# Unbalanced Optimal Transport : Models, Numerical Methods, Applications

Lenaic Chizat

# THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

**Préparée à l'Université Paris-Dauphine**

# Transport optimal de mesures positives : modèles, méthodes numériques, applications

**École Doctorale de Dauphine — ED 543**

Spécialité Sciences

**COMPOSITION DU JURY :**

Gabriel PEYRÉ
CNRS-Ecole Normale Supérieure
Directeur de thèse

Julie DELON
Université Paris-Descartes
Présidente du jury

Jean-David BENAMOU
INRIA-Université Paris-Dauphine
Membre du jury

Bertrand MAURY
Université Paris-Sud
Membre du jury

Nicolas PAPADAKIS
Institut Mathématique de Bordeaux
Membre du jury

Francesco ROSSI
Aix Marseille Université
Membre du jury

François-Xavier VIALARD
Université Paris-Dauphine
Membre du jury

**Soutenue le 10/11/2017**
**par Lénaïc CHIZAT**

Dirigée par **Gabriel PEYRÉ**

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

# Contents

## Remerciements

Je souhaite adresser de sincères remerciements à tous ceux qui ont, de façon plus ou moins directe, rendu cette thèse possible.

Aux rapporteurs de cette thèse, Giuseppe Savaré et Benedikt Wirth ainsi qu'aux membres du jury pour le temps qu'ils ont accordé à la lecture de ce manuscrit.

À Gabriel, dont le dynamisme proverbial, la curiosité et la bienveillance resteront une source d'inspiration pour ma carrière scientifique. À François-Xavier pour les sujets fertiles vers lesquels il a su m'orienter et les heures d'explorations que l'on a partagé qui m'ont donné goût à la persévérance. À Bernhard, qui fût un partenaire de choix pour mes premiers pas de chercheur. À Simone pour son recul et sa bonne humeur constante. À la joyeuse équipe des doctorants de Dauphine de la génération *6ème étage*. À Maxime pour beaucoup de choses, Raphaël pour sa générosité, à mon jumeau de thèse Quentin, à Isabelle, sage flegmatique, à Luca le maestro des dîners sociaux, à Jonathan pour nos joutes rhétoriques et à Camille qui les imite tous (mais je ne suis pas si naïf). Aussi à Daniela, Olivier, Gilles et Thomas, Isabelle, Marie et César pour leur disponibilité et leur aide précieuse aux projets que j'ai portés. Aux doctorants de l'ENS : à Paul qui a supporté mes monologues, à Gwendoline qui prend la relève, à Aude toi qui est une pro, à Jean débordant d'idées. À Cyril, Maxence, Jessica, Zaïna et Bénédicte pour l'accueil chaleureux qu'ils nous ont offert au sein du DMA. À Mokaplan, pour les Mokacredis sympathiques en équipe et pour tous les néologismes de qualité que ce mot nous inspire. À Guillaume et Jean-David qui forment l'âme des lieux, à Yann pour sa conversation éclectique, à Vincent qui me fait l'honneur d'ajouter cet ouvrage à sa riche collection personnelle, à Simon pour l'atmosphère paisible de son bureau, à Lucas et ses énigmes et à Irène pour les solutions.

À tous mes professeurs qui ont patiemment instillé, chez moi comme chez tant d'autres, des graines de connaissances qui ont lentement éclos en passion du savoir. En particulier à Mme Bia Seguin, Mme Galotti et M. Arsac pour leur influence déterminante. À Hélène qui a assisté à mes péripéties depuis les premières loges. À mes amis de longue date que je vois grandir, de plus ou moins loin, mais toujours avec tendresse : Arnaud, Lucas, Charles, Alberto, Alice, Yohann, Alexandre, Mike, Clémentine. À mes parents, à qui je dois toutes les libertés. À ma grande sœur et mon grand frère, Lorédane et Erwan, à ceux qui ont rejoint la troupe, Claudie et Émeric. Enfin à Shivangi, ma plus précieuse découverte de ces années de thèse.

# Introduction (English)

The core problem in the theory of optimal transport can be described in simple terms: given two distributions of mass $\mu$ and $\nu$ on some spaces $X$ and $Y$, and the knowledge of the cost $c(x,y) \in \mathbb{R}$ for moving a unit of mass located at $x \in X$ to $y \in Y$, find among all *plans* that describe a method to move the mass from $\mu$ to $\nu$, the one that minimizes the total cost of transport. Obviously, this question is meaningful only if $\mu$ and $\nu$ are distributions with identical total mass. *The object of this thesis is to define, study and numerically solve similar problems that remain meaningful when the distributions have unequal masses— namely* unbalanced *distributions—and preserve some essential properties of the original optimal transport problem.*

## Origin of optimal transport theory and essential facts

The first formalization of the optimal transport problem generally could be traced back to a treatise of Monge [115] dating from 1781. Motivated by the problem of land leveling, he studied the case of a distance cost $c(x,y) = |y-x|$ on a Euclidean space which led him to the discovery of rich geometrical objects [171]. This problem was rekindled in the 1940s by Kantorovich [91], who stated it in its modern form as a linear programming problem on probability measures:

$$C_c(\mu,\nu) \coloneqq \min \left\{ \int_{X\times Y} c(x,y) \mathrm{d}\gamma(x,y) \; ; \; \gamma \in \mathscr{P}(X\times Y), (\pi^x_\# \gamma, \pi^y_\# \gamma) = (\mu,\nu) \right\} \tag{1}$$

where $\pi^x_\#$ and $\pi^y_\#$ are operators that return the marginals of a measure on the product space $X \times Y$. $X$ and $Y$ are typically compact metric or Polish[1] spaces and $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$ are probability measures.

Kantorovich and his co-authors have established many of the essential facts: existence of optimal plans $\gamma$, dual formulation in terms of prices and optimality conditions. They also realized that the optimal cost $C_c(\mu,\nu)$ is a relevant quantity to compare probability measures. In particular, if $d$ is a distance on $X = Y$ and $p \geq 1$, then $W_p(\mu,\nu) \coloneqq C_{d^p}(\mu,\nu)^{1/p}$ defines (modulo conditions on moments) a distance on $\mathscr{P}(X)$, nowadays known as the *p-Wasserstein* distance. This distance has the nice property to reproduce the structure of $X$ on $\mathscr{P}(X)$: if $(X,d)$ is a Polish or a geodesic space, so is $(\mathscr{P}(X), W_p)$.

The fact that many operations behave naturally on $(\mathscr{P}(X), W_p)$—interpolation of probability measures, differentiation of a path, barycenters of a family, to name a few—partly explains the success of the theory, along with the striking connections to well established problems that are discovered every now and then, such as Euler flows [26] or evolution PDEs as gradient flows [120].

A final noteworthy feature is the variational characterizations of optimal transport interpolations, known as the Benamou-Brenier formula [12], or the *dynamic formulation*. If the cost is of the form $c(x,y) = L(y-x)$ for a strictly convex nonnegative function $L$ on $\mathbb{R}^d$, one has

$$C_c(\mu,\nu) = \min \left\{ \int_0^1 \int_{\mathbb{R}^d} L(v_t) \mathrm{d}\rho_t \mathrm{d}t \; : \; \partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0, (\rho_0, \rho_1) = (\mu,\nu) \right\} \tag{2}$$

[1] a topological space is Polish when it is separable and completely metrizable

where $(\rho_t)_{t\in[0,1]}$ is a weakly continuous family of probability measures interpolating between $\mu$ and $\nu$, advected by the velocity field $v_t : \mathbb{R}^d \to \mathbb{R}^d$ at all time [89].

This thesis was greatly influenced by the monographs [161, 162, 142, 5] and we suggest readers to refer to these monographs to taste the beauty of the optimal transport theory, beyond the basic facts that we have just reviewed.

### Use in applied fields and (sometimes) undesirable mass constraint

Because it enables to lift a metric between points on a space, to a metric between distributions of mass on that space, optimal transport appears as a natural tool in many domains where objects like histograms, probabilities, distributions of mass or densities are manipulated. In particular, optimal transport is becoming increasingly popular in applied fields, a popularity fueled by regular improvements in the numerical methods in various contexts. In the domain of shape processing, it has been used as a tool for registration [82], segmentation [146] or restoration [18]. Popular applications in image processing are color manipulation [132, 54, 131], reflectance interpolation [19] and denoising [97]. In statistical learning, it has proved to be efficient as a metric between histograms [125, 51], in image retrieval in computer vision [140], in semi-supervised learning [153] or in domain adaptation [49].

In most of these applications, the fact that optimal transport requires equality of mass between the two distributions is a drawback rather than a feature, and a more desirable behavior would be obtained by mixing displacement as well as local adjustments of mass. It is thus not surprising that many authors in applied fields have considered extensions of the optimal transport problem that allow for variation of mass. In the following few paragraphs we will see specific applications where it is meaningful to consider other models than pure transport.

**Shapes and image processing** Variational methods are common for shapes processing [167, 141, 168]. Among them, optimal transport stands out by being convex. It has however some flaws on the modelization side. For instance, transport maps have no regard for the topology of the shapes. Moreover, when manipulated with optimal transport, shapes or images are generally represented as densities on $\mathbb{R}^2$ or $\mathbb{R}^3$ and a normalization leading to unit mass is required before using optimal transport. A more flexible model, which is similar to the popular method of metamorphosis in the field of diffeomorphic registration [157], is to explain variations of mass by local growth or shrinkage. This idea led to the introduction of specific unbalanced optimal transport models [106, 105] and was the original motivation of our own work.

**Histogram processing** For image color manipulation, optimal transport is used as a tool to define deformation maps in the color space. Even though here the global mass constraint is naturally satisfied, it has been observed that better results where obtained when this constraint was relaxed [68]. The rationale, that can be generalized to other applications, is that when working with histograms on geometric domains, it is more desirable to preserve the modes (local density maxima) rather than the mass.

**Statistical learning** In statistical learning, optimal transport has been used as a metric between features and it has been observed that relaxing the marginal constraints improves the accuracy of label prediction [74]. In this context, relaxing the mass constraint is understood as a regularization that can be tuned with a parameter.

**Gradient flows** A popular use of optimal transport is to study evolution partial differential equations (PDE). Prominent examples of evolution PDEs can be recovered as gradient flows of some functionals for the Wasserstein metric $W_2$. For instance, the heat equation is the gradient flow of the entropy [90], but non-linear PDEs can also be considered [122] as well as highly non-smooth functionals [110]. This approach is both of theoretical interest for studying existence, uniqueness and convergence of solutions, and of practical one for the computations of degenerate flows [28]. So far, this method was only available for mass preserving evolution PDEs, due to the constraint inherent to optimal transport (except when specific boundary conditions are chosen [70]). A theory of unbalanced optimal transport metric paves the way for further applications to evolution PDEs with mass variation (see Chapter 6).

### Unbalanced optimal transport: state-of-the-art

The term *unbalanced* seems to appear first in [11]—where it is suggested to relax the marginal constraints with a $L^2$ penalization—but the idea has been around long before. Kantorovitch already proposed to throw mass away, out of the (compact) domain (see [80]), an idea that was deepened to study evolution PDEs with Dirichlet boundary conditions [70]. More generally, it has been a common trick in applications to add a dummy point that acts as an infinite source or sink of mass, that can be reached at some cost. If the cost for reaching that dummy point is constant, one recovers the *optimal partial transport problem*, studied theoretically in [32] and [69]. This problem consists in transferring a fixed amount of mass $m > 0$ between two measures while minimizing the transport cost. It is shown in [32] that the mass constraint admits a dual formulation as a maximum transport cost.

Independently, a class of optimal transport distances, where the marginal constraint is relaxed with a total variation penalization[2] has been studied [129, 130]. Many properties are proved or recovered, such as the duality between bounded Lipschitz and the unbalanced Kantorovich-Rubinstein norms [83] and also one dynamic formulation. In Chapter 2, we show that these models are equivalent to the optimal partial transport problem and they are recovered as a limit case of our general framework.

When I started the preparation of this thesis, the theoretical study of unbalanced optimal transport was centered around the optimal partial transport problem. As this approach was not always satisfying for practitioners, other models, tailored for specific applications, have been introduced. The relationships between the various approaches were not known and a systematic study to define the range of available tools was lacking.

[2] In this thesis, total variation refers to the total variation *of measures* which boils down to the $L^1$ norm for functions; not to the $L^1$ norm of the gradient.

### Parallel work

Except when explicitly stated, the material of this thesis is my own work or is work produced with collaborators: Simone Di Marino, Gabriel Peyré, Bernhard Schmitzer, Justin Solomon and François-Xavier Vialard. During the preparation of this thesis, other teams of researchers have worked on related subjects. In the following paragraph, I would like to state how these external contributions have influenced this thesis.

The metric corresponding to the quadratic case studied in Chapter 2 has been introduced independantly by several teams (including ours) over a surprisingly short time period [103, 104, 94, 40, 44, 43]. It has been called *Hellinger-Kantorovich* in [103, 104] and *Wasserstein-Fisher-Rao* in our papers [44, 43]. In this thesis, it is simply denoted by $\widehat{W}_2$ and introduced as part of a larger family of metrics $\widehat{W}_p$. Many extremely valuable aspects developed in these other works have been deliberately left aside from this thesis, which focuses on my personal contributions.

I would like to mention that the metric properties (completeness, metrization of weak convergence) of $\widehat{W}_2$ were not present in our communicated work but were treated in [94, 103]. Still, I found relevant to consider similar results for the larger class of metrics $\widehat{W}_p$ in this thesis.

In [43], we introduced semi-coupling formulations for unbalanced optimal transport. Two closely related, alternative, formulations were proposed in [103]. In Chapter 1, these formulations are introduced (with some variants) and their relationships proved. Moreover, large parts of this thesis (in particular Chapters 3 and 6), make an intensive use of the optimal entropy-transport problem, which is a contribution of [103].

## Part I: Models for Unbalanced Optimal Transport

In the first part of this thesis, we introduce various models of unbalanced optimal transport, study their main properties and establish the relationships between them.

### Chapter 1: Formulations and Equivalences

The first approach, considered in Section 1.1, is a *dynamic formulation* that mirrors (2). It consists in looking for time dependent interpolations $(\rho_t)_{t\in[0,1]}$ between two measures $\mu,\nu\in\mathscr{M}_+(\Omega)$, $\Omega\subset\mathbb{R}^d$, that minimize an action functional. As suggested previously [106, 129], we consider the continuity equation with source (also called "with reaction")

$$\partial_t\rho_t+\nabla\cdot(v_t\rho_t)=g_t\rho_t,\qquad(\rho_0,\rho_1)=(\mu,\nu). \tag{3}$$

This equation allows two degrees of freedom for the modification of mass, through the velocity fields $v_t:\Omega\to\mathbb{R}^d$ and the rate of growth fields $g_t:\Omega\to\mathbb{R}$. Keeping in mind that the key property in optimal transport is homogeneity with respect to mass, we introduce a new general class of problems

$$C_L(\mu,\nu):=\min\left\{\int_0^1\int_\Omega L(v_t(x),g_t(x))\,\mathrm{d}\rho_t(x)\,\mathrm{d}t\ ;\ (\rho_t)\text{ solves (3)}\right\} \tag{4}$$

where $L:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$ is a Lagrangian function which is convex, continuous and minimal at $L(0,0)=0$. In order to gain generality and simplify the analysis, we rephrase this problem

in terms of the perspective function $\psi_L$, but its essence remains the same. We show that this problem admits minimizers with a finite cost and derive the dual problem which maximizes a linear cost over sub-solutions to a Hamilton-Jacobi equation with a zeroth-order term. Just as the optimal transport cost (1), we show that $C_L$ entertains close links with weak convergence of measures and that for a $p$-homogeneous Lagrangian, $p \geq 1$, $C_L^{1/p}$ defines a geodesic metric on the space of nonnegative measures. In the limit case $p = 1$, we show that the time variable is superfluous and $C_L$ can be written as an unbalanced Beckmann's minimal flow problem.

In Section 1.2, we suggest a second approach to unbalanced optimal transport. We extend the notion of coupling to overcome the fact that nonnegative measures cannot be coupled in general. Instead of one coupling, we consider a pair $(f_1\gamma, f_2\gamma)$ that we call a pair of *semi-couplings*, because each of $f_i\gamma$ is a measure on $X \times Y$ that satisfies only one marginal constraint. We then choose a function $h_{x,y}(u,v)$ that gives the cost of moving the atom $u\delta_x$ to $v\delta_y$ which should be sublinear (convex and 1-homogeneous) in $(u,v) \in \mathbb{R}_+^2$. It is used to define the semi-coupling formulation between two measures $(\mu,\nu) \in \mathscr{M}_+(X) \times \mathscr{M}_+(Y)$ on compact metric spaces $X, Y$ as

$$C_h(\mu,\nu) := \min \left\{ \int_{X\times Y} h_{x,y}(f_1(x,y), f_2(x,y)) \mathrm{d}\gamma(x,y) \; ; \; (\pi^x_\#(f_1\gamma), \pi^y_\#(f_2\gamma)) = (\mu,\nu) \right\}$$

where the minimum runs over $f_1, f_2 : X \times Y \to \mathbb{R}_+$ and $\gamma \in \mathscr{P}(X \times Y)$. Properties of $C_h$ are quite similar to those of the Kantorovich problem (1): the minimum is attained, there is a dual problem in terms of continuous functions and, for suitable choices of sublinear cost $h$, $C_h$ is continuous under weak convergence of measures. Another nice feature is that if $X = Y$ and $h^{1/p}$ defines a metric on the cone of $X$—namely the space $X \times \mathbb{R}_+$ where the points $X \times \{0\}$ are identified to one point—then $C_h^{1/p}$ defines a metric on $\mathscr{M}_+(\Omega)$. We also introduce the formulations from [103] and make the connection with ours. In particular, we consider a variant of their *optimal lift* formulation,

$$C_c^{\mathfrak{h}}(\mu,\nu) := \min \{ C_c(\bar{\mu},\bar{\nu}) \; ; \; (\mathfrak{h}\bar{\mu}, \mathfrak{h}\bar{\nu}) = (\mu,\nu) \}$$

where $c$ is a cost on $(X \times \mathbb{R}_+) \times (Y \times \mathbb{R}_+)$ and $\mathfrak{h}$ is a partial expectation operator, informally $\mathfrak{h}\bar{\mu}(\mathrm{d}x) = \int_{\mathbb{R}_+} u \mathrm{d}\mu(\mathrm{d}x, \mathrm{d}u)$. We show that when $c$ is continuous, $C_c^{\mathfrak{h}}$ is equivalent to a semi-coupling problem $C_h$ if $h_{x,y}$ is defined as the sublinear regularization of $c((x,\cdot),(y,\cdot))$ for all $(x,y) \in X \times Y$. There is a last formulation, that is not our contribution but that plays an essential role in this thesis, the optimal-entropy transport formulation [103]. It is similar to the Kantorovich formulation but with a relaxed marginal constraint. Informally, it is defined as

$$C_{c,f}(\mu,\nu) := \min_{\gamma \in \mathscr{M}_+(X\times Y)} \left\{ \int_{X\times Y} c \mathrm{d}\gamma + \int_X f\left(\frac{d(\pi^x_\#\gamma)}{d\mu}\right) \mathrm{d}\mu + \int_Y f\left(\frac{d(\pi^y_\#\gamma)}{d\nu}\right) \mathrm{d}\nu \right\} \qquad (5)$$

where $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ is a cost function and $f$ is an *entropy* function, i.e. a convex function on $\mathbb{R}_+$ minimal at 1, used to measure how much the marginals of the plan $\gamma$ deviate from $\mu$ and $\nu$. The terms depending on the marginals of $\gamma$ are so-called $f$-divergence functionals.

