CrossMark

{Euclidean, metric, and Wasserstein} gradient flows: an overview

Filippo Santambrogio[1]

Abstract This is an expository paper on the theory of gradient flows, and in particular of those PDEs which can be interpreted as gradient flows for the Wasserstein metric on the space of probability measures (a distance induced by optimal transport). The starting point is the Euclidean theory, and then its generalization to metric spaces, according to the work of Ambrosio, Gigli and Savaré. Then comes an independent exposition of the Wasserstein theory, with a short introduction to the optimal transport tools that are needed and to the notion of geodesic convexity, followed by a precise description of the Jordan–Kinderlehrer–Otto scheme and a sketch of proof to obtain its convergence in the easiest cases. A discussion of which equations are gradient flows PDEs and of numerical methods based on these ideas is also provided. The paper ends with a new, theoretical, development, due to Ambrosio, Gigli, Savaré, Kuwada and Ohta: the study of the heat flow in metric measure spaces.

✉ Filippo Santambrogio
  filippo.santambrogio@math.u-psud.fr
  http://www.math.u-psud.fr/~santambr

[1] Laboratoire de Mathématiques D'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay Cedex, France

Contents

1 Introduction

Gradient flows, or *steepest descent curves*, are a very classical topic in evolution equations: take a functional F defined on a vector space X, and, instead of looking at points x minizing F (which is related to the static equation $\nabla F(x) = 0$), we look, given an initial point x_0, for a curve starting at x_0 and trying to minimize F as fast as possible (in this case, we will solve equations of the form $x'(t) = -\nabla F(x(t))$). As we speak of gradients (which are element of X, and not of X' as the differential of F should be), it is natural to impose that X is an Hilbert space (so as to identify it with its dual and produce a gradient vector). In the finite-dimensional case, the above equation is very easy to deal with, but also the infinite-dimensional case is not so exotic. Indeed, just think at the evolution equation $\partial_t u = \Delta u$, which is the evolution variant of the static Laplace equation $-\Delta u = 0$. In this way, the heat equation is the gradient flow, in the L^2 Hilbert space, of the Dirichlet energy $F(u) = \frac{1}{2} \int |\nabla u|^2$, of which $-\Delta u$ is the gradient in the appropriate sense (more generally, one could consider equations of the form $\partial_t u = \delta F / \delta u$, where this notation stands for the first variation of F).

But this is somehow classical… The renovated interest for the notion of gradient flow arrived between the end of the twentieth century and the beginning of the twenty-first century, with the work of Jordan et al. [54] and then of Otto [78], who saw a gradient flow structure in some equations of the form $\partial_t \varrho - \nabla \cdot (\varrho \mathbf{v}) = 0$, where the vector field $\mathbf{v}$ is given by $\mathbf{v} = \nabla[\delta F / \delta \varrho]$. This requires to use the space of probabilities ϱ on a given domain, and to endow it with a non-linear metric structure, derived from the theory of optimal transport. This theory, initiated by Monge [77] in the eighteenth century, then developed by Kantorovich in the '40s [55], is now well-established (many

texts present it, such as [84,90,91]) and is intimately connected with PDEs of the form of the *continuity equation* $\partial_t \varrho - \nabla \cdot (\varrho \mathbf{v}) = 0$.

The turning point for the theory of gradient flows and for the interest that researchers in PDEs developed for it was surely the publication of [3]. This celebrated book established a whole theory on the notion of gradient flow in metric spaces, which requires careful definitions because, in the equation $x'(t) = -\nabla F(x(t))$, neither the term x' nor ∇F make any sense in this framework. For existence and—mainly— uniqueness results, the notion of geodesic convexity (convexity of a functional F defined on a metric space X, when restricted to the geodesic curves of X) plays an important role. Then the theory is particularized in the second half of [3] to the case of the metric space of probability measures endowed with the so-called Wasserstein distance coming from optimal transport, whose differential structure is widely studied in the book. In this framework, the geodesic convexity results that McCann obtained in [72] are crucial to make a bridge from the general to the particular theory.

It is interesting to observe that, finally, the heat equation turns out to be a gradient flow in two different senses: it is the gradient flow of the Dirichlet energy in the L^2 space, but also of the entropy $\int \varrho \log(\varrho)$ in the Wasserstein space. Both frameworks can be adapted from the particular case of probabilities on a domain $\Omega \subset \mathbb{R}^d$ to the more general case of metric measure spaces, and the question whether the two flows coincide, or under which assumptions they do, is natural. It has been recently studied by Ambrosio, Gigli, Savaré and new collaborators (Kuwada and Ohta) in a series of papers [6,48,50], and has been the starting point of recent researches on the differential structure of metric measure spaces.

The present survey, which is an extended, updated, and English version of a Bourbaki seminar given by the author in 2013 ([83]; the reader will also remark that most of the extensions are essentially taken from [84]), aims at giving an overview of the whole theory. In particular, among the goals, there is at the same time to introduce the tools for studying metric gradient flows, but also to see how to deal with Wasserstein gradient flows without such a theory. This could be of interest for applied mathematicians, who could be more involved in the specific PDEs that have this gradient flow form, without a desire for full generality; for the same audience, a section has been added about numerical methods inspired from the so-called JKO (Jordan–Kinderlehrrer– Otto) scheme, and one on a list of equations which fall into these framework. De facto, more than half of the survey is devoted to the theory in the Wasserstein spaces.

The paper is organized as follows: after this introduction, In Sect. 2 we describe the theory in the Euclidean case, and present those which are the good definitions which can be translated into a metric setting; Sect. 3 is devoted to the general metric setting, as in the first half of [3], and is quite expository (only the key ingredients to obtain the proofs are sketched); Sect. 4 is the longest one and contains the Wasserstein setting: after an introduction to optimal transport and to the Wasserstein distances, there in an informal presentation of the equations that can be obtained as gradient flows, a presentation of the different tools to prove convergence of the discrete scheme (called JKO scheme) to a solution of the corresponding equation, a discussion of the functionals which have geodesic convexity properties, a short discussion about other equations and functionals which fit the framework and possible variants, and finally a short section about numerics. Last but not least, in Sect. 5 we give a very short

presentation of the fascinating topic of heat flows in arbitraty metric measure spaces, with reference to the interesting implications that this has in the differential structure of these spaces.

This survey is meant to be suitable for readers with different backgrounds and interests. In particular, the reader who is mainly interested in gradient flows in the Wasserstein space and in PDE applications can decide to skip Sects. 3, part of 4.5 and 5, which deal on the contrary with key objects for the—very lively at the moment—subject of analysis on metric measure spaces.

This is an expository paper where, of course, no new result is presented. Moreover, the presentation is inspired from different sources (several articles and books by the author, in particular [84], even if Sects. 3 and 5 are not detailed there, but also [82,83]); yet, some topics are presented under a different point of view, and/or for the first time. In particular, the reader will find new presentation ingredients in some parts of Sects. 2.1 (some classical observations on gradient flows in $\mathbb{R}^n$, which are presented via a JKO-compatible approach, yet not usually detailed in optimal-transport-oriented texts), 4.5 (the explicit role of geodesic convexity for uniqueness results), and 4.7 (the discussion about numerical methods).

2 From Euclidean to metric

2.1 Gradient flows in the Euclidean space

Before dealing with gradient flows in general metric spaces, the best way to clarify the situation is to start from the easiest case, i.e. what happens in the Euclidean space $\mathbb{R}^n$. Most of what we will say stays true in an arbitrary Hilbert space, but we will stick to the finite-dimensional case for simplicity.

Here, given a function $F : \mathbb{R}^n \to \mathbb{R}$, smooth enough, and a point $x_0 \in \mathbb{R}^n$, a gradient flow is just defined as a curve $x(t)$, with starting point at $t = 0$ given by x_0, which moves by choosing at each instant of time the direction which makes the function F decrease as much as possible. More precisely, we consider the solution of the *Cauchy Problem*

$$\begin{cases} x'(t) = -\nabla F(x(t)) & \text{for } t > 0, \\ x(0) = x_0. \end{cases} \tag{2.1}$$

This is a standard Cauchy problem which has a unique solution if ∇F is Lipschitz continuous, i.e. if $F \in C^{1,1}$. We will see that existence and uniqueness can also hold without this strong assumption, thanks to the variational structure of the equation.

A first interesting property is the following, concerning uniqueness and estimates. We will present it in the case where F is convex, which means that it could be non-differentiable, but we can replace the gradient with the subdifferential. More precisely, we can consider instead of (2.1), the following differential inclusion: we look for an absolutely continuous curve $x : [0, T] \to \mathbb{R}^n$ such that

$$\begin{cases} x'(t) \in -\partial F(x(t)) & \text{for a.e. } t > 0, \\ x(0) = x_0, \end{cases} \tag{2.2}$$

where $\partial F(x) = \{p \in \mathbb{R}^n : F(y) \geq F(x) + p \cdot (y - x)$ for all $y \in \mathbb{R}^n\}$. We refer to [80] for all the definitions and notions from convex analysis that could be needed in the following, and we recall that, if F is differentiable at x, we have $\partial F(x) = \{\nabla F(x)\}$ and that F is differentiable at x if and only if ∂F is a singleton. Also note that $\partial F(x)$ is always a convex set, and is not empty whenever F is real-valued (or x is in the interior of $\{x : F(x) < +\infty\}$), and we denote by $\partial^\circ F(x)$ its element of minimal norm.

Proposition 2.1 *Suppose that F is convex and let x_1 and x_2 be two solutions of* (2.2). *Then $t \mapsto |x_1(t) - x_2(t)|$ is non-increasing. In particular the Cauchy problem* (2.2) *has a unique solution.*

Proof Consider $g(t) = \frac{1}{2}|x_1(t) - x_2(t)|^2$: we have

$$g'(t) = (x_1(t) - x_2(t)) \cdot \left(x_1'(t) - x_2'(t)\right).$$

Using the fact that for every x_1, x_2, p_1, p_2 with $p_i \in \partial F(x_i)$ we have $(x_1 - x_2) \cdot (p_1 - p_2) \geq 0$, we obtain $g'(t) \leq 0$. This proves the first part of the claim. The second is an easy consequence: if we have $x_1(0) = x_2(0)$, then we find $x_1(t) = x_2(t)$ for any $t > 0$. $\qquad\square$

We can also study the case where F is *semi-convex*, i.e. λ-convex for some $\lambda \in \mathbb{R}$. This means that $x \mapsto F(x) - \frac{\lambda}{2}|x|^2$ is convex: for $\lambda > 0$ this is stronger than convexity, and for $\lambda < 0$ it is weaker. Roughly speaking, λ-convexity corresponds to $D^2 F \geq \lambda I$. The reason for the interest in semi-convex functions lies in the fact that on the one hand, as the reader will see throughout the exposition, the general theory of gradient flows applies very well to this class of functions and that, on the other hand, they are general enough to cover many interesting cases. In particular, on a bounded set, all smooth (C^2 is enough) functions are λ-convex for a suitable $\lambda < 0$.

For λ-convex functions, we can define their subdifferential as follows

$$\partial F(x) = \left\{p \in \mathbb{R}^n : F(y) \geq F(x) + p \cdot (y - x) + \frac{\lambda}{2}|y - x|^2 \text{ for all } y \in \mathbb{R}^n\right\}.$$

This definition is consistent with the above one whenever $\lambda \geq 0$ (and guarantees $\partial F(x) \neq \emptyset$ for $\lambda < 0$). Also, one can check that, setting $\tilde{F}(x) = F(x) - \frac{\lambda}{2}|x|^2$, this definition coincides with $\{p \in \mathbb{R}^n : p - \lambda x \in \partial \tilde{F}(x)\}$. Again, we define $\partial^\circ F$ the element of minimal norm of ∂F.

Remark 2.1 From the same proof of Proposition 2.1, one can also deduce uniqueness and stability estimates in the case where F is λ-convex. Indeed, in this case we obtain $|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|e^{-\lambda t}$, which also proves, if $\lambda > 0$, exponential convergence to the unique minimizer of F. The key point is that, if F is λ-convex it is easy to prove that x_1, x_2, p_1, p_2 with $p_i \in \partial F(x_i)$ provides

$$(x_1 - x_2) \cdot (p_1 - p_2) \geq \lambda |x_1 - x_2|^2.$$

This implies $g'(t) \leq -2\lambda g(t)$ and allows to conclude, by Gronwall's lemma, $g(t) \leq g(0)e^{-2\lambda t}$. For the exponential convergence, if $\lambda > 0$ then F is coercive and admits

a minimizer, which is unique by strict convexity. Let us call it $\bar{x}$. Take a solution $x(t)$ and compare it to the constant curve $\bar{x}$, which is a solution since $0 \in \partial F(\bar{x})$. Then we get $|x_1(t) - \bar{x}| \leq e^{-\lambda t}|x_1(0) - \bar{x}|$.

Another well-known fact about the λ-convex case is the fact that the differential inclusion $x'(t) \in -\partial F(x(t))$ actually becomes, a.e., an equality: $x'(t) = -\partial^\circ F(t)$. More precisely, we have the following.

Proposition 2.2 *Suppose that F is λ-convex and let x be a solution of* (2.2). *Then, for all the times t_0 such that both $t \mapsto x(t)$ and $t \mapsto F(x(t))$ are differentiable at $t = t_0$, the subdifferential $\partial F(x(t_0))$ is contained in a hyperplane orthogonal to $x'(t_0)$. In particular, we have $x'(t) = -\partial^\circ F(x(t))$ for a.e. t.*

Proof Let t_0 be as in the statement, and $p \in \partial F(x(t_0))$. From the definition of subdifferential, for every t we have

$$F(x(t)) \geq F(x(t_0)) + p \cdot (x(t) - x(t_0)) + \frac{\lambda}{2}|x(t) - x(t_0)|^2,$$

but this inequality becomes an equality for $t = t_0$. Hence, the quantity

$$F(x(t)) - F(x(t_0)) - p \cdot (x(t) - x(t_0)) - \frac{\lambda}{2}|x(t) - x(t_0)|^2$$

is minimal for $t = t_0$ and, differentiating in t (which is possible by assumption), we get

$$\frac{d}{dt}F(x(t))_{|t=t_0} = p \cdot x'(t_0).$$

Since this is true for every $p \in \partial F(x(t_0))$, this shows that $\partial F(x(t_0))$ is contained in a hyperplane of the form $\{p : p \cdot x'(t_0) = const\}$.

Whenever $x'(t_0)$ belongs to $\partial F(x(t_0))$ (which is true for a.e. t_0), this shows that $x'(t_0)$ is the orthogonal projection of 0 onto $\partial F(x(t_0))$ and onto the hyperplane which contains it, and hence its element of minimal norm. This provides $x'(t_0) = -\partial^\circ F(x(t_0))$ for a.e. t_0, as the differentiability of x and of $F \circ x$ are also true a.e., since x is supposed to be absolutely continuous and F is locally Lipschitz. □

Another interesting feature of those particular Cauchy problems which are gradient flows is their discretization in time. Actually, one can fix a small time step parameter $\tau > 0$ and look for a sequence of points $(x_k^\tau)_k$ defined through the iterated scheme, called *Minimizing Movement Scheme* (see [1,36]),

$$x_{k+1}^\tau \in \operatorname{argmin}_x F(x) + \frac{|x - x_k^\tau|^2}{2\tau}. \tag{2.3}$$

The convexity assumption on F is not necessary for this part of the analysis. It is enough to suppose that F is l.s.c. and satisfies some lower bounds, for instance $F(x) \geq$

$C_1 - C_2|x|^2$, to guarantee that a minimizer exists for small τ. However, it is true that, if F is λ-convex, then these conditions are satisfied, and we also have uniqueness of the minimizer. Indeed, for $\lambda > 0$ we have strict convexity for every τ, and the sum will be strictly convex for small τ if $\lambda < 0$.

The sequence of points x_k^τ must be considered as the values of the curve $x(t)$ at times $t = 0, \tau, \ldots, k\tau, \ldots$ If we look at the optimality conditions of the recursive minimization we obtain

$$x_{k+1}^\tau \in \text{argmin } F(x) + \frac{|x - x_k^\tau|^2}{2\tau} \quad \Rightarrow \quad \nabla F\left(x_{k+1}^\tau\right) + \frac{x_{k+1}^\tau - x_k^\tau}{\tau} = 0.$$

This last condition, written as

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} = -\nabla F\left(x_{k+1}^\tau\right)$$

is exactly the discrete-time *implicit Euler scheme* for $x' = -\nabla F(x)$! (note that in the convex non-smooth case this becomes $\frac{x_{k+1}^\tau - x_k^\tau}{\tau} \in -\partial F(x_{k+1}^\tau)$).

We recall that, given an ODE $x'(t) = \mathbf{v}(x(t))$ (that we take autonomous for simplicity), with given initial datum $x(0) = x_0$, Euler schemes are time-discretizations where derivatives are replaced by finite differences. We fix a time step $\tau > 0$ and define a sequence x_k^τ. The explicit scheme is given by

$$x_{k+1}^\tau = x_k^\tau + \tau \mathbf{v}\left(x_k^\tau\right), \quad x_0^\tau = x_0,$$

while the implicit scheme is given by

$$x_{k+1}^\tau = x_k^\tau + \tau \mathbf{v}\left(x_{k+1}^\tau\right), \quad x_0^\tau = x_0.$$

This means that x_{k+1}^τ is selected as a solution of an equation involving x_k^τ, instead of being explicitly computable from x_k^τ. The explicit scheme is obviously easier to implement, but enjoys less stability and qualitative properties than the implicit one. Suppose for instance $\mathbf{v} = -\nabla F$: then the quantity $F(x(t))$ decreases in t in the continuous solution, which is also the case for the implicit scheme, but not for the explicit one (which represents the iteration of the gradient method for the minimization of F). Note that the same can be done for evolution PDEs, and that solving the heat equation $\partial_t \varrho = \Delta \varrho_t$ by an explicit scheme is very dangerous: at every step, ϱ_{k+1}^τ would have two degrees of regularity less than ϱ_k^τ, since it would be obtained through $\varrho_{k+1}^\tau = \varrho_k^\tau - \tau \Delta \varrho_k^\tau$.

It is possible to prove that, for $\tau \to 0$, the sequence we found, suitably interpolated, converges to the solution of Problem (2.2). We give here below the details of this argument, as it will be the basis of the argument that we will use in Sect. 4.

First, we define two different interpolations of the points x_k^τ. Let us define two curves $x^\tau, \tilde{x}^\tau : [0, T] \to \mathbb{R}^n$ as follows: first we define

$$\mathbf{v}_{k+1}^\tau := \frac{x_{k+1}^\tau - x_k^\tau}{\tau},$$

then we set

$$x^\tau(t) = x_{k+1}^\tau \qquad \tilde{x}^\tau(t) = x_k^\tau + (t - k\tau)\mathbf{v}_{k+1}^\tau \text{ for } t \in]k\tau, (k+1)\tau].$$

Also set

$$\mathbf{v}^\tau(t) = \mathbf{v}_{k+1}^\tau \text{ for } t \in]k\tau, (k+1)\tau].$$

It is easy to see that $\tilde{x}^\tau$ is a continuous curve, piecewise affine (hence absolutely continuous), satisfying $(\tilde{x}^\tau)' = \mathbf{v}^\tau$. On the contrary, x^τ is not continuous, but satisfies by construction $\mathbf{v}^\tau(t) \in -\partial F(x^\tau(t))$.

The iterated minimization scheme defining x_{k+1}^τ provides the estimate

$$F\left(x_{k+1}^\tau\right) + \frac{|x_{k+1}^\tau - x_k^\tau|^2}{2\tau} \leq F\left(x_k^\tau\right), \tag{2.4}$$

obtained comparing the optimal point x_{k+1}^τ to the previous one. If $F(x_0) < +\infty$ and $\inf F > -\infty$, summing over k we get

$$\sum_{k=0}^{\ell} \frac{|x_{k+1}^\tau - x_k^\tau|^2}{2\tau} \leq \left(F\left(x_0^\tau\right) - F\left(x_{\ell+1}^\tau\right)\right) \leq C. \tag{2.5}$$

This is valid for every ℓ, and we can arrive up to $\ell = \lfloor T/\tau \rfloor$. Now, note that

$$\frac{|x_{k+1}^\tau - x_k^\tau|^2}{2\tau} = \tau \left(\frac{|x_{k+1}^\tau - x_k^\tau|}{2\tau}\right)^2 = \tau|\mathbf{v}_k^\tau|^2 = \int_{k\tau}^{(k+1)\tau} |(\tilde{x}^\tau)'(t)|^2 dt.$$

This means that we have

$$\int_0^T \frac{1}{2}|(\tilde{x}^\tau)'(t)|^2 dt \leq C \tag{2.6}$$

and hence $\tilde{x}^\tau$ is bounded in H^1 and $\mathbf{v}^\tau$ in L^2. The injection $H^1 \subset C^{0,1/2}$ provides an equicontinuity bound on $\tilde{x}^\tau$ of the form

$$|\tilde{x}^\tau(t) - \tilde{x}^\tau(s)| \leq C|t - s|^{1/2}. \tag{2.7}$$

This also implies

$$|\tilde{x}^\tau(t) - x^\tau(t)| \leq C\tau^{1/2}, \tag{2.8}$$

since $x^\tau(t) = \tilde{x}^\tau(s)$ for a certain $s = k\tau$ with $|s - t| \leq \tau$. Finally, we have proven the necessary compactness to obtain the following:

Proposition 2.3 *Let $\tilde{x}^\tau$, x^τ and $\mathbf{v}^\tau$ be constructed as above using the minimizing movement scheme. Suppose $F(x_0) < +\infty$ and $\inf F > -\infty$. Then, up to a subsequence $\tau_j \to 0$ (still denoted by τ), both $\tilde{x}^\tau$ and x^τ converge uniformly to a same curve $x \in H^1$, and $\mathbf{v}^\tau$ weakly converges in L^2 to a vector function $\mathbf{v}$, such that $x' = \mathbf{v}$ and*

1. *if F is λ-convex, we have $\mathbf{v}(t) \in -\partial F(x(t))$ for a.e. t, i.e. x is a solution of* (2.2);
2. *if F is C^1, we have $\mathbf{v}(t) = -\nabla F(x(t))$ for all t, i.e. x is a solution of* (2.1).

Proof Thanks to the estimates (2.6) and (2.7) and the fact that the initial point $\tilde{x}^\tau(0)$ is fixed, we can apply Ascoli–Arzelà's theorem to $\tilde{x}^\tau$ and get a uniformly converging subsequence. The estimate (2.8) implies that, on the same subsequence, x^τ also converges uniformly to the same limit, that we will call $x = [0, T] \to \mathbb{R}^n$. Then, $\mathbf{v}^\tau = (\tilde{x}^\tau)'$ and (2.6) allows to guarantee, up to an extra subsequence extraction, the weak convergence $\mathbf{v}^\tau \rightharpoonup \mathbf{v}$ in L^2. The condition $x' = \mathbf{v}$ is automatical as a consequence of distributional convergence.

To prove 1), we will fix a point $y \in \mathbb{R}^n$ and write

$$F(y) \geq F\left(x^\tau(t)\right) + \mathbf{v}^\tau(t) \cdot \left(y - x^\tau(t)\right) + \frac{\lambda}{2}|y - x^\tau(t)|^2.$$

We then multiply by a positive measurable function $a : [0, T] \to \mathbb{R}_+$ and integrate:

$$\int_0^T a(t) \left(F(y) - F\left(x^\tau(t)\right) - \mathbf{v}^\tau(t) \cdot \left(y - x^\tau(t)\right) - \frac{\lambda}{2}|y - x^\tau(t)|^2 \right) dt \geq 0.$$

We can pass to the limit as $\tau \to 0$, using the uniform (hence L^2 strong) convergence $x^\tau \to x$ and the weak convergence $\mathbf{v}^\tau \rightharpoonup \mathbf{v}$. In terms of F, we just need its lower semi-continuity. This provides

$$\int_0^T a(t) \left(F(y) - F(x(t)) - \mathbf{v}(t) \cdot (y - x(t)) - \frac{\lambda}{2}|y - x(t)|^2 \right) dt \geq 0.$$

From the arbitrariness of a, the inequality

$$F(y) \geq F(x(t)) + \mathbf{v}(t) \cdot (y - x(t)) + \frac{\lambda}{2}|y - x(t)|^2$$

is true for a.e. t (for fixed y). Using y in a dense countable set in the interior of $\{F < +\infty\}$ (where F is continuous), we get $\mathbf{v}(t) \in \partial F(x(t))$.

To prove 2), the situation is easier. Indeed we have

$$-\nabla F\left(x^\tau\right) = \mathbf{v}^\tau(t) = (\tilde{x}^\tau)'(t).$$

The first term in the equality uniformly converges, as a function of t, (since ∇F is continuous and x^τ lives in a compact set) to $-\nabla F(x)$, the second weakly converges to $\mathbf{v}$ and the third to x'. This proves the claim, and the equality is now true for every t as the function $t \mapsto -\nabla F(x(t))$ is uniformly continuous. □

In the above result, we only proved convergence of the curves x^τ to the limit curve x, solution of $x' = -\nabla F(x)$ (or $-x' \in \partial F(x)$), but we gave no quantitative order of convergence, and we will not study such an issue in the rest of the survey neither. On the contrary, the book [3] which will be the basis for the metric case, also provides

explicit estimates; these estimates are usually of order τ. An interesting observation, in the Euclidean case, is that if the sequence x_k^τ is defined by

$$x_{k+1}^\tau \in \mathrm{argmin}_x \quad 2F\left(\frac{x + x_k^\tau}{2}\right) + \frac{|x - x_k^\tau|^2}{2\tau},$$

then we have

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} = -\nabla F\left(\frac{x + x_k^\tau}{2}\right),$$

and the convergence is of order τ^2. This has been used in the Wasserstein case[1] (see Sect. 4 and in particular Sect. 4.7) in [61].

