José A. Carrillo

Optimal Transport and Partial Differential Equations

Mathematical Institute, Notes MT21

Preface

This course will serve as an introduction to optimal transportation theory, its application in the analysis of PDE, and its connections to the macroscopic description of interacting particle systems. The optimal transportation problem started with Gaspard Monge in late XVIII century with his seminal work "Mémoire sur la théorie des déblais et des remblais" and expanded by Leonid Kantorovich with connections to economics. Brenier's dynamical formulation of optimal transport in the 80's-90's gave rise to a flurry of applications of optimal mass transportation theory in PDE theory, geometry, engineering, and lately in data science, that has been increasing in the last 30 years. This course will cover some of the basic notions of transportation metrics between probability measures as well as applications in mean-field limits and PDE as gradient flows or steepest descent in spaces of probability measures.

The main learning outcomes are: Getting familiar with the Monge-Kantorovich problem and transport distances. Derivation of macroscopic models via the mean-field limit and their analysis based on stability of transport distances. Dynamic Interpretation and Geodesic convexity. A brief introduction to gradient flows and examples. Prerequisites: A4 Integration. The short option in Calculus of Variation in Part A and functional analysis courses will ease understanding concepts but not compulsory.

Regarding textbooks to find basic material I advice to look up the general monographs in optimal transport theory [21, 18], the book [12] for basic related material in functional analysis, the lecture notes from summer schools [11, 2, 8], and finally [14] for the mean-field limit and [20] for nonlinear diffusions. Additional material can be found related to courses taught at University of Cambridge [19] and at ETH-Zurich [13]. Further complementary material can also be found in [22, 4, 3].

Contents

Chapter 1
Interacting Particle Systems & PDE

This course is devoted to the analysis of solutions of the following family of Partial Differential Equations

$$\frac{\partial \rho}{\partial t} = \nabla \cdot [\rho \nabla (V + W * \rho)] + \Delta P(\rho), \tag{1.1}$$

where the unknown $\rho(t,\cdot)$ is a time-dependent probability measure on $\mathbb{R}^d$ ($d \geq 1$), $P : [0,\infty) \to \mathbb{R}$ is an increasing function with $P(0) = 0$, $V : \mathbb{R}^d \to \mathbb{R}$ is a confinement potential and $W : \mathbb{R}^d \to \mathbb{R}$ is an interaction potential. The symbols ∇ and Δ denote the gradient and the Laplacian operators and will always be applied to functions, while $\nabla\cdot$ stands for the divergence operator, and will always be applied to vector fields. In the sequel, we identify both the probability measure $\rho(t,\cdot) = \rho_t$ with its Radon-Nikodym density $d\rho_t/dx$ with respect to Lebesgue, and thus, we use the notation $d\rho_t = d\rho(t,x) = \rho(t,x)\,dx$ unless discussing about general probability measures. The basic assumptions on P implies that the last term in (1.1) represents a diffusion term. The interaction potential W is always assumed to be symmetric: $\forall z \in \mathbb{R}^d$, $W(-z) = W(z)$. Finally, the smoothness on the potentials V and W will be specified in each particular case.

We will be interested in understanding the well-posednes and the qualitative properties of solutions to (1.1) given by curves of probability densities, i.e., we are looking for solutions such that $\rho(t,\cdot) \in L^1_+(\mathbb{R}^d)$ for all $t \geq 0$, and even sometimes we will work with curves of probability measures. Sometimes in particular models the measures will not be normalized to unit mass, but we will be assuming that we look for nonnegative integrable solutions with a fixed given mass.

1.1 Aggregation Equation: Granular Flow Models.

Rapid granular flow models were developed to describe dissipative or inelastic collisions between particles by statistical mechanic approaches. A basic model that

triggered the attention of researchers in kinetic theory at the end of the 90's on this type of equations (1.1) with $P = 0$ can be introduced on the real line. Assume we have particles on the real line moving freely until they collide, while they loose part of the relative velocity in each collision. Denopting by v and w the velocities of these particle before collision, and assuming conservation of the momentum but a loss of their relative velocity measured by the restitution coefficient $0 \leq r \leq 1$, we can write the post-collisional velocities by

$$v' = \frac{1}{2}(v+w) + \frac{r}{2}(v-w); \qquad w' = \frac{1}{2}(v+w) - \frac{r}{2}(v-w). \tag{1.2}$$

A more suitable form of (1.2) can be obtained by setting the coefficient of restitution $r = 1-2\bar{r}$, where now $0 \leq \bar{r} \leq 1/2$ is the dissipation parameter. In terms of $\bar{r}$, the dissipative collision reads

$$v' = (1-\bar{r})v + \bar{r}w; \qquad w' = \bar{r}v + (1-\bar{r})w. \tag{1.3}$$

Note that $r = 1$ corresponds to elastic collisions that in one dimension leads to trivial dynamics, swapping of labels for the particles. An integral equation given the evolution of the statistical distribution of the velocities of the particles on the line can be phenomenologically introduced of the form

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = Q_r(f,f), \tag{1.4}$$

usually called a Boltzmann type equation, where the unknown is the statistical distribution $f(t,x,v)$ in position x and velocity v at time $t \geq 0$. The right hand side models the gain and loss of particles with a given velocity v due to collisions with other particles. The dissipative Boltzmann collision operator $Q_r(f,f)$ is usually defined in its weak form, that is, in how it acts on given test functions $\varphi \in C^\infty(\mathbb{R})$

$$< \varphi, Q_r(f,f) >= \int_{\mathbb{R}}\int_{\mathbb{R}} B(|v-w|)f(v)f(w)\Big[\varphi(v') - \varphi(v)\Big]\,dv\,dw, \tag{1.5}$$

with $B(z)$, $z \in [0,\infty)$, being the collision frequency, i. e., the probability of collision of two particles may depend on the relative velocity at which they are colliding. Typical values of the collision frequency are $B(z) = |z|^\gamma$ with $\gamma \geq -1$, being $\gamma = 1$ refered as inelastic hard spheres.

Notice that in order for the right hand side in (1.5) to be well defined, f must belong to some L^p spaces and satisfy certain moments in v bounded depending on the growth of the test functions, but we proceed formally in order to understand further the model. As mentioned earlier, in one–dimension of velocity space an elastic binary collision particles simply exchange their velocities and the Bolzmann collision operator for elastic collisions disappears $Q_1 = 0$. By the symmetry of the collision mechanism (1.3), we can write the collision operator as

$$<\varphi, Q_r(f,f)>=\frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}} B(|v-w|)f(v)f(w)\Big[\varphi(v')+\varphi(w')-\varphi(v)-\varphi(w)\Big]\,dv\,dw. \tag{1.6}$$

Let us now focus on the homogeneous problem, meaning that we assume the initial data is homogeneous in space and we look for solutions only depending on the velocity variable in order to understand just the velocity distribution, i.e., $f(t,v)$ satisfies

$$\frac{\partial f}{\partial t} = Q_r(f,f). \tag{1.7}$$

It is easy to check that the homogeneous Boltzmann equation conserves mass, momentum and dissipates energy, meaning that

$$<1, Q_r(f,f)>=<v, Q_r(f,f)>=0$$

and

$$<v^2, Q_r(f,f)>=-\frac{(1-r)^2}{4}\int_{\mathbb{R}}\int_{\mathbb{R}} B(|v-w|)(v-w)^2 f(v)f(w)\,dv\,dw.$$

These properties mean that solutions to (1.7) should be probability measures conserving their mean and dissipating the kinetic energy by multiplying (1.7) by 1, v and v^2 and integrating in v. Due to translational invariance, let us assume that the mean velocity is zero, i.e.,

$$\int_{\mathbb{R}} v f(t,v)\,dv = 0, \quad \forall t \geq 0. \tag{1.8}$$

Let us look for simpler models, assuming that the inelasticity is small $r \simeq 1$ or equivalentely $\bar{r} \simeq 0$, we approximate the Boltzmann collision operator by expanding in the expression (1.5) to get

$$\varphi(v') - \varphi(v) \simeq \frac{\partial \varphi}{\partial v}(v)(v'-v) = -\bar{r}(v-w)\frac{\partial \varphi}{\partial v}(v).$$

Therefore, we can approximate the collision operator $Q_r(f,f)$ by

$$<\varphi, Q_r(f,f)>\simeq -\bar{r}\int_{\mathbb{R}}\int_{\mathbb{R}} B(|v-w|)(v-w)f(v)f(w)\frac{\partial \varphi}{\partial v}(v)\,dv\,dw. \tag{1.9}$$

The right-hand side of (1.9) is the weak form of a differential operator, thus we can finally write a one-dimensional simplified granular flow model as

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial v}\left[f\left(\frac{\partial W}{\partial v} * f\right)\right], \qquad \text{with } \frac{\partial W}{\partial v} = vB(|v|), \tag{1.10}$$

where the factor $\bar{r}$ is absorbed in the time derivative. Notice that for the typical cases of collision frequencies, $W(v) = \frac{|v|^{\gamma+2}}{\gamma+2}$, $\gamma \geq -1$. Therefore, this simplified granular

flow model corresponds to cases of the general family of PDE (1.1) with $V = 0$, $P = 0$ and convex interaction potentials W.

Intuitively, we should expect concentration in velocity variable as time evolves due to the inelasticity of the interactions, particles will start to decrease their relative velocities until eventually reaching rest state. Is this captured by the simplified model (1.10)? Let us look at the evolution of the variance of the distribution in velocity variable, that is

$$\frac{d}{dt}\int_{\mathbb{R}} |v|^2 f(t,v)\,dv = -\int_{\mathbb{R}}\int_{\mathbb{R}} B(|v-w|)(v-w)^2 f(v)f(w)\,dv\,dw$$

by substituting in (1.9) and symetrizing. In case $\gamma = 0$, we can expand the square and use the conservation of zero mean velocity (1.8) to simplify the right-hand side. Therefore, denoting the variance of $f(t,\cdot)$ by $x(t)$, then it follows the ODE $x'(t) \leq -cx(t)$, and thus $x(t) \to 0$ as $t \to \infty$ exponentially fast for $\gamma = 0$. We conclude that the variance is decreasing and converging to 0 as $t \to \infty$. Let us assume that there is no concentration in finite time, as a consequence as $t \to \infty$ the probability densities $f(t,\cdot) \rightharpoonup \delta_0$ weakly-$*$ as measures as $t \to \infty$. Now the question is how this concentration in velocity happens for the solutions of (1.10), does it really happen in finite or infinite time and if so, can we understand the convergence towards concentration? Is there any typical profile? What is the long-time behavior for other values of $-1 \leq \gamma$?

1.2 Aggregation-Diffusion: McKean-Vlasov Equations.

Consider a confinement potential $V \in C^1$, and a particle that moves in this potential with a large friction such that we can neglect the inertia term. Thus a given particle X_t follow the ODE system $\frac{dX_t}{dt} = -\nabla V(X_t)$. Let us also assume that we perturb this motion stochastically by a Brownian noise added to the system of strength σ. Therefore, the SDE system followed by the particle is given by the Langevin equation

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\sigma}\,dB_t\,, \tag{1.11}$$

where B_t is the standard Brownian motion. Ito's formula implies that the law $\rho(t,\cdot)$ of the random variable X_t satisfies the Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \nabla\cdot(\rho\nabla V) + \sigma\Delta\rho, \tag{1.12}$$

that is a particular case of the general family of PDE (1.1) with general confinement potential V, zero interaction $W = 0$ and linear diffusion $P(\rho) = \sigma\rho$. One can easily observe that convexity properties of V will play an important role in the long time dynamics of this equations (1.11) or equivalently (1.12). In fact, let us take two realizations X_t and Y_t, $t \geq 0$, of the SDE (1.11), meaning two solutions of (1.11) with differential initial data but constructed with the same Brownian motion. This

means that the solutions X_t and Y_t are correlated for $t > 0$ even if we assume them initially independent. Since they are constructed from the same Brownian motion, even if separately both trajectories X_t and Y_t do not have good regularity, it is not difficult to deduce from stochastic analysis theory that the difference $\alpha_t = X_t - Y_t$ is C^1, and it satisfies

$$\frac{d\alpha_t}{dt} = -(\nabla V(X_t) - \nabla V(Y_t)), \qquad t \geq 0,$$

and therefore, we deduce

$$\frac{1}{2}\frac{d}{dt}|\alpha_t|^2 = -(\nabla V(X_t) - \nabla V(Y_t)) \cdot (X_t - Y_t), \qquad t \geq 0.$$

Therefore, if the potential V is uniformly convex, there exists $\lambda > 0$ such that $D^2V \geq \lambda I_d$, then $\frac{1}{2}\frac{d}{dt}|\alpha_t|^2 \leq -\lambda|\alpha_t|^2$, for all $t \geq 0$, and thus $|\alpha_t|^2 \leq |\alpha_t|^2 e^{-2\lambda t}$. Now, take the initial random variables X_0 and Y_0 with finite variance, i.e., $\mathbb{E}[|X_0|^2] < \infty$ and $\mathbb{E}[|Y_0|^2] < \infty$, we can compute the expectation of $|\alpha_t|^2$ as

$$\mathbb{E}[|X_t - Y_t|^2] \leq \mathbb{E}[|X_0 - Y_0|^2]e^{-2\lambda t} \leq 2\left(\mathbb{E}[|X_0|^2 + \mathbb{E}[|Y_0|^2]\right)e^{-2\lambda t}, \qquad t \geq 0.$$

Therefore, two solutions of the SDE converge towards each other exponentially fast in the above sense. It is easy to check by direct inspection that the normalized Gaussian

$$\rho_\infty(x) = \frac{1}{Z}e^{-V(x)/\sigma} \qquad \text{with } Z = \int_{\mathbb{R}^d} e^{-V(x)/\sigma}\,dx,$$

is a stationary state of (1.12). By taking Y_0 the random variable whose distribution is given by ρ_∞, we have shown that all solutions of the Langevin equation (1.11) converge in the sense above to the stationary state (1.11). The Gaussian measure ρ_∞ is usually referred as invariant measure in stochastic analysis. We will see how this convergence translates onto the convergence of solutions $\rho(t,\cdot)$ of the linear Fokker-Planck equation (1.12) towards ρ_∞ in a suitable sense.

Finally, we can also introduce a pairwise interaction potential W between particles and introduce a systems of N interacting particles perturbed by Brownian noise of the form

$$dX_t^i = -\frac{1}{N}\sum_{i\neq j}^{N} \nabla W(X_t^i - X_t^j)dt + \sqrt{2\sigma}\,dB_t^i, \tag{1.13}$$

where B_t^i, $i = 1,\dots,N$, are N independent Brownian motions. Now, it is more difficult to analyse the correlations between the particles and what is the PDE, if any, that gives the typical behavior of one of the particles as $N \to \infty$. The answer to this question is the so-called mean-field limit that allows to identifiy the limiting PDE that satisfies the law of a particle in the large number of particles limit $N \to \infty$. Notice that the interaction potential has the factor $\frac{1}{N}$ in front in the SDE system (3.9), which is crucial to identify a sort of mean-field potential created by the particle en-

semble. It is proven that under certain assumptions on the interaction potential the limiting PDE is given by the McKean-Vlasov equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot [\rho(\nabla W * \rho)] + \sigma \Delta \rho. \tag{1.14}$$

Convexity properties of the interaction potential will give information on the long time asymptotics of both the SDE system (3.9) and the McKean-Vlasov equation (1.14). Let us finally remark that McKean-Vlasov equation (1.14) are ubiquitous in applications in the sciences from sinchronysation to swarming models for collective behavior in mathematical biology, to opinion formation in social sciences or to self-assembly alloys and granular flows in material science, and lately they have found a renewed interest in data science.

1.3 Nonlinear Diffusions.

The most well-known cases of nonlinear diffusions are the homogeneus nonlinearities, $P(\rho) = \rho^m$ with $m > 0$. The flow of gas in an d-dimensional porous medium is described by Darcy's law, pressure proportional to the density of the gas, leading to

$$\frac{\partial u}{\partial t} = \Delta u^m, \qquad (x \in \mathbb{R}^d, t > 0), \tag{1.15}$$

The function u represents the density of the gas in the porous medium and $m > 1$ is a physical constant. This equation can be thought as a nonlinear heat equation in which the thermal conductivity is $m\rho^{m-1}$, and therefore directly proportional to the density for $m > 1$. The porous medium equation degenerates in vacuum, i.e. for $\rho = 0$, leading to the interesting phenomena of free boundaries and finite speed of propagation due to slow diffusion for small values of the density. We refer to [20] for a comprehensive treatment of this problem. The equation for $0 < m < 1$ receives the name of fast diffusion equation since the heat conduction is now inversely proportional to the density, and thus very fast diffusion happens for small values of the density u.

Equation (1.15) has some important explicit solutions that led to the advance of functional analysis and techniques for understading long-time behavior of nonlinear diffusion equations since the 1970's. There are self-similar solutions generalizing the role of the heat kernel for the heat equation. Let us remind that the solution to the Cauchy problem for the heat equation

$$\frac{\partial u}{\partial t} = \sigma \Delta u, \qquad (x \in \mathbb{R}^d, t > 0), \tag{1.16}$$

with initial data a probability measure $\rho_0 = \mu$ can be obtained by the Poisson's formula $\rho_t = K(t, \cdot) * \mu$ donde $K(t,x)$ is the heat kernel given by

$$K(t,x) = (4\pi\sigma t)^{-\frac{d}{2}} \exp\left(-\frac{x^2}{4\sigma t}\right). \tag{1.17}$$

The heat kernel can be understood as the solution with initial data given by a Dirac-delta at the origin, $\rho_0 = \delta_0$, and it is a self-similar solution of the heat equation.

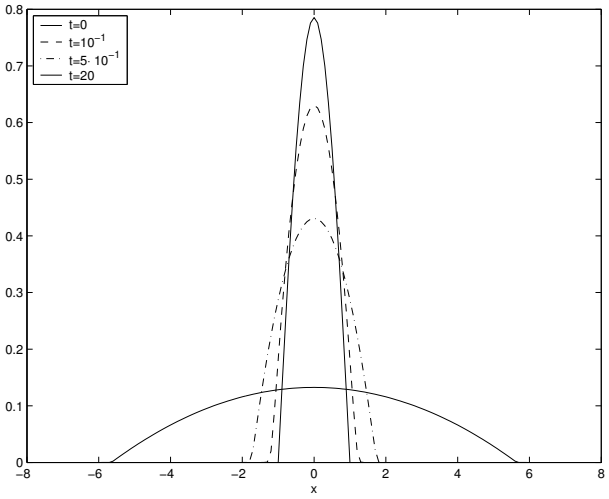

Fig. 1.1 Evolution of the porous medium equation (1.15) with initial data $\frac{\pi}{4}\cos(\frac{\pi}{2}x)$ for $m = 2$.

Generalizations of the heat kernel solution for (1.16) can be obtained for (1.15) by finding the right self-similar change of variables. In fact, if one seeks solutions to (1.15) with the mass-preserving scaling of the form $t^{-d/\alpha}F(xt^{-1/\alpha})$, one can check that the self-similar profile F satisfies the nonlinear equation $\operatorname{div}(xF + \alpha\nabla F^m) = 0$ by choosing $\alpha = d(m-1)+2$ (this fact is an exercise). This equation can be analysed to find that a solution is given by

$$F(\xi) = \left(C - \kappa|\xi|^2\right)_+^{\frac{1}{m-1}}$$

with $\kappa = \frac{m-1}{2m\alpha}$ and $C > 0$ is determined uniquely by imposing unit mass for $m > \frac{d-2}{d}$ due to integrability at ∞, see [20] for details. We will denote these self-similar solutions by $\mathscr{B}(t,x)$ as they were first discovered by G.I. Barenblatt in the 1960's.

We first notice that these solutions are only weak solutions in the distributional sense for (1.15) in the porous medium range $m > 1$, while they are classical solutions in the fast diffusion range $0 < m < 1$. In fact, the Barenblatt solutions $\mathscr{B}(t,x)$ are compactly supported on a ball for $m > 1$, and they only enjoy certain Hölder-regularity for $m > 2$ at the edge of the support, Lipschitz for $m = 2$ at this edge, and they become more and more regular as $m \to 1^+$ at the boundary of the support. In the fast diffusion range $0 < m < 1$, they become instantaneously positive and C^∞ everywhere as for the heat kernel, but with algebraic tails as $|x| \to \infty$ instead. Moreover

as in the case of the heat equation, the initial data of these solutions is a Dirac-delta at the origin in the sense that $\mathscr{B}(t,\cdot) \rightharpoonup \delta_0$ weakly-$*$ as measures as $t \to 0^+$ (this fact is an exercise).

We illustrate in one dimension, Figure 1.1, and in two dimensions, Figure 1.2, the evolution for the porous medium equation with exponent $m = 2$ with certain initial data, and we can observe that solutions become more and more similar to the Barenblatt profiles modulo translation for large times.

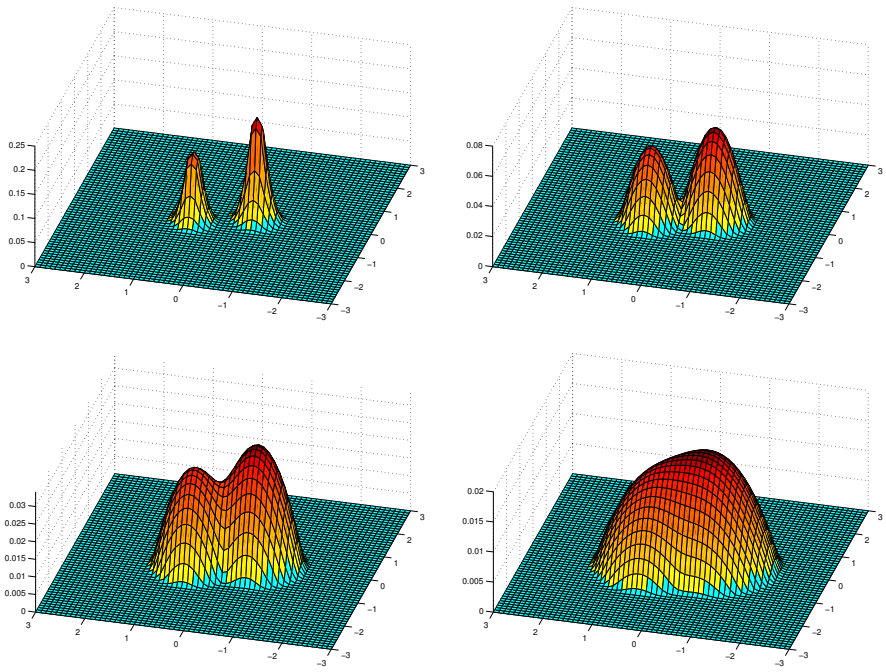

Fig. 1.2 Evolution for the porous medium equation (1.15) for $m = 2$.

Instead of working with (1.15) directly, we will analyse the asymptotic decay towards its equilibrium state of solutions to the (nonlinear) Fokker-Planck type equations

$$\frac{\partial \rho}{\partial t} = \operatorname{div}(x\rho + \nabla \rho^m), \qquad (x \in \mathbb{R}^d, t > 0), \tag{1.18}$$

that corresponds to the choice of $P(\rho) = \rho^m$, with $m > 0$, and the confinement potential $V(x) = |x|^2/2$ in the general family of PDE (1.1).

The connection between the porous medium, the heat, and the fast diffusion equations (1.15) with nonlinear Fokker-Planck equations (1.18) becomes apparent after the following fundamental observation: there exists a time dependent scaling which transforms (1.18) into the porous medium, the heat, and the fast diffusion equations (1.15) while keeping the same initial data. Actually, if u is a solution of (1.15) then

$$\rho(t,x) = e^{dt} u\left(\frac{e^{\alpha t}-1}{\alpha}, e^t x\right) \tag{1.19}$$

is a solution of (1.18) and vice versa, if ρ is a solution of (1.18), then

$$u(t,x) = (1+\alpha t)^{-d/\alpha} \rho\left(\frac{1}{\alpha}\log\left(1+\alpha t\right), \left(1+\alpha t\right)^{-1/\alpha} x\right) \tag{1.20}$$

is a solution of (1.15) (these facts are an exercise). We finally remark that a stationary solution of (1.18) is given by the Barenblatt type formula

$$\rho_\infty(x) = \left(C - \frac{m-1}{2m}|x|^2\right)_+^{\frac{1}{m-1}} \tag{1.21}$$

for a $C>0$ such that ρ_∞ has unit mass. In fact, one can check that this is a stationary solution of (1.18) by noticing that the flux $x\rho + \nabla\rho^m$ is zero,

$$x\rho_\infty + \nabla\rho_\infty^m = \rho_\infty\left(\frac{m}{m-1}\nabla\rho_\infty^{m-1} + x\right) = 0.$$

Notice that the last computation makes sense since ρ_∞^{m-1} is a Lipschitz function.

We point out that $\rho_\infty(x)$ corresponds to $\mathscr{B}(t+\frac{1}{\alpha},x)$ through the change of variables (1.19)–(1.20). As a conclusion, if we are able to derive any property about the asymptotic behavior of $\rho(t,x)$ towards $\rho_\infty(x)$ we can translate it into a result about the asymptotic behavior of $u(t,x)$ towards the Barenblatt profile $\mathscr{B}(t,x)$. More precisely, showing the exponential decay of the solutions to (1.18) towards the stationary state ρ_∞ translates into algebraic decay towards self-similar profiles of the porous medium, the heat, and the fast diffusion equations (1.15) via the change of variables (1.19)–(1.20).

1.4 Nonlinear Aggregation-Diffusion Equations: The Patlak-Keller-Segel model.

The Patlak-Keller-Segel (PKS) equation is widely used in mathematical biology to model the collective motion of cells which are attracted by a self-emitted chemical substance, being the slime mold amoebae *Dictyostelium discoideum* a prototype organism for this behaviour. Moreover, the PKS equation has become a paradigmatic mathematical problem since it shows a concentration-collapse dichotomy: for masses larger than a critical value solutions aggregate their mass, as Dirac-deltas, in finite time while solutions exist globally and disperse collapsing down to zero below this critical mass threshold.

Historically, the first mathematical models in chemotaxis were introduced in 1953 by C. S. Patlak and E. F. Keller and L. A. Segel in 1970 in two dimensions since they were interested in the chemotactic movement of cells in Petri dishes. The

basic model in any dimension reads as

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \Delta \rho - \chi \nabla \cdot [\rho \nabla c] & t > 0\,,\ x \in \mathbb{R}^d\,, \\ c(t,x) = -\dfrac{1}{d\pi} \displaystyle\int_{\mathbb{R}^d} \log|x-y|\, \rho(t,y)\, dy\,, & t > 0\,,\ x \in \mathbb{R}^d\,, \end{cases} \tag{1.22}$$

Here $(t,x) \mapsto \rho(t,x)$ represents the normalized cell density, and $(t,x) \mapsto c(t,x)$ is the concentration of chemo-attractant. The constant $\chi > 0$ is the *sensitivity* of the bacteria to the chemo-attractant. Mathematically, it measures the attractive interaction force between cells, and hence, the strength of the non-linear coupling. Note that (1.22) corresponds to the choice $P(\rho) = \rho$ and $W(x) = -\frac{1}{d\pi} \log|x|$ in the general family of PDE (1.1).

We first remind that a notion of weak solution ρ in the space $C^0\big([0,T); L^1_+(\mathbb{R}^d)\big)$, with fixed $T > 0$, using the symmetry in x, y for the concentration gradient, can be introduced to handle even measure solutions. We shall say that ρ is a weak solution to the system (1.22) if for all test functions $\zeta \in C^2_b(\mathbb{R}^d)$,

$$\begin{aligned} \frac{d}{dt} \int_{\mathbb{R}^d} \zeta(x)\, \rho(t,x)\, dx &= \int_{\mathbb{R}^d} \Delta \zeta(x)\, \rho(t,x)\, dx \\ &\quad - \frac{\chi}{2\,d\,\pi} \iint_{\mathbb{R}^d \times \mathbb{R}^d} [\nabla \zeta(x) - \nabla \zeta(y)] \cdot \frac{x-y}{|x-y|^2}\, \rho(t,s)\, \rho(t,y)\, dx\, dy \end{aligned} \tag{1.23}$$

in the distributional sense in $(0,T)$. Here, the Banach space $C^2_b(\mathbb{R}^d)$ is defined as the set of C^2-functions with bounded second derivatives. Notice that the singularity due to the derivative of the log-kernel dissappears by symmetrization of the term using the mean value theorem. Any weak solution in the previous sense with initial data a probability density function satisfies mass and center of mass conservations, i. e.,

$$\int_{\mathbb{R}^d} \rho(t,x)\, dx = \int_{\mathbb{R}^d} \rho_0(t,x)\, dx = 1 \quad \text{and} \quad \int_{\mathbb{R}^d} x\, \rho(t,x)\, dx = \int_{\mathbb{R}^d} x\, \rho_0(t,x)\, dx = 0,$$

the latter being assumed without loss of generality by translational invariance. In order to check the behavior of the system, we can check the evolution of the variance of the distribution as done in the first example of this section. By taking $\zeta(x) = |x|^2$ as test function in (1.23), we obtain

$$\frac{d}{dt} \int_{\mathbb{R}^d} |x|^2\, \rho(t,x)\, dx = 2d - \frac{\chi}{d\,\pi}\,.$$

Therefore, if $\chi > 2d^2\pi$, the variance of the distribution $\rho(t,x)$ becomes zero in finite time. This means that in finite time, there should be a concentration as a Dirac-delta at the origin contradicting the existence of a weak solution in the sense of (1.23) at that time.

