

# {Euclidean, Metric, and Wasserstein} Gradient Flows: an overview

Filippo Santambrogio\*

## 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, with proof of convergence in the easiest case: the linear Fokker-Planck equation. A discussion of other 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.

**AMS Subject Classification (2010):** 00-02, 34G25, 35K05, 49J45, 49Q20, 49M29, 54E35

**Keywords:** Cauchy problem, Subdifferential, Analysis in metric spaces, Optimal transport, Wasserstein distances, Heat flow, Fokker-Planck equation, Numerical methods, Contractivity, Metric measure spaces

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>From Euclidean to Metric</b>	<b>4</b>
2.1	Gradient flows in the Euclidean space . . . . .	4
2.2	An introduction to the metric setting . . . . .	9
<b>3</b>	<b>The general theory in metric spaces</b>	<b>13</b>
3.1	Preliminaries . . . . .	13
3.2	Existence of a gradient flow . . . . .	14
3.3	Uniqueness and contractivity . . . . .	17

---

\* Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay Cedex, France, [filippo.santambrogio@math.u-psud.fr](mailto:filippo.santambrogio@math.u-psud.fr), <http://www.math.u-psud.fr/~santambr>

<b>4</b>	<b>Gradient flows in the Wasserstein space</b>	<b>19</b>
4.1	Preliminaries on Optimal Transport . . . . .	19
4.2	The Wasserstein distances . . . . .	24
4.3	Minimizing movement schemes in the Wasserstein space and evolution PDEs . . . . .	29
4.4	Geodesic convexity in $\mathbb{W}_2$ . . . . .	32
4.5	Analysis of the Fokker-Planck equation as a gradient flow in $\mathbb{W}_2$ . . . . .	37
4.6	Other gradient-flow PDEs . . . . .	42
4.7	Dirichlet boundary conditions . . . . .	49
4.8	Numerical methods from the JKO scheme . . . . .	51
<b>5</b>	<b>The heat flow in metric measure spaces</b>	<b>56</b>
5.1	Dirichlet and Cheeger energies in metric measure spaces . . . . .	57
5.2	A well-posed gradient flow for the entropy . . . . .	58
5.3	Gradient flows comparison . . . . .	59
	<b>Bibliography</b>	<b>60</b>

## 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  $x$  minimizing  $F$  (which is related to the statical 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 statical 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 20th century and the beginning of the 21st, with the work of Jordan, Kinderlehrer and Otto ([55]) and then of Otto [76], 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  $\varrho$  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 in the 18th century ([75]), then developed by Kantorovich in the '40s ([56]), is now well-established (many texts present it, such as [89, 90, 84]) 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 for sure the publication of [6]. 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  $\nabla F$  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 [6] 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 [69] 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 ([49, 51, 9]), 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-Kinderlehrer-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 and full proofs are given in the easiest case.

The paper is organized as follows: after this introduction, Section 2 exposes the theory in the Euclidean case, and presents which are the good definitions which can be translated into a metric setting; Section 3 is devoted to the general metric setting, as in the first half of [6], and is quite expository (only the key ingredients to obtain the proofs are sketched); Section 4 is the longest one and develops the Wasserstein setting: after an introduction to optimal transport and to the Wasserstein distances, there is an informal presentation of the equations that can be obtained as gradient flows, a discussion of the functionals which have geodesic convexity properties, a quite precise proof of convergence in the linear case of the Fokker-Planck equation, a discussion about the other equations and functionals which fit the framework and about boundary conditions, and finally a short section about numerics. Last but not least, Section 5 gives a very short presentation of the fascinating topic of heat flows in arbitrary 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 gradient flows in the Wasserstein space and in PDE applications can decide to skip sections 3, 4.4 and 5, which deal on the contrary with key objects for the - very lively at the moment - subject of analysis on metric measure spaces.

## 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  $F : \mathbb{R}^n \rightarrow \mathbb{R}$ , 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*

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

This is a standard Cauchy problem which has a unique solution if  $\nabla F$  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  $x : [0, T] \rightarrow \mathbb{R}^n$  such that

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

where  $\partial F(x) = \{p \in \mathbb{R}^n : F(y) \geq F(x) + p \cdot (y - x) \text{ for all } y \in \mathbb{R}^n\}$ . 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  $\partial F$  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 we have  $|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|$  for every  $t$ . In particular this gives uniqueness of the solution of the Cauchy problem.*

*Proof.* Let us consider  $g(t) = \frac{1}{2}|x_1(t) - x_2(t)|^2$  and differentiate it. We have

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

Here we use a basic property of gradient of convex functions, i.e. that for every  $x_1, x_2, p_1, p_2$  with  $p_i \in \partial F(x_i)$ , we have

$$(x_1 - x_2) \cdot (p_1 - p_2) \geq 0.$$

From these considerations, we obtain  $g'(t) \leq 0$  and  $g(t) \leq g(0)$ . This gives the first part of the claim.

Then, if we take two different solutions of the same Cauchy problem, we have  $x_1(0) = x_2(0)$ , and this implies  $x_1(t) = x_2(t)$  for any  $t > 0$ .  $\square$

We can also study the case where  $F$  is semi-convex. We recall that  $F$  semi-convex means that it is  $\lambda$ -convex for some  $\lambda \in \mathbb{R}$  i.e.  $x \mapsto F(x) - \frac{\lambda}{2}|x|^2$  is convex. For  $\lambda > 0$  this is stronger than convexity, and for  $\lambda < 0$  this is weaker. Roughly speaking,  $\lambda$ -convexity corresponds to  $D^2F \geq \lambda I$ . Functions which are  $\lambda$ -convex for some  $\lambda$  are called *semi-convex*. The reason of the interest towards 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  $\lambda$ -convex for a suitable  $\lambda < 0$ .

For  $\lambda$ -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  $\partial F$ .

**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  $\lambda$ -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  $\lambda$ -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  $\lambda$ -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  $\lambda$ -convex and let  $x$  be a solutions 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  $\{p : p \cdot x'(t_0) = \text{const}\}$ .

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.  $\square$

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*,

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

We can forget now the convexity assumption on  $F$ , which are not necessary for this part of the analysis. Indeed, very mild assumptions on  $F$  (l.s.c. and some lower bounds, for instance  $F(x) \geq C_1 - C_2|x|^2$ ) are sufficient to guarantee that these problems admit a solution for small  $\tau$ . The case where  $F$  is  $\lambda$ -convex is covered by these assumptions, and also provides uniqueness of the minimizers. This is evident if  $\lambda > 0$  since we have strict convexity for every  $\tau$ , and if  $\lambda$  is negative the sum will be strictly convex for small  $\tau$ .

We can interpret this sequence of points as the values of the curve  $x(t)$  at times  $t = 0, \tau, 2\tau, \dots, k\tau, \dots$ . It happens that the optimality conditions of the recursive minimization exactly give a connection between these minimization problems and the equation, since we have

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

i.e.

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

This expression 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-discretization where derivatives are replaced by finite differences. We fix a time step  $\tau > 0$  and define a sequence  $x_k^\tau$ . The explicit scheme is given by

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

while the implicit scheme is given by

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

This means that  $x_{k+1}^\tau$  is selected as a solution of an equation involving  $x_k^\tau$ , instead of being explicitly computable from  $x_k^\tau$ . 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,  $\varrho_{k+1}^\tau$  would have two degrees of regularity less than  $\varrho_k^\tau$ , 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  $\tau \rightarrow 0$ , 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 Section 4.

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

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

then we set

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

Also set

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

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,  $x^\tau$  is not continuous, but satisfies by construction  $\mathbf{v}^\tau(t) \in -\partial F(x^\tau(t))$ .

The iterated minimization scheme defining  $x_{k+1}^\tau$  provides the estimate

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

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

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

This is valid for every  $\ell$ , 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

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

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

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

This also implies

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

since  $x^\tau(t) = \tilde{x}^\tau(s)$  for a certain  $s = k\tau$  with  $|s - t| \leq \tau$ .

This provides the necessary compactness to prove the following.

**Proposition 2.3.** *Let  $\tilde{x}^\tau$ ,  $x^\tau$  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  $\tau_j \rightarrow 0$  (still denoted by  $\tau$ ), both  $\tilde{x}^\tau$  and  $x^\tau$  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  $\lambda$ -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,  $x^\tau$  also converges uniformly to the same limit, that we will call  $x = [0, T] \rightarrow \mathbb{R}^n$ . Then,  $\mathbf{v}^\tau = (\tilde{x}^\tau)'$  and (2.6) allow 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

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

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

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

We can pass to the limit as  $\tau \rightarrow 0$ , using the uniform (hence  $L^2$  strong) convergence  $x^\tau \rightarrow x$  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

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

The first term in the equality uniformly converges, as a function of  $t$ , (since  $\nabla F$  is continuous and  $x^\tau$  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.  $\square$

In the above result, we only proved convergence of the curves  $x^\tau$  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 [6] which will be the basis for



the metric case, also provides explicit estimates; these estimates are usually of order  $\tau$ . An interesting observation, in the Euclidean case, is that if the sequence  $x_k^\tau$  is defined by

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

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  $\tau^2$ . This has been used in the Wasserstein case<sup>1</sup> (see Section 4 and in particular Section 4.8) 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, a huge advantage of this discretized formulation is also that it can easily be adapted to metric spaces. Actually, if one has a metric space  $(X, d)$  and a l.s.c. function  $F : X \rightarrow \mathbb{R} \cup \{+\infty\}$  (under suitable compactness assumptions to guarantee existence of the minimum), one can define

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

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

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

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

De Giorgi, in [38], defined what he called *Generalized Minimizing Movements*<sup>2</sup>:

**Definition 2.1.** A curve  $x : [0, T] \rightarrow X$  is called *Generalized Minimizing Movements (GMM)* if there exists a sequence of time steps  $\tau_j \rightarrow 0$  such that the sequence of curves  $x^{\tau_j}$  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 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. Yet, we can also obtain a Hölder-like estimate from the piecewise constant interpolation.

We start from

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

<sup>1</sup>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  $x_k^\tau$  and a point  $x$  (on a sphere, for instance, this means that if  $x_k^\tau$  is the North Pole, then  $y$  must lie in the northern hemisphere).

<sup>2</sup>We prefer not to make any distinction here between Generalized Minimizing Movements and Minimizing Movements.

and

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

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

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

This means that the curves  $x^\tau$  - if we forget that they are discontinuous - are morally equi-hölder continuous with exponent  $1/2$  (up to a negligible error of order  $\sqrt{\tau}$ ), 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  $\omega$  is a continuous function defined on a 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 speed  $\omega'(t)$  has no meaning, unless  $X$  is a vector space. Surprisingly, it is possible to give a meaning to the modulus of the velocity,  $|\omega'| (t)$ .

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

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

provided this limit exists.

In the spirit of Rademacher Theorem, it is possible to prove (see [12]) that, if  $\omega : [0, 1] \rightarrow X$  is Lipschitz continuous, then the metric derivative  $|\omega'| (t)$  exists for a.e.  $t$ . Moreover we have, for  $t_0 < t_1$ ,

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

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

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

It is well-known that every absolutely continuous curve can be reparametrized in time (through a monotone-increasing reparametrization) and become Lipschitz continuous, and the existence of the metric derivative for a.e.  $t$  is also true for  $\omega \in AC(X)$ , 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  $\omega : [0, 1] \rightarrow X$ , let us define

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

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

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

We collect now some more definitions.

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

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

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

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

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

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

In a length space, a curve  $\omega : [t_0, t_1] \rightarrow X$  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.  $\omega$  is a constant-speed geodesic defined on  $[t_0, t_1]$  and joining  $x_0$  and  $x_1$ ,
2.  $\omega \in \text{AC}(X)$  and  $|\omega'(t)| = \frac{d(\omega(t_0), \omega(t_1))}{t_1 - t_0}$  a.e.,
3.  $\omega$  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  $x^\tau : [0, T] \rightarrow X$  such that  $x^\tau(k\tau) = x_k^\tau$  and such that  $x^\tau$  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  $x^\tau$  (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  $\tau \rightarrow 0$ , 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!

If the space  $X$ , the distance  $d$ , and the functional  $F$  are explicitly known, in some cases 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)  $x(t)$ . It will be possible to do so in the framework of probability measures, as it will be discussed in Section 4, but not in general. Indeed, without a little bit of (differential) structure on the space  $X$ , it is essentially impossible 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 [6], 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 Cauchy-Schwartz inequality, for every curve we have

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

Here, the first inequality is an equality if and only if  $x'(r)$  and  $\nabla F(x(r))$  are vectors with opposite directions for a.e.  $r$ , and the second is an equality if and only if their norms are equal. 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 ([6]) 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  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $\lambda$ -convex, the inequality that characterizes the gradient is

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

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  $\lambda$ , it should be noted that the condition EVI should actually be written as  $\text{EVI}_\lambda$ , since it involves a parameter  $\lambda$ , which is a priori arbitrary. Actually,  $\lambda$ -convexity of  $F$  is not necessary to define the  $\text{EVI}_\lambda$  property, but it will be necessary in order to guarantee the existence of curves which satisfy such a condition. The notion of  $\lambda$ -convexity will hence be crucial also in metric spaces, where it will be rather “ $\lambda$ -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

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

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

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  $x : [0, T] \rightarrow X$  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:

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

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* every function  $g : X \rightarrow \mathbb{R}$  such that, for every Lipschitz curve  $x$ , we have

$$|F(x(0)) - F(x(1))| \leq \int_0^1 g(x(t)) |x'(t)| dt.$$

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

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

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:

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

(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 (Section 5) we will see how to define a Sobolev space  $H^1$  on a (measure) metric 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<sup>3</sup> 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  $\lambda$ -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

---

<sup>3</sup>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  $F < +\infty$ .

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

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

and call  $x(\theta)$  any minimizer for this problem, and  $\varphi(\theta)$  the minimal value. It’s clear that, for  $\theta \rightarrow 0^+$ , we have  $x(\theta) \rightarrow x_k^\tau$  and  $\varphi(\theta) \rightarrow F(x_k^\tau)$ , and that, for  $\theta = 1$ , we get back to the original problem with minimizer  $x_{k+1}^\tau$ . Moreover, the function  $\varphi$  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  $\theta$  such that  $\varphi'(\theta)$  exists. Moreover, the optimality condition 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

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

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

We now come back to the function  $\varphi$  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):

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

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

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

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  $F(x(0)) - F(x(t)) \leq \int_0^t |\nabla^- F(x(r))| |x'(r)| dr$  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  $s < t$ :

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

Magically, it happens that the assumption that  $F$  is  $\lambda$ -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 [6, 5]. 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  $\lambda$ -geodesically convex, then we have

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

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  $y \rightarrow x$  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<sup>4</sup>. It is also possible (see [5], for instance) to prove that the slope is an upper gradient.

Let us insist anyway on the fact that the  $\lambda$ -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 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 Section 4), without passing through the general theory and the EDE condition.

---

<sup>4</sup>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.