2.2 An introduction to the metric setting

The iterated minimization scheme that we introduced above has another interesting feature: it even suggests how to define solutions for functions F which are only l.s.c., with no gradient at all!

Even more, this discretized formulation can easily be adapted to metric spaces. Indeed, given a metric space (X, d) and a l.s.c. function $F : X \to \mathbb{R} \cup \{+\infty\}$ (under suitable compactness assumptions to guarantee existence of the minimum), one can define

$$x_{k+1}^\tau \in \mathrm{argmin}_x F(x) + \frac{d(x, x_k^\tau)^2}{2\tau} \tag{2.9}$$

and study the limit as $\tau \to 0$. Then, we use the piecewise constant interpolation

$$x^\tau(t) := x_k^\tau \quad \text{for every } t \in](k-1)\tau, k\tau] \tag{2.10}$$

and study the limit of x^τ as $\tau \to 0$.

De Giorgi, in [36], defined what he called *Generalized Minimizing Movements*[2]:

Definition 2.1 A curve $x : [0, T] \to X$ is called Generalized Minimizing Movements (GMM) if there exists a sequence of time steps $\tau_j \to 0$ such that the sequence of curves x^{τ_j} defined in (2.10) using the iterated solutions of (2.9) uniformly converges to x in $[0, T]$.

The compactness results in the space of curves guaranteeing the existence of GMM are also a consequence of an Hölder estimate that we already saw in the Euclidean

---

[1] The attentive reader can observe that, setting $y := (x + x_k^\tau)/2$, this minimization problem becomes $\min_y 2F(y) + 2|y - x_k^\tau|^2/\tau$. Yet, when acting on a metric space, or simply on a manifold or a bounded domain, there is an extra constraint on y: the point y must be the middle point of a geodesic between x_k^τ and a point x (on a sphere, for instance, this means that if x_k^τ is the North Pole, then y must lie in the northern hemisphere).

[2] We prefer not to make any distinction here between generalized minimizing movements and minimizing movements.

case. Yet, in the general case, some arrangements are needed, as we cannot use the piecewise affine interpolation. We will see later that, in case the segments may be replaced by geodesics, a similar estimate can be obtained. However, we can also obtain a Hölder-like estimate from the piecewise constant interpolation.

We start from

$$F\left(x_{k+1}^{\tau}\right) + \frac{d\left(x_{k+1}^{\tau}, x_k^{\tau}\right)^2}{2\tau} \leq F\left(x_k^{\tau}\right), \tag{2.11}$$

and

$$\sum_{k=0}^{l} d\left(x_{k+1}^{\tau}, x_k^{\tau}\right)^2 \leq 2\tau\left(F\left(x_0^{\tau}\right) - F\left(x_{l+1}^{\tau}\right)\right) \leq C\tau.$$

The Cauchy–Schwartz inequality gives, for $t < s, t \in [k\tau, (k+1)\tau[$ and $s \in [l\tau, (l+1)\tau[$ (hence $|l - k| \leq \frac{|t-s|}{\tau} + 1$),

$$
\begin{aligned}
d\left(x^{\tau}(t), x^{\tau}(s)\right) &\leq \sum_{k=0}^{l} d\left(x_{k+1}^{\tau}, x_k^{\tau}\right) \leq \left(\sum_{k=0}^{l} d\left(x_{k+1}^{\tau}, x_k^{\tau}\right)^2\right)^{1/2}\left(\frac{|t-s|}{\tau}+1\right)^{1/2} \\
&\leq C\left(|t-s|^{1/2}+\sqrt{\tau}\right).
\end{aligned}
$$

This provides a Hölder bound on the curves x^{τ} (with a negligible error of order $\sqrt{\tau}$ which disappears at the limit $\tau \to 0$), and allows to extract a converging subsequence.

Anyway, if we add some structure to the metric space (X, d), a more similar analysis to the Euclidean case can be performed. This is what happens when we suppose that (X, d) is a *geodesic space*. This requires a short discussion about curves and geodesics in metric spaces.

Curves and geodesics in metric spaces We recall that a curve ω is a continuous function defined on an interval, say $[0, 1]$, and valued in a metric space (X, d). As it is a map between metric spaces, it is meaningful to say whether it is Lipschitz or not, but its velocity $\omega'(t)$ has no meaning, unless X is a vector space. Surprisingly, it is possible to give a meaning to the modulus of the velocity (i.e. the speed) $|\omega'|(t)$.

Definition 2.2 If $\omega : [0, 1] \to X$ is a curve valued in the metric space (X, d) we define the metric derivative of ω at time t, denoted by $|\omega'|(t)$ through

$$|\omega'|(t) := \lim_{h \to 0} \frac{d(\omega(t+h), \omega(t))}{|h|},$$

provided this limit exists.

In the spirit of Rademacher Theorem, it is possible to prove (see [9]) that, if $\omega : [0, 1] \to X$ is Lipschitz continuous, then the metric derivative $|\omega'|(t)$ exists for a.e. t. Moreover we have, for $t_0 < t_1$,

$$d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} |\omega'|(s)\, \mathrm{d}s.$$

The same is also true for more general curves, not only Lipschitz continuous.

Definition 2.3 A curve $\omega : [0, 1] \to X$ is said to be *absolutely continuous* whenever there exists $g \in L^1([0, 1])$ such that $d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(s)\, \mathrm{d}s$ for every $t_0 < t_1$. The set of absolutely continuous curves defined on $[0, 1]$ and valued in X is denoted by $\mathrm{AC}(X)$.

It is well-known that every absolutely continuous curve can be reparametrized in time (through a monotone-increasing reparametrization) and become Lipschitz continuous. The existence of the metric derivative for a.e. t is also true for $\omega \in \mathrm{AC}(X)$, via this reparametrization.

Given a continuous curve, we can also define its length, and the notion of geodesic curves.

Definition 2.4 For a curve $\omega : [0, 1] \to X$, let us define

$$\mathrm{Length}(\omega) := \sup \left\{ \sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1})) \ : \ n \geq 1, \ 0 = t_0 < t_1 < \cdots < t_n = 1 \right\}.$$

It is easy to see that all curves $\omega \in \mathrm{AC}(X)$ satisfy $\mathrm{Length}(\omega) \leq \int_0^1 g(t)\mathrm{d}t < +\infty$. Also, we can prove that, for any curve $\omega \in \mathrm{AC}(X)$, we have

$$\mathrm{Length}(\omega) = \int_0^1 |\omega'|(t)\mathrm{d}t.$$

We collect now some more definitions.

Definition 2.5 A curve $\omega : [0, 1] \to X$ is said to be a geodesic between x_0 and $x_1 \in X$ if $\omega(0) = x_0$, $\omega(1) = x_1$ and $\mathrm{Length}(\omega) = \min\{\mathrm{Length}(\tilde{\omega}) \ : \ \tilde{\omega}(0) = x_0, \ \tilde{\omega}(1) = x_1\}$.

A space (X, d) is said to be a *length space* if for every x and y we have

$$d(x, y) = \inf\{\mathrm{Length}(\omega) \ : \ \omega \in \mathrm{AC}(X), \ \omega(0) = x, \ \omega(1) = y\}.$$

A space (X, d) is said to be a *geodesic space* if for every x and y we have

$$d(x, y) = \min\{\mathrm{Length}(\omega) \ : \ \omega \in \mathrm{AC}(X), \ \omega(0) = x, \ \omega(1) = y\},$$

i.e. if it is a length space and there exist geodesics between arbitrary points.

In a length space, a curve $\omega : [t_0, t_1] \to X$ is said to be a *constant-speed geodesic* between $\omega(0)$ and $\omega(1) \in X$ if it satisfies

$$d(\omega(t), \omega(s)) = \frac{|t - s|}{t_1 - t_0} d(\omega(t_0), \omega(t_1)) \quad \text{for all } t, s \in [t_0, t_1].$$

It is easy to check that a curve with this property is automatically a geodesic, and that the following three facts are equivalent (for arbitrary $p > 1$)

1. ω is a constant-speed geodesic defined on $[t_0, t_1]$ and joining x_0 and x_1,
2. $\omega \in AC(X)$ and $|\omega'|(t) = \frac{d(\omega(t_0),\omega(t_1))}{t_1-t_0}$ a.e.,
3. ω solves $\min \left\{ \int_{t_0}^{t_1} |\omega'|(t)^p dt \; : \; \omega(t_0) = x_0, \omega(t_1) = x_1 \right\}$.

We can now come back to the interpolation of the points obtained through the Minimizing Movement scheme (2.9) and note that, if (X, d) is a geodesic space, then the piecewise affine interpolation that we used in the Euclidean space may be helpfully replaced via a piecewise geodesic interpolation. This means defining a curve $x^\tau : [0, T] \to X$ such that $x^\tau(k\tau) = x_k^\tau$ and such that x^τ restricted to any interval $[k\tau, (k+1)\tau]$ is a constant-speed geodesic with speed equal to $d(x_k^\tau, x_{k+1}^\tau)/\tau$. Then, the same computations as in the Euclidean case allow to prove an H^1 bound on the curves x^τ (i.e. an L^2 bound on the metric derivatives $|(x^\tau)'|$) and prove equicontinuity.

The next question is how to characterize the limit curve obtained when $\tau \to 0$, and in particular how to express the fact that it is a gradient flow of the function F. Of course, one cannot try to prove the equality $x' = -\nabla F(x)$, just because neither the left-hand side nor the right-hand side have a meaning in a metric space!

In some particular cases, when the space X, the distance d, and the functional F have some extra structures, it is possible to pass to the limit the optimality conditions of each optimization problem in the discretized scheme, and characterize the limit curves (or the limit curve) $x(t)$. This will be the case in the framework of probability measures, as it will be discussed in Sect. 4, but not in general. Indeed, having some sort of differential structure on the space X is necessary in order to do so. Hence, if we want to develop a general theory for gradient flows in metric spaces, finer tools are needed. In particular, we need to characterize the solutions of $x' = -\nabla F(x)$ (or $x' \in -\partial F(x)$) by only using metric quantities (in particular, avoiding derivatives, gradients, and more generally vectors). The book by Ambrosio–Gigli–Savaré [3], and in particular its first part (the second being devoted to the space of probability measures) exactly aims at doing so.

Hence, what we do here is to present alternative characterizations of gradient flows in the smooth Euclidean case, which can be used as a definition of gradient flow in the metric case, since all the quantities which are involved have their metric counterpart.

The first observation is the following: thanks to the Young inequality, for every curve we have

$$F(x(s))-F(x(t))= \int_s^t -\nabla F(x(r)) \cdot x'(r) \, dr \leq \int_s^t \left(\frac{1}{2}|x'(r)|^2 + \frac{1}{2}|\nabla F(x(r))|^2 \right) dr.$$

Here, this inequality is an equality if and only if $x'(r) = -\nabla F(x(r))$ for a.e. r. Hence, the condition, called EDE (*Energy Dissipation Equality*)

$$F(x(s)) - F(x(t)) = \int_s^t \left(\frac{1}{2}|x'(r)|^2 + \frac{1}{2}|\nabla F(x(r))|^2 \right) dr, \quad \text{for all } s < t$$

(or even the simple inequality $F(x(s))-F(x(t)) \geq \int_s^t \left(\frac{1}{2}|x'(r)|^2 + \frac{1}{2}|\nabla F(x(r))|^2 \right) dr$) is equivalent to $x' = -\nabla F(x)$ a.e., and could be taken as a definition of gradient flow.

In the general theory of Gradient Flows [3] in metric spaces, another characterization, different from the EDE, is proposed in order to cope with uniqueness and stability results. It is based on the following observation: if $F : \mathbb{R}^d \to \mathbb{R}$ is convex, then the inequality

$$F(y) \geq F(x) + p \cdot (y - x) \quad \text{for all } y \in \mathbb{R}^d$$

characterizes (by definition) the vectors $p \in \partial F(x)$ and, if $F \in C^1$, it is only satisfied for $p = \nabla F(x)$. Analogously, if F is λ-convex, the inequality that characterizes the gradient is

$$F(y) \geq F(x) + \frac{\lambda}{2}|x - y|^2 + p \cdot (y - x) \quad \text{for all } y \in \mathbb{R}^d.$$

Hence, we can pick a curve $x(t)$ and a point y and compute

$$\frac{d}{dt}\frac{1}{2}|x(t) - y|^2 = (y - x(t)) \cdot (-x'(t)).$$

Consequently, imposing

$$\frac{d}{dt}\frac{1}{2}|x(t) - y|^2 \leq F(y) - F(x(t)) - \frac{\lambda}{2}|x(t) - y|^2,$$

for all y, will be equivalent to $-x'(t) \in -\partial F(x(t))$. This will provide a second characterization (called EVI, *Evolution Variational Inequality*) of gradient flows in a metric environment. Indeed, all the terms appearing in the above inequality have a metric counterpart (only squared distances and derivatives w.r.t. time appear). Even if we often forget the dependance on λ, it should be noted that the condition EVI should actually be written as EVI_λ, since it involves a parameter λ, which is a priori arbitrary. Actually, λ-convexity of F is not necessary to define the EVI_λ property, but it will be necessary in order to guarantee the existence of curves which satisfy such a condition. The notion of λ-convexity will hence be crucial also in metric spaces, where it will be rather "λ-geodesic-convexity".

The role of the EVI condition in the uniqueness and stability of gradient flows is quite easy to guess. Take two curves, that we call $x(t)$ and $y(s)$, and compute

$$\frac{d}{dt}\frac{1}{2}d(x(t), y(s))^2 \leq F(y(s)) - F(x(t)) - \frac{\lambda}{2}d(x(t), y(s))^2, \qquad (2.12)$$

$$\frac{d}{ds}\frac{1}{2}d(x(t), y(s))^2 \leq F(x(t)) - F(y(s)) - \frac{\lambda}{2}d(x(t), y(s))^2. \qquad (2.13)$$

If one wants to estimate $E(t) = \frac{1}{2}d(x(t), y(t))^2$, summing up the two above inequalities, after a chain-rule argument for the composition of the function of two variables $(t, s) \mapsto \frac{1}{2}d(x(t), y(s))^2$ and of the curve $t \mapsto (t, t)$, gives

$$\frac{d}{dt}E(t) \leq -2\lambda E(t).$$

By Gronwall Lemma, this provides uniqueness (when $x(0) = y(0)$) and stability.

3 The general theory in metric spaces

3.1 Preliminaries

In order to sketch the general theory in metric spaces, first we need to give (or recall) general definitions for the three main objects that we need in the EDE and EVI properties, characterizing gradient flows: the notion of speed of a curve, that of slope of a function (somehow the modulus of its gradient) and that of (geodesic) convexity.

Metric derivative We already introduced in the previous section the notion of metric derivative: given a curve $x : [0, T] \to X$ valued in a metric space, we can define, instead of the velocity $x'(t)$ as a vector (i.e., with its direction, as we would do in a vector space), the speed (i.e. the modulus, or norm, of $x'(t)$) as follows:

$$|x'|(t) := \lim_{h \to 0} \frac{d(x(t), x(t + h))}{|h|},$$

provided the limit exists. This is the notion of speed that we will use in metric spaces.

Slope and modulus of the gradient Many definitions of the modulus of the gradient of a function F defined over a metric space are possible. First, we call *upper gradient* of F every function $g : X \to \mathbb{R}$ such that, for every Lipschitz curve x, we have

$$|F(x(0)) - F(x(1))| \leq \int_0^1 g(x(t))|x'|(t)\mathrm{d}t.$$

If F is Lipschitz continuous, a possible choice is the *local Lipschitz constant*

$$|\nabla F|(x) := \limsup_{y \to x} \frac{|F(x) - F(y)|}{d(x, y)}; \tag{3.1}$$

another is the *descending slope* (we will often say just *slope*), which is a notion more adapted to the minimization of a function than to its maximization, and hence reasonable for lower semi-continuous functions:

$$|\nabla^- F|(x) := \limsup_{y \to x} \frac{[F(x) - F(y)]_+}{d(x, y)}$$

(note that the slope vanishes at every local minimum point). In general, it is not true that the slope is an upper gradient, but we will give conditions to guarantee that it is. Later on (Sect. 5) we will see how to define a Sobolev space H^1 on a metric (measure) space, by using suitable relaxations of the modulus of the gradient of F.

Geodesic convexity The third notion to be dealt with is that of convexity. This can only be done in a geodesic metric space. On such a space, we can say that a function is *geodesically convex* whenever it is convex along geodesics. More precisely, we

require that for every pair $(x(0), x(1))$ there exists[3] a geodesic x with constant speed connecting these two points and such that

$$F(x(t)) \leq (1 - t)F(x(0)) + tF(x(1)).$$

We can also define λ-convex functions as those which satisfy a modified version of the above inequality:

$$F(x(t)) \leq (1 - t)F(x(0)) + tF(x(1)) - \lambda \frac{t(1 - t)}{2}d^2(x(0), x(1)). \quad (3.2)$$

3.2 Existence of a gradient flow

Once fixed these basic ingredients, we can now move on to the notion of gradient flow. A starting approach is, again, the sequential minimization along a discrete scheme, for a fixed time step $\tau > 0$, and then pass to the limit. First, we would like to see in which framework this procedure is well-posed. Let us suppose that the space X and the function F are such that every sub-level set $\{F \leq c\}$ is compact in X, either for the topology induced by the distance d, or for a weaker topology, such that d is lower semi-continuous w.r.t. it; F is required to be l.s.c. in the same topology. This is the minimal framework to guarantee existence of the minimizers at each step, and to get estimates as in (2.11) providing the existence of a limit curve. It is by the way a quite general situation, as we can see in the case where X is a reflexive Banach space and the distance d is the one induced by the norm: in this case there is no need to restrict to the (very severe) assumption that F is strongly l.s.c., but the weak topology allows to deal with a much wider situation.

We can easily understand that, even if the estimate (2.11) is enough to provide compactness, and thus the existence of a GMM, it will never be enough to characterize the limit curve (indeed, it is satisfied by any discrete evolution where x_{k+1}^τ gives a better value than x_k^τ, without any need for optimality). Hence, we will never obtain either of the two formulations—EDE or EVI—of metric gradient flows.

In order to improve the result, we should exploit how much x_{k+1}^τ is better than x_k^τ. An idea due to De Giorgi allows to obtain the desired result, via a "variational interpolation" between the points x_k^τ and x_{k+1}^τ. In order to do so, once we fix x_k^τ, for every $\theta \in]0, 1]$, we consider the problem

$$\min_x \quad F(x) + \frac{d^2(x, x_k^\tau)}{2\theta\tau}$$

and call $x(\theta)$ any minimizer for this problem, and $\varphi(\theta)$ the minimal value. It is clear that, for $\theta \to 0^+$, we have $x(\theta) \to x_k^\tau$ and $\varphi(\theta) \to F(x_k^\tau)$, and that, for $\theta = 1$,

---

[3] Warning: this definition is not equivalent to true convexity along the geodesic, since we only compare intermediate instants t to 0 and 1, and not to other intermediate instants; however, in case of uniqueness of geodesics, or if we required the same condition to be true for all geodesics, then we would recover the same condition. Also, let us note that we will only need the existence of geodesics connecting pairs of points where $F < +\infty$.

we get back to the original problem with minimizer x_{k+1}^τ. Moreover, the function φ is non-increasing and hence a.e. differentiable (actually, we can even prove that it is locally semiconcave). Its derivative $\varphi'(\theta)$ is given by the derivative of the function $\theta \mapsto F(x) + \frac{d^2(x, x_k^\tau)}{2\theta\tau}$, computed at the optimal point $x = x(\theta)$ (the existence of $\varphi'(\theta)$ implies that this derivative is the same at every minimal point $x(\theta)$). Hence we have

$$\varphi'(\theta) = -\frac{d^2(x(\theta), x_k^\tau)}{2\theta^2\tau},$$

which means, by the way, that $d(x(\theta), x_k^\tau)^2$ does not depend on the minimizer $x(\theta)$ for all θ such that $\varphi'(\theta)$ exists. Moreover, the optimality conditions for the minimization problem with $\theta > 0$ easily show that

$$|\nabla^- F|(x(\theta)) \leq \frac{d(x(\theta), x_k^\tau)}{\theta\tau}.$$

This can be seen if we consider the minimization of an arbitrary function $x \mapsto F(x) + cd^2(x, \bar{x})$, for fixed $c > 0$ and $\bar{x}$, and we consider a competitor y. If x is optimal we have $F(y) + cd^2(y, \bar{x}) \geq F(x) + cd^2(x, \bar{x})$, which implies

$$F(x) - F(y) \leq c\left(d^2(y, \bar{x}) - d^2(x, \bar{x})\right) = c\left(d(y, \bar{x}) + d(x, \bar{x})\right)\left(d(y, \bar{x}) - d(x, \bar{x})\right)$$
$$\leq c\left(d(y, \bar{x}) + d(x, \bar{x})\right)d(y, x).$$

We divide by $d(y, x)$, take the positive part and then the lim sup as $y \to x$, and we get $|\nabla^- F|(x) \leq 2cd(x, \bar{x})$.

We now come back to the function φ and use

$$\varphi(0) - \varphi(1) \geq -\int_0^1 \varphi'(\theta)\, d\theta$$

(the inequality is due to the possible singular part of the derivative for monotone functions; actually, we can prove that it is an equality by using the local semiconcave behavior, but this is not needed in the following), together with the inequality

$$-\varphi'(\theta) = \frac{d(x(\theta), x_k^\tau)^2}{2\theta^2\tau} \geq \frac{\tau}{2}|\nabla^- F(x(\theta))|^2$$

that we just proved. Hence, we get an improved version of (2.11):

$$F\left(x_{k+1}^\tau\right) + \frac{d\left(x_{k+1}^\tau, x_k^\tau\right)^2}{2\tau} \leq F\left(x_k^\tau\right) - \frac{\tau}{2}\int_0^1 |\nabla^- F(x(\theta))|^2 d\theta.$$

If we sum up for $k = 0, 1, 2, \ldots$ and then take the limit $\tau \to 0$, we can prove, for every GMM x, the inequality

$$F(x(t)) + \frac{1}{2} \int_0^t |x'|(r)^2 \mathrm{d}r + \frac{1}{2} \int_0^t |\nabla^- F(x(r))|^2 \mathrm{d}r \leq F(x(0)), \qquad (3.3)$$

under some suitable assumptions that we must select. In particular, we need lower-semicontinuity of F in order to handle the term $F(x_{k+1}^\tau)$ (which will become $F(x(t))$ at the limit), but we also need lower-semicontinuity of the slope $|\nabla^- F|$ in order to handle the corresponding term.

This inequality does not exactly correspond to EDE: on the one hand we have an inequality, and on the other we just compare instants t and 0 instead of t and s. If we want equality for every pair (t, s), we need to require the slope to be an upper gradient. Indeed, in this case, we have the inequality $F(x(0)) - F(x(t)) \leq \int_0^t |\nabla^- F(x(r))||x'|(r)\mathrm{d}r$ and, starting from the usual inequalities, we find that (3.3) is actually an equality. This allows to subtract the equalities for s and t, and get, for $s < t$:

$$F(x(t)) + \frac{1}{2} \int_s^t |x'|(r)^2 \mathrm{d}r + \frac{1}{2} \int_s^t |\nabla^- F(x(r))|^2 \mathrm{d}r = F(x(s)).$$

Magically, it happens that the assumption that F is λ-geodesically convex simplifies everything. Indeed, we have two good points: the slope is automatically l.s.c., and it is automatically an upper gradient. These results are proven in [2,3]. We just give here the main idea to prove both. This idea is based on a pointwise representation of the slope as a sup instead of a lim sup: if F is λ-geodesically convex, then we have

$$|\nabla^- F|(x) = \sup_{y \neq x} \left[\frac{F(x) - F(y)}{d(x, y)} + \frac{\lambda}{2} d(x, y) \right]_+. \qquad (3.4)$$

In order to check this, we just need to add a term $\frac{\lambda}{2} d(x, y)$ inside the positive part of the definition of $|\nabla^- F|(x)$, which does not affect the limit as $y \to x$ and shows that $|\nabla^- F|(x)$ is smaller than this sup. The opposite inequality is proven by fixing a point y, connecting it to x through a geodesic $x(t)$, and computing the limit along this curve.

This representation as a sup allows to prove semicontinuity of the slope.[4] It is also possible (see [2], for instance) to prove that the slope is an upper gradient.

Let us insist anyway on the fact that the λ-convexity assumption is not natural nor crucial to prove the existence of a gradient flow. On the one hand, functions smooth enough could satisfy the assumptions on the semi-continuity of F and of $|\nabla^- F|$ and the fact that $|\nabla^- F|$ is an upper gradient independently of convexity; on the other hand the discrete scheme already provides a method, well-posed under much weaker assumptions, to find a limit curve. If the space and the functional allow for it (as it

---

[4] Warning: we get here semi-continuity w.r.t. the topology induced by the distance d, which only allows to handle the case where the set $\{F \leq c\}$ are d-compacts.

will be the case in the next section), we can hope to characterize this limit curve as the solution of an equation (it will be a PDE in Sect. 4), without passing through the general theory and the EDE condition.

3.3 Uniqueness and contractivity

On the contrary, if we think at the uniqueness proof that we gave in the Euclidean case, it seems that some sort of convexity should be the good assumption in order to prove uniqueness. Here we will only give the main lines of the uniqueness theory in the metric framework: the key point is to use the EVI condition instead of the EDE.