This intuition can be made rigorous at certain extent. The Cauchy problem for the PKS equation (1.22) presents the following dichotomy: either L^1-solutions blow-up

in finite time for the super-critical case $\chi > 2d^2\pi$ or rather solutions exist globally in time and spread in space decaying towards a stationary solution in rescaled variables as $t \to \infty$ in the sub-critical case $\chi < 2d^2\pi$. The critical case $\chi = 2d^2\pi$ is also fairly well understood leading to infinite time blow-up or convergence to stationary states depending on the initial data. We refer to the recent survey [9] and the references therein for further details and even more general cases with nonlinear diffusions and general interaction kernels. This example show us that concentration and diffusion phenomena can coexist for the same type of equations depending on just one parameter.

1.5 Nonlinear Aggregation-Diffusion Equations: Phase Transitions in collective behavior models.

The final example arises in collective behavior models for animal swarming. We refer to the survey [10] for details about the modelling and the mean-field limit from interacting particle systems of 2nd order leading to the following localized Cucker-Smale model for aligment for self-propelled particles with noise. Here, f represents the distribution in both space x and velocity v at time t of individuals, and the model features a Cucker-Smale term which aligns the velocity of points nearby in space, a term adding noise in the velocity, and a friction term which relaxes velocities back to norm one leading to

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \nabla_v \cdot \left(\beta(|v|^2 - 1) v f + (v - u_f) f + \sigma \nabla_v f \right) ,$$

where

$$u_f(t,x) = \frac{\int K(x,y) v f(t,y,v)\, dv\, dy}{\int K(x,y) f(t,y,v)\, dv\, dy}.$$

Here $K(x,y)$ is a suitably defined compactly supported localization kernel and β and σ are respectively the self-propulsion force and noise intensities. If we first look for the behavior in the spatially homogeneous case, the model reduces to

$$\partial_t f = \nabla_v \cdot \left(\beta(|v|^2 - 1) v f + (v - u_f) f + \sigma \nabla_v f \right) . \tag{1.24}$$

where

$$u_f(t) = \int_{\mathbb{R}^d} v f(t,v)\, dv, \tag{1.25}$$

and where $f = f(t,v)$ is the velocity distribution at time t. This alignment model can be again recast as the general PDE (1.1) by the changing the notation from $f(t,v)$ to $\rho(t,x)$ with the choices

$$P(\rho) = \sigma\rho, \quad V(x) = \beta \left(\tfrac{|x|^4}{4} - \tfrac{|x|^2}{2} \right) \quad \text{and} \quad W(x) = \tfrac{|x|^2}{2} .$$

The interesting phenomena happening in this particular model is that as soon as one of the potentials, in this case the confinement potential, is not convex, complicated dynamics can happen. In fact, there is a phase transition between unpolarized and polarized motion as the noise intensity σ is varied, for a specific range of the values of β. More precislely, one can analytically prove that, for large noise σ, there is only one isotropic stationary solution, while for small σ, there is an additional infinite family of stationary states parameterized by a unit vector on the sphere, referred to as the polarized equilibria. Moreover the change from one single isotropic stationary state to infintely many steady states happens at a precise threshold critical value of σ_c, depending on β, that is known in dimensions 1 and 2, see Fig. 1.3. These questions are nowadays of current interest in research.

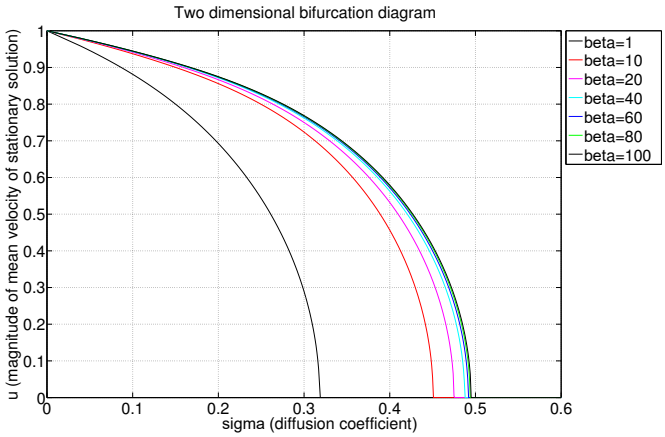

Fig. 1.3 Mean speed u of the stationary state as a function of the diffusion parameter σ for several values of the self-propulsion strength β. There is a continuous bifurcation critical diffusion σ_c from the existence of polarized to unpolarized stationary states.

Chapter 2
Optimal Transportation: The metric side

In this chapter, we will do a short primer on the classical optimal transport and their associated transport distances. Let us start by introducing quickly the basic notation of the objects we are dealing with, probability mesures.

2.1 Functional Analysis tools: measures and weak convergence.

Let us consider the space of continuous functions with zero limit at infinity $C_0(\mathbb{R}^d)$, i.e., $f \in C_0(\mathbb{R}^d)$ if it is continuous and for all $\varepsilon > 0$, there exists $R > 0$ such that $|f(x)| \leq \varepsilon$ for $|x| \geq R$. $C_0(\mathbb{R}^d)$ is a separable Banach space endowed with the uniform norm. We recall a basic notion in measure theory

Definition 2.1. A finite signed measure μ on $\mathbb{R}^d$ is a map that assigns to every Borel subset $A \subset \mathbb{R}^d$ a value $\mu(A) \in \mathbb{R}$ such that

$$\mu\left(\cup_{i\geq 1} A_i\right) = \sum_{i\geq 1} \mu\left(A_i\right) \quad \text{and} \quad \sum_{i\geq 1} |\mu\left(A_i\right)| < \infty$$

hold for every countable disjoint union $A_i \cap A_j = \emptyset$, $i \neq j$. The set of all finite signed measures on $\mathbb{R}^d$ will be denoted by $\mathscr{M}(\mathbb{R}^d)$. It is a Banach space endowed with the norm

$$\|\mu\| = \sup\left\{ \sum_{i\geq 1} |\mu\left(A_i\right)| \, : \, \mathbb{R}^d = \cup_{i\geq 1} A_i \quad \text{with } A_i \cap A_j = \emptyset,\, i \neq j \right\}.$$

Riesz's representation theorem provides a very useful characterization of the set of finite signed measures, every element of the dual Banach space $\mathscr{X}'$ of $\mathscr{X} = C_0(\mathbb{R}^d)$ can be represented in a unique way by a finite signed measure $\mu \in \mathscr{M}(\mathbb{R}^d)$. The weak-$*$ convergence on finite signed measures is then defined based on the dual pairing $(C_0(\mathbb{R}^d), \mathscr{M}(\mathbb{R}^d))$ and its representation

$$<\mu,\varphi>=\int_{\mathbb{R}^d}\varphi(x)\,d\mu(x).$$

We say that the sequence of measures μ_n converges weakly-$*$ to μ if and only if $<\mu_n,\varphi>\to<\mu,\varphi>$ for all $\varphi\in C_0(\mathbb{R}^d)$. This will be denoted by $\mu_n\rightharpoonup\mu$ weakly-$*$. In short, the dual space to $C_0(\mathbb{R}^d)$ is by definition the set of locally finite signed Radon measures in $\mathbb{R}^d$. The set of probability measures $\mathscr{P}(\mathbb{R}^d)$ is defined as the subset of nonnegative finite signed measures such that $\mu(\mathbb{R}^d)=1$.

Let us denote by $\mathscr{L}$ the Lebesgue measure on $\mathbb{R}^d$. When a probability measure $\mu\in\mathscr{P}(\mathbb{R}^d)$ is absolutely continous with respect to the Lebesgue measure, that is it has at least the same zero measures sets, denoted by $\mu\ll\mathscr{L}$, then the measure μ has a density $\rho\in L^1_+(\mathbb{R}^d)$, meaning that by the Radon-Nikodym theorem, it can be represented by the density ρ, i. e.

$$<\mu,\varphi>=\int_{\mathbb{R}^d}\varphi(x)\,d\mu(x)=\int_{\mathbb{R}^d}\varphi(x)\rho(x)\,dx.$$

We will use in these set of notes the notation of measure and its associated density indistinctively unless there is confusion. To finish these measure theory preliminaries, let us introduce another notion of convergence by duality for probability measures. We say that the sequence of measures μ_n narrow or weakly converges to μ if and only if $<\mu_n,\varphi>\to<\mu,\varphi>$ for all $\varphi\in C_b(\mathbb{R}^d)$ or in other words that the measures convergences in the duality with $C_b(\mathbb{R}^d)$. This will also be denoted abusing the notation by $\mu_n\rightharpoonup\mu$. We point out that the dual of $C_b(\mathbb{R}^d)$ can also be characterized in terms of certain set of measures larger than $\mathscr{M}(\mathbb{R}^d)$ but it is a weird space, see [21, Section 1.3] for further details.

Finally, let us remind few Functional Analysis results on the compactness of subsets of measures. Given a dual pair of Banach spaces $(\mathscr{X},\mathscr{X}')$ and its associated duality $<\cdot,\cdot>$, Banach-Alaoglu's theorem asserts that any bounded set in $\mathscr{X}'$ is precompact in the weak-$*$ topology. In practice, this implies that any sequence of probability measures has a weakly-$*$ subsequence towards a nonnegative measure not necessarily being a probability measure. In order for the weak-$*$ limit to be a probability measure, we need an additional property.

Definition 2.2. A sequence μ_n in $\mathscr{P}(\mathbb{R}^d)$ is said to be tight if for every $\varepsilon>0$, there exists $R>0$ such that $\mu_n(\mathbb{R}^d\setminus B_R)\le\varepsilon$ for every n, where B_R is the euclidean ball of radius R centered at the origin.

We refer to [6] for futher details on duality pairings and weak topologies.

Prokhorov's Theorem gives a characterization of weakly-$*$ precompact subsets of probability measures.

Theorem 2.1. *(Prokhorov) Every tight sequence μ_n in $\mathscr{P}(\mathbb{R}^d)$ has a weakly or narrowly convergent subsequence to a limiting probability measure. Conversely, every weakly converging sequence of probability measures $\mu_n\rightharpoonup\mu$ is tight.*

In order to explain better the classical optimal transportation problem, we need some further definitions.

Definition 2.3. Let μ and ν be in $\mathscr{P}(\mathbb{R}^d)$ the space of probability measure in $\mathbb{R}^d$, and T be a measurable map $\mathbb{R}^d \to \mathbb{R}^d$. We say that T *transports* μ onto ν, ν is the *push-forward* or *the image measure* of μ through T, and we denote it by $\nu = T\#\mu$, if for any measurable set $B \subset \mathbb{R}^d$, $\nu(B) = \mu(T^{-1}(B))$.

In fact, the previous definition of pushforward is equivalent to

$$\int_{\mathbb{R}^d} (\zeta \circ T)(x)\, d\mu(x) = \int_{\mathbb{R}^d} \zeta(y)\, d\nu(y) \quad \forall \zeta \in \mathscr{C}_b(\mathbb{R}^d)\,. \tag{2.1}$$

Actually, the change of variables formula (2.1) is true for all $\zeta \in L^1(\mathbb{R}^d)$. We leave this as a warm-up in integration, measure theory and dominated/monotone convergence theorems (this fact is an exercise). The image measure through a map T can also be directly connected to basic probability theory. In fact, a random variable X with law μ is by definition a measurable map $X : (\mathscr{S}, \mathscr{A}, P) \longrightarrow \mathscr{L}$ from a probability space of reference $(\mathscr{S}, \mathscr{A}, P)$ onto the Lebesgue space $\mathscr{L}$ such that the image measures through X of P is μ, i.e. $X\#P = \mu$.

2.2 A brief introduction to optimal transport

Let us first introduce intuitively the optimal transportation problem. Let us assume that the probability measure μ represents the density of frozen fish-and-chips suppliers in the United Kingdom while the probability measure ν represents the density of pubs (it is a good approximation to assume that at least the measure ν has parts which are absolutely continuous with respect to Lebesgue and atomic parts, think about the London area or the Costwolds while the first measure μ might be concentrated in coastal areas in the Southwest of England and Scotland). Assume that the market is in equilibrium, meaning supply=demand, so all the produced frozen fish-and-chips are consumed by the pubs. The question we want to solve is how to find a way of transporting all the product from the suppliers at each specified location x to the the consumers at locations y optimally. The optimality here has to be specified and it should include an estimate of the cost needed for the transportation. The union of frozen fish-and-chips suppliers is overseeing the whole operation of transportation, thus they would like to know how to transport all the frozen fish-and-chips from the suppliers to the consumers minimizing the overall cost of this task. Let us represent by $c(x,y)$ the cost of sending a unit of the product from supplier location $x \in \mathbb{R}^d$ to consumer location $y \in \mathbb{R}^d$, i.e., we define a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0,\infty)$.

The transportation problem was mathematically set up for the first time by Gaspard Monge, a French mathematician and engineer in the late 1700's in his essay "Mémoire sur la théorie des déblais et remblais" in 1781. His transportation problem was very much related to French army's operations and less to fish-and-chips distribution though. He posed the problem in the following way, from all possible ways of transporting the goods from location x to location y, can we find the optimal

one minimizing the total incurred cost? More precisely and in modern mathematical terms, given two probability measures μ and ν, can we find an optimal map T transporting μ onto ν, $\nu = T\#\mu$, minimizing the total cost given by

$$\int_{\mathbb{R}^d} c(x,T(x))\,d\mu(x)\,?$$

This classical problem from Calculus of Variation, sketched in Figure 2.1, is the

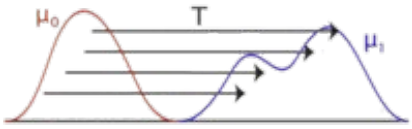

Fig. 2.1 Monge trasnportation problem between two probability measures μ_0 and μ_1. Figure taken from Wikipedia.

so-called Monge transportation problem, that is to find, if possible, the solution to the following minimization problem:

$$I_M := \inf_T \left\{ \int_{\mathbb{R}^d} c(x,T(x)\,d\mu(x) \,:\, \nu = T\#\mu \right\}.$$

He posed this question with the cost given by the distance between the locations $c(x,y) = |x-y|$. It is very easy to see that this problem does not have a solution for general probability measures. In fact, the set of maps pushing one probability measure μ onto ν might be even empty making the classical Monge problem trivially impossible. Take $\mu = \delta_{x_0}$ and $\nu = \frac{1}{2}\delta_{x_0} + \frac{1}{2}\delta_{x_1}$ with $x_0 \neq x_1$ where δ_{x_0} is the Dirac delta measure at x_0. Then $\nu(\{x_1\}) = \frac{1}{2}$ but either $\mu(T^{-1}(\{x_1\})) = 1$ or $\mu(T^{-1}(\{x_1\})) = 0$ depending if $T(x_0) = x_1$ or not. Thus, there is no map pushing forward μ onto ν.

The issue here is that in the classical Monge transportation problem choosing transportation maps is not a good idea. It is a better idea "to split the mass", this is even more advantageous economically. In fact, Leonyd Kantorovich in 1942 realized that a better way to pose the transportation problems lies in the basic idea that for each producer it will be generically more economic to split its production among several consumers that sending all its production to a unique location. He introduced the concept of transportation or transference plan, that is a probability measure $\Pi(x,y)$ on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. The basic meaning of $\Pi(x,y)$ is the number of units of production at location x sent to location y while using fully the total number of produced units by the supplier located at x and fulfilling the total number of units demanded by the consumer located at y. The mathematical statement of the last sentence is translated in the fact that the marginal measures of Π must be μ and ν respectively. Let us denote by $\Gamma(\mu,\nu)$ the set of

all transference plans, that is, the set of joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, i.e.,

$$\iint_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(x)\, d\Pi(x,y) = \int_{\mathbb{R}^d} \varphi(x)\, d\mu(x)$$

and

$$\iint_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(y)\, d\Pi(x,y) = \int_{\mathbb{R}^d} \varphi(y)\, d\nu(y)$$

for all $\varphi \in C_b(\mathbb{R}^d)$. Allowing splitting of the mass, Kantorovich proposed a relaxed variational problem that avoids the problems of the Monge transportation problem: find among all possible transference plans $\Pi \in \Gamma(\mu,\nu)$ an optimal one minimizing the total cost

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi(x,y)\,.$$

More precisely, the relaxed Monge-Kantorovich transportation problem consists in finding, if possible, the solutions to the minimization problem:

$$I_K := \inf_{\Pi \in \Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi(x,y) \right\}.$$

Let us remark that the product measure $\mu \times \nu$ always belongs to $\Gamma(\mu,\nu)$, and thus $\Gamma(\mu,\nu) \neq \emptyset$. Proving that the infimum in the Kantorovich formulation of the transportation problem is achieved, and thus there is a minimum, is the main objective of the next section. Kantorovich received the Nobel Prize in Economics in 1975 "for his contributions to the theory of optimum allocation of resources."

In fact, let us check that the Kantorovich formulation is really a relaxed variational problem of the Monge transportation problem. Given any measurable map T transporting μ onto ν, $\nu = T\#\mu$, let us define the transference plan $\Pi_T = (1_{\mathbb{R}^d} \times T)\#\mu$ as the element in $\mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(x,y)\, d\Pi_T(x,y) = \int_{\mathbb{R}^d} \psi(x,T(x))\, d\mu(x)$$

for all $\psi \in C_b(\mathbb{R}^d \times \mathbb{R}^d)$. It is easy to check that $\Pi_T \in \Gamma(\mu,\nu)$, and thus $I_K \leq I_M$. Conversely, if there is an optimal transference plan of the form Π_{T_o} for certain T_o for the Kantorovich problem, then T_o is an optimal map for the Monge problem. Sufficient conditions for this to happen will be discussed in Chapter 4. For the time being, let us just say that for the quadratic cost $c(x,y) = |x-y|^2$ and whenever $\nu \ll \mathscr{L}$, then there is an optimal map achieving the infimum in the Monge and the Kantorovich transportation problems.

The beauty and strength of the Monge and Kantorovich problems is that they allowed for natural interpolation between probability measures. Assume that an optimal map T_o for the Monge problem exists between μ and ν. We can define the curves of measures

$$\mu_t = (1-t)\mu + tT_o\#\mu = ((1-t)1_{\mathbb{R}^d} + tT_o)\#\mu\,,$$

for $0 \le t \le 1$. Then, we can use μ_t as a morphing between the two probability measures, see Figure 2.2 for an example of this construction between two characteristic sets suitably normalized.

Interpolation measure between the Pac-Man and the Ghost probability measures

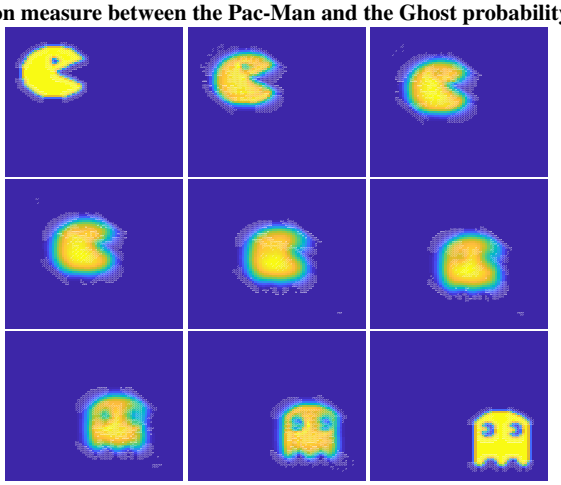

Fig. 2.2 Computation of a interpolation measure by the Monge-Kantorovich problem with quadratic cost between Pac-Man and the Ghost characteristic sets suitably normalized.

We refer to the link

`https://figshare.com/projects/Primal_dual_methods_for_Wasserstein_gradient_flows/59474`

to see the video for this simulation.

2.3 The Kantorovich Formulation and Duality. The Brenier Theorem.

Classical calculus of variations deals with the problem of finding the extrema of functionals $I : X \mapsto \mathbb{R} \cup \{+\infty\}$ defined on a given metric space X of functions and possibly considered over a nonempty subset $K \subset X$. The main goal is to find minimizers of such functionals, that is, functions $f \in K$ such that $I[f] \le I[g]$ for all $g \in K$. Even in situations where variations of the possible minimizer lead to necessary conditions for f to be satisfied, the so-called Euler-Lagrange conditions, it is important to know apriori if minimizers exist for the functional I. The first necessary assumption on I is that the functional I must be bounded below, if not there is nothing to be proven, this means that

$$I_* := \inf_f \{I[f] : f \in K \subset X\} > -\infty.$$

This shows the existence of a minimizing sequence, that is, a sequence $f_n \in K$ such that $I[f_n] \to I_*$. Notice that it is not even clear that if there is $f \in X$ achieving the infimum, then f does belong to K. The direct method of the calculus of variations is an adapted version for general metric spaces of the classical Weierstrass criterion for the existence of extremal points of continuous functions in compact sets in finite dimensions. It can be summarized as "compactness + semi-continuity" leads to existence of nontrivial minimization problems.

Definition 2.4. A functional $I : X \mapsto \mathbb{R} \cup \{+\infty\}$ on a metric space X is said to be lower semi-continuous (l.s.c), if for every sequence $f_n \in X$ such that $f_n \to f$, we have $I[f] \leq \liminf_n I[f_n]$.

Theorem 2.2 (Direct Method of Calculus of Variations). *A lower semi-continuous functional $I : X \mapsto \mathbb{R} \cup \{+\infty\}$ defined on a metric space X achieves its infimum in any compact subset $K \subset X$ where I is bounded from below, that is, there exists $f_o \in K$ such that $I[f_o] = \min\{I[f] : f \in K\}$.*

Proof. Since I is bounded below in K, there exists a minimizing sequence in K, that is, a sequence $f_n \in K$ such that $I[f_n] \to I_*$ with $I_* = \inf_f \{I[f] : f \in K \subset X\} > -\infty$. Since K is a compact subset of X, then f_n has a convergent subsequence to a limiting function $f_o \in K$. Without loss of generality, we can assume that the minimizing sequence is convergent to $f_o \in K$ with $I[f_o] \geq I_*$ by its definition. By virtue of lower semi-continuity we deduce that $I[f_o] \leq \liminf_n I[f_n] = I_*$, and therefore the infimum of the functional in K is achieved at f_o, $I_* = I[f_o]$, and actually the infimum is a minimum. □

A direct application of the previous theorem leads to the existence of optimal transference plans.

Theorem 2.3 (Existence of optimal transference plans). *Assume that the cost function $c : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0,\infty)$ is lower semi-continuous. Given two probability measures μ and ν, then there exists an optimal transference plan, that is, there exists a $\Pi_o \in \Gamma(\mu,\nu)$ achieving the infimum in the Kantorovich formulation of optimal transport*

$$I_* := \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi_o(x,y) = \min_{\Pi \in \Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi(x,y) \right\}.$$

Proof. Since μ and ν are probability measures, then for any $\varepsilon > 0$ there exists $R > 0$ such that

$$\mu(\mathbb{R}^d \setminus B_R) \leq \varepsilon \qquad \text{and} \qquad \nu(\mathbb{R}^d \setminus B_R) \leq \varepsilon\,,$$

and thus

$$\begin{aligned} \Pi((\mathbb{R}^d \times \mathbb{R}^d) \setminus (B_R \times B_R)) &\leq \Pi(\mathbb{R}^d \times (\mathbb{R}^d \setminus B_R)) + \Pi((\mathbb{R}^d \setminus B_R) \times \mathbb{R}^d) \\ &= \mu(\mathbb{R}^d \setminus B_R) + \nu(\mathbb{R}^d \setminus B_R) \leq 2\varepsilon\,, \end{aligned}$$

for all $\Pi \in \Gamma(\mu,\nu)$. Hence the set of transference plans $\Gamma(\mu,\nu)$ is tight in $\mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$. By Prokhorov's theorem, the closure of the set of transference plans in the weak topology is compact. By definition of the convergence in the weak or narrow topology, it is easy to check that the set of transference plans $\Gamma(\mu,\nu)$ is closed. Therefore, we consider the functional

$$I[\Pi] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi(x,y)$$

defined on the compact set of all transference plans $\Pi \in \Gamma(\mu,\nu)$, w.r.t. the weak topology of measures. On the other hand, since c is a l.s.c. bounded from below function in $\mathbb{R}^d \times \mathbb{R}^d$, it can be approximated by an increasing sequence of continuous and bounded functions c_n in $\mathbb{R}^d \times \mathbb{R}^d$ (this statement is an exercise). Monotone convergence theorem implies that

$$I_n[\Pi] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c_n(x,y)\, d\Pi(x,y) \nearrow I[\Pi]$$

for all $\Pi \in \Gamma(\mu,\nu)$. Notice that since c_n is continuous and bounded, the functionals I_n are trivially continuous in the weak topology. Moreover, $I[\Pi] = \sup_n I_n[\Pi]$, and thus I is l.s.c in the weak topology as a supremum of continuous functionals in the weak topology (this statement is an exercise). We now have all the ingredients to repeat the same argument of the direct method of the calculus of Variations Theorem 2.2, but with the weak topology instead of the metric topology to obtain the announced result. □

Notice that if we use that the set of probability measures $\mathscr{P}(\mathbb{R}^d)$ endowed with the weak topology is metrizable, the previous result can be considered a direct application of Theorem 2.2.

The previous theorem gives a rough answer to the existence of optimal transference plans but much more can be obtained by realizing that the Kantorovich reformulation of the transportation problem is a linear optimization problem under convex constraints, given by linear equalities or inequalities. Therefore, this is the place in which convex analysis and duality in optimization plays an important role. Kantorovich realized this and he introduced duality together with an economic interpretation of the dual variables as shadow prices. The idea is to include the constraints on the marginals as Lagrange multipliers rewriting the minimization problem with constraints as an inf − sup optimization problem without constraints. More precisely, let us express the constraint $\Pi \in \Gamma(\mu,\nu)$ as follows: if $\Pi \in \mathscr{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$, then take two functions $\varphi, \psi \in C_b(\mathbb{R}^d)$, acting as Lagrange multipliers, to have

$$R = \begin{cases} 0 & \text{if } \Pi \in \Gamma(\mu,\nu) \\ +\infty & \text{otherwise} \end{cases},$$

with R defined by

$$R := \sup_{\varphi,\psi\in C_b(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \varphi(x)\, d\mu(x) + \int_{\mathbb{R}^d} \psi(y)\, d\nu(y) - \int_{\mathbb{R}^d\times\mathbb{R}^d} (\varphi(x)+\psi(y))\, d\Pi(x,y) \right\}.$$

Hence, we can remove the constraint on Π if we add the quantity R to $I[\Pi]$, since if the constraint is satisfied we are not adding anything and if not the infinity values will be avoided by the minimization. Therefore the Kantorovich problem is equivalent to the following inf $-$ sup problem: finding $\Pi \in \mathscr{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$ such that

$$\inf_{\Pi}\sup_{\varphi,\psi} \left\{ \int_{\mathbb{R}^d} \varphi(x)\, d\mu(x) + \int_{\mathbb{R}^d} \psi(y)\, d\nu(y) + \int_{\mathbb{R}^d\times\mathbb{R}^d} (c(x,y) - \varphi(x) - \psi(y))\, d\Pi(x,y) \right\}.$$

Notice that we have also relaxed the mass constraint on Π too.

Assume now that the inf and sup can be exchanged, that is, the inf $-$ sup problem is equivalent to the sup $-$ inf problem. This is not always possible, the main tool in finite dimensional convex analysis is called the Rockafellar theorem. The more general Fenchel-Rockafellar duality theorem is needed in order to show this rigorously, this is outside the scope of this course, we refer to [21, 18] for further information. The exchange of infimum and supremum is true for the Kantorovich reformulation of the transportation problem under the assumption of a l.s.c. cost function c. Now, coming back to the sup $-$ inf problem written as

$$\sup_{\varphi,\psi} \left\{ \int_{\mathbb{R}^d} \varphi\, d\mu + \int_{\mathbb{R}^d} \psi\, d\nu + \inf_{\Pi} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} (c(x,y) - \varphi(x) - \psi(y))\, d\Pi(x,y) \right) \right\},$$

we again notice that the infimum problem can be written as a constraint on the pair of functions (φ,ψ) by realizing that

$$S = \begin{cases} 0 & \text{if } \varphi(x)+\psi(y) \le c(x,y) \text{ on } \mathbb{R}^d\times\mathbb{R}^d \\ -\infty & \text{otherwise} \end{cases},$$

with

$$S := \inf_{\Pi} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} (c(x,y) - \varphi(x) - \psi(y))\, d\Pi(x,y) \right),$$

(this statement is an exercise). Therefore, the sup $-$ inf can be rewritten as an optimization problem with constraints:

$$J_* := \sup_{\varphi,\psi\in C_b(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \varphi\, d\mu + \int_{\mathbb{R}^d} \psi\, d\nu \,:\, \varphi(x)+\psi(y)) \le c(x,y) \right\}.$$

This the the so-called dual optimization problem to the Kantorovich problem. It is easy to observe that $J_* \le I_*$ just by integrating the constraint $\varphi(x)+\psi(y)) \le c(x,y)$ against the measure $\Pi(x,y)$, and thus $J_* < +\infty$. In order to cope with probability measures in the whole space $\mathbb{R}^d$, we need to further relax the dual optimization problem by considering

$$J_* := \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi], \quad \text{with} \quad J[\varphi,\psi] := \int_{\mathbb{R}^d} \varphi\, d\mu + \int_{\mathbb{R}^d} \psi\, d\nu$$

and

$$\Phi_c := \left\{ (\varphi,\psi) \in L^1(d\mu)\times L^1(d\nu) : \varphi(x)+\psi(y) \le c(x,y) \text{ a.e. w.r.t. } \mu\times\nu \right\}.$$

It is not difficult to check that $J_* \le I_*$ still holds for this relaxed problem (this statement is an exercise). Let us now state the duality theorem in full generality whose proof is outside the scope of this basic course, see [21, 18] for details.

