that morphs the first probability measure into the second one while minimizing the cost[1], induces a notion of distance between the probability measures that retain geometric properties of the underlying space (i.e. the one upon which the measures are defined).

In the past thirty years, it has been discovered that spaces of probability measures, when endowed with this metric, exhibit rich geometric properties [1]. This perspective has led to significant developments in partial differential equations [3], and differential geometry [10], [4].

Recently, many researchers in mathematical statistics and optimization have sought to leverage these geometric insights to develop new approaches to regression, density fitting, computer vision [11]; and to variational inference and sampling [5].

In this project, we will first develop the main intuitions behind the geometric properties of measure spaces, highlighting key results. We will then explore their relevance to mathematical and computational statistics through concrete derivations and implementations. We will assume that the reader has been already exposed to the classical results in optimal transport (the first two chapters of [6] are sufficient), which within themselves require the reader to be comfortable in measure theory and to know the main techniques of functional analysis. More advanced topics can be found in [2]. It is also necessary to have an intuitive understanding of basic concepts in Riemannian geometry.

To unify the notation and recall basic concepts upon which the work is based, we include a preliminary section on metric geometry, optimal transport and RKHS.

[1] Note that we are being sloppy here for the sake of exposition: the distance actually comes in general from the optimal coupling, where couplings between two probability measures are relaxation of transport maps.

1 Preliminaries

The objectives of this section are to recall some fundamental ideas and to fix the notation for the subsequent sections.

1.1 Metric geometry

Metric geometry is the field of mathematics that studies abstraction of key ideas from differential geometry, relying only on the intrinsic notion of distance. This approach enables us to define geodesics and other concepts in purely metric terms, offering analogies with classical differential structures while operating in a more general setting. All the proofs in this section can be verified in [5].

This approach will be of fundamental importance as in the following sections we will deal with the pseudo-Riemannian structure that arises from spaces of probability measures when endowed with the Wasserstein distance. As a toy example, at the end of this subsection, we will see that we can endow a very expressive space of probability measures with a different metric that allows some differential constructions as well, but is not of much interest due to its limited ability to retain properties of the space upon which the measures are defined.

Let us fix a set $\mathcal{S}$, together with a function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$.

Definition 1.1. We say that $(\mathcal{S}, d)$ is a *metric space* if it satisfies the following conditions:

1. *positive definiteness.* We have that for any $x_0, x_1 \in \mathcal{S}$, the metric satisfies $d(x_0, x_1) \geq 0$; and moreover $d(x_0, x_1) = 0 \iff x_0 = x_1$.
2. *symmetry.* We have that for any $x_0, x_1 \in \mathcal{S}$, the metric satisfies $d(x_0, x_1) = d(x_1, x_0)$.
3. *triangle inequality.* Usually the hardest to check, requires that for any $x_0, x_1, x_2 \in \mathcal{S}$, the metric satisfies $d(x_0, x_2) \leq d(x_0, x_1) + d(x_1, x_2)$.

For the rest of the section the couple $(\mathcal{S}, d)$ will represent a metric space. With only this notion, we can already define *paths* and *lengths*.

Definition 1.2. A *path* in $\mathcal{S}$ is a continuous map $\omega : I \to \mathcal{S}$, where $I \subseteq \mathbb{R}$ is an interval.

Definition 1.3. The *length* $L(\omega) \in \mathbb{R} \cup \{+\infty\}$ of a path $\omega : I \to \mathbb{R}$ is defined as

$$L(\omega) := \sup \sum_{i=1}^{n-1} d(\omega(t_i), \omega(t_{i+1})),$$

where the supremum is taken over all $n \geq 1$ and all $n-$tuples $t_1 < ... < t_n$. Once we have notions of smoothness and derivatives it can be shown that under regularity conditions, if d is induced by a norm $\|\cdot\|_d$, this cumbersome definition is equivalent to $\int_I \|\frac{d\omega}{dt}\|_d \, dt$, which has to be interpreted as the limit $n \to +\infty$.

A path is called *rectifiable* whenever it has finite length. We have the following proposition:

Proposition 1.4. For any rectifiable path $\omega : I \to \mathbb{R}$, the function $t \mapsto l(t) := L(\omega_{(-\infty,t]})$ is continuous on I.

As in the familiar Euclidean settings, two paths $\omega_1 : I_1 \to \mathcal{S}$ and $\omega_2 : I_2 \to \mathcal{S}$ are *equivalent* if there exists a continuous, non-decreasing and surjective function $\varphi : I_1 \to I_2$ such that $\omega_1 = \omega_2 \circ \varphi$. In this case, ω_2 is called a *reparametrization* of ω_1 (and vice-versa by symmetry), and it is trivial to check that $L(\omega_1) = L(\omega_2)$.

We say that a path $\omega : [a, b] \to \mathcal{S}$ has *constant speed* if for all $a \le s \le t \le b$,

$$L(\omega_{[s,t]}) = \frac{t - s}{b - a} L(\omega), \tag{1}$$

and we have the following classical result:

Proposition 1.5. Any rectifiable path $\omega : [a, b] \to \mathcal{S}$ has a constant-speed reparametrization $\bar{\omega} : [0, 1] \to \mathcal{S}$.

Let us now move towards the focus of this subsection, which mainly regards our geometric intuition. In Euclidean spaces $(\mathbb{R}^d, \|\cdot\|_2)$, for fixed $x_0, x_1 \in \mathbb{R}^d$, we think of the *midpoint* between them as the linear interpolation, i.e. the point in the middle of the segment joining them, $x_{\frac{1}{2}} := \frac{x_0 + x_1}{2}$.

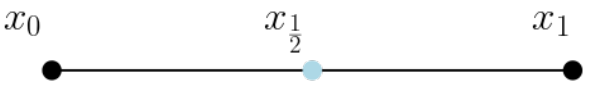

Figure 1: Midpoint in $\mathbb{R}^2$.

In a Euclidean embedded submanifold, such as $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$, we also have an intuition about midpoints, but being the space not linear we cannot trivially interpolate between two points and hope that the result will lie on the submanifold. In order to lift the idea of *midpoint* we need to lift the concept of segment into submanifolds, and this translates into *geodesics*. In general, for fixed x_0 and x_1, a geodesic is a path between them which minimizes the length. Furthermore, we can look at the constant speed geodesic $\omega : [0, 1] \to \mathbb{S}^{d-1}$, with $\omega(0) = x_0, \omega(1) = x_1$, and define the midpoint as $x_{\frac{1}{2}} := \omega\left(\frac{1}{2}\right)$.

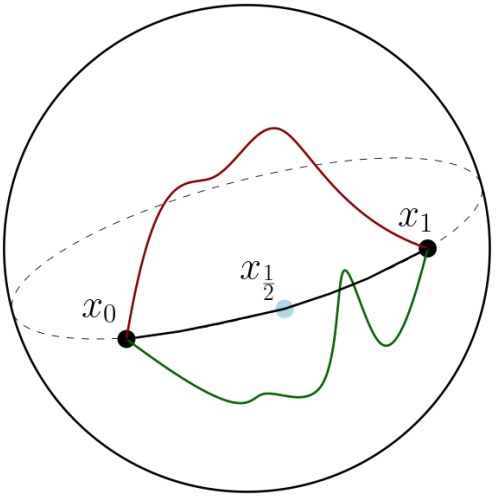

Figure 2: Midpoint in $\mathbb{S}^2$. The red and green paths are not geodesics.

This is the intuition we are going to abstract in the metric geometry framework. Given $x_0, x_1 \in \mathcal{S}$, a path $\omega : [a, b] \to \mathcal{S}$ is said to *connect* x_0 to x_1 if $\omega(a) = x_0$ and $\omega(b) = x_1$. By construction of the length function, $d(x_0, x_1) \le L(\omega)$ for any path ω connecting x_0 to x_1.

Definition 1.6. The space $\mathcal{S}$ is called a *length space* if for all $x, y \in \mathcal{S}$, we have

$$d(x_0, x_1) = \inf_{\omega} L(\omega), \tag{2}$$

where the infimum is taken over all paths ω connecting x_0 to x_1.

And moreover we have the following.

Definition 1.7. A length space is said to be a *geodesic space* if for all $x, y \in \mathcal{S}$, the infimum on the right hand side of (2) is attained.

Accordingly, we can formally define the notion of geodesic.

Definition 1.8. A *geodesic* between x_0 and x_1 is any path $\omega : [0, 1] \to \mathcal{S}$ attaining the infimum in (2).

It follows from this definition that if $\omega : [0, 1] \to \mathcal{S}$ is a geodesic, $d(\omega(s), \omega(t)) = L(\omega_{[s,t]})$, for all $0 \leq s \leq t \leq 1$. To see this, just notice that

$$\begin{aligned} d(\omega(0), \omega(1)) = L(\omega_{[0,1]}) &\geq L(\omega_{[0,s]}) + L(\omega_{[s,t]}) + L(\omega_{[t,1]}) \\ &\geq d(\omega(0), \omega(s)) + d(\omega(s), \omega(t)) + d(\omega(t), \omega(1)), \end{aligned}$$

and that by the triangle inequality

$$d(\omega(0), \omega(1)) \leq d(\omega(0), \omega(s)) + d(\omega(s), \omega(t)) + d(\omega(t), \omega(1)),$$

and so each inequality is an equality. Together with (1), this yields the following useful characterization of *constant speed geodesics*.

Proposition 1.9. A path $\omega : [0, 1] \to \mathcal{S}$ is a *constant speed geodesic* if and only if

$$d(\omega(s), \omega(t)) = (t - s)d(\omega(0), \omega(1)),$$

for all $0 \leq s \leq t \leq 1$.

Proof. Suppose ω is a constant speed geodesic. By definition, its length satisfies

$$L(\omega_{[s,t]}) = (t - s)L(\omega),$$

where $L(\omega) = d(\omega(0), \omega(1))$. Since ω is a geodesic, by our previous discussion we have

$$d(\omega(s), \omega(t)) = L(\omega_{[s,t]}),$$

which gives the desired identity.

Conversely, if

$$d(\omega(s), \omega(t)) = (t - s)d(\omega(0), \omega(1))$$

for all $0 \leq s \leq t \leq 1$, then ω moves at constant speed along a shortest path, implying that it is a geodesic. □

So, accordingly to our previous discussion, for any two points $x_0, x_1 \in \mathcal{S}$ we define the *midpoint* of (x_0, x_1) as any $x_{\frac{1}{2}} \in \mathcal{S}$ that satisfies

$$d(x_0, x_{\frac{1}{2}}) = d(x_{\frac{1}{2}}, x_1) = \frac{1}{2} d(x_0, x_1).$$

As anticipated, we end this chapter with an example.

Example 1.10. Let us take $\mathcal{P}_2^{ac}(\lambda)$ to be the family of probability measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$ and whose second moment is finite[2]. Given $\mu_0, \mu_1 \in \mathcal{P}_2^{ac}(\lambda)$ we can write $\mu_0 = f_0\, d\lambda$ and $\mu_1 = f_1\, \lambda$, and we can leverage on the fact that $\|\cdot\|_{L_2(\lambda)}$ is a metric in functions space to extend it to this measures space in the following way:

$$\|\mu_0 - \mu_1\|_{L_2(\lambda)} := \int_{\mathbb{R}^d} |f_0(x) - f_1(x)|^2\, d\lambda(x) = \|f_0 - f_1\|_{L_2(\lambda)}.$$

Formally, we have the following result,

Proposition 1.11. The space $(\mathcal{P}_2^{ac}(\lambda), \|\cdot\|_{L_2(\lambda)})$ is a metric space.

Let us now show that this is a geodesic space by exhibiting a constant speed geodesic between two generic μ_0, μ_1, accordingly with Definition 1.7.