In Section 1.3, we study the relationship between the dynamic and coupling formulation and show a counterpart of the Benamou-Brenier formula (2) in the unbalanced setting. Indeed, if one

defines $h_{x,y}(a,b)$ as the l.s.c. sublinear regularization of the *minimal path cost* that is required to join the points $(x,a)$ to $(y,b)$ by an absolutely continuous path in $\Omega \times \mathbb{R}_+$, as measured by the Lagrangian $L$, then it holds $C_L(\mu,\nu) = C_h(\mu,\nu)$ for all pairs of nonnegative measures. This completes the web of relationships between all the formulations of unbalanced optimal transport.

### Chapter 2: Unbalanced Optimal Transport Metrics

The second chapter is devoted to the study of specific models. We consider the family of $p$-homogeneous Lagrangians

$$L_{p,\alpha}(v,g) = \left(\tfrac{1}{\alpha}|v|\right)^p + \left(\tfrac{1}{p}|g|\right)^p$$

with $p \geq 1$ and $\alpha > 0$. The quantities $\widehat{W}_{p,\alpha} = C_{L_{p,\alpha}}^{1/p}$ define a new family of metrics that turn $\mathscr{M}_+(\Omega)$ into a complete geodesic space, with the topology of weak convergence. This family form the natural counterpart of the $p$-Wasserstein metrics to the unbalanced setting. We show that $\alpha$ is a scale parameter, in the sense that, if $s_\#$ the pushforward map by the linear rescaling $s : x \mapsto \alpha x$, then $s_\# : (\mathscr{M}_+(\Omega), \widehat{W}_{p,1}) \to (\mathscr{M}_+(\alpha\Omega), \widehat{W}_{p,\alpha})$ is an isometry. We then study the limits of the $\widehat{W}_{p,\alpha}$ geodesics when $\alpha$ goes to 0 and $\infty$. We show that when $\alpha \downarrow 0$, we recover the geodesics for a class of *pure growth* metrics whose expression $\left(\int_X |\mu^{1/p} - \nu^{1/p}|^p\right)^{1/p}$ is a homogenized version of the $L^p$ metrics. They contain the total variation and the Hellinger metrics as particular cases. In the opposite limit $\alpha \uparrow \infty$, we recover the classical $W_p$ geodesics if $\mu(\Omega) = \nu(\Omega)$ and a generalized version of those otherwise.

The next section considers the quadratic case $\widehat{W}_2$ for which an explicit formula for the minimal path cost is available. It follows, letting $\alpha = 1$, an explicit semi-coupling formulation $\widehat{W}_2^2(\mu,\nu) = C_h(\mu,\nu)$ with the sublinear cost

$$h_{x,y}(a,b) = a + b - 2\sqrt{ab}\cos(\min\{\mathrm{dist}(x,y), \pi/2\}).$$

This static reformulation enables the use of efficient numerical methods as considered later in Chapter 3. For the case $p = 1$, we recover the traditional Bounded Lipschitz metric. Finally, we consider Lagrangians of the form $L(v,g) = (|v|/\alpha)^p + \frac{1}{2}|g|$ which are reminiscent of the problems considered in [129, 130]. We prove that these models are equivalent to the optimal partial transport problem [69, 32], with a dual parametrization.

## Part II: Numerical Methods

The second part of this thesis focuses on numerical schemes for solving unbalanced optimal transport and related variational problems. Incidentally, our analysis suggests frameworks and improvements for solving other problems related to optimal transport.

### Chapter 3 : Scaling Algorithms for Optimal Couplings

This chapter is devoted to the definition and study of a class of *scaling algorithms* to solve problems of the form

$$\min\left\{\int_{X\times Y} c\cdot \mathrm{d}\gamma + F_1(\pi^x_\#\gamma) + F_2(\pi^y_\#\gamma)\ ;\ \gamma \in \mathscr{M}_+(X\times Y)^n\right\} \tag{6}$$

that we refer to as the *generic formulation*. In this problem, the unknown is a family of $n$ couplings (possibly $n > 1$), $c : X \times Y \to \bar{\mathbb{R}}^n$ is a family of cost functions and $F_1$, $F_2$ are convex, l.s.c, simple functionals, typically "local functionals".

In Section 3.1, we underline the generality of (6) by showing that it covers many known problems, as well as new ones: computation of optimal transport plans, barycenters, gradient flows, and their respective unbalanced counterparts. We also prove well-posedness and existence of minimizers for typical cases of this generic formulation (6).

In Section 3.2, we introduce the structure of the numerical scheme. It involves (i) adding a regularization term to (6) in the form of a Bregman divergence, and (ii) performing alternate bloc maximization on the dual problem. We prove a primal-dual relationship for Bregman divergences which, combined to a result of Beck [10], implies that the algorithm produces a sequence of iterates $(\gamma^{(\ell)})_{\ell\in\mathbb{N}}$ that converges to the optimal *regularized* plan at a $O(1/\ell)$ rate as measured in Bregman distance.

In Section 3.3, we consider more specifically the entropic regularization as initially considered by Cuturi [51]. This corresponds to adding, to the objective functional (6), the term $\varepsilon H(\gamma|\gamma^{(0)})$ which is the relative entropy (also called Kullback-Leibler divergence) of the unknown plan $\gamma$ w.r.t. a reference plan $\gamma^{(0)}$. This leads to a new class of simple *scaling* algorithms. In the discrete setting, with an initialization $b^{(0)} \in \mathbb{R}^Y_{++}$, it produces the sequence of iterates

$$a^{(\ell+1)} = \frac{\operatorname{prox}^H_{F_1/\varepsilon}(K(b^{(\ell)}))}{K(b^{(\ell)})}, \qquad b^{(\ell+1)} = \frac{\operatorname{prox}^H_{F_2/\varepsilon}(K^T(a^{(\ell+1)}))}{K^T(a^{(\ell+1)})} \tag{7}$$

where $K := \exp(-c/\varepsilon)\gamma^{(0)}$ acts on $a$ and $b$ as a matrix vector product. This algorithm is practical whenever the proximal operator with respect to the relative entropy

$$\operatorname{prox}^H_{F_1/\varepsilon} : \bar{s} \in \mathbb{R}^X_+ \mapsto \arg\min_s \; F_1(s)/\varepsilon + H(s|\bar{s})$$

is explicit or simple to compute for $F_1$ and $F_2$. We suggest a stabilization method that allows for the first time to reach small regularization parameters without sacrificing speed and make informal comments on efficient implementation as well as possible generalizations.

In Section 3.4, we propose to use a method to accelerate convergence, known as the *successive over-relaxation* method which, given an acceleration parameter $\theta \in [1,2[$, modifies the iterates as

$$a^{(\ell+1)} = (a^{(\ell)})^{(1-\theta)} \left( \frac{\operatorname{prox}^H_{F_1/\varepsilon}(K(b^{(\ell)}))}{K(b^{(\ell)})} \right)^{\theta}$$

and similarly for $b^{(\ell)}$. This idea is new in the context of optimal transport, and in numerical experiments we observed that it improves convergence speed by up to orders of magnitude. We give a detailed local convergence analysis that exhibits the best choice of $\theta$ and predicts the acceleration factor (these results are well known in the context of resolution of linear systems).

Finally, in Section 3.5, we consider scaling algorithms (7) in the continuous setting. Our main contribution is a proof that the fixed point mapping underlying the iterations (7) is non-expansive in $L^\infty_{++}(X)$ and $L^\infty_{++}(Y)$ for the *Thompson* metric, a metric originating from non-linear Perron-Frobenius theory. In some particular cases of practical importance (computation of barycenters,

gradient flows, optimal plans for $\widehat{W}_2$), we prove that the *global* convergence rate is linear for this metric, in the infinite dimensional setting.

### Chapter 4: Solving Dynamic Formulation

In this short chapter, we propose a numerical method for solving dynamic formulations of the form (4), or more generally, after the change of variables $(\omega_t, \zeta_t) = (v_t\rho_t, g_t\rho_t)$, problems of the form

$$\min\left\{\int_0^1\int_\Omega f(\rho_t(x),\omega_t(x),\zeta_t(x))\mathrm{d}x\mathrm{d}t\ ;\ \partial_t\rho_t+\nabla\cdot\zeta_t=\omega_t \text{ and boundary conditions}\right\}$$

where $f$ is a l.s.c. convex function whose $L^2$-proximal operator $\bar{z}\mapsto \arg\min_z\{f(z)+|z-\bar{z}|^2\}$ is easy to compute. We adopt the approach described in [123] that deals with the balanced case, extend it to the unbalanced setting and make new comments. After discretization on staggered grids, we obtain a non-smooth convex minimization problem that can be tackled with operator splitting methods, such as Douglas-Rachford's algorithm.

In Section 4.2 we derive the method for fast computation of the projection maps in the spectral domain, considering various boundary conditions, and the discretized proximal maps for various action functionals. In Section 4.3, we display numerical geodesics for models introduced in Chapter 2 as well as other models of unbalanced optimal transport considered in the literature on 1-D and 2-D domains. This illustrates the various behaviors that can be expected from those models.

## Part III: Applications

### Chapter 5: Illustrations for Scaling Algorithms

This chapter is the applied and illustrative counterpart of Chapter 3. We review several problems that fit in the framework of the scaling algorithms, derive the explicit form of the iterates and display illustrations. In Section 5.1, we solve optimal entropy-transport problems (5) with four choices of $f$-divergences: exact marginal constraints, relative entropy, total variation, and range constraint. We display 1-D and 2-D experiments as well as a color transfer experiment where the unbalanced framework comes naturally. In Section 5.2, we consider unbalanced optimal transport barycenter problems with the same choices of $f$-divergences. This is applied to the computation of geodesics for $\widehat{W}_2$ and to a comparison between balanced and unbalanced barycenters in 1-D and 2-D domains. In Section 5.3, we explain how to use scaling algorithms for computing time discretized gradient flows w.r.t. optimal transport or unbalanced optimal transport metrics. In particular, we derive an algorithm for solving the challenging Wasserstein gradient flow of the functional "total variation of the gradient".

### Chapter 6: Gradient Flow of Hele-Shaw Type

The chapter evolves around an evolution partial differential equation (PDE) of Hele-Shaw type, that has been studied in mathematical biology as a mechanical model for tumor growth. This PDE models the evolution of a positive density of, say, malignant cells $(\rho_t)_{t\in[0,T]}$ on a domain

$\Omega \subset \mathbb{R}^d$ that multiply under a maximum density constraint. When the constraint is saturated, this generates a positive pressure $p_t(x)$ that diminishes the rate of growth and advects the density through Darcy's law. Formally, a solution is a weakly continuous paths $(\rho_t)_{t\in[0,T]}$ starting form $\rho_0 \in L^1_+(\Omega)$ with $\rho_0 \leq 1$, that weakly solves

$$\begin{cases} \partial_t \rho_t - \nabla \cdot (\rho_t \nabla p_t) = \Phi(p_t)\rho_t \\ p_t(1-\rho_t) = 0 \\ 0 \leq \rho_t \leq 1 \end{cases} \tag{8}$$

where $\Phi(p)$ is the rate of growth, which is a decreasing function of the pressure $p$. Focusing on the case $\Phi(p) = 4(\lambda - p)_+$ with $\lambda > 0$ we show existence and uniqueness of solutions to (8) by showing that this system characterizes gradient flows of the function

$$G(\rho) = \begin{cases} -\lambda\rho(\Omega) & \text{if } \rho \ll \mathscr{L}^d \text{ and } \frac{d\rho}{d\mathscr{L}^d} \leq 1, \\ +\infty & \text{otherwise} \end{cases} \tag{9}$$

for the metric $\widehat{W}_2$ introduced in Chapter 2. More precisely, we show that any minimizing movement is a solution to (8) and that if $\Omega$ is convex, solutions to (8) are, in turn, $\mathrm{EVI}_{(-2\lambda)}$ solutions of gradient flow. The latter is a characterization of gradient flows in metric spaces that guarantees in particular uniqueness [5].

This approach has the following advantages: (i) it suggests a simple interpretation of solutions to (8) as *the most efficient way for a density to gain mass under a maximum density constraint*, where efficiency is measured by the metric $\widehat{W}_2$, (ii) it allows to improve over the existing theoretical results (contrary to [113], we make no regularity assumption on the initial data) and (iii) it leads up to a numerical scheme thanks to the time discretized gradient flow.

In Section 6.3, we proceed with the spatial discretization and derive the scaling algorithm for solving the gradient flow time steps. We show that the numerical scheme is consistent throughout, in that it recovers a solution to (8) when all parameters tend to their limit, successively. We then derive explicit spherical solutions, that we use to assess the precision of the numerical method as well as the convergence of the scheme, in Section 6.4. We conclude with numerical illustrations in 1-D and 2-D.

### Chapter 7 : Algorithmic Approach for Matrices

In this last chapter, we consider an extension of the entropy-transport problem to allow optimal transport of measures whose values are positive semidefinite (PSD) matrices. Our objective is to find a compromise between geometrical faithfulness and algorithmic efficiency, relying on the ideas of Chapter 3.

We propose to solve an entropy-transport-like problem

$$\min \int_{X\times Y} c(x,y)\cdot \mathrm{d}\gamma(x,y) + \int_X H_q(\pi^x_\#\gamma|\mu) + \int_Y H_q(\pi^y_\#\gamma|\nu)$$

where $\gamma$ is now a measure on a product space $X \times Y$ that takes PSD values, and the cost function $c$ is matrix valued. The Von-Neumann quantum entropy $H_q(P|Q)$ is a sublinear function that captures the dissimilarity between two PSD matrices $P, Q \in S^d_+$ (the integral of a sublinear function

of measures is defined in the section *Notation*). We propose a quantum-entropic regularization of this convex optimization problem, which can be solved efficiently in the discrete case with an iterative scaling algorithm. This algorithm boils down to a specific version of (7) in the scalar case.

The crux of this approach lies in the algebraic properties of the quantum relative entropy, that allows to painlessly derive a scaling algorithm even in this non-commutative setting. We detail a simple adaptation of the model to add a "fixed trace" constraint, which is alike a conservation of mass constraint and also extend this formulation and the algorithm to compute barycenters of a collection of input tensor fields.

# Introduction (Français)

Le problème qui est au cœur de la théorie du transport optimal peut se formuler en ces termes simples : étant données deux distributions de masse $\mu$ et $\nu$ sur des espaces $X$ et $Y$, et étant donné un coût $c(x,y) \in \mathbb{R}$ pour déplacer une unité de masse située au point $x \in X$ vers le point $y \in Y$, déterminer parmi tous les *plans* de transport qui décrivent une manière de déplacer toute la masse de $\mu$ vers celle de $\nu$, celui qui minimise le coût total de transport. Il apparaît immédiatement que cette question n'a de sens que si $\mu$ et $\nu$ sont des distributions dont la masse totale est identique. *L'objet de cette thèse est de définir, étudier, et résoudre numériquement des problèmes similaires qui sont restent bien définis quand les masses totales diffèrent et qui préservent les propriétés caractéristiques du transport optimal classique.*

## Théorie du transport optimal

La première formalisation du problème du transport optimal remonte à un mémoire de Gaspard Monge [115] en 1781. Motivé par des questions pratiques de déplacement de déblais, il a étudié le cas d'un coût de transport de la forme $c(x,y) = |y-x|$ dans un espace Euclidien, ce qui l'a mené à la découverte de structures géométriques riches [171]. Ce problème connu un regain d'intérêt dans les années 1940 sous l'impulsion de Kantorovich [91] qui lui a donné sa forme moderne, en tant que problème de programmation linéaire sur l'espace des mesures de probabilité :

$$C_c(\mu,\nu) := \min \left\{ \int_{X\times Y} c(x,y) \mathrm{d}\gamma(x,y) \; ; \; \gamma \in \mathscr{P}(X\times Y), (\pi^x_\# \gamma, \pi^y_\# \gamma) = (\mu,\nu) \right\} \tag{10}$$

où $\pi^x_\#$ et $\pi^y_\#$ sont des opérateurs qui associent à chaque mesure sur l'espace produit $X \times Y$ ses marginales. $X$ et $Y$ sont typiquement des espaces métriques compacts ou Polonais[3] et $\mu \in \mathscr{P}(X)$ et $\nu \in \mathscr{P}(Y)$ sont des mesures de probabilité.

Kantorovich et ses coauteurs ont établi un certain nombre de résultats fondamentaux : existence de plans $\gamma$ optimaux, formulation duale en termes de prix et conditions d'optimalité. Ils ont aussi remarqué que le coût de transport optimal $C_c(\mu,\nu)$ est une quantité pertinente pour comparer des mesures de probabilité. En particulier, si $d$ est une distance sur $X = Y$ et $p \geq 1$, alors $W_p(\mu,\nu) := C_{d^p}(\mu,\nu)^{1/p}$ définit (modulo des conditions sur les moments) une distance sur $\mathscr{P}(X)$, aujourd'hui connue sous le nom de distance de *p-Wasserstein*. Cette distance a la propriété agréable de reproduire la structure de $X$ sur $\mathscr{P}(X)$ : si $(X,d)$ est un espace Polonais ou géodésique, alors $(\mathscr{P}(X), W_p)$ l'est aussi.

Le succès de la théorie du transport optimal s'explique en partie par le fait que de nombreuses opérations, telles que l'interpolation de mesures de probabilité, la différentiation d'un chemin, la définition de barycentres, se comportent bien dans $(\mathscr{P}(X), W_p)$. On peut aussi l'attribuer aux connections surprenantes qui existent entre le transport optimal et des problèmes bien établis tels que l'équation d'Euler [26] ou encore les équations au dérivées partielles (EDP) d'évolution [120].

[3] un espace topologique est Polonais s'il est séparable et complètement métrisable

Nous nous contenterons de mentionner une dernière caractéristique du transport optimal : la formulation variationnelle des interpolations, connue sous le nom de formule de *Benamou-Brenier* [12], ou *formulation dynamique*. Lorsque la fonction de coût est de la forme $c(x,y) = L(y-x)$ pour une fonction $L$ positive et strictement convexe sur $\mathbb{R}^d$, on a

$$C_c(\mu,\nu) = \min\left\{ \int_0^1 \int_{\mathbb{R}^d} L(v_t)\mathrm{d}\rho_t \mathrm{d}t : \partial_t \rho_t + \nabla\cdot(v_t\rho_t) = 0, (\rho_0,\rho_1) = (\mu,\nu) \right\} \quad (11)$$

où $(\rho_t)_{t\in[0,1]}$ est une famille faiblement continue de mesures de probabilité qui interpole entre $\mu$ et $\nu$, sous l'action d'un champ de vitesses $v_t : \mathbb{R}^d \to \mathbb{R}^d$ dépendant du temps [89].

Nous nous limiterons à ces quelques fait basiques et orientons le lecteur intéressé par la théorie générale du transport optimal vers les ouvrages de référence [161, 162, 142, 5].