### 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 [5]), 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  $\ell^\infty$  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 Section 3.2.
- The EVI condition is in general too strong in order to get existence (in the example above of the  $\ell^\infty$  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 [37] that, if  $F$  is such that from every starting point  $x_0$  there exists an  $\text{EVI}_\lambda$  gradient flow, then  $F$  is necessarily  $\lambda$ -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

$$\begin{aligned} \frac{d}{dt}d(x(t), y(t_0))^2 \Big|_{t=t_0} &\leq -\lambda d(x(t_0), y(t_0))^2 + 2F(y(t_0)) - 2F(x(t_0)) \\ \frac{d}{ds}d(x(t_0), y(s))^2 \Big|_{s=t_0} &\leq -\lambda d(x(t_0), y(t_0))^2 + 2F(x(t_0)) - 2F(y(t_0)) \end{aligned}$$

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

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

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 [6], 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 [41]). Here, we will just present this extra convexity assumption, needed for the existence of EVI gradient flows developed in [6].

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

$$\begin{aligned} F(x(t)) &\leq (1-t)F(x_0) + tF(x_1) - \lambda \frac{t(1-t)}{2} d^2(x_0, x_1), \\ d^2(x(t), y) &\leq (1-t)d^2(x_0, y) + td^2(x_1, y) - t(1-t)d^2(x_0, x_1). \end{aligned}$$

In other words, we require  $\lambda$ -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 Section 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  $\tau$  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 existence of an EVI gradient flow, has been done in view of the applications to the case of the Wasserstein spaces, that will 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- $\lambda$ -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  $\omega$  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  $\lambda$ -convex,

but almost, and this is the theory developed in [36]. Such a paper studies the case where  $F$  satisfies some sort of  $\omega$ -convexity for a “modulus of convexity”  $\omega$ . More precisely, this means

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

on generalized geodesics  $x_t$  (note that in the case  $\omega(s) = s$  we come back to (3.2)). The function  $\omega$  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  $L^\infty$  constraints on the density (using the fact that the gradient  $\nabla u$  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<sup>5</sup>, in what concerns applied mathematics) of the theory of gradient flows in metric spaces is for sure that of evolution PDEs in the space of measures. This topic is inspired from the work by Jordan, Kinderlehrer and Otto ([55]), 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 not mean that proofs in [55] 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 C. Villani ([89, 90]) and in the book on gradient flows by Ambrosio, Gigli and Savaré [6]<sup>6</sup>; 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 in 1781 ([75]): given two densities of mass  $f, g \geq 0$  on  $\mathbb{R}^d$ , with  $\int f = \int g = 1$ , find a map  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$  pushing the first one onto the other, i.e. such that

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

<sup>5</sup>This is for sure 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.

<sup>6</sup>Lighter versions exist, such as [11], or the recent *User's Guide to Optimal Transport* ([5]), 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 Section 3), and finally for one third with metric spaces with curvature bounds (that we will briefly sketch in Section 5).

and minimizing the quantity

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

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  $\mu$  on  $X$  (measures will indeed replace the densities  $f$  and  $g$  in the most general formulation of the problem) through a measurable map  $T : X \rightarrow Y$ : it is the measure denoted by  $T_{\#}\mu$  on  $Y$  and characterized by

$$(T_{\#}\mu)(A) = \mu(T^{-1}(A)) \quad \text{for every measurable set } A,$$

$$\text{or } \int_Y \phi d(T_{\#}\mu) = \int_X \phi \circ T d\mu \quad \text{for every measurable function } \phi.$$

The problem of Monge has stayed with no solution (does a minimizer exist? how to characterize it?... ) till the progress made in the 1940s with the work by Kantorovich ([56]). 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<sup>7</sup> and a cost function  $c : X \times X \rightarrow [0, +\infty]$ . 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

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

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  $\pi_0$  and  $\pi_1$  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  $\mu$ : 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  $\mu$  onto  $\nu$  (i.e.  $T_{\#}\mu = \nu$ ) and the functional takes the form  $\int c(x, T(x)) d\mu(x)$ , thus generalizing Monge's problem.

---

<sup>7</sup>Most of the results that we present stay true without this assumptions, anyway, and we refer in particular to [6] or [89] for details, since a large part of the analysis of [84] is performed under this simplifying assumption.

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

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  $\nu$  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]).

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  $\gamma$  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  $\gamma$  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  $\gamma$  and the sup in  $\phi, \psi$ :

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

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 [89], where the proof is based on Fenchel-Rockafellar duality (see, for instance, [43]). Anyway, we insist that in this case it is true that  $\inf \sup = \sup \inf$ .

Afterwards, one can re-write the inf in  $\gamma$  as a constraint on  $\phi$  and  $\psi$ , 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  $\mu$  and  $\nu$  and the cost function  $c : X \times X \rightarrow [0, +\infty]$  we consider the problem

$$(DP) \quad \max \left\{ \int_X \phi \, d\mu + \int_X \psi \, d\nu : \phi, \psi \in C(X), \phi(x) + \psi(y) \leq c(x, y) \text{ for all } (x, y) \in X \times X \right\}. \quad (4.3)$$

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  $\chi : X \rightarrow \overline{\mathbb{R}}$  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  $\psi$  is  $c$ -concave if there exists  $\chi$  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  $(\phi, \psi)$  in the maximization problem (DP), one can always replace it with  $(\phi, \phi^c)$ , and then with  $(\phi^{cc}, \phi^c)$ , 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  $\phi$ . This is the same as saying that  $\psi^{cc} = \psi$  for any  $c$ -concave function  $\psi$ , 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

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

where the max on the right hand side is realized. In particular the minimum value of (KP) is a convex function of  $(\mu, \nu)$ , as it is a supremum of linear functionals.

**Definition 4.2.** The functions  $\phi$  realizing the maximum in (4.4) are called Kantorovich potentials for the transport from  $\mu$  to  $\nu$  (and will be often denoted by the symbol  $\varphi$  instead of  $\phi$ ).

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

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

and

$$\phi \in \Psi_c(X) \Rightarrow \phi^c = -\phi. \quad (4.6)$$

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

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

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

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

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  $\gamma$  and a Kantorovich potential  $\varphi$  and write

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

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

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

and, if  $\varphi$  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  $(x_0, y_0) \in \text{spt}\gamma$ , which means that  $\gamma$  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 ([23, 24]). 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  $\mu$ -a.e. differentiability of the potential  $\varphi$ . Since  $\varphi$  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  $\mu$  and  $\nu$  probability measures on a domain  $\Omega \subset \mathbb{R}^d$  there exists an optimal transport plan  $\gamma$  for the quadratic cost  $\frac{1}{2}|x - y|^2$ . It is unique and of the form  $(\text{id}, T)_\# \mu$ , provided  $\mu$  is absolutely continuous. Moreover there exists also at least a Kantorovich potential  $\varphi$ , and the gradient  $\nabla\varphi$  is uniquely determined  $\mu$ -a.e. (in particular  $\varphi$  is unique up to additive constants if the density of  $\mu$  is positive a.e. on  $\Omega$ ). The optimal transport map  $T$  and the potential  $\varphi$  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: 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  $\varphi$  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  $\mu$  and  $\nu$ : hence the uniqueness of the Kantorovich potential is true even if it  $\nu$  (and not  $\mu$ ) has positive density a.e. (since one can retrieve  $\varphi$  from  $\varphi^c$  and viceversa).

We stress that Theorem 4.2 admits a converse implication and that any gradient of a convex function is indeed optimal between  $\mu$  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  $\mu$  and another measure  $\nu$ , 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  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , 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  $\Omega$  is unbounded we need to restrict our analysis to the following set of probabilities

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

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

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

(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

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

The quantities that we obtain in this way are called Wasserstein distances<sup>8</sup>. 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.

---

<sup>8</sup>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 [55, 76], even if other names have been suggested, such as Monge-Kantorovich distances, Kantorovich-Rubinstein... and we will stick to this terminology.



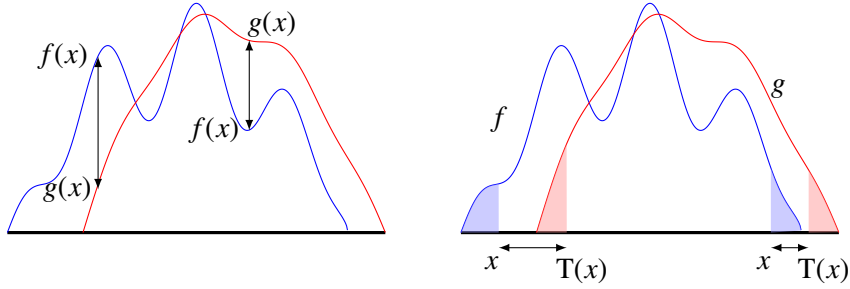


Figure 1: “Vertical” vs “horizontal” distances (the transport  $T$  is computed in the picture on the right using 1D considerations, imposing equality between the blue and red areas under the graphs of  $f$  and  $g$ ).

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  $h \rightarrow \infty$ ) and depends much more on the displacement than on the shapes of  $f$  and  $g$ .

We now analyze some properties of these distances. Most proofs can be found in [84], chapter 5, or in [89] or [6].

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  $\Omega$  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  $\Omega$  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  $\phi$ , 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  $\mu, \varrho$  and  $\nu$ , and suppose that  $\mu$  and  $\varrho$  are absolutely continuous. Let  $T$  be the optimal transport from  $\mu$  to  $\varrho$  and  $S$  the optimal one from  $\varrho$  to  $\nu$ . Then  $S \circ T$  is an admissible transport from  $\mu$  to  $\nu$ , since  $(S \circ T)_\# \mu = S_\#(T_\# \mu) = S_\# \varrho = \nu$ . We have

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

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).  $\square$

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  $\mu$  and a sequence  $(\mu_n)_n$  in  $\mathbb{W}_p(X)$ , the following are equivalent:*

- $\mu_n \rightarrow \mu$  according to  $W_p$ ;
- $\mu_n \rightarrow \mu$  and  $m_p(\mu_n) \rightarrow m_p(\mu)$ ;
- $\int_X \phi d\mu_n \rightarrow \int_X \phi d\mu$  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  $\phi$  such that  $|\phi(x)| \leq A + Bd(x, x_0)^p$  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  $\mu_t$  of a family of particles initially distributed according to  $\mu_0$  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 [6]) 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 [6] (but the reader can also find alternative proofs in Chapter 5 of [84], in the case where  $\Omega$  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  $\mu_t$  and  $\mu_{t+h}$  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  $h \rightarrow 0$ .

Once we know about curves in their generality, it is interesting to think about geodesics. The following result is a characterization of geodesics in  $\mathbb{W}_p(\Omega)$  when  $\Omega$  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  $\mu$  and  $\nu$  belong to  $\mathbb{W}_p(\Omega)$ , and  $\gamma$  is an optimal transport plan from  $\mu$  to  $\nu$  for the cost  $c_p(x, y) = |x - y|^p$ , then the curve*

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

where  $\pi_t : \Omega \times \Omega \rightarrow \Omega$  is given by  $\pi_t(x, y) = (1 - t)x + ty$ , is a constant-speed geodesic from  $\mu$  to  $\nu$ . In the case  $p > 1$  all the constant-speed geodesics are of this form, and, if  $\mu$  is absolutely continuous, then there is only one geodesic and it has the form

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

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

The above theorem may be adapted to the case where the Euclidean domain  $\Omega$  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 [71].

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 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  $\mu$  (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

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

selects constant-speed geodesics connecting  $\mu$  to  $\nu$ , and hence allows to find the optimal transport between these two measures. This is what is usually called *Benamou-Brenier formula* ([14]).

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  $(\varrho_t, E_t)$  where  $E_t = \mathbf{v}_t \varrho_t$ , thus obtaining the following minimization problem

$$\min \left\{ \int_0^1 \int_{\Omega} \frac{|E_t|^p}{\varrho_t^{p-1}} dx dt : \partial_t \varrho_t + \nabla \cdot E_t = 0, \varrho_0 = \mu, \varrho_1 = \nu \right\}. \quad (4.8)$$

We need to use the properties of the function  $f_p : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ , 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 [14]).

### 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:

$$\varrho_{k+1}^\tau \in \operatorname{argmin}_{\varrho} F(\varrho) + \frac{W_2^2(\varrho, \varrho_k^\tau)}{2\tau}. \quad (4.9)$$

Note that we denote now the measures on  $\Omega$  by the letter  $\varrho$  instead of  $\mu$  or  $\nu$  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 Section 2, we saw that the equation  $x' = -\nabla F(x)$  corresponds to a sequence of minimization problems with the squared distance  $|x - x_k^\tau|^2$  (if we change the exponent here we can 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  $p$ , and in the Wasserstein space  $\mathbb{W}_p$  the distance 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 (4.9). 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  $G : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$  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  $\chi$  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  $\varrho$ . In order to understand this notion, the easiest possibility is to analyze some examples.

The three main classes of examples are the following functionals<sup>9</sup>

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

where  $f : \mathbb{R} \rightarrow \mathbb{R}$  is a convex superlinear function (and the functional  $F$  is set to  $+\infty$  if  $\varrho$  is not absolutely continuous w.r.t. the Lebesgue measure) and  $V : \Omega \rightarrow \mathbb{R}$  and  $W : \mathbb{R}^d \rightarrow \mathbb{R}$  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 (4.9), 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  $c : \Omega \times \Omega \rightarrow \mathbb{R}$  be a continuous cost function. Then the functional  $\varrho \mapsto \mathcal{T}_c(\varrho, \nu)$  is convex, and its subdifferential at  $\varrho_0$  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  $\varphi$  from  $\varrho_0$  to  $\nu$  up to additive constants, then we also have  $\frac{\delta \mathcal{T}_c(\cdot, \nu)}{\delta \varrho}(\varrho_0) = \varphi$ .*

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  $\Omega$  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  $\varrho$  and shows convexity. Standard considerations from convex analysis allow to identify the subdifferential as the set of functions  $\varphi$  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  $\phi$ , in which case the first variation is of course  $\phi$ ; 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  $\phi$ .

Once we know how to compute first variations, we come back to the optimality conditions for the minimization problem (4.9). Which are these optimality conditions? roughly speaking, we should have

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

(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  $\varphi$  is the Kantorovich potential associated with the transport from  $\varrho_{k+1}^\tau$  to  $\varrho_k^\tau$  (and not viceversa).

---

<sup>9</sup>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  $\chi$  which make the corresponding functional finite.

More precise statements and proofs of this optimality conditions will be presented in the next section. Here we look at the consequences we can get. Actually, if we combine the fact that the above sum is constant, and that we have  $T(x) = x - \nabla\varphi(x)$  for the optimal  $T$ , we get

$$-\mathbf{v}(x) := \frac{T(x) - x}{\tau} = -\frac{\nabla\varphi(x)}{\tau} = \nabla\left(\frac{\delta F}{\delta\varrho}\right)(x). \quad (4.10)$$

We will denote 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  $\varrho_{k+1}^\tau$  to  $\varrho_k^\tau$ , it is better to view it rather as a backward velocity (which justifies the minus sign).