Theorem 2.4. *Given two probability measures $\mu,\nu \in \mathscr{P}(\mathbb{R}^d)$, and a lower semi-continuous cost function $c : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0,\infty)$, then there is no duality gap $J_* = I_*$.*

Let us know focus on the particular but important case of the euclidean cost function $c(x,y) = \frac{1}{2}|x-y|^2$ and show the existence of maximizers to the dual optimization problem. We first introcuce some basic concepts of convex analysis. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}\cup\{+\infty\}$, we say that it is proper if f is not identically $+\infty$. Given a proper function, we define its Legendre-Fenchel transform f^* as

$$f^*(y) = \sup_{x\in\mathbb{R}^d} (x\cdot y - f(x)) \qquad \text{for all } y\in\mathbb{R}^d.$$

Notice that the Legendre-Fenchel transform of $\frac{1}{p}|x|^p$ is $\frac{1}{q}|x|^q$ with $\frac{1}{p}+\frac{1}{q}=1$, $1<p<\infty$. Similarly, given a function $\varphi : \mathbb{R}^d \mapsto \mathbb{R}\cup\{-\infty\}$, we say that it is proper if φ is not identically $-\infty$ and its c-transform is defined as

$$\varphi^c(y) = \inf_{x\in\mathbb{R}^d} (\tfrac{1}{2}|x-y|^2 - \varphi(x)) \qquad \text{for all } y\in\mathbb{R}^d.$$

We define c-concave functions as functions that are the c-transform of some function.

Let us remark that since f^* is defined as the supremum of affine functions on y then f^* is a convex function. It is important to notice that the Legendre-Fenchel transform f^* induces a duality between l.s.c. proper convex functions. More precisely, one can prove that a proper function is convex and l.s.c. if and only if there exists g proper function with $f = g^*$, in which case $f^{**} = f$. It is a classical result in convex analysis that convex functions are locally Lipschitz and a.e. differentiable in the interior of the set where they are finite. We refer to [17] as a good source of convex analysis results, a summary can be found in [21, Chapter 2] and [18, Section 1.6].

It is easy to check by definition that for a proper function $\varphi : \mathbb{R}^d \mapsto \mathbb{R}\cup\{-\infty\}$, then

$$\tfrac{1}{2}|y|^2 - \varphi^c(y) = \left(\tfrac{1}{2}|x|^2 - \varphi(x)\right)^*. \tag{2.2}$$

Notice we infer that $\frac{1}{2}|y|^2 - \varphi^c(y)$ is convex for a c-concave function $f = \varphi^c$. In particular, this implies that if f is continuous and concave then $f^{cc} = f$. In fact, this last result is more general: $f^{cc} = f$ characterizes the set of c-concave functions, see [18].

With these notions at hand, let us check that in order to solve the dual optimization problem, we can restrict ourselves to pairs of c-concave functions. Let us also denote by $\mathscr{P}_2(\mathbb{R}^d)$ the set of probability measures with bounded second moment, i.e.,

$$\mathscr{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathscr{P}(\mathbb{R}^d) \,:\, \int_{\mathbb{R}^d} |x|^2 \, d\mu(x) < \infty \right\}.$$

Given two probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, let us denote by M the quantity

$$M := \int_{\mathbb{R}^d} |x|^2 \, d\mu(x) + \int_{\mathbb{R}^d} |x|^2 \, d\nu(x).$$

Lemma 2.1. *Given two probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, then for any $a \in \mathbb{R}$ and any $(\varphi, \psi) \in \Phi_c$, one can change the values of (φ, ψ) on a zero measure set with respect to $\mu \times \nu$ such that $\varphi(x) + \psi(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$. Moreover, for the new pair, denoted the same for simplicity, we have $J[\varphi^{cc} - a, \varphi^c + a] \geq J[\varphi, \psi]$ and $\varphi^{cc}(x) + \varphi^c(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$.*

Furthermore, if there exists $(C_x, C_Y) \in L^1(d\mu) \times L^1(d\nu)$ such that $\varphi^{cc} \leq C_x$ and $\varphi^c \leq C_Y$ and $J[\varphi, \psi] > -\infty$, then $(\varphi^{cc} - a, \varphi^c + a) \in \Phi_c$.

Proof. Since $J[\varphi^{cc} - a, \varphi^c + a] = J[\varphi^{cc}, \varphi^c]$ for all $a \in \mathbb{R}$, we are reduced to show it for $a = 0$. Since the value of

$$J[\varphi, \psi] := \int_{\mathbb{R}^d} \varphi \, d\mu + \int_{\mathbb{R}^d} \psi \, d\nu$$

does not change by changing the values of (φ, ψ) on a zero measure set with respect to $\mu \times \nu$, then we can set $(\varphi, \psi) = (-\infty, -\infty)$ whenever the inequality $\varphi(x) + \psi(y) \leq c(x,y)$ is not satisfied. Therefore, we can assume the inequality $\varphi(x) + \psi(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$. Let us remark that by definition of the c-transform we get

$$\varphi^c(y) = \inf_{x \in \mathbb{R}^d} \left(\tfrac{1}{2}|x-y|^2 - \varphi(x) \right) \geq \psi(y)$$

since $\varphi(x) + \psi(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$, and thus $\varphi^c(y) \geq \psi(y)$ for all $y \in \mathbb{R}^d$. Similarly, one can prove that

$$\varphi^{cc}(x) = \inf_{y \in \mathbb{R}^d} \sup_{z \in \mathbb{R}^d} \left(\tfrac{1}{2}|x-y|^2 - \tfrac{1}{2}|y-z|^2 + \varphi(z) \right) \geq \varphi(x)$$

by choosing $z = x$. By definition we have

$$\varphi^{cc}(x) + \varphi^c(y) = \inf_{z \in \mathbb{R}^d} \left(\tfrac{1}{2}|x-z|^2 - \varphi^c(z) + \varphi^c(y) \right) \leq c(x,y),$$

where the last inequality holds by choosing $z = y$. For the furthermore part of the lemma, one only needs to show the integrability statement: $(\varphi^{cc}, \varphi^c) \in L^1(d\mu) \times L^1(d\nu)$. Note that by assumption

$$\int_{\mathbb{R}^d}(C_x-\varphi^{cc})\,d\mu+\int_{\mathbb{R}^d}(C_y-\varphi^{c})\,d\nu\le\tilde{M}-J[\varphi,\psi]$$

with $\tilde{M}$ given by

$$\tilde{M}=\int_{\mathbb{R}^d}C_x\,d\mu+\int_{\mathbb{R}^d}C_y\,d\nu\,.$$

Since $C_x-\varphi^{cc}\ge 0$ and $C_y-\varphi^{c}\ge 0$, then it follows that $C_x-\varphi^{cc}\in L^1(d\mu)$ and $C_y-\varphi^{c}\in L^1(d\nu)$, and thus by the assumption $(C_x,C_Y)\in L^1(d\mu)\times L^1(d\nu)$, we obtain $(\varphi^{cc},\varphi^{c})\in L^1(d\mu)\times L^1(d\nu)$ as desired. □

We now get an upperbound on maximizing sequences.

Lemma 2.2. *Given two probability measures $\mu,\nu\in\mathscr{P}_2(\mathbb{R}^d)$, then there exists a maximizing sequence $(\varphi_k,\psi_k)\in\Phi_c$ for the dual optimization problem, that is, $J[\varphi_k,\psi_k]\nearrow J_*$ such that $\varphi_k(x)+\psi_k(y)\le c(x,y)$, $\varphi_k(x)\le|x|^2$ and $\psi_k(y)\le|y|^2$ for all $x,y\in\mathbb{R}^d$ and $k\in\mathbb{N}$.*

Proof. Notice that $0=J[0,0]\le J_*\le I^*\le M$ since $\frac{1}{2}|x-y|^2\le|x|^2+|y|^2$. Therefore, there exists a maximizing sequence composed by proper functions $(\varphi_k,\psi_k)\in\Phi_c$. Using the first part of Lemma 2.1, we can assume without loss of generality that $\varphi_k(x)+\psi_k(y)\le c(x,y)$ for all $x,y\in\mathbb{R}^d$ and all $k\in\mathbb{R}^d$. We define the sequence

$$a_k=\inf_{y\in\mathbb{R}^d}(|y|^2-\varphi_k^c(y))\,.$$

Let us first show that $a_k\in\mathbb{R}$. Since $(\varphi_k,\psi_k)\in\Phi_c$, then $\varphi_k(x)\le c(x,y)-\psi_k(y)$ for all $y\in\mathbb{R}^d$. Since ψ_k is a proper function, there exists b_o (possibly depending on k), such that $\varphi_k(x)\le c(x,y_o)+b_o$. Then

$$\varphi_k^c(y_o)=\inf_{x\in\mathbb{R}^d}(\tfrac{1}{2}|x-y_o|^2-\varphi_k(x))\ge-b_o\,,$$

and thus $a_k\le|y_o|^2-\varphi_k^c(y_o)\le|y_o|^2+b_o<+\infty$. Similarly, we also have

$$|y|^2-\varphi_k^c(y)=\sup_{x\in\mathbb{R}^d}(|y|^2-\tfrac{1}{2}|x-y|^2+\varphi_k(x))\ge\sup_{x\in\mathbb{R}^d}(-|x|^2+\varphi_k(x))\ge-|x_o|^2+\varphi_k(x_o)$$

for any $x_o\in\mathbb{R}^d$ and for all $y\in\mathbb{R}^d$, where again we used $\frac{1}{2}|x-y|^2\le|x|^2+|y|^2$. Since φ_k is proper, then we have $a_k\ge-|x_o|^2+\varphi_k(x_o)>-\infty$ for some $x_o\in\mathbb{R}^d$. With this at hand, the new pair $(\tilde{\varphi}_k,\tilde{\psi}_k):=(\varphi_k^{cc}-a_k,\varphi_k^c+a_k)$ is well defined and due to Lemma 2.1 it satisfies $J[\tilde{\varphi}_k,\tilde{\psi}_k]\ge J[\varphi_k,\psi_k]$ and $\tilde{\varphi}_k(x)+\tilde{\psi}_k(y)\le c(x,y)$ a.e. w.r.t. $\mu\times\nu$.

Therefore, we only need to show the integrability $(\tilde{\varphi}_k,\tilde{\psi}_k)\in L^1(d\mu)\times L^1(d\nu)$ to deduce that $(\tilde{\varphi}_k,\tilde{\psi}_k)\in\Phi_c$ by the last part of Lemma 2.1 and finish the proof. Clearly by definition of a_k, we get $\tilde{\psi}_k(y)=\varphi_k^c(y)+a_k\le|y|^2$. By definition of $\varphi_k^{cc}(x)$, we deduce

$$\tilde{\varphi}_k(x)-|x|^2=\inf_{y\in\mathbb{R}^d}(\tfrac{1}{2}|x-y|^2-\varphi_k^c(y)-a_k-|x|^2)\le\inf_{y\in\mathbb{R}^d}(|y|^2-\varphi_k^c(y)-a_k)=0$$

due to $\frac{1}{2}|x-y|^2\le|x|^2+|y|^2$ again. □

We finally can arrive to show the existence of maximizers for the dual optimization problem.

Theorem 2.5. *Given two probability measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, then there exists $(\varphi_o, \psi_o) \in \Phi_c$ such that $J_* = J[\varphi_o, \psi_o]$, and thus*

$$J[\varphi_o, \psi_o] = J_* = \max_{(\varphi,\psi)\in\Phi_c} J[\varphi, \psi]\,.$$

Furthermore, it can be chosen such that $(\varphi_o, \psi_o) = (\eta_o^{cc}, \eta_o^c)$ with $\eta_o \in L^1(d\mu)$ and satisfying the inequalities $\varphi_o(x) + \psi_o(y) \le c(x,y)$, $\varphi_o(x) \le |x|^2$ and $\psi_o(y) \le |y|^2$ for all $x, y \in \mathbb{R}^d$.

Proof. Notice again that $0 = J[0,0] \le J_* \le I^* \le M$ by the assumption, therefore using Lemma 2.2 we have a maximing sequence $(\varphi_k, \psi_k) \in \Phi_c$ satisfying $J[\varphi_k, \psi_k] \nearrow J_*$ such that $\varphi_k(x) + \psi_k(y) \le c(x,y)$, $\varphi_k(x) \le |x|^2$ and $\psi_k(y) \le |y|^2$ for all $x, y \in \mathbb{R}^d$ and $k \in \mathbb{N}$. Take $l \in \mathbb{N}$ and we define the cut-off sequence of functions $(\varphi_k^{(l)}, \psi_k^{(l)})$ as

$$\begin{aligned}\varphi_k^{(l)}(x) &= \max\{\varphi_k(x) - |x|^2, -l\} + |x|^2\\ \psi_k^{(l)}(y) &= \max\{\psi_k(y) - |y|^2, -l\} + |y|^2\,.\end{aligned}$$

It is easy to check that both sequences are decreasing in $l \in \mathbb{N}$ converging as $l \to \infty$ at all points to the original pair (φ_k, ψ_k), that is, $\varphi_k \le \varphi_k^{(l+1)} \le \varphi_k^{(l)}$ and $\psi_k \le \psi_k^{(l+1)} \le \psi_k^{(l)}$ with $\varphi_k^{(l)} \to \varphi_k$ and $\psi_k^{(l)} \to \psi_k$ as $l \to \infty$. Moreover, $-l \le \varphi_k^{(l)}(x) - |x|^2 \le 0$ and $-l \le \psi_k^{(l)}(y) - |y|^2 \le 0$ for all $x, y \in \mathbb{R}^d$ and $k, l \in \mathbb{N}$. Moreover, one can also check that

$$\begin{aligned}\varphi_k^{(l)}(x) + \psi_k^{(l)}(y) &\le \max\{\varphi_k(x) + \psi_k(y) - |x|^2 - |y|^2, -l\} + |x|^2 + |y|^2\\ &\le \max\{c(x,y) - |x|^2 - |y|^2, -l\} + |x|^2 + |y|^2\,,\end{aligned} \tag{2.3}$$

for all $x, y \in \mathbb{R}^d$ and $k, l \in \mathbb{N}$.

For each fixed $l \in \mathbb{N}$, the sequence $\varphi_k^{(l)}(x) - |x|^2$ is bounded in $L^\infty(\mathbb{R}^d)$, and therefore bounded in $L^p(d\mu)$, $1 \le p \le \infty$ since the $L^p(d\mu)$-norms are monotone in p for a probability measure μ. Without loss of generality, we can assume the existence of $\varphi^{(l)}(x) - |x|^2 \in L^2(d\mu)$ such that $\varphi_k^{(l)}(x) - |x|^2 \rightharpoonup \varphi^{(l)}(x) - |x|^2$ weakly in $L^2(d\mu)$. Since $|x|^2 \in L^1(d\mu)$, then $\varphi^{(l)} \in L^1(d\mu)$, since $L^2(d\mu) \subset L^1(d\mu)$ and $|x|^2 \in L^1(d\mu)$. Moreover, since $L^\infty(d\mu) \subset L^2(d\mu)$, then we can use 1 as a test function for the weak convergence $\varphi_k^{(l)}(x) - |x|^2 \rightharpoonup \varphi^{(l)}(x) - |x|^2$ to get

$$\int_{\mathbb{R}^d} \varphi^{(l)}(x)\, d\mu(x) = \lim_{k\to\infty} \int_{\mathbb{R}^d} \varphi_k^{(l)}(x)\, d\mu(x)\,. \tag{2.4}$$

By a diagonalization argument and after extraction of subsequences, we can assume that the above arguments apply for the same subsequence k for all $l \in \mathbb{N}$. Since the

weak convergence preserves the ordering, we conclude that the limiting sequence $\varphi^{(l)} \in L^1(d\mu)$ satisfies $\varphi^{(l+1)} \leq \varphi^{(l)} \leq |x|^2$ with $|x|^2 \in L^1(d\mu)$. Let us denote by φ_o the pointwise limit of the sequence $\varphi^{(l)}$, then the monotone convergence theorem implies that the pointwise limits of φ_o satisfies

$$\int_{\mathbb{R}^d} \varphi_o(x)\, d\mu(x) = \lim_{l\to\infty} \int_{\mathbb{R}^d} \varphi^{(l)}(x)\, d\mu(x)\,. \tag{2.5}$$

An analogous procedure can be done with the sequence $\psi_k^{(l)}(x)$, to define the limiting function ψ_o.

The pair (φ_o, ψ_o) is the our candidate maximiser. We need to show that $(\varphi_o, \psi_o) \in \Phi_c$. We first observe that

$$J_* = \lim_{k\to\infty} J[\varphi_k, \psi_k] \leq \lim_{k\to\infty} J[\varphi_k^{(l)}, \psi_k^{(l)}] = J[\varphi^{(l)}, \psi^{(l)}]$$

since $\varphi_k \leq \varphi_k^{(l)}$, $\psi_k \leq \psi_k^{(l)}$, and (2.4). Then,

$$J_* \leq \lim_{l\to\infty} J[\varphi^{(l)}, \psi^{(l)}] = J[\varphi_o, \psi_o]$$

due to (2.5). Hence, if $(\varphi_o, \psi_o) \in \Phi_c$ then (φ_o, ψ_o) maximises $J[\varphi, \psi]$ and is a solution to the dual optimization problem. By taking the limit $k \to \infty$ and then $l \to \infty$ in (2.3), we get $\varphi_o(x) + \psi_o(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$. Notice here that we use that weak limits preserve ordering. Moreover, since $\varphi^{(l)} \leq |x|^2$ and $\psi^{(l)} \leq |x|^2$ then $\varphi_o(x) \leq |x|^2$ and $\psi_o(y) \leq |y|^2$ for all $x, y \in \mathbb{R}^d$. Finally, integrability follows from

$$0 \leq \int_{\mathbb{R}^d} (|x|^2 - \varphi_o(x))\, d\mu(x) + \int_{\mathbb{R}^d} (|y|^2 - \psi_o(y))\, d\nu(y) \leq -J[\varphi_o, \psi_o] + M \leq -J_* + M\,,$$

where we used $\varphi_o(x) + \psi_o(y) \leq c(x,y)$ in the first inequality, and then since $|x|^2 - \varphi_o \geq 0$ and $|x|^2 - \psi_o \geq 0$, both integrals are finite and thus, $|x|^2 - \varphi_o(x) \in L^1(d\mu)$ and $|y|^2 - \psi_o(y) \in L^1(d\nu)$. Hence, $\varphi_o(x) \in L^1(d\mu)$ and $\psi_o(y) \in L^1(d\nu)$ since $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ finalizing the proof of the claim $(\varphi_o, \psi_o) \in \Phi_c$. A further application of the double c-transform trick in Lemma 2.1 shows the additional statement in the form of the obtained maximizer (this statement is an exercise). □

The pair of functions (φ_o, ψ_o) achieving the maximum are called Kantorovich potentials for the dual optimization problem, and they can be assumed to be c-concave functions without loss of generality. In fact, given a maximizer of the dual optimization problem (φ_o, ψ_o), it is not difficult to show that it is equal μ and ν-a.e. respectively to c-concave Kantorovich potentials. We now take advantage further of their definitions as c-transforms of a given $\eta \in L^1(d\mu)$. We insisted to do all the previous computations with c-concave functions to show that this proof has the potential to be generalizable to a family of costs functions much larger than the quadratic cost. Let us check that the Kantorovich c-concave potentials are in fact more regular than simply integrable functions taking advantage of its particular form.

Corollary 2.1. *Any c-concave Kantorovich potentials for the dual optimization problem are locally Lipschitz in the interior of the set where they are finite. Furthermore, the Kantorovich potentials can be chosen of the form* (φ_o, φ_o^c) *with* φ_o *c-concave and satisfying the inequalities* $\varphi_o(x) + \varphi_o^c(y) \leq c(x,y)$, $\varphi_o(x) \leq |x|^2$ *and* $\varphi_o^c(y) \leq |y|^2$ *for all* $x, y \in \mathbb{R}^d$.

Proof. Since $(\varphi_o, \psi_o) = (\eta_o^{cc}, \eta_o^c)$ with $\eta_o \in L^1(d\mu)$, then each Kantorovich potential is the c-transform of some $\eta_o \in L^1(d\mu)$. Since $\frac{1}{2}|y|^2 - \varphi^c(y)$ is convex for a c-concave function φ due to (2.2), and convex functions are locally Lipschitz continuous and a.e. differentiable in the interior of the set wherever they are finite, we obtain the same property for φ. A further application of the double c-transform trick in Lemma 2.1 shows the additional statement in the form of the obtained maximizer (this statement is an exercise) using that for c-concave functions $f^{cc} = f$. □

The previous corollary asserts that c-concave Kantorovich potentials are a.e. differentiable with respect to the Lebesgue measure in the interior of the set wherever they are finite. In fact, from the duality Theorem 2.4 and the existence of minimizers and maximizers of the primal and the dual optimization problems in Theorems 2.3 and 2.5, we deduce that given Π_o optimal transference plan and a φ_o concave Kantorovich potential, then

$$\begin{aligned} J_* = I_* &= \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\Pi_o(x,y) = \int_{\mathbb{R}^d} \varphi_o(x)\, d\mu(x) + \int_{\mathbb{R}^d} \varphi_o^c(y)\, d\nu(x) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} (\varphi_o(x) + \varphi_o^c(y))\, d\Pi_o(x,y). \end{aligned}$$

Since $\varphi_o(x) + \varphi_o^c(y) \leq c(x,y)$ for all $x, y \in \mathbb{R}^d$ from Theorem 2.5 and Corollary 2.1, then one expects $\varphi_o(x) + \varphi_o^c(y) = c(x,y)$ Π_o-a.e. Let us define the support of the measure Π_o as:

Definition 2.5. The support of a measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ is defined as the smallest closed set in which μ is not zero, i.e.

$$\begin{aligned} spt(\mu) &:= \bigcap \{A : A \text{ is closed and } \mu(\mathbb{R}^d \setminus A) = 0\} \\ &= \{x \in \mathbb{R}^d : \mu(B_r) > 0 \text{ for all } r > 0\}. \end{aligned}$$

Therefore, one can prove that $\varphi_o(x) + \varphi_o^c(y) = c(x,y)$ on $spt(\Pi_o)$. A full proof of this fact needs the Knott-Smith optimality criteria using that φ_o is c-concave that we refer to [21]. Now, given $(x_o, y_o) \in spt(\Pi_o)$ and using the definition of the c-transform $\varphi_o^c(y_o)$, the function $x \mapsto \varphi_o(x) - c(x, y_o)$ achieves its minimum at $x = x_o$. Assuming the Kantorovich potential is differentiable at x_o, we deduce that $\nabla\varphi_o(x_o) = x_o - y_o$. This implies that y_o is uniquely determined in terms of x_o if the Kantorovich potential φ_o is differentiable at x_o by $y_o = x_o - \nabla\varphi_o(x_o)$ for $(x_o, y_o) \in spt(\Pi_o)$. Since $\varphi_o \in L^1(d\mu)$ and $J_* \in \mathbb{R}$, then φ_0 is finite μ-a.e. Since convex functions are differentiable a.e. on the closure of set of points where they are finite (this is a consequence of Alexandrov's theorem, an advanced result in con-

vex analysis, see [22]) and if we further assume that $\mu \ll \mathscr{L}$, then the Kantorovich potential is μ-a.e. differentiable.

Therefore, if the measure $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ is absolutely continuous, then the Kantorovich potential is μ-a.e. differentiable and its gradient is defined uniquely μ-a.e. by the relation $y_o = x_o - \nabla\varphi_o(x_o)$. Therefore, we have shown that any optimal transference plan in the Kantorovich reformulation of the optimal transport problem with quadratic cost can be characterized as $\Pi_o = (1_{\mathbb{R}^d} \times T)\#\mu$ with $T(x) = x - \nabla\varphi_o(x)$, and therefore the optimal transference plan is unique μ-a.e. since it only depends on the values of $\nabla\varphi_o$ μ-a.e.

To make the last statements completely rigorous, one can use disintegration of measures that in the case of probability measures $\Pi \in \Gamma(\mu,\nu)$ reads as: given $\Pi \in \Gamma(\mu,\nu)$ and any test function $\zeta \in C_b(\mathbb{R}^d \times \mathbb{R}^d)$, we can find a unique μ-a.e. defined family of probabilty measures $\mu_x \in \mathscr{P}(\mathbb{R}^d)$, $x \in \mathbb{R}^d$, supported inside $\{x\} \times \mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \zeta(x,y)\, d\Pi(x,y) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \zeta(x,y)\, d\mu_x(y)\, d\mu(x).$$

This is also the precise definition of conditional law in probability theory. If X is a random variable with law μ and Y is a random variable with law ν, $\Pi \in \Gamma(\mu,\nu)$ represents the law of a coupling (X,Y) between X and Y. Then, μ_x represents the law of the conditional probability of the random variable Y subject to knowing $X = x$. With the disintegration of measures at hand, see [3] for a proof, we have previously shown that by disintegrating the measure Π_o with respect to μ then $\mu_x = \delta_{y=T(x)}$ since $T(x)$ is the only point on the support of Π_o for x μ-a.e. The above considerations can now be stated as the following result which is due to Yann Brenier in a more general form.

Theorem 2.6. *[Monge finally meets Kantorovich] Given two probability measures $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ with $\mu \ll \mathscr{L}$, then there exists an unique optimal transference plan for the quadratic cost of the form $\Pi_o = (1_{\mathbb{R}^d} \times T)\#\mu \in \Gamma(\mu,\nu)$ achieving the infimum in the Kantorovich formulation of optimal transport*

$$\int_{\mathbb{R}^d} |x - T(x))|^2\, d\mu(x) = \min_{\Pi \in \Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2\, d\Pi(x,y) \right\}.$$

Moreover, this map is given by $T(x) = x - \nabla\varphi_o(x)$ defined uniquely μ-a.e. where $\varphi_o(x)$ is a c-concave Kantorovich potential.

The previous theorem finally connects the Monge transportation problem to the Kantorovich reformulation by showing the the infimum on the Monge problem is achieved and coincides with the minimum of the Kantorovich reformulation for the quadratic cost if $\mu \ll \mathscr{L}$. Notice also that the optimal transport map can be chosen as $T = \nabla\Psi$ with Ψ a convex function by taking $\Psi = \frac{1}{2}|x|^2 - \varphi_o(x)$. The uniqueness part is not proven here and we refer to the literature. All the results presented for the quadratic cost can be similarly generalized to costs functions of the form $c(x,y) = h(x-y)$ with h strictly convex. For instance, $h(s) = |s|^p$, $1 < p < \infty$. We refer to [21, 22, 18].

2.4 Transport distances between measures: properties.

The goal of this section is to introduce transport distances based on the optimal transport introduced in the previous section. Let us take simple cases first. Assume that $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ are just two Dirac Deltas at two different points $\mu = \delta_{x_o}$ and $\nu = \delta_{x_1}$ with $x_o, x_1 \in \mathbb{R}^d$ and $x_o \neq x_1$. Then, it is easy to see that the norm introduced in the set of finite signed measures is $\|\delta_{x_o} - \delta_{x_1}\| = 2$ no matter how close the two points are. Notice that in fact the norm introduced in Definition 2.1 is just the total variation norm between measures. It is clearly not a good distance if we think about how close or how far are δ_{x_o}, δ_{x_1} in terms of the distance between the points where they are concentrated on. Now, let us take any $\Pi \in \Gamma(\delta_{x_o}, \delta_{x_1})$. It is clear that the only possible transference plan is the product measure $\delta_{x_o} \times \delta_{x_1}$, for instance using the disintegration of measures theorem. Therefore, any map that sends x_o onto x_1 is an optimal map for the optimal mass transportation problem for any l.s.c. cost function $c(x,y)$ and therefore the optimal cost is $c(x_0, x_1)$. A desirable property of the cost function satisfied by all the basic costs $c(x,y) = |x-y|^p$, $1 \leq p < \infty$, is that the cost is continuous and has zero value for $x = y$. Thus we could consider the value of the cost transporting δ_{x_o} onto δ_{x_1} as a measure of the distance between the probability measures δ_{x_o} and δ_{x_1}. Moreover, it is a measure that is continuous as $x_o \to x_1$. These ideas lead to the following definition.

Definition 2.6. The Wasserstein distance between μ and ν, d_p, $1 \leq p < \infty$ can be defined by

$$d_p^p(\mu, \nu) = \inf_{\Pi \in \Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, d\Pi(x,y) \right\},$$

i.e., by the p-th root of the value of the optimum in the Kantorovich reformulation of the mass transport problem with cost $c(x,y) = |x-y|^p$, $1 \leq p < \infty$.