### Utilisation dans les application et contrainte de masse

Le transport optimal permet de construire, à partir d'une métrique sur un espace donné, une métrique entre les mesures de probabilité sur cet espace. Cela en fait un outil de choix dans beaucoup d'applications où des objets tels que des histogrammes, des distributions de masse ou des densités sont manipulées. En conséquence, et sous l'impulsion des progrès récurrents apportés aux méthodes numériques, le transport optimal est de plus en plus utilisé dans des domaines appliqués. En traitement de formes, il sert d'outil pour le recalage [82], la segmentation [146] ou la restauration [18]. En traitement d'images, il est couramment utilisé pour la manipulation de couleurs [132, 54, 131], l'ajustement de réflectances [19] et le débruitage [97]. En apprentissage statistique, son efficacité a été observée comme métrique entre histogrammes [125, 51], en classification d'images [140], en apprentissage semi-supervisé [153] et pour l'adaptation de domaine [49].

Dans la plupart de ces applications, la contrainte d'égalité des masses apparaît comme une limite plutôt qu'un avantage et une formulation du transport optimal autorisant un dosage de déplacement ainsi que d'ajustement de la masse peut sembler plus attrayante. C'est donc sans surprise que de nombreux auteurs ont proposé des extensions du transport optimal autorisant des variations de masse. Dans les paragraphes qui suivent, nous passons en revue quelques applications où ces extensions sont pertinentes.

**Traitement d'images et de formes** Dans la jungle de méthodes variationnelles utilisées pour le traitement de formes [167, 141, 168], le transport optimal a l'avantage d'être convexe mais a des défauts d'un point de vue modélisation. Par example, les applications de transport n'attachent aucune importance à la topologie des formes. De plus, les images ou formes sont généralement représentées par des densités sur $\mathbb{R}^2$ ou $\mathbb{R}^3$ et une étape de normalisation est nécessaire avant d'utiliser le transport optimal. Un modèle plus flexible, partageant des similarités avec la méthode des métamorphoses pour le recalage difféomorphique [157], est d'expliquer les variations globales de masse par des ajustements locaux. Cette idée a motivé l'introduction de modèles de transport non-équilibré spécifiques [106, 105] et est la motivation d'origine de notre travail.

**Traitement d'histogrammes** Dans le cadre de la manipulation des couleurs d'images, le transport optimal permet de définir des déformations dans l'espace des couleurs qui minimisent la distorsion globale. Dans ce contexte de meilleurs résultats ont été obtenus en relâchant la contrainte de masse [68]. L'idée, qui se généralise à d'autres applications, est que lorsque l'on traite d'histogrammes sur des domaines géométriques, il est parfois plus pertinent de préserver les modes (maxima locaux de densité) plutôt que la masse.

**Apprentissage statistique** En apprentissage statistique, lorsque le transport optimal est utilisé comme une métrique entre descripteurs, il a été observé que relâcher les contraintes de marginales améliore la précision de la prédiction d'étiquettes [74]. Dans ce contexte, cette relaxation peut s'interprététer comme une régularisation, dont l'importance peut se régler avec un paramètre.

**Flots de gradient** Le transport optimal est aussi utilisé pour étudier des EDP d'évolution : certaines de ces équations admettent en effet une caractérisation comme flot de gradient pour la métrique de Wasserstein $W_2$. Par exemple, l'équation de la chaleur est le flot du gradient de l'entropie [90], mais des EDP non linéaires peuvent aussi être vues sous ce prisme [122] ainsi que des gradients de fonctions dégénérées [110]. Cette approche comporte un intérêt à la fois théorique pour étudier l'existence, l'unicité et la convergence de solutions et un intérêt pratique pour le calcul numérique des solutions d'EDP dégénérées [28]. Cette méthode n'était jusqu'à présent possible que pour des EDP qui conservent la masse, à cause de la contrainte associée au transport optimal (à l'exception des cas où des conditions aux bords spécifiques sont choisies [70]). Une théorie du transport entre mesures positives permet de l'étendre à certaines EDP qui ne conservent pas la masse (voir chapitre 6).

### Transport optimal non-équilibré : état de l'art

Le terme anglais *unbalanced* (que nous traduisons par *non-équilibré*) semble apparaître pour la première fois dans [11], où une relaxation du problème du transport optimal avec une pénalisation $L^2$ est étudiée, mais l'idée date de bien avant. Kantorovich proposait déjà de considérer le transport entre mesures de masse arbitraire et de rejeter une portion de masse hors du domaine (compact) [80], une idée qui a été approfondie pour étudier des EDP avec des conditions aux bords de Dirichlet [70]. Plus généralement, une astuce répandue dans les applications est d'ajouter un point fictif qui joue le rôle d'une source (ou puit) de masse infinie, qui peut être atteint à un coût donné. Si le coût pour rejoindre ce point fictif est constant sur tout le domaine, alors on retrouve le problème du *transport partiel optimal*, étudié théoriquement dans [70] et [69]. Ce problème consiste en la recherche du transfert d'une quantité $m > 0$ de masse entre deux mesures qui minimise le coût de transport. Il est montré dans [32] que cette contrainte de masse admet une formulation duale sous la forme d'un problème de transport optimal modifié où apparaît un coût maximum au delà duquel tout transport est interdit.

Indépendamment, une classe de distances de transport optimal, où les contraintes de marginales sont relaxées avec une pénalisation Variation Totale[4] a été étudiée par [129, 130]. De nom-

[4] dans cette thèse, la Variation Totale fait référence la Variation Totale des mesures, qui s'identifie à la norme $L^1$ des

breuses propriétés sont prouvées ou retrouvées, telles que la dualité entre la norme Bounded Lipschitz et les normes de Kantorovich-Rubinstein [83] ainsi qu'une formulation dynamique. Dans le chapitre 2, nous retrouvons ces modèles comme cas limite de notre cadre général et montrons qu'ils sont équivalents au problème du transport partiel optimal.

Lorsque j'ai commencé la préparation de ma thèse, l'étude théorique du transport entre mesures positives était ainsi limité au problème de transport partiel. Cette approche n'était pas toujours satisfaisante en pratique et donc d'autres modèles, ajustés à des applications spécifiques, avaient vu le jour. Les relations entre les différents points de vue n'étaient pas connues et une analyse systématique de ces modèles manquait.

### Travaux parallèles

Sauf mention explicite du contraire, le contenu de cette thèse est issu de mon propre travail ou de mon travail avec mes collaborateurs : Simone Di Marino, Gabriel Peyré, Bernhard Schmitzer, Justin Solomon et François-Xavier Vialard. Pendant la préparation de cette thèse, d'autres équipes ont travaillé sur des sujets similaires et je souhaite clarifier dans quelle mesure ces contributions extérieures ont influencé ce manuscrit.

La métrique de transport non-équilibré correspondant au cas quadratique étudiée au chapitre 2 a été introduite indépendamment par différentes équipes (dont la notre) dans un laps de temps étonnamment court [103, 104, 94, 40, 44, 43]. Cette métrique a reçu le nom *Hellinger-Kantorovich* dans [103, 104] et *Wasserstein-Fisher-Rao* dans [44, 43]. Dans cette thèse, elle est simplement notée $\widehat{W}_2$ et introduite comme un cas particulier d'une famille de métriques $\widehat{W}_p$. De nombreuses contributions importantes apportées par ces autres équipes ont été délibérément mises de côté dans cette thèse pour mettre l'accent sur mes propres contributions.

Je souligne le fait que les propriétés métriques (complétude, métrisation de la convergence faible) de $\widehat{W}_2$ n'étaient pas présentes dans nos travaux précédemment communiqués mais avaient été traités dans [94, 103]. Néanmoins, il m'a semblé pertinent de traiter ces propriétés pour la classe plus large de métriques $\widehat{W}_p$.

Dans [43], nous avons introduit une formulation du transport non-équilibré sous forme de "semi-couplages". Deux formulations alternatives ont été proposées dans [103]. Dans le chapitre 1, ces formulations sont introduites (avec quelques variantes) et des équivalences montrées. Par ailleurs, nous faisons une utilisation intensive de la formulation de transport-entropie, en particulier dans les chapitres 3 et 6 : il s'agit d'une contribution de [103].

## Partie I : Modèles de transport optimal entre mesures positives

Dans la première partie de cette thèse, nous introduisons différents modèles de transport optimal non-équilibré, étudions leurs principales propriétés et établissons des équivalences entre des modèles.

fonctions, et non la norme $L^1$ du gradient.

### Chapitre 1 : Formulations et équivalences

La première approche considérée dans la section 1.1 est une *formulation dynamique* qui s'inspire de (11). Elle consiste en la recherche d'une interpolation $(\rho_t)_{t\in[0,1]}$ dépendant du temps entre deux mesures $\mu,\nu\in\mathscr{M}_+(\Omega)$, $\Omega\subset\mathbb{R}^d$ qui minimise une fonctionnelle d'action. Comme suggéré précédemment [106, 129], nous considérons l'équation de continuité avec source (aussi dite "avec réaction")

$$\partial_t\rho_t+\nabla\cdot(v_t\rho_t)=g_t\rho_t,\qquad(\rho_0,\rho_1)=(\mu,\nu). \tag{12}$$

Cette équation autorise deux degrés de liberté pour la modification de la masse, à travers les champs de vitesse $v_t:\Omega\to\mathbb{R}^d$ et de croissance $g_t:\Omega\to\mathbb{R}$. En se rappelant que l'homogénéité par rapport à la masse est une propriété essentielle du transport optimal, on définit une nouvelle classe de problèmes

$$C_L(\mu,\nu):=\min\left\{\int_0^1\int_\Omega L(v_t(x),g_t(x))\,\mathrm{d}\rho_t(x)\,\mathrm{d}t\ ;\ (\rho_t)\text{ solves (3)}\right\} \tag{13}$$

où $L:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$ est un Lagrangien convexe, continu et minimal en $L(0,0)=0$. Afin de gagner en généralité et de simplifier l'analyse, on reformule ce problème à l'aide de la fonction perspective $\psi_L$, mais l'essence du problème est inchangée. Nous montrons que ce problème admet des minimiseurs avec un coût fini et formulons le problème dual, qui consiste en la maximisation d'un coût linéaire sur l'espace des sous-solutions d'une équation de Hamilton-Jacobi avec un terme d'ordre zero. De façon similaire au coût de transport optimal classique (10), nous montrons que $C_L$ est lié à la convergence faible des mesures et que pour un Lagrangien $p$-homogène, $p\geq 1$, $C_L^{1/p}$ définit une métrique géodésique sur l'espace des mesures positives. Dans le cas limite où $p=1$, on montre que la variable temporelle est superflue et que $C_L$ peut se ré-écrire comme un problème de flot minimal de Beckmann.

Dans la section 1.2, nous suggérons une deuxième approche pour le transport de mesures positives. Nous étendons la notion de couplage pour contourner le fait que le couplage entre deux mesures positives n'existe pas en général. Au lieu d'un unique couplage, on considère une paire $(f_1\gamma,f_2\gamma)$ que nous appelons une paire de *semi-couplages*, parce que l'on impose à chaque $f_i\gamma$, qui est une mesure sur $X\times Y$, de satisfaire une seule contrainte de marginale. On choisit ensuite une fonction $h_{x,y}(u,v)$ qui définit le coût pour déplacer un atome $u\delta_x$ vers $v\delta_y$, qui doit être sous-linéaire (c'est-à-dire convexe et 1-homogène) en $(u,v)\in\mathbb{R}_+^2$. Cette fonction permet de définir une formulation *semi-couplage* du transport optimal entre deux mesures $(\mu,\nu)\in\mathscr{M}_+(X)\times\mathscr{M}_+(Y)$ sur des espaces métriques compacts $X,Y$ :

$$C_h(\mu,\nu):=\min\left\{\int_{X\times Y}h_{x,y}(f_1(x,y),f_2(x,y))\mathrm{d}\gamma(x,y)\ ;\ (\pi^x_\#(f_1\gamma),\pi^y_\#(f_2\gamma))=(\mu,\nu)\right\}$$

où le minimum est recherché sur l'ensemble des fonctions $f_1,f_2:X\times Y\to\mathbb{R}_+$ et mesures $\gamma\in\mathscr{P}(X\times Y)$. Les propriétés de $C_h$ sont similaires à celles du problème de Kantorovich (10) : le minimum est atteint, on peut formuler un problème dual sur l'espace des fonctions continues et, pour certains choix de coûts sous-linéaires $h$, $C_h$ est continu pour la convergence faible des mesures. Une autre propriété séduisante est que si $X=Y$ et $h^{1/p}$ définit une métrique sur le

cône de $X$ (l'espace $X \times \mathbb{R}_+$ où les points $X \times \{0\}$ sont identifiés à un seul point), alors $C_h^{1/p}$ définit une métrique sur $\mathscr{M}_+(\Omega)$. On introduit également les formulations proposées par [103] et faisons la connexion avec les nôtres. En particulier, nous considérons une variante de leur formulation où l'on recherche un *relèvement optimal*

$$C_c^{\mathfrak{h}}(\mu,\nu) \coloneqq \min\{C_c(\bar{\mu},\bar{\nu})\ ;\ (\mathfrak{h}\bar{\mu},\mathfrak{h}\bar{\nu}) = (\mu,\nu)\}$$

où le coût $c$ est un coût sur $(X \times \mathbb{R}_+) \times (Y \times \mathbb{R}_+)$ et $\mathfrak{h}$ est un opérateur d'espérance conditionnelle, informellement $\mathfrak{h}\bar{\mu}(\mathrm{d}x) = \int_{\mathbb{R}_+} u\mathrm{d}\mu(\mathrm{d}x,\mathrm{d}u)$. On montre que si $c$ est continu, alors $C_c^{\mathfrak{h}}$ est équivalent à un problème de semi-couplage $C_h$ où $h_{x,y}$ est défini comme la régularisation sous-linéaire de $c((x,\cdot),(y,\cdot))$ pour tous $(x,y) \in X \times Y$. Une dernière formulation, qui n'est pas une contribution personnelle, joue un rôle essentiel dans la suite de cette thèse : la formulation de *transport-entropie* [103], qui est similaire à la formulation de Kantorovich mais où les contraintes de marginales sont relâchées. Informellement, elle est définie par

$$C_{c,f}(\mu,\nu) \coloneqq \min_{\gamma \in \mathscr{M}_+(X\times Y)} \left\{ \int_{X\times Y} c\mathrm{d}\gamma + \int_X f\left(\frac{d(\pi^x_\#\gamma)}{d\mu}\right)\mathrm{d}\mu + \int_Y f\left(\frac{d(\pi^y_\#\gamma)}{d\nu}\right)\mathrm{d}\nu \right\} \tag{14}$$

où $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ est une fonction de coût et $f$ est une fonction d'entropie, c'est-à-dire une fonction convexe sur $\mathbb{R}_+$ minimale en 1, qui sert à quantifier l'écart entre les marginales du plan $\gamma$ et $\mu$ et $\nu$. Les termes dépendant des marginales de $\gamma$ sont appelés des $f$-divergences.

Dans la section 1.3, nous étudions la relation entre les formulations dynamiques et de couplage, et nous montrons un analogue de la formule de Benamou-Brenier (11) dans le cadre non-équilibré. Plus précisément, si l'on définit $h_{x,y}(a,b)$ comme la régularisation sous-linéaire et semi-continue inférieure du *coût de chemin minimal* que l'on obtient en cherchant à joindre $(x,a)$ à $(y,b)$ par des chemins absolument continus dans $\Omega \times \mathbb{R}_+$, alors on a $C_L(\mu,\nu) = C_h(\mu,\nu)$ pour toute paire de mesures positives. Ce résultat parachève la description des différents modèles de transport optimal de mesures positives et des liens qui les unissent.

### Chapitre 2 : Métriques de transport optimal non-équilibré

Le deuxième chapitre est dédié à l'étude de modèles spécifiques. On considère une famille de Lagrangiens $p$-homogènes

$$L_{p,\alpha}(v,g) = \left(\tfrac{1}{\alpha}|v|\right)^p + \left(\tfrac{1}{p}|g|\right)^p$$

avec $p \geq 1$ et $\alpha > 0$. Les quantités $\widehat{W}_{p,\alpha} = C_{L_{p,\alpha}}^{1/p}$ définissent une nouvelle famille de métrique qui font de $\mathscr{M}_+(\Omega)$ un espace géodésique complet, muni de la topologie faible. Cette famille est le pendant naturel des métriques de $p$-Wasserstein au cadre non-équilibré. Nous montrons que $\alpha$ est un paramètre d'échelle, dans la mesure où si $s_\#$ est l'application qui à une mesure associe la mesure image par la transformation linéaire $s : x \mapsto \alpha x$, alors $s_\# : (\mathscr{M}_+(\Omega), \widehat{W}_{p,1}) \to (\mathscr{M}_+(\alpha\Omega), \widehat{W}_{p,\alpha})$ est une isométrie. Nous étudions ensuite la limite des géodésiques de $\widehat{W}_{p,\alpha}$ quand $\alpha$ tend vers 0 et $\infty$. On montre que quand $\alpha \downarrow 0$, on retrouve les géodésiques d'une classe de métrique de pure "croissance" dont l'expression $\left(\int_X |\mu^{1/p} - \nu^{1/p}|^p\right)^{1/p}$ est une version homogénéisée des métriques $L^p$. Les métriques de Hellinger et de Variation Totale font notamment

partie de cette famille. Pour l'autre limite $\alpha \uparrow \infty$, on obtient les géodésiques de $W_p$ classique si $\mu(\Omega) = \nu(\Omega)$, et une généralisation de celles-ci sinon.

La section suivante s'intéresse au cas quadratique $\widehat{W}_2$ pour lequel le coût de chemin minimal est explicite. Il s'ensuit une formulation "semi-couplages" explicite, c'est à dire $\widehat{W}_2^2(\mu,\nu) = C_h(\mu,\nu)$ avec $\alpha = 1$ et le coût sous-linéaire

$$h_{x,y}(a,b) = a + b - 2\sqrt{ab}\cos(\min\{\mathrm{dist}(x,y), \pi/2\}).$$

C'est formulation statique ouvre la voie à des méthodes numériques efficaces telles que celles considérées au chapitre 3. Pour le cas $p = 1$, on retrouve le cas bien connu de la métrique Bounded Lipschitz. Finalement, on considère des Lagrangiens de la forme $L(v,g) = (|v|/\alpha)^p + \frac{1}{2}|g|$ qui rappellent les problèmes considérés dans [129, 130]. On montre que ces modèles sont équivalents au problème du transport partiel optimal [69, 32] avec une paramétrisation duale.

## Partie II : Méthodes numériques

Dans la deuxième partie de cette thèse, nous nous intéressons aux schémas numériques pour résoudre les problèmes de transport non-équilibrés et des problèmes variationnels liés. Incidemment, notre analyse suggère un cadre d'étude et des améliorations pour résoudre d'autres problèmes liés au transport optimal classique.

### Chapitre 3 : Algorithmes de *scaling* pour les couplages optimaux

Ce chapitre est dédié à la définition et l'étude d'une classe d'algorithmes de *scaling* pour résoudre des problèmes de la forme

$$\min\left\{ \int_{X\times Y} c\cdot \mathrm{d}\gamma + F_1(\pi^x_\#\gamma) + F_2(\pi^y_\#\gamma)\ ;\ \gamma \in \mathscr{M}_+(X\times Y)^n \right\} \tag{15}$$

que nous appelons la *formulation générique*. Dans ce problème, l'inconnue est une famille de $n$ couplages (potentiellement $n > 1$), $c : X \times Y \to \bar{\mathbb{R}}^n$ est une famille de fonctions de coût et $F_1, F_2$ sont des fonctionnelles convexes, semi-continues inférieurement et simples, typiquement des fonctionnelles "locales".