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

$$\partial_t \varrho - \nabla \cdot \left( \varrho \nabla \left( \frac{\delta F}{\delta \varrho} \right) \right) = 0. \quad (4.11)$$

This is a PDE where the velocity field in the continuity equation depends on the density  $\varrho$  itself. An interesting example is the case where we use  $F(\varrho) = \int f(\varrho(x))dx$ , 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*  $\partial_\varrho - \Delta \varrho = 0$ . Using  $F(\varrho) = \int f(\varrho(x))dx + \int V(x)d\varrho(x)$ , we would have the *Fokker-Planck Equation*  $\partial_\varrho - \Delta \varrho - \nabla \cdot (\varrho \nabla V) = 0$ . We will see later which other interesting PDEs can be obtained in this way.

Many possible proofs can be built for the convergence of the above iterated minimization scheme. In particular, one could follow the general theory developed in [6], 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 [6], 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  $\lambda$ -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 [6] (devoted to the general metric case) are not necessary; if the second part of [6] is exactly concerned with Wasserstein spaces and with the characterization of the limit as  $\tau \rightarrow 0$  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 [6] (which is an advantage of this general theory).



- the  $\lambda$ -convexity assumption is in general not crucial in what concerns existence (but the general theory in [6] has been built under this assumption, as we saw in Section 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 section 4.4). On the other hand, uniqueness results are almost never possible without some kind of  $\lambda$ -convexity (or weaker versions of it, as in [36]) of the functional.

For the reader who wants an idea of how to prove convergence of the scheme to a solution of the PDE independently of the EDE/EVI theory, here are some sketchy ideas. Everything will be developed in details in Section 4.5 in a particular case.

The main point lies in the interpolation of the sequence  $\varrho_k^\tau$  (and of the corresponding velocities  $\mathbf{v}_k^\tau$ ). Indeed, two possible interpolations turn out to be useful: on the one hand, we can define an interpolation  $(\varrho^\tau, \mathbf{v}^\tau)$  which is piecewise constant, as in (2.10) (to define  $\mathbf{v}^\tau$  we use  $\nabla\varphi/\tau$ ); on the other hand, we can connect each measure  $\varrho_k^\tau$  to  $\varrho_{k+1}^\tau$  by using a piecewise geodesic curve  $\tilde{\varrho}^\tau$ , where geodesics are chosen according to the Wasserstein metric, using Theorem (4.7) to get an explicit expression. This second interpolation is a Lipschitz curve in  $\mathbb{W}_2(\Omega)$ , and has an explicit velocity field, that we know thanks to Theorem (4.7): we call it  $\tilde{\mathbf{v}}^\tau$  and it is related to  $\mathbf{v}^\tau$ . The advantage of the second interpolation is that  $(\tilde{\varrho}^\tau, \tilde{\mathbf{v}}^\tau)$  satisfies the continuity equation. On the contrary, the first interpolation is not continuous, the continuity equation is not satisfied, but the optimality conditions at each time step provide a connection between  $\mathbf{v}^\tau$  and  $\varrho^\tau$  ( $\mathbf{v}^\tau = -\nabla(\frac{\delta F}{\delta \varrho}(\varrho^\tau))$ ), which is not satisfied with  $\tilde{\varrho}^\tau$  and  $\tilde{\mathbf{v}}^\tau$ . It is possible to prove that the two interpolations converge to the same limit as  $\tau \rightarrow 0$ , and that the limit will satisfy a continuity equation with a velocity vector field  $\mathbf{v} = -\nabla(\frac{\delta F}{\delta \varrho}(\varrho))$ , which allows to conclude.

#### 4.4 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, for sure it could be interesting to check whether the general theory of Section 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  $\lambda$ -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 [69].

**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

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



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

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

Similarly, if  $V$  or  $W$  are  $\lambda$ -convex we get  $\lambda$ -geodesical convexity. Note that in the case of  $\mathcal{V}$  it is easy to see that the  $\lambda$ -convexity of  $V$  is also necessary for the  $\lambda$ -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  $T : \Omega \rightarrow \Omega$  is a map smooth enough<sup>10</sup> and injective, and  $\det(DT(x)) \neq 0$  for a.e.  $x \in \Omega$ , if we set  $\varrho_T := T_{\#}\varrho$ , then  $\varrho_T$  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  $A$  be a  $d \times d$  matrix such that its eigenvalues  $\lambda_i$  are all real and larger than  $-1$  (for instance this is the case when  $A$  is symmetric and  $I + A \geq 0$ ). Then  $[0, 1] \ni t \mapsto g(t) := \det(I + tA)^{1/d}$  is concave.*

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  $\Omega$  is convex and take  $1 < p < \infty$ . Then  $\mathcal{F}$  is geodesically convex in  $\mathbb{W}_2$ .*

*Proof.* Let us consider two measures  $\mu_0, \mu_1$  with  $\mathcal{F}(\mu_0), \mathcal{F}(\mu_1) < +\infty$ . They are absolutely continuous and hence there is a unique constant-speed geodesic  $\mu_t$  between them, which has the form  $\mu_t = (T_t)_{\#}\mu_0$ , where  $T_t = \text{id} + t(T - \text{id})$ . Note that we have  $T_t(x) = x - t\nabla\varphi(x)$ , where  $\varphi$  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  $\mu_t$  is absolutely continuous and we can write its density  $\varrho_t$  as  $\varrho_t(y) = \varrho(T_t^{-1}(y)) / \det(I + tA(T_t^{-1}(y)))$ , where  $A = -D^2\varphi$  and  $\varrho$  is the density of  $\mu$ , and hence

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

where we used the change of variables  $y = T_t(x)$  and  $dy = \det DT_t(x) dx = \det(I + tA(x)) dx$ .

From Lemma 4.9 we know that  $\det(I + tA(x)) = g(t, x)^d$  for a function  $g : [0, 1] \times \Omega$  which is concave in  $t$ . It is a general fact that the composition of a convex and decreasing function with a concave one gives a convex function. This implies that

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

<sup>10</sup>We need at least  $T$  to be countably Lipschitz, i.e.  $\Omega$  may be written as a countable union of measurable sets  $(\Omega_i)_{i \geq 0}$  with  $\Omega_0$  negligible and  $T|_{\Omega_i}$  Lipschitz continuous for every  $i \geq 1$ .

is convex (if  $\varrho(x) \neq 0$  this uses the assumption on  $f$  and the fact that  $t \mapsto g(t, x)/\varrho(x)^{\frac{1}{d}}$  is concave; if  $\varrho(x) = 0$  then this function is simply zero). 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  $\mathcal{F}$ .

Let us see some easy examples of convex functions satisfying the assumptions of Theorem 4.10:

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

**Convexity on generalized geodesics.** It is quite disappointing to note that the functional  $\mu \mapsto W_2^2(\mu, \nu)$  is not, in general, displacement convex. This seems contrary to the intuition because usually squared distances are nice convex functions<sup>11</sup>. However, we can see that this fails from the following easy example. Take  $\nu = \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(-1,0)}$  and  $\mu_t = \frac{1}{2}\delta_{(t,a)} + \frac{1}{2}\delta_{(-t,-a)}$ . If  $a > 1$ , the curve  $\mu_t$  is the geodesic between  $\mu_{-1}$  and  $\mu_1$  (because the optimal transport between these measures sends  $(a, -1)$  to  $(a, 1)$  and  $(-a, 1)$  to  $(-a, -1)$ ). Yet, if we compute  $W_2^2(\mu_t, \nu)$  we have

$$W_2^2(\mu_t, \nu) = a^2 + \min\{(1-t)^2, (1+t)^2\}.$$

But this function is not convex! (see Figure 2)

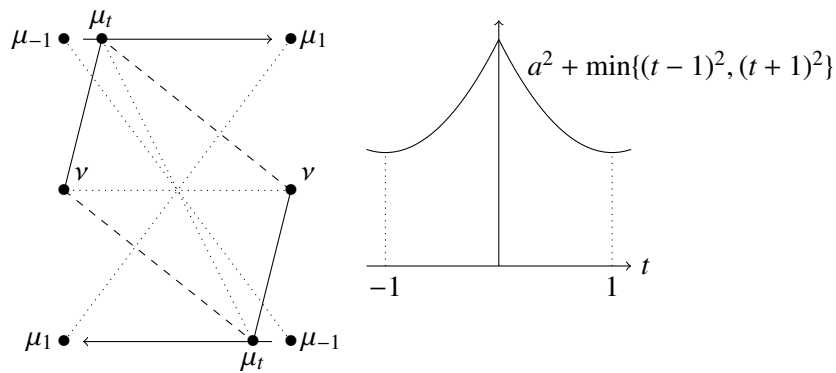


Figure 2: The value of  $W_2^2(\mu_t, \nu)$ .

<sup>11</sup> Actually, this is true in normed spaces, but not even true in Riemannian manifolds, as it depends on curvature properties.

The lack of geodesic convexity of this easy functional<sup>12</sup> 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.** *If we fix an absolutely continuous probability  $\varrho \in \mathcal{P}(\Omega)$ , for every pair  $\mu_0, \mu_1 \in \mathcal{P}(\Omega)$  we call generalized geodesic between  $\mu_0$  and  $\mu_1$  with base  $\varrho$  in  $\mathbb{W}_2(\Omega)$  the curve  $\mu_t = ((1-t)T_0 + tT_1)_\# \varrho$ , where  $T_0$  is the optimal transport map (for the cost  $|x-y|^2$ ) from  $\varrho$  to  $\mu_0$ , and  $T_1$  from  $\varrho$  to  $\mu_1$ .*

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

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

and hence we have the desired convexity along this curve. Moreover, similar considerations to those we developed in this section show that all the functionals that we proved to be geodesically convex are also convex along generalized geodesics. For the case of functionals  $\mathcal{V}$  and  $\mathcal{W}$  this is easy, while for the case of the functional  $\mathcal{F}$ , Lemma 4.9 has to be changed into “ $t \mapsto \det((1-t)A + tB)^{1/d}$  is concave, whenever  $A$  and  $B$  are symmetric and positive-definite” (the proof is similar). We do not develop these proofs here, and we refer to [6] 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 see 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  $\mu_0, \mu_1$  two other probabilities, with  $T_0, T_1$  the optimal transports from  $\nu$  to  $\mu_0$  and  $\mu_1$ , respectively, then the curve

$$\mu_t := ((1-t)T_0 + tT_1)_\# \nu \tag{4.12}$$

connects  $\mu_0$  to  $\mu_1$  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 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  $\lambda$ -geodesic convexity of the Entropy functional  $\varrho \mapsto \int \varrho \log(\varrho) d\text{Vol}$  (where  $\varrho$  is absolutely continuous w.r.t. the volume measure  $\text{Vol}$  and densities are computed accordingly) on a manifold characterizes a lower bound on its Ricci curvature:

<sup>12</sup>By the way, this functional can even be proven to be somehow geodetically concave, as it is shown in [6], Theorem 7.3.2.

**Proposition 4.11.** *Let  $M$  be a compact manifold of dimension  $d$  and  $\text{Vol}$  its volume measure. Let  $\mathcal{E}$  be the entropy functional defined via  $\mathcal{E}(\varrho) = \int \varrho \log \varrho \, d\text{Vol}$  for all measures  $\varrho \ll \text{Vol}$  (set to  $+\infty$  on non-absolutely continuous measures). Then  $\mathcal{E}$  is  $\lambda$ -geodesically convex in the Wasserstein space  $\mathbb{W}_2(M)$  if and only if the Ricci curvature  $\text{Ric}_M$  satisfies  $\text{Ric}_M \geq \lambda$ . 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 Section 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] and Lott and Villani, [63].

**Remark 4.2.** *The reader who followed the first proofs of this section has for sure observed that it is easy, in the space  $\mathbb{W}_2(\Omega)$ , to produce  $\lambda$ -geodesically convex functionals which are not geodesically convex (with  $\lambda < 0$ , of course), of the form  $\mathcal{V}$  (just take a  $\lambda$ -convex function  $V$  which is not convex), but that Theorem 4.10 only provides geodesic convexity (never provides  $\lambda$ -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  $\lambda$ -convexity is a crucial tool to establish uniqueness and stability results in gradient flows is not only true in the abstract theory of Section 3, where we saw that the EVI condition (intimately linked to  $\lambda$ -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 (4.11). We will see this fact via some formal considerations, starting from an interesting lemma:

**Lemma 4.12.** *Suppose that  $F : \mathbb{W}_2(\Omega) \rightarrow \mathbb{R} \cup \{+\infty\}$  is  $\lambda$ -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  $\varphi$  is the Kantorovich potential in the transport from  $\varrho_0$  to  $\varrho_1$  and  $\psi = \varphi^c$  is the Kantorovich potential from  $\varrho_1$  to  $\varrho_0$ .

*Proof.* Let  $\varrho_t$  be the geodesic curve in  $\mathbb{W}_2(\Omega)$  connecting  $\varrho_0$  to  $\varrho_1$  and  $g(t) := F(\varrho_t)$ . The assumption of  $\lambda$ -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.  $\square$

With the above lemma in mind, we consider two curves  $\varrho_t^0$  and  $\varrho_t^1$ , 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  $\varphi$  is the Kantorovich potential in the transport from  $\varrho_t^0$  to  $\varrho_t^1$  and  $\psi = \varphi^c$ . Indeed, a rigorous proof is provided in [6] 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  $\phi, \psi$  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  $\varrho_t^0$  and  $\varrho_t^1$  are solutions of (4.11) 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  $F : \mathbb{W}_2(\Omega) \rightarrow \mathbb{R} \cup \{+\infty\}$  is  $\lambda$ -geodesically convex and that the two curves  $\varrho_t^0$  and  $\varrho_t^1$  are solutions of (4.11). Then, setting  $d(t) := \frac{1}{2} W_2^2(\varrho_t^0, \varrho_t^1)$ , we have*

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

*This implies uniqueness of the solution of (4.11) for fixed initial datum, stability, and exponential convergence as  $t \rightarrow +\infty$  if  $\lambda > 0$ .*

## 4.5 Analysis of the Fokker-Planck equation as a gradient flow in $\mathbb{W}_2$

This section will be a technical parenthesis providing proof details in a particular case study, that of the Fokker-Planck equation, which is the gradient flow of the functional

$$J(\varrho) = \int_{\Omega} \varrho \log \varrho + \int_{\Omega} V d\varrho,$$

where  $V$  is a  $C^1$  function<sup>13</sup> on a domain  $\Omega$ , that we will choose compact for simplicity. The initial measure  $\varrho_0 \in \mathcal{P}(\Omega)$  is taken such that  $J(\varrho_0) < +\infty$ .