The situation concerning these two different notions of gradient flows (EVI and EDE) in abstract metric spaces has been clarified by Savaré (in an unpublished note, but the proof can also be found in [2]), who showed that

- All curves which are gradient flows in the EVI sense also satisfy the EDE condition.
- The EDE condition is not in general enough to guarantee uniqueness of the gradient flow. A simple example: take $X = \mathbb{R}^2$ with the ℓ^∞ distance

$$d((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| \vee |x_2 - y_2|,$$

  and take $F(x_1, x_2) = x_1$; we can check that any curve $(x_1(t), x_2(t))$ with $x_1'(t) = -1$ and $|x_2'(t)| \leq 1$ satisfies EDE.
- On the other hand, existence of a gradient flow in the EDE sense is quite easy to get, and provable under very mild assumption, as we sketched in Sect. 3.2.
- The EVI condition is in general too strong in order to get existence (in the example above of the ℓ^∞ norm, no EVI gradient flow would exist), but always guarantees uniqueness and stability (w.r.t. initial data).

Also, the existence of EVI gradient flows is itself very restricting on the function F: indeed, it is proven in [35] that, if F is such that from every starting point x_0 there exists an EVI$_\lambda$ gradient flow, then F is necessarily λ-geodesically-convex.

We provide here an idea of the proof of the contractivity (and hence of the uniqueness) of the EVI gradient flows.

Proposition 3.1 *If two curves $x, y : [0, T] \rightarrow X$ satisfy the EVI condition, then we have*

$$\frac{d}{dt} d(x(t), y(t))^2 \leq -2\lambda d(x(t), y(t))^2$$

and $d(x(t), y(t)) \leq e^{-\lambda t} d(x(0), y(0))$.

The second part of the statement is an easy consequence of the first one, by Gronwall Lemma. The first is (formally) obtained by differentiating $t \mapsto d(x(t), y(t_0))^2$ at $t = t_0$, then $s \mapsto d(x(t_0), y(s))^2$ at $s = t_0$. The EVI condition allows to write

$$\frac{d}{dt}d(x(t), y(t_0))^2_{|t=t_0} \le -\lambda d(x(t_0), y(t_0))^2 + 2F(y(t_0)) - 2F(x(t_0))$$

$$\frac{d}{ds}d(x(t_0), y(s))^2_{|s=t_0} \le -\lambda d(x(t_0), y(t_0))^2 + 2F(x(t_0)) - 2F(y(t_0))$$

and hence, summing up, and playing with the chain rule for derivatives, we get

$$\frac{d}{dt}d(x(t), y(t))^2 \le -2\lambda d(x(t), y(t))^2.$$

If we want a satisfying theory for gradient flows which includes uniqueness, we just need to prove the existence of curves which satisfy the EVI condition, accepting that this will probably require additional assumptions. This can still be done via the discrete scheme, adding a compatibility hypothesis between the function F and the distance d, a condition which involves some sort of convexity. We do not enter the details of the proof, for which we refer to [3], where the convergence to an EVI gradient flow is proven, with explicit error estimates. These a priori estimates allow to prove that we have a Cauchy sequence, and then allow to get rid of the compactness part of the proof (by the way, we could even avoid using compactness so as to prove existence of a minimizer at every time step, using almost-minimizers and the in the Ekeland's variational principle [39]). Here, we will just present this extra convexity assumption, needed for the existence of EVI gradient flows developed in [3].

This assumption, that we will call C^2G^2 (*Compatible Convexity along Generalized Geodesics*) is the following: suppose that, for every pair (x_0, x_1) and every $y \in X$, there is a curve $x(t)$ connecting $x(0) = x_0$ to $x(1) = x_1$, such that

$$F(x(t)) \le (1-t)F(x_0) + tF(x_1) - \lambda\frac{t(1-t)}{2}d^2(x_0, x_1),$$

$$d^2(x(t), y) \le (1-t)d^2(x_0, y) + td^2(x_1, y) - t(1-t)d^2(x_0, x_1).$$

In other words, we require λ-convexity of the function F, but also the 2-convexity of the function $x \mapsto d^2(x, y)$, along a same curve which is not necessarily the geodesic. This second condition is automatically satisfied, by using the geodesic itself, in the Euclidean space (and in every Hilbert space), since the function $x \mapsto |x - y|^2$ is quadratic, and its Hessian matrix is $2I$ at every point. We can also see that it is satisfied in a normed space if and only if the norm is induced by a scalar product. It has been recently pointed out by Gigli that the sharp condition on the space X in order to guarantee existence of EVI gradient flows is that X should be *infinitesimally Hilbertian* (this will be made precise in Sect. 5).

Here, we just observe that C^2G^2 implies $(\lambda + \frac{1}{\tau})$-convexity, along those curves, sometimes called *generalized geodesics* (consider that these curves also depend on a third point, sort of a base point, typically different from the two points that should be connected), of the functional that we minimize at each time step in the minimizing movement scheme. This provides uniqueness of the minimizer as soon as τ is small enough, and allows to perform the desired estimates.

Also, the choice of this C^2G^2 condition, which is a technical condition whose role is only to prove the existence of an EVI gradient flow, has been done in view of the

applications to the case of the Wasserstein spaces, that wil be the object of the next section. Indeed, in these spaces the squared distance is not in general 2-convex along geodesics, but we can find some adapted curves on which it is 2-convex, and many functionals F stay convex on these curves.

We finish this section by mentioning a recent extension to some non-λ-convex functionals. The starting point is the fact that the very use of Gronwall's lemma to prove uniqueness can be modified by allowing for a weaker condition. Indeed, it is well-known that, whenever a function ω satisfies an Osgood condition $\int_0^1 \frac{1}{\omega(s)} ds = +\infty$, then $E' \leq \omega(E)$ together with $E(0) = 0$ implies $E(t) = 0$ for $t > 0$. This suggests that one could define a variant of the EVI definition for functions which are not λ-convex, but almost, and this is the theory developed in [34]. Such a paper studies the case where F satisfies some sort of ω-convexity for a "modulus of convexity" ω. More precisely, this means

$$F(x_t) \leq (1-t)F(_0) + tF(x_1) - \frac{|\lambda|}{2}\Big[(1-t)\omega(t^2 d(x_0, x_1)^2) \\ + t\omega((1-t)^2 d(x_0, x_1)^2)\Big],$$

on generalized geodesics x_t (note that in the case $\omega(s) = s$ we come back to (3.2)). The function ω is required to satisfy an Osgood condition (and some other technical conditions). Then, the EVI condition is replaced by

$$\frac{d}{dt}\frac{1}{2}d(x(t), y)^2 \leq F(y) - F(x(t)) + \frac{|\lambda|}{2}\omega(d(x(t), y)^2),$$

and this allows to produce a theory with existence and uniqueness results (via a variant of Proposition 3.1). In the Wasserstein spaces (see next section), a typical case of functionals which can fit this theory are functionals involving singular interaction kernels (or solutions to elliptic PDEs, as in the Keller–Segel case) under L^∞ constraints on the density (using the fact that the gradient ∇u of the solution of $-\Delta u = \varrho$ is not Lipschitz when $\varrho \in L^\infty$, but is at least log-lipschitz).

4 Gradient flows in the Wasserstein space

One of the most exciting applications (and maybe the only one,[5] in what concerns applied mathematics) of the theory of gradient flows in metric spaces is that of evolution PDEs in the space of measures. This topic is inspired from the work by Jordan, Kinderlehrer and Otto [54], who had the intuition that the heat and the Fokker–Planck equations have a common variational structure in terms of a particular distance on the probability measures, the so-called Wasserstein distance. Yet, the theory has only become formal and general with the work by Ambrosio, Gigli and Savaré (which does

---

[5] This is surely exaggerated, as we could think for instance at the theory of geometrical evolutions of shapes and sets, even if it seems that this metric approach has not yet been generalized in this framework.

not mean that proofs in [54] were not rigorous, but the intuition on the general structure still needed to be better understood).

The main idea is to endow the space $\mathcal{P}(\Omega)$ of probability measures on a domain $\Omega \subset \mathbb{R}^d$ with a distance, and then deal with gradient flows of suitable functionals on such a metric space. Such a distance arises from optimal transport theory. More details about optimal transport can be found in the books by Villani [90,91] and in the book on gradient flows by Ambrosio et al. [3][6]; a recent book by the author of this survey is also available [84].

4.1 Preliminaries on optimal transport

The motivation for the whole subject is the following problem proposed by Monge [77]: given two densities of mass f, $g \geq 0$ on $\mathbb{R}^d$, with $\int f = \int g = 1$, find a map $T : \mathbb{R}^d \to \mathbb{R}^d$ pushing the first one onto the other, i.e. such that

$$\int_A g(x)\mathrm{d}x = \int_{T^{-1}(A)} f(y)\mathrm{d}y \quad \text{for any Borel subset } A \subset \mathbb{R}^d \qquad (4.1)$$

and minimizing the quantity

$$\int_{\mathbb{R}^d} |T(x) - x| f(x)\,\mathrm{d}x$$

among all the maps satisfying this condition. This means that we have a collection of particles, distributed with density f on $\mathbb{R}^d$, that have to be moved, so that they arrange according to a new distribution, whose density is prescribed and is g. The movement has to be chosen so as to minimize the average displacement. The map T describes the movement, and $T(x)$ represents the destination of the particle originally located at x. The constraint on T precisely accounts for the fact that we need to reconstruct the density g. In the sequel, we will always define, similarly to (4.1), the image measure of a measure μ on X (measures will indeed replace the densities f and g in the most general formulation of the problem) through a measurable map $T : X \to Y$: it is the measure denoted by $T_{\#}\mu$ on Y and characterized by

$$(T_{\#}\mu)(A) = \mu(T^{-1}(A)) \quad \text{for every measurable set } A,$$
$$\text{or } \int_Y \phi \, \mathrm{d}(T_{\#}\mu) = \int_X \phi \circ T \, \mathrm{d}\mu \quad \text{for every measurable function } \phi.$$

The problem of Monge has stayed with no solution (does a minimizer exist? how to characterize it?…) till the progress made in the 1940 s with the work by Kantorovich

---

[6] Lighter versions exist, such as [8], or the recent *User's Guide to Optimal Transport* [2], which is a good reference for many topics in this survey, as it deals for one half with optimal transport (even if the title suggests that this is the only topic of the guide), then for one sixth with the general theory of gradient flows (as in our Sect. 3), and finally for one third with metric spaces with curvature bounds (that we will briefly sketch in Sect. 5).

[55]. In the Kantorovich's framework, the problem has been widely generalized, with very general cost functions $c(x, y)$ instead of the Euclidean distance $|x - y|$ and more general measures and spaces.

Let us start from the general picture. Consider a metric space X, that we suppose compact for simplicity[7] and a cost function $c : X \times X \to [0, +\infty]$. For simplicity of the exposition, we will suppose that c is continuous and symmetric: $c(x, y) = c(y, x)$ (in particular, the target and source space will be the same space X).

The formulation proposed by Kantorovich of the problem raised by Monge is the following: given two probability measures $\mu, \nu \in \mathcal{P}(X)$, consider the problem

$$(KP) \quad \min \left\{ \int_{X \times X} c \, d\gamma \; : \; \gamma \in \Pi(\mu, \nu) \right\}, \tag{4.2}$$

where $\Pi(\mu, \nu)$ is the set of the so-called *transport plans*, i.e.

$$\Pi(\mu, \nu) = \{ \gamma \in \mathcal{P}(X \times X) : (\pi_0)_\# \gamma = \mu, \; (\pi_1)_\# \gamma = \nu, \}$$

where π_0 and π_1 are the two projections of $X \times X$ onto its factors. These probability measures over $X \times X$ are an alternative way to describe the displacement of the particles of μ: instead of saying, for each x, which is the destination $T(x)$ of the particle originally located at x, we say for each pair (x, y) how many particles go from x to y. It is clear that this description allows for more general movements, since from a single point x particles can a priori move to different destinations y. If multiple destinations really occur, then this movement cannot be described through a map T. It can be easily checked that if $(id, T)_\# \mu$ belongs to $\Pi(\mu, \nu)$ then T pushes μ onto ν (i.e. $T_\# \mu = \nu$) and the functional takes the form $\int c(x, T(x)) d\mu(x)$, thus generalizing Monge's problem.

The minimizers for this problem are called *optimal transport plans* between μ and ν. Should γ be of the form $(id, T)_\# \mu$ for a measurable map $T : X \to X$ (i.e. when no splitting of the mass occurs), the map T would be called *optimal transport map* from μ to ν.

This generalized problem by Kantorovich is much easier to handle than the original one proposed by Monge: for instance in the Monge case we would need existence of at least a map T satisfying the constraints. This is not verified when $\mu = \delta_0$, if ν is not a single Dirac mass. On the contrary, there always exists at least a transport plan in $\Pi(\mu, \nu)$ (for instance we always have $\mu \otimes \nu \in \Pi(\mu, \nu)$). Moreover, one can state that (KP) is the relaxation of the original problem by Monge: if one considers the problem in the same setting, where the competitors are transport plans, but sets the functional at $+\infty$ on all the plans that are not of the form $(id, T)_\# \mu$, then one has a functional on $\Pi(\mu, \nu)$ whose relaxation (in the sense of the largest lower-semicontinuous functional smaller than the given one) is the functional in (KP) (see for instance Section 1.5 in [84]).

---

[7] Most of the results that we present stay true without this assumptions, anyway, and we refer in particular to [3] or [90] for details, since a large part of the analysis of [84] is performed under this simplifying assumption.

Anyway, it is important to note that an easy use of the Direct Method of Calculus of Variations (i.e. taking a minimizing sequence, saying that it is compact in some topology—here it is the weak convergence of probability measures—finding a limit, and proving semicontinuity, or continuity, of the functional we minimize, so that the limit is a minimizer) proves that a minimum does exist. As a consequence, if one is interested in the problem of Monge, the question may become "does this minimizer come from a transport map T?" (note, on the contrary, that directly attacking by compactness and semicontinuity Monge's formulation is out of reach, because of the non-linearity of the constraint $T_\# \mu = \nu$, which is not closed under weak convergence).

Since the problem (*KP*) is a linear optimization under linear constraints, an important tool will be duality theory, which is typically used for convex problems. We will find a dual problem (*DP*) for (*KP*) and exploit the relations between dual and primal.

The first thing we will do is finding a formal dual problem, by means of an inf-sup exchange.

First express the constraint $\gamma \in \Pi(\mu, \nu)$ in the following way : notice that, if γ is a non-negative measure on $X \times X$, then we have

$$\sup_{\phi, \psi} \int \phi \, d\mu + \int \psi \, d\nu - \int (\phi(x) + \psi(y)) \, d\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu, \nu) \\ +\infty & \text{otherwise} \end{cases}.$$

Hence, one can remove the constraints on γ by adding the previous sup, since if they are satisfied nothing has been added and if they are not one gets $+\infty$ and this will be avoided by the minimization. We may look at the problem we get and interchange the inf in γ and the sup in ϕ, ψ:

$$\min_{\gamma \geq 0} \int c \, d\gamma + \sup_{\phi, \psi} \left(\int \phi \, d\mu + \int \psi \, d\nu - \int (\phi(x) + \psi(y)) \, d\gamma \right)$$

becomes

$$\sup_{\phi, \psi} \int \phi \, d\mu + \int \psi \, d\nu + \inf_{\gamma \geq 0} \int (c(x, y) - (\phi(x) + \psi(y))) \, d\gamma.$$

Obviously it is not always possible to exchange inf and sup, and the main tools to do this come from convex analysis. We refer to [84], Section 1.6.3 for a simple proof of this fact, or to [90], where the proof is based on Flenchel–Rockefeller duality (see, for instance, [41]). Anyway, we insist that in this case it is true that inf sup = sup inf.

Afterwards, one can re-write the inf in γ as a constraint on ϕ and ψ, since one has

$$\inf_{\gamma \geq 0} \int (c(x, y) - (\phi(x) + \psi(y))) \, d\gamma$$

$$= \begin{cases} 0 & \text{if } \phi(x) + \psi(y) \leq c(x, y) \text{ for all } (x, y) \in X \times X \\ -\infty & \text{otherwise} \end{cases}.$$

This leads to the following dual optimization problem: given the two probabilities μ and ν and the cost function $c : X \times X \to [0, +\infty]$ we consider the problem

$$(DP) \quad \max \left\{ \int_X \phi \, d\mu + \int_X \psi \, d\nu \; : \; \phi, \psi \in C(X), \; \phi(x) + \psi(y) \leq c(x, y) \right.$$

$$\left. \text{for all } (x, y) \in X \times X \right\}. \tag{4.3}$$

This problem does not admit a straightforward existence result, since the class of admissible functions lacks compactness. Yet, we can better understand this problem and find existence once we have introduced the notion of c-transform (a kind of generalization of the well-known Legendre transform).

Definition 4.1 Given a function $\chi : X \to \overline{\mathbb{R}}$ we define its *c-transform* (or c-conjugate function) by

$$\chi^c(y) = \inf_{x \in X} c(x, y) - \chi(x).$$

Moreover, we say that a function ψ is *c-concave* if there exists χ such that $\psi = \chi^c$ and we denote by $\Psi_c(X)$ the set of c-concave functions.

It is quite easy to realize that, given a pair (ϕ, ψ) in the maximization problem (DP), one can always replace it with (ϕ, ϕ^c), and then with (ϕ^{cc}, ϕ^c), and the constraints are preserved and the integrals increased. Actually one could go on but it is possible to prove that $\phi^{ccc} = \phi^c$ for any function ϕ. This is the same as saying that $\psi^{cc} = \psi$ for any c-concave function ψ, and this perfectly recalls what happens for the Legendre transform of convex functions (which corresponds to the particular case $c(x, y) = -x \cdot y$).

A consequence of these considerations is the following well-known result:

Proposition 4.1 *We have*

$$\min (KP) = \max_{\phi \in \Psi_c(X)} \int_X \phi \, d\mu + \int_X \phi^c \, d\nu, \tag{4.4}$$

where the max on the right hand side is realized. In particular the minimum value of (KP) *is a convex function of* (μ, ν), *as it is a supremum of linear functionals.*

Definition 4.2 The functions ϕ realizing the maximum in (4.4) are called *Kantorovich potentials* for the transport from μ to ν (and will be often denoted by the symbol φ instead of ϕ).

Notice that any c-concave function shares the same modulus of continuity of the cost c. Hence, if c is uniformly continuous (which is always the case whenever c is continuous and X is compact), one can get a uniform modulus of continuity for the functions in $\Psi_c(X)$. This is the reason why one can prove existence for (DP) (which is the same of the right hand side problem in Proposition 4.1), by applying Ascoli–Arzelà's Theorem.

We look at two interesting cases. When $c(x, y)$ is equal to the distance $d(x, y)$ on the metric space X, then we can easily see that

$$\phi \in \Psi_c(X) \Longleftrightarrow \phi \text{ is a 1-Lipschitz function} \tag{4.5}$$

and

$$\phi \in \Psi_c(X) \Rightarrow \phi^c = -\phi. \tag{4.6}$$

Another interesting case is the case where $X = \Omega \subset \mathbb{R}^d$ and $c(x, y) = \frac{1}{2}|x - y|^2$. In this case we have

$$\phi \in \Psi_c(X) \Longrightarrow x \mapsto \frac{x^2}{2} - \phi(x) \text{ is a convex function.}$$

Moreover, if $X = \mathbb{R}^d$ this is an equivalence.

A consequence of (4.5) and (4.6) is that, in the case where $c = d$, Formula 4.4 may be re-written as

$$\min (KP) = \max (DP) = \max_{\phi \in \mathrm{Lip}_1} \int_X \phi \, d(\mu - \nu). \tag{4.7}$$

We now concentrate on the quadratic case when X is a domain $\Omega \subset \mathbb{R}^d$, and look at the existence of optimal transport maps T. From now on, we will use the word *domain* to denote the closure of a bounded and connected open set.

The main tool is the duality result. If we have equality between the minimum of (KP) and the maximum of (DP) and both extremal values are realized, one can consider an optimal transport plan γ and a Kantorovich potential φ and write

$$\varphi(x) + \varphi^c(y) \le c(x, y) \text{ on } X \times X \text{ and } \varphi(x) + \varphi^c(y) = c(x, y) \text{ on } \mathrm{spt}\gamma.$$

The equality on $\mathrm{spt}\gamma$ is a consequence of the inequality which is valid everywhere and of

$$\int c \, d\gamma = \int \varphi \, d\mu + \int \varphi^c \, d\nu = \int (\varphi(x) + \varphi^c(y)) \, d\gamma,$$

which implies equality γ-a.e. These functions being continuous, the equality passes to the support of the measure. Once we have that, let us fix a point $(x_0, y_0) \in \mathrm{spt}\gamma$. One may deduce from the previous computations that

$$x \mapsto \varphi(x) - \frac{1}{2}|x - y_0|^2 \quad \text{is maximal at } x = x_0$$

and, if φ is differentiable at x_0, one gets $\nabla\varphi(x_0) = x_0 - y_0$, i.e. $y_0 = x_0 - \nabla\varphi(x_0)$. This shows that only one unique point y_0 can be such that $(x_0, y_0) \in \mathrm{spt}\gamma$, which means that γ is concentrated on a graph. The map T providing this graph is of the form $x \mapsto x - \nabla\varphi(x) = \nabla u(x)$ (where $u(x):=\frac{x^2}{2} - \varphi(x)$ is a convex function). This shows

the following well-known theorem, due to Brenier ([19,20], see also [45–47,71]). Note that this also gives uniqueness of the optimal transport plan and of the gradient of the Kantorovich potential. The only technical point to let this strategy work is the μ-a.e. differentiability of the potential φ. Since φ has the same regularity of a convex function, and convex function are locally Lipschitz, it is differentiable Lebesgue-a.e., which allows to prove the following:

Theorem 4.2 *Given μ and ν probability measures on a domain $\Omega \subset \mathbb{R}^d$ there exists an optimal transport plan γ for the quadratic cost $\frac{1}{2}|x-y|^2$. It is unique and of the form $(id, T)_{\#}\mu$, provided μ is absolutely continuous. Moreover there exists also at least a Kantorovich potential φ, and the gradient $\nabla\varphi$ is uniquely determined μ-a.e. (in particular φ is unique up to additive constants if the density of μ is positive a.e. on Ω). The optimal transport map T and the potential φ are linked by $T(x) = x - \nabla\varphi(x)$. Moreover, the optimal map T is equal a.e. to the gradient of a convex function u, given by $u(x):=\frac{x^2}{2} - \varphi(x)$.*

Actually, the existence of an optimal transport map is true under weaker assumptions (see [47]): we can replace the condition of being absolutely continuous with the condition "$\mu(A) = 0$ for any $A \subset \mathbb{R}^d$ such that $\mathcal{H}^{d-1}(A) < +\infty$" or with any condition which ensures that the non-differentiability set of φ is negligible (and convex function are more regular than locally Lipschitz functions).

In Theorem 4.2 only the part concerning the optimal map T is not symmetric in μ and ν: hence the uniqueness of the Kantorovich potential is true even if it ν (and not μ) has positive density a.e. (since one can retrieve φ from φ^c and viceversa).

We stress that Theorem 4.2 admits a converse implication and that any gradient of a convex function is indeed optimal between μ and its image measure. Moreover, Theorem 4.2 can be translated in very easy terms in the one-dimensional case $d = 1$: given a non-atomic measure μ and another measure ν, both in $\mathcal{P}(\mathbb{R})$, there exists a unique monotone increasing transport map T such that $T_{\#}\mu = \nu$, and it is optimal for the quadratic cost.

Finally, the same kind of arguments could be adapted to prove existence and uniqueness of an optimal map for other costs, in particular to costs of the form $c(x, y) = h(x - y)$ for a strictly convex function $h : \mathbb{R}^d \to \mathbb{R}$, which includes all the costs of the form $c(x, y) = |x - y|^p$ with $p > 1$. In the one-dimensional case it even happens that the same monotone increasing map is optimal for every $p \geq 1$ (and it is the unique optimal map for $p > 1$)!

4.2 The Wasserstein distances

Starting from the values of the problem (KP) in (4.2) we can define a set of distances over $\mathcal{P}(X)$.

We mainly consider costs of the form $c(x, y) = |x - y|^p$ in $X = \Omega \subset \mathbb{R}^d$, but the analysis can be adapted to a power of the distance in a more general metric space X. The exponent p will always be taken in $[1, +\infty[$ (we will not discuss the case $p = \infty$) in order to take advantage of the properties of the L^p norms. When Ω is unbounded we need to restrict our analysis to the following set of probabilities

$$\mathcal{P}_p(\Omega) := \left\{ \mu \in \mathcal{P}(\Omega) \; : \; \int_\Omega |x|^p \, \mathrm{d}\mu(x) < +\infty \right\}.$$

In a metric space, we fix an arbitrary point $x_0 \in X$, and set

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \; : \; \int_X d(x, x_0)^p \, \mathrm{d}\mu(x) < +\infty \right\}$$

(the finiteness of this integral does not depend on the choice of x_0).

The distances that we want to consider are defined in the following way: for any $p \in [1, +\infty[$ set

$$W_p(\mu, \nu) = \left(\min (KP) \text{ with } c(x, y) = |x - y|^p \right)^{1/p}.$$

The quantities that we obtain in this way are called Wasserstein distances.[8] They are very important in many fields of applications and they seem a natural way to describe distances between equal amounts of mass distributed on a same space.

It is interesting to compare these distances to L^p distances between densities (a comparison which is meaningful when we consider absolutely continuous measures on $\mathbb{R}^d$, for instance). A first observation is the very different behavior of these two classes of distances. We could say that, if L^p distances can be considered "vertical", Wasserstein distances are instead "horizontal". This consideration is very informal, but is quite natural if one associates with every absolutely continuous measure the graph of its density. To compute $||f - g||_{L^p}$ we need to look, for every point x, at the distance between $f(x)$ and $g(x)$, which corresponds to a vertical displacement between the two graphs, and then integrate this quantity. On the contrary, to compute $W_p(f, g)$ we need to consider the distance between a point x and a point $T(x)$ (i.e. an horizontal displacement on the graph) and then to integrate this, for a particular pairing between x and $T(x)$ which makes a coupling between f and g.