Dans la section 3.1, nous mettons en avant la généralité de (15) en montrant que cette formulation couvre de nombreux problèmes connus ou nouveaux : calcul de plans de transport optimaux, de barycentres, de flots de gradient, et leurs équivalents non-équilibrés. On prouve aussi que pour les cas typiques, la formulation générique (15) est bien posées et admet un minimiseur.

Dans la section 3.2, on introduit la structure du schéma numérique. Il implique (i) d'ajouter un terme de régularisation à (6) sous la forme d'une divergence de Bregman, et (ii) de réaliser une méthode de maximisation alternée par blocs sur le problème dual. On prouve une relation primal-duale pour les divergence de Bregman qui, une fois combinée aux résultats de Beck [10], implique que l'algorithme génère une séquence d'itérées $(\gamma^{(\ell)})_{\ell\in\mathbb{N}}$ qui converge vers le transport optimal *régularisé* à un taux $O(1/\ell)$, mesuré en terme de divergence de Bregman.

Dans la section 3.3, on considère le cas de la régularisation entropique, tel que proposé par Cuturi [51]. Cette régularisation se traduit par l'ajout, à la fonctionnelle (15), du terme $\varepsilon H(\gamma|\gamma^{(0)})$

qui est l'entropie relative (aussi appellée divergence de Kullback-Leibler) du plan $\gamma$ par rapport à un mesure de référence $\gamma^{(0)}$. Cela conduit à une nouvelle classe d'algorithmes de *scaling*. Dans le cas discret, avec une initialisation $b^{(0)} \in \mathbb{R}^Y_{++}$, on obtient la séquence

$$a^{(\ell+1)} = \frac{\mathrm{prox}^H_{F_1/\varepsilon}(K(b^{(\ell)}))}{K(b^{(\ell)})}, \qquad b^{(\ell+1)} = \frac{\mathrm{prox}^H_{F_2/\varepsilon}(K^T(a^{(\ell+1)}))}{K^T(a^{(\ell+1)})} \tag{16}$$

où $K \coloneqq \exp(-c/\varepsilon)\gamma^{(0)}$ agit sur $a$ et $b$ comme un produit matrice-vecteur. Cet algorithme est intéressant en pratique du moment que l'opérateur proximal par rapport à l'entropie relative

$$\mathrm{prox}^H_{F_1/\varepsilon} : \bar{s} \in \mathbb{R}^X_+ \mapsto \arg\min_s \; F_1(s)/\varepsilon + H(s|\bar{s})$$

est explicite ou calculable facilement pour $F_1$ et $F_2$. On suggère une méthode de stabilisation qui permet d'atteindre des petits paramètres de régularisation sans sacrifier la faible complexité des itérations et faisons des commentaires informels pour une implémentation efficace ainsi que sur des généralisations possibles.

Dans la section 3.4, on propose d'utiliser une méthode de surrelaxation successive pour accélérer la convergence de l'algorithme. Étant donné un paramètre d'accélération $\theta \in [1,2[$, cette méthode produit les itérées

$$a^{(\ell+1)} = (a^{(\ell)})^{(1-\theta)} \left( \frac{\mathrm{prox}^H_{F_1/\varepsilon}(K(b^{(\ell)}))}{K(b^{(\ell)})} \right)^{\theta}$$

et de même pour $b^{(\ell)}$. Cette idée est nouvelle dans le contexte du transport optimal et permet, dans les expériences numériques, une convergence bien plus rapide dans certains cas. Nous faisons l'analyse détaillée de la convergence locale de ce nouvel algorithme, qui nous révèle le meilleur choix de $\theta$ et prédit le facteur d'accélération (ces résultats sont bien connus pour la résolution de systèmes linéaires).

Finalement, dans la section 3.5, on considère des algorithmes de scaling (16) dans le cadre continu. Notre principale contribution est une preuve que l'itération de point fixe sous-jacente à l'algorithme (16) est non-expansive dans $L^\infty_+(X)$ et $L^\infty_+(Y)$ pour la métrique de Thompson, une métrique issue de la théorie de Perron-Frobenius non linéaire. Dans certains cas particulier importants (calcul de barycentres, flots de gradient, plans optimaux pour $\widehat{W}_2$), on montre que la convergence globale est linéaire pour cette métrique, en dimension infinie.

### Chapitre 4 : Résoudre la formulation dynamique

Dans ce court chapitre, on propose une méthode numérique pour résoudre des problèmes de la forme (13), ou plus généralement, après le changement de variables $(\omega_t, \zeta_t) = (v_t\rho_t, g_t\rho_t)$, des problèmes de la forme

$$\min \left\{ \int_0^1 \int_\Omega f(\rho_t(x), \omega_t(x), \zeta_t(x)) \mathrm{d}x\mathrm{d}t \; ; \; \partial_t\rho_t + \nabla\cdot\zeta_t = \omega_t \text{ et conditions aux bords} \right\}$$

où $f$ est une fonction convexe semi-continue inférieurement dont l'opérateur $L^2$-proximal $\bar{z} \mapsto \arg\min_z\{f(z) + |z-\bar{z}|^2\}$ se calcule facilement. On adopte l'approche décrite dans [123] qui

traite du cas du transport équilibré et l'étend au cadre non-équilibré. Après discrétisation du problème sur des grilles en quinconce, on obtient un problème de minimisation convexe non-lisse qui peut se résoudre avec des méthodes d'éclatement d'opérateurs, telles que l'algorithme de Douglas-Rachford.

Dans la section 4.2, on détaille comment calculer les opérateurs de projection de manière efficace dans le domaine spectral, pour différentes conditions aux bords. On exhibe aussi les opérateurs proximaux pour certains choix de fonctionnelles d'action. Dans la section 4.3, on représente des résultats numériques pour le calcul de géodésiques pour certains modèles traités dans le chapitre 2 ou bien introduits par ailleurs dans la littérature, sur des domaines 1-D et 2-D. Ces résultats illustrent les différents comportement qui sont associés aux différents modèles de transport non-équilibré.

## Partie III : Applications

### Chapitre 5 : : Illustrations des algorithmes de scaling

Ce chapitre est le pendant appliqué et illustré du chapitre 3. On passe en revue quelques problèmes qui entrent dans le cadre des algorithmes de *scaling*, on construit la forme explicite des itérées et on propose des illustrations. Dans la section 5.1, on résout des problèmes d'entropie-transport (14) pour quatre choix de $f$-divergences : contraintes de marginales exactes, entropie relative, Variation Totale et contrainte d'intervalle. On représente en particulier une expérience de transfert de couleurs où le cadre non-équilibré se révèle utile. Dans la section 5.2, on considère le calcul de barycentres avec les mêmes choix de $f$-divergences. On applique l'algorithme au calcul de géodésiques pour $\widehat{W}_2$ et à la comparaison entre les barycentres classiques et non-équilibrés. Dans la section 5.3, on détaille comment les algorithmes de scaling permettent de résoudre des flots de gradient pour des métriques de transport optimal classiques ou non-équilibrées. En particulier, on propose une résolution du problème de flot de gradient pour la fonctionnelle "Variation Totale du gradient".

### Chapitre 6 : Flot de gradients de type Hele-Shaw

Ce chapitre évolue autour d'une EDP d'évolution de type Hele-Shaw qui a été étudiée en biologie mathématique en tant que modèle mécanique de croissance de tumeur. Cette EDP modélise l'évolution d'une densité positive $(\rho_t)_{t\in[0,T]}$ de cellules malignes sur un domaine $\Omega \subset \mathbb{R}^d$ qui se démultiplient sous une contrainte de densité maximale. Quand la contrainte est saturée, une pression $p_t(x)$ strictement positive apparaît, fait diminuer le taux de croissance et provoque un déplacement des cellules suivant la loi de Darcy. Formellement, une solution est un chemin absolument continu $(\rho_t)_{t\in[0,T]}$ dans l'espace des mesures qui part de $\rho_0 \in L^1_+(\Omega)$ avec $\rho_0 \leq 1$ et est une solution faible de

$$\begin{cases} \partial_t \rho_t - \nabla\cdot(\rho_t \nabla p_t) = \Phi(p_t)\rho_t \\ p_t(1-\rho_t) = 0 \\ 0 \leq \rho_t \leq 1 \end{cases} \tag{17}$$

où $\Phi(p)$ est le taux de croissance, qui est une fonction décroissante de la pression $p$. Nous nous concentrons sur le cas $\Phi(p) = 4(\lambda - p)_+$ avec $\lambda > 0$ et montrons l'existence et l'unicité de

solutions de (17) en montrant que ce système caractérise les flots de gradient de la fonction

$$G(\rho) = \begin{cases} -\lambda\rho(\Omega) & \text{if } \rho \ll \mathscr{L}^d \text{ and } \frac{d\rho}{d\mathscr{L}^d} \leq 1, \\ +\infty & \text{otherwise} \end{cases} \tag{18}$$

pour la métrique $\widehat{W}_2$ introduite au chapitre 2. Plus précisément, on montre que tout mouvement minimisant est une solution de (17) et que si $\Omega$ est convexe, alors les solutions de (17) sont à leur tour des solutions $\mathrm{EVI}_{(-2\lambda)}$ du flot de gradient (l'acronyme EVI désigne une caractérisation exigeante des flots de gradient dans un espace métrique qui garantit en particulier l'unicité des solutions [5]).

Cette approche pour étudier (17) possède les avantages suivants : (i) elle suggère une interprétation simple des solutions de (17) comme *la façon la plus efficace pour une densité pour gagner de la masse sous une contrainte de densité maximale*, où la notion d'efficacité est déterminée par la métrique $\widehat{W}_2$, (ii) elle permet d'améliorer les résultats théoriques (au contraire de [113], nous ne faisons pas d'hypothèse de régularité sur la donnée initiale) et (iii) elle mène à un schéma numérique pour simuler l'EDP, à l'aide du flot de gradient discrétisé en temps.

Dans la section 6.3, on procède à la discrétisation spatiale et déterminons l'algorithme de *scaling* qui permet de résoudre les étapes temporelles du flot de gradient. On montre que le schéma numérique est consistent dans son ensemble, dans le sens où une solution continue de (17) est obtenue quand tous les paramètres tendent vers leur limite successivement. On construit ensuite les solutions sphériques explicites qui nous servent de référence pour évaluer la précision de la méthode numérique ainsi que la convergence du schéma dans la section 6.4. On conclue avec des illustrations numériques en 1-D et 2-D.

### Chapitre 7 : Approche algorithmique pour les matrices

Dans ce dernier chapitre, on considère une extension du problème de transport-entropie qui permet de considérer le transport optimal de mesures dont le co-domaine est l'espace des matrices semi-définies positives (SDP). Notre objectif est de trouver un compromis entre bon comportement conforme à la géométrie du problème et efficacité algorithmique, en s'inspirant des idées développées dans le chapitre 3.

On propose de résoudre un problème similaire aux problèmes d'entropie-transport

$$\min \int_{X\times Y} c(x,y)\cdot \mathrm{d}\gamma(x,y) + \int_X H_q(\pi^x_\#\gamma|\mu) + \int_Y H_q(\pi^y_\#\gamma|\nu)$$

où $\gamma$ est maintenant une mesure sur l'espace produit $X \times Y$ qui prend des valeurs SDP et la fonction de coût $c$ est à valeurs matricielles. L'entropie quantique de Von-Neumann $H_q(P|Q)$ est une fonction sous-linéaire qui quantifie la dissimilarité entre deux matrices SDP $P, Q \in S^d_+$ (l'intégrale d'une fonction sous-linéaire de mesures est définie dans la section *Notation*). On propose une régularisation de ce problème de minimisation convexe à l'aide de l'entropie quantique. Ce nouveau problème se résout de manière efficace à l'aide d'un algorithme de *scaling*. Cet algorithme se ramène à une version spécifique de (16) dans le cas scalaire.

Le point crucial de cette approche repose dans les propriétés algébriques de l'entropie relative quantique qui permet de définir un algorithme de *scaling* même dans ce cadre non-commutatif.

On propose une adaptation simple du modèle où l'on ajoute une contrainte de trace qui peut s'interpréter comme une contrainte de conservation de masse. Finalement, on propose un algorithme pour calculer le barycentre d'une collection de champs de tenseurs.

# Notation

**Abstract ambient spaces.** $X, Y, Z$ denote abstract ambient spaces (typically topological spaces or compact metric spaces). Measurability is always understood w.r.t. the Borel $\sigma$-algebra. For a product space $X \times Y \times Z \times \dots$, we denote $\pi^x, \pi^y, \dots$ the marginalization maps $\pi^x(x,y,z,\dots) = x$. Sometimes, more complex marginalizations are used, but the notation remains consistent. For instance $\pi^{(x,z)}(x,y,z,\dots) = (x,z)$. We denote by id the identity map.

**Domains** We denote $\Omega$ the closure of a Lipschitz, connected, bounded domain of $\mathbb{R}^d$, $\partial\Omega$ its boundary and $\Omega^\varepsilon := \Omega + B(0,\varepsilon)$ where $B^d(x,\varepsilon)$ is the closed ball of radius $\varepsilon$ centered at $x$ in $\mathbb{R}^d$. The geodesic distance between two points $(x_0,x_1) \in \Omega$ is denoted $\mathrm{dist}(x_0,x_1)$ (and equals $|x_1 - x_0|$ if $\Omega$ is convex). For a function $f : ]a,b[ \times \Omega \to \mathbb{R}$, we denote $\partial_t f$ its partial derivative w.r.t. the first variable, by $\nabla f$ its gradient w.r.t. the variable in $\Omega$ and if $f$ is valued in $\mathbb{R}^d$, then $\nabla \cdot f$ denotes the divergence operator w.r.t. the variable in $\Omega$. These operator are, generally understood in the sense of distribution and act on measures as well.

**Spaces of measures.** $\mathscr{M}(X;E)$ is the set of finite Radon measures on a topological space taking their values in $E$ ($E$ is typically $\mathbb{R}_+$ or a subset of a finite dimensional vector space). For conciseness, we write $\mathscr{M}(X) = \mathscr{M}(X;\mathbb{R})$ and $\mathscr{M}_+(X) = \mathscr{M}(X;\mathbb{R}_+)$. For a measure $\mu \in \mathscr{M}(X;\mathbb{R}^n)$ and a norm $|\cdot|$ on $\mathbb{R}^n$, we denote $|\mu| \in \mathscr{M}_+(X)$ its variation and $\mathrm{d}\mu/\mathrm{d}\nu$ its density w.r.t. another measure $\nu \in \mathscr{M}_+(X)$. The pushforward map associated to a Borel map $T$ is written $T_\#$. For a measure $\mu \in \mathscr{M}_+(X \times \mathbb{R}_n)$, the partial expectation operator is denoted $\mathfrak{h}$ and defined as $\mathfrak{h}\mu(\mathrm{d}x) = \int_{\mathbb{R}^n} u \cdot \mu(\mathrm{d}x, \mathrm{d}u)$. The notation $\otimes$ is used for product of measures and for measures generated by disintegration as well. We tried to keep consistent notations for the optimal transport costs and their generalizations throughout the thesis:

- $C_c$ is a *standard* optimal transport cost with the ground cost $c$;
- $C_L$ is a *dynamic* unbalanced optimal transport cost with Lagrangian $L$;
- $C_c^{\mathfrak{h}}$ is a *optimal lift* transport cost with a cost on the lifted space $c$;
- $C_{c,f}$ is an optimal *entropy-transport* cost, with cost $c$ and entropy function $f$.

**Spaces of functions.** We denote by $\mathscr{C}(X;E)$ the set of continuous and bounded functions on a set $X$ and by $\mathscr{C}^k(\Omega;E)$ the set of $k$ times continuously differentiable functions on $\Omega$ (more rigorously, restrictions to $\Omega$ of $\mathscr{C}^k$ functions on $\mathbb{R}^d$ endowed with the norm $\|f\| = \inf\{\|\tilde{f}\|_{\mathscr{C}^k} \ ; \ \tilde{f} \text{ extends } f\}$). For $p \in [1,\infty]$, we denote by $L^p(X)$ or $L^p(\mu)$ the usual spaces of (equivalence classes of) measurable functions on a measured space $(X,\mu)$. Also, $L^p_+(X)$ and $L^p_{++}(X)$ denote the subset of $L^p(X)$ of functions which are nonnegative and positive, respectively. The Sobolev subspace of $L^2(\mathbb{R}^d)$ of functions with $L^2$ distributional gradient is denoted $H^1(\mathbb{R}^d)$.

**Convex analysis.** For a function $f : V \to \mathbb{R} \cup \{\infty\}$, $\operatorname{dom} f$ is the domain of a function (the set of points where it is finite), $\operatorname{im} f$ its range, $\partial f$ is its (convex) subdifferential. Its convex conjugate is $f^*$. We use also $A^*$ for the adjoint of a linear operator. For a convex set $C$, we denote $\iota_C$ its convex indicator, that is the function worth 0 on $C$ and $\infty$ outside. The brackets $\langle \cdot, \cdot \rangle$ denotes a duality pairing between paired spaces. The pairings that are considered in this thesis are, with $E$ a Euclidean space (and $X$ compact for the first case):

$$\mathscr{C}(X;E), \text{ sup norm topology} \leftrightarrow \mathscr{M}(X,E), \text{ weak topology}$$
$$L^\infty(X;R), \text{ ess-sup norm topology} \leftrightarrow L^1(X), \text{ weak* topology.}$$

The notation "$\cdot$" is reserved for multiplication or the standard inner product on $\mathbb{R}^n$. Appendix A is devoted to the essential facts of convex analysis used in this thesis.

**Families, discrete notations** Families index by a set $I$ are denoted $(x_i)_{i \in I}$ or, if the set $I$ is clear from context, simply $(x_i)$. For two vectors $a, b \in \mathbb{R}^n$, we denote by $a \odot b$ their elementwise product and by $a \oslash b$ their elementwise division with the convention $0/0 = 0$. For $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$, we define $(u \oplus v)_{i,j} = u_i + v_j$ for $i = 1, \dots, n$ and $j = 1, \dots, m$.

**Perspective functions** We denote by $\psi_f(a|b)$ the perspective of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ (Definition A.0.11). This notation with a vertical bar allows to distinguish the "depth parameter" which is always a real, from to the other arguments which can be vectors. This notation is consistent with the standard notation for $f$-divergences (a.k.a. Csiszár divergences), a concept recovered in the case when $f$ is an entropy function.

**Homogeneous functions of measures** In this thesis, homogeneous functions of measures are ubiquitous and in order to maintain consistent and lightweight notation throughout, we decided to adopt a somehow unusual (but not new [56]) notation. For a positively homogeneous function $\psi : E \to F$ between Euclidean spaces and a measure $\mu \in \mathscr{M}(\Omega;E)$, we denote by $\psi(\mu)$ the measure in $\mathscr{M}(\Omega;F)$ defined by

$$\psi(\mu)(\mathrm{d}x) := \psi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\right)\lambda(\mathrm{d}x).$$

where $\lambda \in \mathscr{M}_+(\Omega)$ is any measure that dominates $\mu$. This definition does not depend on the choice of $\lambda$. In the case where $\psi = \psi_L(\cdot|\cdot) : E \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is the perspective function of a convex function $L : E \to \mathbb{R} \cup \{\infty\}$, then using the Lebesgue decomposition $\mu = \sigma\rho + \mu^\perp$, it follows

$$\psi_L(\mu|\rho) = L(\sigma)\rho + L'_\infty(\mu^\perp)$$

where $L'_\infty$ is the recession function of $f$, itself a sublinear functional.