Notice that the Wasserstein distance d_p is finite for any measures $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$, being

$$\mathscr{P}_p(\mathbb{R}^d) := \left\{ \mu \in \mathscr{P}(\mathbb{R}^d) \, : \, \int_{\mathbb{R}^d} |x|^p \, d\mu(x) < \infty \right\},$$

$1 \leq p < \infty$. The classical Monge-Kantorovich problem was posed for the case of the Euclidean distance, $p = 1$, and it is usually refered as the Monge-Kantorovich distance. Another name in the engineering and applied mathematical sciences used is the earth movers distance alluding to the origin of the Monge problem. The name Wasserstein was used due to classical papers popularizing the use of d_2 in Partial Differential Equations, and it has kept that name for the last 20+ years. However, historically attributing to Wasserstein the name of this distance is not wrong but not completely fair either. Many people call the distances d_p as transport distances too. More information about other appearances in the literature of these transport distances can be read in the summer school notes in [11].

From a probabilistic point of view, the Wasserstein distance d_p can be alternatively defined as

$$d_p^p(\mu,\nu) = \inf_{(X,Y)\in\tilde{\Gamma}} \mathbb{E}\left[|X-Y|^p\right], \tag{2.6}$$

where $\tilde{\Gamma}$ is the set of all possible couplings of random variables (X,Y) with laws μ and ν respectively, i.e., $X,Y : (\mathscr{S},\mathscr{A},P) \longrightarrow \mathscr{L}$ measurable maps from a probability space of reference $(\mathscr{S},\mathscr{A},P)$ onto the Lebesgue space $\mathscr{L}$ and $(X,Y) : (\mathscr{S},\mathscr{A},P) \longrightarrow \mathscr{L}\times\mathscr{L}$ such that the laws or image measures are $X\#P = \mu$, $Y\#P = \nu$, and $(X,Y)\#P = \Pi$ with $\Pi \in \Gamma(\mu,\nu)$.

In order to prove the triangle inequality, we need some preliminary results.

Lemma 2.3. *[Gluing lemma] Given probability measures $\mu,\nu,\omega \in \mathscr{P}(\mathbb{R}^d)$, $\Pi_1 \in \Gamma(\mu,\nu)$ and $\Pi_2 \in \Gamma(\nu,\omega)$, there exists a measure $\gamma \in \mathscr{P}(\mathbb{R}^{3d})$ such that $P_{12}\#\gamma = \Pi_1$ and $P_{23}\#\gamma = \Pi_2$ being P_{12} and P_{23} the projections maps into the first and the last two variables respectively, i.e., $P_{12}(x,y,z) = (x,y)$ and $P_{23}(x,y,z) = (y,z)$ for all $x,y,z \in \mathbb{R}^d$.*

Proof. By the disintegration of measures, we can write

$$\Pi_1(A\times B) = \int_B \nu_y^1(A)\, d\nu(y) \qquad \text{and} \qquad \Pi_2(B\times C) = \int_B \nu_y^2(C)\, d\nu(y)$$

for some family of probability measures ν_y^i, $i_1,2$, and any Borel sets A,B,C in $\mathbb{R}^d$. We define $\gamma \in \mathscr{P}(\mathbb{R}^{3d})$ given by

$$\gamma(A\times B\times C) = \int_B \nu_y^1(A)\nu_y^2(C)\, d\nu(y).$$

It is easy to check that $\gamma(A\times B\times\mathbb{R}^d) = \Pi_1(A\times B)$ and $\gamma(\mathbb{R}^d\times B\times C) = \Pi_2(B\times C)$ as desired.

Proposition 2.1. *The distance d_p is a metric on $\mathscr{P}_p(\mathbb{R}^d)$.*

Proof. Since the cost $c(x,y) = |x-y|^p$ is nonnegative and symetric, it is easy to see that the optimal value is nonnegative and that the distance is symmetric on its arguments $d_p^p(\mu,\nu) = d_p^p(\nu,\mu)$. For the last statement, notice that $\Pi \in \Gamma(\mu,\nu)$ if and only if $S\#\Pi \in \Gamma(\nu,\mu)$ with $S : \mathbb{R}^d\times\mathbb{R}^d \mapsto \mathbb{R}^d\times\mathbb{R}^d$ given by $S(x,y) = (y,x)$. Now, if $\mu = \nu$, we can take $\Pi(x,y) = \delta_x(y)\mu(x) \in \Gamma(\mu,\nu)$ to obtain that

$$0 \le d_p^p(\mu,\mu) \le \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p\, d\Pi(x,y) = 0$$

since $x = y$ Π-a.e. Now, if $d_p^p(\mu,\mu) = 0$ then there exists $\Pi(x,y) \in \Gamma(\mu,\nu)$ such that $x = y$ Π-a.e. Hence, for any test function $\zeta \in C_b(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} \zeta(x)\, d\mu(x) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \zeta(x)\, d\Pi(x,y) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \zeta(y)\, d\Pi(x,y) = \int_{\mathbb{R}^d} \zeta(y)\, d\nu(y),$$

and thus $\mu = \nu$ by the Riesz representation theorem. The only remaining property to show is the triangular inequality. Let $\mu,\nu,\omega \in \mathscr{P}(\mathbb{R}^d)$, $\Pi_1 \in \Gamma(\mu,\nu)$ and $\Pi_2 \in$

$\Gamma(\nu,\omega)$ optimal transference plans by Theorem 2.3. Lemma 2.3 implies there exists a measure $\gamma\in\mathscr{P}(\mathbb{R}^{3d})$ such that $P_{12}\#\gamma=\Pi_1$ and $P_{23}\#\gamma=\Pi_2$. We define $\Pi_3=P_{13}\#\gamma$ being $P_{13}(x,y,z)=(x,z)$ for all $x,y,z\in\mathbb{R}^d$. One can check that $\Pi_3\in\Gamma(\mu,\omega)$ (this statement is an exercise). Using the definition of the distance d_p, the definition of γ, the triangle inequality, the Minkowski inequality for L^p spaces, and the optimality of Π_1 and Π_2, we obtain

$$\begin{aligned}
d_p(\mu,\omega) &\le \left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-z|^p\,d\Pi_3(x,y)\right)^{\frac{1}{p}} = \left(\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}|x-z|^p\,d\gamma(x,y,z)\right)^{\frac{1}{p}}\\
&\le \left(\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}(|x-y|+|y-z|)^p\,d\gamma(x,y,z)\right)^{\frac{1}{p}}\\
&\le \left(\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}|x-y|^p\,d\gamma(x,y,z)\right)^{\frac{1}{p}} + \left(\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}|y-z|^p\,d\gamma(x,y,z)\right)^{\frac{1}{p}}\\
&= \left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^p\,d\Pi_1(x,y)\right)^{\frac{1}{p}} + \left(\int_{\mathbb{R}^d\times\mathbb{R}^d}|y-z|^p\,d\Pi_2(y,z)\right)^{\frac{1}{p}}\\
&= d_p(\mu,\nu)+d_p(\nu,\omega)\,,
\end{aligned}$$

as desired. □

Finally, let us remark that the sequence of metrics $d_p(\mu,\nu)$ is nondecreasing in p, $1\le p<\infty$. This is a simple consequence of the Hölder's inequality for L^p-spaces. This allows to define a quantity that we call the ∞-Wasserstein distance as

$$d_\infty(\mu,\nu):=\lim_{p\nearrow\infty} d_p(\mu,\nu).$$

This is at least a metric on the set of compactly supported probability measures. We will not discuss much more on this interesting transport distance and refer to the literature for more details. By the monotone property of the distances d_p, we deduce that for compactly supported probability measures, the topology induced by d_p gets finer as p increases.

Notice also that if $\mu,\nu\in\mathscr{P}(\mathbb{R}^d)$ are both supported on a ball $\bar{B}_R$ then

$$d_p(\mu,\nu)\le(2R)^{(p-1)/p}d_1(\mu,\nu)^{1/p}\,. \tag{2.7}$$

This is due to the fact that for any $\Pi\in\Gamma(\mu,\nu)$ we have $\Pi(\mathbb{R}^d\times(\mathbb{R}^d\setminus\bar{B}_R))=\nu(\mathbb{R}^d\setminus\bar{B}_R)=0$ and $\Pi((\mathbb{R}^d\setminus\bar{B}_R)\times\mathbb{R}^d)=\mu(\mathbb{R}^d\setminus\bar{B}_R)=0$. Thus, we deduce $\Pi((\mathbb{R}^d\times\mathbb{R}^d)\setminus(\bar{B}_R\times\bar{B}_R))=0$, and therefore, $spt(\Pi)\subset\bar{B}_R\times\bar{B}_R$. Take now the optimal transference plan $\Pi_o\in\Gamma(\mu,\nu)$ for the distance d_1, then

$$\begin{aligned}
d_p^p(\mu,\nu) &\le \int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^p\,d\Pi_o(x,y) = \int_{\bar{B}_R\times\bar{B}_R}|x-y|^p\,d\Pi_o(x,y)\\
&\le (2R)^{(p-1)}\int_{\bar{B}_R\times\bar{B}_R}|x-y|\,d\Pi_o(x,y) = (2R)^{(p-1)}d_1(\mu,\nu)\,,
\end{aligned}$$

leading to the desired inequality.

Let us focus now in understading the notion of convergence in transport metrics d_p. We will denote by $\mathrm{Lip}(\mathbb{R}^d)$ the set of Lipschitz functions on $\mathbb{R}^d$ and by $W^{1,\infty}(\mathbb{R}^d)$ the set of bounded and Lipschitz functions on $\mathbb{R}^d$.

Corollary 2.2 (Convergence of averages with d_p). *Given probability measures $\mu,\nu\in\mathscr{P}_p(\mathbb{R}^d)$ and $\varphi\in\mathrm{Lip}(\mathbb{R}^d)$ with Lipschitz constant L, then we have*

$$\left|\int_{\mathbb{R}^d}\varphi(x)\,d\mu(x)-\int_{\mathbb{R}^d}\varphi(x)\,d\nu(x)\right|\le L d_p(\mu,\nu).$$

Proof. Since $d_p(\mu,\nu)$ is nondecreasing in p, we can reduce to show the statement for d_1. Let $\Pi_o(x,y)$ an optimal plan between $\mu,\nu\in\mathscr{P}_1$ for d_1. Then

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|d\Pi_o(x,y)=d_1(\mu,\nu),$$

and we can write

$$\int_{\mathbb{R}^d}\varphi(x)\,d\mu(x)-\int_{\mathbb{R}^d}\varphi(x)\,d\nu(x)=\int_{\mathbb{R}^d\times\mathbb{R}^d}(\varphi(x)-\varphi(y))\,d\Pi_o(x,y).$$

Using that φ is Lipschitz with constant L and estimating, we get

$$\begin{aligned}\left|\int_{\mathbb{R}^d}\varphi(x)\,d\mu(x)-\int_{\mathbb{R}^d}\varphi(x)\,d\nu(x)\right|&\le\int_{\mathbb{R}^d\times\mathbb{R}^d}|\varphi(x)-\varphi(y)|\,d\Pi_o(x,y)\\&\le L\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|\,d\Pi_o(x,y)\le L d_1(\mu,\nu)\end{aligned}$$

giving the assertion. □

In the noticeable case of the Monge-Kantorovich distance d_1, the previous corollary is a characterization by duality. More precisely, as a consequence of Fenchel-Rockafellar's duality principle, one can deduce the Kantorovich-Rubinstein theorem [21, Theorem 1.14] giving that

$$d_1(\mu,\nu)=\sup\left\{\left|\int_{\mathbb{R}^d}\varphi(x)\,d(\mu-\nu)(x)\right|,\varphi\in\mathrm{Lip}(\mathbb{R}^d),\|\varphi\|_{\mathrm{Lip}(\mathbb{R}^d)}\le 1\right\}. \tag{2.8}$$

Another classical distance between measures , not necessarily probabilty measures, is the so-called Bounded Lipschitz (BL) distance, that is defined as

$$\|\mu-\nu\|_{BL}=\sup\left\{\left|\int_{\mathbb{R}^d}\varphi(x)\,d(\mu-\nu)(x)\right|,\varphi\in W^{1,\infty}(\mathbb{R}^d),\|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)}\le 1\right\},$$

that is, the dual $W^{1,\infty}(\mathbb{R}^d)$-norm. The convergence in BL-distance is equivalent to weak convergence of measures by the Pormanteau theorem. We will see next that the topology in d_1 is finer than the topology induced by the BL-distance in $\mathbb{R}^d$.

However, we start by showing that on compact sets of $\mathbb{R}^d$ the convergence in any transport distance is also equivalent to weak convergence of measures.

Proposition 2.2. *Given a sequence of probability measures $\mu_n \in \mathscr{P}_p(\mathbb{R}^d)$ and $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ supported on some euclidean ball $\bar{B}_R$. Then $d_p(\mu_n,\mu) \to 0$ if and only if μ_n converges weakly to μ.*

Proof. We first notice that due to (2.7) and that $d_1(\mu_n,\mu) \le d_p(\mu_n,\mu)$, we are reduced to show the result for d_1. The charaterization (2.8) implies directly that if $d_1(\mu_n,\mu) \to 0$ then μ_n converges weakly to μ. Here, we can use the Pormanteau theorem, a result that can be seen in basic courses of probability, or an argument by density to go from Lipschitz to bounded continuous functions. Conversely, let us take the subsequence μ_{n_k} such that

$$\limsup_{n\to\infty} d_1(\mu_n,\mu) = \lim_{k\to\infty} d_1(\mu_{n_k},\mu)\,.$$

By the characterization (2.8), we have the existence of a sequence of 1-Lispschitz functions φ_k such that

$$d_1(\mu_{n_k},\mu) \le \int_{\mathbb{R}^d} \varphi_k(x)\, d(\mu_{n_k}-\mu)(x) + \frac{1}{k} = \int_{\bar{B}_R} (\varphi_k(x)-\varphi_k(0))\, d(\mu_{n_k}-\mu)(x) + \frac{1}{k}\,,$$

since $\mu_n \in \mathscr{P}_p(\mathbb{R}^d)$ and $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ supported on $\bar{B}_R$. The sequence of functions $\tilde{\varphi}_k(x) := \varphi_k(x) - \varphi_k(0)$ is now 1-Lipschitz and bounded defined on $\bar{B}_R$, then by Ascolí-Arzelá theorem, there exists a further subsequence, that we denote with the same index, converging uniformly to a 1-Lispchitz function $\tilde{\varphi}$. Hence, we conclude that

$$\begin{aligned}
\limsup_{n\to\infty} d_1(\mu_n,\mu) &\le \limsup_{k\to\infty} \int_{\bar{B}_R} \tilde{\varphi}_k(x)\, d(\mu_{n_k}-\mu)(x) \\
&\le \limsup_{k\to\infty} \int_{\bar{B}_R} (\tilde{\varphi}_k(x)-\tilde{\varphi}(x))\, d\mu_{n_k}(x) + \limsup_{k\to\infty} \int_{\bar{B}_R} \tilde{\varphi}(x)\, d(\mu_{n_k}-\mu)(x) \\
&\quad - \limsup_{k\to\infty} \int_{\bar{B}_R} (\tilde{\varphi}_k(x)-\tilde{\varphi}(x))\, d\mu(x) \\
&\le 2 \limsup_{k\to\infty} \|\tilde{\varphi}_k(x)-\varphi(x)\|_{L^\infty(\bar{B}_R)} + \limsup_{k\to\infty} \int_{\bar{B}_R} \tilde{\varphi}(x)\, d(\mu_{n_k}-\mu)(x) = 0\,,
\end{aligned}$$

where we used in the last line the weak convergence of μ_n towards μ, and thus $d_1(\mu_n,\mu) \to 0$ as $n \to \infty$. □

Now, we come back to the whole space to study the notion of d_p convergence.

Theorem 2.7. *Given a sequence of probability measures $\mu_n \in \mathscr{P}_p(\mathbb{R}^d)$ and $\mu \in \mathscr{P}_p(\mathbb{R}^d)$. Then $d_p(\mu_n,\mu) \to 0$ if and only if μ_n converges weakly to μ and*

$$\int_{\mathbb{R}^d} |x|^p\, d\mu_n(x) \to \int_{\mathbb{R}^d} |x|^p\, d\mu(x) \quad \text{as } n \to \infty.$$

Proof. The necessary implication is a consequence of Corollary 2.2 together with the triangle inequality. In fact, observing that

$$\int_{\mathbb{R}^d} |x|^p \, d\mu_n(x) = d_p^p(\mu_n, \delta_0) \qquad \text{and} \qquad \int_{\mathbb{R}^d} |x|^p \, d\mu(x) = d_p^p(\mu, \delta_0)\,,$$

we get by the triangle inequality for d_p that

$$|d_p(\mu_n, \delta_0) - d_p(\mu, \delta_0)| \le d_p(\mu_n, \mu) \to 0$$

as $n \to \infty$.

Conversely, let us truncate the p-th moment by defining $\phi_R(x) = \min(|x|, R)^p$ which is continuous and bounded. Therefore, by weak convergence of the sequence μ_n towards μ and the convergence of the p-th moments, we get

$$\int_{\mathbb{R}^d} (|x|^p - \phi_R(x)) \, d\mu_n(x) \to \int_{\mathbb{R}^d} (|x|^p - \phi_R(x)) \, d\mu(x) \quad \text{as } n \to \infty.$$

Now, we can take R large enough such that

$$\int_{\mathbb{R}^d} (|x|^p - \phi_R(x)) \, d\mu(x) = \int_{|x|>R} (|x|^p - R^p) \, d\mu(x) \le \int_{|x|>R} |x|^p \, d\mu(x) \le \frac{\varepsilon}{2}\,,$$

for a given fixed $\varepsilon > 0$, since $\mu \in \mathscr{P}_p(\mathbb{R}^d)$. Therefore, for n large enough, we also have

$$\int_{\mathbb{R}^d} (|x|^p - \phi_R(x)) \, d\mu_n(x) \le \varepsilon\,.$$

For $0 < b < a$ and $p \ge 1$, it is easy to check that $a^p + b^p \le (a+b)^p$. We can infer that for $|x| > R$ then $(|x| - R)^p \le |x|^p - R^p = |x|^p - \phi_R(x)$. So for n large enough

$$\int_{|x|>R} (|x| - R)^p \, d\mu_n(x) \le \varepsilon \qquad \text{and} \qquad \int_{|x|>R} (|x| - R)^p \, d\mu(x) \le \varepsilon\,.$$

Let us consider the euclidean projection onto the ball $\bar{B}_R$ denoted by P_R, this map is continuous, leaves invariant $\bar{B}_R$ and otherwise $|x - P_R(x)| = |x| - R$. Hence, we deduce that

$$d_p^p(\mu, P_R\#\mu) \le \int_{\mathbb{R}^d} |x - P_R(x)|^p \, d\mu(x) = \int_{|x|>R} (|x| - R)^p \, d\mu(x) \le \varepsilon\,,$$

and analogously $d_p^p(\mu_n, P_R\#\mu_n) \le \varepsilon$. Since $\mu_n \rightharpoonup \mu$ weakly, it is easy to check that $P_R\#\mu_n \rightharpoonup P_R\#\mu$ weakly, and thus using the characterization of the convergence for measures supported in B_R in Proposition 2.2, we get $d_p(P_R\#\mu_n, P_R\#\mu) \to 0$ as $n \to \infty$. We conclude by estimating using the triangle inequality as follows:

$$\begin{aligned} d_p(\mu_n, \mu) &\le d_p(\mu_n, P_R\#\mu_n) + d_p(P_R\#\mu_n, P_R\#\mu) + d_p(\mu, P_R\#\mu) \\ &\le 2\varepsilon^{1/p} + d_p(P_R\#\mu_n, P_R\#\mu)\,, \end{aligned}$$

for n large enough. Taking the limit $n \to \infty$, we get

$$\limsup_{n\to\infty} d_p(\mu_n, \mu) \leq 2\varepsilon^{1/p}$$

and then, taking the limit $\varepsilon \to 0$, we finally obtain $d_p(\mu_n, \mu) \to 0$ as $n \to \infty$ as desired. □

We end this section by making a summary of the main properties of the d_p distances.

Proposition 2.3 (d_p-properties). *The space $(\mathscr{P}_p(\mathbb{R}^d), d_p)$ is a complete metric space, $1 \leq p < \infty$. Moreover, the following properties of the distance d_p hold:*

i) **Optimal transference plan:** *The infimum in the definition of the distance d_p is achieved at a joint probability measure Π_o called an optimal transference plan satisfying:*

$$d_p^p(\mu, \nu) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Pi_o(x,y).$$

ii) **Convergence of measures:** *Given μ_n and μ in $\mathscr{P}_p(\mathbb{R}^d)$, $1 \leq p < \infty$, the following assertions are equivalent:*

a) $d_p(\mu_n, \mu)$ tends to 0 as n goes to infinity.
b) μ_n tends to μ weakly and

$$\int_{\mathbb{R}^d} |x|^p \, d\mu_n(x) \to \int_{\mathbb{R}^d} |x|^p \, d\mu(x) \text{ as } n \to +\infty.$$

iii) **Lower semicontinuity:** *d_p is weakly-* lower semicontinuous in each argument, $1 \leq p < \infty$.*

iv) **Moments as distances:** *If $\mu \in \mathscr{P}_p(\mathbb{R}^d)$, then*

$$d_p^p(\mu, \delta_a) = \int_{\mathbb{R}^d} |x|^2 \, d\mu(x).$$

v) **Convexity:** *Given f_1, f_2, g_1 and g_2 in $\mathscr{P}_p(\mathbb{R}^d)$ and α in $[0,1]$, then*

$$d_p^p(\alpha f_1 + (1-\alpha) f_2, \alpha g_1 + (1-\alpha) g_2) \leq \alpha d_p^p(f_1, g_1) + (1-\alpha) d_p^p(f_2, g_2).$$

As a simple consequence, given f, g and h in $\mathscr{P}_p(\mathbb{R}^d)$, then

$$d_p(h * f, h * g) \leq d_p(f, g)$$

where $*$ stands for the convolution of measures in $\mathbb{R}^d$.

vi) **Additivity with respect to convolution of d_2:** *Given f_1, f_2, g_1 and g_2 in $\mathscr{P}_2(\mathbb{R}^d)$ with equal mean values, then*

$$d_2^2(f_1 * f_2, g_1 * g_2) \leq d_2^2(f_1, g_1) + d_2^2(f_2, g_2).$$

Proof. Most of the properties have been shown except Property *iii)* and the last two. Properties *iii)* and *v)* are left to the reader (these statements are exercises). Notice here that the convolution between measures is defined as usual by duality on test functions, meaning $f * g$ for $f, g \in \mathscr{P}(\mathbb{R}^d)$ is defined as the measure $f * g \in \mathscr{P}(\mathbb{R}^d)$ such that

$$\int_{\mathbb{R}^d} \zeta(x)\, d(f*g)(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \zeta(x+y)\, df(x)\, dg(y) \quad \forall \zeta \in \mathscr{C}_b(\mathbb{R}^d)\,.$$

Property *vi)* is a direct consequence of the definition of d_2 in terms of random variables. Let (X_1, Y_1), (X_2, Y_2) be two independent pairs of random variables, and let f_i (resp. g_i) be the laws of X_i (resp. Y_i) $i = 1, 2$. Suppose moreover that X_i and Y_i have the same mean value, namely $\mathbb{E}[X_i] = \mathbb{E}[Y_i]$ $i = 1, 2$. If the pairs (X_1, Y_1), (X_2, Y_2) realize the optimal transference plans, then for $i = 1, 2$

$$d_2^2(f_i, g_i) = \mathbb{E}\left[|X_i - Y_i|^2\right].$$

Notice that the law of the random variable $X_1 + X_2$ is given by $f_1 * f_2$ since X_1 and X_2 are independent (this statement is an exercise), then

$$\begin{aligned} d_2^2(f_1 * f_2, g_1 * g_2) &\leq \mathbb{E}\left[|(X_1 + X_2) - (Y_1 + Y_2)|^2\right] \\ &= \mathbb{E}\left[|X_1 - Y_1|^2\right] + \mathbb{E}\left[|X_2 - Y_2|^2\right] + 2\,\mathbb{E}\left[(X_1 - Y_1) \cdot (X_2 - Y_2)\right] \\ &= d_2^2(f_1, g_1) + d_2^2(f_2, g_2) \end{aligned}$$

In fact, the term $\mathbb{E}\left[(X_1 - Y_1) \cdot (X_2 - Y_2)\right]$ is equal to zero due to the independence of the pairs, and to the equality of the mean values. □

2.5 One-dimensional Wasserstein metric

Given a probability measure in one dimension, $\mu \in \mathscr{P}(\mathbb{R})$, we define as usual its distribution function $F : \mathbb{R} \mapsto [0, 1]$ as

$$F(x) = \int_{-\infty}^{x} d\mu(x) = \mu((-\infty, x])\,.$$

Notice that the definite integral is just a notation here made precise by the second equality. With this definition, F is a monotone nondecreasing right-continuous function with $F(-\infty) = 0$ and $F(+\infty) = 1$, in fact it is a càdlàg function. In principle, F does not have an inverse since it can have plateaus, but we can define its generalized inverse or pseudo-inverse function $\mathbb{X} : [0, 1] \mapsto \mathbb{R} \cup \{\pm\infty\}$ as follows:

$$\mathbb{X}(\eta) = \inf_{x \in \mathbb{R}} \{x \in \mathbb{R} : F(x) \geq \eta\}$$

for all $\eta \in [0,1]$. The infimum is a minimum as soon as the set is not empty (otherwise is $+\infty$) and bounded from below (otherwise it is $-\infty$), thanks to the right continuity of F. Note that $\mathbb{X}(\eta) \leq x \iff \eta \leq F(x)$ or equivalently $\mathbb{X}(\eta) > x \iff \eta > F(x)$. By definition $\mathbb{X}(\eta)$ is a nondecreasing function and $F(\mathbb{X}(\eta)) \geq \eta$ and $\mathbb{X}(F(x)) \leq x$. If F is increasing and continuous then $F(\mathbb{X}(\eta)) = \eta$. We remind the reader some basic properties about monotone funcitions that we will use below. A monotone function can have only a countable number of discontinuities and they are jump discontinuities if they exist. Moreover, the set of possible discontinuities has zero Lebesgue measure. In our case, this implies that both F and $\mathbb{X}$ have a countable number of jump discontinuities and plateaus, note that a plateau for F is a jump discontinuity for $\mathbb{X}$ and viceversa, and both sets are of zero Lebesgue measure.

Proposition 2.4. *Given $\mu \in \mathscr{P}(\mathbb{R})$, and $\mathbb{X}$ the pseudo-inverse of its distribution function F, then $\mathbb{X}\#\mathscr{L} = \mu$. Moreover, given $\mu, \nu \in \mathscr{P}(\mathbb{R})$, and $\mathbb{X}$ and $\mathbb{Y}$ their corresponding pseudo-inverses with distribution functions F and G, then the measure $\gamma_m := (\mathbb{X}, \mathbb{Y})\#\mathscr{L}$ belongs to the admissible set $\Gamma(\mu, \nu)$ and $\gamma_m((-\infty, a] \times (-\infty, b]) = min(F(a), G(b))$. Furthermore, if $\mu \in \mathscr{P}(\mathbb{R})$ is atomless, then $F\#\mu = \mathscr{L}$, and as a consequence, for every $l \in [0,1]$, the set $\{x \in \mathbb{R} : F(x) = l\}$ is μ-negligible.*

Proof. Let us clarify that by $\mathscr{L}$ we mean the Lebesgue measure on the interval of definition of the pseudo-inverses $[0,1]$. We first realize that

$$\mathscr{L}(\{\eta \in [0,1] : \mathbb{X}(\eta) \leq x\}) = \mathscr{L}(\{\eta \in [0,1] : \eta \leq F(x)\}) = F(x)\,,$$

which implies that $\mathbb{X}\#\mathscr{L}$ and μ coincide by definition on the intervals $(-\infty, x]$ for all $x \in \mathbb{R}$. Then the two measures $\mathbb{X}\#\mathscr{L} = \mu$, since this family of intervals generate the whole Borel σ-algebra on the real line. We proceed similarly to prove the second one by computing

$$\begin{aligned}\gamma_m((-\infty, a] \times (-\infty, b]) &= \mathscr{L}(\{\eta \in [0,1] : \mathbb{X}(\eta) \leq a \text{ and } \mathbb{Y}(\eta) \leq b\}) \\ &= \mathscr{L}(\{\eta \in [0,1] : \eta \leq F(a) \text{ and } \eta \leq G(b)\}) \\ &= min(F(a), G(b))\,.\end{aligned}$$

Since μ is atomless then F is a continuous function. Hence, for all $a \in (0,1)$ the set $F^{-1}((-\infty, a]) = (-\infty, x_a]$ with $F(x_a) = a$. Hence, $\mu(F^{-1}([0,a])) = F(x_a) = a$ giving the first part of the last statement. The second part is by contradiction, otherwise if one of these sets of the form $\{x \in \mathbb{R} : F(x) = l\}$ has μ positive measure, then this will mean that the Lebesgue measure should have an atom at l by the first part of the last statement.

The mass transference plan γ_m is called the monotone mass transference plan.