A first example where we can see the very different behavior of these two ways of computing distances is the following: take two densities f and g supported on $[0, 1]$, and define g_h as $g_h(x) = g(x - h)$. As soon as $|h| > 1$, the L^p distance between f and g_h equals $(||f||_{L^p}^p + ||g||_{L^p}^p)^{1/p}$, and does not depend on the "spatial" information consisting in $|h|$. On the contrary, the W_p distance between f and g_h is of the order of $|h|$ (for $h \to \infty$) and depends much more on the displacement than on the shapes of f and g (Fig. 1).

We now analyze some properties of these distances. Most proofs can be found in [84], Chapter 5, or in [90] or [3].

---

[8] They are named after L. Vaserstein (whose name is sometimes spelled Wasserstein), but this choice is highly debated, in particular in Russia, as the role he played in these distances is very marginal. However, this is now the standard name used in Western countries, probably due to the terminology used in [54,78], even if other names have been suggested, such as Monge–Kantorovich distances, Kantorovich–Rubinstein… and we will stick to this terminology.
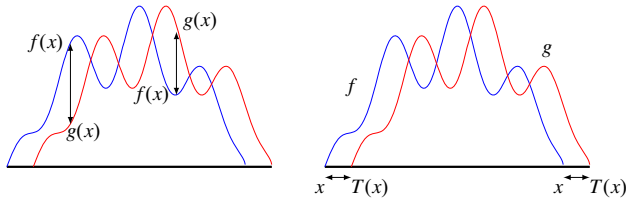
Fig. 1 "Vertical" versus "horizontal" distances (in the particular case where $g = f_h$)

First we underline that, as a consequence of Hölder (or Jensen) inequalities, the Wasserstein distances are always ordered, i.e. $W_{p_1} \leq W_{p_2}$ if $p_1 \leq p_2$. Reversed inequalities are possible only if Ω is bounded, and in this case we have, if set $D = \text{diam}(\Omega)$, for $p_1 \leq p_2$,

$$W_{p_1} \leq W_{p_2} \leq D^{1-p_1/p_2} W_{p_1}^{p_1/p_2}.$$

This automatically guarantees that, if the quantities W_p are distances, then they induce the same topology, at least when Ω is bounded. But first we should check that they are distances…

Proposition 4.3 *The quantity W_p defined above is a distance over $\mathcal{P}_p(\Omega)$.*

Proof First, let us note that $W_p \geq 0$. Then, we also remark that $W_p(\mu, \nu) = 0$ implies that there exists $\gamma \in \Pi(\mu, \nu)$ such that $\int |x - y|^p \, d\gamma = 0$. Such a $\gamma \in \Pi(\mu, \nu)$ is concentrated on $\{x = y\}$. This implies $\mu = \nu$ since, for any test function ϕ, we have

$$\int \phi \, d\mu = \int \phi(x) \, d\gamma = \int \phi(y) \, d\gamma = \int \phi \, d\nu.$$

We need now to prove the triangle inequality. We only give a proof in the case $p > 1$, with absolutely continuous measures.

Take three measures μ, ϱ and ν, and suppose that μ and ϱ are absolutely continuous. Let T be the optimal transport from μ to ϱ and S the optimal one from ϱ to ν. Then $S \circ T$ is an admissible transport from μ to ν, since $(S \circ T)_{\#}\mu = S_{\#}(T_{\#}\mu) = S_{\#}\varrho = \nu$. We have

$$W_p(\mu, \nu) \leq \left(\int |S(T(x)) - x|^p \, d\mu(x) \right)^{\frac{1}{p}} = ||S \circ T - id||_{L^p(\mu)}$$

$$\leq ||S \circ T - T||_{L^p(\mu)} + ||T - id||_{L^p(\mu)}.$$

Moreover,

$$||S \circ T - T||_{L^p(\mu)} = \left(\int |S(T(x)) - T(x)|^p \, d\mu(x) \right)^{\frac{1}{p}} = \left(\int |S(y) - y|^p d\varrho(y) \right)^{\frac{1}{p}}$$

and this last quantity equals $W_p(\varrho, \nu)$. Moreover, $||T - id||_{L^p(\mu)} = W_p(\mu, \varrho)$, whence

$$W_p(\mu, \nu) \leq W_p(\mu, \varrho) + W_p(\varrho, \nu).$$

This gives the proof when $\mu, \varrho \ll \mathcal{L}^d$ and $p > 1$. For the general case, an approximation is needed (but other arguments can also apply, see, for instance, [84], Section 5.1).□

We now give, without proofs, two results on the topology induced by W_p on a general metric space X.

Theorem 4.4 *If X is compact, for any $p \geq 1$ the function W_p is a distance over $\mathcal{P}(X)$ and the convergence with respect to this distance is equivalent to the weak convergence of probability measures.*

To prove that the convergence according to W_p is equivalent to weak convergence one first establish this result for $p = 1$, through the use of the duality formula in the form (4.7). Then it is possible to use the inequalities between the distances W_p (see above) to extend the result to a general p.

The case of a noncompact space X is a little more difficult. As we said, the distance is only defined on a subset of the whole space of probability measures, to avoid infinite values. Set, for a fixed reference point x_0 which can be chosen to be 0 in the Euclidean space,

$$m_p(\mu) := \int_X d(x, x_0)^p d\mu(x).$$

In this case, the distance W_p is only defined on $\mathcal{P}_p(X) := \{\mu \in \mathcal{P}(X) : m_p(\mu) < +\infty\}$. We have

Theorem 4.5 *For any $p \geq 1$ the function W_p is a distance over $\mathcal{P}_p(X)$ and, given a measure μ and a sequence $(\mu_n)_n$ in $\mathbb{W}_p(X)$, the following are equivalent:*

- *$\mu_n \to \mu$ according to W_p;*
- *$\mu_n \rightharpoonup \mu$ and $m_p(\mu_n) \to m_p(\mu)$;*
- *$\int_X \phi \, d\mu_n \to \int_X \phi \, d\mu$ for any $\phi \in C^0(X)$ whose growth is at most of order p (i.e. there exist constants A and B depending on ϕ such that $|\phi(x)| \leq A + B d(x, x_0)^p$ for any x).*

After this short introduction to the metric space $\mathbb{W}_p := (\mathcal{P}_p(X), W_p)$ and its topology, we will focus on the Euclidean case, i.e. where the metric space X is a domain $\Omega \subset \mathbb{R}^d$, and study the curves valued in $\mathbb{W}_p(\Omega)$ in connections with PDEs.

The main point is to identify the absolutely continuous curves in the space $\mathbb{W}_p(\Omega)$ with solutions of the continuity equation $\partial_t \mu_t + \nabla \cdot (\mathbf{v}_t \mu_t) = 0$ with L^p vector fields $\mathbf{v}_t$. Moreover, we want to connect the L^p norm of $\mathbf{v}_t$ with the metric derivative $|\mu'|(t)$.

We recall that standard considerations from fluid mechanics tell us that the continuity equation above may be interpreted as the equation ruling the evolution of the

density μ_t of a family of particles initially distributed according to μ_0 and each of which follows the flow

$$\begin{cases} y'_x(t) = \mathbf{v}_t(y_x(t)) \\ y_x(0) = x. \end{cases}$$

The main theorem in this setting (originally proven in [3]) relates absolutely continuous curves in $\mathbb{W}_p$ with solutions of the continuity equation:

Theorem 4.6 *Let $(\mu_t)_{t \in [0,1]}$ be an absolutely continuous curve in $\mathbb{W}_p(\Omega)$ (for $p > 1$ and $\Omega \subset \mathbb{R}^d$ an open domain). Then for a.e. $t \in [0, 1]$ there exists a vector field $\mathbf{v}_t \in L^p(\mu_t; \mathbb{R}^d)$ such that*

- *the continuity equation $\partial_t \mu_t + \nabla \cdot (\mathbf{v}_t \mu_t) = 0$ is satisfied in the sense of distributions,*
- *for a.e. t we have $||\mathbf{v}_t||_{L^p(\mu_t)} \leq |\mu'|(t)$ (where $|\mu'|(t)$ denotes the metric derivative at time t of the curve $t \mapsto \mu_t$, w.r.t. the distance W_p);*

Conversely, if $(\mu_t)_{t \in [0,1]}$ is a family of measures in $\mathcal{P}_p(\Omega)$ and for each t we have a vector field $\mathbf{v}_t \in L^p(\mu_t; \mathbb{R}^d)$ with $\int_0^1 ||\mathbf{v}_t||_{L^p(\mu_t)} \, dt < +\infty$ solving $\partial_t \mu_t + \nabla \cdot (\mathbf{v}_t \mu_t) = 0$, then $(\mu_t)_t$ is absolutely continuous in $\mathbb{W}_p(\Omega)$ and for a.e. t we have $|\mu'|(t) \leq ||\mathbf{v}_t||_{L^p(\mu_t)}$.

Note that, as a consequence of the second part of the statement, the vector field $\mathbf{v}_t$ introduced in the first part must a posteriori satisfy $||\mathbf{v}_t||_{L^p(\mu_t)} = |\mu'|(t)$.

We will not give the proof of this theorem, which is quite involved. The main reference is [3] (but the reader can also find alternative proofs in Chapter 5 of [84], in the case where Ω is compact). Yet, if the reader wants an idea of the reason for this theorem to be true, it is possible to start from the case of two time steps: there are two measures μ_t and μ_{t+h} and there are several ways for moving the particles so as to reconstruct the latter from the former. It is exactly as when we look for a transport. One of these transports is optimal in the sense that it minimizes $\int |T(x) - x|^p d\mu_t(x)$ and the value of this integral equals $W_p^p(\mu_t, \mu_{t+h})$. If we call $\mathbf{v}_t(x)$ the "discrete velocity of the particle located at x at time t, i.e. $\mathbf{v}_t(x) = (T(x) - x)/h$, one has $||\mathbf{v}_t||_{L^p(\mu_t)} = \frac{1}{h} W_p(\mu_t, \mu_{t+h})$. We can easily guess that, at least formally, the result of the previous theorem can be obtained as a limit as $h \to 0$.

Once we know about curves in their generality, it is interesting to think about geodesics. The following result is a characterization of geodesics in $W_p(\Omega)$ when Ω is a convex domain in $\mathbb{R}^d$. This procedure is also known as *McCann's displacement interpolation.*

Theorem 4.7 *If $\Omega \subset \mathbb{R}^d$ is convex, then all the spaces $\mathbb{W}_p(\Omega)$ are length spaces and if μ and ν belong to $\mathbb{W}_p(\Omega)$, and γ is an optimal transport plan from μ to ν for the cost $c_p(x, y) = |x - y|^p$, then the curve*

$$\mu^\gamma(t) = (\pi_t)_\# \gamma,$$

where $\pi_t : \Omega \times \Omega \to \Omega$ is given by $\pi_t(x, y) = (1 - t)x + ty$, is a constant-speed geodesic from μ to ν. In the case $p > 1$ all the constant-speed geodesics are of this

form, and, if μ is absolutely continuous, then there is only one geodesic and it has the form

$$\mu_t = [T_t]\#\mu, \quad \text{where } T_t := (1-t)id + tT$$

where T is the optimal transport map from μ to ν. In this case, the velocity field $\mathbf{v}_t$ of the geodesic μ_t is given by $\mathbf{v}_t = (T - id) \circ (T_t)^{-1}$. In particular, for $t = 0$ we have $\mathbf{v}_0 = -\nabla\varphi$ and for $t = 1$ we have $\mathbf{v}_1 = \nabla\psi$, where φ is the Kantorovich potential in the transport from μ to ν and $\psi = \varphi^c$.

The above theorem may be adapted to the case where the Euclidean domain Ω is replaced by a Riemannian manifold, and in this case the map T_t must be defined by using geodesics instead of segments: the point $T_t(x)$ will be the value at time t of a constant-speed geodesic, parametrized on the interval $[0, 1]$, connecting x to $T(x)$. For the theory of optimal transport on manifolds, we refer to [73].

Using the characterization of constant-speed geodesics as minimizers of a strictly convex kinetic energy, we can also infer the following interesting information.

- Looking for an optimal transport for the cost $c(x, y) = |x - y|^p$ is equivalent to looking for a constant-speed geodesic in $\mathbb{W}_p$ because from optimal plans we can reconstruct geodesics and from geodesics (via their velocity field) it is possible to reconstruct the optimal transport;
- constant-speed geodesics may be found by minimizing $\int_0^1 |\mu'|(t)^p dt$;
- in the case of the Wasserstein spaces, we have $|\mu'|(t)^p = \int_\Omega |\mathbf{v}_t|^p d\mu_t$, where $\mathbf{v}$ is a velocity field solving the continuity equation together with μ (this field is not unique, but the metric derivative $|\mu'|(t)$ equals the minimal value of the L^p norm of all possible fields).

As a consequence of these last considerations, for $p > 1$, solving the kinetic energy minimization problem

$$\min\left\{\int_0^1 \int_\Omega |\mathbf{v}_t|^p d\varrho_t \, dt \; : \; \partial_t\varrho_t + \nabla \cdot (\mathbf{v}_t\varrho_t) = 0, \; \varrho_0 = \mu, \; \varrho_1 = \nu\right\} \quad (4.8)$$

selects constant-speed geodesics connecting μ to ν, and hence allows to find the optimal transport between these two measures. This is what is usually called *Benamou–Brenier formula* (see [11] for the quadratic case, and [21] for a general introduction).

On the other hand, this minimization problem in the variables $(\varrho_t, \mathbf{v}_t)$ has non-linear constraints (due to the product $\mathbf{v}_t\varrho_t$) and the functional is non-convex (since $(t, x) \mapsto t|x|^p$ is not convex). Yet, it is possible to transform it into a convex problem. For this, it is sufficient to switch variables, from $(\varrho_t, \mathbf{v}_t)$ into (ϱ_t, E_t) where $E_t = \mathbf{v}_t\varrho_t$, thus obtaining the following minimization problem

$$\min\left\{\int_0^1 \int_\Omega \frac{|E_t|^p}{\varrho_t^{p-1}} \, dxdt \; : \; \partial_t\varrho_t + \nabla \cdot E_t = 0, \; \varrho_0 = \mu, \; \varrho_1 = \nu\right\}. \quad (4.9)$$

We need to use the properties of the function $f_p : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, defined through

$$f_p(t, x) := \sup_{(a,b) \in K_q} (at + b \cdot x) = \begin{cases} \frac{1}{p} \frac{|x|^p}{t^{p-1}} & \text{if } t > 0, \\ 0 & \text{if } t = 0, x = 0 \\ +\infty & \text{if } t = 0, x \neq 0, \text{ or } t < 0, \end{cases}$$

where $K_q := \{(a, b) \in \mathbb{R} \times \mathbb{R}^d \ : \ a + \frac{1}{q}|b|^q \leq 0\}$ and $q = p/(p-1)$ is the conjugate exponent of p. In particular, f_p is convex, which makes the above minimization problem convex, and also allows to write what we formally wrote as $\int_0^1 \int_\Omega \frac{|E_t|^p}{\varrho_t^{p-1}} \, dx dt$ (an expression which implicitly assumes $\varrho_t, E_t \ll \mathcal{L}^d$) into the form

$$\mathcal{B}_p(\varrho, E) := \sup \left\{ \int a \, d\varrho + \int b \cdot dE \ : \ (a, b) \in C(\Omega \times [0, 1]; K_q) \right\}.$$

Both the convexity and this last expression will be useful for numerical methods (as it was first done in [11]).

4.3 Minimizing movement schemes in the Wasserstein space and evolution PDEs

Thanks to all the theory which has been described so far, it is natural to study gradient flows in the space $\mathbb{W}_2(\Omega)$ (the reason for choosing the exponent $p = 2$ will be clear in a while) and to connect them to PDEs of the form of a continuity equation. The most convenient way to study this is to start from the time-discretized problem, i.e. to consider a sequence of iterated minimization problems:

$$\varrho_{k+1}^\tau \in \operatorname{argmin}_\varrho F(\varrho) + \frac{W_2^2(\varrho, \varrho_k^\tau)}{2\tau}. \tag{4.10}$$

This iterated minimization scheme is called *JKO scheme* (after Jordan et al. [54]).

Note that we denote now the measures on Ω by the letter ϱ instead of μ or ν because we expect them to be absolutely continuous measures with nice (smooth) densities, and we want to study the PDE they solve. The reason to focus on the case $p = 2$ can also be easily understood. Indeed, from the very beginning, i.e. from Sect. 2, we saw that the equation $x' = -\nabla F(x)$ corresponds to a sequence of minimization problems involving the squared[9] distance $|x - x_k^\tau|^2$, and in the Wasserstein space $\mathbb{W}_p$ the distance

---

[9] If we change the exponent we could consider

$$\min_x \ F(x) + \frac{1}{p} \cdot \frac{|x - x_k^\tau|^p}{\tau^{p-1}},$$

but this gives raise to the equation $x' = -|\nabla F(x)|^{q-2} \nabla F(x)$, where $q = p/(p - 1)$ is the conjugate exponent of p.

is defined as the power $1/p$ of a transport cost; only in the case $p = 2$ the exponent goes away and we are lead to a minimization problem involving $F(\varrho)$ and a transport cost of the form

$$\mathcal{T}_c(\varrho, \nu):= \min \left\{ \int c(x, y) \, d\gamma \; : \; \gamma \in \Pi(\varrho, \nu) \right\},$$

for $\nu = \varrho_k^\tau$.

In the particular case of the space $\mathbb{W}_2(\Omega)$, which has some additional structure, if compared to arbitrary metric spaces, we would like to give a PDE description of the curves that we obtain as gradient flows, and this will pass through the optimality conditions of the minimization problem (4.10). In order to study these optimality conditions, we introduce the notion of first variation of a functional. This will be done in a very sketchy and formal way (we refer to Chapter 7 in [84] for more details).

Given a functional $G : \mathcal{P}(\Omega) \to \mathbb{R}$ we call $\frac{\delta G}{\delta \varrho}(\varrho)$, if it exists, the unique (up to additive constants) function such that $\frac{d}{d\varepsilon} G(\varrho + \varepsilon \chi)_{|\varepsilon=0} = \int \frac{\delta G}{\delta \varrho}(\varrho) d\chi$ for every perturbation χ such that, at least for $\varepsilon \in [0, \varepsilon_0]$, the measure $\varrho + \varepsilon \chi$ belongs to $\mathcal{P}(\Omega)$. The function $\frac{\delta G}{\delta \varrho}(\varrho)$ is called *first variation* of the functional G at ϱ. In order to understand this notion, the easiest possibility is to analyze some examples.

The three main classes of examples are the following functionals[10]

$$\mathcal{F}(\varrho)= \int f(\varrho(x)) dx, \quad \mathcal{V}(\varrho) = \int V(x) d\varrho, \quad \mathcal{W}(\varrho)=\frac{1}{2} \int \int W(x - y) d\varrho(x) d\varrho(y),$$

where $f : \mathbb{R} \to \mathbb{R}$ is a convex superlinear function (and the functional F is set to $+\infty$ if ϱ is not absolutely continuous w.r.t. the Lebesgue measure for the semicontinuity of these functionals see, for instance, [24]) and $V : \Omega \to \mathbb{R}$ and $W : \mathbb{R}^d \to \mathbb{R}$ are regular enough (and W is taken symmetric, i.e. $W(z) = W(-z)$, for simplicity). In this case it is quite easy to realize that we have

$$\frac{\delta \mathcal{F}}{\delta \varrho}(\varrho) = f'(\varrho), \quad \frac{\delta \mathcal{V}}{\delta \varrho}(\varrho) = V, \quad \frac{\delta \mathcal{W}}{\delta \varrho}(\varrho) = W * \varrho.$$

It is clear that the first variation of a functional is a crucial tool to write optimality conditions for variational problems involving such a functional. In order to study the problem (4.10), we need to complete the picture by understanding the first variation of functionals of the form $\varrho \mapsto \mathcal{T}_c(\varrho, \nu)$. The result is the following:

Proposition 4.8 *Let* $c : \Omega \times \Omega \to \mathbb{R}$ *be a continuous cost function. Then the functional* $\varrho \mapsto \mathcal{T}_c(\varrho, \nu)$ *is convex, and its subdifferential at* ϱ_0 *coincides with the set of Kantorovich potentials* $\{\varphi \in C^0(\Omega) \; : \; \int \varphi \, d\varrho_0 + \int \varphi^c \, d\nu = \mathcal{T}_c(\varrho, \nu)\}$. *Moreover, if there is a unique c-concave Kantorovich potential* φ *from* ϱ_0 *to* ν *up to additive constants, then we also have* $\frac{\delta \mathcal{T}_c(\cdot, \nu)}{\delta \varrho}(\varrho_0) = \varphi$.

---

[10] Note that in some cases the functionals that we use are actually valued in $\mathbb{R} \cup \{+\infty\}$, and we restrict to a suitable class of perturbations χ which make the corresponding functional finite.

Even if a complete proof of the above proposition is not totally trivial (and Chapter 7 in [84] only provides it in the case where Ω is compact), one can guess why this is true from the following considerations. Start from Proposition 4.1, which provides

$$\mathcal{T}_c(\varrho, \nu) = \max_{\phi \in \Psi_c(X)} \int_{\Omega} \phi \, d\varrho + \int_{\Omega} \phi^c \, d\nu.$$

This expresses $\mathcal{T}_c$ as a supremum of linear functionals in ϱ and shows convexity. Standard considerations from convex analysis allow to identify the subdifferential as the set of functions φ attaining the maximum. An alternative point of view is to consider the functional $\varrho \mapsto \int \phi \, d\varrho + \int \phi^c \, d\nu$ for fixed ϕ, in which case the first variation is of course ϕ; then it is easy to see that the first variation of the supremum may be obtained (in case of uniqueness) just by selecting the optimal ϕ.

Once we know how to compute first variations, we come back to the optimality conditions for the minimization problem (4.10). Which are these optimality conditions? roughly speaking, we should have

$$\frac{\delta F}{\delta \varrho}(\varrho_{k+1}^\tau) + \frac{\varphi}{\tau} = const \tag{4.11}$$

(where the reasons for having a constant instead of 0 is the fact that, in the space of probability measures, only zero-mean densities are considered as admissible perturbations, and the first variations are always defined up to additive constants). Note that here φ is the Kantorovich potential associated with the transport from ϱ_{k+1}^τ to ϱ_k^τ (and not viceversa).

Let us look at the consequences we can get from this optimality condition. Actually, if we combine (4.11) with the fact that the optimal map T satisfies $T(x) = x - \nabla\varphi(x)$, we get

$$-\mathbf{v}(x) := \frac{T(x) - x}{\tau} = -\frac{\nabla\varphi(x)}{\tau} = \nabla\left(\frac{\delta F}{\delta \varrho}(\varrho)\right)(x). \tag{4.12}$$

We denoted by $-\mathbf{v}$ the ratio $\frac{T(x)-x}{\tau}$. Why? because, as a ratio between a displacement and a time step, it has the meaning of a velocity, but since it is the displacement associated to the transport from ϱ_{k+1}^τ to ϱ_k^τ, it is better to view it rather as a backward velocity (which justifies the minus sign).

Since here we have $\mathbf{v} = -\nabla\left(\frac{\delta F}{\delta \varrho}(\varrho)\right)$, this suggests that at the limit $\tau \to 0$ we will find a solution of

$$\partial_t \varrho - \nabla \cdot \left(\varrho \, \nabla\left(\frac{\delta F}{\delta \varrho}(\varrho)\right)\right) = 0. \tag{4.13}$$

This is a PDE where the velocity field in the continuity equation depends on the density ϱ itself.

We can see many interesting examples.

First, suppose $F(\varrho) = \int f(\varrho(x)) dx$, with $f(t) = t \log t$. In such a case we have $f'(t) = \log t + 1$ and $\nabla(f'(\varrho)) = \frac{\nabla\varrho}{\varrho}$: this means that the gradient flow equation associated to the functional F would be the *heat equation*

$$\partial_t \varrho - \Delta \varrho = 0.$$

Using $F(\varrho) = \int f(\varrho(x)) \, dx + \int V(x) \, d\varrho(x)$, we would have the *Fokker–Planck Equation*

$$\partial_t \varrho - \Delta \varrho - \nabla \cdot (\varrho \nabla V) = 0.$$

Changing the function f we can obtain other forms of diffusion. For instance, if one uses

$$F(\varrho) = \frac{1}{m-1} \int \varrho^m(x) \, dx + \int V(x) \varrho(x) \, dx,$$

the equation we obtain is

$$\partial_t \varrho - \Delta(\varrho^m) - \nabla \cdot (\varrho \nabla V) = 0.$$

When $m > 1$ this equation is called *Porous Media Equation* (see [89] for a complete monography) and models the diffusion of a fluid into a material whose porosity slows down the diffusion. Among the properties of this equation there is a finite-speed propagation effect, different from linear diffusion (if ϱ_0 is compactly supported the same stays true for ϱ_t, $t > 0$, while this is not the case for the heat equation). Note that linear diffusion can be obtained as a limit $m \to 1$ in the following way: in the functional F, we can replace the term $\frac{1}{m-1} \int \varrho^m(x) \, dx$ with $\frac{1}{m-1} \int (\varrho^m(x) - \varrho(x)) \, dx$, since they only differ by a constant term (the total mass of ϱ is fixed). Then, we just observe that we have

$$\lim_{m \to 1} \frac{\varrho^m - \varrho}{m-1} = \varrho \log \varrho,$$

which provides the entropy in the limit.