# Part I.

# Models for Unbalanced Optimal Transport

# Chapter 1.

# Formulations and Equivalences

In this chapter, we propose a unified theory of unbalanced optimal transport, based on models that fall into two categories: the dynamic formulations and the coupling formulations.
In Section 1.1, we introduce a new general framework for dynamic formulations of unbalanced optimal transport problems. The two basic concepts are (i) the continuity equation with source and (ii) an action functional based on the choice of a convex Lagrangian. A variational problem allows then to define interpolations between nonnegative measures with a mixture of mass displacement and mass growth. We study standard properties of this variational problem, show that it can be used to define geodesic metrics, and that the minimum enjoys continuity properties w.r.t. weak convergence.
In Section 1.2, we propose another class of formulations based on the new concept of *semi-couplings*. It enters in the definition of a transport cost between nonnegative measures, which can be used to generate metrics that metrize weak convergence. We also mention two formulations due to [103] and show their connexions to the semi-coupling problem: on the one hand, every optimal entropy-transport problems can be recast as a semi-coupling problem (this is implicit from the results in [103]), on the other hand, we prove that their *optimal lift* formulation can be also recast as a semi-coupling problem involving the *sublinear regularization* of the cost function.
In Section 1.3, we prove that any dynamic problem admits a semi-coupling formulation, thus completing the web of relationships between these models and unifying the theory of unbalanced optimal transport.
The content of this chapter is based on a submitted paper [43]. Many results have been added or improved specifically for this thesis.

# 1.1. Dynamic formulations

## 1.1.1. Introduction

**Motivation** In this section, we describe a first approach to unbalanced optimal transport. Inspired by the *dynamic* formulation of classical optimal transport, we look for the *interpolation* (or path) between two measures $\mu \in \mathscr{M}_+(\Omega)$ and $\nu \in \mathscr{M}_+(\Omega)$ that minimizes an *action* functional. What is meant by interpolation, is a family of measures $(\rho_t)_{t\in[0,1]}$ indexed by a parameter $t$ akin to time, which satisfies $\rho_0 = \mu$ and $\rho_1 = \nu$. An action functional is an integral functional summing up the "effort" required to generate the path $(\rho_t)$.

**Setting** The technical setting in that chapter is that of measures on a domain $\Omega \subset \mathbb{R}^d$ which is the closure of an open bounded and connected set with Lipschitz boundary. These regularity conditions allow us to join every pair of points by an absolutely continuous path of finite length, and also to have for any pair of points $(x,y) \in \Omega^2$, continuity as $\varepsilon \downarrow 0$ of the distance in $\Omega^\varepsilon$.

The original motivation of this work was to build interpolations between shapes or images (represented as measures) through convex variational problems. This explains the choice of a compact setting as well as the differences that can be found in our choices of definitions or interpretations, compared to a "PDE" approach. For instance, we do not assume a priori that solutions to the continuity equation are weakly continuous.

## 1.1.2. Continuity equation with source

In the theory of standard optimal transport, time variations of the interpolation $\partial_t \rho_t$ are exclusively explained by displacement, and are thus generated by the flux (minus the divergence) of some *momentum* $\omega_t \in \mathscr{M}(\Omega;\mathbb{R}^d)$. This implies that only the flux of mass through the boundary of $\Omega$ can explain variations of the total mass $\rho_t(\Omega)$—in particular, $\rho_t(\Omega)$ is constant if no-flux boundary conditions are enforced. In a general model of unbalanced optimal transport, it is desirable to take into account local variations of mass, described by a source $\zeta_t \in \mathscr{M}(\Omega)$. The continuity equation with source defined hereafter enforces the mass balance of this dynamic: at any point in time and space, the variation of mass equals the (signed) source of mass minus the divergence of the flux:

$$\partial_t \rho + \nabla \cdot \omega = \zeta. \tag{1.1.1}$$

This definition makes sense for smooth fields, but we typically assume much less regularity. We do not even assume a priori that the time marginals admit a density with respect to Lebesgue and thus consider triplets of finite measures $(\rho,\omega,\zeta)$ with $\rho \in \mathscr{M}_+([0,T]\times\Omega)$, $\omega \in \mathscr{M}([0,T]\times\Omega;\mathbb{R}^d)$ and $\zeta \in \mathscr{M}([0,T]\times\Omega)$. The rigorous definition, with no-flux boundary conditions, is as follows.

**Definition 1.1.1** (Continuity equation with source)**.** *For $(\mu,\nu) \in \mathscr{M}_+(\Omega)^2$, a triplet of measures $(\rho,\omega,\zeta)$ is said to satisfy the continuity equation between $\mu$ and $\nu$ in the sense of distributions on $[0,T]$ and we write $(\rho,\omega,\zeta) \in \mathrm{CE}_0^T(\mu,\nu)$ if*

$$\int_0^T \int_\Omega \partial_t \varphi \,\mathrm{d}\rho + \int_0^T \int_\Omega \nabla\varphi \cdot \mathrm{d}\omega + \int_0^T \int_\Omega \varphi \,\mathrm{d}\zeta = \int_\Omega \varphi_T \,\mathrm{d}\nu - \int_\Omega \varphi_0 \,\mathrm{d}\mu, \tag{1.1.2}$$

*for all* $\varphi \in \mathscr{C}^1([0,T]\times\Omega)$ *with the notation* $\varphi_t := \varphi(t,\cdot)$. *One can remove the time boundary constraints by testing against* $\varphi \in \mathscr{C}^1_c(]0,T[\times\Omega)$.

With this continuity equation, taking a "time slice" of $\rho$ is permitted and variations of the total mass $\rho_t(\Omega)$ are described by the time marginals of $\zeta$, as formalized in the next lemma.

**Lemma 1.1.2** (Variation of total mass)**.** *If* $(\rho,\omega,\zeta)$ *solves the continuity equation, then* $\rho$ *admits a disintegration with respect to the Lebesgue measure in time* $\rho = \rho_t \otimes \mathrm{d}t$ *and* $t \to \rho_t(\Omega)$ *is of bounded variation, with distributional derivative* $\pi^t_\# \zeta \in \mathscr{M}([0,T])$.

*Proof.* Testing (1.1.2) against functions $\varphi$ which are constant in space, one finds that

$$\int_0^T \varphi'(t)\,\mathrm{d}(\pi^t_\#\rho) + \int_0^T \varphi(t)\,\mathrm{d}(\pi^t_\#\zeta) = \varphi(T)\,\nu(\Omega) - \varphi(0)\,\mu(\Omega),$$

where $\pi^t : (t,x) \mapsto t$. Since $\pi^t_\#\zeta \in \mathscr{M}([0,T])$ is a finite measure, one has $\pi^t_\#\rho = m\mathrm{d}t$ for a function $m : [0,T] \to \mathbb{R}_+$ of bounded variation. By the disintegration theorem [5, Thm. 5.3.1], one can always write $\rho = \tilde\rho_t \otimes \pi^t_\#\rho$ with $\tilde\rho_t \in \mathscr{P}(\Omega)$ for all $t$, so the claim follows by posing $\rho_t = m(t)\tilde\rho_t$ which satisfies $\rho_t(\Omega) = m(t)$. □

When the momentum and source variables are disintegrable w.r.t. Lebesgue in time, a finer result is possible: the interpolation $(\rho_t)$ is weakly continuous and solves a *weak* formulation of the continuity equation. The following is adapted from [142, 5].

**Proposition 1.1.3** (Weak formulation)**.** *Assume that the time marginal of* $\omega$ *and* $\zeta$ *admit densities w.r.t. Lebesgue, so that one can disintegrate them as* $\omega = \omega_t \otimes \mathrm{d}t$ *and* $\zeta = \zeta_t \otimes \mathrm{d}t$. *Then every distributional solution of the continuity equation* (1.1.2) *is* $\mathrm{d}t$*-a.e. equal to a weakly continuous curve* $(\tilde\rho_t)$ *that satisfies* $\mathrm{d}t$*-a.e.*

$$\frac{d}{dt}\left(\int_\Omega \psi\,\mathrm{d}\tilde\rho_t\right) = \int_\Omega \nabla\psi\cdot\mathrm{d}\omega_t + \int_\Omega \psi\mathrm{d}\zeta_t \tag{1.1.3}$$

*for all* $\psi \in \mathscr{C}^1(\Omega)$. *Reciprocally, any weakly continuous curve which solves* (1.1.3) *and such that* $(\tilde\rho_0,\tilde\rho_T) = (\mu,\nu)$ *solves* (1.1.2)*.*

*Proof.* Let $(\rho,\omega,\zeta)$ be a distributional solution of which the time marginals are absolutely continuous w.r.t. Lebesgue and let us test it against functions $\varphi$ of the form $\varphi(t,x) = a(t)\psi(x)$ with $\psi \in \mathscr{C}^1(\Omega)$ and $a \in \mathscr{C}^1_c(]0,T[)$. We get

$$\int_0^T a'(t)\int_\Omega \psi(x)\,\mathrm{d}\rho_t\mathrm{d}t + \int_0^T a(t)\left(\int_\Omega \nabla\psi(x)\cdot\mathrm{d}\omega_t + \int_\Omega \psi(x)\,\mathrm{d}\zeta_t\right)\mathrm{d}t = 0$$

so the map $\rho_t(\psi) : t \mapsto \int_\Omega \psi\,\mathrm{d}\rho_t$ admits a distributional derivative in $L^1(]0,T[)$ as written in (1.1.3). Now we argue as in [5, Lem. 8.1.2]: let $Z$ be a dense countable subset of $\mathscr{C}^1(\Omega)$ and let $L_Z := \cap_{\psi\in Z} L_\psi$ be the intersection of the sets $L_\psi$ of Lebesgue points of $t \mapsto \rho_t(\psi)$. As a countable union of zero measure sets, $]0,T[\setminus L_Z$ is of zero Lebesgue measure. The restriction of

the curve $(\rho_t)$ to $L_Z$ provides a uniformly continuous family of bounded functionals on $\mathscr{C}^1(\Omega)$ since

$$|\rho_t(\psi)-\rho_s(\psi)| \le \|\psi\|_{\mathscr{C}^1(\Omega)} \int_s^t (|\omega_r|(\Omega)+|\zeta_r|(\Omega))\mathrm{d}r, \quad \forall s,t \in L_Z, \forall \psi \in \mathscr{C}^1(\Omega).$$

Therefore, it admits a unique extension to a continuous curve $(\tilde{\rho}_t)_{t\in[0,T]}$ in $\mathscr{C}^1(\Omega)^*$. Yet, the set $\{\rho_t\}_{t\in L_Z}$ is weakly sequentially compact in $\mathscr{M}_+(\Omega)$ because $\{\rho_t(\Omega)\}_{t\in L_Z}$ is bounded by Lemma 1.1.2 and $\Omega$ is compact. So the extension curve $\tilde{\rho}_t$ actually belongs to $\mathscr{M}_+(\Omega)$ for all $t \in [0,T]$.

For the reciprocal statement, remark as in [142] that weak solutions satisfy (1.1.2) for any $\varphi$ of the form $\varphi(t,x) = a(t)\psi(x)$ and the linear space generated by such separable functions is dense in $\mathscr{C}^1([0,T]\times\Omega)$. □

**Example 1.1.4.** *There exists weak solutions to the continuity equation between any pair $\mu,\nu \in \mathscr{M}_+(\Omega)$. For instance, with $T=1$, one can easily verify that* (1.1.3) *holds for the following triplets:*

- linear interpolation: $\rho_t = (1-t)\mu + t\nu$, $\omega = 0$ *and* $\zeta_t = \nu - \mu$*;*
- coupling-based interpolation: *if $\mu(\Omega) = \nu(\Omega)$, choose a family of differentiable paths $t \mapsto e_t(x,y)$ of uniformly bounded velocity linking any pair of points $(x,y) \in \Omega^2$, then let $\gamma \in \Pi(\mu,\nu)$ be a coupling and define $\rho_t = (e_t)_\# \gamma$, $\omega_t = (e_t)_\#(\gamma \partial_t e_t)$ and $\zeta = 0$.*
- displacement interpolation: *same as above, but choosing $\gamma$ as an optimal coupling for the cost $c(x,y) = |y-x|^p$, with $p \ge 1$ and $t \mapsto e_t(x,y)$ a constant speed parametrization of a geodesic in $\Omega$;*

*One sees that "deformations" of the form $(0,\zeta_t)$ correspond to the Banach space structure (in this case $\zeta_t$ is called the* weak* differential *of the path) and those of the form $(\omega_t,0)$ are those appearing in classical optimal transport. In the dynamic formulation of unbalanced optimal transport, we combine these two types of deformations.*

When the momentum and the source are generated by specified velocity and growth fields that have some regularity, the initial value problem has a unique solution in the form of a flow. This result is taken from [108, Prop. 3.6] and adapted to our compact setting.

**Theorem 1.1.5** (Uniqueness)**.** *Let $v_t : \Omega \to \mathbb{R}^d$ be a* velocity *field that is Lipschitz continuous in $x$, uniformly in $t$, and satisfies Neumann boundary conditions on $\partial\Omega$ and consider its flow $Y_t$ on $[0,T]$. Let $g_t : \Omega \to \mathbb{R}$ be a* rate of growth *field that is Lipschitz continuous in $x$ and bounded, uniformly in $t$, and consider the total growth $G_t(x) = \exp(\int_0^t g_s(Y_s(x))\mathrm{d}s)$. Then $\rho_t = (Y_t)_\#(G_t\rho_0)$ is the unique weakly continuous solution to the continuity equation starting from $\rho_0 \in \mathscr{M}_+(\Omega)$ and such that $(\omega_t,\zeta_t) = (v_t\rho_t, g_t\rho_t)$.*

In the context of Theorem 1.1.5, the continuity equation can be rewritten with the more standard variables of speed and rate of growth:

$$\partial_t \rho_t + \nabla\cdot(v_t\rho_t) = g_t\rho_t \tag{1.1.4}$$

Let us prove some additional results that are used in the proofs of this chapter. We first study the stability of solutions of the continuity equation under linear time or space rescaling.

**Proposition 1.1.6** (Linear scaling)**.** *Let $(\rho,\omega,\zeta)$ solve the continuity equation on $[0,T]\times\Omega$ between $\mu,\nu\in\mathscr{M}_+(\Omega)$ and let $s:[0,T]\times\Omega\to[0,\alpha T]\times\beta\Omega$ be a time and space scaling $s(t,x)=(\alpha t,\beta x)$ with $\alpha,\beta>0$. Then $s_\#(\alpha\rho,\beta\omega,\zeta)$ solves the continuity equation between $\bar{s}_\#\mu$ and $\bar{s}_\#\nu$ on $[0,\alpha T]$, where $\bar{s}(x)=\beta x$.*

*Proof.* Let $\psi\in\mathscr{C}^1([0,\alpha T]\times\beta\Omega)$ and define $\varphi:=\psi\circ s$ so that $\partial_t\varphi=\alpha(\partial_t\psi)\circ s$ and $\nabla\varphi=\beta(\nabla\psi)\circ s$. Since (1.1.2) holds for $(\rho,\omega,\zeta)$, it follows

$$\begin{aligned}\int_{\beta\Omega}\psi_{\alpha T}\,\mathrm{d}(\bar{s}_\#\nu)-\int_{\beta\Omega}\psi_0\,\mathrm{d}(\bar{s}_\#\mu)&=\int_\Omega\varphi_T\,\mathrm{d}\nu-\int_\Omega\varphi_0\,\mathrm{d}\mu\\&=\int_0^T\int_\Omega\partial_t\varphi\,\mathrm{d}\rho+\int_0^T\int_\Omega\nabla\varphi\cdot\mathrm{d}\omega+\int_0^T\int_\Omega\varphi\,\mathrm{d}\zeta\\&=\int_0^T\int_\Omega\alpha\partial_t\psi\circ s\,\mathrm{d}\rho+\int_0^T\int_\Omega\beta(\nabla\psi)\circ s\cdot\mathrm{d}\omega+\int_0^T\int_\Omega\varphi\,\mathrm{d}\zeta\end{aligned}$$

and thus $s_\#(\alpha\rho,\beta\omega,\zeta)$ satisfies (1.1.2) with the boundary conditions announced. □

In the more specific case when the time marginals admit a density, a time reparametrization is possible. The proof is a direct adaptation of [5, Lem. 8.1.3] since in this case the weak formulation of Proposition 1.1.3 holds (it also follows by adapting the previous proof).

**Proposition 1.1.7** (Time rescaling)**.** *Let $\mathrm{t}:[0,T']\to[0,T]$ be a strictly increasing absolutely continuous map with absolutely continuous inverse. Then $(\rho_t,\omega_t,\zeta_t)_{t\in[0,T]}$ is a weak solution of the continuity equation if and only if $(\rho_t\circ\mathrm{t},\mathrm{t}'(\omega_t\circ\mathrm{t}),\mathrm{t}'(\omega_t\circ\mathrm{t}))$ also is, on $[0,T']$.*

The following proposition explains how to extend the time range of solutions. Combined with the linearity property of the continuous equation, this allows to concatenate solutions in time, a result known as the glueing lemma.

**Proposition 1.1.8** (Extension)**.** *Let $(\rho,\omega,\zeta)$ be a solution to the continuity equation* (1.1.2) *on $[0,T]\times\Omega$ between $\mu$ and $\nu$, and extend these measures by $0$ on $\mathbb{R}\setminus[0,T]$. Then $(\rho,\omega,\zeta+\mu\otimes\delta_0-\nu\otimes\delta_T)$ solves the continuity equation on any interval $]a,b[\supset[0,T]$.*

*Proof.* It is direct by plugging into (1.1.2). □

The last proposition is classical for distributional solutions of PDEs.

**Proposition 1.1.9** (Smoothing)**.** *Let $(\rho,\omega,\zeta)$ be a distributional solution to the continuity equation* (1.1.2) *on $]0,T[\times\Omega$, and let $r_\varepsilon:\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}$ be a mollifier supported on $B(0,\frac{\varepsilon}{2})\times B^d(0,\frac{\varepsilon}{2})$. Then $(r_\varepsilon*\rho,r_\varepsilon*\omega,r_\varepsilon*\zeta)$ is a classical solution of* (1.1.1) *on $]\frac{\varepsilon}{2},T-\frac{\varepsilon}{2}[\times\mathbb{R}^d$.*

### 1.1.3. Action minimizing problems on $\mathscr{M}_+(\Omega)$

#### Definitions

In order to select an interpolation among all the solutions we choose the interpolation that minimizes an *action*. We pose an action that is the integral over time and space of a convex Lagrangian function $L(v,g)$ which is a function of the instantaneous "deformation" $(v,g)$ (speed

and rate of growth) applied to a particle of unit mass[1]. This leads to a functional of the form

$$\int_0^1 \int_\Omega L(v_t(x), g_t(x)) \mathrm{d}\rho_t(x) \mathrm{d}t.$$

However, defining the variational problem directly in terms of the Lagrangian has drawbacks: (i) in the variables $(v,g)$, the continuity equation (1.1.4) is not linear so this leads to a non-convex problem (ii) sometimes, the velocity and growth fields are not well defined because the momentum $\omega$ and the source $\zeta$ have a singular part w.r.t. $\rho$ (this is allowed by Lagrangians which are not superlinear).