Proposition 2.5. *Given $\mu, \nu \in \mathscr{P}(\mathbb{R})$. Assume that $\mu \in \mathscr{P}(\mathbb{R})$ is atomless, and $\mathbb{X}$ and $\mathbb{Y}$ their corresponding pseudo-inverses with distribution functions F and G, then there exists a unique μ-a.e. defined nondecreasing map $T_m : \mathbb{R} \mapsto \mathbb{R}$ such that $\nu = T_m\#\mu$ given by $T_m = \mathbb{Y} \circ F$.*

Proof. Notice that the map $T_m = \mathbb{Y} \circ F$ is well defined with values on $\mathbb{R}$ as soon as $F(x) \in (0,1)$. Since the sets $\{x \in \mathbb{R} : F(x) = 0\}$ and $\{x \in \mathbb{R} : F(x) = 1\}$ are μ-negligible, then $T_m = \mathbb{Y} \circ F$ is well defined μ-a.e. The fact that T_m is nondecreasing is obvious by composition of nondecreasing function. Using Proposition 2.4, we have that $\mathbb{Y}\#\mathscr{L} = \nu$ and $F\#\mu = \mathscr{L}$ since μ is atomless, and thus $\nu = T_m\#\mu$ by the definition of push-forward.

Let us know prove the uniqueness part. Consider any monotone nondecreasing map T such that $\nu = T\#\mu$. From the monotonicity, we deduce that $(-\infty, x] \subset T^{-1}((-\infty, T(x)])$. Thus, we have

$$F(x) = \mu((-\infty,x]) \leq \mu(T^{-1}((-\infty,T(x)])) = \nu((-\infty,T(x)]) = G(T(x)),$$

and thus by definition of pseudo-inverse $T(x) \geq \mathbb{Y}(F(x))$. Assume that the inequality is strict now, there there exists $\varepsilon_o > 0$ such that $G(T(x) - \varepsilon) \geq F(x)$ for every $\varepsilon \in (0, \varepsilon_o)$. By monotonicity again, we have $T^{-1}((-\infty, T(x) - \varepsilon]) \subset (-\infty, x]$ and then, $G(T(x) - \varepsilon) \leq F(x)$. Thus, we get that $G(T(x) - \varepsilon) = F(x)$ for all $\varepsilon \in (0, \varepsilon_o)$. Then $F(x)$ is the value that $G(x)$ takes on an interval that is constant. We know that the set of plateaus on G is countable, so then it is the set of possible values that F takes on those intervals. The last statement in Proposition 2.4 says that each of these sets is negligible μ-a.e. and thus it is the case for a countable union of them. Therefore the set of points x such that $T(x) > \mathbb{Y}(F(x))$ is μ negligible, and thus $T = T_m$ μ-a.e. □

The first main result of this section characterizes the monotone plan and map between two one dimensional probability measures.

Proposition 2.6. *Let $\gamma \in \Gamma(\mu, \nu)$ be a transport plan between the probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R})$. Assume that it satisfies the property*

$$(x,y),(x',y') \in spt(\gamma) \text{ and } x < x' \Longrightarrow y \leq y',$$

then $\gamma = \gamma_m$. In particular, there is a unique γ satisfying the previous property.

Proof. We just need to prove that

$$\gamma((-\infty,a] \times (-\infty,b]) = min(F(a), G(b))$$

according to Proposition 2.4. Consider the sets $A = (-\infty, a] \times (b, +\infty)$ and $B = (a, +\infty) \times (-\infty, b]$. By assumption it is not possible to have both $\gamma(A) > 0$ and $\gamma(B) > 0$, otherwise we contradict the assumption. Since these two sets have empty intersection with $(-\infty, a] \times (-\infty, b]$ and at least, one of them has zero γ measure, then

$$\gamma((-\infty,a] \times (-\infty,b]) = min(\gamma((-\infty,a] \times (-\infty,b] \cup A), \gamma((-\infty,a] \times (-\infty,b] \cup B)).$$

It is easy to see that $\gamma((-\infty, a] \times (-\infty, b] \cup A) = \gamma((-\infty, a] \times \mathbb{R}) = F(a)$. Analogously for $\gamma((-\infty, a] \times (-\infty, b] \cup B) = G(b)$, obtaining the desired result. □

The previous result allows to find the solution to the transportation problem in one dimension and leads to a more general concept that characterizes optimal transport in higher dimensions, the concept of cyclical monotone sets. We just finish this section by stating the theorem without a proof that we refer to [21, 18].

Theorem 2.8. *Given two probability measures $\mu,\nu \in \mathscr{P}(\mathbb{R})$, $h : \mathbb{R} \mapsto [0,\infty)$ a strictly convex function, and the cost function of the form $c(x,y) = h(x-y)$. Assume that the Kantorovich problem associate to the cost c among these two measures is finite, i.e.*

$$I_* := \min_{\Pi\in\Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}\times\mathbb{R}} c(x,y)\, d\Pi(x,y) \right\} < +\infty .$$

Then the infimum is achieved uniquely by γ_m. If $\mu \in \mathscr{P}(\mathbb{R})$ is atomless, this optimal plan is induced by the map T_m. Moreover, if we assume plain convexity for h then γ_m is an optimal transport plan but the uniqueness is not guaranteed. Finally, in all cases the optimal cost can be expressed as

$$I_* = \int_0^1 h(\mathbb{X}(\eta) - \mathbb{Y}(\eta))\, d\eta ,$$

where $\mathbb{X}$ and $\mathbb{Y}$ are the pseudoinverses of $\mu,\nu \in \mathscr{P}(\mathbb{R})$.

Note that the previous theorem implies that the Wasserstein distance d_p, $1 \le p \le \infty$, between two one dimensional probability measures is given by the L^p-norm of the difference of their corresponding pseudoinverses functions. In particular, for the Monge-Kantorovich distance, we have

$$d_1(\mu,\nu) = \int_0^1 |\mathbb{X}(\eta) - \mathbb{Y}(\eta)|\, d\eta = \int_{\mathbb{R}} |F(x) - G(x)|\, dx,$$

(this last equality is an exercise).

Finally, let us see how to connect again to some of the PDE models we saw in the first chapter. Consider the one dimensional PDE

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(\rho V'\right), \tag{2.9}$$

with V a C^2 confinement potential such that V is uniformly convex, $V'(x) \ge \lambda > 0$, with global minimum at zero. Let us consider smooth positive probability measure solutions of (2.9), and let us denote by $F(t,x)$ the distribution function associated to the solution $\rho(t,x)$ of (2.9), and $\mathbb{X}(t,\eta)$ its pseudo inverse. By the definition of pseudo-inverse function we have

$$F(t,\mathbb{X}(t,\eta)) = \eta. \tag{2.10}$$

Differentiating (2.10) with respect to η gives

$$\frac{\partial F}{\partial x}\Big|_{x=\mathbb{X}} \frac{\partial \mathbb{X}}{\partial \eta} = 1, \tag{2.11}$$

and twice gives

$$\frac{\partial \rho}{\partial x}\Big|_{x=\mathbb{X}}\Big(\frac{\partial \mathbb{X}}{\partial \eta}\Big)^2+\rho(t,\mathbb{X})\frac{\partial^2 \mathbb{X}}{\partial \eta^2}=1. \tag{2.12}$$

Differentiating (2.10) with respect to t gives

$$\frac{\partial F}{\partial t}\Big|_{x=\mathbb{X}}+\frac{\partial F}{\partial x}\Big|_{x=\mathbb{X}}\frac{\partial \mathbb{X}}{\partial t}=0 \tag{2.13}$$

Then we collect from (2.11)-(2.12) that

$$\frac{\partial F}{\partial t}\Big|_{x=\mathbb{X}}=\int_{-\infty}^{\mathbb{X}}\frac{\partial \rho}{\partial t}(t,x)\,dx=\int_{-\infty}^{\mathbb{X}}\frac{\partial}{\partial x}\left(\rho V'\right)\,dx=\left[\rho(t,x)V'(x)\right]_{x=\mathbb{X}}=V'(\mathbb{X})\rho(t,\mathbb{X}),$$

which, in light of (2.13), leads us to the following evolution problem

$$\mathbb{X}_t=-V'(\mathbb{X}),\ \eta\in(0,1),t>0, \tag{2.14}$$

Therefore, we can deduce that if ρ_1 and ρ_2 are two such solutions of (2.9) and if their corresponding pseudoinverses are $\mathbb{X}_1$ and $\mathbb{X}_2$, then

$$\frac{1}{2}\frac{d}{dt}\int_0^1(\mathbb{X}_1(t,\eta)-\mathbb{X}_2(t,\eta))^2\,d\eta=$$
$$-\int_0^1(\mathbb{X}_1(t,\eta)-\mathbb{X}_2(t,\eta))(V'(\mathbb{X}_1(t,\eta))-V'(\mathbb{X}_2(t,\eta)))\,d\eta\,.$$

If the confinement potential V is uniformly convex, then $V''(x)\geq\lambda>0$ and using Theorem 2.8, we get

$$\frac{d}{dt}d_2^2(\rho_1(t),\rho_2(t))\leq-2\lambda d_2^2(\rho_1(t),\rho_2(t)),$$

for all $t\geq 0$. Gronwall's lemma implies that

$$d_2^2(\rho_1(t),\rho_2(t))\leq e^{-2\lambda t}d_2^2(\rho_1(0),\rho_2(0)).$$

Therefore, for a uniformly convex confinement potential, the flow of the PDE (2.9) is a contraction in the d_2 metric. Let us finally point out that $\mathbb{X}=0$ is a solution of (2.14) that corresponds to a Dirac Delta at the minimum of the potential V in original variables. In fact, one can show that δ_0 is a stationary solution to (2.9) (this statement is left as a simple exercise). Therefore, by taking the stationary solution δ_0 as one of the solutions, then $d_2^2(\rho(t),\delta_0)\leq e^{-2\lambda t}d_2^2(\rho(0),\delta_0)$ for all solutions of (2.9). So, solutions of (2.9) concentrate in infinite time to the Delta Dirac at the origin. We again see that the convexity of the potential is essential to discuss the long time asymptotics of (2.9).

We leave as an exercise to compute the equation satisfied by the pseudoinverse of the solutions of the linear Fokker-Planck equation (1.12) in one dimension and to draw conclusions about the asymptotic behavior.

We finish this section by pointing out that this approach can be turned into an effective numerical method to compute solutions of PDEs in one dimension of the general form (1.1). We showcase this in Figures 2.3 and 2.4, where we show the evolution of the pseudoinverse function associated to the solution of the PKS model (1.22) in one dimension in sub- and supercritical cases. We see how the numerical method is able to capture the diffusion of the solution in the subcritical case and the concentration in the supercritical case leading to a Dirac Delta forming in finite time according to the numerical simulations.

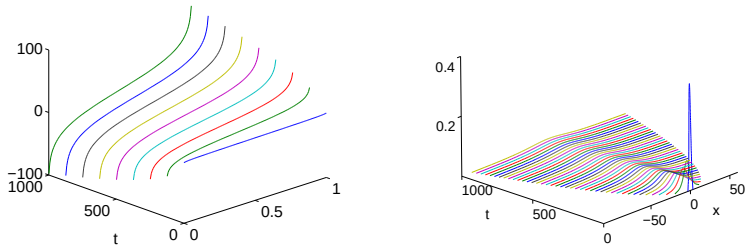

Fig. 2.3 Solution of the equation for the pseudoinverse associated to the PKS (1.22) in one dimension in a subcritical case.

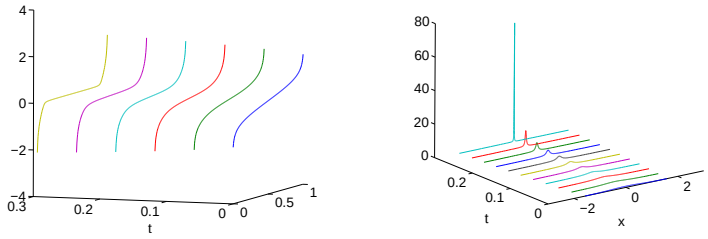

Fig. 2.4 Solution of the equation for the pseudoinverse associated to the PKS (1.22) in one dimension in a supercritical case.

Chapter 3
Mean Field Limit & Couplings

We start this chapter by studying in detail a linear continuity equation resulting from eliminating the nonlinearities in (1.1). We will focus on stability estimates for this linear equation in transport distances. We will see in the rest of this chapter how to take advantage of these estimates to derive the mean-field limit for nonlocal interaction potentials and then in next chapter we will use convexity properties of the potentials to discuss detailed properties of the gradient flows.

3.1 Measures sliding down a convex potential

Let us consider the particular case of (1.1) with $W = 0$ and $U = 0$, that is the linear continuity equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) \,, \tag{3.1}$$

for the evolution of a probability density in a velocity field given by $u = -\nabla V$ where $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a C^2 function with bounded second derivatives on $\mathbb{R}^d$ such that $D^2V(x) \geq \lambda I_d$ for $\lambda > 0$. We can assume without loss of generality that this function has a unique global minimum at zero whose value is zero by changing variables in x and adding a constant to the potential V if necessary. The first step we want to do is to give a sense to unique weak solutions to (3.1) with initial data a probability measure. In fact we will be looking for solutions to (3.1) as curves in $\rho \in C([0,T], \mathscr{P}_p(\mathbb{R}^d)$ continuous with the topology induced by d_p making $\mathscr{P}_p(\mathbb{R}^d)$ a complete metric space, $1 \leq p < \infty$. We say that $\rho \in C([0,T], \mathscr{P}_p(\mathbb{R}^d))$ is a solution to (1.1) with initial data $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ if for any $\psi \in C_o^\infty([0,\infty) \times \mathbb{R}^d)$ we have

$$\begin{aligned} \int_0^T \int_{\mathbb{R}^d} \frac{\partial \psi}{\partial t} \, d\rho(t) \, dt + \int_{\mathbb{R}^d} \psi(0) \, d\mu \\ = \int_0^T \int_{\mathbb{R}^d} (\nabla \psi \cdot \nabla V) \, d\rho(t) \, dt + \int_{\mathbb{R}^d} \psi(T) \, d\rho(T). \end{aligned} \tag{3.2}$$

Let us define $\Phi_{s,t}(x)$ to be the general solution of the finite dimensional gradient flow

$$\begin{cases} \dfrac{dr}{d\tau} = u(r) = -\nabla V(r) \ \text{ in } \ s < \tau < t, \\ r(s) = x \in \mathbb{R}^d. \end{cases}$$

Notice that due to the assumptions on V, the velocity field u is globally Lispchitz with constant $L > 0$ and the solutions exists globally since the velocity field has linear growth $|u(x)| \leq C(1+|x|)$, and thus the Cauchy-Lispchitz theory tells me that $\Phi_{s,t}(x)$ are a family of diffeomorphisms from $\mathbb{R}^d$ onto itself. Let us call the flow map associated to the finite dimensional gradient flow to the particular case of $s = 0$ that we denoted by $\Phi_t(x)$. Moreover, due to the linear growth of u, then the flow map has linear growth in x, meaning for all $T > 0$, there exists $C(T) > 0$ such that

$$\Phi_t(x) \leq C(T)(1+|x|), \qquad 0 \leq t \leq T, x \in \mathbb{R}^d.$$

Let us use a duality argument to find that solutions to (3.1) are unique and their explicit solution by the method of characteristics. In fact, let us consider the Cauchy problem

$$\begin{cases} \dfrac{\partial \psi}{\partial t} - (\nabla V \cdot \nabla \psi) = \dfrac{\partial \psi}{\partial t} + (u \cdot \nabla \psi) = 0 \ \text{ in } \ t < T, x \in \mathbb{R}^d \\ \psi(T,x) = \varphi(x) \in C_o^\infty(\mathbb{R}^d), \end{cases}$$

that has a unique classical solution given by $\psi(t,x) = \varphi(\Phi_{t,T}(x))$ by the method of characteristics. By linearity of (3.1), we are reduced to show that the unique solution to (3.2) with initial data $\mu = 0$ is the zero solution. Assume that $\mu = 0$ in (3.2) and take as test function $\psi(t,x) = \varphi(\Phi_{t,T}(x))$ in (3.2), then we deduce that

$$\int_{\mathbb{R}^d} \psi(T,x)\, d\rho(T)(x) = \int_{\mathbb{R}^d} \varphi(x)\, d\rho(T)(x) = 0,$$

for all $\varphi(x) \in C_o^\infty(\mathbb{R}^d)$, and thus $\rho(T) = 0$. Thus, the solution to (3.2) is unique. Moreover, by direct inspection we can check that $\rho(t) = \Phi_t \# \mu$ is a weak solution to (3.2) for all $T > 0$. Actually, we can obtain by Definition 2.3 of push foward that

$$\begin{aligned} \int_0^T \int_{\mathbb{R}^d} \left[\frac{\partial \psi}{\partial t} - (\nabla \psi \cdot \nabla V) \right] d\rho(t)\, dt &= \int_0^T \int_{\mathbb{R}^d} \left[\frac{\partial \psi}{\partial t} - (\nabla \psi \cdot \nabla V) \right] (t, \Phi_t(x))\, d\mu(x)\, dt \\ &= \int_0^T \int_{\mathbb{R}^d} \frac{d}{dt} \left[\psi(t, \Phi_t(x)) \right] d\mu(x)\, dt \\ &= \int_{\mathbb{R}^d} \left[\psi(T, \Phi_T(x)) - \psi(0,x) \right] d\mu(x) \\ &= \int_{\mathbb{R}^d} \psi(T,x)\, d\rho(T)(x) - \int_{\mathbb{R}^d} \psi(0,x)\, d\mu(x). \end{aligned}$$

It is an exercise to show that $\rho(t) = \Phi_t \# \mu \in C([0,T], \mathscr{P}_p(\mathbb{R}^d))$ if $\mu \in \mathscr{P}_p(\mathbb{R}^d)$. Therefore, $\rho(t) = \Phi_t \# \mu$ is the unique weak solution to (3.2) in $C([0,T], \mathscr{P}_p(\mathbb{R}^d))$ with initial data $\mu \in \mathscr{P}_p(\mathbb{R}^d)$.

Let us know see that we can obtain stability in d_1 for weak solutions to (3.1). Since the velocity field $u(x) = -\nabla V(x)$ is globally Lispchitz with constant $L > 0$, it is straightforward to show that the flow map Φ_t is also Lipschitz with constant e^{Lt}. Actually by the definition of flow map, we have

$$\Phi_t(x) - \Phi_t(y) = x - y - \int_0^t [\nabla V(\Phi_s(x)) \nabla V(\Phi_s(y))]\, ds,$$

so we can estimate it as

$$|\Phi_t(x) - \Phi_t(y)| \leq |x-y| + L \int_0^t |\Phi_s(x) - \Phi_s(y)|\, ds.$$

Gronwall's Lemma implies the claim that Φ_t is Lipschitz with constant e^{Lt}. Given $\varphi \in \mathrm{Lip}(\mathbb{R}^d)$ with $\|\varphi\|_{\mathrm{Lip}(\mathbb{R}^d)} \leq 1$, we get

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi(x)\, d(\Phi_t \# \mu_1 - \Phi_t \# \mu_2)(x) &= \int_{\mathbb{R}^d} \varphi(\Phi_t(x))\, d(\mu_1 - \mu_2)(x) \\ &= \int_{\mathbb{R}^d} (\varphi(\Phi_t(x)) - \varphi(\Phi_t(y)))\, d\Pi_o(x,y)\,, \end{aligned}$$

where $\Pi_o \in \Gamma(\mu_1, \mu_2)$ is an optimal plan for the d_1 distance. Estimating we infer that

$$\left| \int_{\mathbb{R}^d} \varphi(x)\, d(\Phi_t \# \mu_1 - \Phi_t \# \mu_2)(x) \right| \leq e^{Lt} \int_{\mathbb{R}^d} |x-y|\, d\Pi_o(x,y) = e^{Lt} d_1(\mu_1, \mu_2)\,.$$

Now, we use the characterization of the Monge-Kantorovich distance by Rubinstein-Kantorovich duality (2.8), to deduce that

$$d_1(\Phi_t \# \mu_1, \Phi_t \# \mu_2) \leq e^{Lt} d_1(\mu_1, \mu_2)\,,$$

showing the well-posedness of solutions $\rho(t) = \Phi_t \# \mu \in C([0,T], \mathscr{P}_1(\mathbb{R}^d))$ in d_1. Much more can be obtained by studying carefully the evolution of transport distances between two solutions. Let us check that we have been very rough in the previous estimates in d_1 in this particular case of measures sliding down a convex potential.

Notice that if $\mu = \delta_{x_0}$, the unique weak solutions to (3.1) is $\rho(t) = \delta_{x_0(t)}$ where $x_0(t)$ is the solution to the finite dimensional gradient flow

$$\begin{cases} \dfrac{dx_0}{dt} = -\nabla V(x_0(t)) \ \text{ in } \ t > 0 \\ x_0(0) = x_0 \in \mathbb{R}^d \end{cases}$$

Observe also that $\rho_\infty = \delta_0$ is a stationary solution since 0 is the unique minimum of the potential V. It seems intuitive that the uniform convexity of V controls the rate of convergence towards the equilibrium ρ_∞ for all weak solutions of (3.1). It is easy to check that given two solutions $x_1(t)$ and $x_2(t)$ of $\frac{dr}{d\tau} = -\nabla V(r)$, we have

$$d_2(\delta_{x_1(t)}, \delta_{x_2(t)}) \leq e^{-\lambda t} d_2(\delta_{x_1(0)}, \delta_{x_2(0)}).$$

This is left as an exercise. This shows that all Dirac Delta (particle) solutions to (3.1) converge exponentially fast to the steady state $\rho_\infty = \delta_0$. It is possible to prove a convergence much more general than this for general initial data in $\mathscr{P}_2(R^d)$.

Theorem 3.1 (Asymptotic Behavior $W = U = 0$). *Given $V \in C^2(\mathbb{R}^d)$ such that $D^2V(x) \geq \lambda I$ in $\mathbb{R}^d$ with $\lambda > 0$ and $|D^2V(x)| \leq C$ with global minimum at 0. Given any two weak solutions $\rho_1(t)$ and $\rho_2(t)$ of* (3.1) *in $C([0,T], \mathscr{P}_2(\mathbb{R}^d))$, we have*

$$d_2(\rho_1(t), \rho_2(t)) \leq e^{-\lambda t} d_2(\rho_1(0), \rho_2(0)),$$

and as a consequence,

$$d_2(\rho_1(t), \rho_\infty) = d_2(\rho_1(t), \delta_0) \leq e^{-\lambda t} d_2(\rho_1(0), \delta_0).$$

Proof. Let us take Π_o the optimal transference plan between $\rho_1(0)$ and $\rho_2(0)$ for the d_2 distance. Let us consider the two solutions $\rho_1(t)$ and $\rho_2(t)$ given by $\rho_1(t) = \Phi_t \# \rho_1(0)$ and $\rho_2(t) = \Phi_t \# \rho_2(0)$. Define $\Pi_t = (\Phi_t \times \Phi_t) \# \Pi_o$, it is clear that $\Pi_t \in \Gamma(\rho_1(t), \rho_2(t))$, then

$$d_2^2(\rho_1(t), \rho_2(t)) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Pi_t(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_t(x) - \Phi_t(y)|^2 \, d\Pi_o(x,y). \tag{3.3}$$

We claim that

$$\left. \frac{d}{dt} \right|_{0^+} d_2^2(\rho_1(t), \rho_2(t))/2 \leq -\int_{\mathbb{R}^d \times \mathbb{R}^d} (x-y) \cdot (\nabla V(x) - \nabla V(y)) \, d\Pi_o(x,y).$$

For this it suffices to justify the exchange of the integral and the time derivative on the right hand side, since we can substract

$$d_2^2(\rho_1(0), \rho_2(0)) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Pi_o(x,y)$$

in each side of (3.3), divide by t and take the limit as $t \to 0^+$. Notice that

$$\begin{aligned} \left| \frac{d}{dt} |\Phi_t(x) - \Phi_t(y)|^2 \right| &= 2|(\Phi_t(x) - \Phi_t(y)) \cdot (\nabla V(\Phi_t(x)) - \nabla V(\Phi_t(y)))| \\ &\leq C |\Phi_t(x) - \Phi_t(y)|^2 \end{aligned}$$

by the assumptions on V. Using the flow map equation, it is also easy to check that $|\Phi_t(x)| \leq C(T)|x|$ for $0 \leq t \leq T$, and thus

$$|(\Phi_t(x) - \Phi_t(y)) \cdot (\nabla V(\Phi_t(x)) - \nabla V(\Phi_t(y)))| \leq C(|x| + |y|)^2 .$$

Therefore, we can apply dominated convergence to show that

$$\frac{d}{dt}\bigg|_{0^+} \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_t(x) - \Phi_t(y)|^2 \, d\Pi_o(x,y) = -\int_{\mathbb{R}^d \times \mathbb{R}^d} (x-y) \cdot (\nabla V(x) - \nabla V(y)) \, d\Pi_o(x,y) .$$

Finally, using the uniform convexity of V we get

$$\begin{aligned} \frac{d}{dt}\bigg|_{0^+} d_2^2(\rho_1(t), \rho_2(t)) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{d}{dt}\bigg|_{0^+} |\Phi_t(x) - \Phi_t(y)|^2 \, d\Pi_o(x,y) \\ &= -2\lambda \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\Pi_o(x,y) = -2\lambda d_2^2(\rho_1(0), \rho_2(0)). \end{aligned}$$

Since this inequality derived at time 0 can be done at any arbitrary time $t_o \geq 0$, we obtain

$$\frac{d^+}{dt} d_2^2(\rho_1(t), \rho_2(t)) \leq -2\lambda \, d_2^2(\rho_1(t), \rho_2(t))$$

for all $t \geq 0$. Integrating in time, we deduce the first statement. The last part is an immediate consequence of choosing one of the solutions as the stationary solution $\rho_\infty = \delta_0$.

In summary, we have proven that for weak solutions of the (3.1), we obtain a strict contraction in d_2. Moreover, we have obtained a kind of semiflow in time in the metric space $\mathscr{P}_2(R^d)$ for solutions of the PDE (3.1) leading to the exponential convergence towards the unique steady state ρ_∞. Moreover, we could have obtained the existence of this unique steady state as the unique fixed point of this semiflows since they are strict contractions. We will now focus on this chapter in taking advantage of the rough estimate of stability in d_1 obtained above in nonlinear situations, and we will come back to the more refined estimates using the detailed structure of the equation in the last chapter about gradient flows.

3.2 Dobrushin approach: existence, stability, and derivation of the Aggregation Equation.

Let us assume in this section that $W \in C^2(\mathbb{R}^d)$ with bounded second derivatives is an interaction potential for the aggregation equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot [\rho(\nabla W * \rho)] . \tag{3.4}$$

that corresponds to (1.1) with $U = 0$ and $V = 0$. We want to show the well-posedness of solutions to (3.4) in $C([0,T],\mathscr{P}_1(\mathbb{R}^d))$ with initial data $\mu \in \mathscr{P}_1(\mathbb{R}^d)$. In order to do this, we will proceed by a fixed point argument to get the existence and uniqueness of solutions to (3.4) as a first step.

Given $\rho \in C([0,T],\mathscr{P}_1(\mathbb{R}^d))$, we define the associated velocity field $u(\rho)(t,x) = -\nabla W * \rho$. By the assumptions on W, we have that there exists $C > 0$ such that $|\nabla W(x)| \leq C(1+|x|)$ and ∇W is globally Lipschitz with Lipschitz constant $L > 0$. As a consequence, we have that the associated velocity field $u(\rho)(t,x) = -\nabla W * \rho$ satisfies

$$|u(\rho)(t,x)| \leq C \int_{\mathbb{R}^d} (1+|x-y|) d\rho(t)(y) \leq CM(\rho)(1+|x|),$$

for all $0 \leq t \leq T$ and $x \in \mathbb{R}^d$ with

$$M(\rho) := \max_{0\leq t\leq T} \int_{\mathbb{R}^d} |x| d\rho(t)(x).$$

Moreover, the velocity field is also globally Lispchitz since

$$|u(\rho)(t,x) - u(\rho)(t,y)| \leq \int_{\mathbb{R}^d} |\nabla W(x-z) - \nabla W(y-z)|\, d\rho(t)(z) \leq L|x-y|$$

for all $0 \leq t \leq T$ and $x, y \in \mathbb{R}^d$. Moreover, it is continuous in time since

$$|u(\rho)(t,x) - u(\rho)(s,x)| \leq \left| \int_{\mathbb{R}^d} \nabla W(x-y)\, d(\rho(t) - \rho(s))(y) \right| \leq L d_1(\rho(t),\rho(s)),$$

since $\nabla W(x-y)$ is Lipschitz in y with constant L and the d_1 characterization in (2.8). Since $u(\rho)$ satisfies the linear growth condition and is globally Lispchitz in x and continuous in t, we can apply the Cauchy-Lispchitz theory to have a well defined flow map associated to $u(\rho)(t,x)$ satisfying

$$\begin{cases} \dfrac{dr}{d\tau} = u(\rho)(\tau,r) = -(\nabla W * \rho(\tau))(\tau,r) \ \text{ in } \ 0 < \tau < t, \\ r(0) = x \in \mathbb{R}^d. \end{cases}$$

We denote the flow map by $\Phi_t(\rho)$. The following lemma summarizes properties of this flow map.

Lemma 3.1. *Given $\rho \in C([0,T],\mathscr{P}_1(\mathbb{R}^d))$ and their associated velocity field $u(\rho)$ and flow map $\Phi_t(\rho)$, then the following properties hold:*

i) Linear growth of the flow map: for all $T > 0$, there is constant $C(T) > 0$ depending only on $M(\rho)$ and T such that

$$|\Phi_t(\rho)(x)| \leq C(T)(1+|x|)$$

for all $0 \leq t \leq T$ and $x \in \mathbb{R}^d$.

ii) Lipschitz in x:

$$|\Phi_t(\rho)(x) - \Phi_t(\rho)(y)| \leq e^{Lt}|x-y|$$

for all $t \geq 0$ and $x, y \in \mathbb{R}^d$.

iii) Continuity in t: for all $T > 0$, there is constant $C(T) > 0$ depending only on $M(\rho)$ and T such that

$$|\Phi_t(\rho)(x) - \Phi_s(\rho)(x)| \leq C(T)(1+|x|)|t-s|$$

for all $0 \leq t, s \leq T$ and $x \in \mathbb{R}^d$.