Proposition 1.12. The constant speed geodesic between μ_0 and μ_1 is the *linear interpolation measure*

$$\mu_t := h(t)\, d\lambda = [(1-t)f_0 + tf_1]\, d\lambda.$$

Proof. Obviously, we have $h(0) = f_0$ and $h(1) = f_1$. Furthermore, $\frac{dh}{dt} = f_1 - f_0$, which is constant with respect to t. Finally,

$$L(h) = \int_0^1 \left\| \frac{dh}{dt} \right\|_{L_2(\lambda)} dt = \int_0^1 \|f_1 - f_0\|_{L_2(\lambda)}\, dt = \|f_1 - f_0\|_{L_2(\lambda)},$$

which concludes our proof. □

We can thus define our midpoint between μ_0 and μ_1 as $\mu_{\frac{1}{2}} := h\left(\frac{1}{2}\right) d\lambda$. Unfortunately this construction does not retain at all the geometry of the underlying space, as the midpoint is a *vertical interpolation* between the probability measures, rather than a *horizontal interpolation*[3], as the following visualization shows.

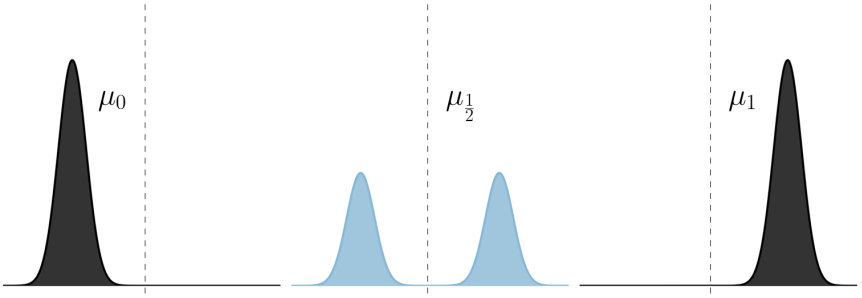

Figure 3: μ_0 and μ_1 are set to be $N(-5, 1)$ and $N(5, 1)$.

1.2 Monge and Kantorovich problems

The heart of this manuscript relies in the ability of the Wasserstein distance on measures spaces to retain properties of the underlying space, i.e. the one upon which the measures

[2]This is just a technical condition to ensure finiteness.

[3]That will be achieved with the Wasserstein distance instead.

are defined. The right way to introduce this notion of distance is by addressing the optimal transport problem, as the Wasserstein metric arises as a byproduct of this prolific theory.

Although we assume that the reader has already been exposed to all of these results, we still list them to unify notation and for the sake of completeness.

We are going to denote with $\mathcal{P}$ the set of probability measures on $\mathbb{R}^d$ and with $\mathcal{P}_2$ the subset of probability measures with finite second moment.

Definition 1.13. Given $\mu_0 \in \mathcal{P}$ and $T : \mathbb{R}^d \to \mathbb{R}$, we define the *push forward measure* as

$$T\#\mu_0 := \mu_0(T^{-1}(A)), \text{ for any } A \subseteq \mathbb{R}^d.$$

Remarkably, $\mu_1 = T\#\mu_0$ if and only if $\int_{\mathbb{R}^d} \varphi \, d\mu_1 = \int_{\mathbb{R}^d} \varphi \circ T \, d\mu_0$ for any $\varphi : \mathbb{R}^d \to \mathbb{R}$ measurable and bounded, as shown in [6].

We have the following theorem which is useful in computations.

Theorem 1.14. Let $\mu_0, \mu_1 \in \mathcal{P}$, with $\mu_0, \mu_1 \ll \lambda$, where λ is the Lebesgue measure on $\mathbb{R}^d$. If $f = \frac{d\mu_0}{d\lambda}$ and $g = \frac{d\mu_1}{d\lambda}$ are the Radon-Nykodim derivatives, and $T : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism such that $\mu_1 = T\#\mu_0$, we have that

$$f(x) = |\det J_{T(x)}| g(T(x)).$$

Proof. It follows from the standard change of variables formula. Indeed, for any $\varphi : \mathbb{R}^d \to \mathbb{R}$ measurable and bounded,

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi \circ T(x) f(x) \, dx &= \int_{\mathbb{R}^d} \varphi(y) g(y) \, dy \\ &= \int_{\mathbb{R}^d} \varphi \circ T(x) g(T(x)) |\det J_{T(x)}| \, dx, \end{aligned}$$

which concludes because φ was arbitrary and T was bijective. □

Definition 1.15. We will call *canonical projections* the functions

$$\begin{aligned} \pi_X : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad \pi_X[(x, y)] = x \\ \pi_Y : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad \pi_Y[(x, y)] = y \end{aligned}$$

Sometimes we will abuse of this notation without specifying the projections' domain and codomain, but it will always be clear from the context.

Definition 1.16. Given $\mu_0 \in \mathcal{P}$ and $T : \mathbb{R}^d \to \mathbb{R}$, we denote the set of *couplings* as

$$\Gamma(\mu_0, \mu_1) := \{\gamma \in \mathcal{P} \times \mathcal{P} : \ \pi_X\#\gamma = \mu_0, \ \pi_Y\#\gamma = \mu_1\}.$$

We have the machinery to define the fundamental optimization problems.

Definition 1.17. Given $\mu_0, \mu_1 \in \mathcal{P}$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the *Monge's problem* as

$$(\text{MP}) = \inf_{T:\mathbb{R}^d \to \mathbb{R}} \left\{ \int_{\mathbb{R}^d} c(T(x), x) \, d\mu_0(x) : T\#\mu_0 = \mu_1 \right\}$$

This formally translates that we aim to move a probability measure to another one by minimizing a given cost function. We can relax the need of a transport map by only requiring a restriction in terms of couplings.

Definition 1.18. Given $\mu_0, \mu_1 \in \mathcal{P}$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the *Kantorovich's problem* as

$$(\mathrm{KP}) = \inf_{\gamma \in \mathcal{P} \times \mathcal{P}} \left\{ \int_{\mathbb{R}^d} c(x,y)\, d\gamma(x,y) : \gamma \in \Gamma(\mu_0, \mu_1) \right\}.$$

Remarkably the set of couplings is never empty: it is enough to consider $\mu_0 \otimes \mu_1 \in \mathcal{P} \times \mathcal{P}$. Moreover, due to Prokhorov and Banach-Alaoglu's theorems, the infimum is actually reached, besides very pathological situations [6].

Furthermore any transport map T such that $T\#\mu_0 = \mu_1$ induces a coupling $\gamma_T := (id, T)\#\mu_0 \in \Gamma(\mu_0, \mu_1)$, as can be checked into [6].

In the following we are going to work only with $c(x,y) = \|x - y\|^2$, as this choice is the most natural and the most studied, and furthermore we can anticipate that this choice will enable us to define the tangent space (in a fixed point of our Wasserstein space) as a Hilbert space. Accordingly to this choice, to avoid issues with $+\infty$, we will restrict ourselves to $\mathcal{P}_2$.

We have the important theorem, due to Brenier.

Theorem 1.19. Given $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$ and $\mu_1 \in \mathcal{P}_2$, then there exists a unique optimizer $\bar{\gamma}$ in (KP). In addition, there exists $T = \nabla\varphi$, with $\varphi : \mathbb{R}^d \to \mathbb{R}$ convex, such that $\bar{\gamma} = (id, T)\#\mu_0$ and T is the unique optimizer in (MP).

1.3 Reproducing kernel Hilbert spaces

In this subsection we review the main definitions of RKHS theory and state the important *representer theorem*. This partially interrupts the flow, and may consequently be skipped for the time being, as its relevance is only devoted to the fourth section.

Definition 1.20. A Hilbert space $\mathcal{H}$ is a complete, possibly infinite-dimensional linear space endowed with a inner product.

In the following we will only work with functions defined on $\mathbb{R}^d$. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces.

A norm in $\mathcal{H}$ can be naturally defined from the given inner product, as $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$. Our norm will always be assumed to be the one arising from the latter. Furthermore, we always assume that $\mathcal{H}$ is separable (contains a countable dense subset) so that $\mathcal{H}$ has a countable orthonormal basis[4].

Example 1.21. Let us consider the class of square integrable functions on the interval $[a, b]$, denoted by $L^2[a, b]$. We define the inner product as

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx$$

[4]This is just a technical condition used in the proof of the *representer theorem*.

This produces the correct norm:

$$\|f\| = \left(\int_a^b f^2(x)\, dx \right)^{1/2}$$

It can be checked that this space is complete, so it is a Hilbert space. However, there is one problem with the functions in this space. Consider trying to evaluate the function $f(x)$ at the point $x = k$. There exists a function g in the space defined as follows:

$$g(x) = \begin{cases} c & \text{if } x = k \\ f(x) & \text{otherwise} \end{cases}$$

Because it differs from f only at one point, g is clearly still square-integrable, and moreover, $\|f - g\| = 0$. However, we can set the constant c (or, more generally, the value of $g(x)$ at any finite number of points) to an arbitrary real value. What this means is that a condition on the integrability of the function is not strong enough to guarantee that we can use it predictively, since prediction requires evaluating the function at a particular data value. This characteristic is what will differentiate reproducing kernel Hilbert spaces from ordinary Hilbert spaces, as we discuss in the following.

Definition 1.22. An *evaluation functional* over a Hilbert space of functions $\mathcal{H}$ is a linear functional

$$\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$$

that evaluates each function in the space at the point t, or

$$\mathcal{F}_t[f] = f(t) \quad \text{for all } f \in \mathcal{H}.$$

Definition 1.23. A Hilbert space $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) if the evaluation functionals are bounded, i.e., if for all t there exists some $M > 0$ such that

$$|\mathcal{F}_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H},$$

which means that the norm topology is stronger than the topology induced by pointwise convergence.

While this condition might seem obscure or specific, it is actually quite general and is the weakest possible condition that ensures us both the existence of an inner product and the ability to evaluate each function in the space at every point in the domain.

In practice, it is difficult to work with this definition directly. We would like to establish an equivalent notion that is more useful in practice. To do this, we will need the *reproducing kernel* from which the reproducing kernel Hilbert space takes its name.

Firstly, from the definition of the reproducing kernel Hilbert space, we get that evaluations are linear and bounded functionals, hence we can apply the Riesz representation theorem.

Proposition 1.24. If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $K_t \in \mathcal{H}$ (called the *representer of* t) with the *representing property*

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t) \quad \text{for all } f \in \mathcal{H}.$$

This allows us to represent our linear evaluation functional by taking the inner product with an element of $\mathcal{H}$. Since K_t is a function in $\mathcal{H}$, by the representing property, for each $x \in X$ we can write

$$K_t(x) = \langle K_t, K_x \rangle_{\mathcal{H}}.$$

We take this to be the definition of reproducing kernel in $\mathcal{H}$.

Definition 1.25. The *reproducing kernel of* $\mathcal{H}$ is a function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined by

$$K(t, x) := K_t(x).$$

Definition 1.26. A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *reproducing kernel* if it is symmetric, i.e. $K(x, y) = K(y, x)$, and positive definite:

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, \ldots, t_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Having this general notion of a reproducing kernel is important because it allows us to define an RKHS in terms of its reproducing kernel, rather than attempting to derive the kernel from the definition of the function space directly. The following theorem formally establishes the relationship between the RKHS and a reproducing kernel.

Proposition 1.27. A RKHS defines a corresponding reproducing kernel. Conversely, a reproducing kernel defines a unique RKHS.

Let us move to the main result for this subsection. Let us suppose having N data points $(x_i, y_i)_{i=1,\ldots,N}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We want to find the function in $\mathcal{H}$ that best interpoles the data while not being too complex, which formally is translated into:

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|^2_{\mathcal{H}}$$

for a loss function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a constant $\lambda \in \mathbb{R}$. Optimizing through a vector space with an infinite number of dimension is a priori undoable on a computer, but we have the following theorem, known as *representer theorem* which allows us to move directly to a finite dimension optimization problem.