These two concerns are fixed if one considers the change of variables $(\omega,\zeta) := (v\rho, g\rho)$. In physical terms, this corresponds to switching from intensive to extensive variables. Notice that one has (for $v,g,\rho$ positive real numbers)

$$L(v,g)\rho = \rho L(\omega/\rho, \zeta/\rho) = \psi_L(\omega,\zeta|\rho)$$

where $\psi_L$ is the so-called *perspective function* of $L$. This construction, detailed in Appendix A, has the property that if $L$ is convex in $(v,g)$, then $\psi_L$ is sublinear jointly in all variables, see Definition A.0.11.

**Assumption 1.1.10.** *The Lagrangian $L : \mathbb{R}^d \times \mathbb{R}$ is convex, continuous and admits the unique minimum $L(0,0) = 0$. We denote by $\psi_L$ its perspective function.*

The proof of the following is in Appendix A, Proposition A.0.13.

**Proposition 1.1.11** (Conjugate of $\psi_L$)**.** *The functions $\psi_L$ and $\iota_{Q_L}$, the convex indicator of the set*

$$Q_L := \{(a,b,c) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \,;\, a + L^*(b,c) \leq 0\}$$

*form a pair of convex conjugate functions.*

Note that it is natural to consider Lagrangians which are separable in $v$ and $g$, i.e. of the form $L_1(v) + L_2(g)$ in which case one has $L^*(b,c) = L_1^*(b) + L_2^*(c)$. The action functional is defined as the integral of $\psi_L$ over time and space.

**Definition 1.1.12** (Action functional)**.** *The action functional is defined as*

$$\mathbb{A}(\rho,\omega,\zeta) := \int_0^1 \int_\Omega \psi_L(\omega,\zeta|\rho)$$

*where we use the notation described in the section* Notation *for the image of a measure by a homogeneous function. When L is superlinear in all directions, i.e. $L'_\infty = \infty$, this reduces to*

$$\mathbb{A}(\rho,\omega,\zeta) = \begin{cases} \int_0^1 \int_\Omega L(v,g)\mathrm{d}\rho & \text{if } \omega,\zeta \ll \rho \text{ and } (\omega,\zeta) = (v\rho, g\rho), \\ \infty & \text{otherwise.} \end{cases}$$

The dynamic formulation of unbalanced optimal transport is defined as a minimization of the action over the solutions to the continuity equation between two endpoints $\mu,\nu \in \mathscr{M}_+(\Omega)$.

**Definition 1.1.13** (Dynamic formulation)**.** *For $\mu,\nu$ two measures in $\mathscr{M}_+(\Omega)$, define*

$$\boxed{C_L(\mu,\nu) := \inf\left\{\mathbb{A}(\rho,\omega,\zeta) \,;\, (\rho,\omega,\zeta) \in \mathrm{CE}_0^1(\mu,\nu)\right\}.}$$

[1] A continuous, multiplicative dependency in $(t,x)$ is also practicable but we do not consider it here for ease of reading. The dependance in $x$ is treated in our article [43].

### Variational properties

In most cases, we can guarantee a finite cost $C_L(\mu,\nu)$.

**Proposition 1.1.14** (Finite cost)**.** *Let $\mu,\nu\in\mathscr{M}_+(\Omega)$.*

- *if $L(0,\cdot)$ has at most a polynomial growth then $C_L(\mu,\nu)<\infty$.*
- *If $\mu$ and $\nu$ are different from $0$ then $C_L(\mu,\nu)<\infty$;*

*Proof.* For the first case, consider a family $(\rho_t,0,\omega_t)_{t\in[0,1]}$ with $\rho_t=t^\alpha\mu$ and $\omega_t=\alpha t^{\alpha-1}\mu$ where $\alpha>0$ is such that $L(0,g)=o(g^\alpha)$ (using Landau $o$ notation). One has

$$\mathbb{A}(\rho,0,\omega)=\int_0^1\int_\Omega L(0,\frac{\alpha}{t})t^\alpha\,\mathrm{d}\mu\,\mathrm{d}t<\mu(\Omega)\int_0^1 L(0,\frac{\alpha}{t})t^\alpha\,\mathrm{d}t<\infty$$

so $C_L(0,\mu)<\infty$. For any $\mu,\nu\in\mathscr{M}_+(\Omega)$, one builds similarly an interpolation that goes through $0$ at $t=\frac{1}{2}$ and the conclusion is the same.

For the second case, first assume that $\mu(\Omega)=\nu(\Omega)$. Take the coupling-based interpolation from Example 1.1.4: let $\gamma=\mu\otimes\nu$ be the product coupling and let $(\rho_t,\omega_t,\zeta_t)=((e_t)_\#\gamma,(e_t)_\#(\gamma\partial_t e_t),0)$ for $t\in[0,1]$, where $(t\mapsto e_t(x,y))$ is a family of constant speed absolutely continuous paths of finite length between pairs of points $(x,y)\in\Omega^2$. There is a uniform bound on the velocities $|\partial_t e_t|<M$ and since a uniform bound on a relative density is preserved by pushforward (the pushforward operator is order preserving), one has $|\mathrm{d}\omega_t/\mathrm{d}\rho_t|<M$. Thus the action is finite:

$$\mathbb{A}(\rho,\omega,\zeta)=\int_0^1\int_\Omega L\left(\frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t},0\right)\mathrm{d}((e_t)_\#\gamma)<L(M,0)\;\mu(\Omega)<\infty.$$

Finally if $\alpha:=\nu(\Omega)/\mu(\Omega)\neq 1$ then we define an admissible interpolation in two phases: during time $[0,\frac{1}{2}]$ define a uniform growth $\rho_t=\alpha^{2t}\mu$ and $\omega_t=\partial_t\rho_t$ and during time $[\frac{1}{2},1]$ define an interpolation as above (using $\bar{e}_t=e_{2t-1}$). The associated action is finite because the rate of growth $2\log\alpha$ is constant in the first phase, and $|\partial_t\bar{e}_t|<2M$ in the second. □

The dynamic problem admits a dual formulation which domain are the set of subsolutions to a Hamilton-Jacobi equation. In contrast to classical optimal transport, this equation involves a zeroth order term.

**Theorem 1.1.15** (Duality)**.** *If $C_L(\mu,\nu)<\infty$, there exists a minimizing triplet $(\rho,\omega,\zeta)$ and*

$$C_L(\mu,\nu)=\sup_{\varphi\in\mathscr{C}^1([0,1]\times\Omega)}\left\{\int_\Omega\varphi_1\mathrm{d}\nu-\int_\Omega\varphi_0\,\mathrm{d}\mu\;;\;\partial_t\varphi+L^*(\nabla\varphi,\varphi)\leq 0\right\}\tag{1.1.5}$$

*where the constraint (with an abuse of notations) stands for all $(t,x)\in[0,1]\times\Omega$.*

*Proof.* The right-hand side of (1.1.5) can be written as

$$-\inf_{\varphi\in\mathscr{C}^1([0,1]\times\Omega)}F(A\varphi)+G(\varphi)$$

where $A : \varphi \mapsto (\partial_t \varphi, \nabla \varphi, \varphi)$, is a bounded linear operator from $\mathscr{C}^1([0,1] \times \Omega)$ to $\mathscr{C}([0,1] \times \Omega)^{d+2}$, $F$ is the convex indicator of the closed set of triplets of continuous functions $(a,b,c)$ such that $a + L^*(b,c) \leq 0$ everywhere and $G : \varphi \mapsto \int_\Omega \varphi(0,\cdot) \mathrm{d}\rho_0 - \int_\Omega \varphi(1,\cdot) \mathrm{d}\rho_1$ is linear. $F$ and $G$ are convex, proper and l.s.c. functionals.

Besides, the assumption that $L$ attains its unique minimum at $L(0,0) = 0$ implies that $L^*(0,0) = 0$ and $L^*$ is differentiable at $(0,0)$ of differential $(0,0)$. It follows that $|L^*(0,c)|/c \to 0$ when $c \to 0$, so there exists $\varepsilon > 0$ such that $L^*(0, \theta\varepsilon/2) < \varepsilon$ for $\theta \in [-1,1]$ and thus the function $\varphi : (t,x) \mapsto -\varepsilon t + \varepsilon/2$ is such that $F(A\varphi) + G(\varphi) < +\infty$ and $F$ is continuous at $A\varphi$. Then, by Fenchel-Rockafellar duality (Appendix A), (1.1.5) is equal to

$$\min_{\mu \in \mathscr{M}([0,1]\times\Omega)^{1+d+1}} G^*(-A^*\mu) + F^*(\mu).$$

By Theorem A.0.14 (Appendix A) we have $F^* = \mathbb{A}$, and by direct computations, $G^* \circ (-A^*)$ is the convex indicator of $\mathrm{CE}_0^1(\mu,\nu)$. □

A first interesting property, which is true but trivial with standard optimal transport, is that the dynamic cost is itself a sublinear function jointly in both its arguments.

**Corollary 1.1.16** (Sublinearity of $C_L$)**.** *For $\alpha > 0$ and $\mu, \nu, \bar{\mu}, \bar{\nu} \in \mathscr{M}_+(\Omega)$, it holds*

$$C_L(\alpha\mu, \alpha\nu) = \alpha C_L(\mu,\nu) \quad \textit{and} \quad C_L(\mu+\bar{\mu}, \nu+\bar{\nu}) \leq C_L(\mu,\nu) + C_L(\bar{\mu},\bar{\nu}).$$

*Proof.* The 1-homogeneity of $C_L$ is a consequence of the homogeneity of $\psi_L$ and the stability of the continuity equation under multiplication by a scalar. The subadditivity is also inherited from the subadditivity of $\psi_L$ and the stability of the continuity equation by sum of solutions. Alternatively, one may notice that by Theorem 1.1.15, $C_L$ is the conjugate function of the indicator of a convex set and, as such, sublinear. □

### Sufficient optimality and uniqueness conditions

We now leverage tools from convex analysis in order to provide a useful condition ensuring uniqueness of geodesics. We used these optimality conditions in [44] to prove the form of the geodesics of $\widehat{W}_2$ (see Chapter 2) between certain kinds of atomic measures. Although this result might have its own interest since we exhibited the explicit form of the dual variables, I chose to not reproduce it here because the proof is quite lengthy and the result less instructive than the ones in Section 1.3.

**Lemma 1.1.17** (Subdifferential)**.** *Let $(\rho,\omega,\zeta)$ be a triplet of finite action and write the Lebesgue decompositions $\omega = v\rho + \omega^\perp$ and $\zeta = g\rho + \omega^\perp$. The subdifferential of $\mathbb{A}$ at $(\rho,\omega,\zeta)$ is the set of triplets $(a,b,c) \in \mathscr{C}([0,1]\times\Omega)$ satisfying*

- $a + L^*(b,c) \leq 0$ *everywhere*
- $a + L^*(b,c) = 0$ *and* $(b,c) \in \partial L(v,g)$ $\rho$*-a.e.*
- $(b,c) = \partial L'_\infty(\frac{\omega^\perp}{\lambda}, \frac{\zeta^\perp}{\lambda})$ $\lambda$*-a.e. where* $\lambda = |\omega^\perp| + |\zeta^\perp|$.

*Proof.* Write $\mu = (\rho, \omega, \zeta)$ and let $\bar{\mu}$ be another triplet. Let $m = \rho + |\omega| + |\zeta|$ and let $m^\perp \in \mathscr{M}_+([0,1]\times\Omega)$ be singular to $m$ and such that $\bar{\mu} \ll m + m^\perp$. For $\phi = (a,b,c) \in \partial\mathbb{A}(\mu)$, it holds

$$\begin{aligned}\mathbb{A}(\bar{\mu}) - \mathbb{A}(\mu) &= \int \left( \psi_L(\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}m}) - \psi_L(\frac{\mathrm{d}\mu}{\mathrm{d}m}) \right) \mathrm{d}m + \int \psi_L(\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}m^\perp})\mathrm{d}m^\perp \\ &\geq \langle \phi, (\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}m} - \frac{\mathrm{d}\mu}{\mathrm{d}m})m \rangle + \langle \phi, (\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}m^\perp})m^\perp \rangle \\ &= \langle \phi, \bar{\mu} - \mu \rangle\end{aligned}$$

where the inequality holds for any $\bar{\mu}$ if and only if $\phi(t,x)$ is in the subdifferential of $\psi_L(\frac{\mathrm{d}\mu}{\mathrm{d}m}(t,x))$ for $m$ almost every $(t,x)$ (for the first term) and $\phi(t,x)$ is in the set $\mathrm{dom}(\psi_L^*)$ everywhere (for the second term). The expression for $\partial\psi_L$ is proved, e.g. in [46]. □

**Theorem 1.1.18** (Sufficient optimality and uniqueness condition)**.** *If $(\rho,\omega,\zeta)$ solves the continuity equation between $\mu,\nu \in \mathscr{M}_+(\Omega)$ and there exists $\varphi \in \mathscr{C}^1([0,T]\times\Omega)$ such that*

$$(\partial_t\varphi, \nabla\varphi, \varphi) \in \partial\mathbb{A}(\rho,\omega,\zeta)$$

*then $(\rho,\omega,\zeta)$ is a minimizer for $C_L(\mu,\nu)$. If moreover $L$ is strictly convex and superlinear and $(\omega,\zeta) = (v\rho, g\rho)$ for some velocity/growth fields $(v,g)$ satisfying the regularity assumptions of Theorem 1.1.5, then it is the unique minimizer.*

*Proof.* With the notations of the proof of Theorem 1.1.15 and writing $\sigma = (\rho,\omega,\zeta)$, Fenchel-Rockafellar duality theorem (see Appendix A) also gives the following result: $(\sigma,\varphi)$ is a pair of optimizers for the primal and dual problems if and only if $A\varphi \in \partial F^*(\sigma)$ and $-A^*\sigma \in \partial G(\varphi)$. The first condition is our hypothesis and the second is satisfied since for all $\psi \in \mathscr{C}^1([0,1]\times\Omega)$,

$$\langle -A^*\sigma, \psi - \varphi \rangle = \langle \sigma, A\phi \rangle - \langle \sigma, A\psi \rangle = G(\psi) - G(\phi)$$

as $\sigma$ solves the continuity equation. This shows that $\sigma$ is a minimizer for $C_L(\mu,\nu)$.

For uniqueness, consider another minimizer $\tilde{\sigma} = (\tilde{\rho},\tilde{\omega},\tilde{\zeta})$ solving the continuity equation between $\mu,\nu$. It holds

$$\begin{aligned}\mathbb{A}(\tilde{\sigma}) &\geq \int_0^1 \int_\Omega \partial_t\varphi \,\mathrm{d}\tilde{\rho} + \int_0^1 \int_\Omega \nabla\varphi \cdot \mathrm{d}\tilde{\omega} + \int_0^1 \int_\Omega \varphi \,\mathrm{d}\tilde{\zeta} \\ &= \int_\Omega \varphi_1 \,\mathrm{d}\nu - \int_\Omega \varphi_0 \,\mathrm{d}\mu = \mathbb{A}(\tilde{\sigma})\end{aligned}$$

where we used successively: the fact that $\mathbb{A}$ is the conjugate of the indicator of a set to which $(\partial\varphi,\nabla\varphi,\varphi)$ belongs, the fact that $\tilde{\sigma}$ solves the continuity equation and the optimality of $\varphi$ in the dual problem. So, the first inequality is an equality, and hence $\lambda$ a.e., $(\partial_t\varphi,\nabla\varphi,\varphi)(t,x) \in \partial\mathbb{A}\left(\frac{\mathrm{d}\tilde{\sigma}}{\mathrm{d}\lambda}(t,x)\right)$ for $\lambda$ such that $\tilde{\sigma} \ll \lambda$. Since $L$ is assumed superlinear, $\omega = v\rho$ and $\zeta = g\rho$ for some $(v,g)$ and the strict convexity of $L$ implies that $(\partial L)^{-1} = \partial L^*$ is at most single valued so $(v,g)$ is the unique element in $\partial L^*(\nabla\varphi,\varphi)$. It follows from the characterization of $\partial\mathbb{A}$ that

$$\tilde{\sigma} = (\tilde{\rho}, v\tilde{\rho}, g\tilde{\rho}).$$

Thus, $\rho$ and $\tilde{\rho}$ are both solutions of $\partial_t\rho + \nabla\cdot(v\rho) = g\rho$ with initial condition $\mu$. This equation has a unique solution if the assumptions of Theorem 1.1.5 are satisfied. □

### Time reparametrizations

The following result shows that if the Lagrangian is (positively) $p$-homogeneous, i.e. if for all $\lambda \geq 0$ it holds $L(\lambda v, \lambda g) = \lambda^p L(v,g)$ for all $(v,g) \in \mathbb{R}^d \times \mathbb{R}$, then the minimizing interpolation has a constant "deformation rate" in $[0,1]$ as measured by the action. The case $p > 1$ is classical but the case $p = 1$ requires more care.

**Proposition 1.1.19** (Constant speed minimizers)**.** *If $L$ is $p$-homogeneous for $p > 1$, then minimizers $(\rho, \omega, \zeta)$ of* (1.1.13) *can be disintegrated in time w.r.t. Lebesgue and satisfy*

$$C_L(\rho_s, \rho_t) = |t-s|^p C_L(\mu, \nu). \tag{1.1.6}$$

*and the same holds true for some minimizers if $p = 1$. Moreover, for $p \geq 1$, one has for any $T > 0$,*

$$C_L(\mu,\nu)^{\frac{1}{p}} = \inf \left\{ \int_0^T \left( \int_\Omega \psi_L(\omega_t, \zeta_t | \rho_t) \right)^{\frac{1}{p}} \mathrm{d}t \right\} \tag{1.1.7}$$

*where the infimum runs over solutions to the continuity equation between $\mu, \nu$ on $[0,T]$.*

*Proof.* We first treat the case $p > 1$. By Proposition 1.1.14, we know that $C(\mu,\nu)$ is finite. Moreover, since $L$ is superlinear, any feasible $(\rho, \omega, \zeta)$ satisfy $\omega, \zeta \ll \rho$ and satisfies the weak formulation of Proposition 1.1.3. Let us denote $\bar{C}$ the infimum in (1.1.7) taken with $T = 1$ (the fact that this value does not change with $T$ is a consequence of Proposition 1.1.6). One may argue exactly as in [61, Thm. 5.4] to show the inequality $C_L \leq \bar{C}$. The reverse inequality follows from Hölder inequality and is exact if and only if $\psi_L(\rho_t, \omega_t, \zeta_t) = C_L(\mu,\nu)$ $\mathrm{d}t$-a.e. for any minimizer $(\rho_t, \omega_t, \zeta_t)$. This constant speed property, combined with the fact that $(\rho, \omega, \zeta)$, after time rescaling, remains minimizing between any pair of intermediate times $0 \leq s < t \leq 1$ (otherwise one could improve the action on $[0,1]$ by glueing), leads to (1.1.6).