We stress that the first term of the functional is defined as

$$\mathcal{E}(\varrho) := \begin{cases} \int_{\Omega} \varrho(x) \log \varrho(x) dx & \text{if } \varrho \ll \mathcal{L}^d, \\ +\infty & \text{otherwise,} \end{cases}$$

where we identify the measure  $\varrho$  with its density, when it is absolutely continuous. This functional is l.s.c. for the weak topology of measures (for general references on the semicontinuity of convex functionals on the space of measures we refer to Chapter 7 in [84] or to [27]), which is equivalent, on the compact domain  $\Omega$ , to the  $W_2$  convergence. Semi-continuity allows to establish the following:

<sup>13</sup>The equation would also be meaningful for  $V$  only Lipschitz continuous, but we prefer to stick to the  $C^1$  case for simplicity.

**Proposition 4.14.** *The functional  $J$  has a unique minimum over  $\mathcal{P}(\Omega)$ . In particular  $J$  is bounded from below. Moreover, for each  $\tau > 0$  the following sequence of optimization problems recursively defined is well-posed*

$$\varrho_{(k+1)}^\tau \in \operatorname{argmin}_{\varrho} \quad J(\varrho) + \frac{W_2^2(\varrho, \varrho_{(k)}^\tau)}{2\tau}, \quad (4.13)$$

which means that there is a minimizer at every step, and this minimizer is unique.

*Proof.* Just apply the direct method, noting that  $\mathcal{P}(\Omega)$  is compact for the weak convergence, which is the same as the convergence for the  $W_2$  distance (again, because  $\Omega$  is compact), and for this convergence  $\mathcal{F}$  is l.s.c. and the other terms are continuous. This gives at the same time the existence of a minimizer for  $J$  and of a solution to each of the above minimization problems (4.13). Uniqueness comes from the fact that all the functionals are convex (in the usual sense) and  $\mathcal{F}$  is strictly convex.  $\square$

**Optimality conditions at each time step.** A first preliminary result we need is the following :

**Lemma 4.15.** *Any minimizer  $\widehat{\varrho}$  in (4.13) must satisfy  $\widehat{\varrho} > 0$  a.e.*

*Proof.* Consider the measure  $\widehat{\varrho}$  with constant positive density  $c$  in  $\Omega$  (i.e.  $c = |\Omega|^{-1}$ ). Let us define  $\varrho_\varepsilon$  as  $(1 - \varepsilon)\widehat{\varrho} + \varepsilon\widehat{\varrho}$  and compare  $\widehat{\varrho}$  to  $\varrho_\varepsilon$ .

By optimality of  $\widehat{\varrho}$ , we may write

$$\mathcal{E}(\widehat{\varrho}) - \mathcal{E}(\varrho_\varepsilon) \leq \int_{\Omega} V d\varrho_\varepsilon - \int_{\Omega} V d\widehat{\varrho} + \frac{W_2^2(\varrho_\varepsilon, \varrho_{(k)}^\tau)}{2\tau} - \frac{W_2^2(\widehat{\varrho}, \varrho_{(k)}^\tau)}{2\tau}. \quad (4.14)$$

The two differences in the right hand side may be easily estimated (by convexity, for instance) so that we get (for a constant  $C$  depending on  $\tau$ , of course):

$$\int_{\Omega} f(\widehat{\varrho}) - f(\varrho_\varepsilon) \leq C\varepsilon$$

where  $f(t) = t \log t$  (set to 0 in  $t = 0$ ). Write

$$A = \{x \in \Omega : \widehat{\varrho}(x) > 0\}, \quad B = \{x \in \Omega : \widehat{\varrho}(x) = 0\}.$$

Since  $f$  is convex we write, for  $x \in A$ ,  $f(\widehat{\varrho}(x)) - f(\varrho_\varepsilon(x)) \geq (\widehat{\varrho}(x) - \varrho_\varepsilon(x))f'(\varrho_\varepsilon(x)) = \varepsilon(\widehat{\varrho}(x) - \widehat{\varrho}(x))(1 + \log \varrho_\varepsilon(x))$ . For  $x \in B$  we simply write  $f(\widehat{\varrho}(x)) - f(\varrho_\varepsilon(x)) = -\varepsilon c \log(\varepsilon c)$ . This allows to write

$$-\varepsilon c \log(\varepsilon c)|B| + \varepsilon \int_A (\widehat{\varrho}(x) - c)(1 + \log \varrho_\varepsilon(x)) dx \leq C\varepsilon$$

and, dividing by  $\varepsilon$ ,

$$-c \log(\varepsilon c)|B| + \int_A (\widehat{\varrho}(x) - c)(1 + \log \varrho_\varepsilon(x)) dx \leq C \quad (4.15)$$

Note that the always have

$$(\widehat{\varrho}(x) - c)(1 + \log \varrho_\varepsilon(x)) \geq (\widehat{\varrho}(x) - c)(1 + \log c)$$

(just distinguish between the case  $\widehat{\varrho}(x) \geq c$  and  $\widehat{\varrho}(x) \leq c$ ). Thus, we may write

$$-c \log(\varepsilon c)|B| + \int_A (\widehat{\varrho}(x) - c)(1 + \log c) dx \leq C.$$

Letting  $\varepsilon \rightarrow 0$  provides a contradiction, unless  $|B| = 0$ .  $\square$

**Remark 4.3.** If, after proving  $|B| = 0$ , we go on with the computations (as it is actually done in Chapter 8 in [84]), we can also obtain  $\log \widehat{\varrho} \in L^1$ , which is a stronger condition than just  $\widehat{\varrho} > 0$  a.e.

We can now compute the first variation and give optimality conditions on the optimal  $\varrho_{(k+1)}^\tau$ .

**Proposition 4.16.** *The optimal measure  $\varrho_{(k+1)}^\tau$  in (4.13) satisfies*

$$\log(\varrho_{(k+1)}^\tau) + V + \frac{\varphi}{\tau} = \text{const. a.e.} \quad (4.16)$$

where  $\varphi$  is the (unique) Kantorovich potential from  $\varrho_{(k+1)}^\tau$  to  $\varrho_{(k)}^\tau$ . In particular,  $\log \varrho_{(k+1)}^\tau$  is Lipschitz continuous. If  $T_k^\tau$  is the optimal transport from  $\varrho_{(k+1)}^\tau$  to  $\varrho_{(k)}^\tau$ , then it satisfies

$$\mathbf{v}_{(k)}^\tau := \frac{id - T_k^\tau}{\tau} = -\nabla \left( \log(\varrho_{(k+1)}^\tau) + V \right) \text{ a.e.} \quad (4.17)$$

*Proof.* Take the optimal measure  $\widehat{\varrho} := \varrho_{(k+1)}^\tau$ . We can check that all the functionals involved in the minimization admit a first variation at  $\widehat{\varrho}$ . For the linear term it is straightforward, and for the Wasserstein term we can apply Proposition 4.8. The uniqueness of the Kantorovich potential is guaranteed by the fact that  $\widehat{\varrho}$  is strictly positive a.e. on the domain  $\Omega$  (Lemma 4.15). For the entropy term, the integrability of  $\log \widehat{\varrho}$  provided in the previous remark allows (computations are left to the reader) to differentiate under the integral sign for smooth perturbations.

The first variation of  $J$  is hence

$$\frac{\delta J}{\delta \varrho} = f'(\varrho) + V = \log(\varrho) + 1 + V.$$

Using carefully (details are in Chapter 8 of [84]) the optimality conditions, we obtain Equation (4.16), which is valid a.e. since  $\widehat{\varrho} > 0$ . In particular, this implies that  $\varrho_{(k+1)}^\tau$  is Lipschitz continuous, since we have

$$\varrho_{(k+1)}^\tau(x) = \exp \left( C - V(x) - \frac{\varphi(x)}{\tau} \right).$$

Then, one differentiates and gets the equality

$$\nabla \varphi = \frac{id - T_k^\tau}{\tau} = -\nabla \left( \log(\varrho_{(k+1)}^\tau) + V \right) \text{ a.e.}$$

and this allows to conclude. □

**Interpolation between time steps and uniform estimates** Let us collect some other tools

**Proposition 4.17.** *For any  $\tau > 0$ , the sequence of minimizers satisfies*

$$\sum_k \frac{W_2^2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} \leq C := 2(J(\varrho_0) - \inf J).$$

*Proof.* This is just the standard estimate in the minimizing movement scheme, corresponding to what we presented in (2.5) in the Euclidean case. □

Let us define two interpolations between the measures  $\varrho_{(k)}^\tau$ .

With this time-discretized method, we have obtained, for each  $\tau > 0$ , a sequence  $(\varrho_{(k)}^\tau)_k$ . We can use it to build at least two interesting curves in the space of measures:

- first we can define some piecewise constant curves, i.e.  $\varrho_t^\tau := \varrho_{(k+1)}^\tau$  for  $t \in ]k\tau, (k+1)\tau[$ ; associated to this curve we also define the velocities  $\mathbf{v}_t^\tau = \mathbf{v}_{(k+1)}^\tau$  for  $t \in ]k\tau, (k+1)\tau[$ , where  $\mathbf{v}_{(k)}^\tau$  is defined as in (4.17):  $\mathbf{v}_{(k)}^\tau = (id - T_k^\tau)/\tau$ , taking as  $T_k^\tau$  the optimal transport from  $\varrho_{(k+1)}^\tau$  to  $\varrho_{(k)}^\tau$ ; we also define the momentum variable  $E^\tau = \varrho^\tau \mathbf{v}^\tau$ ;
- then, we can also consider the densities  $\widetilde{\varrho}_t^\tau$  that interpolate the discrete values  $(\varrho_{(k)}^\tau)_k$  along geodesics:

$$\widetilde{\varrho}_t^\tau = \left( \frac{k\tau - t}{\tau} \mathbf{v}_{(k)}^\tau + id \right)_\# \varrho_{(k)}^\tau, \quad \text{for } t \in ](k-1)\tau, k\tau[; \quad (4.18)$$

the velocities  $\widetilde{\mathbf{v}}_t^\tau$  are defined so that  $(\widetilde{\varrho}^\tau, \widetilde{\mathbf{v}}^\tau)$  satisfy the continuity equation and  $\|\widetilde{\mathbf{v}}_t^\tau\|_{L^2(\widetilde{\varrho}_t^\tau)} = |(\widetilde{\varrho}^\tau)'|(t)$ . To do so, we take

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

as before, we define a momentum variable:  $\widetilde{E}^\tau = \widetilde{\varrho}^\tau \widetilde{\mathbf{v}}^\tau$ .

After these definitions we consider some a priori bounds on the curves and the velocities that we defined. We start from some estimates which are standard in the framework of Minimizing Movements.

Note that the velocity (i.e. metric derivative) of  $\widetilde{\varrho}^\tau$  is constant on each interval  $]k\tau, (k+1)\tau[$  and equal to

$$\frac{W_2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} = \frac{1}{\tau} \left( \int |id - T_k^\tau|^2 d\varrho_{(k+1)}^\tau \right)^{1/2} = \|\mathbf{v}_{k+1}^\tau\|_{L^2(\varrho_{(k+1)}^\tau)},$$

which gives

$$\|\widetilde{\mathbf{v}}_t^\tau\|_{L^2(\widetilde{\varrho}_t^\tau)} = |(\widetilde{\varrho}^\tau)'|(t) = \frac{W_2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} = \|\mathbf{v}_t^\tau\|_{L^2(\varrho_t^\tau)},$$

where we used the fact that the velocity field  $\widetilde{\mathbf{v}}^\tau$  has been chosen so that its  $L^2$  norm equals the metric derivative of the curve  $\widetilde{\varrho}^\tau$ .

In particular we can obtain

$$\begin{aligned} |E^\tau|([0, T] \times \Omega) &= \int_0^T dt \int_\Omega |\mathbf{v}_t^\tau| d\varrho_t^\tau = \int_0^T \|\mathbf{v}_t^\tau\|_{L^1(\varrho_t^\tau)} dt \leq \int_0^T \|\mathbf{v}_t^\tau\|_{L^2(\varrho_t^\tau)} dt \\ &\leq T^{1/2} \int_0^T \|\mathbf{v}_t^\tau\|_{L^2(\varrho_t^\tau)}^2 dt = T^{1/2} \sum_k \tau \left( \frac{W_2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} \right)^2 \leq C. \end{aligned}$$

The estimate on  $\widetilde{E}^\tau$  is completely analogous

$$|\widetilde{E}^\tau|([0, T] \times \Omega) = \int_0^T dt \int_\Omega |\widetilde{\mathbf{v}}_t^\tau| d\widetilde{\varrho}_t^\tau \leq T^{1/2} \int_0^T \|\widetilde{\mathbf{v}}_t^\tau\|_{L^2(\widetilde{\varrho}_t^\tau)}^2 dt = T^{1/2} \sum_k \tau \left( \frac{W_2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} \right)^2 \leq C.$$



This gives compactness of  $E^\tau$  and  $\widetilde{E}^\tau$  in the space of vector measures on space-time, for the weak convergence. As far as  $\widetilde{\varrho}^\tau$  is concerned, we can obtain more than that. Consider the following estimate, for  $s < t$

$$W_2(\widetilde{\varrho}_t^\tau, \widetilde{\varrho}_s^\tau) \leq \int_s^t |(\widetilde{\varrho}^\tau)'|(r) dr \leq (t-s)^{1/2} \left( \int_s^t |(\widetilde{\varrho}^\tau)'|(r)^2 dr \right)^{1/2}.$$

From the previous computations, we have again

$$\int_0^T |(\widetilde{\varrho}^\tau)'|(r)^2 dr = \sum_k \tau \left( \frac{W_2(\varrho_{(k+1)}^\tau, \varrho_{(k)}^\tau)}{\tau} \right)^2 \leq C,$$

and this implies

$$W_2(\widetilde{\varrho}_t^\tau, \widetilde{\varrho}_s^\tau) \leq C(t-s)^{1/2}, \quad (4.19)$$

which means that the curves  $\widetilde{\varrho}^\tau$  are uniformly Hölder continuous. Since they are defined on  $[0, T]$  and valued in  $\mathbb{W}_2(\Omega)$  which is compact, we can apply the Ascoli Arzelà Theorem. This implies that, up to subsequences, we have

$$E^\tau \rightharpoonup E \text{ in } \mathcal{M}^d([0, T] \times \Omega), \quad \widetilde{E}^\tau \rightharpoonup \widetilde{E} \text{ in } \mathcal{M}^d([0, T] \times \Omega); \quad \widetilde{\varrho}^\tau \rightarrow \varrho \text{ uniformly for the } W_2 \text{ distance.}$$

The limit curve  $\varrho$ , from the uniform bounds on  $\widetilde{\varrho}^\tau$ , is both  $\frac{1}{2}$ -Hölder continuous and absolutely continuous in  $\mathbb{W}_2$ . As far as the curves  $\varrho^\tau$  are concerned, they also converge uniformly to the same curve  $\varrho$ , since  $W_2(\varrho_t^\tau, \widetilde{\varrho}_t^\tau) \leq C\sqrt{\tau}$  (a consequence of (4.19), of the fact that  $\widetilde{\varrho}^\tau = \varrho^\tau$  on the points of the form  $k\tau$  and of the fact that  $\varrho^\tau$  is constant on each interval  $]k\tau, (k+1)\tau[$ ).