Theorem 1.28. The minimizer over the RKHS $\mathcal{H}$, of the regularized empirical loss functional

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|^2_{\mathcal{H}}$$

can be represented by the expression

$$f^*(x) = \sum_{i=1}^{n} c_i K(x_i, x),$$

for some n-tuple $(c_1, \ldots, c_n) \in \mathbb{R}^n$. Hence, minimizing over the (possibly infinite-dimensional) Hilbert space boils down to minimizing over $\mathbb{R}^d$.

For the proof we suggest reading [7].

In the vector-valued setting, where we consider functions $\varphi : \mathbb{R}^d \to \mathbb{R}^d$, the RKHS accordingly becomes $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$ (with d copies); and representer theorem extends componentwise.

That is, each component $\varphi^{(j)}(x)$ of the minimizer $\varphi^* \in \mathcal{H}^d$ can be written as:

$$\varphi^{(j)}(x) = \sum_{i=1}^{N} c_i^{(j)} K(x_i, x), \quad \text{for } j = 1, \ldots, d.$$

Thus, the full vector field $\varphi^*(x) \in \mathbb{R}^d$ has the form:

$$\varphi^*(x) = \sum_{i=1}^{N} K(x_i, x)\, c_i, \quad \text{with } c_i \in \mathbb{R}^d.$$

This implies that optimization over the infinite-dimensional space $\mathcal{H}^d$ reduces to optimization over N vectors in $\mathbb{R}^d$.

2 Wasserstein spaces

In this section we are going to use the metric induced by (KP) on the space of measures and study its properties. We will show that the geometry induced by this metric lifts the geometry of the underlying space, as we wanted.

It turns out that this space also has a natural differential structure that leads not only to defining geodesics but also to notions of tangent spaces and gradients. These objects will lead to defining Wasserstein gradient flows, which are tools of fundamental importance that will play a key role in the next sections.

In this section we will also have a digression on functional analysis, which is required to formally address evolution of measures subject to gradient fields induced by Wasserstein gradients.

2.1 Pseudo-Riemannian geometry

As hinted in the previous section, the inf in (KP) is really a min, besides pathological situations. Recalling that we restricted to the quadratic cost, we can thus define the following.

Definition 2.1. Given two probability measures $\mu_0, \mu_1 \in \mathcal{P}_2$, we define the *Wasserstein distance* between them as

$$\mathcal{W}_2(\mu_0, \mu_1) := \min_{\gamma} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, d\gamma(x, y) \,|\, \gamma \in \Gamma(\mu_0, \mu_1) \right)^{\frac{1}{2}}.$$

This is indeed a metric, as the next result shows.

Theorem 2.2. $\mathcal{W}_2 : \mathcal{P}_2 \times \mathcal{P}_2 \to \mathbb{R}$ is a metric on $\mathcal{P}_2$.

Proof. We need to check *positive definiteness*, *symmetry* and *triangle inequality*. Let us fix $\mu_0, \mu_1, \mu_2 \in \mathcal{P}_2$.

1. *positive definiteness.*
   Trivially $\mathcal{W}_2(\mu_0, \mu_1) \geq 0$. Moreover, $\mathcal{W}_2(\mu_0, \mu_0) = 0$ since $\bar{\gamma} = (id, id)\#\mu_0$ is optimal. For the last direction, notice that if $\mathcal{W}_2(\mu_0, \mu_1) = 0$, there exists $\bar{\gamma} \in \Gamma(\mu_0, \mu_1)$ optimal and such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\bar{\gamma}(x, y) = 0$, which implies that $x = y$, $\bar{\gamma}$−a.e.; and if we take any $\varphi : \mathbb{R}^d \to \mathbb{R}$ continuous,

   $$\int_{\mathbb{R}^d} \varphi \, d\mu_0 = \int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(x) \, d\bar{\gamma}(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(y) \, d\bar{\gamma}(x, y) = \int_{\mathbb{R}^d} \varphi(y) \, d\mu_1(y).$$

   and thus $\mu_0 = \mu_1$ because φ is arbitrary.

2. *symmetry.*
   Let $S(x, y) := (y, x)$. Then if $\gamma \in \Gamma(\mu_0, \mu_1)$, we have $S\#\gamma \in \Gamma(\mu_1, \mu_0)$, and S does not change the integral with respect to γ because the Euclidean distance is symmetric.

3. *triangle inequality*.
   Let $\bar{\gamma}_{01} \in \Gamma(\mu_0, \mu_1)$ and $\bar{\gamma}_{12} \in \Gamma(\mu_1, \mu_2)$ be optimal. By Dudley's lemma (in [11]),

there exists a composition of couplings Λ such that $\pi_{(X,Y)}\#\Lambda = \gamma_{01}$, $\pi_{(Y,Z)}\#\Lambda = \gamma_{12}$ and $\pi_{(X,Z)}\#\Lambda \in \Gamma(\mu_0, \mu_2)$. Then we can compute:

$$\begin{aligned}\mathcal{W}_2(\mu_0, \mu_2) &\leq \|x_0 - x_2\|_{L_2(\pi_{X,Z}\#\Lambda)} = \|x_0 - x_2\|_{L_2(\Lambda)} \\ &\leq \|x_0 - x_1\|_{L_2(\Lambda)} + \|x_1 - x_2\|_{L_2(\Lambda)} \\ &= \|x_0 - x_1\|_{L_2(\bar{\gamma}_{01})} + \|x_1 - x_2\|_{L_2(\bar{\gamma}_{12})} \\ &= \mathcal{W}_2(\mu_0, \mu_1) + \mathcal{W}_2(\mu_1, \mu_2),\end{aligned}$$

as we wanted to show.

□

Remarkably, when the optimal coupling is induced (uniquely) by an optimal map, the distance simplifies to $\mathcal{W}(\mu_0, \mu_1) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|T_{\mu_0 \to \mu_1}(x) - x\|_2^2 \, d\mu_0(x)\right)^{\frac{1}{2}}$. From now on, for fixed $\mu_0, \mu_1 \in \mathcal{P}_2$, we are going to assume the existence of a unique optimal transport plan[5], as this will not let us lose much and simplifies the exposition.

Recalling our definitions from the first section, we can show that $(\mathcal{P}_2, \mathcal{W}_2)$ is a geodesic space via the characterization of Proposition 1.9 by exhibiting geodesics between any two fixed $\mu_0, \mu_1 \in \mathcal{P}_2$.

Proposition 2.3. Given any $\mu_0, \mu_1 \in \mathcal{P}_2$, the constant speed geodesic with respect to the Wasserstein distance is $\mu_t := T_t\#\mu_0$, where $T_t(x) := (1-t)x + tT_{\mu_0 \to \mu_1}(x)$.

Proof. Let $\bar{\gamma} \in \Gamma(\mu_0, \mu_1)$ be an optimal coupling for W_p. Set $\pi_t(x, y) := (1-t)x + ty$, so that

$$\begin{cases}(\pi_0)\#\bar{\gamma} = \mu_0 \\ (\pi_1)\#\bar{\gamma} = \mu_1\end{cases}$$

Define $\mu_t := (\pi_t)\#\bar{\gamma}$ and let $\gamma_{s,t} := (\pi_s, \pi_t)\#\bar{\gamma} \in \Gamma(\mu_s, \mu_t)$. Then for any $0 \leq s \leq t \leq 1$

$$\begin{aligned}W_p(\mu_s, \mu_t) &\leq \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |z - z'|^p \, d\gamma_{s,t}(z, z')\right)^{1/p} = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |\pi_s(x, y) - \pi_t(x, y)|^p \, d\bar{\gamma}(x, y)\right)^{1/p} \\ &= (t-s)\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \, d\bar{\gamma}(x, y)\right)^{1/p} = (t-s)W_p(\mu_0, \mu_1).\end{aligned}$$

Applying this bound on the intervals $[0, s]$, $[s, t]$, and $[t, 1]$, we get

$$W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1) \leq [s + (t-s) + 1 - t]W_p(\mu_0, \mu_1) = W_p(\mu_0, \mu_1).$$

Note that the converse inequality always holds, by the triangle inequality. Hence, all inequalities are equalities and we deduce that

$$W_p(\mu_s, \mu_t) = (t-s)W_p(\mu_0, \mu_1) \qquad \forall\, 0 \leq s \leq t \leq 1,$$

as wanted to show. □

Again, we can reasonably define the midpoint as $\mu_{\frac{1}{2}} := T_{\frac{1}{2}}\#\mu_0$ and this geometry encodes the horizontal interpolation we were looking for, as the following illustration shows.

[5] This occurs, for instance, when $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$, as shown in Brenier's theorem.

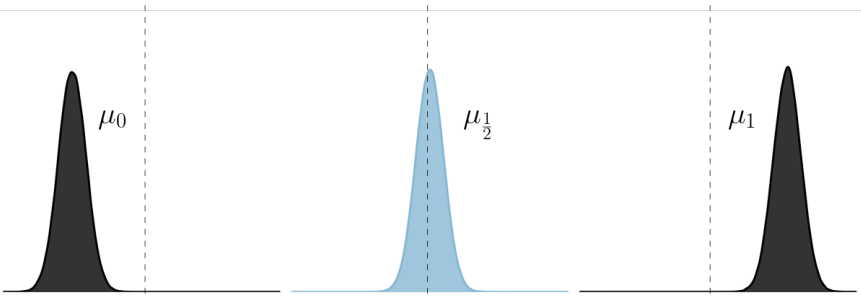

Figure 4: μ_0 and μ_1 are set to be $N(-5, 1)$ and $N(5, 1)$.

Moreover, we can see that the underlying space's geometry is preserved by looking at the isometry $(x, \|\cdot\|) \mapsto (\delta_x, \mathcal{W}_2)$:

Proposition 2.4. Given any $x_0, x_1 \in \mathbb{R}^d$, we have that

$$\|x_1 - x_0\|_2 = \mathcal{W}_2(\delta_{x_0}, \delta_{x_1}).$$

Proof. Obviously we notice that $\Gamma(\delta_{x_0}, \delta_{x_1}) = \delta_{(x_0,x_1)}$, and the cost induced by the coupling (which is actually induced by the transport map that sends $x \mapsto x_1, \ \forall x \in \mathbb{R}^d$), is

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\delta_{(x_0,x_1)} = \|x_0 - x_1\|^2,$$

which concludes our argument. □

Moreover, the transport structure of the Wasserstein space induces not only a metric space, but a manifold-like geometry, referred to in the literature as *Otto calculus*. In order to fix ideas, let us take as a toy example the usual Riemannian geometry framework on $\mathbb{S}^2$.

Example 2.5. Let us consider a generic $x_0 \in \mathbb{S}^2$. Now, for any $x_1 \in \mathbb{S}^2$ (not equal and not antipodal to x_0), there exists a unique constant speed geodesic ω, i.e. the smaller section of the maximal circumference through x_0 and x_1. By smoothness, it is well defined and unique the vector $\omega'(0)$, which tells a moving particle starting in x_0 where to go if it wants to follow ω.

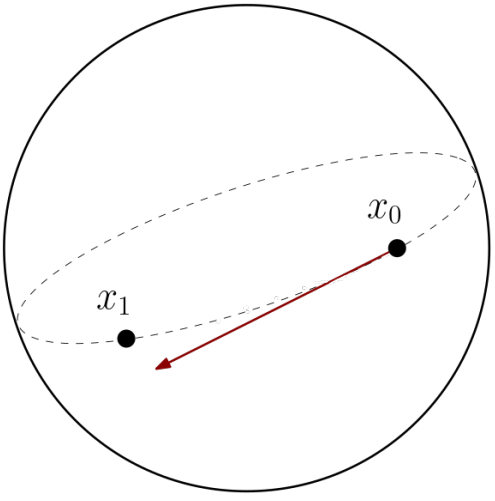

Figure 5: The vector $\omega'(0)$ is in red.