If $p = 1$, let $(\rho, \omega, \zeta)$ be a minimizer of (1.1.13). Extend it in time by $(\mu, 0, 0)$ for $t < 0$ et $(\nu, 0, 0)$ for $t > 1$ as explained in Proposition 1.1.8 and smooth it by convolution with a mollifier $r_\varepsilon$ which only depend on time and supported on $]-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}[$. We call $(\rho^\varepsilon, \omega^\varepsilon, \zeta^\varepsilon)$ the result, which solves the continuity equation between $\mu$ and $\nu$ on $[-\frac{\varepsilon}{2}, 1+\frac{\varepsilon}{2}]$ (Proposition 1.1.9) and satisfies

$$\int_{-\frac{\varepsilon}{2}}^{1+\frac{\varepsilon}{2}} \int_\Omega \psi_L(\omega^\varepsilon, \zeta^\varepsilon | \rho^\varepsilon) \leq \int_0^1 \int_\Omega \psi_L(\omega, \zeta | \rho) = C_L(\mu,\nu)$$

according to Lemma 1.1.23. Posing $s : (t,x) \mapsto ((t+\frac{\varepsilon}{2})/(1+\varepsilon), x)$, one also has

$$\begin{aligned} C_L(\mu,\nu) &\leq \int_0^1 \int_\Omega \psi_L(s_\# \omega^\varepsilon, s_\# \zeta^\varepsilon | s_\# \rho^\varepsilon / (1+\varepsilon)) \\ &= \int_0^1 \int_\Omega \psi_L(s_\# \omega^\varepsilon, s_\# \zeta^\varepsilon | s_\# \rho^\varepsilon) \\ &= \int_{-\frac{\varepsilon}{2}}^{1+\frac{\varepsilon}{2}} \int_\Omega \psi_L(\omega^\varepsilon, \zeta^\varepsilon | \rho^\varepsilon) \leq C_L(\mu,\nu) \end{aligned}$$

So there exists minimizers which can be disintegrated with respect to $\mathrm{d}t$ and the rest of the proof goes through as for the case $p > 1$, except that the equality case in Hölder inequality is no longer relevant. □

We used in the previous proof that fact that the action with a 1-homogeneous Lagrangian is invariant by time reparametrization, and the associated action is independent of $\rho$. From these properties, it follows that time can completely by removed from the variational problem in that case. This gives a steady state problem that provides a generalization of Beckmann's problems [142, Chap. 4].

**Proposition 1.1.20** (Static formulation for 1-homogeneous Lagrangians)**.** *If $L$ is 1-homogeneous (and thus sublinear), then one has*

$$\begin{aligned} C_L(\mu,\nu) &= \min\left\{ \int_\Omega L(\omega,\zeta)\ ;\ \nu-\mu = \zeta - \nabla\cdot\omega,\ (\omega,\zeta)\in \mathscr{M}(\Omega;\mathbb{R}^d)\times\mathscr{M}(\Omega)\right\} \\ &= \sup\left\{ \int_\Omega \varphi\,\mathrm{d}(\nu-\mu)\ ;\ L^*(\nabla\varphi,\varphi)<\infty,\ \varphi\in\mathscr{C}^1(\Omega)\right\}. \end{aligned}$$

*where the equation $\nu-\mu=\zeta-\nabla\cdot\omega$ is understood in the distributional sense, with no-flux boundary conditions on $\partial\Omega$.*

*Proof.* In order to show that the infimum and the supremum problems are equal, we may apply Fenchel-Rockafellar duality. We do not give all the details (see the proof of Theorem 1.1.15) but we may emphasize that the bounded linear operator appearing in the duality is $A:\mathscr{C}^1(\Omega)\to\mathscr{C}(\Omega)^d\times\mathscr{C}(\Omega)$ and the qualification constraint is satisfied because $A(\varphi)$ is in the interior of $L^*$ for $\varphi=0$. This additionally shows that the minimum, let us write it $C_S(\mu,\nu)$, is attained. In order to show $C_S(\mu,\nu)=C_L(\mu,\nu)$, first take a feasible couple for the static problem $(\bar\omega,\bar\zeta)$ and define $(\rho_t,\omega_t,\zeta_t)=((1-t)\mu+t\nu,\bar\omega,\bar\zeta)$ for $t\in[0,1]$, which satisfies the continuity equation between $\mu$ and $\nu$. One has

$$C_L(\mu,\nu)\le\int_0^1\int_\Omega \psi_L(\omega_t,\zeta_t|\rho_t)=\int_0^1\int_\Omega L(\omega_t,\zeta_t)=\int_\Omega L(\bar\omega,\bar\zeta)$$

and it follows $C_L(\mu,\nu)\le C_S(\mu,\nu)$. Conversely, let $\bar\varphi\in\mathscr{C}^1(\Omega)$ be feasible in the static supremum problem and let $\varphi:(t,x)\mapsto\bar\varphi(x)$. Since $L$ is sublinear, $L^*$ is the indicator of a (convex) set so $\varphi$ satisfies the constraint from the dual formulation in Theorem 1.1.15 and

$$\int_\Omega\bar\varphi\,\mathrm{d}(\nu-\mu)=\int_\Omega\varphi_1\,\mathrm{d}\nu-\int_\Omega\varphi_0\,\mathrm{d}\mu\le C_L(\mu,\nu).$$

By taking the supremum, it follows $C_S(\mu,\nu)\le C_L(\mu,\nu)$. □

The situation where the Lagrangian is 1-homogeneous only w.r.t. the velocity has been studied in detail in the recent work [147] where static formulations are derived and connected to a general framework of "unbalanced $W_1$" models [148]. This provides a rich development of the theory considered in this part of the thesis and the associated models are computationally advantageous in some cases.

### Other technical results

We finish this subsection on action minimizing problems with technical results. We begin with deriving a bound of the total variation of triplets $(\rho,\omega,\zeta)$ with bounded action. Such bounds

can be easily derived for a quadratic Lagrangian (see our article [44]) but the framework here is more general so we rely on an application of Grönwall's lemma.

**Proposition 1.1.21** (Bounded mass)**.** *Let $S \subset \mathscr{M}_+(\Omega)$ be a bounded set, $M > 0$ be a constant and $\mathscr{A}$ the set of all $(\rho, \omega, \zeta)$ solutions to the continuity equation between $\mu \in S$ and $\nu \in \mathscr{M}_+(\Omega)$ on $[0,1]$, such that $\int_0^1 \int_\Omega \psi_L(\omega, \zeta | \rho) < M$. Then*

$$\nu(\Omega), \quad \rho([0,1]\times\Omega), \quad |\omega|([0,1]\times\Omega) \quad \textit{and} \quad |\zeta|([0,1]\times\Omega)$$

*are uniformly bounded in $\mathscr{A}$.*

*Proof.* Consider the "reduced" Lagrangians $L_v$ and $L_g$ on $\mathbb{R}$, defined as the convex regularizations of $a \mapsto \inf_{|v|=|a|, g\in\mathbb{R}} L(v,g)$ and $a \mapsto \inf_{|g|=|a|, v\in\mathbb{R}^d} L(v,g)$, respectively (they are equal to $L(v,0)$ and $L(0,g)$ in the examples of the next chapter). These functions are also nonnegative, continuous, convex, and admit a unique minimizer at 0. Let $(\rho,\omega,\zeta) \in \mathscr{A}$ and consider $f : t \mapsto \rho_t(\Omega)$. The distributional derivative of $f$ is bounded by $\sigma := \pi^t_\# |\zeta| \in \mathscr{M}_+([0,1])$, by Lemma 1.1.2 and one can write the Lebesgue decomposition $\sigma = g f \mathrm{d}t + \sigma^\perp$ with $\sigma^\perp \in \mathscr{M}([0,1])$ and $(gf) \in L^1([0,1])$. We want to show that there is a bound on the maximum value of $f$. To this end, let $b \in [0,1]$, if it exists, be such that $f(b) > f(0)$ and let $a := \sup_{t\in[0,b]} f(t) \leq f(0)$. One has, thanks to the sublinearity of $\psi_L$,

$$M > \int_0^1 \int_\Omega \psi_{L_g}(|\zeta| \,|\, \rho) \geq \int_0^1 \psi_{L_g}(\sigma | f \mathrm{d}t) \geq f(0) \int_a^b L_g(g)\mathrm{d}t + (L_g')_\infty \sigma^\perp([a,b]).$$

Thus, taking any affine lower bound $y = \alpha \cdot g + \beta$ of $L_g$ with $\alpha > 0$ and $\beta < 0$, it follows

$$\int_a^b g \mathrm{d}t \leq (M/f(0) - \beta(b-a))/\alpha \quad \text{and} \quad \sigma^\perp([a,b]) \leq M/(L_g)'_\infty.$$

We conclude by Grönwall's lemma that since for all $t \in [a,b]$

$$f(t) \leq f(0) + \sigma([a,t]) \leq (f(0) + \sigma^\perp([a,b])) + \int_a^t g(t) f(t) \mathrm{d}t,$$

then

$$f(b) \leq (f(0) + \sigma([a,b])) e^{\int_a^b g(t)\mathrm{d}t} \leq (\mu(\Omega) + \frac{M}{(L_g)'_\infty}) e^{(\frac{M}{\mu(\Omega)} - \beta)/\alpha}.$$

so the mass $\rho_t(\Omega)$ is uniformly bounded, and so is $\rho([0,1]\times\Omega)$. It follows easily

$$M > L_g(|\zeta|([0,1]\times\Omega)/\rho([0,1]\times\Omega)) \text{ and } M > L_v(|\omega|([0,1]\times\Omega)/\rho([0,1]\times\Omega))$$

from which we deduce the uniform bound on the total variations of $\zeta$ and $\omega$. □

Let us state clearly in which situations it can be guaranteed that the velocity and growth fields are well-defined.

**Lemma 1.1.22** (Finite superlinear cost)**.** *If $L$ is superlinear, then any feasible $(\rho,\omega,\zeta)$ for (1.1.13) satisfies $\omega = (v_t\rho_t) \otimes \mathrm{d}t$ and $\zeta = (g_t\rho_t) \otimes \mathrm{d}t$ with $v_t \in L^1_{\rho_t}(\Omega;\mathbb{R}^d)$ and $g_t \in L^1_{\rho_t}(\Omega;\mathbb{R})$ where $(\rho_t)$ is the time disintegration of $\rho$ w.r.t. Lebesgue.*

*Proof.* By the very definition of $\psi_L$, a superlinear Lagrangian and a finite action implies that the singular parts of $\omega$ and $\zeta$ w.r.t. $\rho$ vanish. We know also by Lemma 1.1.2 that $\rho$ can be disintegrated in time w.r.t. Lebesgue. □

Finally, the monotony of sublinear functionals with respect to smoothing is a convenient technical tool.

**Lemma 1.1.23.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a l.s.c. sublinear function. Then, for any measure $\mu \in \mathscr{M}(\mathbb{R}^d; E)$ with compact support and mollifier $r_\varepsilon : \mathbb{R}^d \to \mathbb{R}_+$, it holds*

$$\int_{\mathbb{R}^d} f(r_\varepsilon * \mu) \leq \int_{\mathbb{R}^d} f(\mu).$$

*Proof.* Let $Q \subset \mathbb{R}^d$ be the domain of $f^*$ which is convex and closed and let $K$ be the set of continuous functions on $\mathbb{R}^d$ taking their values in $Q$. For any $\phi \in K$, it holds

$$\int \phi \mathrm{d}(r_\varepsilon * \mu) = \int (r_\varepsilon * \phi) \mathrm{d}\mu \leq \int f(\mu)$$

because $\int f(\mu) = \sup_{\phi \in K} \int \phi \mathrm{d}\mu$ (see Theorem A.0.14), and the convexity of $K$ implies that $r_\varepsilon * \phi \in K$. Since this holds for any $\phi \in K$, the inequality follows. □

### 1.1.4. Geometric and topological properties

The following two results are inspired from [61] where another extension of optimal transport is considered. We recall that a sequence of (signed) measures $\mu_n$ on a metric space converges weakly to $\mu$ if for all real valued, continuous and bounded function $\varphi$ on $X$, it holds $\int_X \varphi \mathrm{d}\mu_n \to \int_X \varphi \mathrm{d}\mu$.

**Proposition 1.1.24** (Lower semicontinuity)**.** *Let $(\mu_n), (\nu_n)$, $n \in \mathbb{N}$ be sequences in $\mathscr{M}_+(\Omega)$ that converge weakly to $\mu, \nu$ and let $(L_n)$ be a sequence of Lagrangian functions that converge pointwise and increasingly to L, also an admissible Lagrangian. Then*

$$C_L(\mu, \nu) \leq \liminf_{n \to \infty} C_{L_n}(\mu_n, \nu_n).$$

*Proof.* Similarly as in [61], let us choose a sequence $(\rho_n, \omega_n, \zeta_n)$ of minimizers for $C_{L_n}(\mu_n, \nu_n)$. By Proposition 1.1.21, their mass is bounded so let us extract a subsequence that attains the lim inf (again indexed by $n$), weakly converges to $(\rho, \omega, \zeta)$, and belongs to $\mathrm{CE}_0^1(\mu, \nu)$. For any $m < n$, it holds

$$\int_0^1 \int_\Omega \psi_{L_m}(\omega_n, \zeta_n | \rho_n) \leq \int_0^1 \int_\Omega \psi_{L_n}(\omega_n, \zeta_n | \rho_n) = C_{L_n}(\mu_n, \nu_n).$$

Taking the limit $n \to \infty$, one has, by lower-semicontinuity that for all $m \in \mathbb{N}$,

$$\int_0^1 \int_\Omega \psi_{L_m}(\omega, \zeta | \rho) \leq \liminf_{n \to \infty} C_{L^n}(\mu_n, \nu_n).$$

The result follows by passing to the limit $m \to \infty$. □

We are now in position to prove that dynamic formulations with $p$-homogeneous Lagrangians give a geodesic space structure to $\mathscr{M}_+(\Omega)$. A metric space $(X,d)$ is said *geodesic* if any pair of points $(x_0,x_1) \in X^2$ can be connected by a path $t \in [0,1] \to x_t \in X$ such that $d(x_t,x_s) = |t-s|\,d(x_0,x_1)$ called a *minimal, constant speed geodesic*. Remember that it is one of the striking features of standard optimal transport to give a geodesic structure to the space of probability measures, other than the classical linear structure. We show here an analogue of this property for the space of nonnegative measures.

**Theorem 1.1.25** (Metric property)**.** *If $L$ is (positively) $p$-homogeneous and symmetric with respect to the origin then $C_L^{1/p}$ is a metric and $(\mathscr{M}_+(\Omega), C_L^{1/p})$ is a complete geodesic metric space.*

*Proof.* Let $\mu,\nu \in \mathscr{M}_+(\Omega)$. It is clear that $C_L^{1/p}(\mu,\nu)$ is finite by Proposition 1.1.14. The symmetry property comes from the symmetry of $L$ and the fact that $(\rho,\omega,\zeta) \in \mathrm{CE}_0^1(\mu,\nu) \Leftrightarrow (\rho,-\omega,-\zeta) \in \mathrm{CE}_0^1(\nu,\mu)$. It is clear that $C_L(\mu,\mu) = 0$ and conversely, if $C_L(\mu,\nu) = 0$, then $\omega = \zeta = 0$ which implies $\mu = \nu$. Finally, the triangle inequality follows from the characterization (1.1.7), which also proves that any pair of points can be joined by a constant speed minimizing geodesic. For the completeness property, consider a sequence $\mu_n \in \mathscr{M}_+(\Omega)$ that satisfies the Cauchy property. By Lemma 1.1.21, $\mu_n$ is bounded, so it admits a subsequence $\mu_{n_k}$ that weakly converges to $\mu \in \mathscr{M}_+(\Omega)$. By the lower-semicontinuity property of Proposition 1.1.24, one has for all $m \in \mathbb{N}$ that $C_L^{1/p}(\mu_m,\mu) \leq \liminf C_L^{1/p}(\mu_m,\mu_{n_k})$ and it follows, using the Cauchy property, $\limsup_{m\to\infty} C_L^{1/p}(\mu_m,\mu) \leq \limsup_{n,m\to\infty} C_L^{1/p}(\mu_m,\mu_n) = 0$ so $\mu_n$ converges to $\mu$ for the metric $C_L^{1/p}$. □

Finally, we study the relationship that $C_L$ shares with the weak topology.

**Theorem 1.1.26** (Weak convergence)**.** *If $(\mu_n)$ and $(\nu_n)$, $n \in \mathbb{N}$ converge weakly to $\mu$ and $\nu$ respectively, then $C_L(\mu,\nu) \leq \liminf C_L(\mu_n,\nu_n)$. Moreover, if $L$ has a polynomial growth in the second argument $g$, then*

$$\mu_n \text{ converges weakly to } \mu \quad \Leftrightarrow \quad C_L(\mu_n,\mu) \to 0.$$

*Proof of the l.s.c. property.* It can be seen from the duality formula (1.1.5): $C_L$ is the supremum of a nonempty family of continuous linear functionals. It is thus weakly l.s.c. □

*Proof of $\Leftarrow$.* If $C_L(\mu_n,\mu) \to 0$, by Proposition 1.1.21, $\mu_n(\Omega)$ is a bounded sequence. Let $\bar{\mu}$ be a weak cluster point. Taking the suitable subsequence, one has $C_L(\bar{\mu},\mu) \leq \lim C_L(\mu_n,\mu) = 0$ so $\bar{\mu} = \mu$. Since this holds for any cluster point, $\mu_n \rightharpoonup \mu$. □

*Proof of $\Rightarrow$.* Let $\varepsilon > 0$ and let $(W_k)_{k\in I}$ be a finite Borel partition of $\Omega$ of diameter smaller than $\varepsilon$ and such that $\mu(\partial W_k) = 0$ for all $k \in I$. The fact that such a partition exists is called the mosaic lemma ( it relies on the fact that given a fixed center, the set of balls whose boundaries have a nonzero mass is at most countable). By the Portmanteau lemma, $\mu_n(W_k) \to \mu(W_k)$ for all $k \in I$. Let us now build a triplet $(\rho,\omega,\zeta)$ which has a vanishing action, by defining it on each $W_k$. If $\mu(W_k) = 0$, let $\omega|_{W_k} = 0$ and $\zeta_t|_{W_k} = (1-t)^\alpha \mu|_{W_t}$ where $\alpha$ is the exponent that appears in the proof of Proposition 1.1.14. The associated cost is proportional to $\mu_n(W_k)$. Otherwise, if

$\mu(W_k) > 0$, assume that $\mu_n(W_k) > 0$ too since this is eventually the case and define the same interpolation as the second part of Proposition 1.1.14 : adjustment of mass by uniform growth on $[0, \frac{1}{2}]$ and interpolation with the product coupling on $[\frac{1}{2}, 1]$. The associated cost is proportional to $\mu(W_k) \cdot L(2\varepsilon, 0)$ for the transport part and $L(0, 2\log[\mu(\Omega)/\mu_n(\Omega)])$ for the growth part. Summing all these local contributions, one obtains a bound on the action

$$\mathbb{A}(\rho, \omega, \zeta) \leq \text{cst} \times \left( \sum_{k \in I_0} \mu_n(W_k) + \sum_{k \in I \setminus I_0} \mu(W_k). \left( L(2\varepsilon, 0)) + L(0, 2\log \frac{\mu(\Omega)}{\mu_n(\Omega)}) \right) \right)$$

where $I_0 := \{k \in I \,;\, \mu(W_k) = 0\}$. This upper bounds tends to $\text{cst} \times L(2\varepsilon, 0).\mu(\Omega)$ as $n \to \infty$. Since $\varepsilon > 0$ is arbitrary, it follows $C_L(\mu_n, \mu) \to 0$. Note that a similar proof can be used to show the weak continuity of Wasserstein metrics in the compact setting. □

## 1.2. Coupling formulations

In this section, we describe the second approach to unbalanced optimal transport, inspired by the *coupling* formulation of optimal transport introduced by Kantorovich. In this formulation, the notion of time dynamic disappears: one directly looks for a pairwise matching of infinitesimal particles of mass. Among all the admissible pairings, we select the one that minimizes an integral cost. In the case of unbalanced optimal transport, there are several ways to formalize this idea and we review three variants of coupling formulations: the semi-coupling formulation, that we introduced in [43] and that will be our guideline in this section, and the optimal lifting and optimal entropy-transport problems, both introduced in [103].