Let us now prove that  $\widetilde{E} = E$ .

**Lemma 4.18.** *Suppose that we have two families of vector measures  $E^\tau$  and  $\widetilde{E}^\tau$  such that*

- $\widetilde{E}^\tau = \widetilde{\varrho}^\tau \widetilde{\mathbf{v}}^\tau$ ;  $E^\tau = \varrho^\tau \mathbf{v}^\tau$ ;
- $\widetilde{\mathbf{v}}_t^\tau = \mathbf{v}_t^\tau \circ ((k\tau - t)\mathbf{v}_{(k)}^\tau + id)^{-1}$ ;  $\widetilde{\varrho}^\tau = ((k\tau - t)\mathbf{v}_{(k)}^\tau + id)_\# \varrho^\tau$ ;
- $\int \int |\mathbf{v}^\tau|^2 d\varrho^\tau \leq C$  (with  $C$  independent of  $\tau$ );
- $E^\tau \rightharpoonup E$  and  $\widetilde{E}^\tau \rightharpoonup \widetilde{E}$  as  $\tau \rightarrow 0$

Then  $\widetilde{E} = E$ .

*Proof.* It is sufficient to fix a Lipschitz function  $f : [0, T] \times \Omega \rightarrow \mathbb{R}^d$  and to prove  $\int f \cdot dE = \int f \cdot d\widetilde{E}$ . To do that, we write

$$\int f \cdot d\widetilde{E}^\tau = \int_0^T dt \int_\Omega f \cdot \widetilde{\mathbf{v}}_t^\tau d\widetilde{\varrho}^\tau = \int_0^T dt \int_\Omega f \circ ((k\tau - t)\mathbf{v}^\tau + id) \cdot \mathbf{v}_t^\tau d\varrho^\tau,$$

which implies

$$\left| \int f \cdot d\widetilde{E}^\tau - \int f \cdot dE^\tau \right| \leq \int_0^T dt \int_\Omega |f \circ ((k\tau - t)\mathbf{v}^\tau + id) - f| |\mathbf{v}_t^\tau| d\varrho^\tau \leq \text{Lip}(f)\tau \int_0^T \int_\Omega |\mathbf{v}_t^\tau|^2 d\varrho^\tau \leq C\tau.$$

This estimate proves that the limit of  $\int f \cdot d\widetilde{E}^\tau$  and  $\int f \cdot dE^\tau$  is the same, i.e.  $E = \widetilde{E}$ .  $\square$

**Relation between  $\varrho$  and  $E$**  We can obtain the following:

**Proposition 4.19.** *The pair  $(\varrho, E)$  satisfies, in distributional sense*

$$\partial_t \varrho + \nabla \cdot E = 0, \quad E = -\nabla \varrho - \varrho \nabla V,$$

with no-flux boundary conditions on  $\partial\Omega$ . In particular we have found a solution to

$$\begin{cases} \partial_t \varrho - \Delta \varrho - \nabla \cdot (\varrho \nabla V) = 0, \\ (\nabla \varrho + \varrho \nabla V) \cdot \mathbf{n} = 0, \\ \varrho(0) = \varrho_0, \end{cases}$$

where the initial datum is to be intended in the following sense: the curve  $t \mapsto \varrho_t$  is (absolutely) continuous in  $\mathbb{W}_2$ , and its initial value is  $\varrho_0$ .

*Proof.* First, consider the weak convergence  $(\tilde{\varrho}^\tau, \tilde{E}^\tau) \rightharpoonup (\varrho, E)$  (which is a consequence of  $\tilde{E} = E$ ). The continuity equation  $\partial_t \tilde{\varrho}^\tau + \nabla \cdot \tilde{E}^\tau = 0$  is satisfied by construction in the sense of distributions, and this passes to the limit. Hence,  $\partial_t \varrho + \nabla \cdot E = 0$ . The continuity in  $\mathbb{W}_2$  and the initial datum pass to the limit because of the uniform  $C^{0,1/2}$  bound in (4.19).

Then, use the convergence  $(\varrho^\tau, E^\tau) \rightharpoonup (\varrho, E)$ . Actually, using the optimality conditions of Proposition 4.16 and the definition of  $E^\tau = \mathbf{v}^\tau \varrho^\tau$ , we have, for each  $\tau > 0$ ,  $E^\tau = -\nabla \varrho^\tau - \varrho^\tau \nabla V$  (note that  $\varrho^\tau$  is Lipschitz continuous for fixed  $\tau > 0$ , which allows to write this equality, exploiting in particular  $\nabla \varrho^\tau = \varrho^\tau \nabla (\log(\varrho^\tau))$ , which would have no meaning for less regular measures; this regularity could be, of course, lost in the limit  $\tau \rightarrow 0$ ). It is not difficult to pass this condition to the limit either (but now  $\nabla \varrho$  will have to be intended in the sense of distributions). Take  $f \in C_c^1([0, T] \times \Omega; \mathbb{R}^d)$  and test:

$$\int f \cdot dE^\tau = - \int f \cdot \nabla \varrho^\tau - \int f \cdot \nabla V \varrho^\tau = \int \nabla \cdot f d\varrho^\tau - \int f \cdot \nabla V \varrho^\tau.$$

These terms pass to the limit as  $\varrho^\tau \rightharpoonup \varrho$ , using the assumption  $V \in C^1$ , since all the test functions above are continuous. This gives  $\int f \cdot dE = \int (\nabla \cdot f) d\varrho - \int f \cdot \nabla V d\varrho$ , which implies  $E = -\nabla \varrho - \varrho \nabla V$ .  $\square$

## 4.6 Other gradient-flow PDEs

We saw in the previous section the example, and a detailed analysis, of the Fokker-Planck equation. The main reason to choose such example is its simplicity, because it is a linear equation. This allows easily to pass to the limit all the terms in the relation between  $E$  and  $\varrho$ . Yet, many other important equations can be obtained as gradient flows in  $\mathbb{W}_2$ , choosing other functionals. We will see some of them here, without entering into details of the proof. We stress that handling non-linear terms is often difficult and requires ad-hoc estimates. We will discuss some of the ideas that one should apply. Note on the other hand that, if one uses the abstract theory of [6], there is no need to produce these ad-hoc estimates: after developing a general (and hard) theory for general metric spaces, the second part of [6] explains which are the curves that one finds as gradient flows in  $\mathbb{W}_2$ , with the relation between the velocity field  $\mathbf{v}$  and the derivatives (with an ad-hoc notion of subdifferential in the Wasserstein space) of the functional  $F$ . This automatically gives the desired result as a part of a larger theory.

We will discuss five classes of PDEs in this section: the *porous media equation*, the *Keller-Segel equation*, more general *diffusion*, *advection and aggregation* equations, a model for *crowd motion* with density constraints, and the *flow of the squared sliced Wasserstein distance*  $\text{SW}_2^2$ .

**Porous Media Equation** This equation models the diffusion of a substance into a material whose properties are different than the void, and which slows down the diffusion. If one considers the case of particles which are advected by a potential and subject to this kind of diffusion, the PDE reads

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

for an exponent  $m > 1$ . One can formally check that this is the equation of the gradient flow of the energy

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

(set to  $+\infty$  for  $\varrho \notin L^m$ ). Indeed, the first variation of the first part of the functional is  $\frac{m}{m-1} \varrho^{m-1}$ , and  $\varrho \nabla \left( \frac{m}{m-1} \varrho^{m-1} \right) = m \varrho \cdot \varrho^{m-2} \nabla \varrho = \nabla(\varrho^m)$ .

Note that, in the discrete step  $\min_{\varrho} F(\varrho) + \frac{W_2^2(\varrho, \varrho_0)}{2\tau}$ , the solution  $\varrho$  satisfies

$$\begin{cases} \frac{m}{m-1} \varrho^{m-1} + V + \frac{\varphi}{\tau} = C & \varrho - \text{a.e.} \\ \frac{m}{m-1} \varrho^{m-1} + V + \frac{\varphi}{\tau} \geq C & \text{on } \{\varrho = 0\}. \end{cases}$$

This allows to express  $\varrho^{m-1} = \frac{m-1}{m} (C - V - \varphi/\tau)_+$  and implies that  $\varrho$  is compactly supported if  $\varrho_0$  is compactly supported, as soon as  $V$  has some growth conditions. This fact contradicts the usual infinite propagation speed that one finds in linear diffusion models (Heat and Fokker-Planck equation).

The above analysis works in the case  $m > 1$ : the fact that the usual Fokker-Planck equation can be obtained for  $m \rightarrow 1$  can be seen in the following way: nothing changes if we define  $F$  via  $F(\varrho) = \frac{1}{m-1} \int (\varrho^m(x) - \varrho(x)) dx + \int V(x) \varrho(x) dx$ , since the mass  $\int \varrho(x) dx = 1$  is a given constant. Yet, then it is easy to guess the limit, since

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

which provides the entropy that we already used for the Heat equation.

It is also interesting to consider the case  $m < 1$ : the function  $\varrho^m - \varrho$  is no longer convex, but it is concave and the negative coefficient  $1/(m-1)$  makes it a convex function. Unfortunately, it is not superlinear at infinity, which makes it more difficult to handle. But for  $m \geq 1 - 1/d$  the functional  $F$  is still displacement convex. The PDE that we get as a gradient flow is called *Fast diffusion* equation, and it has different (and opposite) properties in terms of diffusion rate than the porous media one.

From a technical point of view, proving compactness of the minimizing movement scheme for these equations is not very easy, since one needs to pass to the limit the non-linear term  $\Delta(\varrho^m)$ , which means proving strong convergence on  $\varrho$  instead of weak convergence. The main ingredient is a sort of  $H^1$  bound in space, which comes from the fact that we have

$$\int_0^T \int_{\Omega} |\nabla(\varrho^{m-1/2})|^2 dx dt \approx \int_0^T \int_{\Omega} \varrho \frac{|\nabla \varphi|^2}{\tau^2} dx dt = \int_0^T |\varrho'(t)|^2 dt \leq C \quad (4.20)$$

(but one has to deal with the fact that this is not a full  $H^1$  bound, and the behavior in time has to be controlled, via some variants of the classical Aubin-Lions Lemma, [13]).

**Diffusion, advection and aggregation** Consider a more general case where the movement is advected by a potential determined by the superposition of many potentials, each created by one particle. For instance, given a function  $W : \mathbb{R}^d \rightarrow \mathbb{R}$ , the particle located at  $x$  produces a potential  $W(\cdot - x)$  and, globally, the potential is given by  $V(y) = \int W(y - x)d\varrho(x)$ , 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

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

When  $W$  is convex, for instance in the quadratic case  $\iint |x - y|^2 d\varrho(x)d\varrho(y)$ , this gives raise to a general aggregation behavior of the particles, and as  $t \rightarrow \infty$  one expects  $\varrho_t \rightarrow \delta_{x_0}$  (the point  $x_0$  depending on the initial datum  $\varrho_0$ : in the quadratic example above it is the barycenter of  $\varrho_0$ ). If  $W$  is not smooth enough, the aggregation into a unique point can also occur in finite time, see [30].

Note that these equations are both non-linear (the term in the divergence is quadratic in  $\varrho$ ) and non-local. It is rare to see these non-local aggregation terms alone in the equation, as they are often coupled with diffusion or other terms. This is why we do not provide specific references except [30]. We also note that from the technical point of view this kind of nonlinearity is much more compact than the previous ones, since the convolution operator transforms weak convergence into strong one, provided  $W$  is regular enough.

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

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

This gives the equation

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

Among the mathematical interest for this family of equations, we stress that they are those where more results (in terms of stability, and convergence to equilibrium) can be proven, due to the fact that conditions to guarantee that  $F$  is displacement convex are well-known (Section 4.4). See in particular [31, 32] for physical considerations and convergence results on this equation.

**Keller-Segel** An interesting model in mathematical biology (see [57, 58] for the original modeling) is the following: a population  $\varrho$  of bacteria evolves in time, following diffusion and advection by a potential. The potential is given by the concentration  $u$  of a chemo-attractant nutrient substance, produced by the bacteria themselves. This kind of phenomenon is also known under the name of *chemotaxis*. More precisely, bacteria move (with diffusion) in the direction where they find more nutrient, i.e. in the direction of  $\nabla u$ , where the distribution of  $u$  depends on their density  $\varrho$ . The easiest model uses linear diffusion and supposes that the distribution of  $u$  is related to  $\varrho$  by the condition  $-\Delta u = \varrho$ , with Dirichlet

boundary conditions  $u = 0$  on  $\partial\Omega$ . This gives the system

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

The parameter  $\alpha$  stands for the attraction intensity of bacteria towards the chemo-attractant. By scaling, instead of using probability measures  $\varrho \in \mathcal{P}(\Omega)$  one can set  $\alpha = 1$  and play on the mass of  $\varrho$  (indeed, the non-linearity is only in the term  $\varrho \nabla u$ , which is quadratic in  $\varrho$ ).

Alternative equations can be considered for  $u$ , such as  $-\Delta u + u = \varrho$  with Neumann boundary conditions. On the contrary, the boundary conditions on  $\varrho$  must be of no-flux type, to guarantee conservation of the mass. This system can also be set in the whole space, with suitable decay conditions at infinity. Note also that often the PDE condition defining  $u$  as the solution of a Poisson equation is replaced, when  $\Omega = \mathbb{R}^2$ , by the explicit formula

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

There is some confusion in higher dimension, as the very same formula does not hold for the Poisson equation (the logarithmic kernel should indeed be replaced by the corresponding Green function), and there two alternatives: either keep the fact that  $u$  solves  $-\Delta u = \varrho$ , or the fact that it derives from  $\varrho$  through (4.21).

One can see that this equation is the gradient flow of the functional

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

Indeed, the only non-standard computation is that of 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

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

It is interesting to note that this Dirichlet term is indeed (up to the coefficient  $-1/2$ ) the square of the  $H^{-1}$  norm of  $\varrho$ , since  $\|u\|_{H_0^1} = \|\nabla u\|_{L^2} = \|\varrho\|_{H^{-1}}$ . We will call it the  $H^{-1}$  term.

It is also possible to replace linear diffusion with non-linear diffusion of porous media type, replacing the entropy  $\int \varrho \log \varrho$  with a power-like energy  $\int \varrho^m$ .

Note that the variational problem  $\min F(\varrho) + \frac{W_2^2(\varrho, \varrho_0)}{2\tau}$  requires some assumption to admit existence of minimizers, as unfortunately the Dirichlet term has the wrong sign. 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.