It is also interesting to consider the case $m < 1$: the function $\varrho^m - \varrho$ is concave, but the negative coefficient $1/(m-1)$ finally gives a convex function (which is, unfortunately, not superlinear at infinity, hence more difficult to handle). The PDE that we get as a gradient flow is called *Fast diffusion* equation, and has opposite properties in terms of diffusion rate than the porous medium one.

In the above equations, it is also possible to get rid of the diffusion part and just study the *advection equation*

$$\partial_t \varrho - \nabla \cdot (\varrho \nabla V) = 0,$$

where V is a given potential. This equation is the gradient flow of

$$F(\varrho) = \int V \, d\varrho.$$

It is not difficult to see that the solution of this equation, where particles do not interact with each other, is obtained in the following way: for every initial point x solve the

Euclidean gradient flow $y_x'(t) = -\nabla V(y_x(t))$ with $y_x(0) = x$, and call Y_t the map which associates with every initial point x the point $y_x(t)$. Then we have $\varrho_t = (Y_t)_{\#}\varrho_0$.

A more interesting case can be observed when the particles are advected by a potential determined by the superposition of many potentials, each created by one particle. For instance, given a function $W : \mathbb{R}^d \to \mathbb{R}$, the particle located at x produces a potential $W(\cdot - x)$ and, globally, the potential is given by $V(y) = \int W(y-x)\,\mathrm{d}\varrho(x)$, i.e. $V = W * \varrho$. The equation, if every particle follows $-\nabla V$ is

$$\partial_t \varrho - \nabla \cdot (\varrho\, ((\nabla W) * \varrho)) = 0,$$

where we used $\nabla(W * \varrho) = (\nabla W) * \varrho$. If W is even (i.e. the interaction between x and y is the same as between y and x), then this is the gradient flow of the functional

$$F(\varrho) = \frac{1}{2} \iint W(x - y)\,\mathrm{d}\varrho(x)\mathrm{d}\varrho(y).$$

When $W(z)$ is an increasing function of $|z|$, for instance in the quadratic case which gives $F(\varrho) = \iint |x - y|^2 \mathrm{d}\varrho(x)\mathrm{d}\varrho(y)$, the equation gives rise to a general aggregation behavior of the particles. When $t \to \infty$ one expects $\varrho_t \rightharpoonup \delta_{x_0}$ (the point x_0 depending on the initial datum ϱ_0: in the quadratic example above it is the barycenter of ϱ_0). If W is not smooth enough, the aggregation into a unique point can even occur in finite time, see [26].

Most often, the above aggregation energy is studied together with an internal energy and a confining potential energy, using the functional

$$F(\varrho) = \int f(\varrho(x))\,\mathrm{d}x + \int V(x)\,\mathrm{d}\varrho(x) + \frac{1}{2} \iint W(x - y)\,\mathrm{d}\varrho(x)\mathrm{d}\varrho(y). \quad (4.14)$$

This gives the following *diffusion–advection–interaction equation*

$$\partial_t \varrho - \nabla \cdot \left(\varrho \left[\nabla(f'(\varrho)) + \nabla V + (\nabla W) * \varrho \right] \right) = 0$$

(see in particular [27,28] for physical modelling and convergence results).

We will see in Sect. 4.6 some other variants of these equations. The reader can also look at Section 8.4.2 in [84]. For the moment, we just wanted to show how the iterated variational scheme called JKO is related to the different PDEs. In the next section we will some tools to prove the convergence of this scheme (without full details).

4.4 How to prove convergence of the JKO scheme

Many possible proofs can be built for the convergence of the JKO scheme. In particular, one could follow the general theory developed in [3], i.e. checking all the assumptions to prove existence and uniqueness of an EVI gradient flow for the functional F in the space $\mathbb{W}_2(\Omega)$, and then characterizing the velocity field that Theorem 4.6 associates with the curve obtained as a gradient flow. In [3], it is proven, under suitable conditions,

that such a vector field $\mathbf{v}_t$ must belong to what is defined as the *Wasserstein sub-differential* of the functional F, provided in particular that F is λ-convex. Then, the *Wasserstein sub-differential* is proven to be of the desired form (i.e. composed only of the gradient of the first variation of F, when F admits a first variation).

This approach has the advantage to use a general theory and to adapt it to the scopes of this particular setting. On the other hand, some considerations seem necessary:

- the important point when studying these PDEs is that the curves $(\varrho_t)_t$ obtained as a limit are true weak solutions of the continuity equations; from this point of view, the notions of EDE and EVI solutions and the formalism developed in the first part of the book [3] (devoted to the general metric case) are not necessary; if the second part of [3] is exactly concerned with Wasserstein spaces and with the characterization of the limit as $\tau \to 0$ as the solution of a PDE, we can say that the whole formalism is sometimes too heavy.
- after using optimal transport theory to select a suitable distance in the discrete scheme above and a suitable interpolation, the passage to the limit can be done by classical compactness techniques in the space of measures and in functional analysis; on the other hand, there are often some difficulties in handling some non-linear terms, which are not always seen when using the theory of [3] (which is an advantage of this general theory).
- the λ-convexity assumption is in general not crucial in what concerns existence (but the general theory in [3] has been built under this assumption, as we saw in Sect. 3).
- as far as uniqueness of the solution is concerned, the natural point of view in the applications would be to prove uniqueness of the weak solution of the equation (or, possibly, to define a more restrictive notion of solution of the PDE for which one can prove uniqueness), and this is a priori very different from the EDE or EVI notions. To this aim, the use of the Wasserstein distance can be very useful, as one can often prove uniqueness by differentiating in time the distance between two solutions, and maybe apply a Gronwall lemma (and this can be done independently of the EVI notion; see for instance the end of Sect. 4.5). On the other hand, uniqueness results are almost never possible without some kind of λ-convexity (or weaker versions of it, as in [34]) of the functional.

Yet, we prefer here to present the ideas to build a more concrete proof, following for instance what is done in Chapter 8 of [84]. For simplicity, we will describe the construction in the model case where F is given by (4.14).

As we said, the final goal is to produce a curve $(\varrho_t)_t$ which is a solution (in the distributional sense on $\mathbb{R}^d$, which is the same as imposing suitable no-flux boundary conditions on $\partial\Omega$) of the PDE

$$\partial_t \varrho_t + \nabla \cdot (\varrho_t \mathbf{v}_t) = 0, \quad \text{where we require } \mathbf{v}_t = -\nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right).$$

We will set $E = \varrho \mathbf{v}$ (the variable E is called *momentum*, in physics) and require at least that E is a finite vector measure over $\Omega \times [0, T]$, acting on test functions ϕ

via $\langle E, \phi \rangle := \int dt \int \phi(t, x) \cdot \mathbf{v}_t \, d\varrho_t$. This condition is equivalent to $\int_0^T ||\mathbf{v}_t||_{L^1(\varrho_t)} dt < +\infty$.

We start from the discrete scheme (4.10), which defines a sequence $(\varrho_k^\tau)_k$. We also define a sequence of velocities $\mathbf{v}_k^\tau = (id - T)/\tau$, taking as T the optimal transport from ϱ_k^τ to ϱ_{k-1}^τ. The considerations in the previous sections guarantee that we have

$$\mathbf{v}_k^\tau = -\nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_k^\tau) \right).$$

This is proven thanks to the optimality conditions in (4.10) and the only delicate point is to guarantee the uniqueness of the Kantorovich potential in the transport from ϱ_k^τ to ϱ_{k-1}^τ. In some cases it is possible to prove a priori that the minimizers of $F(\varrho) + W_2^2(\varrho, \nu)$ have strictly positive density a.e. whatever is ν (this is typically true when $f'(0) = -\infty$, as it is the case for the entropy functional, and more generally in case of linear or fast diffusion), which provides uniqueness up to additive constants of the potential. In Section. 8.3 of [84] full details are provided for the Fokker–Planck case, and it is indeed proved that the minimizers of the JKO scheme are strictly positive a.e., in this case. When positivity of the minimizer is not available, then one can obtain the same optimality conditions by first approximating ν with strictly positive densities, and then passing to the limit.[11]

Then, we build at least two interesting curves in the space of measures:

- first we can define some piecewise constant curves, i.e. $\overline{\varrho}_t^\tau := \varrho_{k+1}^\tau$ for $t \in]k\tau, (k+1)\tau]$; associated with this curve we also define the velocities $\overline{\mathbf{v}}_t^\tau = \mathbf{v}_{k+1}^\tau$ for $t \in]k\tau, (k+1)\tau]$ and the momentum variable $\overline{E}^\tau = \overline{\varrho}^\tau \overline{\mathbf{v}}^\tau$;
- then, we can also consider the densities $\widehat{\varrho}_t^\tau$ that interpolate the discrete values $(\varrho_k^\tau)_k$ along geodesics:

$$\widehat{\varrho}_t^\tau = \left(id - (k\tau - t)\mathbf{v}_k^\tau\right)_\sharp \varrho_k^\tau, \quad \text{for } t \in](k-1)\tau, k\tau[; \qquad (4.15)$$

the velocities $\widehat{\mathbf{v}}_t^\tau$ are defined so that $(\widehat{\varrho}^\tau, \widehat{\mathbf{v}}^\tau)$ satisfy the continuity equation, taking

$$\widehat{\mathbf{v}}_t^\tau = \mathbf{v}_t^\tau \circ \left(id - (k\tau - t)\mathbf{v}_k^\tau\right)^{-1};$$

moreover, as before, we define: $\widehat{E}^\tau = \widehat{\varrho}^\tau \widehat{\mathbf{v}}^\tau$.

After these definitions we look for a priori bounds on the curves and the velocities that we defined. We already know that we have

$$\sum_k \tau \left(\frac{W_2(\varrho_k^\tau, \varrho_{k-1}^\tau)}{\tau} \right)^2 \leq C, \qquad (4.16)$$

---

[11] Note that, for fixed $\tau > 0$, the minimizers usually satisfy strong regularity estimates, and the convergence of the minimizers is strong enough to pass the optimality condition to the limit, obtaining "there exists a Kantorovich potential φ such that $\delta F/\delta \varrho + \varphi/\tau$ is constant".

which is the discrete version of an H^1 estimate in time. As for $\widehat{\varrho}_t^\tau$, it is an absolutely continuous curve in the Wasserstein space and its velocity on the time interval $[(k-1)\tau, k\tau]$ is given by the ratio $W_2(\varrho_{k-1}^\tau, \varrho_k^\tau)/\tau$. Hence, the L^2 norm of its velocity on $[0, T]$ is given by

$$\int_0^T |(\widehat{\varrho}^\tau)'|^2(t)\mathrm{d}t = \sum_k \frac{W_2^2(\varrho_k^\tau, \varrho_{k-1}^\tau)}{\tau}, \tag{4.17}$$

and, thanks to (4.16), it admits a uniform bound independent of τ. In our case, thanks to results on the continuity equation and the Wasserstein metric, this metric derivative is also equal to $||\widehat{\mathbf{v}}_t^\tau||_{L^2(\widehat{\varrho}_t^\tau)}$. This gives compactness of the curves $\widehat{\varrho}^\tau$, as well as Hölder estimates (since $H^1 \subset C^{0,1/2}$). The characterization of the velocities $\overline{\mathbf{v}}^\tau$ and $\widehat{\mathbf{v}}^\tau$ allows to deduce bounds on these vector fields from the bounds on $W_2(\varrho_{k-1}^\tau, \varrho_k^\tau)/\tau$.

Considering all these facts, one obtains the following situation (see also [37,82]):

- The norm $\int ||\overline{\mathbf{v}}_t^\tau||_{L^2(\overline{\varrho}_t^\tau)}^2 \mathrm{d}t$ is τ-uniformly bounded. This quantity is equal to $\mathcal{B}_2(\overline{\varrho}^\tau, \overline{E}^\tau)$.
- In particular, the bound is valid in L^1 as well, which implies that $\overline{E}^\tau$ is bounded in the space of measures over $[0, T] \times \Omega$.
- The very same estimates are true for $\widehat{\mathbf{v}}^\tau$ and $\widehat{E}^\tau$.
- The curves $\widehat{\varrho}^\tau$ are bounded in $H^1([0, T], \mathbb{W}_2(\Omega))$ and hence compact in $C^0([0, T], \mathbb{W}_2(\Omega))$.
- Up to a subsequence, one has $\widehat{\varrho}^\tau \to \varrho$, as $\tau \to 0$, uniformly according to the W_2 distance.
- From the estimate $W_2(\overline{\varrho}_t^\tau, \widehat{\varrho}_t^\tau) \leq C\tau^{1/2}$ one gets that ϱ^τ converges to the same limit ϱ in the same sense.
- If we denote by E a weak limit of $\widehat{E}^\tau$, since $(\widehat{\varrho}^\tau, \widehat{E}^\tau)$ solves the continuity equation, by linearity, passing to the weak limit, also (ϱ, E) solves the same equation.
- It is possible to prove (see Section 8.3 in [84]) that the weak limits of $\widehat{E}^\tau$ and $\overline{E}^\tau$ are the same.
- From the semicontinuity of $\mathcal{B}_2$ and the bound $\mathcal{B}_2(\overline{\varrho}^\tau, \overline{E}^\tau) \leq C$, one gets $\mathcal{B}_2(\varrho, E) < +\infty$, which means that E is absolutely continuous w.r.t. ϱ and has an L^2 density, so that we have for a.e. time t a measure E_t of the form $\varrho_t \mathbf{v}_t$.
- It is only left to prove that one has $\mathbf{v}_t = -\nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right) \varrho_t$-a.e. and for a.e. t. This means proving

$$\overline{E}^\tau = -\overline{\varrho}^\tau \nabla \left(\frac{\delta F}{\delta \varrho}(\overline{\varrho}^\tau) \right) \Rightarrow E = -\varrho \nabla \left(\frac{\delta F}{\delta \varrho}(\varrho) \right)$$

as a limit as $\tau \to 0$. It is crucial in this step to consider the limit of $(\overline{\varrho}^\tau, \overline{E}^\tau)$ instead of $(\widehat{\varrho}^\tau, \widehat{E}^\tau)$.

This last step is the most critical one in many cases. It requires passing to the limit (in the sense of distributions) the terms involving $\frac{\delta F}{\delta \varrho}$ on a sequence $\varrho_j \rightharpoonup \varrho$ (we don't care here where this sequence comes from). We can see what happens with the main class of functionals that we introduced so far.

The easiest case is that of the functional $\mathcal{V}$: we have

$$\varrho \nabla \left(\frac{\delta \mathcal{V}}{\delta \varrho} (\varrho) \right) = \varrho \nabla V.$$

This term is linear in ϱ and $\varrho_j \rightharpoonup \varrho$ obviously implies $\varrho_j \nabla V \rightharpoonup \varrho \nabla V$ as soon as $V \in C^1$ (so that ∇V is a continuous function, which is exactly the functional space whose dual is the space of measures). In case $\nabla V \notin C^0$ it is possible to handle this term as soon as suitable bounds on ϱ_j provide a better weak convergence (for instance, if the functional F is composed of $\mathcal{V}$ plus other terms which impose bounds on the L^p norm of ϱ, then it is enough to suppose $\nabla V \in L^{p'}$).

Another easy case is that of the functional $\mathcal{W}$: we have

$$\varrho \nabla \left(\frac{\delta \mathcal{W}}{\delta \varrho} (\varrho) \right) = \varrho \, (\nabla W) * \varrho.$$

If we take a sequence $\varrho_j \rightharpoonup \varrho$ and a vector test function $\xi \in C_c^\infty$, then we have

$$\int \xi \cdot (\nabla W) * \varrho_j \, d\varrho_j \to \int \xi \cdot (\nabla W) * \varrho \, d\varrho$$

as soon as $W \in C^1$, because ξ is fixed and $(\nabla W) * \varrho_j$ uniformly converges to $(\nabla W) * \varrho$ (this is a typical regularization effect of the convolution). As before, if W is less regular, it is possible to compensate this with stronger bounds on ϱ_j.

The most difficult case is that of the functional $\mathcal{F}$. In the case of the entropy $\mathcal{F}(\varrho) = \int f(\varrho)$ with $f(t) = t \log t$ then everything works fine because, again, the corresponding term is linear:

$$\varrho \nabla \left(\frac{\delta \mathcal{F}}{\delta \varrho} (\varrho) \right) = \varrho \frac{\nabla \varrho}{\varrho} = \nabla \varrho.$$

Then, the convergence of this term in the sense of distributions is straightforward when $\varrho_j \to \varrho$. By the way, the entropy term $\mathcal{F}$ is also enough to obtain suitable bounds to handle V or W which are only Lipschitz, as in this case we need to turn the weak convergence $\varrho_j \rightharpoonup \varrho$ from the sense of measures to the L^1 sense, which is exactly possible because the superlinear bound $\int \varrho_j \log(\varrho_j) \le C$ provides equi-integrability for ϱ_j.

Yet, for other functions f, there is no more linearity, and we need stronger bounds. For instance, for $f(t) = t^m / (m-1)$, we have

$$\varrho \nabla \left(\frac{\delta \mathcal{F}}{\delta \varrho} (\varrho) \right) = \varrho \frac{m}{m-1} \nabla (\varrho^{m-1}) = \nabla (\varrho^m).$$

This means that weak convergence of ϱ_j is no more enough, but one also needs weak convergence of ϱ^m. This could be possible to obtain if one had strong convergence, i.e. stronger bounds (if possible, Sobolev bounds on ϱ_j).

The main ingredient is indeed an H^1 bound in space, which comes from the fact that we have

$$\int_0^T \int_\Omega |\nabla(\varrho^{m-1/2})|^2 \, dx dt \approx \int_0^T \int_\Omega |\nabla(\varrho^{m-1})|^2 \, d\varrho_t dt \approx \int_0^T \int_\Omega \frac{|\nabla\varphi|^2}{\tau^2} \, d\varrho_t dt$$

$$= \int_0^T |\varrho'|(t)^2 dt \le C. \tag{4.18}$$

Two delicate points arise: first, this is not a full H^1 bound, in the sense that it is only concerned with space derivatives, and not with time derivatives and, second, we presented here an estimate on $||\varrho^{m-1/2}||_{H^1}$, but this could not be enough to prove strong convergence for ρ^m. The first point may be handled via some variants of the so-called *Aubin–Lions* lemma (see [10]), which interpolates between compactness in space and in time (roughly speaking: in our situation we have an L^2 bound in time with values in a strong space norm, H^1, but also a stronger bound in time, H^1, valued in the W_2 distance, which is weaker as it metrizes weak convergence; this information together can provide strong compactness). For the second point, finer tools on the JKO scheme, together with stronger assumptions on the initial data, can be used. For instance, it is possible to prove L^∞ bounds on ρ if $\rho_0 \in L^\infty$ and $V \in C^2$ (see Section 7.4.1 in [84]), which is enough to transform the H^1 estimates on $\varrho^{m-1/2}$ into estimates on ϱ^m. Another possibility is the use of estimates obtained via a technique called *flow interchange* (where the optimality of a density ϱ in the JKO scheme is tested against its evolution via the gradient flow of another functional, see [66]), which allows to modify the exponent of ϱ in (4.18).

4.5 Geodesic convexity in $\mathbb{W}_2$

Even if we insisted on the fact that the most natural approach to gradient flows in $\mathbb{W}_2$ relies on the notion of weak solutions to some PDEs and not on the EDE/EVI formulations, surely it could be interesting to check whether the general theory of Sect. 3 could be applied to some model functionals on the space of probabilities, such as $\mathcal{F}$, $\mathcal{V}$ or $\mathcal{W}$. This requires to discuss their λ-convexity, which is also useful because it can provide uniqueness results. As we now know the geodesics in the Wasserstein space, the question of which functionals are geodesically convex is easy to tackle. The notion of geodesic convexity in the $\mathbb{W}_2$ space, also called *displacement convexity*, has first been introduced by McCann in [72].

Displacement convexity of $\mathcal{F}$, $\mathcal{V}$ and $\mathcal{W}$ It is not difficult to check that the convexity of V is enough to guarantee geodesic convexity of $\mathcal{V}$, since

$$\mathcal{V}(\mu_t) = \int V \, d((1-t)id + tT)_\# \mu = \int V\big((1-t)x + tT(x)\big) \, d\mu,$$

as well as the convexity of W guarantees that of $\mathcal{W}$:

$$
\begin{aligned}
\mathcal{W}(\mu_t) &= \int W(x-y) \, d\left(\left((1-t)id+tT\right)_{\#}\mu\right) \otimes \left(\left((1-t)id+tT\right)_{\#}\mu\right)(x,y) \\
&= \int W\left((1-t)x+tT(x), (1-t)y+tT(y)\right) d(\mu \otimes \mu).
\end{aligned}
$$

Similarly, if V or W are λ-convex we get λ-geodesical convexity. Note that in the case of $\mathcal{V}$ it is easy to see that the λ-convexity of V is also necessary for the λ-geodesical convexity of $\mathcal{V}$, while the same is not true for W and $\mathcal{W}$.

The most interesting displacement convexity result is the one for functionals depending on the density. To consider these functionals, we need some technical facts.

The starting point is the computation of the density of an image measure, via standard change-of-variable techniques: if $T : \Omega \to \Omega$ is a map smooth enough[12] and injective, and $\det(DT(x)) \neq 0$ for a.e. $x \in \Omega$, if we set $\varrho_T := T_{\#}\varrho$, then ϱ_T is absolutely continuous with density given by

$$
\varrho_T = \frac{\varrho}{\det(DT)} \circ T^{-1}.
$$

Then, we underline an interesting computation, which can be proven as an exercise.

Lemma 4.9 *Let A, B be two $d \times d$ symmetric and positive-definite matrices. Then $[0,1] \ni t \mapsto g(t) := \det((1-t)A + tB)^{1/d}$ is a concave function*

We can now state the main theorem.

Theorem 4.10 *Suppose that $f(0) = 0$ and that $s \mapsto s^d f(s^{-d})$ is convex and decreasing. Suppose that Ω is convex. Then $\mathcal{F}$ is geodesically convex in $\mathbb{W}_2$.*

Proof Take two measures μ_0, μ_1 with $\mathcal{F}(\mu_0), \mathcal{F}(\mu_1) < +\infty$ (which implies that they are absolutely continuous). There is a unique constant-speed geodesic μ_t between them, given by $\mu_t = (T_t)_{\#}\mu_0$, where $T_t = id + t(T - id)$. Note that we have $T_t(x) = x - t\nabla\varphi(x)$, where φ is such that $\frac{x^2}{2} - \varphi$ is convex. This implies that $D^2\varphi \leq I$ and T_t is, for $t < 1$, the gradient of a strictly convex function, hence it is injective. Moreover $\nabla\varphi$ is countably Lipschitz, and so is T_t. From the formula for the density of the image measure, we know that μ_t is absolutely continuous and we can write its density ϱ_t as $\varrho_t(y) = \varrho(T_t^{-1}(y)) / \det(I + tM(T_t^{-1}(y)))$, where $M = -D^2\varphi$ and ϱ is the density of μ, and hence

$$
\mathcal{F}(\mu_t) = \int f\left(\frac{\varrho(T_t^{-1}(y))}{\det(I + tM(T_t^{-1}(y)))}\right) dy = \int f\left(\frac{\varrho(x)}{\det(I + tM(x))}\right) \det(I + tA(x)) \, dx,
$$

---

[12] We need at least T to be countably Lipschitz, i.e. Ω may be written as a countable union of measurable sets $(\Omega_i)_{i \geq 0}$ with Ω_0 negligible and $T \llcorner \Omega_i$ Lipschitz continuous for every $i \geq 1$.

where we used the change of variables $y = T_t(x)$ and $\mathrm{d}y = \det DT_t(x)\,\mathrm{d}x = \det(\mathrm{I} + tM(x))\,\mathrm{d}x$.

Lemma 4.9 applied to $A = \mathrm{I}$ and $B = \mathrm{I} + M$ implies $\det(\mathrm{I} + tM(x)) = g(t,x)^d$ for a function $g : [0,1] \times \Omega$, concave in t. The composition of a convex and decreasing function with a concave one gives a convex function, which implies that

$$t \mapsto f\left(\frac{\varrho(x)}{g(t,x)^d}\right) g(t,x)^d$$

is convex. Finally, we proved convexity of $t \mapsto \mathcal{F}(\mu_t)$. $\square$

Remark 4.1 Note that the assumption that $s \mapsto s^d f(s^{-d})$ is convex and decreasing implies that f itself is convex (the reader can check it as an exercise), a property which can be useful to establish, for instance, lower semicontinuity of $\mathcal{F}$.

Here are some examples of convex functions satisfying the assumptions of Theorem 4.10:

- the entropy function $f(t) = t \log t$ satisfies these assumptions: $s^d f(s^{-d}) = -d \log s$ is convex and decreasing;
- all the power functions $f(t) = t^q$ for $q > 1$ also satisfy the same assumptions, since $s^d f(s^{-d}) = s^{-d(q-1)}$ is convex and decreasing;
- the function $f(t) = -t^m$ is convex for $m < 1$, and if we compute $s^d f(s^{-d}) = -t^{m(1-d)}$ we get a convex and decreasing function as soon as $1 - \frac{1}{d} \leq m < 1$. Note that in this case f is not superlinear, which requires some attention for the semicontinuity of $\mathcal{F}$.

Finally, it is interesting to observe that the geodesic convexity of higher-order functionals such as $\varrho \mapsto \int |\nabla \varrho|^p$ generally fails, or is a very delicate matter, while these functionals are among the most standard examples of convex functionals in the usual sense. See [29] for some examples of first-order geodesically convex functionals (in dimension one).