Proof. The flow map satisfies

$$\Phi_t(\rho)(x) = x + \int_0^t u(s, \Phi_s(\rho)(x))\, ds$$

for all $0 \leq t \leq T$ and $x \in \mathbb{R}^d$. The first statement uses the linear growth of u to show

$$|\Phi_t(\rho)(x)| \leq |x| + \int_0^t |u(s, \Phi_s(\rho)(x))|\, ds \leq |x| + C(T)\int_0^t (1 + |\Phi_s(\rho)(x)|)\, ds\,.$$

A direct application of Gronwall's lemma implies the claim. Estimating again from the flow map equation, we get

$$\begin{aligned} |\Phi_t(\rho)(x) - \Phi_t(\rho)(y)| &\leq |x-y| + \int_0^t |u(s, \Phi_s(\rho)(x)) - u(s, \Phi_s(\rho)(y))|\, ds \\ &\leq |x-y| + L\int_0^t |\Phi_s(\rho)(x) - \Phi_s(\rho)(y)|\, ds \end{aligned}$$

due to the Lispchitz property of $u(\rho)$. Thus, another direct application of Gronwall's lemma leads to the second claim. The last claim follows a similar proof as the first one using the estimate on the linear growth of $u(\rho)$ and $\Phi_t(\rho)$. □

We now need an estimate between two different flow maps from two given curves in $C([0,T], \mathscr{P}_1(\mathbb{R}^d))$. We will endow the space $C([0,T], \mathscr{P}_1(\mathbb{R}^d))$ with the metric

$$\mathscr{D}_{1,T}(\rho_1, \rho_2) := \max_{0 \leq t \leq T} d_1(\rho_1(t), \rho_2(t))$$

that makes it a complete metric space for all $T > 0$.

Lemma 3.2. *Given $\rho_i \in C([0,T], \mathscr{P}_1(\mathbb{R}^d))$, and their associated velocity field $u^i = u(\rho_i)$ and flow map $\Phi_t^i = \Phi_t(\rho_i)$, then*

$$|\Phi_t^1(x) - \Phi_t^2(x)| \leq L\int_0^t e^{L(t-s)} d_1(\rho_1(s), \rho_2(s))\, ds\,,$$

for all $0 \leq t \leq T$ and $x \in \mathbb{R}^d$, and as a consequence

$$d_1(\Phi_t^1 \# \mu, \Phi_t^2 \# \mu) \leq (e^{Lt} - 1)\, \mathscr{D}_{1,T}(\rho_1, \rho_2)$$

for all $0 \le t \le T$ *and for any* $\mu \in \mathscr{P}_1(\mathbb{R}^d)$.

Proof. The flow map associated to each velocity field u^i satisfies

$$\Phi_t^i(\rho)(x) = x + \int_0^t u^i(s, \Phi_s^i(x))\, ds$$

for all $0 \le t \le T$ and $x \in \mathbb{R}^d$, $i = 1, 2$. Taking the difference and estimating we get

$$\begin{aligned}
|\Phi_t^1(x) - \Phi_t^2(x)| &\le \int_0^t |u^1(s, \Phi_s^1(x)) - u^2(s, \Phi_s^2(x))|\, ds \\
&\le \int_0^t |u^1(s, \Phi_s^1(x)) - u^1(s, \Phi_s^2(x))|\, ds \\
&\quad + \int_0^t |u^1(s, \Phi_s^2(x)) - u^2(s, \Phi_s^2(x))|\, ds \\
&\le L \int_0^t |\Phi_s^1(x) - \Phi_s^2(x)|\, ds + \int_0^t |u^1(s, \Phi_s^2(x)) - u^2(s, \Phi_s^2(x))|\, ds
\end{aligned}$$

due to the Lipschitz property of $u(\rho)$. We now proceed with the last term similarly to the continuity in time of $u(\rho)$ above. Let us denote by $z = \Phi_s^2(x)$, then

$$|u^1(s,z) - u^2(s,z)| \le \left| \int_{\mathbb{R}^d} \nabla W(z-y)\, d(\rho_1(s) - \rho_2(s))(y) \right| \le L d_1(\rho_1(s), \rho_2(s)),$$

since $\nabla W(z-y)$ is Lipschitz in y with constant L and the d_1 characterization in (2.8). Collecting terms we have obtained

$$|\Phi_t^1(x) - \Phi_t^2(x)| \le L \int_0^t |\Phi_s^1(x) - \Phi_s^2(x)|\, ds + L \int_0^t d_1(\rho_1(s), \rho_2(s))\, ds$$

for all $0 \le t \le T$ and $x \in \mathbb{R}^d$. An application of Gronwall's lemma leads to the claim (this is an exercise in the problem sheet). We now take the transference plan $(\Phi_t^1 \times \Phi_t^2)\#\mu \in \Gamma(\Phi_t^1\#\mu, \Phi_t^2\#\mu)$ as candidate transference plan to estimate

$$d_1(\Phi_t^1\#\mu, \Phi_t^2\#\mu) \le \int_{\mathbb{R}^d} |\Phi_t^1(x) - \Phi_t^2(x)|\, d\mu(x) \le L \int_0^t e^{L(t-s)} d_1(\rho_1(s), \rho_2(s))\, ds,$$

having used the first claim in the last inequality. The second statement is a direct consequence of taking the maximum outside in the last integral. □

With these ingredients, we can put together a Banach fixed point argument following a similar strategy to the Picard's theorem in the Cauchy-Lipschitz theory.

Theorem 3.2. *Given* $W \in C^2(\mathbb{R}^d)$ *with bounded second derivatives, there exists a unique global in time weak solution* ρ *in* $C([0,\infty), \mathscr{P}_1(\mathbb{R}^d))$ *to the aggregation equation* (3.4) *with initial data* $\mu \in \mathscr{P}_1(\mathbb{R}^d)$.

Proof. Let us consider $T > 0$ to be chosen later and the complete metric space $X = C([0,T], \mathscr{P}_1(\mathbb{R}^d))$ endowed with the distance $\mathscr{D}_{1,T}$. Define the map $F : X \longrightarrow X$

defined by $F(\rho) = \Phi_t(\rho)\#\mu$ with $\Phi_t(\rho)$ being the flow map associated to $u(\rho)$. By repeating the same arguments as in Section 3.1, one can check that $\tilde{\rho} = F(\rho)$ is the unique weak solution in X to the linear problem

$$\frac{\partial \tilde{\rho}}{\partial t} + \nabla \cdot [\tilde{\rho} u(\rho)] = 0, \tag{3.5}$$

with initial data $\mu \in \mathscr{P}_1(\mathbb{R}^d)$ (this last statement is left as an exercise). To show existence and local uniqueness of solution to (3.4), we are reduced to show the existence and uniqueness of a fixed point of the map F. Notice that by Lemma 3.2, the map F satisfies

$$\mathscr{D}_{1,T}(F(\rho_1), F(\rho_2)) \leq (e^{LT} - 1)\, \mathscr{D}_{1,T}(\rho_1, \rho_2)$$

and therefore by choosing T small enough depending only on L, we have that F is a strict contraction in X. By the Banach fixed point Theorem, we deduce the existence of a unique fixed point of F, and therefore of unique local solution of (3.4). Since the time of existence of this unique local solution only depends on the Lipschitz constant L, we can extend the solution recursively in a unique way for all times, as usually done in the Picard's theorem for ODEs. Details are left to be filled as an exercise. □

Let us now prove a result that is due to Dobrushin about stability of solutions leading to well-posedness for solutions to (3.4) in $\mathscr{P}_1(\mathbb{R}^d)$.

Theorem 3.3 (Dobrushin Stability Estimate). *Given $W \in C^2(\mathbb{R}^d)$ with bounded second derivatives. Let us consider two solutions ρ_i, $i = 1,2$, in $C([0,\infty), \mathscr{P}_1(\mathbb{R}^d))$ to the aggregation equation* (3.4)*, then*

$$d_1(\rho_1(t), \rho_2(t)) \leq e^{2Lt} d_1(\rho_1(0), \rho_2(0)) \tag{3.6}$$

for all $t \geq 0$.

Proof. To simplify notation, let us denote by μ and ν the initial data $\rho_1(0)$ and $\rho_2(0)$ respectively, and by $\Phi_t^i = \Phi_t(\rho_i)$ the flow maps of both solutions, $i = 1,2$. We can use Lemma 3.2 to estimate

$$\begin{aligned} d_1(\rho_1(t), \rho_2(t)) &= d_1(\Phi_t^1\#\mu, \Phi_t^2\#\nu) \leq d_1(\Phi_t^1\#\mu, \Phi_t^2\#\mu) + d_1(\Phi_t^2\#\mu, \Phi_t^2\#\nu) \\ &\leq \int_{\mathbb{R}^d} |\Phi_t^1(x) - \Phi_t^2(x)|\, d\mu(x) + d_1(\Phi_t^2\#\mu, \Phi_t^2\#\nu) \\ &\leq L \int_0^t e^{L(t-s)} d_1(\rho_1(s), \rho_2(s))\, ds + d_1(\Phi_t^2\#\mu, \Phi_t^2\#\nu) \end{aligned}$$

for all $t \geq 0$.

Given $\Pi_o \in \Gamma(\mu,\nu)$ optimal for the d_1 distance, we define the probability measure $(\Phi_t^2 \times \Phi_t^2)\#\Pi_o$. It is easy to check that $(\Phi_t^2 \times \Phi_t^2)\#\Pi_o \in \Gamma(\Phi_t^2\#\mu, \Phi_t^2\#\nu)$, and thus

$$d_1(\Phi_t^2\#\mu,\Phi_t^2\#\nu)\le\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|\Phi_t^2(x)-\Phi_t^2(y)|\,\Pi_o(x,y)|\le e^{Lt}d_1(\mu,\nu)$$

for all $t\ge 0$, due to Lemma 3.1.

Collecting terms, we deduce that $x(t)=e^{-Lt}\,d_1(\rho_1(t),\rho_2(t))$ satisfies

$$x(t)\le d_1(\mu,\nu)+L\int_0^t x(s)\,ds\,,$$

for all $t\ge 0$. Gronwall's Lemma implies the claim. □

Now that we have obtained the Dobrushin stability estimate in (3.6), we obtain as a simple consequence the uniqueness and continuous dependence of global weak solutions in $C([0,\infty),\mathscr{P}_1(\mathbb{R}^d))$ to (3.4) with respect to initial data in $\mathscr{P}_1(\mathbb{R}^d)$. Let us now discuss a different viewpoint on the aggregation equation. Let us start with N interacting particles in $\mathbb{R}^d$ following the system of ODEs

$$\frac{dX_t^i}{dt}=-\frac{1}{N}\sum_{i\neq j}^{N}\nabla W(X_t^i-X_t^j)\,, \tag{3.7}$$

with initial data X_o^i, $i=1,\dots,N$. Since $W\in C^2(\mathbb{R}^d)$ with bounded second derivatives, the system of ODEs (3.7) has a unique globally defined solution. Associated to this global solution, we can define the empirical measure

$$\mu^N(t)=\frac{1}{N}\sum_{i=1}^{N}\delta_{X_t^i}\,.$$

Let us define the velocity field associated to μ^N as $u^N(t,x)=u(\mu^N)(t,x)=-(\nabla W*\mu^N(t))(x)$. By direct inspection, one can check that $\frac{dX_t^i}{dt}=u^N(t,X_t^i)$, $i=1,\dots,N$, since W is symmetric implies that $\nabla W(0)=0$. Moreover, if the associated flow maps are denoted by $\Phi_t^N=\Phi_t(\mu^N)$, then $X_t^i=\Phi_t^N(X_o^i)$, $i=1,\dots,N$. It is left as an exercise to check that $\mu^N\in C([0,\infty),\mathscr{P}_1(\mathbb{R}^d))$ is the unique weak solution to (3.4) with initial data $\mu^N(0)$. We have just proved the following result.

Corollary 3.1 (Empirical measures). *Given $W\in C^2(\mathbb{R}^d)$ with bounded second derivatives and any initial data of the form*

$$\mu^N(0)=\frac{1}{N}\sum_{i=1}^{N}\delta_{X_o^i}\,.$$

with X_o^i, $i=1,\dots,N$. Then the unique weak solution in $C([0,\infty),\mathscr{P}_1(\mathbb{R}^d))$ to (3.4) *with initial data in $\mu^N(0)$ is given by*

$$\mu^N(t)=\frac{1}{N}\sum_{i=1}^{N}\delta_{X_t^i}\,,$$

where X_t^i, $i = 1,\dots,N$, is the unique solution to (3.7) *with initial data X_o^i, $i = 1,\dots,N$.*

For our equation (3.4), the empirical measures are weak solutions to the equation (3.4) with "particles " initial data for all N. Using now the Dobrushin stability estimate (3.6), we can reinterpret this estimate as a proof of derivation of the PDE (3.4) from the particle dynamics (3.7). This is precisely the question of mean-field limit problem: given the dynamics of particles specified by the system (3.7), can we identify the limit as $N \to \infty$ of their empirical measures as the probability measure ρ of finding particles at a particular location x at time t? If so, can we identify the law giving the evolution of ρ? In other words, can we identify ρ as the solution of a PDE? If this is possible, it is said that the PDE obtained is the mean-field PDE associated to the dynamical system (3.7). The name of mean-field comes from the intuition that in this scaling limit, with respect to N, particles in (3.4) feel a mean velocity field associated to many particles in the limit $N \to \infty$.

Corollary 3.2 (Mean Field Limit). *Given $W \in C^2(\mathbb{R}^d)$ with bounded second derivatives and take a sequence of empirical measures initially of the form*

$$\mu^N(0) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_o^i}.$$

with X_o^i, $i = 1,\dots,N$, such that $d_1(\mu^N(0),\mu) \to 0$ as $N \to \infty$ with $\mu \in \mathscr{P}_1(\mathbb{R}^d)$. Define the sequence of empirical measures $\mu^N(t)$ by

$$\mu^N(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i},$$

where X_t^i, $i = 1,\dots,N$, is the unique solution to (3.7) *with initial data X_o^i, $i = 1,\dots,N$. Then $d_1(\mu^N(t),\rho(t)) \to 0$, for all $t > 0$, as $N \to \infty$ with ρ being characterized as the unique weak solution in $C([0,\infty),\mathscr{P}_1(\mathbb{R}^d))$ to* (3.4) *with initial data in μ.*

Proof. This result is a direct application of the Dobrushin estimate in Theorem 3.3 for the solutions given by the empirical measure $\mu^N(t)$ and the solution with initial data $\mu \in \mathscr{P}_1(\mathbb{R}^d)$ given by Theorem (3.2). Actually, (3.6) implies

$$d_1(\mu^N(t),\rho(t)) \le e^{2Lt} d_1(\mu^N(0),\rho(0)) = e^{2Lt} d_1(\mu^N(0),\mu). \tag{3.8}$$

Since the right-hand side of (3.8) converges to 0 as $N \to \infty$ by assumption, the left-hand side does so too finishing the proof. □

Remark 3.1. In order to have a full proof of the mean-field derivation, one needs to show that the set of empirical measures is dense on $\mathscr{P}_1(\mathbb{R}^d)$ that one can find in [14, Subsection 1.4.4].

These Dobrushin stability estimates can be generalized to the case of SDEs. In fact, given the Langevin equations

$$dX_t^i = -\frac{1}{N}\sum_{i\neq j}^{N} \nabla W(X_t^i - X_t^j)dt + \sqrt{2\sigma}\, dB_t^i, \tag{3.9}$$

where B_t^i, $i = 1,\dots,N$, are N independent Brownian motions. Now, it is more difficult to analyse the correlations between the particles and what is the PDE, if any, that gives the typical behavior of one of the particles as $N \to \infty$. In fact, one can define the empirical measures associated to the Langevin system (3.9) but they are no longer solutions of a PDE in $\mathbb{R}^d$. They are random variables in the set of probability measures. However, an approach by stability estimates is possible when $W \in C^2(\mathbb{R}^d)$ with bounded second derivatives. Sznitmann introduced in [7] the so-called coupling method based on stability estimates to be able to derive the mean-field limit for (3.9). He showed that the mean-field limit of (3.9) is characterized by the solution of the McKean-Vlasov equation (1.14) that we recall here:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot [\rho(\nabla W * \rho)] + \sigma \Delta \rho.$$

The details of the proof can be found in [7] for the interested reader that we do not pursue here due to lack of time.

3.3 Boltzmann Equation in the Maxwellian approximation: Tanaka Theorem.

Let us model the evolution of the statistical ensemble in velocity of a system of point particles colliding inelastically and assumed homogeneous in space. The microscopic dynamics can be described with the following hypotheses:

1. The particles interact via *binary* collisions. More precisely, the gas is rarefied enough so that collisions between 3 or more particles can be neglected.
2. These binary collisions are localized in space and time. In particular, all the particles are considered as point particles, even if they describe macroscopic objects.
3. Collisions preserve mass and momentum, but dissipate a fraction $1-e$ of the kinetic energy in the impact direction, where the inelasticity parameter $e \in [0,1]$ is called *restitution coefficient*:

$$\begin{cases} v' + v'_* = v + v_*, \\ |v'|^2 + |v'_*|^2 - |v|^2 - |v_*|^2 = -\dfrac{1-e^2}{2}\,|(v-v_*)\cdot\omega|^2 \le 0, \end{cases} \tag{3.10}$$

with $\omega \in \mathbb{S}^{d-1}$ being the impact direction.

Remark 3.2. Taking $e = 1$ in both (3.11) and (3.12) yields the classical energy-conservative elastic collision dynamics, as illustrated in Fig. 3.1. Notice the possible confusion of notation between the restitution coefficient e and the number

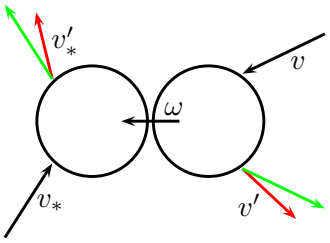

Fig. 3.1 Geometry of the inelastic collision in the physical space (green is elastic, red is inelastic).

e. I decided to keep it as it is since this is the standard notation in books about granular materials and inelastic Boltzmann equations. To make it clear in the statements in this section, I will use a different notation for the exponential function.

Using these conservations, one has the following two possible parametrizations (see also Fig. 3.2) of the post-collisional velocities, as a function of the pre-collisional ones:

- The ω-representation or reflection map, given for $\omega \in \mathbb{S}^{d-1}$ by

$$\begin{aligned} v' &= v - \frac{1+e}{2}\left((v-v_*)\cdot\omega\right)\omega, \\ v'_* &= v_* + \frac{1+e}{2}\left((v-v_*)\cdot\omega\right)\omega. \end{aligned} \tag{3.11}$$

- The σ-representation or swapping map, given for $\sigma \in \mathbb{S}^{d-1}$ by

$$\begin{aligned} v' &= \frac{v+v_*}{2} + \frac{1-e}{4}(v-v_*) + \frac{1+e}{4}|v-v_*|\sigma, \\ v'_* &= \frac{v+v_*}{2} - \frac{1-e}{4}(v-v_*) - \frac{1+e}{4}|v-v_*|\sigma. \end{aligned} \tag{3.12}$$

The geometry of collisions is more complex than the classical elastic collisions case. Indeed, fixing $v, v_* \in \mathbb{R}^3$, denote by

$$\Omega_\pm := \frac{v+v_*}{2} \pm \frac{1-e}{4}(v_*-v), \qquad O := \frac{v+v_*}{2} = \frac{v'+v'_*}{2}.$$

Then if $u := v - v_*$ is the *relative velocity*, one has

$$|\Omega_- - v'| = |\Omega_+ - v'_*| = \frac{1+e}{4}|u|,$$

namely $v' \in \mathscr{S}\left(\Omega_-, |u|(1+e)/4\right)$ and $v'_* \in \mathscr{S}\left(\Omega_+, |u|(1+e)/4\right)$, where $\mathscr{S}(x,r)$ is the sphere centered in x and of radius r (see also Fig. 3.2).

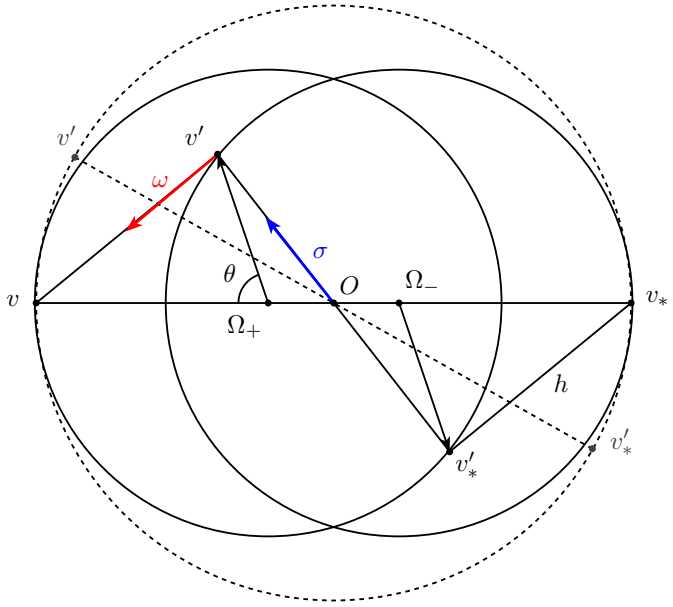

Fig. 3.2 Geometry of the inelastic collision in the phase space (dashed lines represent the elastic case).

Using the microscopic hypotheses (1–2–3), one can define an inelastic collision operator $Q_e(f,f)$ acting on a probability density of particles $f(t,v)$ in its weak form as

$$\int_{\mathbb{R}^d} Q_e(f,f)(v)\,\psi(v)\,dv = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} B f_* f\,\left(\psi' - \psi\right)\,d\sigma\,dv\,dv_*, \tag{3.13}$$

where the collision kernel is typically of the form $B(|u|,\cos\theta) = \Phi(|u|)b(\cos\theta)$, and θ is the angle between σ and u. The Maxwell simplification in the modelling is to assume that the collision frequency of particles is just constant. We will assume in the rest of this section that $B = 1$. We here follow the notation $f = f(v)$, $f_* = f(v_*)$, $f' = f(v')$, and $f'_* = f(v'_*)$ for simplicity. In the Maxwellian approximation, the inelastic collision operator $Q_e(f,f)$ simplies to $Q_e(f,f) = Q_e^+(f,f) - f$ with $Q_e^+(f,f)$ defined by duality as the probability measure satisfying

$$\int_{\mathbb{R}^3} Q_e^+(f,f)(v)\,\psi(v)\,dv = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} f_* f\,\psi'\,d\sigma\,dv\,dv_*. \tag{3.14}$$

Notice that this definition makes sense for any probability measure f. We are interested in properties of the solutions to the homogeneous Boltzmann equation in the Maxwellian approximation given by

$$\frac{\partial f}{\partial t} = Q_e(f,f) = Q_e^+(f,f) - f. \tag{3.15}$$

The basic properties of solutions to (3.15) are conservation of mass and mean velocity and disipation of the kinetic energy.

In this section, we will analyse the behavior of solutions to (3.15) as curves of probability measures in velocity space. Observe though that we will use the notation as if they were densities as it is customary in kinetic equations. We wil not attempt to develop a full well-posedness theory of solutions in these notes but let us focus in understanding the main properties of the gain part of the collision operator that one can use to build the theory of well-posedness and to study the long-time asymptotic properties of the solutions.

Let us first reinterpret the gain operator: given a probability measure f on $\mathbb{R}^3$, the gain operator is the probability measure $Q_e^+(f,f)$ defined by

$$(\varphi, Q_e^+(f,f)) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f(v)\, f(v_*)\, (\varphi, \mathscr{U}_{v,v_*})\, dv\, dv_*$$

where $\mathscr{U}_{v,v_*}$ is the uniform probability distribution on the sphere $\mathscr{S}(\Omega_-, |u|(1+e)/4)$. In probabilistic terms, the gain operator is defined as an expectation:

$$Q_e^+(f,f) = \mathbb{E}[\mathscr{U}_{V,V_*}]$$

where V and V_* are independent random variables with law f.

Theorem 3.4 (Contraction of $Q_e^+(f,f)$ in d_2). *Given f and g in $\mathscr{P}_2(\mathbb{R}^3)$ with equal mean velocity, then*

$$d_2(Q_e^+(f,f), Q_e^+(g,g)) \le \sqrt{\frac{3+e^2}{4}}\, d_2(f,g).$$

Proof. The main steps of the proof can be summarized as follows: Let us take two independent pairs of random variables (V,X) and (W,Y) such that V and W have law f and X and Y have law g.

Step 1.- Convexity of d_2^2 in Theorem 2.3 implies

$$d_2^2(Q_e^+(f,f), Q_e^+(g,g)) = d_2^2(\mathbb{E}[\mathscr{U}_{V,W}], \mathbb{E}[\mathscr{U}_{X,Y}]) \le \mathbb{E}\left[d_2^2(\mathscr{U}_{V,W}, \mathscr{U}_{X,Y})\right]. \tag{3.16}$$

Here, the independency of the pairs of random variables has been used.

Step 2.- The d_2^2 distance between the uniform distributions on the sphere with center O and radius r, $\mathscr{U}_{O,r}$, and on the sphere with center O' and radius r', $\mathscr{U}_{O',r'}$, in $\mathbb{R}^3$ is bounded by $|O'-O|^2 + (r'-r)^2$.

This is an estimate over the euclidean cost of transporting one sphere onto the other made by explicitly constructing a transport map T, $\mathscr{U}_{O',r'} = T\#\mathscr{U}_{O,r}$. Then, the transport plan $\Pi_T = (1_{\mathbb{R}^d} \times T)\#\mathscr{U}_{O,r}$ given by

$$\iint_{\mathbb{R}^3\times\mathbb{R}^3} \eta(v,w)\, d\Pi_T(v,w) = \int_{\mathbb{R}^3} \eta(v,T(v))\, d\mathscr{U}_{O,r}(v)$$

for all test functions $\eta(v,w)$, is used in the definition of d_2 to conclude

$$d_2^2(\mathscr{U}_{O,r},\mathscr{U}_{O',r'}) \leq \int_{\mathbb{R}^3} |v-T(v)|^2\, d\mathscr{U}_{O,r}(v). \tag{3.17}$$

Precisely, we define the map $T:\mathbb{R}^3 \longrightarrow \mathbb{R}^3$ transporting the sphere of center O and radius $r>0$ onto the sphere with center $O'\neq O$ and radius $r'>r$ in the following way: consider the point $\Omega\in\mathbb{R}^3$ given by

$$\Omega = O - \frac{r}{r'-r}(O'-O).$$

Then we let T be the dilation with factor $\frac{r'}{r}$ centered at Ω, that is, we let $T(v) = \Omega + \frac{r'}{r}(v-\Omega)$. The other cases, $O'=O$ or $r'=r$, are done by simple translations or dilations. We show in Figure 3.3 a sketch of the construction of the map T in the case of non-interior spheres.

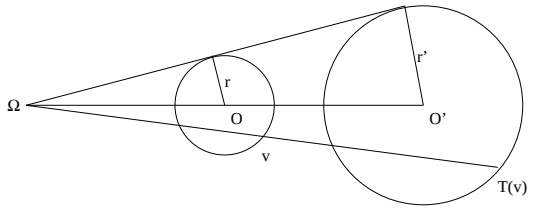

Fig. 3.3 Scheme of the transport map between spheres.

Inserting this definition of the map T in (3.17), we deduce

$$d_2^2(\mathscr{U}_{O,r},\mathscr{U}_{O',r'}) \leq \left(\frac{r'-r}{r}\right)^2 \int_{\mathbb{R}^3} |v-\Omega|^2\, d\mathscr{U}_{O,r}(v)$$

that can be computed explicitly, giving

$$d_2^2(\mathscr{U}_{O,r},\mathscr{U}_{O',r'}) \leq |O'-O|^2 + (r'-r)^2$$

and finishing the proof.

Step 3.- We now estimate the right-hand side of (3.16) by using the formulas of the center Ω_- and radii of the spheres given above to deduce

$$\begin{aligned} d_2^2(Q_e^+(f,f),Q_e^+(g,g)) \leq{}& \frac{5-2e+e^2}{8}\,\mathbb{E}\left[|V-X|^2\right] + \frac{(1+e)^2}{8}\,\mathbb{E}\left[|W-Y|^2\right] \\ &+ \frac{1-e^2}{4}\,\mathbb{E}\left[(V-X)\cdot(W-Y)\right] \end{aligned}$$

where the Cauchy-Schwartz inequality has been used.