Now, for fixed x_0, if we take the linear combination of the vectors $\omega'(0)$ as x_1 varies in $\mathbb{S}^2$, we obtain a plane. We denote such plane as the tangent space $\mathcal{T}_{x_0}\mathbb{S}^2$. We are going to use the same exact idea to lift the concept of tangent space to our measures space.

Under our assumptions, for fixed $\mu_0 \in \mathcal{P}_2$, we can identify uniquely any $\mu_1 \in \mathcal{P}_2$ with $T_{\mu_0\to\mu_1}$. Furthermore, since we showed that $(\mathcal{P}_2, \mathcal{W}_2)$ is a geodesic space, we are motivated to define a *tangent space* like structure in the following way[6] accordingly to our example:

Definition 2.6. Given $\mu_0 \in \mathcal{P}_2$, we define the *tangent space* at μ_0 as

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 := \overline{\left\{\zeta(T_{\mu_0\to\mu_1} - id) : \mu_1 \in \mathcal{P}_2;\ \zeta > 0\right\}}^{L^2(\mu_0)}.$$

The geometric intuition is straightforward: in this metric space, the role of the tangent vector is played by $\frac{dT_t}{dt} = \frac{d}{dt}\left[(1-t)id + tT_{\mu_0\to\mu_1}\right] = T_{\mu_0\to\mu_1} - id$. Then, by our assumption on finiteness of second moments, every such map satisfies $T_{\mu_0\to\mu_1} - id \in L_2(\mu_0)$. We hence take the closure with respect to $\|\cdot\|_{L_2(\mu_0)}$.

An equivalent, useful definition given in [5], is the following:

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 = \overline{\{\nabla\varphi |\, \varphi : \mathbb{R}^d \to \mathbb{R} \text{ compactly supported and smooth}\}}^{L_2(\mu_0)} \tag{3}$$

As a subset of $L_2(\mu_0)$, the tangent space inherits the inner product:

$$\langle f, g\rangle_{\mu_0} = \int_{\mathbb{R}^d} f(x)g(x)\, d\mu_0(x),\ \ f, g \in L_2(\mu_0),$$

and this makes it clear why we chose to work with $\mathcal{W}_2$ instead of any other $\mathcal{W}_p$, $p \in [1, +\infty]$. Though not obvious from the definition, $\mathcal{T}_{\mu_0}\mathcal{P}_2$ is a linear space, as shown in [1]. Motivated by this differential structure, we can define the following tools:

Definition 2.7. Given $\mu_0 \in \mathcal{P}_2$ we define a formal *exponential map* as

$$exp_{\mu_0} : \mathcal{T}_{\mu_0}\mathcal{P}_2 \to \mathcal{P}_2, \qquad exp_{\mu_0}(T) = (T + id)\#\mu_0.$$

Moreover, we define a *logarithm map* as its left inverse, projecting onto $\mathcal{T}_{\mu_0}\mathcal{P}_2$

In particular if $\mu_0 \ll \lambda$, Brenier's theorem yields that

$$log_{\mu_0}(\mu_1) = T_{\mu_0\to\mu_1} - id.$$

In general, for fixed $\mu_0 \in \mathcal{P}_2$, the exponential map takes a generic transformation as input and interprets it as a *tangent vector* to μ_0, and by our identification $\mu_1 \longleftrightarrow T_{\mu_0\to\mu_1}$ it outputs the unique measure that induces the constant speed geodesic that produces that tangent vector. The logarithm map does the inverse.

Finally, we can also define a notion of *curvature*: given $\mu_0, \mu_1, \mu_2 \in \mathcal{P}_2$, we can write:

$$W_2^2(\mu_0, \mu_1) \le \int_{\mathbb{R}^d} \|T_{\mu_2\to\mu_0} - T_{\mu_2\to\mu_1}\|^2\, d\mu_2 = \|log_{\mu_2}(\mu_0) - log_{\mu_2}(\mu_1)\|^2_{\mathcal{T}_{\mu_2}\mathcal{P}_2}.$$

In the language of differential geometry, this suggests that the Wasserstein space exhibits a non-negative sectional curvature at any absolutely continuous probability measure μ_2. And indeed, this space is itself a geodesic space with non-negative curvature, as is shown in [1].

[6] Due to [1].

2.2 Evolution of measures

Now that we have described the geometry of this space, we are interested in studying the evolution of probability measures when a family of time-dependent vector fields is acting on the underlying space. More precisely let $v_t : \mathbb{R}^d \to \mathbb{R}^d$, where $t \geq 0$ represents time. The motion of particles in the underlying space is described by ODEs:

$$\dot{X}_t = v_t(X_t), \quad t \geq 0, \tag{4}$$

where X_t represents the position of a particle at time t.

Let $\mu_t(x)$ represent the probability distribution of particles at time t, where $\mu_t \in \mathcal{P}_2$. The distribution of μ_t will accordingly evolve over time, and in the context of this setup, we will show that this evolution is governed by the *continuity equation*, which expresses the conservation of probability as the underlying particles move through the space.

Preliminarly, we define $\partial_t \mu_t$ as the measure that satisfies $\int_{\mathbb{R}^d} g \, d(\partial_t \mu_t) = \partial_t \mathbb{E}[g(X_t)]$ for any test function $g : \mathbb{R}^d \to \mathbb{R}$ (where test means *smooth and compactly supported*).

Theorem 2.8. Let $(v_t)_{t\geq 0} : \mathbb{R}^d \to \mathbb{R}^d$ be a time dependent vector field with $v_t \in L_1(\mathbb{R}^d)$ for any $t \geq 0$, and suppose that particles evolve according to (4). Then $X_t \sim \mu_t$, where μ_t satisfies

$$\int_{\mathbb{R}^d} g \, d(\partial \mu_t) = \int_{\mathbb{R}^d} \langle \nabla g, v_t \rangle_2 \, d\mu_t, \tag{5}$$

for every test function g.

Proof. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a test function. Then,

$$\begin{aligned}
\int_{\mathbb{R}^d} g \, d(\partial \mu_t) &= \partial_t \int_{\mathbb{R}^d} g(x) \, d\mu_t \quad \text{by definition} \\
&= \partial_t \int_{\mathbb{R}^d} g(X_t(\omega)) \, d\mathbb{P}[\omega] \quad \text{by addressing an underlying probability space} \\
&= \int_{\mathbb{R}^d} \partial_t g(X_t(\omega)) \, d\mathbb{P}[\omega] \quad \text{by dominated convergence, since } \operatorname{supp}(g) \text{ is compact and } v_t \in L_1(\mathbb{R}^d) \\
&= \int_{\mathbb{R}^d} \langle \nabla g(X_t(\omega)), \dot{X}_t(\omega) \rangle_2 \, d\mathbb{P}[\omega] \quad \text{multivariate calculus} \\
&= \mathbb{E}[\langle \nabla g(X_t), \dot{X}_t \rangle_2] \quad \text{returning to the compact notation} \\
&= \mathbb{E}[\langle \nabla g(X_t), v_t(X_t) \rangle_2] \quad \text{by hypothesis} \\
&= \int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 \, d\mu_t(x), \quad \text{by definition of expectation}
\end{aligned}$$

which concludes our proof. □

Whenever $d\mu_t = f_t \, d\lambda$, with $f_t \in C^1(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$, then (5) is equivalent to

$$\partial_t f_t + \langle \nabla, (f_t v_t) \rangle_2 = 0 \tag{6}$$

in weak sense. To see this, fix a test function g, and notice that

$$\begin{aligned}
\int_{\mathbb{R}^d} g \, d(\partial_t \mu_t) &= \partial_t \int_{\mathbb{R}^d} g(x) f_t(x) \, d(x) \quad \text{by definition} \\
&= \int_{\mathbb{R}^d} g(x) \partial_t f_t(x) \, d(x). \quad \text{by dominated convergence}
\end{aligned}$$

On the other side,

$$\int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 \, d\mu_t = \int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 f_t(x) \, dx \text{ by definition}$$
$$= -\int_{\mathbb{R}^d} g(x) \langle \nabla, (f_t(x) v_t(x)) \rangle_2 \, dx, \text{ integration by parts with } f \text{ compactly supported}$$

which implies, by identification, that $\partial_t f_t + \langle \nabla, (f_t v_t) \rangle_2 = 0$.

In this flavor, up to defining $\langle \nabla, (\mu_t v_t) \rangle_2$ as the distribution that satisfies

$$\int_{\mathbb{R}^d} g(x) \, d(\langle \nabla, \mu_t v_t \rangle_2)(x) = -\int_{\mathbb{R}^d} \langle \nabla g(x), v_t(x) \rangle_2 \, d\mu_t(x),$$

for any test function g; even when μ_t does not admit density with respect to the Lebesgue measure, we say that (5) is equivalent to the weak generalized continuity equation (or weak continuity equation for brevity)

$$\partial_t \mu_t + \langle \nabla, (\mu_t v_t) \rangle_2 = 0. \tag{7}$$

Every *nice* curve of probability measures can be interpreted as a fluid moving along a time varying vector field. And by *nice* we mean:

Definition 2.9. A curve $t \mapsto \mu_t \in \mathcal{P}_2$ is said to be *absolutely continuous* if at every time, the metric derivative is finite, i.e., if

$$\text{for all } t, \ |\dot{\mu}|(t) := \lim_{s \to t} \frac{\mathcal{W}_2(\mu_s, \mu_t)}{|s - t|} < +\infty.$$

Theorem 2.10. Let $t \mapsto \mu_t$ be an absolutely continuous curve of measures. Then:

1. For any vector field $(\tilde{v})$ that satisfies the weak continuity equation, we have

$$|\dot{\mu}|(t) \leq \|\tilde{v}_t\|_{L_2(\mu_t)}.$$

2. Conversely, there exists a unique choice of vector field $(v_t)_{t \geq 0}$ that satisfies the weak continuity equation and

$$\|v_t\|_{L_2(\mu_t)} \leq |\dot{\mu}_t|.$$

Moreover, $v_t = \nabla \psi_t$ for $\psi : \mathbb{R}^d \to \mathbb{R}$ and

$$v_t = \lim_{\delta \to 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - id}{\delta}.$$

For a proof, see [1].

The theorem just states that starting from a curve $t \mapsto \mu_t$ we can identify any infinitesimal displacement as a constant speed geodesic between μ_t and $\mu_{t+\delta}$ and accordingly find uniquely the vector field v_t that induces it.

2.3 First variation of functionals

The goal of this and the following subsections is to understand how functionals $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R},\ \mu_t \mapsto \mathcal{F}(\mu_t)$ evolve as the underlying space is subject to a family of time dependent vector fields. This will lead us to a notion of gradient flow in $(\mathcal{P}_2, \mathcal{W}_2)$, which will be our main tool for applications.

This subsection is specifically devoted to an informal description of the tools required to work with functionals. Although its abstract nature may momentarily disrupt the flow of the exposition, we have chosen to include it here rather than in the preliminaries, so as to provide sufficient motivation drawn from the surrounding context.

We want to work our way towards a rigorous definition of *differential operator* associated to $\mathcal{F}$.

The first obstacle come from the non-linearity of probability measures (the sum of two probability measures is no longer a measure), so that the direct differentiation does not make sense. Indeed, we typically understand differentiation of a smooth function f on a Euclidean space as:

$$f(x+h) - f(x) = [(\delta f)(x)](h) + o(h), \qquad h \to 0,\ x, h \in \mathbb{R}^d,$$

where $[(\delta f)(x)]$ is a linear and bounded functional on $\mathbb{R}^d$, i.e., a matrix.

However, say the functional $\mathcal{F}$ admits an extension over the space of *signed measures* $\mathcal{M}$ (on $\mathbb{R}^d$), and that $\mathcal{F}$ satisfies the necessary regularity properties ensuring its differentiation. Then we may write:

$$\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \epsilon[\delta\mathcal{F}(\mu)](\chi) + o(\epsilon), \qquad \mu, \chi \in \mathcal{M},\ \epsilon \to 0, \tag{8}$$

where $[\delta\mathcal{F}(\mu)]$ is a continuous linear functional on the space $\mathcal{M}$ of finite signed measures on $\mathbb{R}^d$.