When we use non-linear diffusion with  $m > 2$  the existence of a solution is quite easy. Sophisticated functional inequalities allow to handle smaller exponents, and even the linear diffusion case in dimension 2, provided  $\alpha \leq 8\pi$ . We refer to [18] and to the references therein for details on the analysis of this equation. Some of the technical difficulties are similar to those of the porous media equation, when passing to the limit non-linear terms. In [18], the  $H^{-1}$  term is treated in terms of its logarithmic kernel,

and ad-hoc variables symmetrization tricks are used. Note however that the nonlinear diffusion case is easier, as  $L^m$  bounds on  $\varrho$  translate into  $W^{2,m}$  bounds on  $u$ , and hence strong compactness for  $\nabla u$ .

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

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

or other variants with different boundary conditions. This equation can also be studied as a gradient flow in two variables, using the distance  $W_2$  on  $\varrho$  and the  $L^2$  distance on  $u$ ; see [19].

**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.

Let us explain the modeling, starting from the discrete case. Suppose that we have a population of particles such that each of them, if alone, would follow its own velocity  $\mathbf{u}$  (which could a priori depend on time, position, on the particle itself...). Yet, these particles are modeled by rigid disks that cannot overlap, hence, the actual velocity cannot always be  $\mathbf{u}$ , in particular if  $\mathbf{u}$  tends to concentrate the masses. We will call  $\mathbf{v}$  the actual velocity of each particle, and the main assumption of the model is that  $\mathbf{v} = P_{\text{adm}(q)}(\mathbf{u})$ , where  $q$  is the particle configuration,  $\text{adm}(q)$  is the set of velocities that do not induce (for an infinitesimal time) overlapping starting from the configuration  $q$ , and  $P_{\text{adm}(q)}$  is the projection on this set.

The simplest example is the one where every particle is a disk with the same radius  $R$  and center located at  $q_i$ . In this case we define the admissible set of configurations  $K$  through

$$K := \{q = (q_i)_i \in \Omega^N : |q_i - q_j| \geq 2R \text{ for all } i \neq j\}.$$

In this way the set of admissible velocities is easily seen to be

$$\text{adm}(q) = \{\mathbf{v} = (\mathbf{v}_i)_i : (\mathbf{v}_i - \mathbf{v}_j) \cdot (q_i - q_j) \geq 0 \text{ for all } (i, j) \text{ with } |q_i - q_j| = 2R\}.$$

The evolution equation which has to be solved to follow the motion of  $q$  is then

$$q'(t) = P_{\text{adm}(q(t))} \mathbf{u}(t) \quad (4.22)$$

(with  $q(0)$  given). Equation (4.22), not easy from a mathematical point of view, was studied by Maury and Venel in [66, 67].

We are now interested in the simplest continuous counterpart of this microscopic model (without pretending that it is any kind of homogenized limit of the discrete case, but only an easy re-formulation in a density setting). In this case the particles population will be described by a probability density  $\varrho \in \mathcal{P}(\Omega)$ , the constraint becomes a density constraint  $\varrho \leq 1$  (we define the set  $K = \{\varrho \in \mathcal{P}(\Omega) : \varrho \leq 1\}$ ), the set of admissible velocities will be described by the sign of the divergence on the saturated region  $\{\varrho = 1\}$ :  $\text{adm}(\varrho) = \{\mathbf{v} : \Omega \rightarrow \mathbb{R}^d : \nabla \cdot \mathbf{v} \geq 0 \text{ on } \{\varrho = 1\}\}$ ; we will consider a projection  $P$ , which will be

either the projection in  $L^2(\mathcal{L}^d)$  or in  $L^2(\varrho)$  (this will turn out to be the same, since the only relevant zone is  $\{\varrho = 1\}$ ). Finally, we solve the equation

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

The main difficulty is the fact that the vector field  $\mathbf{v} = P_{\text{adm}(\varrho_t)} \mathbf{u}_t$  is neither regular (since it is obtained as an  $L^2$  projection, and may only be expected to be  $L^2$  a priori), nor it depends regularly on  $\varrho$  (it is very sensitive to small changes in the values of  $\varrho$ : passing from a density 1 to a density  $1 - \varepsilon$  completely modifies the saturated zone, and hence the admissible set of velocities and the projection onto it).

In [64] these difficulties have been overpassed in the case  $\mathbf{u} = -\nabla D$  (where  $D : \Omega \rightarrow \mathbb{R}$  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

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

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, instead of considering the divergence of vector fields which are only supposed to be  $L^2$ , it is more convenient to give a better description of  $\text{adm}(\varrho)$  by duality:

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

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

$$\begin{aligned} \mathbf{u} &= \mathbf{v} + \nabla p, \quad \mathbf{v} \in \text{adm}(\varrho), \quad \int \mathbf{v} \cdot \nabla p = 0, \\ p &\in \text{press}(\varrho) := \{p \in H^1(\Omega), p \geq 0, p(1 - \varrho) = 0\}, \end{aligned}$$

where  $\text{press}(\varrho)$  is the space of functions  $p$  used as test functions in the dual definition of  $\text{adm}(\varrho)$ , which play the role of a pressure affecting the movement. The two cones  $\nabla \text{press}(\varrho)$  (defined as the set of gradients of elements of  $\text{press}(\varrho)$ ) and  $\text{adm}(\varrho)$  are in duality for the  $L^2$  scalar product (i.e. one is defined as the set of vectors which make a negative scalar product with all the elements of the other). This allows for an orthogonal decomposition  $\mathbf{u}_t = \mathbf{v}_t + \nabla p_t$ , and gives the alternative expression of Equation (4.23), i.e.

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

More details can be found in [64, 81, 65]. In particular, in [64] it is explained how to handle the nonlinearities when passing to the limit. Two sources of nonlinearity are observed: the term  $\varrho \nabla p$  is easy to consider, since it is actually equal to  $\nabla p$  (as we have  $p = 0$  on  $\{\varrho \neq 1\}$ ); on the other hand, we need to deal with the equality  $p(1 - \varrho) = 0$  and pass it to the limit. This is done by obtaining strong compactness on  $p$ , from a bound on  $\int_0^T \int_\Omega |\nabla p|^2$ , similarly to (4.20). It is important to observe that transforming  $\varrho \nabla p$  into  $\nabla p$  is only possible in the case of one only phase  $\varrho$ , while the multi-phasic case presents extra-difficulties (see, for instance [29]). These difficulties can be overpassed in the presence of diffusion, which provides extra  $H^1$  bounds and hence strong convergence for  $\varrho$  (for crowd motion with diffusion,



see [74]). Finally, the uniqueness question is a tricky one: if the potential  $D$  is  $\lambda$ -convex then this is easy, but considerations from the DiPerna-Lions theory ([40, 4]) suggest that uniqueness should be true under weaker assumptions; nevertheless this is not really proven. On the other hand, in [39] uniqueness is proven under very mild assumptions on the drift (or, equivalently, on the potential), provided some diffusion is added.

**Sliced Wasserstein distance** Inspired by a construction proposed in [77] for image inpainting, M. Bernot proposed around 2008 an interesting scheme to build a transport map between two given measures which, if not optimal, at least had some monotonicity properties obtained in an isotropic way.

Consider two measures  $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$  and project them onto any one-dimensional direction. For every  $e \in \mathbb{S}^{d-1}$  (the unit sphere of  $\mathbb{R}^d$ ), we take the map  $\pi_e : \mathbb{R}^d \rightarrow \mathbb{R}$  given by  $\pi_e(x) = x \cdot e$  and look at the image measures  $(\pi_e)_\# \mu$  and  $(\pi_e)_\# \nu$ . They are measures on the real line, and we call  $T_e : \mathbb{R} \rightarrow \mathbb{R}$  the monotone (optimal) transport between them. The idea is that, as far as the direction  $e$  is concerned, every point  $x$  of  $\mathbb{R}^d$  should be displaced of a vector  $\mathbf{v}_e(x) := (T_e(\pi_e(x)) - \pi_e(x))e$ . To do a global displacement, consider  $\mathbf{v}(x) = \int_{\mathbb{S}^{d-1}} \mathbf{v}_e(x) d\mathcal{H}^{d-1}(e)$ , where  $\mathcal{H}^{d-1}$  is the uniform Hausdorff measure on the sphere.

There is no reason to think that  $id + \mathbf{v}$  is a transport map from  $\mu$  to  $\nu$ , and indeed in general it is not. But if one fixes a small time step  $\tau > 0$  and uses a displacement  $\tau \mathbf{v}$  getting a measure  $\mu_1 = (id + \tau \mathbf{v})_\# \mu$ , then it is possible to iterate the construction. One expects to build in this way a sequence of measures  $\mu_n$  that converges to  $\nu$ , but this has not yet been rigorously proven. From the empirical point of view, the transport maps that are obtained in this way are quite satisfactory, and have been tested in particular in the discrete case (a finite number of Dirac masses with equal mass, i.e. the so-called assignment problem).

For the above construction, discrete in time and used for applications to images, there are not many references (essentially, the only one is [78], see also [20] for a wider discussion). A natural continuous counterpart exists: simply consider, for every absolutely continuous measure  $\varrho \in \mathcal{P}(\mathbb{R}^d)$ , the vector field  $\mathbf{v} = \mathbf{v}_{(\varrho)}$  that we defined above as a velocity field depending on  $\varrho$  (absolute continuity is just required to avoid atoms in the projections). Then, we solve the equation

$$\partial_t \varrho_t + \nabla \cdot (\varrho_t \mathbf{v}_{(\varrho_t)}) = 0. \quad (4.25)$$

It happens that this equation has a gradient flow structure: it is indeed the gradient flow in  $\mathbb{W}_2$  of a functional related to a new distance on probability measures, induced by this construction.

This distance can be defined as follows (see [78]): given two measures  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ , we define

$$SW_2(\mu, \nu) := \left( \int_{\mathbb{S}^{d-1}} W_2^2((\pi_e)_\# \mu, (\pi_e)_\# \nu) d\mathcal{H}^{d-1}(e) \right)^{1/2}.$$

This quantity could have been called “projected Wasserstein distance” (as it is based on the behavior through projections), but since in [78] it is rather called “sliced Wasserstein distance”, we prefer to keep the same terminology.

The fact that  $SW_2$  is a distance comes from  $W_2$  being a distance. The triangle inequality may be proven using the triangle inequality for  $W_2$  (see Section 5.1) and for the  $L^2$  norm. Positivity and symmetry are evident. The equality  $SW_2(\mu, \nu) = 0$  implies  $W_2^2((\pi_e)_\# \mu, (\pi_e)_\# \nu) = 0$  for all  $e \in \mathbb{S}^{d-1}$ . This means  $(\pi_e)_\# \mu = (\pi_e)_\# \nu$  for all  $e$  and it suffices to prove  $\mu = \nu$ .

It is evident from its definition, and from the fact that the maps  $\pi_e$  are 1-Lipschitz (which implies  $W_2^2((\pi_e)_\# \mu, (\pi_e)_\# \nu) \leq W_2^2(\mu, \nu)$ ) that we have  $SW_2(\mu, \nu) \leq W_2(\mu, \nu)$ . Moreover, the two distances also



induce the same topology, at least on compact sets. Indeed, the identity map from  $\mathbb{W}_2$  to  $(\mathcal{P}(\Omega), SW_2)$  is continuous (because of  $SW_2 \leq W_2$ ) and bijective. Since the space where it is defined is compact, it is also a homeomorphism. One can also prove more, i.e. an inequality of the form  $W_2 \leq CSW_2^\beta$  for a suitable exponent  $\beta \in ]0, 1[$ . Chapter 5 in [20] proves this inequality with  $\beta = (2(d+1))^{-1}$ .

The interest in the use of this distance is the fact that one has a distance on  $\mathcal{P}(\Omega)$  with very similar qualitative properties as  $W_2$ , but much easier to compute, since it only depends on one-dimensional computations (obviously, the integral over  $e \in \mathbb{S}^{d-1}$  is discretized in practice, and becomes an average over a large number of directions). We remark anyway an important difference between  $W_2$  and  $SW_2$ : the latter is not a geodesic distance. On the contrary, the geodesic distance associated to  $SW_2$  (i.e. the minimal lenght to connect two measures) is exactly  $W_2$ .

If we come back to gradient flows, it is not difficult to see that the equation (4.25) is the gradient flow of  $F(\varrho) := \frac{1}{2}SW_2^2(\varrho, \nu)$  and can be studied as such. Existence and estimates on the solution of this equation are proven in [20], and the nonlinearity of  $\mathbf{v}_{(\varrho)}$  is quite easy to deal with. On the other hand, many, natural and useful, questions are still open: is it true that  $\varrho_t \rightarrow \nu$  as  $t \rightarrow \infty$ ? Can we define (at least under regularity assumptions on the initial data) the flow of the vector field  $\mathbf{v}_{(\varrho_t)}$ , and what is the limit of this flow as  $t \rightarrow \infty$ ? The idea is that it should be a transport map between  $\varrho_0$  and  $\nu$  and, if not the optimal transport map, at least a “good” one, but most of the related questions are open.

## 4.7 Dirichlet boundary conditions

For sure, the attentive reader has already noted that all the equations that have been identified as gradient flows for the distance  $W_2$  on a bounded domain  $\Omega$  are always accompanied by Neumann boundary conditions. This should not be surprising. Wasserstein distances express the movement of masses when passing from a configuration to another, and the equation represents the conservation of mass. It means that we are describing the movement of a collection  $\varrho$  of particles, bound to stay inside a given domain  $\Omega$ , and selecting their individual velocity  $\mathbf{v}$  in way which is linked to the global value of a certain functional  $F(\varrho)$ . It is natural in this case to have boundary conditions which write down the fact that particles do not exit the domain, and the pointwise value of the density  $\varrho$  on  $\partial\Omega$  is not particularly relevant in this analysis. Note that “do not exit” does not mean “those on the boundary stay on the boundary”, which is what happens when solutions are smooth and the velocity field  $\mathbf{v}$  satisfies  $\mathbf{v} \cdot \mathbf{n} = 0$ . Yet, the correct Neumann condition here is rather  $\varrho \mathbf{v} \cdot \mathbf{n} = 0$  a.e., which means that particles could enter from  $\partial\Omega$  into the interior, but immediately after it happens 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  $\Omega$ , then we would have a violation of the Neumann condition, since there would be (intuitively) 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 [44] 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  $\partial\Omega$ . Their definition is as follows: given two finite positive measures  $\mu, \nu \in \mathcal{M}_+(\overset{\circ}{\Omega})$  (not necessarily probabilities, not necessarily with the same mass), we define

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

Then, we set

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

The index  $b$  stands for the special role played by the boundary. Informally, this means that the transport from  $\mu$  to  $\nu$  may be done usually (with a part of  $\gamma$  concentrated on  $\overset{\circ}{\Omega} \times \overset{\circ}{\Omega}$ ), or by moving some mass from  $\mu$  to  $\partial\Omega$  (using  $\gamma \llcorner (\overset{\circ}{\Omega} \times \partial\Omega)$ ), then moving from one point of the boundary to another point of the boundary (this should be done by using  $\gamma \llcorner (\partial\Omega \times \partial\Omega)$ , but since this part of  $\gamma$  does not appear in the constraints, then we can forget about it, and the transport is finally free on  $\partial\Omega$ ), and finally from  $\partial\Omega$  to  $\nu$  (using  $\gamma \llcorner (\partial\Omega \times \overset{\circ}{\Omega})$ ).