Step 4.- Finally, we take both pairs (V,X) and (W,Y) as independent pairs of variables with each of them being an optimal couple for the $d_2(f,g)$ to obtain

$$\begin{aligned} d_2^2(Q_e^+(f,f),Q_e^+(g,g)) &\leq \frac{3+e^2}{4}\, d_2^2(f,g) + \frac{1-e^2}{4}\,\mathbb{E}\left[(V-X)\cdot(W-Y)\right] \\ &= \frac{3+e^2}{4}\, d_2^2(f,g)\,, \end{aligned}$$

due to independency and having equal mean velocity. □

As a consequence, one can deduce the following property for solutions of the Boltzmann equation in the maxwellian approximation (3.15). We assume here the existence and uniqueness of solutions to (3.15) as continuous curves in $\mathscr{P}_2(\mathbb{R}^3)$ that can be obtained similarly to the previous section based on the estimates on the contraction of d_2 in the previous theorem (this is left as an exercise for $e \in [0,1)$). Since the mean velocity of solutions to (3.15) is conserved, we can assume without loss of generality that solutions have zero mean velocity.

Theorem 3.5 (Contraction in d_2). *If f_1 and f_2 are two solutions to* (3.15) *with respective initial data f_1^0 and f_2^0 in $\mathscr{P}_2(\mathbb{R}^3)$ and zero mean velocity, then*

$$d_2(f_1(t),f_2(t)) \leq \exp(-\alpha t)\, d_2(f_1^0,f_2^0)$$

for all $t \geq 0$ with $\alpha = \frac{1-e^2}{8}$.

Proof. Duhamel's formula for (3.15) reads as

$$f_i(t) = \exp(-t)\, f_i^0 + \int_0^t \exp(-(t-s))\, Q_e^+(f_i(s),f_i(s))\, ds, \qquad i = 1,2.$$

As before, the convexity of the squared Wasserstein distance in Theorem 2.3 and the contraction of the gain operator in Theorem 3.4 imply

$$\begin{aligned} d_2^2(f_1(t),f_2(t)) &\leq \exp(-t)\, d_2^2(f_1^0,f_2^0) \\ &\quad + \int_0^t \exp(-(t-s))\, d_2^2\big(Q_e^+(f_1(s),f_1(s)),Q_e^+(f_2(s),f_2(s))\big)\, ds \\ &\leq \exp(-t)\, d_2^2(f_1^0,f_2^0) + \frac{3+e^2}{4}\int_0^t \exp(-(t-s))\, d_2^2(f_1(s),f_2(s))\, ds. \end{aligned}$$

Gronwall's lemma concludes the proof. □

Notice that Theorem 3.5 does not give a strict contraction for the classical Boltzmann equation for Maxwell molecules when $e = 1$. However, one can improve this result by studying the cases of equality in the contraction estimate showing that in fact one converges in d_2 to the Maxwellian equilibria in d_2. This together with the non strict contraction is called the Tanaka theorem for the Boltzmann equation.

Chapter 4
An introduction to Gradient Flows

This chapter is devoted to a brief and partly informal introduction to gradient flows in the space of probability measures. The objective is to illustrate by means of the most basic examples the main ideas of this approach. This is complemented by formal computations for developing some of the intuitions for applications of this theory in many areas of modelling from biological problems to problems in big data or social sciences.

4.1 Brenier's Theorem and Dynamic Interpretation of optimal transport.

Let us consider $u(t,x)$ a bounded smooth vector field in $\mathbb{R}^d$ meaning that u is bounded and globally Lispchitz in x and continuous in t, we can apply the Cauchy-Lispchitz theory to have a well defined flow map associated to $u(t,x)$ satisfying

$$\begin{cases} \dfrac{dr}{ds} = u(s,r) \qquad \text{in } \ 0 \le s \le 1, \\ r(0) = x \in \mathbb{R}^d. \end{cases}$$

We denote the flow map by Φ_t. Reproducing the proof in the first section of Chapter 3, one can show that given $\rho_0 \in \mathscr{P}_2(\mathbb{R}^d)$, the unique weak solution in $C([0,1],\mathscr{P}_2(\mathbb{R}^d))$ of the continuity equation

$$\partial_s \rho + \nabla \cdot (\rho u) = 0 \qquad \text{in } (0,1)\times\mathbb{R}^d \tag{4.1}$$

is given by $\rho(s) = \Phi_s \# \rho_0 \in C([0,1],\mathscr{P}_2(\mathbb{R}^d))$. Given a pair of a curve of probability measures and a velocity field (ρ,u) satisfying the continuity equation (4.1) in the distributional sense, we can define its action as

$$\mathscr{A}[\rho,u] := \int_0^1 \int_{\mathbb{R}^d} |u(s,x)|^2 \, d\rho(s)(x)\, ds\,.$$

The following remarkable formula is due to Benamou and Brenier giving an alternative characterization of the d_2 distance in terms of the path joining two probability measures through the continuity equation (4.1) with the minimal kinetic energy.

Theorem 4.1. *Given probability measures $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$, then*

$$d_2^2(\mu,\nu) = \inf_{(\rho,u)} \left\{ \mathscr{A}[\rho,u] \,:\, (\rho,u) \text{ satisfying (4.1) and } \rho_0 = \mu \text{ and } \rho_1 = \nu \right\}$$

holds.

Proof. We give a formal proof since a full proof goes beyond the scope of this introductory course. Given an admissible pair (ρ,u) of a curve in $C([0,1],\mathscr{P}_2(\mathbb{R}^d))$ and a smooth velocity field satisfying (4.1), $\rho_0 = \mu$, and $\rho_1 = \nu$, we discussed above that the curve can be expressed in terms of the flow map as $\rho(s) = \Phi_s \# \rho_0$. In particular, Φ_1 is a transport map from μ to ν. By definition of push-forward and Holder's inequality we obtain

$$\begin{aligned}
\mathscr{A}[\rho,u] &= \int_0^1 \int_{\mathbb{R}^d} |u(s,x)|^2 \, d\rho(s)(x)\, ds = \int_0^1 \int_{\mathbb{R}^d} |u(s,\Phi_s(x))|^2 \, d\rho_0(x)\, ds \\
&= \int_0^1 \int_{\mathbb{R}^d} \left| \frac{d\Phi_s(x)}{ds} \right|^2 d\rho_0(x)\, ds = \int_{\mathbb{R}^d} \int_0^1 \left| \frac{d\Phi_s(x)}{ds} \right|^2 ds\, d\rho_0(x) \\
&\geq \int_{\mathbb{R}^d} \left| \int_0^1 \frac{d\Phi_s(x)}{ds}\, ds \right|^2 d\rho_0(x) = \int_{\mathbb{R}^d} |\Phi_1(x) - x|^2 \, d\rho_0(x) \geq d_2^2(\mu,\nu)\,.
\end{aligned}$$

Hence, $d_2^2(\mu,\nu)$ is less or equal than the infimum in the statement.

To show equality, assume that the target measure $\mu \ll \mathscr{L}$, then we can use Brenier's Theorem 2.6 to have a well defined transport map leading to the optimal cost for $d_2^2(\mu,\nu)$, i.e., $\nu = T\#\mu$ and the optimal transference plan is of the form $\Pi_o = (1_{\mathbb{R}^d} \times T)\#\mu \in \Gamma(\mu,\nu)$. Then, define $T_s(x) = (1-s)x + sT(x)$ and choose the velocity field such that $\frac{dT_s(x)}{ds} = u(s,T_s(x))$. Then, one can easily checked that all the above inequalities become identities, and thus the infimum is achieved. We leave as an exercise to show that $u(s,x)$ is well defined by proving that $T_s(x)$ in invertible and Lispchitz for $0 \leq s < 1$ using that $T = \nabla\varphi$ with φ convex. □

This dynamic interpretation of the transport distance has been crucial both from the theoretical and numerical viewpoints. It has led to connections to fluid mechanics and to computational transport tools based on optimization methods and numerical approximation of PDEs. It is also crucial to interpret the family of general PDE (1.1) as gradient flows as we will see in the last section. It was a key element for an interpretation of the tangent to a curve of probability measures as introduced by Otto in the seminal work [16] and the nowadays known as Otto's calculus. We do not have time to cover this aspect of the theory.

4.2 McCann's Displacement Convexity: Internal, Interaction and Confinement Energies.

We will start by constructing geodesics between probabilty measures with general transport distances d_p, $1 \leq p < \infty$.

Lemma 4.1. *[Geodesics] Given probability measures $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$, $1 \leq p < \infty$. Given $\Pi \in \Gamma(\mu,\nu)$ an optimal plan for $d_p(\mu,\nu)$, define $\mu_t = \mathscr{T}_t \# \Pi$ with the map $\mathscr{T}_t : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ given by $\mathscr{T}_t(x,y) = (1-t)x + ty$. Then, μ_t is a constant speed d_p geodesic curve joining μ to ν, that is,*

$$d_p(\mu_s,\mu_t) = |t-s| d_p(\mu_0,\mu_1) = |t-s| d_p(\mu,\nu) \quad \textit{for all } 0 \leq s \leq t \leq 1\,.$$

Proof. Since $\Pi \in \Gamma(\mu,\nu)$, it is obvious that $\mu_0 = \mu$ and $\mu_1 = \nu$. Moreover, taking the plan $\Pi_{s,t} = (\mathscr{T}_s, \mathscr{T}_t)\#\Pi \in \Gamma(\mu_s,\mu_t)$, we get

$$\begin{aligned} d_p^p(\mu_s,\mu_t) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, d\Pi_{s,t}(x,y) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} |\mathscr{T}_s(x) - \mathscr{T}_t(y)|^p \, d\Pi(x,y) = |t-s|^p d_p^p(\mu,\nu). \end{aligned}$$

If we use this estimate now on the intervals $[0,s]$, $[s,t]$ and $[t,1]$, we get

$$d_p(\mu,\mu_s) + d_p(\mu_s,\mu_t) + d_p(\mu_t,\nu) \leq (s + t - s + 1 - t) d_p(\mu,\nu) = d_p(\mu,\nu)\,.$$

Notice that the reverse inequality is always true due to the triangular inequality since d_p is a metric, then all the inequalities in between must be equalities, and thus the claim of the Lemma is true. □

Notice that any optimal coupling for $d_p(\mu,\nu)$ generates a constant speed geodesic joining the measures. In case the optimal transference plan is given by an optimal map as in Brenier's Theorem 2.6, i.e., the optimal transference plan is of the form $\Pi_o = (1_{\mathbb{R}^d} \times T)\#\mu \in \Gamma(\mu,\nu)$, then the geodesic is given by $\mu_t = T_t \# \mu$ with $T_t(x) = (1-t)x + tT(x)$ and $\nu = T\#\mu$. In other words, the geodesic is obtained by pushing-forward the density through the linear interpolant of the identity map and the optimal transport map T between the measures μ and ν. Remember we already discussed an application of these interpolants between measures in "image processing" in Figure 2.2.

We have already seen that convexity properties of functionals are very important to understand the dynamics of PDEs of the form (1.1) in various particular cases. Based on the geodesics in transport distances, we can now introduce a notion of convexity that plays an important role in the understanding of gradient flows in probability measures as we will see in the next section.

Definition 4.1 (Displacement Convexity). We say that a functional $\mathscr{F} : \mathscr{P}_2(\mathbb{R}^d) \longrightarrow \mathbb{R} \cup \{+\infty\}$ is d_2-convex or displacement convex, if the one dimensional function

$\mathscr{F}[\mu_t]$ is convex in $t \in [0,1]$ for all d_2 geodesics μ_t joining any two measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, that is,

$$\mathscr{F}[\mu_t] \leq (1-t)\mathscr{F}[\mu_0] + t\mathscr{F}[\mu_1]$$

for all d_2 geodesics μ_t.

Assume that $U : [0,\infty) \to \mathbb{R}$ is a $C([0,\infty),\mathbb{R}) \cap C^2((0,\infty),\mathbb{R})$ function with $U(0) = 0$, $V : \mathbb{R}^d \to \mathbb{R}$ is a bounded below confinement potential and $W : \mathbb{R}^d \to \mathbb{R}$ is a bounded below interaction potential as defined in Chapter 1. Associated to the PDE (1.1), we define the following functionals: internal, confinement, and interaction energy $\mathscr{U}, \mathscr{V}, \mathscr{W} : \mathscr{P}_2(\mathbb{R}^d) \longrightarrow \mathbb{R} \cup \{+\infty\}$, as

$$\mathscr{U}(\rho) = \int_{\mathbb{R}^d} U(\rho)\, dx, \tag{4.2}$$

if $\mu \ll \mathscr{L}$ with density ρ and $\mathscr{U} = +\infty$ otherwise,

$$\mathscr{V}[\mu] = \int_{\mathbb{R}^d} V(x)\, d\mu(x), \tag{4.3}$$

and

$$\mathscr{W}[\rho] = \frac{1}{2}\int_{\mathbb{R}^d \times \mathbb{R}^d} W(x-y)\, d\mu(x)\, d\mu(y). \tag{4.4}$$

Lemma 4.2 (Convexity of confinement and interaction energies). *If $V : \mathbb{R}^d \to \mathbb{R}$ is convex then the confinement energy $\mathscr{V}$ is d_2-convex. If $W : \mathbb{R}^d \to \mathbb{R}$ is convex then the confinement energy $\mathscr{W}$ is d_2-convex. Morever, if V is strictly convex then $\mathscr{V}$ is strictly d_2-convex., and if W is strictly convex then $\mathscr{W}$ is strictly d_2-convex unless the geodesic joining the measures is a translation of a given measure.*

Proof. Given probability measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, any d_2-geodesic joining them is of the form $\mu_t = \mathscr{T}_t \# \Pi$ with $\Pi \in \Gamma(\mu,\nu)$ an optimal plan for $d_2(\mu,\nu)$. We can now compute the values of the confinement and interaction energies on the geodesic to get

$$\mathscr{V}[\mu_t] = \int_{\mathbb{R}^d} V(x)\, d\mu_t(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} V((1-t)x + ty)\, d\Pi(x,y)$$

and

$$\begin{aligned}\mathscr{W}[\mu_t] &= \frac{1}{2}\int_{\mathbb{R}^d \times \mathbb{R}^d} W(x-y)\, d\mu_t(x)\, d\mu_t(y) \\ &= \frac{1}{2}\int_{\mathbb{R}^d \times \mathbb{R}^d}\int_{\mathbb{R}^d \times \mathbb{R}^d} W((1-t)(x-y) + t(z-w))\, d\Pi(x,z)\, d\Pi(y,w).\end{aligned}$$

Using the convexity of V and W in the integrands above implies immediately the first statements of the lemma. The strictly convex claims are an exercise. □

We now focus on the internal energy. Since the internal energy is infinite unless the measure is absolutely continuous with respect to the Lebesgue measure, we

can reduce to the case of a geodesic joining two absolutely continuous measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ with respect to Lebesgue, otherwise there is nothing to prove. In this case, we can use Brenier's Theorem 2.6 to write $\nu = \nabla\varphi\#\mu$ with φ a convex function and the geodesic as $\mu_t = T_t\#\mu$ with $T_t(x) = (1-t)x + t\nabla\varphi(x)$. Let us denote by $\rho_t(x)$ the density of the measure μ_t with respect to Lebesgue. If the convex function φ were C^2 and the map $\nabla\varphi$ a diffeomorphism, we could use change of variables to write the internal energy as

$$\begin{aligned}\mathscr{U}[\mu_t] &= \int_{\mathbb{R}^d} U(\rho_t(x))\,dx = \int_{\rho_t>0} \frac{U(\rho_t(x))}{\rho_t(x)}\rho_t(x)\,dx \\ &= \int_{\rho_0>0} \frac{U((\rho_t\circ T_t)(x))}{(\rho_t\circ T_t)(x)}\rho_0(x)\,dx = \int_{\rho_0>0} U\left(\frac{\rho_0(x)}{\det\nabla T_t(x)}\right)\det\nabla T_t(x)\,dx \\ &= \int_{\mathbb{R}^d} U\left(\frac{\rho_0(x)}{\det((1-t)I_d + tD^2\varphi(x))}\right)\det((1-t)I_d + tD^2\varphi(x))\,dx\,.\end{aligned}$$

In these identities, we formally used the change of variables formula in the push-forward as Exercise 5 in Problem Sheet 1. Now, convex analysis again comes to help us. Since φ is a convex function, it is differentiable almost everywhere and it has distributional second derivatives in the Aleksandrov sense with a hessian matrix $D^2\varphi$ that is a symmetric and nonnegative measure. Moreover, the previous change of variables formula makes sense, we refer to [21, Chapter 4] for further details. Using the notation $D(x,t) = \det((1-t)I_d + tD^2\varphi(x))^{1/d}$, we have finally shown that

$$\mathscr{U}[\mu_t] = \int_{\mathbb{R}^d} U\left(\frac{\rho_0(x)}{D(x,t)^d}\right) D(x,t)^d\,dx.$$

We leave as an exercise to show the following lemma.

Lemma 4.3. *Let Λ be a nonnegative symmetric matrix and $v(t) = \det((1-t)I_d + t\Lambda)^{1/d}$. Then v is concave on $t \in [0,1]$ and strictly concave unless $\Lambda = \lambda I_d$.*

Applying this to $D(x,t)$, we deduce that $D(x,t)$ is concave in t for all $x \in \mathbb{R}^d$. Moreover, defining $G(x,s) = s^d U(\rho_0(x)s^{-d})$ for $s > 0$, we can write the internal energy of the geodesic as

$$\mathscr{U}[\mu_t] = \int_{\mathbb{R}^d} G(x, D(x,t))\,dx.$$

Assume that the function $g(s) = s^d U(s^{-d})$, $s > 0$, is convex and nonincreasing, then it is left as an exercise to show that the map $t \longrightarrow G(x, D(x,t))$ is a convex function in t for all $x \in \mathbb{R}^d$, and thus the internal energy of the geodesic is convex in t. We have shown the so-called McCann's condition for displacement convexity of the internal energy.

Theorem 4.2. *[McCann's condition] Assume $U : [0,\infty) \to \mathbb{R}$ is a $C([0,\infty),\mathbb{R}) \cap C^2((0,\infty),\mathbb{R})$ function with $U(0) = 0$ such that $s^d U(s^{-d})$, $s > 0$, is convex and nonincreasing, then the internal energy $\mathscr{U}$ is d_2-convex.*

Particular important choices of internal energies satisfying the McCann's condition are the Boltzmann entropy with $U(s) = s\log(s)$ and the power-law case $U(s) = s^m$ for all $m \geq 1 - \frac{1}{d}$, $m \neq 1$. We leave as an exercise to check that U satisfies the McCann's condition if and only if $P(s) \geq 0$ and $(1 - \frac{1}{d})P(s) \leq sP'(s)$ for all $s > 0$ with P defined from U by $sU''(s) = P'(s)$ and $P(0) = 0$.

We will now learn how to obtain these conditions from the dynamic interpretation seen in the previous section in a formal way by computing optimality conditions. Given $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, we first obtain the optimality condition for the geodesic equations in the fluid dynamical formulation of the d_2 distance by the Benamou-Brenier formula in Theorem 4.1. We insert the continuity equation

$$\partial_s \rho + \nabla \cdot (\rho u) = 0 \qquad \text{in } (0,1) \times \mathbb{R}^d \tag{4.5}$$

inside the minimization problem as a Lagrange multiplier in its weak form using a test function $\psi \in C_o^\infty([0,1] \times \mathbb{R}^d)$. As a result, we get the unconstrained minimization problem

$$\begin{aligned}\frac{1}{2} d_2^2(\mu,\nu) = \inf_{(\rho,u)} \sup_{\psi} \Big\{ & \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} |u_s(x)|^2 \rho_s(x) \, \mathrm{d}x \mathrm{d}s \\ & - \int_0^1 \int_{\mathbb{R}^d} \left[\rho_s(x) \partial_s \psi(s,x) + \rho_s(x)(u_s(x) \cdot \nabla \psi(s,x)) \right] \, \mathrm{d}x \mathrm{d}s \\ & + \int_{\mathbb{R}^d} \psi(1,x) \, \mathrm{d}\rho_1(x) - \int_{\mathbb{R}^d} \psi(0,x) \, \mathrm{d}\rho_0(x) \Big\}.\end{aligned}$$

The factor $1/2$ is for convenience for the computations below. Applying a formal minimax principle, swapping infimum and supremum, and thus taking first an infimum with respect to u we obtain the optimality condition $u = \nabla \psi$, and the following formal characterization of the distance

$$\begin{aligned}\frac{1}{2} d_2^2(\mu,\nu) = \sup_{\psi} \inf_{\rho} \Big\{ & - \frac{1}{2} \int_0^1 \int_{\mathbb{R}^d} |\nabla \psi|^2 \rho \, \mathrm{d}x \mathrm{d}s - \int_0^1 \int_{\mathbb{R}^d} \rho \partial_s \psi \, \mathrm{d}x \mathrm{d}s \\ & + \int_{\mathbb{R}^d} \psi(1,x) \, \mathrm{d}\rho_1(x) - \int_{\mathbb{R}^d} \psi(0,x) \, \mathrm{d}\rho_0(x) \Big\},\end{aligned}$$

which provides the further optimality condition

$$\partial_s \psi + \frac{1}{2} |\nabla \psi|^2 = 0. \tag{4.6}$$

We thus end up with a coupled system of differential equations in $(0,1) \times \mathbb{R}^d$ as optimality conditions for the geodesic curves in the dynamic formulation

$$\begin{cases} \partial_s \rho + \nabla \cdot (\rho \nabla \psi) = 0, \\ \partial_s \psi + \frac{1}{2} |\nabla \psi|^2 = 0, \end{cases} \tag{4.7}$$

together with the boundary conditions $\rho_0 = \mu$ and $\rho_1 = \nu$. Let us now use the formal equations (4.7) for the geodesics associated to the distance d_2 to compute the conditions under which the internal energy functional is displacement convex. Assuming (ρ_s, ψ_s) is a smooth solution of (4.7), which decays sufficiently at infinity, we proceed by integration by parts to obtain the following formulas:

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathscr{U}[\rho] = -\int_{\mathbb{R}^d} P(\rho)\Delta\psi \ \mathrm{d}x,$$

with P defined by $P'(r) = rU''(r)$ and $P(0) = 0$. Furthermore, one can further compute to obtain

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}s^2}\mathscr{U}[\rho] = &\int_{\mathbb{R}^d} (P'(\rho)\rho - P(\rho))(\Delta\psi)^2 \ \mathrm{d}x \\ &+ \int_{\mathbb{R}^d} P(\rho)(-\nabla\psi\cdot\nabla\Delta\psi + \frac{1}{2}\Delta|\nabla\psi|^2) \ \mathrm{d}x.\end{aligned}$$

The Bochner formula implies that

$$\begin{aligned}-\nabla\psi\cdot\nabla\Delta\psi + \frac{1}{2}\Delta|\nabla\psi|^2 &= |D^2\psi|^2 = \mathrm{trace}\left[(D^2\psi)^2\right] \\ &\geq \frac{1}{d}(\Delta\psi)^2 = \frac{1}{d}\left[\mathrm{trace}(D^2\psi)\right]^2,\end{aligned}$$

the last inequality using the fact that $D^2\psi$ is a symmetric matrix. Assuming that $P(\rho) \geq 0$, we can estimate it as

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\mathscr{U}[\rho] \geq \int_{\mathbb{R}^d} (P'(\rho)\rho - (1-1/d)P(\rho))(\Delta\psi)^2 \ \mathrm{d}x.$$

Therefore, under the displacement McCann's condition, $P(s) \geq 0$ and $(1-\frac{1}{d})P(s) \leq sP'(s)$ for all $s > 0$ with P defined from U by $sU''(s) = P'(s)$ and $P(0) = 0$, the functional $\mathscr{U}$ is convex along the geodesics of the distance d_2 based on these formal computations. This is a very useful procedure to guess the convexity properties of functionals. Let us do a similar computation for the confinement energy $\mathscr{V}$. The formulas of the first and second derivatives along geodesics satisfying (4.7) are

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathscr{V}[\rho] = \int_{\mathbb{R}^d} \rho\nabla V\cdot\nabla\psi \ \mathrm{d}x,$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\mathscr{V}[\rho] = \int_{\mathbb{R}^d} \rho\,(D^2V\,\nabla\psi)\cdot\nabla\psi \ \mathrm{d}x.$$

Again, we observe that if V is convex, $D^2V \geq 0$ and we have displacement convexity of $\mathscr{V}$. Moreover, this computation shows that if V is 2-uniform convex, i.e., $D^2V \geq \lambda I_d$ for $\lambda > 0$, then

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\mathscr{V}[\rho] \geq \lambda \int_{\mathbb{R}^d} \rho\,|\nabla\psi|^2\ \mathrm{d}x = \lambda d_2^2(\mu,\nu)\,.$$

This leads to a definition of 2-uniform displacement convexity for functionals in probability measures. We leave as exercise to use the same procedure for analysing the convexity of the interaction energies $\mathscr{W}$.

In fact, this approach can be generalized to find formal optimality conditions for many different variants of transport distances defined by the dynamical formulation introduced in Theorem 4.1. Examples of these variants are nonlinear continuity equations with different mobility functions, nonlocal mobilities, fractional diffusions, the Landau equation in plasma physics, the relativistic heat equation and many other partial differential equations can be connected in this way to variants of these transport distances.

4.3 Gradient Flows: the differential viewpoint.

Before attempting to construct an abstract argument in a context fraught with perils of nonsmoothness, infinite dimensions, and degenerate convexity, it is instructive to recall basic ideas about gradient flows. The setting will be so simple that not only are the results well-known, they could all be deduced by a good sophomore calculus student. Fix $E \in C^2(\mathbb{R}^d)$ and consider solutions of the ordinary differential equation

$$\frac{dx_t}{dt} = -\nabla E(x_t) \tag{4.8}$$

corresponding to *steepest descent* or *gradient flow* on the energy (entropy) landscape determined by E. Solutions satisfy

$$\frac{d}{dt}E(x_t) = -|\nabla E(x_t)|^2\,, \tag{4.9}$$

and thus the energy decays along the curves x_t solutions to (4.8). Moreover, the energy E is a strict Liapunov functional in the sense that $\frac{d}{dt}E(x_t) = 0$ if and only if x_t is a critical point of E.

Lemma 4.4 (Bounding contraction / expansion rates). *Fix $\lambda \in \mathbb{R}$. If $E \in C^2(\mathbb{R}^d)$ satisfies $D^2E(x) \geq \lambda I_d$ throughout $\mathbb{R}^d$, and the curves x_t and $t \in [0,\infty) \longrightarrow y_t \in \mathbb{R}^d$ both solve the differential equation* (4.8)*, then $|x_t - y_t| \leq e^{-\lambda t}|x_0 - y_0|$.*

Proof. Set $f(t) = |x_t - y_t|^2/2$. Then

$$\begin{aligned} f'(t) &= - < x_t - y_t, \nabla E(x_t) - \nabla E(y_t) > \\ &= - < x_t - y_t, \int_0^1 D^2E[(1-s)x_t + sy_t]\,(y_t - x_t)ds > \leq -2\lambda f(t)\int_0^1 ds. \end{aligned}$$

Gronwall's inequality (integration) implies the desired result: $f(t) \leq e^{-2\lambda t} f(0)$. □

Corollary 4.1 (Contraction in a convex valley). *Taking $\lambda = 0$ in the preceding proposition implies $|x_t - y_t|$ is monotone nonincreasing as a function of $t \in [0,\infty)$.*

Proof. Obviously $|x_t - y_t| \le |x_0 - y_0|$. Since the equation is autonomous, time translation invariance implies $|x_{T+t} - y_{T+t}| \le |x_T - y_T|$ for all $t, T \ge 0$. □

If $\lambda > 0$, more can be achieved. The convexity of E is said to be 2-*uniform*, and we have shown that the solution map $x_0 \in \mathbb{R}^d \longrightarrow X_t(x_0) = x_t$ of the initial value problem (4.8) defines a uniform contraction on $\mathbb{R}^d$ for each $t > 0$. The C^2 smoothness of E ensures that the solution map is well-defined locally in space and time; the map is globally defined for all future times since x_t is constrained to lie in the level set $\{x \mid E(x) \le E(x_0)\}$, whose compactness follows from the coercivity of $E(x) \ge E(x_0) + < \nabla E(x_0), x - x_0 > + \lambda |x - x_0|^2/2$. Since $\mathbb{R}^d$ is complete, the contraction mapping principle dictates that this map has a unique fixed point $X_t(x_\infty) = x_\infty \in \mathbb{R}^d$, and each solution curve $x_t = X_t(x_0)$ must converge to x_∞ in the long time limit $t \to \infty$. The quantity estimated is the decay rate of the slope $|\nabla E(x_t)| \to 0$, that we can call the information.

Lemma 4.5 (Entropy production and information decay rate). *Let $E \in C^2(\mathbb{R}^d)$ satisfy $D^2E(x) \ge \lambda I_d > 0$ throughout $\mathbb{R}^d$. Then any solution $t \in [0,\infty) \longrightarrow x_t \in \mathbb{R}^d$ of* (4.8) *satisfies $|\nabla E(x_t)| \le e^{-\lambda t} |\nabla E(x_0)|$.*

Proof. Let $f(t) := |\nabla E(x_t)|^2/2$. Then

$$f'(t) = < \nabla E(x_t), D^2E(x_t)\, \dot{x}_t > = - < \nabla E(x_t), D^2E(x_t) \nabla E(x_t) > \le -2\lambda f(t),$$

and Gronwall's inequality proves the desired estimate: $f(t) \le e^{-2\lambda t} f(0)$. □

While the conclusions of these two lemmas are not immediately comparable, the following consequence (4.10) of 2-uniform convexity relates them. It shows that information dominates the altitude or *relative entropy* $E(x) - E(x_\infty)$, which in turn dominates horizontal distance squared. Thus in its limited range of validity — $\lambda > 0$ and $y_t := x_\infty$ — and apart from constants, Proposition 4.5 trumps Proposition 4.4. On the other hand, (4.11) also shows that if information remains bounded, then convergence in the weakest sense, namely of distance (unsquared), also implies convergence in the stronger sense of relative entropy.