It turns out that every continuous linear functional on $\mathbb{R}^d$ has a particular form. This result, known as *Kantorovich-Rubinstein duality*, states that the dual space of $\mathcal{M}$ can the identified with the space of bounded continuous functions:

$$\mathcal{M}^* \simeq C_b(\mathbb{R}^d),$$

in the sense that for any linear functional G acting on measures defined on $\mathcal{M}$, there exists $g \in C_b(\mathbb{R}^d)$ such that:

$$G(\chi) = \int_{\mathbb{R}^d} g \, d\chi, \qquad \forall \chi \in \mathcal{M}.$$

Therefore, with a slight abuse of notation, i.e. identifying the map $[\delta\mathcal{F}(\mu)]$ with its Rieszs representative living in $C_b(\mathbb{R}^d)$, we can rewrite (8) as:

$$\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \int_{\mathbb{R}^d} \epsilon[\delta\mathcal{F}(\mu)] \, d\chi + o(\epsilon), \qquad \mu, \chi \in \mathcal{M},\ \epsilon \to 0. \tag{9}$$

2.4 Wasserstein gradient flows

Coming back to probability measures, by Theorem 2.10, we know that given a regular enough flow μ_t we can associate to it a unique vector field v_t such that (5) (a.k.a. the weak continuity equation) holds. Furthermore we can write in weak sense

$$\mu_t = \mu_0 + t\,\partial_t\mu_t + o(t), \quad t \to 0,$$

and by substituting this into (9), we get

$$\lim_{t\to 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_0)}{t} = \int_{\mathbb{R}^d} [\delta\mathcal{F}(\mu_0)]\, d(\partial_t \mu_t),$$

Now, assuming vanishing boundary conditions (which will be the case in functional of interest), we get:

$$\begin{aligned}\lim_{t\to 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_0)}{t} &= \int_{\mathbb{R}^d} [\delta\mathcal{F}(\mu_0)]\, d(\partial_t \mu_t) \\ &= \int_{\mathbb{R}^d} \langle (\nabla[\delta\mathcal{F}(\mu_0)])(x), v_t(x)\rangle_2 \, d\mu_t(x) = \langle (\nabla[\delta\mathcal{F}(\mu_0)]), v_t\rangle_{L_2(\mu_t)},\end{aligned}$$

by the same steps of Theorem 2.8.

Moreover, since $\nabla[\delta\mathcal{F}(\mu)]$ is the gradient of a bounded function, by the characterization in (3), we know that $\nabla[\delta\mathcal{F}(\mu_0)] \in \mathcal{T}_{\mu_0}\mathcal{P}_2$.

We have informally proven the following[7]:

Theorem 2.11. Let $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ be a functional with bounded first variation. Then, the Wasserstein gradient of $\mathcal{F}$ is the vector field defined by:

$$\nabla_W \mathcal{F}(\mu_0) = \nabla[\delta\mathcal{F}(\mu_0)],$$

where $[\delta\mathcal{F}(\mu_0)] \in C_b(\mathbb{R}^d)$ is tacitly the Rieszs representative of the first variation of $\mathcal{F}$ at μ_0, while ∇ is the usual Euclidean gradient.

Example 2.12. Given a potential $V : \mathbb{R}^d \to \mathbb{R}$, we can define the *potential energy* as

$$\mathcal{V}(\mu) := \int_{\mathbb{R}^d} V\, d\mu, \quad \text{for any } \mu \in \mathcal{P}_2.$$

Then,

$$\partial_t \mathcal{V}(\mu_t) = \int_{\mathbb{R}^d} V\, d(\partial_t \mu_t),$$

and thus we can identify $\delta\mathcal{V}(\mu) = V$, for any $\mu \in \mathcal{P}_2$. Therefore,

$$\nabla_{W_2} \mathcal{V}(\mu) = \nabla V, \quad \text{for any } \mu \in \mathcal{P}_2.$$

Example 2.13. Given $\mu \in \mathcal{P}_2^{ac}$, with $d\mu = f\, d\lambda$, we can define the *entropy functional* as

$$\mathrm{Ent}(\mu) := \int_{\mathbb{R}^d} f \log(f)\, d\lambda.$$

Then,

$$\partial_t \mathrm{Ent}(\mu_t) = \int_{\mathbb{R}^d} (\partial_t f_t \log(f_t) + \partial_t f_t)\, d\lambda = \int_{\mathbb{R}^d} \partial_t f_t (\log(f_t) + 1)\, d\lambda,$$

and therefore we can identify $\delta\mathrm{Ent}(\mu) = \log(f) + 1$. Then we can write

$$\nabla_{W_2} \mathrm{Ent}(\mu) = \nabla \log f.$$

[7] For a rigorous proof, take a look at [1].

Now, Theorem 2.11 is showing that the first variation of $\mathcal{F}$ naturally leads to a gradient structure in the Wasserstein space, indeed we can expand $\mathcal{F}(\mu_t)$ to the first order:

$$\mathcal{F}(\mu_{t+h}) = \mathcal{F}(\mu_t) + h\langle \nabla_{\mathcal{W}} \mathcal{F}(\mu_t), v_t \rangle_{L_2(\mu_t)} + o(h), \quad h \to 0.$$

We can now finally define the Wasserstein gradient flow of a functional.

Informally, a gradient flow in the Wasserstein space is a curve of measures $(\mu_t)_{t\geq 0}$ such that the tangent vector to the curve at t equals $-\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)$. Recalling that the tangent vector governs the evolution of $(\mu_t)_{t\geq 0}$ via the continuity equation (5), we arrive at the following definition.

Definition 2.14. Let $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ a functional. Then $(\mu_t)_{t\geq 0}$ is called the *Wasserstein gradient flow* of $\mathcal{F}$ if it solves the following PDE in the weak sense:

$$\partial_t \mu_t = \langle \nabla, (\mu_t \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)) \rangle_2. \tag{10}$$

Unsurprisingly, this gradient flow yields a principled approach for dynamically, smoothly evolving a probability measure in the Wasserstein space, with the aim of minimizing the objective functional $\mathcal{F}$:

$$\partial_t \mathcal{F}(\mu_t) = \langle \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t), v_t \rangle_{L_2(\mu_t)} = -\|\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)\|^2_{L_2(\mu_t)}.$$

Clearly, in order to have theoretical and quantitative guarantees about the convergence to a minima we need to address the convexity of the functional of interest. This will be studied in the next section for a specific case of interest.

3 Variational inference

Some of the most interesting theoretical guarantees that can be derived via Wasserstein spaces are in the context of *variational inference*. In order the introduce the latter, let us start from a concrete problem.

Let us suppose there is a probability measure $\pi \in \mathcal{P}_2$ that we would like to sample from or to visualize, for instance. Unfortunately we only know the measure up to a normalizing constant, i.e. we have access to $\tilde{\pi} = k\pi$ where $k = \frac{1}{\pi(\mathbb{R}^d)}$ is not solvable analytically and too computationally demanding.

A possible approach to overcome this problem is to define a reasonable notion of *divergence* $\mathcal{D}(\cdot\|\pi)$ between measures that enables us to approximate π with a measure μ^* from a fixed subset $Q \subseteq \mathcal{P}_2$ of known measures, by minimizing the divergence under the restriction of remaining in the subset (without the need to compute k in the process), as in classical constrained optimization problems. This approach is known in literature as *variational inference*.

$$\mu^* := \arg\min_{\mu \in Q} \mathcal{D}(\nu\|\pi).$$

In a nutshell, we anticipate that our divergence will be the *Kullback-Leibler* $\mathcal{D}_{KL}(\cdot\|\pi)$. Accordingly, we will study the convexity of this functional over the Wasserstein space to get important theoretical and quantitative guarantees for convergence to an optimal measure when addressing the problem via Wasserstein gradient flows. It is important to remark that our work in this section will be *theoretical*, we will not show any discretized algorithm usable in applications; but instead we will remain in the context of differential calculus (*"if we were able to implement continuity on computers, this section would guarantee the convergence of a Wasserstein gradient flow to a global minima in the context of variational inference"*). This approach still gives us a groundwork for when we will have to discretize and implement an algorithm.

As a last note, this framework not only provides a solid theoretical foundation for approximation and sampling, but it also lies at the heart of many generative models used in practice. A notable example is the class of *Variational Autoencoders* (VAEs), which aim to learn a generative process for complex data starting from a simple latent distribution.

3.1 Kullback-Leibler divergence

We start by properly defining our notion of divergence between measures. To avoid cumbersome notation and pathological issues, from now on we restrict ourselves to measures which are absolutely continuous with respect to the Lebesgue measure λ, although a number of results hold even in more general settings.

Definition 3.1. Given two probability measures $\mu, \pi \ll \lambda$ on $\mathbb{R}^d$, with $d\mu = g d\lambda$, $d\pi = f d\lambda$, the *Kullback-Leibler divergence* is defined as[8]:

$$\mathcal{D}_{KL}(\mu\|\pi) := \int_{\mathbb{R}^d} \log\left(\frac{g(x)}{f(x)}\right) g(x)\, dx = \int_{\mathbb{R}^d} \log\left(\frac{g}{f}\right) g\, d\lambda.$$

[8] Showing also the Riemannian notation for coherency with the literature.

The KL divergence quantifies how much the measure μ differs from the target measure π. Let us start by stating a property which is reasonable to ask for when dealing with notions of divergences.

Theorem 3.2. Given two probability measures $\mu, \pi \ll \lambda$ on $\mathbb{R}^d$, with $d\mu = g d\lambda$, $d\pi = f d\lambda$, we have that $\mathcal{D}_{KL}(\mu||\pi) \geq 0$, and $\mathcal{D}_{KL}(\mu||\pi) = 0$ if and only if $f = g$ $\lambda-$a.e.

Proof. Define the function $\varphi(x) = x \log x$, which is convex on $(0, \infty)$. Then we can write the Kullback-Leibler divergence as

$$\mathcal{D}_{\mathrm{KL}}(\mu||\pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{f}\right) g \, d\lambda = \int_{\mathbb{R}^d} \varphi\left(\frac{g}{f}\right) f \, d\lambda.$$

Now recall Jensen's inequality: if φ is convex and X is an integrable random variable, then

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]),$$

with equality if and only if X is almost surely constant.

Apply this inequality to the convex function φ, we get

$$\int \varphi\left(\frac{g}{f}\right) f \, d\lambda \geq \varphi\left(\int \frac{g}{f} f \, d\lambda\right) = \varphi\left(\int g \, d\lambda\right) = \varphi(1) = 0.$$

Therefore,

$$\mathcal{D}_{\mathrm{KL}}(\mu||\pi) \geq 0,$$

with equality if and only if $\frac{g}{f} = 1$ λ-almost everywhere, that is $f = g$ λ-a.e. □

Unfortunately, in general $\mathcal{D}_{KL}(\nu||\pi) \neq \mathcal{D}_{KL}(\pi||\nu)$, which already shows that the KL divergence is not a metric.

From now on, we are going to focus on the situation in which $\pi = f \, d\lambda$ with $f(x) \propto e^{-V(x)}$. This is very common in applications such as Bayesian statistics, complex systems and statistical mechanics.