Convexity on generalized geodesics When studying displacement convexity a bad surprise arrives when considering the functional $\mu \mapsto W_2^2(\mu, \nu)$. This functional is not, in general, displacement convex, contrary to the intuition stipulating that squared distances should be nice convex functions.[13] However, we can see that this fails from the following easy example. Choose $a, b, c \in \mathbb{R}^2$ to be the vertices of an equilateral triangle: $a = (0,0)$, $b = (1,0)$ and $c = (1/2, \sqrt{3}/2)$. Set $\nu = \frac{1}{2}\delta_a + \frac{1}{2}\delta_b$ and $\mu_0 = \frac{1}{2}\delta_a + \frac{1}{2}\delta_c$ and $\mu_0 = \frac{1}{2}\delta_b + \frac{1}{2}\delta_c$. It is not difficult to see that the geodesic between μ_0 and μ_1 is given by $\mu_t = \frac{1}{2}\delta_{(t,0)} + \frac{1}{2}\delta_c$ and that

$$W_2^2(\mu_t, \nu) = \frac{1}{2} + \frac{1}{2}\min\left\{(1-t)^2, t^2\right\}$$

---

[13] Actually, this is true in normed spaces, but not even true in Riemannian manifolds, as it depends on curvature properties.

Fig. 2 An example where $t \mapsto W_2^2(\mu_t, \nu)$ is not convex
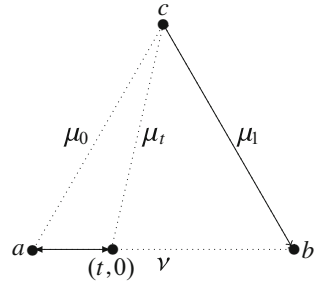
(this is computed by observing that in the transport between μ_t and ν we can send the atom at $(t, 0)$ to the closest atom of ν and the atom at c to the other one, and that we cannot do better, see Fig. 2). But this function is not convex!

The lack of geodesic convexity of this easy functional[14] is a problem for many issues, and in particular for the C^2G^2 condition, and an alternate notion has been proposed, namely that of convexity along generalized geodesics.

Definition 4.3 Given an absolutely continuous probability $\varrho \in \mathcal{P}(\Omega)$, for every pair $\mu_0, \mu_1 \in \mathcal{P}(\Omega)$ we call *generalized geodesic* in $\mathbb{W}_2(\Omega)$ between μ_0 and μ_1 with base ϱ the curve $\mu_t = ((1 - t)T_0 + tT_1)_{\#}\varrho$, where T_i is the optimal transport map (for the cost $|x - y|^2$) from ϱ to μ_i ($i = 0, 1$).

It is clear that $t \mapsto W_2^2(\mu_t, \varrho)$ satisfies

$$
\begin{aligned}
W_2^2(\mu_t, \varrho) &\leq \int |((1 - t)T_0(x) + tT_1(x)) - x|^2 \mathrm{d}\varrho(x) \\
&\leq (1 - t)\int |T_0(x) - x|^2 \mathrm{d}\varrho(x) + t\int |T_1(x) - x|^2 \mathrm{d}\varrho(x) \\
&= (1 - t)W_2^2(\mu_0, \varrho) + tW_2^2(\mu_1, \varrho).
\end{aligned}
$$

This provides the desired convexity along the curve μ_t. Moreover, it is possible, by using similar arguments as those already used in this section, to show that all the functionals $\mathcal{V}, \mathcal{W}$ and $\mathcal{F}$ are also convex along generalized geodesics under the same assumptions. For the case of functionals $\mathcal{V}$ and $\mathcal{W}$ this is easy, while for the case of the functional $\mathcal{F}$ we use again Lemma 4.9 (without restricting to the case $A = \mathrm{I}$). We do not develop these proofs here, and we refer to [3] or Chapter 7 in [84] for more details.

Of course, we could wonder whether the assumption C^2G^2 is satisfied in the Wasserstein space $\mathbb{W}_2(\Omega)$ for these functionals $\mathcal{F}, \mathcal{V}$ and $\mathcal{W}$: actually, if one looks closer at this questions, it is possible to guess that the very definition of C^2G^2 has been given on purpose in order to face the case of Wasserstein spaces. Indeed, if we fix $\nu \in \mathcal{P}(\Omega)$ and take μ_0, μ_1 two other probabilities, with T_0, T_1 the optimal transports from ν to μ_0 and μ_1, respectively, then the curve

---

[14] By the way, this functional can even be proven to be somehow geodesically concave, as it is shown in [3], Theorem 7.3.2.

$$\mu_t := ((1-t)T_0 + tT_1)_\# \nu \tag{4.19}$$

connects μ_0 to μ_1 and can be used in C^2G^2.

Displacement convexity and curvature conditions In the proof of the geodesic convexity of the functional $\mathcal{F}$ we strongly used the Euclidean structure of the geodesics in the Wasserstein space. The key point was the form of the intermediate map T_t: a linear interpolation between the identity map id and the optimal map T, together with the convexity properties of $t \mapsto \det(I + tA)^{1/d}$. On a Riemannian manifold, this would completely change as the geodesic curves between x and $T(x)$, which are no more segments, could concentrate more or less than what happens in the Euclidean case, depending on the curvature of the manifold (think at the geodesics on a sphere connecting points close to the North Pole to points close to the South Pole: they are much farther from each other at their middle points than at their endpoints). It has been found (see [79]) that the condition of λ-geodesic convexity of the Entropy functional $\varrho \mapsto \int \varrho \log(\varrho) \mathrm{d}\mathrm{Vol}$ (where ϱ is absolutely continuous w.r.t. the volume measure Vol and densities are computed accordingly) on a manifold characterizes a lower bound on its Ricci curvature:

Proposition 4.11 *Let M be a compact manifold of dimension d and* Vol *its volume measure. Let $\mathcal{E}$ be the entropy functional defined via $\mathcal{E}(\varrho) = \int \varrho \log \varrho \, \mathrm{d}\mathrm{Vol}$ for all measures $\varrho \ll$ Vol (set to $+\infty$ on non-absolutely continuous measures). Then $\mathcal{E}$ is λ-geodesically convex in the Wasserstein space $\mathbb{W}_2(M)$ if and only if the Ricci curvature Ric_M satisfies $\mathrm{Ric}_M \geq \lambda$. In the case $\lambda = 0$, the same equivalence is true if one replaces the entropy function $f(t) = t \log t$ with the function $f_N(t) = -t^{1-1/N}$ with $N \geq d$.*

This fact will be the basis (we will see it in Sect. 5) of a definition based on optimal transport of the notion of Ricci curvature bounds in more abstract spaces, a definition independently proposed in two celebrated papers by Sturm [87,88] and Lott and Villani [65].

Remark 4.2 The reader who followed the first proofs of this section has surely observed that it is easy, in the space $\mathbb{W}_2(\Omega)$, to produce λ-geodesically convex functionals which are not geodesically convex (with $\lambda < 0$, of course), of the form $\mathcal{V}$ (just take a λ-convex function V which is not convex), but that Theorem 4.10 only provides geodesic convexity (never provides λ-convexity without convexity) for functionals of the form $\mathcal{F}$: this is indeed specific to the Euclidean case, where the optimal transport has the form $T(x) = x - \nabla\varphi(x)$; in Riemannian manifolds or other metric measure spaces, this can be different!

Geodesic convexity and uniqueness of gradient flows The fact that λ-convexity is a crucial tool to establish uniqueness and stability results in gradient flows is not only true in the abstract theory of Sect. 3, where we saw that the EVI condition (intimately linked to λ-convexity) provides stability. Indeed, it can also be observed in the concrete case of gradient flows in $\mathbb{W}_2(\Omega)$, which take the form of the PDE (4.13). We will see this fact via some formal considerations, starting from an interesting lemma:

Lemma 4.12 *Suppose that* $F : \mathbb{W}_2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ *is* λ*-geodesically convex. Then, for every* $\varrho_0, \varrho_1 \in \mathcal{P}(\Omega)$ *for which the integrals below are well-defined, we have*

$$\int \nabla\varphi \cdot \nabla\left(\frac{\delta F}{\delta \varrho}(\varrho_0)\right) d\varrho_0 + \int \nabla\psi \cdot \nabla\left(\frac{\delta F}{\delta \varrho}(\varrho_1)\right) d\varrho_1 \geq \lambda W_2^2(\varrho_0, \varrho_1),$$

where φ *is the Kantorovich potential in the transport from* ϱ_0 *to* ϱ_1 *and* $\psi = \varphi^c$ *is the Kantorovich potential from* ϱ_1 *to* ϱ_0.

Proof Let ϱ_t be the geodesic curve in $\mathbb{W}_2(\Omega)$ connecting ϱ_0 to ϱ_1 and $g(t) := F(\varrho_t)$. The assumption of λ-convexity means (see the definition in (3.2))

$$g(t) \leq (1-t)g(0) + tg(1) - \frac{\lambda}{2}t(1-t)W_2^2(\varrho_0, \varrho_1).$$

Since the above inequality is an equality for $t = 0, 1$, we can differentiate it at these two points thus obtaining

$$g'(0) \leq g(1) - g(0) - \frac{\lambda}{2}W_2^2(\varrho_0, \varrho_1), \quad g'(1) \geq g(1) - g(0) + \frac{\lambda}{2}W_2^2(\varrho_0, \varrho_1),$$

which implies

$$g'(0) - g'(1) \leq -\lambda W_2^2(\varrho_0, \varrho_1).$$

Then, we compute the derivative of g, formally obtaining

$$g'(t) = \int \frac{\delta F}{\delta \varrho}(\varrho_t)\partial_t\varrho_t = -\int \frac{\delta F}{\delta \varrho}(\varrho_t)\nabla \cdot (\mathbf{v}_t\varrho_t) = \int \nabla\left(\frac{\delta F}{\delta \varrho}(\varrho_t)\right) \cdot \mathbf{v}_t \, d\varrho_t,$$

and, using $\mathbf{v}_0 = -\nabla\varphi$ and $\mathbf{v}_1 = \nabla\psi$, we obtain the claim. □

With the above lemma in mind, we consider two curves ϱ_t^0 and ϱ_t^1, with

$$\partial_t\varrho_t^i + \nabla \cdot (\varrho_t^i\mathbf{v}_t^i) = 0 \quad \text{for } i = 0, 1,$$

where the vector fields $\mathbf{v}_t^i$ are their respective velocity fields provided by Theorem 4.6. Setting $d(t) := \frac{1}{2}W_2^2(\varrho_t^0, \varrho_t^1)$, it is natural to guess that we have

$$d'(t) = \int \nabla\varphi \cdot \mathbf{v}_t^0 d\varrho_t^0 + \int \nabla\psi \cdot \mathbf{v}_t^1 d\varrho_t^1,$$

where φ is the Kantorovich potential in the transport from ϱ_t^0 to ϱ_t^1 and $\psi = \varphi^c$. Indeed, a rigorous proof is provided in [3] or in Section 5.3.5 of [84], but one can guess it from the duality formula

$$d(t) = \max\left\{\int \phi \, d\varrho_t^0 + \int \psi \, d\varrho_t^1 \ : \ \phi(x) + \psi(y) \leq \frac{1}{2}|x - y|^2\right\}.$$

As usual, differentiating an expression written as a max involves the optimal functions ϕ, ψ in such a max, and the terms $\partial_t \varrho_t^i$ have been replaced by $-\nabla \cdot (\varrho_t^i \mathbf{v}_t^i)$ as in the proof Lemma 4.12.

When the two curves ϱ_t^0 and ϱ_t^1 are solutions of (4.13) we have $\mathbf{v}_t^i = -\nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t^i) \right)$, and Lemma 4.12 allows to obtain the following:

Proposition 4.13 *Suppose that $F : \mathbb{W}_2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is λ-geodesically convex and that the two curves ϱ_t^0 and ϱ_t^1 are solutions of (4.13). Then, setting $d(t) := \frac{1}{2} W_2^2(\varrho_t^0, \varrho_t^1)$, we have*

$$d'(t) \leq -2\lambda d(t).$$

This implies uniqueness of the solution of (4.13) for fixed initial datum, stability, and exponential convergence as $t \to +\infty$ if $\lambda > 0$.

4.6 Other gradient-flow PDEs and variants

In Sect. 4.3 we saw some examples of evolution PDEs that can be obtained as gradient flows for suitable functionals F in the Wasserstein space of probability measures on a domain Ω, endowed with the distance W_2. Here we would like to give some more examples and present some variants (the reader can also look at Section 8.4.2 in [84]).

Keller–Segel A model in mathematical biology, first introduced in [56,57], has received much attention due to its interesting mathematical properties: consider a population ϱ of bacteria, which move advected by a potential and subject to diffusion. The potential is given by the concentration u of a nutrient, produced by the bacteria themselves, which attracts them by a chemical signal (this is why it is called *chemo-attractant*, and this kind of phenomenon is also known under the name of *chemotaxis*). More precisely, the drift is given by the gradient ∇u, but the distribution of u depends on their density ϱ. The easiest model has linear diffusion and supposes that u is related to ϱ by the condition $-\Delta u = \varrho$, with Dirichlet boundary conditions $u = 0$ on $\partial \Omega$. This gives the system

$$\begin{cases} \partial_t \varrho + \alpha \nabla \cdot (\varrho \nabla u) - \Delta \varrho = 0, \\ -\Delta u = \varrho, \\ u = 0 \text{ on } \partial\Omega, \ \varrho(0, \cdot) = \varrho_0, \ \partial_\mathbf{n} \varrho - \varrho \partial_\mathbf{n} u = 0 \text{ on } \partial\Omega. \end{cases} \tag{4.20}$$

The parameter α tunes the attraction intensity of bacteria towards the chemo-attractant.

This system is also studied in the whole space, where (to impose suitable decay at infinity), the function u is defined through the explicit formula

$$u(x) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} \log(|x - y|)\varrho(y) \, \mathrm{d}y. \tag{4.21}$$

One can see that System (4.20) is the gradient flow of the functional

$$F(\varrho) = \int_\Omega \varrho \log \varrho - \frac{1}{2} \int_\Omega |\nabla u_\varrho|^2, \quad \text{where } u_\varrho \in H_0^1(\Omega) \text{ solves } -\Delta u_\varrho = \varrho.$$

To check this fact, we only need to compute the first variation of the Dirichlet term $-\frac{1}{2} \int |\nabla u_\varrho|^2$. Suppose $\varrho_\varepsilon = \varrho + \varepsilon \chi$ and set $u_{\varrho+\varepsilon\chi} = u_\varrho + \varepsilon u_\chi$. Then

$$\frac{d}{d\varepsilon} \left(-\frac{1}{2} \int |\nabla u_{\varrho+\varepsilon\chi}|^2 \right)_{|\varepsilon=0} = -\int \nabla u_\varrho \cdot \nabla u_\chi = \int u_\varrho \Delta u_\chi = -\int u_\varrho \chi.$$

Note that, due to the wrong sign in the Dirichlet term, the variational problem $\min F(\varrho) + \frac{W_2^2(\varrho, \varrho_0)}{2\tau}$ requires some assumptions when proving the existence of minimizers. In particular, it would be possible that the infimum is $-\infty$, or that the energy is not l.s.c. because of the negative sign. In the case of dimension 2, sophisticated functional inequalities allow to handle the case $\alpha \leq 8\pi$. We refer to [15] and to the references therein for details on the analysis of this equation.

We also remark that the above model, coupling a parabolic equation on ϱ and an elliptic equation on u, implicitly assumes that the configuration of the chemo-attractant instantaneously follows that of ϱ. More sophisticated models can be expressed in terms of the so-called *parabolic–parabolic Keller–Segel* equation, in the form

$$\begin{cases} \partial_t \varrho + \alpha \nabla \cdot (\varrho \nabla u) - \Delta \varrho = 0, \\ \partial_t u - \Delta u = \varrho, \\ u = 0 \text{ on } \partial\Omega, \; \varrho(0, \cdot) = \varrho_0, \; \partial_{\mathbf{n}} \varrho - \varrho \partial_{\mathbf{n}} u = 0 \text{ on } \partial\Omega. \end{cases} \tag{4.22}$$

or other variants with different boundary conditions. Equation (4.22) can also be studied as a gradient flow in two variables, using the distance W_2 on ϱ and an L^2 distance on u; see [16]. In this case, we use

$$F(\varrho, u) := \int \varrho(x) \log(\varrho(x)) \mathrm{d}x - \alpha \int u(x) \varrho(x) \mathrm{d}x + \frac{\alpha}{2} \int |\nabla u|^2 \mathrm{d}x.$$

The JKO schemes becomes

$$(\varrho_{k+1}^\tau, u_{k+1}^\tau) \in \mathrm{argmin}_{(\varrho,u)} \, F(\varrho, u) + \frac{W_2^2(\varrho, \varrho_k^\tau)}{2\tau} + \alpha \frac{||u - u_k^\tau||_{L^2}^2}{2\tau}$$

(note that we multiplied the L^2 part of the distance times a coefficient α in order to adjust the coefficients in the parabolic equation on u). At a very formal level, the reader can see that this implies

$$\log(\varrho) - \alpha u + \frac{\varphi}{\tau} = const, \qquad \text{(from the optimality of } \varrho),$$

which gives the first equation in (4.22), and

$$-\varrho - \Delta u + \frac{u - u_k^\tau}{\tau} = 0, \qquad \text{(from the optimality of } u),$$

which gives the second equation in (4.22).

Crowd motion The theory of Wasserstein gradient flows has interestingly been applied to the study of a continuous model of crowd movement under density constraints.

This theory has first been introduced in a microscopic setting (each individual in the crowd being a rigid sphere), but we will not develop this aspect here (the interested reader can refer to the papers by Maury and Venel [69,70]). We are now interested in the simplest continuous counterpart of this microscopic model. In this case the particles population will be described by a probability density $\varrho \in \mathcal{P}(\Omega)$, and we are given a so-called "spontaneous" velocity field **u** representing the velocity each particle would follow if the other particles were not present. Yet, the crowd is supposed to be incompressible in the sense that it is subject to a maximal density constraint $\varrho \leq 1$. Then **u** must be adapted so that it belongs to a suitable set of admissible velocities, those which preserve this constraint. (This depends on the sign of the divergence of **v** on the saturated region $\{\varrho = 1\}$): $\mathrm{adm}(\varrho) = \big\{\mathbf{v} : \Omega \to \mathbb{R}^d :$ $\nabla \cdot \mathbf{v} \geq 0$ on $\{\varrho = 1\}\big\}$. The main axiom of the model (presented, for instance, in [67]) is that the density ϱ will be advected by a vector field $\mathbf{v}_t$ which is not **u** but its projection, at each time t, on $\mathrm{adm}(\varrho_t)$. Hence, we look for solutions of the equation

$$\partial_t \varrho_t + \nabla \cdot \big(\varrho_t \big(P_{\mathrm{adm}(\varrho_t)} \mathbf{u}_t \big) \big) = 0. \tag{4.23}$$

The main difficulty is the fact that the vector field $\mathbf{v} = P_{\mathrm{adm}(\varrho_t)} \mathbf{u}_t$ is not smooth (since it is only expected to be L^2), and it has a non-smooth dependance on ϱ (since passing from a density 1 to a density $1 - \varepsilon$ completely modifies the saturated zone, **v** is very sensitive to small changes in ϱ).

In [67] these difficulties have been overpassed in the case $\mathbf{u} = -\nabla V$ (where $V : \Omega \to \mathbb{R}$ is a given Lipschitz function) and the existence of a solution (with numerical simulations) is proven via a gradient flow method. Indeed, (4.23) turns out to be the gradient flow in $\mathbb{W}_2$ of the energy

$$F(\varrho) = \begin{cases} \int V \, d\varrho & \text{if } \varrho \in K; \\ +\infty & \text{if } \varrho \notin K, \end{cases}$$

where we define the set $K = \{\varrho \in \mathcal{P}(\Omega) : \varrho \leq 1\}$. The case $\mathbf{u} = -\nabla V$ is already quite satisfying for applications, since a typical case is obtained when $V(x) = \mathrm{dist}(x, \Gamma)$ and $\Gamma \subset \partial\Omega$ is a part of the boundary playing the role of an exit.

We do not enter into the details of the study of this equation, but we just make a little bit more precise the definitions above. Actually, it is more convenient to give a better description of $\mathrm{adm}(\varrho)$ by duality:

$$\mathrm{adm}(\varrho) = \left\{ \mathbf{v} \in L^2(\varrho) \ : \ \int \mathbf{v} \cdot \nabla p \leq 0 \quad \forall p \in H^1(\Omega) \ : \ p \geq 0, \ p(1 - \varrho) = 0 \right\}.$$

In this way we characterize $\mathbf{v} = P_{\mathrm{adm}(\varrho)}(u)$ through

$$\mathbf{u} = \mathbf{v} + \nabla p, \quad \mathbf{v} \in \mathrm{adm}(\varrho), \quad \int \mathbf{v} \cdot \nabla p = 0,$$

$$p \in \mathrm{press}(\varrho) := \{p \in H^1(\Omega), \ p \geq 0, \ p(1 - \varrho) = 0\},$$

where $\mathrm{press}(\varrho)$ is the space of "pressure" functions p (used as test functions in the dual definition of $\mathrm{adm}(\varrho)$). The two cones $\nabla\mathrm{press}(\varrho)$ and $\mathrm{adm}(\varrho)$ are in duality (i.e. one is defined as the set of vectors with negative scalar product with the elements of the other) for the L^2 scalar product. This provides a unique orthogonal decomposition $\mathbf{u}_t = \mathbf{v}_t + \nabla p_t$, and allows to re-write Eq. (4.23) as

$$\begin{cases} \partial_t \varrho_t + \nabla \cdot \big(\varrho_t(\mathbf{u}_t - \nabla p_t)\big) = 0, \\ 0 \leq \varrho \leq 1, \ p \geq 0, \ p(1 - \varrho) = 0. \end{cases} \tag{4.24}$$

More details can be found in [67,68,81]. Variants where the motion of the crowd also undergoes diffusion are also studied, as in [38,76] (in particular, [76] provides interesting uniqueness results) as well as variants with several species (see [25]).

Dirichlet boundary conditions So far, all the equations that we dealt with were always accompanied by Neumann-type boundary conditions: they are based on the continuity equation, which represents the conservation of mass, and we are describing the motion of a population ϱ of particles forced to stay inside a given domain Ω. These boundary conditions exactly impose the fact that particles do not exit the domain. Note that this does not mean that particles on the boundary should stay on the boundary forever (this is what formally happens under the condition $\mathbf{v} \cdot \mathbf{n} = 0$); the boundary condition here is rather $\varrho\mathbf{v} \cdot \mathbf{n} = 0$: if at a given instant of time the particles enter from $\partial\Omega$ into the interior, then immediately after there will be (locally) no mass on the boundary, and the condition is not violated, hence. On the contrary, should some mass go from the interior to outside Ω, then there would be indeed a violation, since there would be some mass $\varrho > 0$ on the boundary with velocity directed outwards.

Anyway, we see that Dirichlet conditions do not find their translation into W_2 gradient flows! To cope with Dirichlet boundary conditions, Figalli and Gigli defined in [42] a sort of modified Wasserstein distance, with a special role played by the boundary $\partial\Omega$, in order to study the heat equation $\partial_t\varrho = \Delta\varrho$ with Dirichlet b.c. $\varrho = 1$ on $\partial\Omega$.

Given two finite positive measures $\mu, \nu \in \mathcal{M}_+(\overset{\circ}{\Omega})$ (not necessarily with the same mass), set

$$\Pi b(\mu, \nu) = \{\gamma \in \mathcal{M}_+(\overline{\Omega} \times \overline{\Omega}) \ : \ (\pi_x)_\# \gamma \llcorner \overset{\circ}{\Omega} = \mu, \ (\pi_y)_\# \gamma \llcorner \overset{\circ}{\Omega} = \nu\}.$$

Then, define

$$Wb_2(\mu, \nu) := \sqrt{\inf\left\{\int_{\overline{\Omega} \times \overline{\Omega}} |x - y|^2 \, d\gamma, \ \gamma \in \Pi b(\mu, \nu)\right\}}.$$

The index b reminds the special role of the boundary. Informally, this means that the transport from μ to ν may be done either usually, or by moving some mass to/from $\partial\Omega$, the transport being free on $\partial\Omega$.

In [42] the authors prove that Wb_2 is a distance, that the space $\mathcal{M}_+(\overset{\circ}{\Omega})$ is a geodesic space (even if Ω is not convex) and they study the gradient flow, for this distance, of the functional $F(\varrho) = \int(\varrho\log\varrho - \varrho)\,\mathrm{d}x$ (note that here the term $-\int\varrho$ is important, because the total mass is not fixed), which corresponds to the heat equation with the particular boundary condition $\varrho = 1$ on $\partial\Omega$ (the value 1 is the value which minimizes $t \mapsto f(t) = t\log t - t$).

We observe that this kind of distances with free transport on the boundary were already present in [17,18], but for the Wasserstein distance W_1, and with no applications to gradient flows.

Reaction–diffusion equations A class of equations which are widely studied in applied mathematics and in particular mathematical biology are PDEs including both a divergence term and a source term, which injects or removes some mass into the system, thus getting an equation of the form

$$\partial_t\varrho_t + \nabla\cdot(\varrho_t\mathbf{v}_t) = f_t,$$

where $\mathbf{v}_t$ and f_t could depend on ϱ_t. When the divergence term is of the form $\Delta(g(\varrho_t))$, then we face a diffusion behavior, while the source term f, usually depending on ϱ, stands for internal reactions which provide growth or reproduction of the particle population ϱ. This is why these equations are usually called *reaction–diffusion* equations. Yet, it is clear that they cannot be dealt with with the theory of Wasserstein gradient flows, because of the reaction term, which is not in conservative form (and, by the way, the mass could not be preserved in time).