In [44] the authors prove that  $Wb_2$  is a distance, that the space  $\mathcal{M}_+(\overset{\circ}{\Omega})$  is always a geodesic space, independently of convexity or connectedness properties of  $\Omega$  (differently from what happens with  $\Omega$ , since here the transport is allowed to “teleport” from one part of the boundary to another, either to pass from one connected component to another or to follow a shorter path going out of  $\Omega$ ), and they study the gradient flow, for this distance, of the functional  $F(\varrho) = \int (\varrho \log \varrho - \varrho) dx$ . Note that in the usual study of the entropy on  $\mathcal{P}(\Omega)$  one can decide to forget the term  $-\int \varrho$ , which is anyway a constant because the total mass is fixed. Here this term becomes important (if the function  $f(t) = t \log t - t$  is usually preferred to  $t \log t$ , it is because its derivative is simpler,  $f'(t) = \log t$ , without changing its main properties).

With this choice of the functional and of the distance, the gradient flow that Figalli and Gigli obtain is the Heat equation with the particular boundary condition  $\varrho = 1$  on  $\partial\Omega$ . One could wonder where the constant 1 comes from, and a reasonable explanation is the following: if the transport on the boundary is free of charge, then automatically the solution selects the value which is the most performant for the functional, i.e. the constant  $t$  which minimizes  $f(t) = t \log t - t$ . In this way, changing the linear part and using  $F(\varrho) = \int (\varrho \log \varrho - c\varrho) dx$  could change the constant on the boundary, but the constant 0 is forbidden for the moment. It would be interesting to see how far one could go with this approach and which Dirichlet conditions and which equations could be studied in this way, but this does not seem to be done at the moment.

Moreover, the authors explain that, due to the lack of geodesic convexity of the entropy w.r.t.  $Wb_2$ , the standard abstract theory of gradient flows is not able to provide uniqueness results (the lack of convexity is due in some sense to the possible concentration of mass on the boundary, in a way similar to what happened in [64] when dealing with the door on  $\partial\Omega$ ). On the other hand, standard Hilbertian results on the Heat equation can provide uniqueness for this equation, as the authors smartly remark in [44].

We observe that this kind of distances with free transport on the boundary were already present in [22, 21], but in the case of the Wasserstein distance  $W_1$ , and the analysis in those papers was not made for applications to gradient flows, which are less natural to study with  $p = 1$ . We can also point out a nice duality formula:

$$Wb_1(\mu, \nu) := \min \left\{ \int |x - y| d\gamma : \gamma \in \Pi b(\mu, \nu) \right\} = \sup \left\{ \int u d(\mu - \nu) : u \in \text{Lip}_1(\Omega), u = 0 \text{ on } \partial\Omega \right\}.$$

In the special case  $p = 1$  and  $\mu(\overset{\circ}{\Omega}) = \nu(\overset{\circ}{\Omega})$ , we also obtain

$$Wb_1(\mu, \nu) = \mathcal{T}_c(\mu, \nu), \quad \text{for } c(x, y) = \min\{|x - y|, d(x, \partial\Omega) + d(y, \partial\Omega)\}.$$

The cost  $c$  is a pseudo-distance on  $\Omega$  where moving on the boundary is free. This kind of distance has also been used in [28] (inspired by [21]) to model free transport costs on other lower dimensional sets, and not only the boundary (with the goal to model, for instance, transportation networks, and optimize their shape). It is interesting to see the same kind of ideas appear for so different goals.

#### 4.8 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. 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  $\nu$  (to be taken equal to  $\varrho_k^\tau$ ) and suitable  $F$  (including the  $\tau$  factor). We will not consider the convergence as  $\tau \rightarrow 0$  of the iterations to solutions of the evolution equation.

We will present two methods. One, essentially taken from [16], is based on the Benamou-Brenier formula first introduced in [14] as a numerical tool for optimal transport. This method is well-suited for the case where the energy  $F(\varrho)$  used in the gradient flow is a convex function of  $\varrho$ . For instance, it works for functionals of the form  $F(\varrho) = \int f(\varrho(x))dx + \int Vd\varrho$  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 [72, 60] and [62] for 3D implementation) and translates 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  $\int f(\varrho(x))dx$  is only admissible if  $f$  satisfies McCann's condition, the term  $\int Vd\varrho$  needs  $V$  to be convex, but interaction terms such as  $\int \int W(x-y) d\varrho(x) d\varrho(y)$  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 (4.8) which reads, in the quadratic case, as

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

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

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

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  $m$  for  $(\varrho, E)$  and  $A$  for  $(a, b)$ . We also set  $G(\phi) := \int \phi_1 d\nu - \int \phi_0 d\mu$ . Then the problem becomes

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

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

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

which means that we look for a pair  $(m, (A, \phi))$  (actually, a triple, but  $A$  and  $\phi$  play together the role of the second variable) where  $m$  minimizes for fixed  $(A, \phi)$  and  $(A, \phi)$  maximizes for fixed  $m$ . 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. 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)

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

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

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

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

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

(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  $m$ , finding the optimal  $\phi$  (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  $\phi$  and  $m$  find the optimal  $A$  (which amounts to a pointwise minimization problem, in this case a projection on the convex set  $K_2$ ); finally update  $m$  by going in the direction of the gradient descent, i.e. replacing  $m$  with  $m - r(A - \nabla_{t,x}\phi)$  (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,  $\mu$ ), but the other is not, and a penalization on the final  $\varrho_1$  is added, of the form  $\tau F(\varrho_1)$ . Inspired from the considerations above, the saddle point below allows to treat the problem

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

by formulating it as

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

where we re-inserted the integration signs to underline the difference between integrals in space-time (with  $m, A$  and  $\phi$ ) and in space only (with  $\phi_0, \phi_1, \varrho_1, V$  and  $\lambda$ ). The role of the variable  $\lambda$  is to be dual to  $\varrho_1$ , 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  $m$  and  $\varrho_1$ , and do not require further explications. The first three steps consist in the optimization of  $\phi$  (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  $\lambda$  (the minimization of  $f^*(\lambda) + \frac{\tau}{2} |\phi_1(x) + \lambda + V(x)|^2 - \varrho_1(x)\lambda$ , 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  $\varrho_{k+1}^\tau$  equal to the optimizer  $\varrho_1$  and the functions  $f$  and  $V$  must include the scale factor  $\tau$ . The time-space  $[0, 1] \times \Omega$  has to be discretized but the evolution in time is infinitesimal (due to the small time scale  $\tau$ ), 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  $k \dots$

The interested reader can consult [16] 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

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

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

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

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  $\mu$  itself is absolutely continuous, with a density that we will denote by  $\varrho_0$ ) is given by<sup>14</sup>

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

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 [35] that we cannot approximate

<sup>14</sup>This same formula takes into account the possibility that  $\nabla u$  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.

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 [42], but this needed many more linear constraints; Oudet and Mérigot [73] decided to test convexity on a grid different (and less refined) than that where the functions are defined. . . These methods give somehow satisfactory answers for functionals involving  $u$  and  $\nabla u$ , but are not able to handle terms involving the Monge-Ampère operator  $\det(D^2 u)$ .

The method proposed in [17], that we will roughly present here, does not really use a prescribed mesh. The idea is the following: suppose that  $\mu$  is a discrete measure of atomic type, i.e. of the form  $\sum_j a_j \delta_{x_j}$ . A convex defined on its support  $S := \{x_j\}_j$  will be a function  $u : S \rightarrow \mathbb{R}$  such that at each point  $x \in S$  the subdifferential

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

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

$$\int_B \det(D^2 u(x)) dx = |\partial u(B)|$$

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  $\int f(\varrho(x)) dx$  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  $(u, P) : S \rightarrow \mathbb{R} \times \mathbb{R}^d$  where  $P(x) \in \partial u(x)$  for every  $x \in S$ . For every such pair  $(u, P)$  we need 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  $\Omega$  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  $\nabla u$ ) as soon as  $F$  is displacement convex in  $\varrho$  (actually, convexity on generalized geodesics based at  $\mu$  should be the correct notion).

Then we are lead to study the variational problem

$$\min_{u, P} \frac{1}{2} \sum_j a_j |P(x_j) - x_j|^2 + \mathcal{V}(G(u, P)) + \mathcal{W}(G(u, P)) + \mathcal{F}(G^{ac}(u, P)) \quad (4.26)$$

under the constraints  $P(x_j) \in A_j := \partial u(x_j) \cap \Omega$ . Note that we should incorporate the scale factor  $\tau$  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  $\tau$  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  $\psi_j$ , 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<sup>15</sup> given by the values  $\psi_j$ . It is not difficult to see that also in this case cells are convex polyhedra. And it is also easy to see that, if  $\varrho$  is an absolutely continuous measure on  $\Omega$  and  $\mu = \sum_j a_j \delta_{x_j}$ , then finding an optimal transport map from  $\varrho$  to  $\mu$  is equivalent to finding values  $\psi_j$  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  $\Omega$ ).

---

<sup>15</sup>If the points  $x_j$  are the locations of some ice-cream sellers, we can think that  $\psi_j$  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.



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 very easily their volumes (incidentally, also find some points  $P$  belonging to them, which is useful so as to satisfy the constraints of Problem (4.26)), as well as the derivatives of their volumes (which depend on the measures of each faces) w.r.t. the values  $\psi_j$ . For the applications to semi-discrete<sup>16</sup> 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 [72, 60, 62] 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  $\varrho_0$  with a finite number of Dirac masses located at points  $x_j$ , to fix  $\tau > 0$  small, then to solve (4.26) 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 [17].

## 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.

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  $\varrho_0$  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  $\Omega$  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.

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

---

<sup>16</sup>One measure being absolutely continuous, the other atomic.



the Lebesgue measure in the Euclidean space). Such a measure is needed in order to define the integral  $\int |\nabla \varrho|^2 dx$ , and also the entropy  $\int \varrho \log \varrho dx$ . Roughly speaking, we need to define “ $dx$ ”.

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: [5, 9, 10, 7, 34, 49, 51, 52, 53, 54, 63, 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 [52, 53, 34, 85]). Moreover, the recent works by Ambrosio, Gigli and Savaré ([9, 7]) 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 [54], 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 Hajlasz, 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

$$g = \text{const} \cdot M[|\nabla f|],$$

where  $M[u]$  denotes the maximal function of  $u$ :  $M[u](x) := \sup_{r>0} \int_{B(x,r)} u$  (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 ([34]), using the relaxation starting from Lipschitz functions, or to Shanmuganlingam ([85]), based on the inequality

$$|f(x(0)) - f(x(1))| \leq \int_0^1 |\nabla f(x(t))| |x'(t)| dt$$

required to hold on almost all curves, in a suitable sense. The recent paper [7] resents a classification of the various weak notions of modulus of the gradient in a MMS and analyzes their equivalence. On the contrary, here we will only choose one unique definition for  $\int |\nabla f|^2 dm$ , 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  $\mathcal{D}(f) := \int |\nabla f|^2(x) dm$ . Then, by relaxation, we define the *Cheeger Energy*<sup>17</sup>  $C(f)$ :

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

<sup>17</sup>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 for quadratic forms.

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

Defining the flow of  $C$  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 [25]). This brings very general existence and uniqueness results.

## 5.2 A well-posed gradient flow for the entropy

A second step (first developed in [49] and then generalized in [9]) 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  $\mathcal{E}(f) := \int f \log f \, dm$  and we look at its gradient flow in  $\mathbb{W}_2$  in the EDE sense. In order to apply the general theory of Section 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  $\lambda$ -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 Section 4.4, 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  $\mathcal{E} : \mathcal{P}(X) \rightarrow \mathbb{R} \cup \{+\infty\}$  defined through*

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

*is  $K$ -geodesically convex in the space  $\mathbb{W}_2(X)$ . In this case we say that  $(X, d, m)$  satisfies the condition<sup>18</sup>  $CD(K, \infty)$ .*

Note that the EVI formulation is not available in this case, as we do not have the geodesic convexity of the squared Wasserstein. 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 [49] 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  $\mu(0) = \bar{\mu}$ .*

<sup>18</sup>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”).

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:

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

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  $C$  (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.  $(C, L^2) \Rightarrow (\mathcal{E}, W_2)$ ) 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 estimates

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

This computation is based on a strategy essentially developed in [51] and on a lemma by Kuwada. The initial proof, contained in [51], is valid for Alexandroff spaces<sup>19</sup>. The generalization of the same result to arbitrary MMS satisfying  $CD(K, \infty)$  is done in [9].

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

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

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

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

(where the slope of  $\mathcal{E}$  is computed for the  $W_2$  distance<sup>20</sup>). Also, if we consider the curve  $\varrho_t = f_t \cdot \mathcal{H}^d$ , it happens that  $\varrho_t$  is an AC curve in the space  $\mathbb{W}_2(X)$  and

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

these three estimates imply that  $\varrho_t$  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 [10] for the proofs):

<sup>19</sup>These spaces, see [26], 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.

<sup>20</sup>As in (5.1).

- 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  $C$  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 [10] a definition of MMS having a *Riemannian 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 Section 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 a 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 spaces, with a weak convergence of the measures), but not after a simple observation: also the class of convex, or  $\lambda$ -convex, functions is stable under uniform (or even weak) convergence! but, of course, proving this stability is not a trivial fact.