As we hinted in the introduction, the optimization problem

$$\mu^* = \arg\min_{\mu \in \mathcal{P}_2^{ac}} \mathcal{D}_{KL}(\mu||\pi) \tag{11}$$

where $d\mu = g d\lambda$, $d\pi = f d\lambda = k\tilde{f} d\lambda = k d\tilde{\pi}$ (and we have direct access to $\tilde{f}$), does not require us to compute the normalizing constant k. Indeed, a direct calculation yields

$$\mathcal{D}_{\mathrm{KL}}(\mu||\pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{k\tilde{f}}\right) g \, d\lambda = \int_{\mathbb{R}^d} \left[\log\left(\frac{g}{\tilde{f}}\right) - \log(k)\right] g \, d\lambda.$$

Separating the terms:

$$\mathcal{D}_{\mathrm{KL}}(\mu||\pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{\tilde{f}}\right) g \, d\lambda - \log(k) \int_{\mathbb{R}^d} g \, d\lambda,$$

but $\int g \, d\lambda = 1$, and thus:

$$\mathcal{D}_{\mathrm{KL}}(\mu||\pi) = \mathcal{D}_{\mathrm{KL}}(\mu||\tilde{\pi}) - \log(k)$$

which implies

$$\arg\min_{\mu \in \mathcal{P}_2^{ac}} \mathcal{D}_{KL}(\mu||\pi) = \arg\min_{\mu \in \mathcal{P}_2^{ac}} \mathcal{D}_{KL}(\mu||\tilde{\pi}),$$

as wanted to prove.

We will show that whenever V is convex, then $\mathcal{D}_{KL}(\cdot||\pi)$ is *geodesically convex* (in $(\mathcal{P}_2^{ac}, \mathcal{W}_2)$): a property that brings important results, as the next subsection will show.

3.2 Geodesic convexity

Let us recall some general knowledge about convex functions from analysis.

Definition 3.3. A function $V : \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* if

$$V((1-t)x_0 + tx_1) \leq (1-t)V(x_0) + tV(x_1), \qquad \forall x_0, x_1 \in \mathbb{R}^d, \ \forall t \in [0,1].$$

In short, a function is convex if for any two fixed points in the domain, if we project the segment joining them to the epigraph the curve we obtain lies uniformly below the convex interpolation of the images of the point.

We can strength this definition to requiring a *uniform rate of convexity*, as shown in the following defintion.

Definition 3.4. A function $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be *α-convex* (or *α-strongly convex*) for some $\alpha > 0$ if

$$V((1-t)x_0 + tx_1) \leq (1-t)V(x_0) + tV(x_1) - \frac{\alpha}{2}t(1-t)\|x_1 - x_0\|^2, \qquad \forall x_0, x_1 \in \mathbb{R}^d, \ \forall t \in [0,1].$$

We have the following useful characterization.

Proposition 3.5. Given a function $V : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we have

$$\begin{aligned} V \text{ is } \alpha\text{−convex} &\iff V - \alpha\|\cdot\|^2 \text{ is convex} \\ &\iff D^2V \geq \alpha Id_d. \end{aligned}$$

Proof. V is α−convex if and only if

$$\begin{aligned} V((1-t)x_0 + tx_1) \leq (1-t)\left(V(x_0) - \frac{\alpha}{2}\|x_0\|^2\right) + t\left(V(x_1) - \frac{\alpha}{2}\|x_1\|^2\right) \\ + \frac{t\alpha}{2}\|x_1\|^2 + (1-t)\frac{\alpha}{2}\|x_0\|^2 - \frac{\alpha}{2}t(1-t)\|x_1 - x_0\|^2 \\ + \frac{\alpha}{2}\|(1-t)x_0 + tx_1\|^2 - \frac{\alpha}{2}\|(1-t)x_0 + tx_1\|^2 \end{aligned}$$

which if true if and only if $V - \frac{\alpha\|\cdot\|^2}{2}$ is convex, because the terms in gray add up to zero. The last condition is trivially equivalent to requiring $D^2V - \alpha Id_d \geq 0$, by differentiating twice. □

In order to extend the concept of convexity to functionals we have to use the same strategy of the first section: instead of segments, we look at geodesics.

Definition 3.6. Let $(\mathcal{X}, d)$ be a geodesic space. A functional $\mathcal{F} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be *geodesically convex* (or *displacement convex*) if for every pair of points $x_0, x_1 \in \mathcal{X}$, and every constant-speed geodesic $(x_t)_{t\in[0,1]}$ joining them, the map $t \mapsto V(x_t)$ is convex. That is,

$$V(x_t) \leq (1-t)V(x_0) + tV(x_1), \qquad \forall t \in [0,1].$$

Definition 3.7. Let $(\mathcal{X}, d)$ be a geodesic space. A functional $\mathcal{F} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be *α-geodesically convex* for some $\alpha > 0$ if for every pair of points $x_0, x_1 \in \mathcal{X}$, and every constant-speed geodesic $(x_t)_{t\in[0,1]}$ joining them, the map $t \mapsto V(x_t)$ satisfies

$$V(x_t) \leq (1-t)V(x_0) + tV(x_1) - \frac{\alpha}{2}t(1-t)\|x_1 - x_0\|^2, \qquad \forall t \in [0,1].$$

For us, the geodesic space upon which the functionals are acting will be $(\mathcal{P}_2, \mathcal{W}_2)$.

We can already show that the entropy functional, introduced in Example 2.12, is geodesically convex.

Proposition 3.8. $\mathrm{Ent}(\mu)$ is geodesically convex.

Proof. Let $\mu_0 = f_0\, d\lambda$ and $\mu_1 = f_1\, d\lambda$ be two absolutely continuous probability measures in $\mathcal{P}_2$. Consider the Wasserstein geodesic $(\mu_t)_{t\in[0,1]}$ defined by

$$\mu_t := ((1-t)\mathrm{id} + tT_{\mu_0\to\mu_1})_{\#}\mu_0.$$

Clearly $\mu_t = f_t\, d\lambda$ is absolutely continuous for each $t \in [0,1]$. Using the change of variables formula from Theorem 1.14, the density f_t satisfies:

$$f_t((1-t)id + tT_{\mu_0\to\mu_1}))(x) = \frac{f_0(x)}{\det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x))},$$

and we can define $y = g(x) := ((1-t)id + tT_{\mu_0\to\mu_1})(x)$ (which is differentiable and bijective on $\mathbb{R}^d$, μ_0-a.s.), so that our change of variables reads $f_t(g(x)) = \frac{f_0(x)}{\det(\nabla g(x))}$.

The entropy at time t is:

$$\begin{aligned}
\mathrm{Ent}(\mu_t) &= \int_{\mathbb{R}^d} f_t(y)\log f_t(y)\, dy \\
&= \int_{\mathbb{R}^d} f_t(g(x))\log f_t(g(x))\det(\nabla g(x))\, dx \\
&= \int_{\mathbb{R}^d} \frac{f_0(x)}{\det(\nabla g(x))}\log\left(\frac{f_0(x)}{\det(\nabla g(x))}\right)\det(\nabla g(x))\, dx \\
&= \int f_0(x)\log\left(\frac{f_0(x)}{\det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x))}\right)dx,
\end{aligned}$$

This can be split as:

$$\mathrm{Ent}(\mu_t) = \int f_0(x)\log f_0(x)\, dx - \int f_0(x)\log\det((1-t)Id_d + t\nabla T_{\mu_0\to\mu_1})(x))\, dx.$$

The first term is constant in t, and the second term is concave in t, since $A \mapsto \log\det A$ is concave on the space of positive definite matrices, and $t \mapsto (1-t)id + t\nabla T_{\mu_0\to\mu_1})(x)$ is affine in t.

Therefore, $t \mapsto \mathrm{Ent}(\mu_t)$ is convex, and we conclude:

$$\mathrm{Ent}(\mu_t) \le (1-t)\mathrm{Ent}(\mu_0) + t\mathrm{Ent}(\mu_1),$$

proving that the entropy functional is geodesically convex. □

Additionally, we have an important result regarding the potential energy functional $\mathcal{V}(\mu) = \int_{\mathbb{R}^d} V\, d\mu$ defined in Example 2.13.

Theorem 3.9. *V is $\alpha-$convex if and only if $\mathcal{V}$ is $\alpha-$geodesically convex.*

Proof. ($\Rightarrow$) We claim that if $\mu_0, \mu_1 \in \mathcal{P}_2$, and $\bar{\gamma} \in \Gamma(\mu_0, \mu_1)$ is the optimal coupling between the two, then $t \mapsto \mu_t$ is convex, where μ_t is the constant speed geodesic.

$$\begin{aligned}\int_{\mathbb{R}^d} V\, d\mu_t &= \int_{\mathbb{R}^d \times \mathbb{R}^d} V((1-t)x_0 + tx_1)\, d\bar{\gamma}(x_0, x_1) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \left[(1-t)V(x_0) + tV(x_1) - \frac{\alpha}{2}t(1-t)\|x_1 - x_0\|^2\right] d\bar{\gamma}(x_0, x_1) \\ &= (1-t)\int_{\mathbb{R}^d} V\, d\mu_0 + t\int_{\mathbb{R}^d} V\, d\mu_1 - \frac{\alpha}{2}t(1-t)\mathcal{W}_2^2(\mu_0, \mu_1).\end{aligned}$$

($\Leftarrow$) For the other direction, it is sufficient to apply the $\alpha-$geodesically convexity of $\mathcal{V}$ to δ_{x_0} and δ_{x_1} to get the $\alpha-$convexity of V. □

We can finally show the main result for this subsection.

Theorem 3.10. If $\pi = f\, d\lambda$, with $f = ke^{-V}$ for some normalizing constant k, and $V : \mathbb{R}^d \to \mathbb{R}$ is convex, then $\mathcal{D}_{KL}(\cdot\|\pi)$ is geodesically convex. Accordingly, if V is $\alpha-$convex, then $\mathcal{D}_{KL}(\cdot\|\pi)$ is $\alpha-$geodesically convex.

Proof. Let $\pi = f\, d\lambda$ with $f = ke^{-V}$ for some normalizing constant $k > 0$, and assume $V : \mathbb{R}^d \to \mathbb{R}$ is convex. For any absolutely continuous probability measure $\mu \ll \lambda$ with density $g = \frac{d\mu}{d\lambda}$, we can write the Kullback–Leibler divergence as:

$$\mathcal{D}_{KL}(\mu\|\pi) = \int_{\mathbb{R}^d} g \log\left(\frac{g}{f}\right) d\lambda = \int_{\mathbb{R}^d} g \log g\, d\lambda + \int_{\mathbb{R}^d} Vg\, d\lambda + \log k.$$

That is,

$$\mathcal{D}_{KL}(\mu\|\pi) = \mathrm{Ent}(\mu) + \mathcal{V}(\nu) + \log k,$$

where $\mathrm{Ent}(\mu)$ is the (negative) entropy relative to the Lebesgue measure; and $\mathcal{V}(\mu) := \int V\, d\mu$ is the potential energy. Now observe the following:

1. The entropy functional $\mu \mapsto \mathrm{Ent}(\mu)$ is geodesically convex by Proposition 3.8.
2. From Theorem 3.9, if V is convex, then $\mathcal{V}(\mu)$ is also geodesically convex.

Since the sum of two geodesically convex functionals remains geodesically convex, we conclude that $\mathcal{D}_{KL}(\cdot\|\pi)$ is geodesically convex.

Moreover, if V is α-convex, then $\mathcal{V}$ is α-geodesically convex (again by Theorem 3.9), and therefore $\mathcal{D}_{KL}(\cdot\|\pi)$ is α-geodesically convex as well. □

The main point of this section is that if the family $\mathcal{Q}$ is geodesically convex (i.e. whenever $\mu_0, \mu_1 \in \mathcal{Q}$, then the constant speed geodesic between them is also in $\mathcal{Q}$), and $\mathcal{V}$ is $\alpha-$geodesically convex then the solution to (11) is unique.

To see this, suppose for the sake of contradiction that there exist two distinct minimizers $\mu_0^*, \mu_1^* \in \mathcal{Q}$ such that $\mathcal{V}(\mu_0^*) = \mathcal{V}(\mu_1^*) = \inf_{\mu\in\mathcal{Q}} \mathcal{V}(\mu)$. Since $\mathcal{Q}$ is geodesically convex, the constant-speed geodesic $(\mu_t^*)_{t\in[0,1]}$ connecting μ_0^* and μ_1^* is contained in $\mathcal{Q}$.