To consider equations of this form, a new distance has been recently introduced. This very interesting distance has been independently introduced by three different groups around the year 2015, and we refer to [31,32,60,63,64] for the different presentations and the different considerations which have been investigated so far. We give here only a very short description of this distance; the names chosen for it were of course different according to the different groups, and we decided to stick to the terminology chosen in [44] where it is called Kantorovich–Fisher–Rao distance to take into account all the contributions.

The starting point is the definition of W_2^2 via (4.8) (in the case $p = 2$), which can be modified as follows

$$\mathrm{KFR}^2(\mu, \nu) = \min\left\{\int_0^1\int_\Omega |\mathbf{v}_t|^2\mathrm{d}\varrho_t\,\mathrm{d}t + \int_0^1\int_\Omega |u_t|^2\mathrm{d}\varrho_t\,\mathrm{d}t \right.$$
$$\left. : \partial_t\varrho_t + \nabla\cdot(\varrho_t\mathbf{v}_t) = \varrho_t u_t,\ \varrho_0 = \mu,\ \varrho_1 = \nu\right\}.$$

This means that one can transform μ into ν either by displacing some masses, with velocity $\mathbf{v}$, and paying for the L^2 norm of $\mathbf{v}$, or by adding/removing some masses, using a scalar field u, and paying as well its L^2 norm. We do not enter into details of

the minimization problem defining KFR, which could also be turned into a convex one by using the variables $(\varrho, \varrho\mathbf{v}, \varrho u)$, and of its equivalent formulations. We only note that the optimal $\mathbf{v}$ and the optimal u are linked by $\mathbf{v} = \nabla u$. Using this fact, one can guess and then prove that the limit as $\tau \to 0$ of the curves obtained by interpolating the solutions of the iterated minimization problems

$$\varrho_{k+1} \in \operatorname{argmin}_\varrho F(\varrho) + \frac{\mathrm{KFR}^2(\varrho, \varrho_k)}{2\tau}$$

will provide a solution of the PDE

$$\partial_t \varrho_t - \nabla \cdot \left(\varrho_t \nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right) \right) = -\varrho_t \frac{\delta F}{\delta \varrho}(\varrho_t).$$

Even if the introduction of the distance KFR has the advantage to allow for reaction terms in the PDE, it is clear that the rigidity which connects the reaction and the diffusion terms is a serious draw-back which prevents to consider a wide class of equations. To cope with this, and consider equations of the form

$$\partial_t \varrho_t - \nabla \cdot \left(\varrho_t \nabla \left(\frac{\delta F}{\delta \varrho}(\varrho_t) \right) \right) = -\varrho_t \frac{\delta G}{\delta \varrho}(\varrho_t),$$

in [44] a splitting scheme is proposed. More precisely, one can define the following iterative algorithm:

$$\varrho_{k+1/2} \in \operatorname{argmin}_\varrho F(\varrho) + \frac{W_2^2(\varrho, \varrho_k)}{2\tau},$$

$$\varrho_{k+1} \in \operatorname{argmin}_\varrho G(\varrho) + \frac{\mathrm{FR}^2(\varrho, \varrho_{k+1/2})}{2\tau},$$

where the distance FR is defined only using the reaction part:

$$\mathrm{FR}^2(\mu, \nu) = \min \left\{ \int_0^1 \int_\Omega |u_t|^2 \mathrm{d}\varrho_t \, \mathrm{d}t \; : \; \partial_t \varrho_t = \varrho_t u_t, \; \varrho_0 = \mu, \; \varrho_1 = \nu \right\}.$$

For this last distance, [15] which is indeed known as Fisher–Rao, or Hellinger, distance, there is an explicit formula:

$$\mathrm{FR}^2(\mu, \nu) = 4 \int \left| \sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}} \right|^2 \mathrm{d}\lambda,$$

---

[15] Note that this second distance does not require the same mass for μ and ν, while the Wasserstein distance in the other minimization problem requires mass equality, but has to be extended to measures with equal mass, but not necessarily probability measures.

where λ is any reference positive measure such that $\mu, \nu \ll \lambda$ (for instance $\lambda = \mu + \nu$; anyway, the functional being one-homogeneous, the result does not depend on the reference measure).

4.7 Numerical methods from the JKO scheme

We present in this section two different numerical methods which have been recently proposed to tackle evolution PDEs which have the form of a gradient flow in $\mathbb{W}_2(\Omega)$ via their variational JKO scheme (the reader can consult [58] for an early treatment of these equations via the JKO scheme). We will only be concerned with discretization methods allowing the numerical treatment of one step of the JKO scheme, i.e. solving problems of the form

$$\min \left\{ F(\varrho) + \frac{1}{2} W_2^2(\varrho, \nu) \ : \ \varrho \in \mathcal{P}(\Omega) \right\},$$

for suitable ν (to be taken equal to ϱ_k^τ) and suitable F (including the τ factor). We will not consider the convergence as $\tau \to 0$ of the iterations to solutions of the evolution equation.

We will present two methods. One, essentially taken from [13], is based on the Benamou–Brenier formula first introduced in [11] as a numerical tool for optimal transport (also see [12]). This method is well-suited for the case where the energy $F(\varrho)$ used in the gradient flow is a convex function of ϱ. For instance, it works for functionals of the form $F(\varrho) = \int f(\varrho(x))\mathrm{d}x + \int V\mathrm{d}\varrho$ and can be used for Fokker–Planck and porous medium equations. The second method is based on methods from semi-discrete optimal transport, essentially developed by Q. Mérigot using computational geometry (see [59,62,74] for 3D implementation) and allows to translate the problem into an optimization problem in the class of convex functions; it is well suited for the case where F is geodesically convex, which means that the term $\int f(\varrho(x))\,\mathrm{d}x$ is only admissible if f satisfies McCann's condition, the term $\int V\mathrm{d}\varrho$ needs V to be convex, but interaction terms such as $\int \int W(x - y)\,\mathrm{d}\varrho(x)\,\mathrm{d}\varrho(y)$ are also allowed, if W is convex.

Augmented Lagrangian methods Let us recall the basis of the Benamou–Brenier method. This amounts to solve the variational problem (4.9) which reads, in the quadratic case, as

$$\min \quad \sup_{(a,b)\in K_2} \int\int a\mathrm{d}\varrho + \int\int b \cdot \mathrm{d}E \ : \ \partial_t \varrho_t + \nabla \cdot E_t = 0, \ \varrho_0 = \mu, \ \varrho_1 = \nu$$

where $K_2 = \{(a, b) \in \mathbb{R} \times \mathbb{R}^d \ : \ a + \frac{1}{2}|b|^2 \leq 0\}$. We then use the fact that the continuity equation constraint can also be written as a sup penalization, by adding to the functional

$$\sup_{\phi \in C^1([0,1]\times\Omega)} -\int\int \partial_t \phi \mathrm{d}\varrho - \int\int \nabla\phi \cdot \mathrm{d}E + \int \phi_1 \mathrm{d}\nu - \int \phi_0 \mathrm{d}\mu,$$

which is 0 in case the constraint is satisfied, and $+\infty$ if not.

It is more convenient to express everything in the space-time formalism, i.e. by writing $\nabla_{t,x}\phi$ for $(\partial_t\phi, \nabla\phi)$ and using the variable $\mathfrak{m}$ for (ϱ, E) and A for (a, b). We also set $G(\phi):= \int \phi_1 d\nu - \int \phi_0 d\mu$. Then the problem becomes

$$\min_{\mathfrak{m}} \sup_{A,\phi} \quad \mathfrak{m} \cdot (A - \nabla_{t,x}\phi) - I_{K_2}(A) + G(\phi),$$

where the scalar product is here an L^2 scalar product, but becomes a standard Euclidean scalar product as soon as one discretizes (in time-space). The function I_{K_2} denotes the indicator function in the sense of convex analysis, i.e. $+\infty$ if the condition $A \in K_2$ is not satisfied, and 0 otherwise.

The problem can now be seen as the search for a saddle-point of the Lagrangian

$$L(\mathfrak{m}, (A, \phi)):=\mathfrak{m} \cdot (A - \nabla_{t,x}\phi) - I_{K_2}(A) + G(\phi),$$

which means that we look for a pair $(\mathfrak{m}, (A, \phi))$ (actually, a triple, but A and ϕ play together the role of the second variable) where $\mathfrak{m}$ minimizes for fixed (A, ϕ) and (A, ϕ) maximizes for fixed $\mathfrak{m}$. This fits the following framework, where the variables are X and Y and the Lagrangian has the form $L(X, Y):=X \cdot \Lambda Y - H(Y)$. In this case one can use a very smart trick, based on the fact that the saddle points of this Lagrangian are the same of the *augmented Lagrangian* $\tilde{L}$ defined as $\tilde{L}(X, Y):=X \cdot \Lambda Y - H(Y) - \frac{r}{2}|\Lambda Y|^2$, whatever the value of the parameter $r > 0$ is (see, for instance, [43]). Indeed, the saddle-point of L are characterized by (we assume all the functions we minimize are convex and all the functions we maximize are concave)

$$\begin{cases} \Lambda Y = 0 & \text{(optimality of } X), \\ \Lambda^t X - \nabla H(Y) = 0 & \text{(optimality of } Y), \end{cases}$$

while those of $\tilde{L}$ are characterized by

$$\begin{cases} \Lambda Y = 0 & \text{(optimality of } X), \\ \Lambda^t X - \nabla H(Y) - r\Lambda^t \Lambda Y = 0 & \text{(optimality of } Y), \end{cases}$$

which is the same since the first equation implies that the extra term in the second vanishes.

In this case, we obtain a saddle point problem of the form

$$\min_{\mathfrak{m}} \max_{A,\phi} \quad \mathfrak{m} \cdot (A - \nabla_{t,x}\phi) - I_{K_2}(A) + G(\phi) - \frac{r}{2}||A - \nabla_{t,x}\phi||^2$$

(where the squared norm in the last term is an L^2 norm in time and space), which is then solved by iteratively repeating three steps: for fixed A and $\mathfrak{m}$, finding the optimal ϕ (which amounts to minimizing a quadratic functional in calculus of variations, i.e. solving a Poisson equation in the space-time $[0, 1] \times \Omega$, with Neumann boundary conditions, homogeneous on $\partial\Omega$ and non-homogeneous on $t = 0$ and $t = 1$, due to

the term G); then for fixed ϕ and $\mathfrak{m}$ find the optimal A (which amounts to a pointwise minimization problem, in this case a projection on the convex set K_2); finally update $\mathfrak{m}$ by going in the direction of the gradient descent, i.e. replacing $\mathfrak{m}$ with $\mathfrak{m} - r(A - \nabla_{t,x}\phi)$ (it is convenient to choose the parameter of the gradient descent to be equal to that the Augmented Lagrangian).

This is what is done in the case where the initial and final measures are fixed. At every JKO step, one is fixed (say, μ), but the other is not, and a penalization on the final ϱ_1 is added, of the form $\tau F(\varrho_1)$. Inspired from the considerations above, the saddle point below allows to treat the problem

$$\min_{\varrho_1} \quad \frac{1}{2}W_2^2(\varrho_1, \mu) + \int f(\varrho_1(x))\mathrm{d}x + \int V\mathrm{d}\varrho_1$$

by formulating it as

$$\min_{\mathfrak{m},\varrho_1} \max_{A,\phi,\lambda} \int \int \mathfrak{m} \cdot (A - \nabla_{t,x}\phi) + \int \varrho_1 \cdot (\phi_1 + \lambda + V) - \int \int I_{K_2}(A) - \int \phi_0 \mathrm{d}\mu$$
$$- \int f^*(\lambda(x))\mathrm{d}x - \frac{r}{2}\int \int |A - \nabla_{t,x}\phi|^2 - \frac{r}{2}\int |\phi_1 + \lambda + V|^2,$$

where we re-inserted the integration signs to underline the difference between integrals in space-time (with $\mathfrak{m}$, A and ϕ) and in space only (with ϕ_0, ϕ_1, ϱ_1, V and λ). The role of the variable λ is to be dual to ϱ_1, which allows to express $f(\varrho_1)$ as $\sup_\lambda \varrho_1\lambda - f^*(\lambda)$.

To find a solution to this saddle-point problem, an iterative procedure is also used, as above. The last two steps are the update via a gradient descent of $\mathfrak{m}$ and ϱ_1, and do not require further explications. The first three steps consist in the optimization of ϕ (which requires the solution of a Poisson problem) and in two pointwise minimization problems in order to find A (which requires a projection on K_2) and λ (the minimization of $f^*(\lambda) + \frac{r}{2}|\phi_1(x) + \lambda + V(x)|^2 - \varrho_1(x)\lambda$, for fixed x).

For the applications to gradient flows, a small time-step $\tau > 0$ has to be fixed, and this scheme has to be done for each k, using $\mu = \varrho_k^\tau$ and setting ϱ_{k+1}^τ equal to the optimizer ϱ_1 and the functions f and V must include the scale factor τ. The time-space $[0, 1] \times \Omega$ has to be discretized but the evolution in time is infinitesimal (due to the small time scale τ), which allows to choose a very rough time discretization. In practice, the interval $[0, 1]$ is only discretized using less than 10 time steps for each k...

The interested reader can consult [13] for more details, examples and simulations.

Optimization among convex functions It is clear that the optimization problem

$$\min_{\varrho} \quad \frac{1}{2}W_2^2(\varrho, \mu) + F(\varrho)$$

can be formulated in terms of transport maps as

$$\min_{T:\Omega\to\Omega} \quad \frac{1}{2}\int_\Omega |T(x) - x|^2\mathrm{d}\mu(x) + F(T_\#\mu).$$

Also, it is possible to take advantage of Brenier's theorem which characterizes optimal transport maps as gradient of convex functions, and recast it as

$$\min_{u \text{ convex} : \nabla u \in \Omega} \frac{1}{2} \int_{\Omega} |\nabla u(x) - x|^2 \mathrm{d}\mu(x) + F((\nabla u)_{\#}\mu).$$

It is useful to note that in the last formulation the convexity of u is not necessary to be imposed, as it would anyway come up as an optimality condition. On the other hand, very often the functional F involves explicitly the density of the image measure $(\nabla u)_{\#}\mu$ (as it is the case for the typical example $\mathcal{F}$), and in this case convexity of u helps in computing this image measure. Indeed, whenever u is convex we can say that the density of $\varrho := (\nabla u)_{\#}\mu$ (if μ itself is absolutely continuous, with a density that we will denote by ϱ_0) is given by[16]

$$\varrho = \frac{\varrho_0}{\det(D^2 u)} \circ (\nabla u)^{-1}.$$

Hence, we are facing a calculus of variations problem in the class of convex functions. A great difficulty to attack this class of problems is how to discretize the space of convex functions. The first natural approach would be to approximate them by piecewise linear functions over a fixed mesh. In this case, imposing convexity becomes a local feature, and the number of linear constraints for convexity is proportional to the size of the mesh. Yet, Choné and Le Meur showed in [33] that we cannot approximate in this way all convex functions, but only those satisfying some extra constraints on their Hessian (for instance, those which also have a positive mixed second derivative $\partial^2 u / \partial x_i \partial x_j$). Because of this difficulty, a different approach is needed. For instance, Ekeland and Moreno-Bromberg used the representation of a convex function as a maximum of affine functions [40], but this needed many more linear constraints; Oudet and Mérigot [75] decided to test convexity on a different (and less refined) grid than that where the functions are defined... These methods give somehow satisfactory answers for functionals involving u and ∇u, but are not able to handle terms involving the Monge–Ampère operator $\det(D^2 u)$.

The method proposed in [14], that we will roughly present here, does not really use a prescribed mesh. The idea is the following: suppose that μ is a discrete measure of atomic type, i.e. of the form $\sum_j a_j \delta_{x_j}$. A convex function defined on its support $S := \{x_j\}_j$ will be a function $u : S \to \mathbb{R}$ such that at each point $x \in S$ the subdifferential

$$\partial u(x) := \{p \in \mathbb{R}^d : u(x) + p \cdot (y - x) \le u(y) \text{ for all } y \in S\}$$

---

[16] This same formula takes into account the possibility that ∇u could be non-injective, i.e. u non-strictly convex, in which case the value of the density could be $+\infty$ due to the determinant at the denominator which would vanish.

is non-empty. Also, the Monge–Ampère operator will be defined by using the sub-differential, and more precisely the equality

$$\int_B \det(D^2 u(x)) \mathrm{d}x = |\partial u(B)|$$

which is valid for smooth convex functions u and arbitrary open sets B. An important point is the fact that whenever f is superlinear, functionals of the form $\int f(\varrho(x)) \mathrm{d}x$ impose, for their finiteness, the positivity of $\det(D^2 u)$, which will in turn impose that the sub-differential has positive volume, i.e. it is non-empty, and hence convexity…

More precisely, we will minimize over the set of pairs $(u, P) : S \to \mathbb{R} \times \mathbb{R}^d$ where $P(x) \in \partial u(x)$ for every $x \in S$. For every such pair (u, P) we weed to define $G(u, P)$ which is meant to be $(\nabla u)_\# \mu$, and define $F(G(u, P))$ whenever F has the form $F = \mathcal{F} + \mathcal{V} + \mathcal{W}$. We will simply define

$$\mathcal{V}(G(u, P)) := \sum_j a_j V(P(x_j)) \quad \text{and} \quad \mathcal{W}(G(u, P)) := \sum_{j,j'} a_j a_{j'} W(P(x_j) - P(x_{j'})),$$

which means that we just use $P_\# \mu$ instead of $(\nabla u)_\# \mu$. Unfortunately, this choice is not adapted for the functional $\mathcal{F}$, which requires absolutely continuous measures, and $P_\# \mu$ is atomic. In this case, instead of concentrating all the mass a_j contained in the point $x_j \in S$ on the unique point $P(x_j)$, we need to spread it, and we will spread it uniformly on the whole subdifferential $\partial u(x_j)$. This means that we also define a new surrogate of the image measure $(\nabla u)_\# \mu$, called $G^{ac}(u, P)$ (where the superscript ac stands for *absolutely continuous*), given by

$$G^{ac}(u, P) := \sum_j \frac{a_j}{|A_j|} \mathcal{L}^d \llcorner A_j,$$

where $A_j := \partial u(x_j) \cap \Omega$ (the intersection with Ω is done in order to take care of the constraint $\nabla u \in \Omega$). Computing $\mathcal{F}(G^{ac}(u, P))$ gives hence

$$\mathcal{F}(G^{ac}(u, P)) = \sum_j |A_j| f\left(\frac{a_j}{|A_j|}\right).$$

It is clear that the discretization of $\mathcal{V}$ and $\mathcal{W}$ in terms of $G(u, P)$ are convex functions of (u, P) (actually, of P, and the constraint relating P and u is convex) whenever V and W are convex; concerning $\mathcal{F}$, it is possible to prove, thanks to the concavity properties of the determinant or, equivalently, to the Brunn–Minkowski inequality (see for instance [86]) that $\mathcal{F}(G^{ac}(u, P))$ is convex in u as soon as f satisfies McCann's condition. Globally, it is not surprising to see that we face a convex variational problem in terms of u (or of ∇u) as soon as F is displacement convex in ϱ (actually, convexity on generalized geodesics based at μ should be the correct notion).

Then, we are lead to study the variational problem

$$\min_{u, P} \quad \frac{1}{2} \sum_j a_j |P(x_j) - x_j|^2 + \mathcal{V}(G(u, P)) + \mathcal{W}(G(u, P)) + \mathcal{F}(G^{ac}(u, P)) \quad (4.25)$$

under the constraints $P(x_j) \in A_j := \partial u(x_j) \cap \Omega$. Note that we should incorporate the scale factor τ in the functional F which means that, for practical purposes, convexity in P is guaranteed as soon as V and W have second derivatives which are bounded from below (they are semi-convex) and τ is small (the quadratic term coming from W_2^2 will always overwhelm possible concavity of the other terms). The delicate point is how to compute the subdifferentials $\partial u(x_j)$, and optimize them (i.e. compute derivatives of the relevant quantity w.r.t. u).

This is now possible, and in a very efficient way, thanks to tools from computational geometry. Indeed, in this context, subdifferentials are exactly a particular case of what are called Laguerre cells, which in turn are very similar to Voronoi cells. We remind that, given some points $(x_j)_j$, their Voronoi cells V_j are defined by

$$V_j := \left\{ x \in \Omega \ : \ \frac{1}{2}|x - x_j|^2 \leq \frac{1}{2}|x - x_{j'}|^2 \ \text{for all } j' \right\}$$

(of course the squares of the distances could be replaced by the distances themselves, but in this way it is evident that the cells V_j are given by a finite number of linear inequalities, and are thus convex polyhedra; the factors $\frac{1}{2}$ are also present only for cosmetic reasons). Hence, Voronoi cells are the cells of points which are closer to one given point $x_j \in S$ than to the others.

In optimal transport a variant of these cells is more useful: given a set of values ψ_j, we look for the cells (called *Laguerre cells*)

$$W_j := \left\{ x \in \Omega \ : \ \frac{1}{2}|x - x_j|^2 + \psi_j \leq \frac{1}{2}|x - x_{j'}|^2 + \psi_{j'} \ \text{for all } j' \right\}.$$

This means that we look at points which are closer to x_j than to the other points $x_{j'}$, up to a correction[17] given by the values ψ_j. It is not difficult to see that also in this case cells are convex polyhedra. And it is also easy to see that, if ϱ is an absolutely continuous measure on Ω and $\mu = \sum_j a_j \delta_{x_j}$, then finding an optimal transport map from ϱ to μ is equivalent to finding values ψ_j such that $\varrho(W_j) = a_j$ for every j (indeed, in this case, the map sending every point of W_j to x_j is optimal, and $-\psi_j$ is the value of the corresponding Kantorovich potential at the point x_j). Finally, it can be easily seen that the Laguerre cells corresponding to $\psi_j := u(x_j) - \frac{1}{2}|x_j|^2$ are nothing but the subdifferentials of u (possibly intersected with Ω).

Handling Laguerre cells from the computer point of view has for long been difficult, but it is now state-of-the-art in computational geometry, and it is possible to compute

---

[17] If the points x_j are the locations of some ice-cream sellers, we can think that ψ_j is the price of an ice-cream at x_j, and the cells W_j will represent the regions where customers will decide to go to the seller j, keeping into account both the price and the distance.

very easily their volumes (incidentally, also find some points P belonging to them, which is useful so as to satisfy the constraints of Problem (4.25)), as well as the derivatives of their volumes (which depend on the measures of each faces) w.r.t. the values ψ_j. For the applications to semi-discrete[18] optimal transport problems, the results are now very fast (with discretizations with up to 10^6 points in some minutes, in 3D; the reader can have a look at [59,62,74] but also to Section 6.4.2 in [84]), and the same tools have been used for the applications to the JKO scheme that we just described.

In order to perform an iterated minimization, it is enough to discretize ϱ_0 with a finite number of Dirac masses located at points x_j, to fix $\tau > 0$ small, then to solve (4.25) with $\mu = \varrho_k^\tau$ and set $\varrho_{k+1}^\tau := G(u, P)$ for the optimal (u, P). Results, proofs of convergence and simulations are in [14].

5 The heat flow in metric measure spaces

In this last section we will give a very sketchy overview of an interesting research topic developed by Ambrosio, Gigli, Savaré and their collaborators, which is in some sense a bridge between

- the theory of gradient flows in $\mathbb{W}_2$, seen from an abstract metric space point of view (which is not the point of view that we underlined the most in the previous section),
- and the current research topic of analysis and differential calculus in metric measure spaces (see, for instance, [49]).

This part of their work is very ambitious, and really aims at studying analytical and geometrical properties of metric measure spaces; what we will see here is only a starting point.

The topic that we will briefly develop here is concerned with the heat flow, and the main observation is the following: in the Euclidean space $\mathbb{R}^d$ (or in a domain $\Omega \subset \mathbb{R}^d$), the heat flow $\partial_t \varrho = \Delta \varrho$ may be seen as a gradient flow in two different ways:

- first, it is the gradient flow in the Hilbert space $L^2(\Omega)$, endowed with the standard L^2 norm, of the functional consisting in the Dirichlet energy $\mathcal{D}(\varrho) = \int |\nabla \varrho|^2 dx$ (a functional which is set to $+\infty$ if $\varrho \notin H^1(\Omega)$); in this setting, the initial datum ϱ_0 could be any function in $L^2(\Omega)$, but well-known properties of the heat equation guarantee $\varrho_0 \geq 0 \Rightarrow \varrho_t \geq 0$ and, if Ω is the whole space, or boundary conditions are Neumann, then $\int \varrho_0 dx = 1 \Rightarrow \int \varrho_t dx = 1$; it is thus possible to restrict to probability densities (i.e. positive densities with mass one);
- then, if we use the functional $\mathcal{E}$ of the previous section (the entropy defined with $f(t) = t \log t$), the heat flow is also a gradient flow in the space $\mathbb{W}_2(\Omega)$.

A natural question arises: is the fact that these two flows coincide a general fact? How to analyze this question in a general metric space? In the Euclidean space this is easy: we just write the PDEs corresponding to each of these flows and, as they are the same PDE, for which we know uniqueness results, then the two flows are the same.

---

[18] One measure being absolutely continuous, the other atomic.

First, we realize that the question is not well-posed if the only structure that we consider on the underlining space is that of a metric space. Indeed, we also need a reference measure (a role played by the Lebesgue measure in the Euclidean space). Such a measure is needed in order to define the integral $\int |\nabla \varrho|^2 \mathrm{d}x$, and also the entropy $\int \varrho \log \varrho \, \mathrm{d}x$. Roughly speaking, we need to define "$\mathrm{d}x$".