Lemma 4.6 (Manifestations of 2-uniform convexity). *Let $0 \le f \in C^2(\mathbb{R})$ satisfy $f(0) = 0$ and $f''(s) > \lambda > 0$ for all $s \in \mathbb{R}$. Then $\lambda s^2 \le 2f(s) \le \lambda^{-1} |f'(s)|^2$ and*

$$f(s) \le s f'(s) - \lambda s^2/2.$$

Proof. Let $g(s) := f(s) - \lambda s^2/2$. Taking two derivatives shows $g(s)$ is convex, so its critical point at the origin must be a minimum: $g(s) \ge g(0) = 0$. This proves the first inequality.

Since $f(s) \ge 0$ is strictly convex, its minimum $f(0) = 0$ is its only critical point. Defining $h(s) := |f'(s)|^2/2 - \lambda f(s)$, we see $h'(s) = f'(s)(f''(s) - \lambda)$ can vanish only where $f'(s)$ does — namely, at zero. Since $h''(0) = f''(0)(f''(0) - \lambda) + 0 f'''(0) >$

0, the unique critical point of $h(s)$ is a strict local minimum; it must be a global minimum since the absence of other critical points ensures that monotonicity of $h(s)$ changes only at zero. Thus $h(s) \geq h(0) = 0$, which establishes the second inequality.

Finally, let $e(s) = sf'(s) - \lambda s^2/2 - f(s)$. Then $e'(s) = s(f''(s) - \lambda)$ vanishes only when $s = 0$. A second derivative $e''(0) = f''(0) - \lambda > 0$ shows this unique critical point of $e(s)$ to be a strict local minimum, hence a global minimum as above: $e(s) \geq e(0) = 0$ to complete the proof of the lemma. □

Corollary 4.2 (Cartoon Log Sobolev, Talagrand, and HWI inequalities). *Suppose $E(x_\infty) \leq E(x) \in C^2(\mathbb{R}^d)$ and $D^2E(x) \geq \lambda I_d > 0$ for all $x \in \mathbb{R}^d$. Then*

$$\frac{\lambda}{2}|x - x_\infty|^2 \;\leq\; E(x) - E(x_\infty) \leq \frac{1}{2\lambda}|\nabla E(x)|^2 \tag{4.10}$$

$$\text{and} \qquad E(x) - E(x_\infty) \leq |x - x_\infty||\nabla E(x)| - \lambda|x - x_\infty|^2/2. \tag{4.11}$$

As a consequence, any solution $t \in [0,\infty) \longrightarrow x_t \in \mathbb{R}^d$ of (4.8) satisfies $E(x_t) - E(x_\infty) \leq e^{-2\lambda t}(E(x_0) - E(x_\infty))$.

Proof. The conclusions of the lemma continue to hold under the relaxed hypothesis $f(s) \geq \lambda$, as is easily seen by replacing λ with $\lambda - 1/n$ and taking a limit $n \to \infty$. Given $x \in \mathbb{R}^d$, the function $f(s) := E(x_\infty + s\frac{x - x_\infty}{|x - x_\infty|}) - E(x_\infty)$ satisfies the hypothesis $f''(s) \geq \lambda$. Setting $s = |x - x_\infty|$ in the conclusion of the lemma, Cauchy-Schwarz yields the desired inequalities (4.10–4.11). Notice that (4.10) together with (4.9) and Gronwall's lemma leads to the exponential decay of the relative energy. □

These intuitions can be applied to particular cases of the general PDE equation (1.1). In particular, let us consider the case of the linear Fokker-Planck equation (1.12) with $W = 0$, $P(\rho) = \sigma\rho$ and V such that $D^2V(x) \geq \lambda I_d$ with $\lambda > 0$, that is,

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \sigma \Delta \rho, \tag{4.12}$$

We already proved in Section 1.2 that

$$d_2(\rho_1(t), \rho_2(t)) \leq e^{-\lambda t} d_2(\rho_1(0), \rho_2(0))$$

for any two solutions of (4.12) based on the definition of d_2 in terms of random variables. However, the deeper reason is that this equation has a "gradient flow" structure in the following sense. Defining the total free energy of the system as

$$\mathscr{F}[\rho] = \sigma \int_{\mathbb{R}^d} \rho \log \rho \, dx + \int_{\mathbb{R}^d} V(x)\rho \, dx, \tag{4.13}$$

we can compute formal variations of the functional around a density $\rho \in L^1_+(\mathbb{R}^d)$ by taking perturbations in the set

$$\mathscr{S} := \{\bar{v} \in L^1(\mathbb{R}^d) \text{with zero mean such that } \rho + \varepsilon\bar{v} \geq 0 \text{ for } \varepsilon > 0 \text{ small enough}\}.$$

By doing so and assuming the necessary conditions to apply the dominated convergence theorem, we obtain

$$\lim_{\varepsilon\to 0}\frac{\mathscr{F}[\rho+\varepsilon\bar{v}]-\mathscr{F}[\rho]}{\varepsilon}=\int_{\mathbb{R}^d}\frac{\delta\mathscr{F}}{\delta\rho}(\rho)\bar{v}\,dx$$

with $\frac{\delta\mathscr{F}}{\delta\rho}(\rho):=\sigma\log(\rho)+V$. Therefore, the linear Fokker-Planck equation can be written as

$$\begin{cases}\partial_s\rho+\nabla\cdot(\rho u)=0 \ \text{ in } \ (0,\infty)\times\mathbb{R}^d\\ u=-\nabla\dfrac{\delta\mathscr{F}}{\delta\rho}\end{cases}, \tag{4.14}$$

where we eliminated the dependence of the variation of $\mathscr{F}$ in ρ to ease the notation, we will do so in the sequel when there is no confusion. The free energy (4.13) is a Liapunov functional for (4.12) since

$$\frac{d}{dt}\mathscr{F}[\rho]=-I[\rho]:=-\int_{\mathbb{R}^d}\left|\nabla\frac{\delta\mathscr{F}}{\delta\rho}\right|^2\rho(x)\,dx, \tag{4.15}$$

at least by formal integration by parts. Notice that this identity resembles the decay of the energy E in the finite dimensional case, and thus the right hand side should be dissipated by the squared norm of the gradient of the energy if this were a real gradient flow. Here, we observe the connection to the dynamical interpretation of the squared Wassertein distance d_2 in Theorem 4.1. The right hand side is the opossite of the kinetic energy associated to the vector field $-\nabla\frac{\delta\mathscr{F}}{\delta\rho}$. As mentioned in the previous section, this is the starting point for a much deeper connection formally introduced by Otto in [16]. The resemblance of the linear Fokker-Planck equation (4.12) to the case of 2-uniform gradient flows in finite dimensions goes further. As we showed in Section 1.2, the function

$$\rho_\infty(x)=\frac{1}{Z}e^{-V(x)/\sigma}\qquad\text{with } Z=\int_{\mathbb{R}^d}e^{-V(x)/\sigma}\,dx,$$

is a steady state of (4.12). Notice that $\frac{\delta\mathscr{F}}{\delta\rho}(\rho_\infty)=\sigma\log(\rho_\infty)+V$ is constant. Therefore, we can define the relative free energy as $\mathscr{F}(\rho|\rho_\infty)=\mathscr{F}[\rho]-\mathscr{F}[\rho_\infty]$ that satisfies

$$\mathscr{F}[\rho|\rho_\infty]=\sigma\int_{\mathbb{R}^d}\eta\log\eta\,\rho_\infty\,dx,$$

with $\eta=\rho/\rho_\infty$. A simple application of Jensen's inequality with respect to the Gaussian measure ρ_∞ using that $x\log x$ is convex gives

$$\mathscr{F}[\rho]-\mathscr{F}[\rho_\infty]\geq\sigma\left(\int_{\mathbb{R}^d}\eta\,\rho_\infty\,dx\right)\log\left(\int_{\mathbb{R}^d}\eta\,\rho_\infty\,dx\right)=0$$

and the equality holds if and only if $\eta=1$. Therefore, the Gaussian ρ_∞ is the global minimum of the functional $\mathscr{F}[\rho]$. Moreover, due to the results of the previous sec-

tion, the free energy functional $\mathscr{F}[\rho]$ is 2-uniform d_2 displacement convex since it is the sum of a displacement convex functional with a 2-uniform displacement functional. So, we are in the best of the worlds, a uniforly convex functional with a global minimum but in the displacement sense in the Wasserstein metric space.

A very detailed theory for gradient flows for 2-uniform displacement convex functionals has been developed [3, 22, 2]. The theory of 2-uniform d_2 gradient flows applies and it leads to the following conclusions for the linear Fokker-Planck equation. One can show the following functional inequalities known as the Log-Sobolev, the Talagrand, and the HWI inequalities:

$$\mathscr{F}[\rho|\rho_\infty] \leq \frac{1}{2\lambda} I[\rho], \tag{4.16}$$

$$d_2(\rho,\rho_\infty) \leq \sqrt{\frac{2}{\lambda}\mathscr{F}[\rho|\rho_\infty]}, \tag{4.17}$$

and

$$\mathscr{F}[\rho|\rho_\infty] \leq d_2(\rho,\rho_\infty)\sqrt{I[\rho]} - \frac{\lambda}{2} d_2^2(\rho,\rho_\infty). \tag{4.18}$$

The name of HWI for (4.18) comes from the H-theorem of the Boltzmann entropy, W for Wasserstein distance, and I for the Fisher information functional $I[\rho]$ as coined by C. Villani. All of them are manifestations of the uniform convexity in this infinite dimensional setting and they correspond to the results in Corollary 4.2 in disguise. These inequalities imply directly a convergence rate towards the steady state ρ_∞ in relative entropy and d_2 sense. Just make use of (4.16) in (4.15) to deduce that

$$\frac{d}{dt}\mathscr{F}[\rho|\rho_\infty] = -I[\rho] \leq -2\lambda\mathscr{F}[\rho|\rho_\infty]$$

giving by Gronwall's Lemma the exponential decay $\mathscr{F}[\rho|\rho_\infty] \leq e^{-2\lambda t}\mathscr{F}[\rho_0|\rho_\infty]$, and then Talagrand's inequality (4.17) to deduce the exponential decay of $d_2(\rho,\rho_\infty)$. However, the 2-uniform displacement convexity of $\mathscr{F}[\rho]$ implies a further consequence, the uniform contraction in d_2 for solutions of (4.12) as proven in Section 1.2. This is a general property for this type of gradient flows [3, 22]. Finally, notice that the general family of PDEs (1.1) introduced in the first chapter of this course can be written formally in the form of a gradient flow as in (4.14) with the free energy given by $\mathscr{F}[\rho] = \mathscr{U}[\rho] + \mathscr{V}[\rho] + \mathscr{W}[\rho]$ for a suitable function U related to to P, we leave this as an exercise.

4.4 Gradient Flows: the metric viewpoint

Let us come back to the case of gradient flows in $\mathbb{R}^d$. Given $E \in C^2(\mathbb{R}^d)$, we consider the gradient flow $\frac{dx_t}{dt} = -\nabla E(x_t)$ for which solutions satisfy (4.9)

$$\frac{d}{dt}E(x_t) = -|\nabla E(x_t)|^2 .$$

The previous formula encodes important ingredients for gradient flows. We observe the enery decays the fastest at each point x_t of the trajectory, since the energy decays the fastest in the direction $-\nabla E(x_t)$ at x_t. To even being able to write this we need the notion of gradient of a function. The theory of gradient flows can be generalized to Hilbert spaces [5]. However, when we want to generalize this theory to the case of metric spaces we do not have this notion of gradient defined properly. Even more our velocity fields might not be even C^1 as we saw in the case of the Barenblatt solution to the porous medium equation in Chapter 1. Therefore, a different generalization of gradient flows is needed. A classical way to construct solutions to the gradient flow (4.8) is by discretizing in time via the implicit Euler scheme: given a time step Δt and an approximation to the solution at time $t_k = k\Delta t$, we find the approximation at time t_{k+1} by solving

$$x_{k+1} = x_k - \Delta t \nabla E(x_{k+1}) .$$

It is easy to see that this identitiy is nothing else than the critical point condition for the following functional

$$E_k(x) = \frac{1}{2\Delta t}|x - x_k|^2 + E(x) ,$$

that is, x_{k+1} is a critical point of the function E_k. Therefore, a natural way to construct x_{k+1} is by looking for a global minimizer of the energy E_k. Assume now that E is a convex function in $\mathbb{R}^d$, then the critical point is equivalent to

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{1}{2\Delta t}|x - x_k|^2 + E(x) \right\}$$

since E_k is uniformly convex for all $\Delta t > 0$ and all k. The previous variational characterization of the implicit Euler scheme for gradient flows of convex functions is useful in two ways: the smoothness of E is only needed to characterize critical points and it encodes again the steepest descent primary property of gradient flows. As mentioned before, the smoothness assumption E in C^1 is very strong, so let us take a function $E : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{\infty\}$ such that is convex and lower semicontinuous. Associated to a convex function, we can define its subdifferential as the set of directions determining all supporting hyperplanes, that is

$$\partial E(x) := \left\{ y \in \mathbb{R}^d : E(z) \geq E(x) + < y, z - x > \text{ for all } z \in \mathbb{R}^d \right\} .$$

Recall that absolutely continuous functions on an interval are a.e. differentiable with respect to Lebesgue. Now we have the ingredients to generalize the notion of gradient flow solution.

Definition 4.2. An absolutely continuous curve $x : [0,\infty) \longrightarrow \mathbb{R}$ is a gradient flow solution with initial data $x(0)$ for the convex and lower semicontinuous energy functional $E : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{\infty\}$ if

$$-\frac{dx}{dt} \in \partial E(x(t)) \text{ for a.e. } t > 0\,.$$

It is not difficult to generalize the variational characterization of the implicit Euler scheme to this setting. This is left as an exercise, that is,

$$-\frac{x_{k+1}-x_k}{\Delta t} \in \partial E(x_{k+1})$$

if and only if

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

This formulation can be easily generalized to Hilbert spaces and it is easy to show that the variational scheme above is well defined, i.e., it has a minimum for all k and Δt. The dificulty lies in showing the convergence as $\Delta t \to 0$ to a solution in the sense of Definition 4.2. We refer to [5, 2] for more details. Let us do a classical example in Hilbert spaces, take $\mathscr{H} = L^2(\mathbb{R}^d)$ and define the energy functional

$$\mathscr{E}[u] := \begin{cases} \dfrac{1}{2}\displaystyle\int_{\mathbb{R}^d} |\nabla u|^2\,dx & \text{if } u \in H^1(\mathbb{R}^d) \\ +\infty & \text{otherwise} \end{cases}. \tag{4.19}$$

It is convex in the classical sense and lower semicontinuous by standard results of weak convergence. We claim that $\partial\mathscr{E}(x) \neq \emptyset$ if and only if $\Delta u \in L^2(\mathbb{R}^d)$, and in that case $\partial\mathscr{E}(x) = \{-\Delta u\}$. Assume $p \in \partial\mathscr{E}(x)$, that is, $p \in L^2(\mathbb{R}^d)$ such that $\mathscr{E}(v) \geq \mathscr{E}(u) + < p, v-u >_2$ with an obvious notation for the L^2-scalar product. Take $v = u + \varepsilon w$ with $w \in H^1(\mathbb{R}^d)$, then this inequality leads to

$$\frac{1}{2}\int_{\mathbb{R}^d} |\nabla u + \varepsilon\nabla w|^2\,dx - \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2\,dx \geq \varepsilon \int_{\mathbb{R}^d} p(x)w(x)\,dx\,.$$

Rearranging the terms and taking $\varepsilon \to 0$, we get

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w\,dx \geq \int_{\mathbb{R}^d} p(x)w(x)\,dx\,,$$

for all $w \in H^1(\mathbb{R}^d)$. Taking $-w$ in the previous inequality, we conclude that

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w\,dx = \int_{\mathbb{R}^d} p(x)w(x)\,dx\,,$$

for all $w \in H^1(\mathbb{R}^d)$, and thus by definition $\Delta u \in L^2(\mathbb{R}^d)$ and $-\Delta u = p$. We leave the converse argument as exercise. Therefore, we can properly say that the heat equation is the gradient flow of the Dirichlet energy (4.19) with respect to the L^2 scalar product. One can actually show that the heat equation is the limit of the implicit Euler scheme defining a piecewise continuous interpolants in time for the L^2-functions constructively obtained by the interative variational scheme

$$u_{k+1} = \arg\min_{x\in L^2(\mathbb{R}^d)} \left\{ \frac{1}{2\Delta t}\|u-u_k\|^2_{L^2(\mathbb{R}^d)} + \mathscr{E}(x) \right\},$$

and showing their convergence to the heat equation as $\Delta t \to 0$. This approach was introduced by Jordan, Kiderlehrer and Otto [15] to derive the linear Fokker-Planck equation (4.12) as the steepest descent of the free energy (4.13) in the Wasserstein d_2 sense. More precisely, they showed that by defining a sequence of measures iteratively by

$$\rho_{k+1} = \arg\min_{\rho\in\mathscr{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\Delta t} d_2^2(\rho,\rho_k) + \mathscr{F}[\rho] \right\}$$

for any fixed $\Delta t > 0$ and any $k \in \mathbb{N}$ starting with a given measure $\rho_0 \in \mathscr{P}_2(\mathbb{R}^d)$, then a suitable interpolant in time of these measures leads to a curve of measures $\rho_{\Delta t}$ converging to the unique solution of (4.12) with initial data $\rho_0 \in \mathscr{P}_2(\mathbb{R}^d)$ as $\Delta t \to 0$. We do not have time to cover this proof in this course, but let us at least show that the variational scheme is well defined in a simpler setting. Let us consider Ω a bounded smooth domain of $\mathbb{R}^d$, and let us take as energy functional just the Boltzmann entropy functional, that is,

$$\mathscr{E}[\rho] = \int_\Omega \rho\log\rho\, dx.$$

The formal d_2-gradient flow of $\mathscr{E}$ is the heat equation in Ω with Neumann boundary conditions. Let us finally show that the variational scheme

$$\begin{aligned}\rho_{k+1} &= \arg\min_{\rho\in\mathscr{P}_2(\Omega)} \left\{ \frac{1}{2\Delta t} d_2^2(\rho,\rho_k) + \mathscr{E}[\rho] \right\} \\ &= \arg\min_{\rho\in\mathscr{P}_2(\Omega)} \left\{ \frac{1}{2\Delta t} d_2^2(\rho,\rho_k) + \int_\Omega \rho\log\rho\, dx \right\}\end{aligned}$$

is well defined for any fixed $\Delta t > 0$ starting from $\rho_0 \in \mathscr{P}(\Omega)$.

Lemma 4.7. *Given $\Delta t > 0$, for any $k \in \mathbb{N}$ the functional*

$$\mathscr{E}_k[\rho] := \frac{1}{2\Delta t} d_2^2(\rho,\rho_k) + \mathscr{E}[\rho]$$

has a minimum in $\mathscr{P}(\Omega)$ for a given $\rho_k \in \mathscr{P}(\Omega)$.

Proof. Taking a suitably normalized Gaussian as ρ and since $x\log x \geq -1$, it is clear that the functional $\mathscr{E}_k$ has a finite infimum in $\mathscr{P}(\Omega)$. Take a minimizing sequence $\rho_n \in \mathscr{P}(\Omega)$ that is

$$\frac{1}{2\Delta t} d_2^2(\rho_n,\rho_k) + \mathscr{E}[\rho_n] \to I := \inf_{\rho\in\mathscr{P}_2(\Omega)} \left\{ \frac{1}{2\Delta t} d_2^2(\rho,\rho_k) + \mathscr{E}[\rho] \right\}.$$

Using again $x\log x \geq -1$ and that Ω is bounded, we deduce that the sequence

$$\int_\Omega \rho_n \log^+ \rho_n \, dx \leq C$$

for $n \in \mathbb{N}$, where $\log^+(x) = \max(0, \log(x))$. Given $M \in \mathbb{N}$, denote by $\rho_n \wedge M := \min(\rho_n, M)$. The sequence of cut-off functions $\rho_n \wedge M$ is bounded in $L^\infty(\Omega)$, thus by Banach-Alaoglu theorem, it is weakly-* compact in $L^\infty(\Omega)$ for each given $M \in \mathbb{N}$. By a standard diagonal argument, we can extract a subsequence, denoted with the same index to simplify the notation, such that $\rho_n \wedge M \rightharpoonup \rho_M$ weakly-$*$ in $L^\infty(\Omega)$ for all $M \in \mathbb{N}$. Define $\bar{\rho} = \sup_M \rho_M$. Notice that $\rho_n \wedge M \leq \rho_n \wedge (M+1)$, and therefore their weak-$*$ limits are also ordered, so ρ_M is an increasing sequence in M a.e. $x \in \mathbb{R}^d$. By monotone convergence theorem, we deduce that

$$\int_\Omega \rho_M \, dx \to \int_\Omega \bar{\rho} \, dx \leq \infty.$$

Furthermore, since $\rho_n \wedge M \rightharpoonup \rho_M$ weakly-$*$ in $L^\infty(\Omega)$, we have the convergence testing against $L^1(\Omega)$ functions, in particular against the constant 1 so

$$\int_\Omega \rho_n \wedge M \, dx \to \int_\Omega \rho_M \, dx,$$

as $n \to \infty$, and since $\rho_n \in \mathscr{P}_2(\Omega)$, then ρ_M is bounded in $L^1(\Omega)$, and in particular $\bar{\rho} \in L^1(\Omega)$ and

$$\|\rho_M - \bar{\rho}\|_{L^1(\Omega)} \to 0$$

as $M \to \infty$. Furthermore, we can estimate

$$\begin{aligned} \|\rho_n - \rho_n \wedge M\|_{L^1(\Omega)} &= \int_\Omega (\rho_n - \rho_n \wedge M) \, dx = \int_{\rho_n \geq M} (\rho_n - M) \, dx \leq \int_{\rho_n \geq M} \rho_n \, dx \\ &\leq \frac{1}{\log(M)} \int_{\rho_n \geq M} \rho_n \log(\rho_n) \, dx \leq \frac{C}{\log(M)}, \end{aligned}$$

for $M \geq 2$, and thus the right hand side goes to 0 as $M \to \infty$. Our claim is that $\rho_n \rightharpoonup \bar{\rho}$ weakly in $L^1(\Omega)$. Take a test function $\varphi \in L^\infty(\Omega)$, then we can estimate

$$\begin{aligned} \left| \int_\Omega \rho_n \varphi \, dx - \int_\Omega \bar{\rho} \varphi \, dx \right| &\leq \|\varphi\|_{L^\infty(\Omega)} \|\rho_n - \rho_n \wedge M\|_{L^1(\Omega)} + \|\varphi\|_{L^\infty(\Omega)} \|\rho_M - \bar{\rho}\|_{L^1(\Omega)} \\ &\quad + \left| \int_\Omega \rho_M \varphi \, dx - \int_\Omega \rho_n \wedge M \varphi \, dx \right|. \end{aligned}$$

We take the limit as $M \to \infty$ and $n \to \infty$ in that order on the right hand side since the first two terms can be made small by taking M large uniformly in n, and the last one can be made small by taking n large enough afterwards. Therefore, $\rho_n \rightharpoonup \bar{\rho}$ weakly in $L^1(\Omega)$ as claimed. Let us also show that $\bar{\rho} \in \mathscr{P}(\Omega)$. Take the set $N_\varepsilon := \{x \in \Omega : dist(x, \partial\Omega) \leq \varepsilon\}$. It is obvious that $|N_\varepsilon| \leq C\varepsilon$ and

$$\int_{N_\varepsilon} \rho_n\,dx \le \int_{N_\varepsilon\cap\{\rho_n\le R\}} \rho_n\,dx + \int_{N_\varepsilon\cap\{\rho_n\ge R\}} \rho_n\,dx$$
$$\le CR\varepsilon + \frac{1}{\log(R)}\int_{N_\varepsilon\cap\{\rho_n\ge R\}} \rho_n\log(\rho_n)\,dx \le CR\varepsilon + \frac{C}{\log(R)}$$

for all $R>0$, taking $R=\varepsilon|\log(\varepsilon)|$ leads to

$$\int_{\Omega\setminus N_\varepsilon} \rho_n\,dx \ge 1-\frac{C}{|\log(\varepsilon)|} \qquad \text{and thus,} \qquad \int_{\Omega\setminus N_\varepsilon} \bar\rho\,dx \ge 1-\frac{C}{|\log(\varepsilon)|}$$

due to $\rho_n \rightharpoonup \bar\rho$ weakly in $L^1(\Omega)$. Letting now $\varepsilon\to 0$, we conclude that $\bar\rho\in\mathscr{P}(\Omega)$ and that ρ_n converges weakly to $\bar\rho$ in $\mathscr{P}(\Omega)$. Due to the property iii) in Proposition 2.3, proven in exercise 6 in Problem sheet 2, we conclude that

$$d_2^2(\bar\rho,\rho_k) \le \liminf_{n\to\infty} d_2^2(\rho_n,\rho_k)\,.$$

Let us know work with the Boltzmann entropy functional to show that it is lower semicontinuous too. Note that for each $s>0$, we have

$$s\log(s) \ge s(w+1)-e^w \text{ for all } w\in\mathbb{R}$$

with equality for $w=\log(s)$. Hence, given any continuous function φ in $\bar\Omega$, we have

$$\liminf_{n\to\infty}\int_\Omega \rho_n\log\rho_n\,dx \ge \liminf_{n\to\infty}\int_\Omega \left(\rho_n(x)(\varphi(x)+1)-e^{\varphi(x)}\right)dx$$
$$= \int_\Omega \left(\bar\rho(x)(\varphi(x)+1)-e^{\varphi(x)}\right)dx\,.$$

Since this is is true for all continuous and bounded functions φ, one can take the supremum in the right hand side. One can prove that this supremum is given by $\mathscr{E}[\rho_\infty]$ by approximating $\log\rho_\infty$ by continuous functions. This result of lower semicontinuity of the entropy is much more general and it can be seen in [1]. Putting together the previous results, we get that

$$I=\liminf_{n\to\infty}\frac{1}{2\Delta t}d_2^2(\rho_n,\rho_k)+\mathscr{E}[\rho_n] \ge \frac{1}{2\Delta t}d_2^2(\bar\rho,\rho_k)+\mathscr{E}[\bar\rho]\,,$$

and thus the infimum of $\mathscr{E}_k$ is a minimum achieved at $\bar\rho$. □

As we discussed earlier, a suitable interpolation of the variational scheme obtained in the previous result leads in the limit $\Delta t\to 0$ to a solution of the heat equation with Neumann boundary conditions. This can been seen in [21, 3, 2, 13] and it has been extended to a very general class of equations of the form (1.1) under certain conditions on the potentials V and W and the nonlinearity U. This is certainly an area of active research still nowadays branching in many different directions in terms of other metrics involved, applications in differential geometry, in stochastic analysis, mathematical finance, machine learning and many other corresponding evolution PDEs that be cast in this framework.

References

1. L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.
2. L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 1–155. Springer, Heidelberg, 2013.
3. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
4. L. Ambrosio and G. Savaré. Gradient flows of probability measures. In *Handbook of differential equations: evolutionary equations. Vol. III*, Handb. Differ. Equ., pages 1–136. Elsevier/North-Holland, Amsterdam, 2007.
5. H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).
6. H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
7. D. L. Burkholder, E. Pardoux, and A. Sznitman. *École d'Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1991. Papers from the school held in Saint-Flour, August 16–September 2, 1989, Edited by P. L. Hennequin.
8. J. A. Carrillo, Y.-P. Choi, and M. Hauray. The derivation of swarming models: mean-field limit and Wasserstein distances. In *Collective dynamics from bacteria to crowds*, volume 553 of *CISM Courses and Lect.*, pages 1–46. Springer, Vienna, 2014.
9. J. A. Carrillo, K. Craig, and Y. Yao. Aggregation-diffusion equations: dynamics, asymptotics, and singular limits. In *Active particles, Vol. 2*, Model. Simul. Sci. Eng. Technol., pages 65–108. Birkhäuser/Springer, Cham, 2019.
10. J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, Model. Simul. Sci. Eng. Technol., pages 297–336. Birkhäuser Boston, Boston, MA, 2010.
11. J. A. Carrillo and G. Toscani. Contractive probability metrics and asymptotic behavior of dissipative kinetic equations. *Riv. Mat. Univ. Parma (7)*, 6:75–198, 2007.
12. L. C. Evans. *Weak convergence methods for nonlinear partial differential equations*, volume 74 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1990.
13. A. Figalli and F. Glaudo. An invitation to optimal transport, wassertein distances and gradient flows. to appear.
14. F. Golse. On the dynamics of large particle systems in the mean field limit. In *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, volume 3 of *Lect. Notes Appl. Math. Mech.*, pages 1–144. Springer, [Cham], 2016.
15. R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.
16. F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.
17. R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.
18. F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

19. M. Thorpe. Introduction to optimal transport. Notes of Course at University of Cambridge. 2018.
20. J. L. Vázquez. *The porous medium equation*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford, 2007. Mathematical theory.
21. C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
22. C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.