By α-geodesic convexity of $\mathcal{V}$, we have for all $t \in (0, 1)$:

$$\mathcal{V}(\mu_t^*) \leq (1-t)\mathcal{V}(\mu_0^*) + t\mathcal{V}(\mu_1^*) - \frac{\alpha}{2}t(1-t)W_2^2(\mu_0^*, \mu_1^*).$$

Since both μ_0^* and μ_1^* are minimizers, the right-hand side equals $\inf_{\mu\in Q}\mathcal{V}(\mu)$ minus a strictly positive term (unless $\mu_0^* = \mu_1^*$). Therefore:

$$\mathcal{V}(\mu_t^*) < \inf_{\mu\in Q}\mathcal{V}(\mu),$$

which contradicts the minimality of μ_0^* and μ_1^*.

We conclude that the minimizer is unique.

Recall that by what we computed in the previous section (examples 2.12, 2.13), we can write the Wasserstein gradient flow as

$$\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi) = \nabla V + \nabla \log f. \tag{12}$$

On top of uniqueness of the optimizer in (11), we have the Poljak–Łojasiewicz inequality that gives us a rate of convergence to the optimal solution when the optimization is addressed via Wasserstein gradient flows.

Theorem 3.11. Let $\pi \propto \exp(-V)$ be a density on $\mathbb{R}^d$, where V is α-convex. Let $Q \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be geodesically convex. Then, the Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot\,\|\,\pi)$ constrained to lie in Q satisfies

$$\mathcal{D}_{KL}(\mu_t\|\pi) - \mathcal{D}_{KL}(\mu^*\|\pi) \leq e^{-2\alpha t}[\mathcal{D}_{KL}(\mu_0\|\pi) - \mathcal{D}_{KL}(\mu^*\|\pi)].$$

3.3 Hints on the JKO scheme

Let us make a brief historical remark. So far, we have discussed gradient flows in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ from a purely theoretical viewpoint. In particular, we have studied the Wasserstein gradient flow of the Kullback–Leibler divergence functional $\mathcal{D}_{KL}(\cdot\|\pi)$, which under suitable convexity assumptions converges uniquely and exponentially fast to the target distribution π.

However, a natural question arises: *how can we actually compute such gradient flows on a computer?* This will be addressed profoundly in the following section via particle methods; but we aim to describe here another very general method, adaptable to other functionals as well: the JKO scheme.

Let us begin with a fundamental idea from the numerical analysis of gradient flows in finite-dimensional spaces. When it is not possible to compute the exact solution of a gradient flow, a common strategy is to use a *time-discretized approximation scheme*.

In Euclidean space, given a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, the gradient flow associated to F is defined by the differential equation

$$\frac{dx}{dt} = -\nabla F(x).$$

To approximate the solution of this flow, a classical method is the *implicit Euler scheme*, which updates the state according to

$$x_{k+1} = x_k - \tau\nabla F(x_{k+1}),$$

where $\tau > 0$ is the time-step size. This is called an *implicit* scheme because the gradient is evaluated at the new point x_{k+1}, rather than the current point x_k.

This update rule can be equivalently reformulated as a minimization problem:

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{1}{2\tau}\|x - x_k\|^2 + F(x) \right\}.$$

This formulation highlights an important interpretation: each step of the flow seeks a balance between staying close to the previous iterate x_k and decreasing the energy $F(x)$. In other words, *gradient descent can be viewed as a sequence of variational minimization steps*.

This variational viewpoint serves as the foundation for the JKO scheme in the Wasserstein space.

Inspired by this observation, Jordan, Kinderlehrer, and Otto in [8] proposed to approximate gradient flows in the Wasserstein space by replacing the Euclidean squared distance with the squared Wasserstein distance $\mathcal{W}_2^2$, and working with functionals $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

The resulting *JKO scheme* with time-step $\tau > 0$ defines a sequence $(\mu_k^\tau)_{k\in\mathbb{N}}$ recursively as:

$$\mu_{k+1}^\tau \in \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \mu_k^\tau) + \mathcal{F}(\mu) \right\}. \tag{13}$$

Each step of this scheme selects a new measure μ_{k+1}^τ by balancing:

1. Proximity to the previous iterate via the Wasserstein distance;
2. Decrease in energy encoded by the functional $\mathcal{F}$.

As $\tau \to 0$, and under mild regularity assumptions, the piecewise constant or piecewise interpolated curves $\mu^\tau(t)$ converge to a continuous-time curve $(\mu_t)_{t\geq 0}$ that solves the Wasserstein gradient flow of $\mathcal{F}$. This convergence has been rigorously established in various settings.

In our case, we take $\mathcal{F}(\mu) = \mathcal{D}_{KL}(\mu\|\pi)$. As seen before this functional admits a representation:

$$\mathcal{D}_{KL}(\mu\|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{f}\right) g\, d\lambda = \mathrm{Ent}(\mu) + \mathcal{V}(\mu) + \log k,$$

with $f = ke^{-V}$ and $d\mu = g\, d\lambda$. Hence, the JKO scheme becomes:

$$\mu_{k+1}^\tau \in \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \mu_k^\tau) + \mathrm{Ent}(\mu) + \int V\, d\mu \right\}. \tag{14}$$

4 Particles variational inference

The results derived in the previous section hold whenever $\mathcal{Q} \subseteq \mathcal{P}_2^{ac}$. In particular there exists a whole branch of variational inference focusing on the case where $\mathcal{Q} = \mathcal{Q}_G$ is the set of gaussian probability measures, and this framework is known as *gaussian variational inference*. We will not address this branch in this manuscript, but we redirect to [5] for a summary of the main results.

Instead, we will focus here on the case in which $\mathcal{Q} = \mathcal{Q}_N$ is the set of *empirical measures*, i.e., $\mu \in \mathcal{Q}_N$ if and only if $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, for $N \in \mathbb{N}$ and $x_i \in \mathbb{R}^d$.

This family is extremely expressive, in fact as $N \to +\infty$ we can always find a sequence of emprical measures that converges weakly to any measure $\pi \in \mathcal{P}_2$, as a consequence of the strong law of large numbers. On top of that we are motivated to look at this measures, as since there is an identification

$$\dot{X}_t = v_t(X_t) \longleftrightarrow \partial_t \mu_t + \langle \nabla, (\mu_t v_t)\rangle_2,$$

we can see the evolution of μ_t not only by looking at the PDE, but also by considering the finite number of particles evolving via the ODE, which is more tractable in general.

Now we have a trade-off in front of us: the results from the previous subsection, especially Theorem 3.11, do not hold a priori in this new setting; and moreover despite there exist versions of the KL divergence allowing for non absolutely continuous measure, no version of the KL divergence can be extended on empirical measures due to logarithmic poles, making it difficult for us to adapt (11).

We will not deal directly with the first problem, as heuristically we can think that the empirical measures approximately behave well as the number of particles N is big enough. To address the second problem we will rely on *reproducing kernel Hilbert spaces*, as done in [9].

It is remarkable to notice that this framework could allow for discretization of the flow easily thanks to the ODE formulation, and subsequently for some manageable implementations.

4.1 Many particles systems

We start by looking at how Wasserstein gradient flows naturally describe a *mean field interaction* when applied to empirical meausures.

The WGF with respect to $\mathcal{F} : \mathcal{Q}_N \to \mathbb{R}$ is the PDE associated to

$$\dot{X}_t = -\nabla_{W_2}\mathcal{F}(\mu_t)(X_t), \tag{15}$$

where μ_t is the law of X_t.

Let us initialize our gradient flow at $\mu_0 := \frac{1}{N}\sum_{i=1}^N \delta_{X_0^i}$ for some $X_0^i \in \mathbb{R}^d$, $i = 1, ..., N$. Particles interactions systems can be summarized in situations where particles $(X_t^1, \dots, X_t^N)$ are subject to dynamics of the form

$$\dot{X}_t^i = V_t^i(X_t^1, \dots, X_t^N), \quad i \in 1, ..., N.$$

Note that in the case of Wasserstein gradient flows, we further have each particle interact with the others only via the distribution μ_t, and that the interactions have the same form for all particles.

This means that the general dynamics is captured by

$$\dot{X}_t^i = V_t^i(X_t^1, ..., X_t^N) = V_t^i(X_t^i, \mu_t) = V_t(X_t^i, \mu_t)$$

which is an equation describing a *mean field interaction* system. These systems are convenient because it is strictly equivalent to describe the dynamics of each particle and that of their distribution, as hinted in the introduction.

Since the Wasserstein gradient flow only moves particles, the weights in μ_0 do not change over time: if the Wasserstein gradient flow is initialized at

$$\mu_0 := \sum_{j=1}^{N} w_0^j \delta_{x_0^j},$$

where $w_0^j \geq 0,\ j \in [N]$ and $\sum_{j=1}^N w_0^j = 1$, then

$$\mu_t := \sum_{j=1}^{N} w_0^j \delta_{x_t^j},$$

where $X_t^1, \ldots, X_t^N$ evolve according to the ODE (15). In particular this holds if all the weights are $\frac{1}{N}$, as is the case for measures in $\mathcal{Q}_N$.

4.2 Stein variational gradient descent

As stated previously, we want to solve

$$\mu^* := \arg\min_{\mu \in \mathcal{Q}_N} \mathcal{D}_{KL}(\mu \| \pi)$$

where $\mathcal{Q}_N$ is the set of empirical measures with a fixed number of particles, and $d\pi = f d\lambda$ with $f \propto e^{-V}$.

We want to approach the problem via Wasserstein gradient flows, i.e. starting from a measure $\mu_0 := \frac{1}{N}\sum_{i=1}^N \delta_{x_0^j}$ and evolving it via the ODE (15). There comes a problem: $\nabla_{W_2}\mathcal{D}_{KL}(\mu_t \| \pi)$ is not well defined when $\mu_t \in \mathcal{Q}_N$.

To overcome this issue, we consider instead the effect of pushing μ forward along a smooth perturbation of the identity. Given a smooth vector field $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ (each component of φ is an element of a RKHS $\mathcal{H}$), define the transport map

$$T_\epsilon(x) := x + \epsilon\varphi(x),$$

and let $\mu_\epsilon := (T_\epsilon)_{\#}\mu$. We then study how the KL divergence changes under this perturbation. It can be shown (see [9]) that:

$$\frac{d}{d\epsilon}\mathcal{D}_{KL}(\mu_\epsilon \| \pi)\Big|_{\epsilon=0} = -\mathbb{E}_{X\sim\mu}\left[\langle \nabla \log f(X), \varphi(X)\rangle_2 + \langle \nabla, \varphi(X)\rangle_2\right].$$

SVGD chooses the direction φ that most decreases the KL divergence per unit time, under the constraint that φ lies in a unit ball of a Reproducing Kernel Hilbert Space (RKHS) of vector-valued functions $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$. That is, we solve:

$$\varphi^* = \arg\max_{\varphi \in \mathcal{H}^d,\, \|\varphi\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_{X\sim\mu}\left[\langle \nabla \log f(X), \varphi(X)\rangle_2 + \langle \nabla, \varphi(X)\rangle_2\right].$$

Since we assume that the target measure π has density $f \propto e^{-V}$ with respect to the Lebesgue measure, we have:

$$\nabla \log f(x) = -\nabla V(x).$$

Substituting this expression into the variational problem, the optimal direction φ^* becomes the constrained maximizer of:

$$\mathbb{E}_{X\sim\mu}\left[-\langle \nabla V(X), \varphi(X)\rangle_2 + \langle \nabla, \varphi(X)\rangle_2\right].$$

This optimization problem fits naturally into the framework of regularized empirical risk minimization in a reproducing kernel Hilbert space. Indeed, maximizing a linear functional under a unit-norm constraint in $\mathcal{H}^d$ is equivalent, via convex duality, to minimizing a regularized objective of the form:

$$\varphi^* = \arg\min_{\varphi\in\mathcal{H}^d} \left\{ \frac{1}{N}\sum_{i=1}^{N} \left[\langle \nabla V(x_i), \varphi(x_i)\rangle - \langle \nabla, \varphi(x_i)\rangle_2\right] + \lambda \|\varphi\|^2_{\mathcal{H}^d} \right\}.$$

This expression matches the structure of the representer theorem (Theorem 1.28), where the empirical loss corresponds to the first-order and divergence terms, and the regularization is given by the RKHS norm. As a result, the optimal solution admits a finite-dimensional representation in terms of the kernel evaluated at the training points.