Hence, we need to consider *metric measure spaces*, (X, d, m), where $m \geq 0$ is a reference measure (usually finite) on the Borel tribe of X. The unexperienced reader should not be surprised: metric measure spaces are currently the new frontier of some branches of geometric analysis, as a natural generalization of Riemannian manifolds. In order not to overburden the reference list, we just refer to the following papers, already present in the bibliography of this survey for other reasons: [2,4,6,7,30,48,50–53,65,85,87].

5.1 Dirichlet and Cheeger energies in metric measure spaces

In order to attack our question about the comparison of the two flows, we first need to define and study the flow of the Dirichlet energy, and in particular to give a suitable definition of such an energy. This more or less means defining the space $H^1(X)$ whenever X is a metric measure space (MMS). This is not new, and many authors studied it: we cite in particular [30,51,52,85]). Moreover, in their recent works, [4,6], Ambrosio, Gigli and Savaré presented some results in this direction, useful for the analysis of the most general case (consider that most of the previous results require a *doubling* assumption and the existence of a *Poincaré inequality*, see also [53], and this assumption on (X, d, m) is not required in their papers). One of the first definition of Sobolev spaces on a MMS had been given by Haiłasz, who used the following definition

$$f \in H^1(X, d, m) \text{ if there is } g \in L^2(X, m)$$
$$\text{such that } |f(x) - f(y)| \leq d(x, y)(g(x) + g(y)).$$

This property characterizes Sobolev spaces in $\mathbb{R}^d$ by choosing

$$g = const \cdot M\left[|\nabla f|\right],$$

where $M[u]$ denotes the maximal function of u: $M[u](x) := \sup_{r>0} \fint_{B(x,r)} u$ (the important point here is the classical result in harmonic analysis guaranteeing $||M[u]||_{L^2} \leq C(d)||u||_{L^2}$) and c is a suitable constant only depending on the dimension d. As this definition is not local, almost all the recent investigations on these topics are rather based on some other ideas, due to Cheeger [30], using the relaxation starting from Lipschitz functions, or to Shanmuganlingam [85], based on the inequality

$$|f(x(0)) - f(x(1))| \leq \int_0^1 |\nabla f(x(t))||x'(t)| \, \mathrm{d}t$$

required to hold on almost all curves, in a suitable sense. In the recent paper [4] the authors present a classification of the various weak notions of modulus of the gradient in a MMS and analyze their equivalence. On the contrary, here we will only choose one unique definition for $\int |\nabla f|^2 \mathrm{d}m$, the one which seems the simplest.

For every Lipschitz function f on X, let us take its *local Lipschitz constant* $|\nabla f|$, defined in (3.1), and set $\mathcal{D}(f) := \int |\nabla f|^2(x) \mathrm{d}m$. Then, by relaxation, we define the *Cheeger Energy*[19] $\mathcal{C}(f)$:

$$\mathcal{C}(f) := \inf \left\{ \liminf_n \ \mathcal{D}(f_n) \ : \ f_n \to f \text{ in } L^2(X, m), \ f_n \in \mathrm{Lip}(X) \right\}.$$

We then define the Sobolev space $H^1(X, d, m)$ as the space of functions such that $\mathcal{C}(f) < +\infty$. This space will be a Banach space, endowed with the norm $f \mapsto \sqrt{\mathcal{C}(f)}$ and the function $f \mapsto \mathcal{C}(f)$ will be convex. We can also define $-\Delta f$ as the element of minimal norm of the subdifferential $\partial \mathcal{C}(f)$ (an element belonging to the dual of $H^1(X, d, m)$). Beware that, in general, the map $f \mapsto -\Delta f$ will not be linear (which corresponds to the fact that the norm $\sqrt{\mathcal{C}(f)}$ is in general not Hilbertian, i.e. it does not come from a scalar product).

Defining the flow of $\mathcal{C}$ in the Hilbert space $L^2(X, m)$ is now easy, and fits well the classical case of convex functionals on Hilbert spaces or, more generally, of monotone maximal operators (see [22]). This brings very general existence and uniqueness results.

5.2 A well-posed gradient flow for the entropy

A second step (first developed in [48] and then generalized in [5,6]) consists in providing existence and uniqueness conditions for the gradient flow of the entropy, w.r.t. the Wasserstein distance W_2. To do so, we consider the functional $\mathcal{E}$, defined on the set of densities f such that $\varrho := f \cdot m$ is a probability measure via $\mathcal{E}(f) := \int f \log f \, \mathrm{d}m$ and we look at its gradient flow in $\mathbb{W}_2$ in the EDE sense. In order to apply the general theory of Sect. 3, as we cannot use the notion of weak solutions of the continuity equation, it will be natural to suppose that this functional $\mathcal{E}$ is λ-geodesically convex for some $\lambda \in \mathbb{R}$. This means, in the sense of Sturm and Lott–Villani, that the space (X, d, m) is a MMS with *Ricci curvature bounded from below*. We recall here the corresponding definition, based on the characteristic property already evoked in the Sect. 4.5, which was indeed a theorem (Proposition 4.11) in the smooth case.

Definition 5.1 A metric measure space (X, d, m) is said to have a Ricci curvature bounded from below by a constant $K \in \mathbb{R}$ in the sense of Sturm and Lott-Villani if the entropy functional $\mathcal{E} : \mathcal{P}(X) \to \mathbb{R} \cup \{+\infty\}$ defined through

$$\mathcal{E}(\varrho) = \begin{cases} \int f \log f \, \mathrm{d}m & \text{if } \varrho = f \cdot m \\ +\infty & \text{if } \varrho \text{ is not absolutely continuous w.r.t. } m \end{cases}$$

---

[19] The name has been chosen because Cheeger also gave a definition by relaxation; moreover, the authors did not wish to call it Dirichlet energy, as generally this name is used fro quadratic forms.

is K-geodesically convex in the space $\mathbb{W}_2(X)$. In this case we say that (X, d, m) satisfies the condition[20] $CD(K, \infty)$.

Note that the EVI formulation is not available in this case, as we do not have the geodesic convexity of the squared Wasserstein distance. Moreover, on a general metric space, the use of generalized geodesics is not always possible. This is the reason why we will define the gradient flow of $\mathcal{E}$ by the EDE condition and not the EVI, but this requires to prove via other methods the uniqueness of such a gradient flow. To do so, Gigli introduced in [48] an interesting strategy to prove uniqueness for EDE flows, which is based on the following proposition.

Proposition 5.1 *If $F : \mathcal{P}(X) \rightarrow \mathbb{R} \cup \{+\infty\}$ is a strictly convex functional (w.r.t. usual convex combinations $\mu_s := (1 - s)\mu_0 + s\mu_1$, which is meaningful in the set of probability measures), such that $|\nabla^- F|$ is an upper gradient for F and such that $|\nabla^- F|^2$ is convex, then for every initial measure $\bar{\mu}$ there exists at most one gradient flow $\mu(t)$ in the EDE sense for the functional F satisfying $\mu(0) = \bar{\mu}$.*

In particular, this applies to the functional $\mathcal{E}$: the strict convexity is straightforward, and the squared slope can be proven to be convex with the help of the formula (3.4) (it is interesting to observe that, in the Euclidean case, an explicit formula for the slope is known:

$$|\nabla^- \mathcal{E}|^2(\varrho) = \int \frac{|\nabla f|^2}{f} \, dx, \tag{5.1}$$

whenever $\varrho = f \cdot \mathcal{L}^d$).

5.3 Gradient flows comparison

The last point to study (and it is not trivial at all) is the fact that every gradient flow of $\mathcal{C}$ (w.r.t. the L^2 distance) is also an EDE gradient flow of $\mathcal{E}$ for the W_2 distance. This one (i.e. $(\mathcal{C}, L^2) \Rightarrow (\mathcal{E}, W_2)$) is the simplest direction to consider in this framework, as computations are easier. This is a consequence of the fact that the structure of gradient flows of convex functionals in Hilbert spaces is much more well understood. In order to do so, it is useful to compute and estimate

$$\frac{d}{dt}\mathcal{E}(f_t), \quad \text{where } f_t \text{ is a gradient flow of } \mathcal{C} \text{ in } L^2(X, m).$$

This computation is based on a strategy essentially developed in [50] and on a lemma by Kuwada. The initial proof, contained in [50], is valid for Alexandroff spaces.[21]

---

[20] The general notation $CD(K, N)$ is used to say that a space has curvature bounded from below by K and dimension bounded from above by N (CD stands for "curvature-dimension").

[21] These spaces, see [23], are metric spaces where triangles are at least as fat as the triangles of a model comparison manifold with constant curvature equal to K, the comparison being done in terms of the distances from a vertex of a triangle to the points of a geodesic connecting the two other vertices. These spaces can be proven to have always an integer dimension $d \in \mathbb{N} \cup \{\infty\}$, and can be considered as MMS whenever $d < \infty$, by endowing them with their Hausdorff measure $\mathcal{H}^d$. Note anyway that the comparison manifold with constant curvature can be, anyway, taken of dimension 2, as only triangles appear in the definition.

The generalization of the same result to arbitrary MMS satisfying $CD(K, \infty)$ is done in [6].

Proposition 5.2 *If f_t is a gradient flow of $\mathcal{C}$ in $L^2(X, \mathcal{H}^d)$, then we have the following equality with the Fisher information:*

$$-\frac{d}{dt}\mathcal{E}(f_t) = \mathcal{C}(2\sqrt{f_t}).$$

Moreover, for every $\varrho = f \cdot \mathcal{H}^d \in \mathcal{P}(X)$ we have

$$\mathcal{C}(2\sqrt{f}) \geq |\nabla^-\mathcal{E}|^2(\varrho)$$

(where the slope of $\mathcal{E}$ is computed for the W_2 distance[22]). Also, if we consider the curve $\varrho_t = f_t \cdot \mathcal{H}^d$, it happens that ϱ_t is an AC curve in the space $\mathbb{W}_2(X)$ and

$$|\varrho'|(t)^2 \leq \mathcal{C}(2\sqrt{f_t}).$$

These three estimates imply that ϱ_t is a gradient flow of $\mathcal{E}$ w.r.t. W_2.

Once this equivalence is established, we can wonder about the properties of this gradient flow. The L^2 distance being Hilbertian, it is easy to see that the C^2G^2 property is satisfied, and hence this flow also satisfies EVI. On the contrary, it is not evident that the same is true when we consider the same flow as the gradient flow of $\mathcal{E}$ for the distance W_2. Indeed, we can check that the following three conditions are equivalent (all true or false depending on the space (X, d, m), which is supposed to satisfy $CD(K, \infty)$; see [7] for the proofs):

- the unique EDE gradient flow of $\mathcal{E}$ for W_2 also satisfies EVI;
- the heat flow (which is at the same time the gradient flow of $\mathcal{E}$ for W_2 and of $\mathcal{C}$ for L^2) depends linearly on the initial datum;
- (if we suppose that (X, d, m) is a Finsler manifold endowed with its natural distance and its volume measure), X is a Riemannian manifold.

As a consequence, Ambrosio, Gigli and Savaré proposed in [7] a definition of MMS having a *Riemanniann ricci curvature* bounded from below by requiring both to satisfy the $CD(K, \infty)$ condition, and the linearity of the heat flow (this double condition is usually written $RCD(K, \infty)$). This is the notion of infinitesimally Hilbertian space that we mentioned at the end of Sect. 3.

It is important to observe (but we will not develop this here) that these notions of Ricci bounds (either Riemannian or not) are stable via measured Gromov–Hausdorff convergence (a notion of convergence similar to the Gromov–Hausdorff convergence of metric spaces, but considering the minimal Wasserstein distance between the images of two spaces via isometric embeddings into the same space). This can be surprising at a first sight (curvature bounds are second-order objects, and we are claiming that they are stable via a convergence which essentially sounds like a uniform convergence of the

---

[22] As in (5.1).

spaces, with a weak convergence of the measures), but not after a simple observation: also the class of convex, or λ-convex, functions is stable under uniform (or even weak) convergence! but, of course, proving this stability is not a trivial fact.

References

1. Ambrosio, L.: Movimenti minimizzanti. Rend. Accad. Naz. Sci. XL Mem. Mat. Sci. Fis. Natur. **113**, 191–246 (1995)
2. Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modelling and Optimisation of Flows on Networks. Lecture Notes in Mathematics, vol. 2062, pp. 1–155 Springer-Verlag (2013)
3. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Spaces of Probability Measures. Lectures in Mathematics. ETH Zurich, Birkhäuser (2005)
4. Ambrosio, L., Gigli, N., Savaré, G.: Density of Lipschitz functions and equivalence of weak gradients in metric measure spaces. Rev. Mat. Iberoam. **29**, 969–986 (2013)
5. Ambrosio, L., Gigli, N., Savaré, G.: Heat flow and calculus on metric measure spaces with Ricci curvature bounded below—the compact case. In: Brezzi, F., Colli Franzone, P., Gianazza, U.P., Gilardi, G. (eds.) Analysis and Numerics of Partial Differential Equations. "INDAM", vol. 4, pp. 63–116. Springer, Berlin (2013)
6. Ambrosio, L., Gigli, N., Savaré, G.: Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. Inv. Math. **195**(2), 289–391 (2014)
7. Ambrosio, L., Gigli, N., Savaré, G.: Metric measure spaces with Riemannian Ricci curvature bounded from below. Duke Math. J. **163**, 1405–1490 (2014)
8. Ambrosio, L., Savaré, G.: Gradient flows of probability measures. In: Dafermos, C.M., Feireisl, E. (eds.) Handbook of Differential Equations: Evolutionary equations, vol. 3. Elsevier, Amsterdam (2007)
9. Ambrosio, L., Tilli, P.: Topics on Analysis in Metric Spaces. Oxford Lecture Series in Mathematics and its Applications, vol. 25. Oxford University Press, Oxford (2004)
10. Aubin, J.-P.: Un théorème de compacité. C. R. Acad. Sci. Paris **256**, 5042–5044 (1963)
11. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**, 375–393 (2000)
12. Benamou, J.-D., Carlier, G.: Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. J. Opt. Theor. Appl. **167**(1), 1–26 (2015)
13. Benamou, J.-D., Carlier, G., Laborde, M.: An augmented Lagrangian approach to Wasserstein gradient flows and applications. ESAIM Proc. **54**, 1–17 (2016)
14. Benamou, J.-D., Carlier, G., Mérigot, Q., Oudet, É.: Discretization of functionals involving the Monge–Ampère operator. Numerische Mathematik **134**(3), 611–636 (2016)
15. Blanchet, A., Calvez, V., Carrillo, J.A.: Convergence of the mass-transport steepest descent scheme for the subcritical Patlak–Keller–Segel model. SIAM J. Numer. Anal. **46**(2), 691–721 (2008)
16. Blanchet, A., Carrillo, J.-A., Kinderlehrer, D., Kowalczyk, M., Laurençot, P., Lisini, S.: A hybrid variational principle for the Keller–Segel system in $\mathbb{R}^2$. ESAIM M2AN **49**(6), 1553–1576 (2015)
17. Bouchitté, G., Buttazzo, G.: Characterization of optimal shapes and masses through Monge–Kantorovich equation. J. Eur. Math. Soc. **3**(2), 139–168 (2001)
18. Bouchitté, G., Buttazzo, G., Seppecher, P.: Shape optimization solutions via Monge–Kantorovich equation. C. R. Acad. Sci. Paris Sér. I Math **324**(10), 1185–1191 (1997)
19. Brenier, Y.: Décomposition polaire et réarrangement monotone des champs de vecteurs. C. R. Acad. Sci. Paris Sér. I Math **305**(19), 805–808 (1987)
20. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**, 375–417 (1991)

21. Brenier, Y.: Extended Monge–Kantorovich theory. In: Optimal transportation and Applications, 91–121, 2003, Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 2–8 (2001)
22. Brezis, H.: Opérateurs maximaux monotones et semi-groupes de contractions dans les spaces de Hilbert. North-Holland Mathematics Studies, Amsterdam (1973)
23. Burago, Yu., Gromov, M., Perelman G.: A. D. Alexandrov spaces with curvatures bounded below. Uspekhi Mat. Nauk **47**, 3–51 (1992); English translation: Russ. Math. Surv. **47**, 1–58 (1992)
24. Buttazzo, G.: Semicontinuity, Relaxation, and Integral Representation in the Calculus of Variations. Longman Scientific and Technical, New York (1989)
25. Cancès, C., Gallouët, T., Monsaingeon, L.: Incompressible immiscible multiphase flows in porous media: a variational approach https://arxiv.org/abs/1607.04009 (2016) **(preprint)**
26. Carrillo, J.-A., Di Francesco, M., Figalli, A., Laurent, T., Slepčev, D.: Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. Duke Math J. **156**(2), 229–271 (2011)
27. Carrillo, J.-A., McCann, R.J., Villani, C.: Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Rev. Mat. Iberoam. **19**, 1–48 (2003)
28. Carrillo, J.-A., McCann, R.J., Villani, C.: Contractions in the 2-Wasserstein length space and thermalization of granular media. Arch. Ration. Mech. Anal. **179**, 217–263 (2006)
29. Carrillo, J.-A., Slepčev, D.: Example of a displacement convex functional of first order. Calc. Var. Partial Differ. Equ. **36**(4), 547–564 (2009)
30. Cheeger, J.: Differentiability of Lipschitz functions on metric measure spaces. Geom. Funct. Anal. **9**, 428–517 (1999)
31. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.-X.: Unbalanced optimal transport: geometry and Kantorovich formulation. arXiv preprint arXiv:1508.05216 (2015)
32. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.-X.: An interpolating distance between optimal transport and Fisher–Rao. arXiv preprint arXiv:1506.06430 (2015)
33. Choné, P., Le Meur, H.: Non-convergence result for conformal approximation of variational problems subject to a convexity constraint. Numer. Funct. Anal. Optim. **5–6**(22), 529–547 (2001)
34. Craig, K.: Nonconvex gradient flow in the Wasserstein metric and applications to constrained nonlocal interactions, preprint available at arXiv:1512.07255
35. Daneri, S., Savaré, G.: Eulerian calculus for the displacement convexity in the Wasserstein distance. SIAM J. Math. Anal. **40**, 1104–1122 (2008)
36. De Giorgi, E.: New problems on minimizing movements. In: Baiocchi, C., Lions, J.L. (eds.) Boundary Value Problems for PDE and Applications, pp. 81–98. Masson, Paris (1993)
37. Di Marino, S., Maury, B., Santambrogio, F.: Measure sweeping processes. J. Convex Anal. **23**(1), 567–601 (2016)
38. Di Marino, S., Mészáros, A.R.: Uniqueness issues for evolution equations with density constraints. Math. Models Methods Appl. Sci. **26**(09), 1761–1783 (2016)
39. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**(2), 324–353 (1974)
40. Ekeland, I., Moreno-Bromberg, S.: An algorithm for computing solutions of variational problems with global convexity constraints. Numerische Mathematik **115**(1), 45–69 (2010)
41. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. Society for Industrial and Applied Mathematics, Classics in Mathematics, Philadelphia (1999)
42. Figalli, A., Gigli, N.: A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. J. Math. Pures Appl. **94**(2), 107–130 (2010)
43. Fortin, M., Glowinski, R.: Augmented Lagrangian Methods Applications to the Numerical Solution of Boundary-Value Problems. North-Holland, Amsterdam (1983)
44. Gallouët, T., Monsaingeon, L.: A JKO splitting scheme for Kantorovich-Fisher–Rao gradient flows, preprint available at arXiv:1602.04457
45. Gangbo, W.: An elementary proof of the polar factorization of vector-valued functions. Arch. Ration. Mech. Anal. **128**, 381–399 (1994)
46. Gangbo, W.: The Monge mass transfer problem and its applications. Contemp. Math. **226**, 79–104 (1999)
47. Gangbo, W., McCann, R.: The geometry of optimal transportation. Acta Math. **177**, 113–161 (1996)
48. Gigli, N.: On the Heat flow on metric measure spaces: existence, uniqueness and stability. Calc. Var. Partial Differ. Equ. **39**, 101–120 (2010)

49. Gigli, N.: Propriétés géométriques et analytiques de certaines structures non lisses. Mémoire HDR, Univ. Nice-Sophia-Antipolis (2011)
50. Gigli, N., Kuwada, K., Ohta, S.I.: Heat flow on Alexandrov spaces. Commun. Pure Appl. Math. **LXVI**, 307–331 (2013)
51. Hajłasz, P.: Sobolev spaces on an arbitrary metric space. Potential Anal. **5**, 403–415 (1996)
52. Hajłasz, P.: Sobolev spaces on metric-measure spaces. Contemp. Math. **338**, 173–218 (2003)
53. Hajłasz, P., Koskela, P.: Sobolev met Poincaré. Mem. Am. Math. Soc. **688**, 1–101 (2000)
54. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
55. Kantorovich, L.: On the transfer of masses. Dokl. Acad. Nauk. USSR **37**, 7–8 (1942)
56. Keller, E.F., Segel, L.A.: Initiation of slide mold aggregation viewed as an instability. J. Theor. Biol. **26**, 399–415 (1970)
57. Keller, E.F., Segel, L.A.: Model for chemotaxis. J. Theor. Biol. **30**, 225–234 (1971)
58. Kinderlehrer, D., Walkington, N.J.: Approximation of parabolic equations using the Wasserstein metric. ESAIM Math. Model. Numer. Anal. **33**(04), 837–852 (1999)
59. Kitagawa, J., Mérigot, Q., Thibert, B.: Convergence of a Newton algorithm for semi-discrete optimal transport. J. Eur. Math. Soc. (2016). Accepted 2017
60. Kondratyev, S., Monsaingeon, L., Vorotnikov, D.: A new optimal transport distance on the space of finite radon measures. arXiv preprint arXiv:1505.07746 (2015)
61. Legendre, G., Turinici, G.: Second order in time schemes for gradient flows in Wasserstein and geodesic metric spaces https://hal.archives-ouvertes.fr/hal-01317769/document (2016) **(preprint)**
62. Lévy, B.: A numerical algorithm for L^2 semi-discrete optimal transport in 3D. ESAIM Math. Model. Numer. Anal **49**(6), 1693–1715 (2015)
63. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. arXiv preprint arXiv:1508.0794 (2015)
64. Liero, M., Mielke, A., Savaré, G.: Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. arXiv preprint arXiv:1509.00068 (2015)
65. Lott, J., Villani, C.: Ricci curvature for metric-measure spaces via optimal transport. Ann. Math. **169**, 903–991 (2009)
66. Matthes, D., McCann, R.J., Savaré, G.: A family of nonlinear fourth order equations of gradient flow type. Commun. Partial Differ. Equ. **34**(10–12), 1352–1397 (2009)
67. Maury, B., Roudneff-Chupin, A., Santambrogio, F.: A macroscopic crowd motion model of gradient flow type. Math. Models Methods Appl. Sci. **20**(10), 1787–1821 (2010)
68. Maury, B., Roudneff-Chupin, A., Santambrogio, F., Venel, J.: Handling congestion in crowd motion modeling. Net. Het. Media **6**(3), 485–519 (2011)
69. Maury, B., Venel, J.: A mathematical framework for a crowd motion model. C. R. Acad. Sci. Paris, Ser. I **346**, 1245–1250 (2008)
70. Maury, B., Venel, J.: A discrete contact model for crowd motion. ESAIM Math. Model. Numer. Anal. **45**(1), 145–168 (2011)
71. McCann, R.J.: Existence and uniqueness of monotone measure preserving maps. Duke Math. J. **80**, 309–323 (1995)
72. McCann, R.J.: A convexity principle for interacting gases. Adv. Math. **128**(1), 153–159 (1997)
73. McCann, R.J.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. **11**(3), 589–608 (2001)
74. Mérigot, Q.: A multiscale approach to optimal transport. Comput. Gr. Forum **30**(5), 1583–1592 (2011)
75. Mérigot, Q., Oudet, É.: Handling convexity-like constraints in variational problems. SIAM J. Numer. Anal. **52**(5), 2466–2487 (2014)
76. Mészáros, A.R., Santambrogio, F.: Advection–diffusion equations with density constraints. Anal. PDE **9–3**, 615–644 (2016)
77. Monge, G.: Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris 666–704 (1781)
78. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**, 101–174 (2011)
79. von Renesse, M.-K., Sturm, K.-T.: Entropic measure and Wasserstein diffusion. Ann. Probab. **37**, 1114–1191 (2009)
80. Rockafellar, R.T.: Convex Anal. Princeton University Press, Princeton (1970)

81. Roudneff-Chupin, A.: Modélisation macroscopique de mouvements de foule, PhD Thesis, Université Paris-Sud (2011)
82. Santambrogio, F.: Gradient flows in Wasserstein spaces and applications to crowd movement, Séminaire Laurent Schwartz no 27, École Polytechnique (2010)
83. Santambrogio, F.: Flots de gradient dans les espaces métriques et leurs applications (d'après Ambrosio–Gigli–Savaré). In: Proceedings of the Bourbaki Seminar (2013) **(in French)**
84. Santambrogio, F.: Optimal Transport for Applied Mathematicians, Progress in Nonlinear Differential Equations and Their Applications no 87, Birkhäuser Basel (2015)
85. Shanmugalingam, N.: Newtonian spaces: an extension of Sobolev spaces to metric measure spaces. Rev. Mat. Iberoam. **16**, 243–279 (2000)
86. Schneider, R.: Convex Bodies: The Brunn–Minkowski Theory, vol. 44. Cambridge University Press, Cambridge (1993)
87. Sturm, K.-T.: On the geometry of metric measure spaces. I. Acta Math. **196**, 65–131 (2006)
88. Sturm, K.-T.: On the geometry of metric measure spaces. II. Acta Math. **196**, 133–177 (2006)
89. Vázquez, J.L.: The Porous Medium Equation. Mathematical Theory. Oxford University Press, Oxford (2007)
90. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics. AMS, Providence (2003)
91. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2008)