## References

- [1] M. AGUEH AND M. BOWLES, One-dimensional numerical algorithms for gradient flows in the  $p$ -Wasserstein spaces, *Acta applicandae mathematicae* 125 (2013), no. 1, 121–134.
- [2] L. AMBROSIO, Movimenti minimizzanti, *Rend. Accad. Naz. Sci. XL Mem. Mat. Sci. Fis. Natur.* 113, 191–246, 1995.
- [3] L. AMBROSIO, Lecture Notes on Optimal Transport Problems, *Mathematical Aspects of Evolving Interfaces*, Springer Verlag, Lecture Notes in Mathematics (1812), 1–52, 2003.
- [4] L. AMBROSIO, Transport equation and Cauchy problem for BV vector fields, *Inventiones Mathematicae*, 158 (2004), 227–260.
- [5] L. AMBROSIO AND N. GIGLI A user’s guide to optimal transport, in *Modelling and Optimisation of Flows on Networks* Lecture Notes in Mathematics, 1–155, 2013.
- [6] L. AMBROSIO, N. GIGLI AND G. SAVARÉ, *Gradient flows in metric spaces and in the spaces of probability measures*. Lectures in Mathematics, ETH Zurich, Birkhäuser, 2005.
- [7] L. AMBROSIO, N. GIGLI AND G. SAVARÉ, Density of Lipschitz functions and equivalence of weak gradients in metric measure spaces, *Rev. Mat. Iberoamericana* 29 (2013) 969-986.
- [8] L. AMBROSIO, N. GIGLI AND G. SAVARÉ, Heat flow and calculus on metric measure spaces with Ricci curvature bounded below - the compact case, *Analysis and Numerics of Partial Differential*

*Equations*, Brezzi F., Colli Franzone P., Gianazza U.P., Gilardi G. (ed.) Vol. 4, “INDAM”, Springer (2013) 63-116

- [9] L. AMBROSIO, N. GIGLI AND G. SAVARÉ, Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below, *Inv. Math.* 195 (2), 289–391, 2014.
- [10] L. AMBROSIO, N. GIGLI AND G. SAVARÉ, Metric measure spaces with Riemannian Ricci curvature bounded from below, *Duke Math. J.* 163 (2014) 1405–1490
- [11] L. AMBROSIO AND G. SAVARÉ, Gradient flows of probability measures, *Handbook of differential equations, Evolutionary equations* **3**, ed. by C.M. Dafermos and E. Feireisl (Elsevier, 2007).
- [12] L. AMBROSIO AND P. TILLI, *Topics on analysis in metric spaces*. Oxford Lecture Series in Mathematics and its Applications (25). Oxford University Press, Oxford, 2004.
- [13] J.-P. AUBIN, Un théorème de compacité. (French). *C. R. Acad. Sci. Paris* 256. pp. 5042–5044
- [14] J.-D. BENAMOU AND Y. BRENIER, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, *Numer. Math.*, 84 (2000), 375–393.
- [15] J.-D. BENAMOU AND G. CARLIER, Augmented Lagrangian Methods for Transport Optimization, Mean Field Games and Degenerate Elliptic Equations, *J. Opt. Theor. Appl.*, (2015), *to appear*.
- [16] J.-D. BENAMOU, G. CARLIER AND M. LABORDE An augmented Lagrangian approach to Wasserstein gradient flows and applications, preprint.
- [17] J.-D. BENAMOU, G. CARLIER, Q. MÉRIGOT AND É. OUDET Discretization of functionals involving the Monge-Ampère operator, *Numerische Mathematik*, to appear.
- [18] A. BLANCHET, V. CALVEZ AND J.A. CARRILLO, Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model, *SIAM J. Numer. Anal.* 46 (2), 691–721, 2008.
- [19] A. BLANCHET, J.-A. CARRILLO, D. KINDERLEHRER, M. KOWALCZYK, P. LAURENÇOT AND S. LISINI, A Hybrid Variational Principle for the Keller-Segel System in  $\mathbb{R}^2$ , *ESAIM M2AN*, Vol. 49, no 6, 1553–1576, 2015.
- [20] N. BONNOTTE *Unidimensional and Evolution Methods for Optimal Transportation*, PhD Thesis, Université Paris-Sud, 2013
- [21] G. BOUCHITTÉ AND G. BUTTAZZO, Characterization of optimal shapes and masses through Monge-Kantorovich equation *J. Eur. Math. Soc.* 3 (2), 139–168, 2001.
- [22] G. BOUCHITTÉ, G. BUTTAZZO AND P. SEPPECHER, Shape optimization solutions via Monge-Kantorovich equation. *C. R. Acad. Sci. Paris Sér. I Math.* 324 (10), 1185–1191, 1997.
- [23] Y. BRENIER, Décomposition polaire et réarrangement monotone des champs de vecteurs. (French) *C. R. Acad. Sci. Paris Sér. I Math.* 305 (19), 805–808, 1987.

- [24] Y. BRENIER, Polar factorization and monotone rearrangement of vector-valued functions, *Communications on Pure and Applied Mathematics* 44, 375–417, 1991.
- [25] H. BREZIS *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland mathematics studies, 1973.
- [26] YU. BURAGO, M. GROMOV AND G. PERELMAN A. D. Alexandrov spaces with curvatures bounded below (Russian), *Uspekhi Mat. Nauk* **47** (1992), 3–51; English translation: *Russian Math. Surveys* **47** (1992), 1–58.
- [27] G. BUTTAZZO, *Semicontinuity, relaxation, and integral representation in the calculus of variations* Longman Scientific and Technical, 1989.
- [28] G. BUTTAZZO, É. OUDET AND E. STEPANOV. Optimal transportation problems with free Dirichlet regions. In *Variational methods for discontinuous structures*, 41–65, vol 51 of *PNLDE*, Birkhäuser, Basel, 2002.
- [29] C. CANCÈS, T. GALLOUËT AND L. MONSAINGEON Incompressible immiscible multiphase flows in porous media: a variational approach, preprint, 2016.
- [30] J.-A. CARRILLO, M. DI FRANCESCO, A. FIGALLI, T. LAURENT AND D. SLEPČEV, Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations *Duke Math. J.* 156 (2), 229–271, 2011.
- [31] J.-A. CARRILLO, R.J. MCCANN AND C. VILLANI, Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates, *Revista Matemática Iberoamericana* 19, 1–48, 2003.
- [32] J.-A. CARRILLO, R.J. MCCANN AND C. VILLANI, Contractions in the 2-Wasserstein length space and thermalization of granular media, *Arch. Rat. Mech. An.* 179, 217–263, 2006.
- [33] J.-A. CARRILLO AND D. SLEPČEV Example of a displacement convex functional of first order *Calculus of Variations and Partial Differential Equations* 36 (4), 547–564, 2009.
- [34] J. CHEEGER Differentiability of Lipschitz functions on metric measure spaces, *Geom. Funct. Anal.* **9** (1999), 428–517.
- [35] P. CHONÉ AND H. LE MEUR, Non-convergence result for conformal approximation of variational problems subject to a convexity constraint, *Numer. Funct. Anal. Optim.* 5-6 (2001), no. 22, 529–547.
- [36] K. CRAIG Nonconvex gradient flow in the Wasserstein metric and applications to constrained non-local interactions, preprint available at <http://arxiv.org/abs/1512.07255>
- [37] S. DANERI AND G. SAVARÉ, Eulerian Calculus for the Displacement Convexity in the Wasserstein Distance. *SIAM J. Math. An.* 40, 1104–1122, 2008.
- [38] E. DE GIORGI, New problems on minimizing movements, *Boundary Value Problems for PDE and Applications*, C. Baiocchi and J. L. Lions eds. (Masson, 1993) 81–98.

- [39] S. DI MARINO AND A. R. MÉSZÁROS Uniqueness issues for evolution equations with density constraints, *Math. Models Methods Appl. Sci.*, to appear (2016)
- [40] R. J. DiPERNA AND P. L. LIONS, Ordinary differential equations, transport theory and Sobolev spaces, *Inventiones mathematicae* 98.3 (1989): 511–548.
- [41] I. EKELAND, On the variational principle, *J. Math. Anal. Appl.*, vol. 47, no 2, 1974, p. 324–353.
- [42] I. EKELAND AND S. MORENO-BROMBERG, An algorithm for computing solutions of variational problems with global convexity constraints, *Numerische Mathematik* 115 (2010), no. 1, 45–69.
- [43] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Classics in Mathematics, Society for Industrial and Applied Mathematics (1999).
- [44] A. FIGALLI AND N. GIGLI, A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions, *J. Math. Pures et Appl.*, 94 (2), 107–130, 2010.
- [45] M. FORTIN AND R. GLOWINSKI, *Augmented Lagrangian methods, Applications to the Numerical Solution of Boundary-Value Problems*, North-Holland (1983).
- [46] W. GANGBO, An elementary proof of the polar factorization of vector-valued functions, *Arch. Rational Mech. Anal.*, 128, 381–399, 1994.
- [47] W. GANGBO, The Monge mass transfer problem and its applications, *Contemp. Math.*, 226, 79–104, 1999.
- [48] W. GANGBO AND R. McCANN, The geometry of optimal transportation, *Acta Math.*, 177, 113–161, 1996.
- [49] N. GIGLI On the Heat flow on metric measure spaces: existence, uniqueness and stability, *Calc. Var. Part. Diff. Eq.* **39** (2010), 101–120.
- [50] N. GIGLI *Propriétés géométriques et analytiques de certaines structures non lisses*, Mémoire HDR, Univ. Nice-Sophia-Antipolis, 2011.
- [51] N. GIGLI, K. KUWADA AND S. I. OHTA, Heat flow on Alexandrov spaces, *Comm. Pure Appl. Math.* Vol. LXVI, 307–331 (2013).
- [52] P. HAJŁASZ, Sobolev spaces on an arbitrary metric space, *Potential Analysis* **5** (1996), 403–415.
- [53] P. HAJŁASZ, Sobolev spaces on metric-measure spaces, *Contemp. Math.* **338** (2003), 173–218.
- [54] P. HAJŁASZ AND P. KOSKELA Sobolev met Poincaré, *Mem. Amer. Math. Soc.* **688** (2000), 1–101.
- [55] R. JORDAN, D. KINDERLEHRER AND F. OTTO, The variational formulation of the Fokker-Planck equation, *SIAM J. Math. Anal.* 29 (1), 1–17, 1998.
- [56] L. KANTOROVICH, On the transfer of masses. *Dokl. Acad. Nauk. USSR*, 37, 7–8, 1942.

- [57] E. F. KELLER AND L. A. SEGEL, Initiation of slime mold aggregation viewed as an instability, *J. Theor. Biol.*, 26, 399–415, 1970.
- [58] E. F. KELLER AND L. A. SEGEL, Model for chemotaxis, *J. Theor. Biol.*, 30, 225–234, 1971.
- [59] D. KINDERLEHRER AND N. J. WALKINGTON, Approximation of parabolic equations using the Wasserstein metric, *ESAIM: Mathematical Modelling and Numerical Analysis* 33 (1999), no. 04, 837–852.
- [60] J. KITAGAWA, Q. MÉRIGOT AND B. THIBERT, Convergence of a Newton algorithm for semi-discrete optimal transport, preprint, 2016
- [61] G. LEGENDRE AND G. TURINICI, Second order in time schemes for gradient flows in Wasserstein and geodesic metric spaces, preprint, 2016.
- [62] B. LÉVY A numerical algorithm for  $L^2$  semi-discrete optimal transport in 3D *ESAIM: Mathematical Modelling and Numerical Analysis* 49 (6), 1693–1715, 2015
- [63] J. LOTT AND C. VILLANI Ricci curvature for metric-measure spaces via optimal transport *Ann. of Math.* 169, 903–991, 2009.
- [64] B. MAURY, A. ROUDNEFF-CHUPIN AND F. SANTAMBROGIO A macroscopic crowd motion model of gradient flow type, *Math. Models and Methods in Appl. Sciences* 20 (10), 1787–1821, 2010.
- [65] B. MAURY, A. ROUDNEFF-CHUPIN, F. SANTAMBROGIO AND J. VENEL, Handling congestion in crowd motion modeling, *Net. Het. Media* 6 (3), 485–519, 2011.
- [66] B. MAURY AND J. VENEL, A mathematical framework for a crowd motion model, *C. R. Acad. Sci. Paris, Ser. I* 346 (2008), 1245–1250.
- [67] B. MAURY AND J. VENEL, A discrete contact model for crowd motion *ESAIM: M2AN* 45 (1), 145–168, 2011.
- [68] R. J. McCANN, Existence and uniqueness of monotone measure preserving maps, *Duke Math. J.*, 80, 309–323, 1995.
- [69] R. J. McCANN A convexity principle for interacting gases. *Adv. Math.* 128 (1), 153–159, 1997.
- [70] R. J. McCANN Exact solutions to the transportation problem on the line. *Proc. Royal Soc. London Ser. A* 455, 1341–1380, 1999.
- [71] R. J. McCANN Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.*, 11 (3), 589–608, 2001.
- [72] Q. MÉRIGOT A multiscale approach to optimal transport. *Computer Graphics Forum* 30 (5) 1583–1592, 2011.
- [73] Q. MÉRIGOT AND É. OUDET Handling convexity-like constraints in variational problems *SIAM Journal on Numerical Analysis*, 52 (5), 2466–2487, 2014



- [74] A. R. MÉSZÁROS AND F. SANTAMBROGIO Advection-diffusion equations with density constraints, *Analysis & PDE* 9-3 (2016), 615–644.
- [75] G. MONGE, Mémoire sur la théorie des déblais et des remblais, *Histoire de l'Académie Royale des Sciences de Paris*, 666–704, 1781.
- [76] F. OTTO, The geometry of dissipative evolution equations: The porous medium equation, *Comm. Partial Differential Equations*, 26, 101–174, 2011.
- [77] F. PITIÉ, A. C. KOKARAM AND R. DAHYOT. Automated colour grading using colour distribution transfer. *Comp. Vis. Image Understanding*, 107 (1–2), 123–137, 2007.
- [78] J. RABIN, G. PEYRÉ, J. DELON AND M. BERNOT. Wasserstein Barycenter and Its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision*, edited by A. M. Bruckstein, B. M. Haar Romeny, A.M. Bronstein, and M. M. Bronstein, Lecture Notes in Computer Science, vol. 6667, 435–446, Springer Berlin Heidelberg, 2012.
- [79] M.-K. VON RENESSE AND K.-T. STURM, Entropic measure and Wasserstein diffusion, *Ann. Probab.* **37** (2009), 1114–1191.
- [80] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, 1970.
- [81] A. ROUDNEFF-CHUPIN, *Modélisation macroscopique de mouvements de foule*, PhD Thesis, Université Paris-Sud, 2011, available at [www.math.u-psud.fr/~roudneff/Images/these\\_roudneff.pdf](http://www.math.u-psud.fr/~roudneff/Images/these_roudneff.pdf)
- [82] F. SANTAMBROGIO Gradient flows in Wasserstein spaces and applications to crowd movement, *Séminaire Laurent Schwartz* no 27, École Polytechnique, 2010.
- [83] F. SANTAMBROGIO, Flots de gradient dans les espaces métriques et leurs applications (d’après Ambrosio-Gigli-Savaré), proceedings of the *Bourbaki Seminar*, 2013 (in French).
- [84] F. SANTAMBROGIO *Optimal Transport for Applied Mathematicians*, Progress in Nonlinear Differential Equations and Their Applications no 87, Birkhäuser Basel (2015).
- [85] N. SHANMUGALINGAM Newtonian spaces: an extension of Sobolev spaces to metric measure spaces, *Rev. Mat. Iberoamericana* **16** (2000), 243–279.
- [86] SCHNEIDER, *Convex bodies: the Brunn-Minkowski theory*, vol. 44, Cambridge Univ. Press, 1993.
- [87] K.-T. STURM On the geometry of metric measure spaces. I. *Acta Math.* 196, 65–131, 2006.
- [88] K.-T. STURM On the geometry of metric measure spaces. II. *Acta Math.* 196, 133–177, 2006.
- [89] C. VILLANI *Topics in Optimal Transportation*. Graduate Studies in Mathematics, AMS, 2003.
- [90] C. VILLANI, *Optimal transport: Old and New*, Springer Verlag, 2008.