By the theory of vector-valued RKHSs and the representer theorem (extended componentwise), the optimal direction φ^* can be expressed as:

$$\varphi^*(x) = \sum_{i=1}^{N} K(x_i, x)c_i, \qquad \text{with } c_i \in \mathbb{R}^d,$$

where K is a positive-definite scalar kernel on $\mathbb{R}^d$.

It turns out (see again [9]) that the explicit form of the optimal φ^* is given by:

$$\varphi^*(x) = \frac{1}{N}\sum_{j=1}^{N} \Big[K(x_j, x)(-\nabla V(x_j)) + \nabla_{x_j} K(x_j, x)\Big].$$

This yields a velocity field that is well-defined and smooth, even when μ is an empirical measure.

Finally, the particles evolve according to the following deterministic ODE:

$$\dot{X}_i(t) = \varphi^*(X_i(t)),$$

which drives each particle in the direction of lower potential (according to V), while maintaining diversity through repulsive interactions induced by the kernel gradient. This defines the SVGD particle system.

4.3 Implementation of SVGD

We now provide a toy implementation of the SVGD particle dynamics corresponding to the Wasserstein gradient flow of the KL divergence over the space of empirical measures Q_N, as discussed in the previous sections.

Let $\pi \in \mathcal{P}_2^{ac}$ be a target probability measure absolutely continuous with respect to the Lebesgue measure λ, and let $f := \frac{d\pi}{d\lambda}$ denote its density. We assume the density has the following form:

$$f(x) \propto e^{-V(x)}, \qquad \text{with } V(x) := \frac{x^4}{4} - \frac{x^2}{2}.$$

The potential V is smooth, and the density f is strictly positive and differentiable, so that π satisfies all the regularity assumptions required to define the Wasserstein gradient of the KL divergence.

We initialize the empirical measure $\mu_0 = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ by sampling N particles from the standard normal distribution. The SVGD flow then evolves the particles according to the dynamics

$$\dot{X}_i = \varphi^*(X_i) = \frac{1}{N}\sum_{j=1}^{N}\Big[-K(X_j, X_i)\nabla V(X_j) + \nabla_{X_j} K(X_j, X_i)\Big],$$

where K is a positive definite kernel. In our implementation, we use the RBF kernel

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{h}\right).$$

Heuristically, setting the bandwith as the median of $\{K(X_j, X_i)\}_{1\le i,j\le N}$ after each iteration, is known to yield good performances for our algorithm. The evolution is implemented via explicit Euler scheme, and the step-size has been empirically set to ?.

MISSING PIC.

The code is available on this repository.

5 Sampling

One of the most interesting applications of the theory of Wasserstein gradient flows is that it gives a geometric interpretation of the Langevin diffusion, as introduced in the seminal work by Jordan, Kinderlehrer, and Otto [8].

Let again $V : \mathbb{R}^d \to \mathbb{R}$ be a smooth potential, and let $d\pi = f d\lambda$ be the probability distribution over $\mathbb{R}^d$ with density $f \propto \exp(-V)$.

Suppose we want to generate samples from the distribution π. Contrary to variational inference, we do not aim to approximate π with a *close* measure, but rather we want to develop a procedure that allows to sample directly from π without the need to compute the normalizing constant.

A common approach to this sampling problem is called **Markov Chain Monte Carlo** (MCMC), where the goal is to design a Markov chain whose stationary distribution is the target distribution π. A standard MCMC algorithm is obtained by discretizing the **Langevin diffusion**, which is the solution to the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. If ∇V is, for example, Lipschitz continuous, then this SDE has a unique strong solution for any initial condition, and its stationary distribution is π.

In short, to have samples from π our strategy is to evolve independently many particles from a generic distribution μ_0 via Langevin diffusion. After a sufficient number of iterations[9] they will be distributed accordingly with π.

The catch is that if we track the evolution of the marginal law $\mu_t := \mathrm{law}(X_t)$ of the Langevin diffusion, then $(\mu_t)_{t\geq 0}$ follows the Wasserstein gradient flow of the Kullback–Leibler divergence $\mathrm{KL}(\cdot\|\pi)$.

This idea is the starting point of a growing literature that connects sampling with optimization, using tools like gradient flows and optimization algorithms to sample efficiently.

5.1 Langevin diffusion as a Wasserstein gradient flow

The main result is summarized in the theorem:

Theorem 5.1. The density of the marginal law $\mu_t := \mathrm{law}(X_t)$ of the Langevin diffusion with potential V is given by the solution to the Fokker–Planck equation

$$\partial_t f_t = \Delta f_t + \langle \nabla, (f_t \nabla V)\rangle_2. \tag{16}$$

But in our framework, the density of μ_t in (16) is exactly the gradient flow of $\mathcal{D}_{KL}(\cdot\|\pi)$! To see this, just notice that we can write (following (12))

$$\nabla_{W_2}\mathcal{D}_{\mathrm{KL}}(\mu_t\|\pi) = \nabla \log f_t + \nabla V,$$

whose flow equation reads

$$\partial_t f_t + \langle \nabla, (f_t \nabla (\log f_t + V))\rangle_2 = 0;$$

[9] This has to be specified in implementations.

and we can expand the divergence term as

$$\langle \nabla, (f_t \nabla \log f_t) \rangle = \langle \nabla, (\nabla f_t) \rangle = \Delta f_t,$$

so the equation becomes

$$\partial_t f_t = \Delta f_t + \langle \nabla, (f_t \nabla V) \rangle_2,$$

which is precisely the Fokker–Planck equation associated with the Langevin dynamics.

As a special case when $V = 0$, we also obtain the following proposition.

Proposition 5.2. If $(\mu_t)_{t\geq 0}$ is the marginal law of a (rescaled) Brownian motion $(\sqrt{2}B_t)_{t\geq 0}$, then $(\mu_t)_{t\geq 0}$ solves the heat equation $\partial_t f_t = \Delta f_t$, and it is the Wasserstein gradient flow of the entropy functional Ent.

We showed in Theorem 3.11 that the strong log-concavity of π implies rapid convergence of the Wasserstein gradient flow of the KL divergence. Therefore, we immediately obtain the following elegant convergence result for the Langevin diffusion.

Proposition 5.3. Let π be an α-strongly log-concave measure, and let $(\mu_t)_{t\geq 0}$ denote the marginal law of the Langevin diffusion with stationary distribution π. Then,

$$\mathcal{D}_{KL}(\mu_t \,\|\, \pi) \leq e^{-2\alpha t}\, \mathcal{D}_{KL}(\mu_0 \,\|\, \pi).$$

To wrap up this section, let us summarize the main ideas. The Wasserstein gradient flow (WGF) of the KL divergence can be viewed as a deterministic process. In contrast, our discussion in this subsection began with the Langevin diffusion, which describes a stochastic dynamic:

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t. \tag{LD}$$

But we have shown that the distribution $\mu_t := \mathrm{law}(X_t)$ satisfies the Fokker–Planck equation:

$$\partial_t \mu_t = \Delta \mu_t + \mathrm{div}(\mu_t \nabla V), \tag{FP}$$

which also characterizes the Wasserstein gradient flow of the Kullback-Leibler divergence $\mathcal{D}_{KL}(\cdot\|\pi)$. It is important to highlight that (FP) is a coarser description than (LD), since (LD) retains detailed information about time correlations within the stochastic process.

In summary, we have the chain relations:

$$\text{(LD)} \Longrightarrow \text{(FP)} \iff \text{(WGF)},$$

meaning that all three formalisms represent the same trajectory in the space of probability distributions.

5.2 Implementation of Lagenvin diffusion

The goal of this subsection is to give a minimal working example that illustrates how Langevin diffusion can be implemented and used for sampling from an absolutely continuous target distribution.

We consider the one-dimensional case $d = 1$, and choose a nonconvex potential

$$V(x) := \frac{x^4}{4} - \frac{x^2}{2},$$

The Langevin diffusion associated with this potential is defined as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. The stationary distribution of this SDE is exactly π, even though the normalizing constant of f is unknown.

Since we cannot simulate continuous-time processes exactly, we discretize the Langevin diffusion using the *Euler–Maruyama scheme*. For a fixed step size $\eta > 0$, we define the discrete-time update:

$$X_{k+1} = X_k - \eta\nabla V(X_k) + \sqrt{2\eta}\,\xi_k, \quad \xi_k \sim \mathcal{N}(0,1).$$

This scheme is straightforward to implement and requires only evaluation of the gradient $\nabla V(x) = x^3 - x$, and generation of Gaussian noise.

In our numerical experiment, we initialize a collection of particles from a standard normal distribution and evolve them independently using the discretized Langevin dynamics. After a sufficient number of steps, the empirical distribution of the particles converges toward the target density $f(x) \propto e^{-V(x)}$.

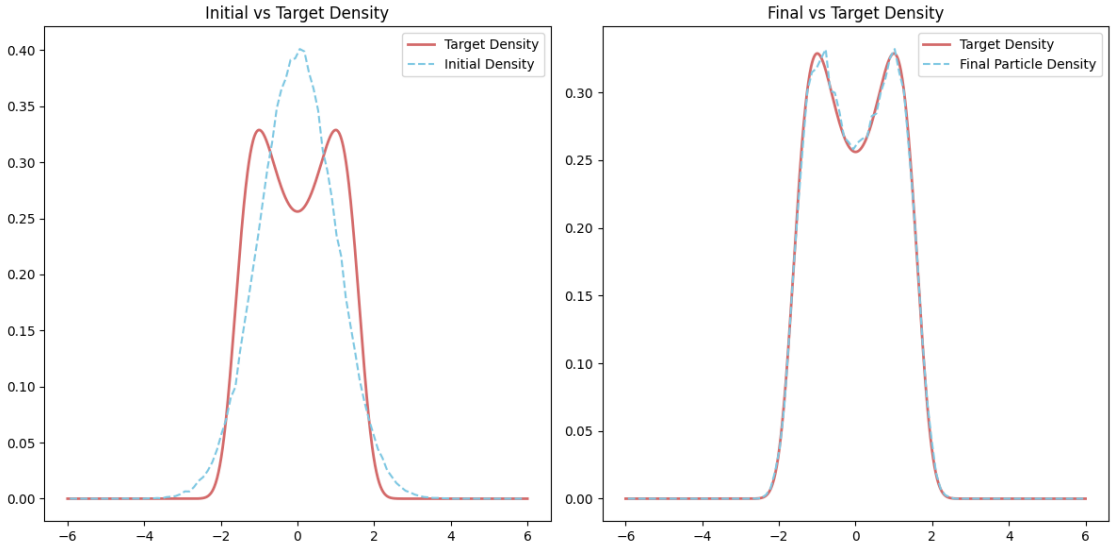

Figure 6: Langevin diffusion, implementation results.

The code is available on this repository.

References

[1] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[2] Luigi Ambrosio, Elia Bruè, and Daniele Semola. *Lectures on Optimal Transport*. 01 2024.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Calculus and heat flow in metric measure spaces and applications to spaces with ricci bounds from below. *Inventiones mathematicae*, 195(2):289–391, February 2013.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Metric measure spaces with riemannian ricci curvature bounded from below. *Duke Mathematical Journal*, 163(7), May 2014.

[5] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024.

[6] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. EMS Press, Berlin, second edition edition, 2023.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[8] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[9] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2019.

[10] John Lott and Cedric Villani. Ricci curvature for metric-measure spaces via optimal transport, 2006.

[11] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.