### 1.2.1. Definition and first properties

Let $X$ and $Y$ be compact metric spaces[2] and consider $\mu \in \mathscr{M}_+(X)$, $\nu \in \mathscr{M}_+(Y)$. Whenever $\mu(X) \neq \nu(Y)$, the set of couplings $\{\gamma \in \mathscr{M}_+(X \times Y)\ ;\ (\pi^x, \pi^y)_\# \gamma = (\mu, \nu)\}$ is empty, because pushforward operators preserve the total mass of nonnegative measures. To overcome this difficulty, one possibility is to describe an unbalanced optimal transport with *two* couplings: this is the semi-coupling formulation that we introduced in [43]. The first coupling describes where the mass goes as it leaves from $\mu$ and the other describes from where it comes as it arrives at $\nu$. In standard optimal transport one imposes that these two couplings match exactly. Here, instead, we take into account the discrepancy between the mass that leaves $\mu$ and the mass that arrives at $\nu$ in the cost function. One thus chooses a function $h_{x_1,x_2}(u_1,u_2)$ which gives the cost of associating a mass $u_1$ located at $x_1$ to a mass $u_2$ at $x_2$. It is natural to require this cost to be sublinear in $(u_1,u_2)$: 1-homogeneous because the amount of effort should be proportional to the amount of mass transported (as in optimal transport) and subbaditive so that it is never more efficient to divide mass into smaller chunks for transporting it between the same endpoints. This property will also appear clearly from the link to the lifted formulation in Section 1.2.3.

**Assumption 1.2.1** (Sublinear cost)**.** *The cost function $(h_{x_1,x_2})_{(x_1,x_2)\in X\times Y}$ is a family of proper sublinear functions $h_{x_1,x_2} : \mathbb{R}^2_+ \to [0,+\infty]$ which is l.s.c. jointly in all variables $(x_1,x_2,u_1,u_2)$.*

For instance, standard optimal transport problems with a l.s.c. cost $c : X \times Y \to [0,\infty]$ correspond to the sublinear costs whose domain is the positive diagonal:

$$h_{x_1,x_2}(u_1,u_2) = \begin{cases} u_1 c(x_1,x_2) & \text{if } u_1 = u_2 \geq 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{1.2.1}$$

Note that each function $h_{x_1,x2}$ is, after the change of variables $(t,s) = (u_1+u_2, u_1-u_2)$, the perspective of the function $\theta \mapsto h_{x_1,x_2}(1-\theta,\theta)$ which thus entirely characterizes $h$ (actually such a remark could be made for any l.s.c. sublinear function of two real variables that is infinite on a linear half-space).

[2] the differentiable structure of $\mathbb{R}^d$ (or manifolds) is no longer needed for static formulations. The compactness assumption is here for technical convenience.

**Definition 1.2.2** (Semi-coupling formulation)**.** *For $(\mu,\nu)\in\mathscr{M}_+(X)\times\mathscr{M}_+(Y)$, we define*

$$\boxed{C_h(\mu,\nu):=\inf_{\gamma_1,\gamma_2\in\mathscr{M}_+(X\times Y)}\left\{\int_{X\times Y}h_{x_1,x_2}(\gamma_1,\gamma_2)\;;\;(\pi^1_\#\gamma_1,\pi^2_\#\gamma_2)=(\mu,\nu)\right\}.}\tag{1.2.2}$$

See the section *Notation* for the definition of homogeneous functions of measures (the same convention is use throughout the thesis). Decomposing w.r.t. a dominating measure, this problem can be rewritten in the less compact way

$$\inf_{f_1,f_2,\gamma}\left\{\int_{X\times Y}h_{x_1,x_2}(f_1(x_1,x_2),f_2(x_1,x_2))\mathrm{d}\gamma(x_1,x_2)\;;\;(\pi^1_\#(f_1\gamma),\pi^2_\#(f_2\gamma))=(\mu,\nu)\right\}$$

where the infimum is over probabilities $\gamma\in\mathscr{P}(X\times Y)$ and densities $f_1,f_2\in L^1_\gamma(X\times Y)$. I might help for the intuition to see how standard optimal transport is recovered by plugging the sublinear cost (1.2.1) in (1.2.2).

**Proposition 1.2.3** (Minimizers)**.** *For any $\mu\in\mathscr{M}_+(X)$ and $\nu\in\mathscr{M}_+(Y)$,* (1.2.2) *admits a minimizer.*

*Proof.* By Theorem A.0.14, the objective functional is weakly l.s.c. on $\mathscr{M}_+(X\times Y)^2$. Since $X$ and $Y$ are compact and the marginals $\mu,\nu$ have finite mass, the (closed) constraint set is weakly sequentially compact. It follows that any minimizing sequence admits a cluster point which is a minimizer (the minimum is not necessarily finite). □

The following result mirrors the celebrated Kantorovich duality theorem, and extends it to the unbalanced case. In Kantorovich duality, the constraint set on the dual variables is a half-space, which is here replaced by a more general convex set.

**Theorem 1.2.4** (Duality)**.** *Let $(h_{x_1,x_2})$ be a sublinear cost and let $Q_h(x_1,x_2)$ be the family of closed convex sets such that $h^*_{x_1,x_2}=\iota_{Q_h(x_1,x_2)}$ for all $(x_1,x_2)\in X\times Y$. Then it holds*

$$C_h(\mu,\nu)=\sup_{\substack{\phi\in\mathscr{C}(X)\\ \psi\in\mathscr{C}(Y)}}\left\{\int_X\phi\,\mathrm{d}\mu+\int_Y\psi\,\mathrm{d}\nu\;;\;(\phi(x),\psi(y))\in Q_h(x,y),\,\forall(x,y)\in X\times Y\right\}.$$

*Proof.* We rewrite the supremum problem as $\sup -F(\xi_1,\xi_2)-G(\xi_1,\xi_2)$ for $(\xi_1,\xi_2)\in\mathscr{C}(X\times Y)^2$ where

$$G:(\xi_1,\xi_2)\mapsto\begin{cases}-\int_X\phi\,\mathrm{d}\mu-\int_Y\psi\,\mathrm{d}\nu & \text{if } \xi_1(x,y)=\phi(x) \text{ and } \xi_2(x,y)=\psi(y)\\ +\infty & \text{otherwise,}\end{cases}$$

and $F$ is the indicator of $\{(\xi_1,\xi_2)\in\mathscr{C}(X\times Y)^2:(\xi_1,\xi_2)(x,y)\in Q_h(x,y),\forall(x,y)\in X\times Y\}$. Note that $F$ and $G$ are convex and proper. Also, given our assumptions, there is a pair of functions $(\xi_1,\xi_2)$ at which $F$ is continuous (for the sup norm topology) and $F$ and $G$ are finite since for all $(x,y)\in X\times Y$, $Q_h(x,y)$ contains the negative orthant $\mathbb{R}_-\times\mathbb{R}_-$. Then Fenchel-Rockafellar duality (Appendix A) states that

$$\sup_{(\xi_1,\xi_2)\in\mathscr{C}(X\times Y)^2}-F(\xi_1,\xi_2)-G(\xi_1,\xi_2)=\min_{(\gamma_1,\gamma_2)\in\mathscr{M}(X\times Y)^2}F^*(\gamma_1,\gamma_2)+G^*(-\gamma_1,-\gamma_2).\tag{1.2.3}$$

Let us compute the conjugate functions. For $G$, we obtain

$$G^*(-\gamma_1,-\gamma_2) = \sup_{\substack{\phi\in\mathscr{C}(X)\\ \psi\in\mathscr{C}(Y)}} -\int_{X\times Y}\phi(x)\mathrm{d}\gamma_1 - \int_{X\times Y}\psi(x)\mathrm{d}\gamma_2 + \int_X \phi(x)\mathrm{d}\mu + \int_Y \psi(y)\mathrm{d}\nu$$
$$= \begin{cases} 0 & \text{if } (\pi^1_\#\gamma_1, \pi^2_\#\gamma_2) = (\mu,\nu) \\ +\infty & \text{otherwise.} \end{cases}$$

On the other hand, $F^*$ is given by Theorem A.0.14 (Appendix A) which states that

$$F^*(\gamma_1,\gamma_2) = \int_{X\times Y} h_{x_1,x_2}(\gamma_1,\gamma_2).$$

Finally, as $F^*$ includes the nonnegativity constraint, the right hand side of (1.2.3) is equal to $C_h(\mu,\nu)$. □

**Proposition 1.2.5** (Lower semicontinuity)**.** *Assume that $(h^{(n)})$ is a sequence of admissible sublinear costs that converge pointwise and increasingly to an admissible sublinear cost $h$ and let $\mu_n \rightharpoonup \mu$ in $\mathscr{M}_+(X)$ and $\nu_n \rightharpoonup \nu$ in $\mathscr{M}_+(Y)$. Then*

$$C_h(\mu,\nu) \le \liminf_{n\to\infty} C_{h^{(n)}}(\mu_n,\nu_n).$$

*Proof.* We follow a similar line of reasoning than in [61, Thm. 5.6]. Let us choose a sequence of minimizers $(\gamma_1^{(n)},\gamma_2^{(n)})$. Each member of these pairs satisfies one coupling constraint, so their mass is bounded and any subsequence (in particular those which attains the lim inf) admit a weak limit $(\gamma_1,\gamma_2)$, which itself satisfies the semi-coupling constraints $(\pi^1_\#\gamma_1,\pi^2_\#\gamma_2) = (\mu,\nu)$. For any $m<n$, it holds

$$\int_{X\times Y} h^{(m)}_{x_1,x_2}(\gamma_1^{(n)},\gamma_2^{(n)}) \le \int_{X\times Y} h^{(n)}_{x_1,x_2}(\gamma_1^{(n)},\gamma_2^{(n)}) = C_{h^{(n)}}(\mu_n,\nu_n).$$

Taking the limit $n\to\infty$, one has, by the lower semicontinuity of the functional, for all $m\in\mathbb{N}$,

$$\int_{X\times Y} h^{(m)}_{x_1,x_2}(\gamma_1,\gamma_2) \le \liminf_{n\to\infty} C_{h^{(n)}}(\mu_n,\nu_n).$$

The result follows by passing to the limit $m\to\infty$. □

### 1.2.2. Geometric and topological properties

As stated in the introduction of this thesis, an essential property of optimal transport is that it can be used to lift a metric from the base space $X$ to the space of probability measures over $X$ [162, Chapter 6]). This property has an analogue in the unbalanced framework, if one takes care of the fact that two particles of zero mass are considered equal, whatever their location in $X$. The *cone* of a space, a standard construction in topology, accounts for this trait.

**Definition 1.2.6** (Cone)**.** *The space $\mathfrak{C}(X)$, called the cone of $X$ is defined as the space $X\times\mathbb{R}_+$ where all the points with zero mass $X\times\{0\}$ collapse to one point, called the* apex*, endowed with the quotient topology.*

In the non-compact setting, it is noted in [103] that the quotient topology is not the relevant choice anymore, but a weakened version of it.

**Theorem 1.2.7** (Metric)**.** *Assume that $X = Y$ and let $h$ be a cost function such that, for some $p \geq 1$ it holds*

$$(x_1, u_1), (x_2, u_2) \mapsto h_{x_1,x_2}(u_1, u_2)^{1/p} \tag{1.2.4}$$

*is a metric for $\mathfrak{C}(X)$. Then $C_h^{1/p}$ defines a metric on $\mathscr{M}_+(X)$.*

**Remark 1.2.8.** *Compatibility with the cone topology implies in particular: (i) for all $x_2 \in \Omega$, $u \geq 0$, $h_{x_1,x_2}(0,u)$ is independent of $x_1 \in \Omega$ and (ii) $h_{x_1,x_2}(0,0) = 0$.*

**Remark 1.2.9.** *We can replace the word "metric" by "extended metric" (i.e. allowing the value $+\infty$) and the proof goes through. By considering the cost in* (1.2.1)*, a corollary is that $C_h^{1/p}$ defines a proper metric on each equivalence class for the relation $\mu \sim \nu \Leftrightarrow \mu(X) = \nu(X)$. The metric property of the Wasserstein distance is thus recovered as a particular case.*

*Proof.* Symmetry and non-negativity are inherited from $h$. Also, if $C_h(\gamma_1, \gamma_2) = 0$, there exists a coupling $\gamma \in \mathscr{M}_+(X^2)$ such that $(\pi^1_\# \gamma, \pi^2_\# \gamma) = (\mu, \nu)$ and $x = y$ $\gamma$-a.e. so $\mu = \nu$. It remains to show the triangle inequality. Fix $\mu_a$, $\mu_b$, $\mu_c \in \mathscr{M}_+(X)$. Take two pairs of minimizers $(\gamma_1^{ab}, \gamma_2^{ab})$ and $(\gamma_1^{bc}, \gamma_2^{bc})$ for (1.2.2). They satisfy the constraints

$$(\pi^1_\# \gamma_1^{ab}, \pi^2_\# \gamma_2^{ab}) = (\mu_a, \mu_b), \qquad (\pi^1_\# \gamma_1^{bc}, \pi^2_\# \gamma_2^{bc}) = (\mu_b, \mu_c).$$

Let $\lambda \in \mathscr{M}_+(X)$ be a reference measure dominating all the marginals on the second factor i.e. such that $\pi^2_\# \gamma_1^{ab}$, $\pi^2_\# \gamma_2^{ab}$, $\pi^1_\# \gamma_1^{bc}$, $\pi^1_\# \gamma_2^{bc} \ll \lambda$ and denote by $\gamma_i^{ab}(\mathrm{d}x|y)$ the disintegration of $\gamma_i^{ab}$ along the second factor w.r.t. $\lambda$. This means that for all $y \in X$, $\gamma_i^{ab}(\mathrm{d}x|y) \in \mathscr{M}_+(X)$ and it holds, for any measurable $f : X^2 \to \mathbb{R}$,

$$\int_{X^2} f \,\mathrm{d}\gamma_i^{ab} = \int_X \left( \int_X f(x,y)\, \gamma_i^{ab}(\mathrm{d}x|y) \right) \lambda(\mathrm{d}y).$$

We define analogously $\gamma_i^{bc}(\mathrm{d}z|y)$ the disintegration w.r.t. $\lambda$ along the first factor and write $\sigma_b = \frac{\mathrm{d}\mu_b}{\mathrm{d}\lambda} \in L^1(\lambda)$ the density of $\mu_b$ w.r.t. $\lambda$. Let us combine the optimal semi-couplings in a suitable way to define $\gamma_1$, $\gamma_2$, $\hat{\gamma} \in \mathscr{M}(X^3)$ via their disintegration w.r.t. $\lambda$ along the second factor:

$$\gamma_1(\mathrm{d}x, \mathrm{d}z|y) := \begin{cases} \gamma_1^{ab}(\mathrm{d}x|y) \otimes [\gamma_1^{bc}(\mathrm{d}z|y) / \sigma_b(y)] & \text{if } \sigma_b(y) > 0, \\ \gamma_1^{ab}(\mathrm{d}x|y) \otimes \delta_y(\mathrm{d}z) & \text{otherwise,} \end{cases}$$

$$\gamma_2(\mathrm{d}x, \mathrm{d}z|y) := \begin{cases} [\gamma_2^{ab}(\mathrm{d}x|y) / \sigma_b(y)] \otimes \gamma_2^{bc}(\mathrm{d}z|y) & \text{if } \sigma_b(y) > 0, \\ \delta_y(\mathrm{d}x) \otimes \gamma_2^{bc}(\mathrm{d}z|y) & \text{otherwise,} \end{cases}$$

$$\hat{\gamma}(\mathrm{d}x, \mathrm{d}z|y) := \begin{cases} \gamma_2^{ab}(\mathrm{d}x|y) \otimes \gamma_1^{bc}(\mathrm{d}z|y) / \sigma_b(y) & \text{if } \sigma_b(y) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The interpretation of $\gamma_1$ is that all mass that leaves $x$ towards $y$, according to $\gamma_1^{ab}(\mathrm{d}x, \mathrm{d}y)$, is distributed over the third factor according to $\gamma_1^{bc}(\mathrm{d}y, \mathrm{d}z)$. In case the mass disappears at $y$, it is

simply dropped at $y$ on the third factor. Then $\gamma_2$ is built analogously for the incoming masses and $\hat{\gamma}$ is a combination of incoming and outgoing masses. For $i = 1,2$ let $\gamma_i^{ac} := \pi^{1,3}_{\#}\gamma_i$ and note that, by construction, $(\pi^1_{\#}\gamma_1^{ac}, \pi^2_{\#}\gamma_2^{ac}) = (\mu_a, \mu_c)$. One has

$$\begin{aligned}\int_{X^3} h_{x,y}(\gamma_1,\hat{\gamma}) &= \int_{\sigma_b>0}\left(\int_{X^2} h_{x,y}(\gamma_1^{ab}(\mathrm{d}x|y),\gamma_2^{ab}(\mathrm{d}x|y))\otimes[\gamma_1^{bc}(\mathrm{d}z|y)/\sigma_b(y)]\right)\lambda(\mathrm{d}y)\\ &\quad+\int_{\sigma_b=0}\left(\int_{X^2} h_{x,y}(\gamma_1^{ab}(\mathrm{d}x|y),0)\otimes\delta_y(\mathrm{d}z)\right)\lambda(\mathrm{d}y)\\ &= \int_{X^2} h_{x,y}(\gamma_1^{ab},\gamma_2^{ab}) = C_h(\mu_a,\mu_b).\end{aligned}$$

Analogously, one finds $\int_{X^3} h_{yz}(\hat{\gamma},\gamma_2) = C_h(\mu_b,\mu_c)$. The main calculation can now be carried out:

$$\begin{aligned}C_h(\mu_a,\mu_c)^{\frac{1}{p}} &\le \left(\int_{X^2} h_{x,z}(\gamma_1^{ac},\gamma_2^{ac})\right)^{\frac{1}{p}} \overset{(1)}{\le} \left(\int_{X^3} h_{x,z}(\gamma_1,\gamma_2)\right)^{\frac{1}{p}}\\ &\overset{(2)}{\le} \left(\int_{X^3}\left[h_{x,y}(\gamma_1,\hat{\gamma})^{\frac{1}{p}} + h_{y,z}(\hat{\gamma},\gamma_2)^{\frac{1}{p}}\right]^p\right)^{\frac{1}{p}}\\ &\overset{(3)}{\le} \left(\int_{X^3} h_{x,y}(\gamma_1,\hat{\gamma})\right)^{\frac{1}{p}} + \left(\int_{X^3} h_{y,z}(\hat{\gamma},\gamma_2)\right)^{\frac{1}{p}}\\ &\overset{(4)}{\le} C_h(\mu_a,\mu_b)^{\frac{1}{p}} + C_h(\mu_b,\mu_c)^{\frac{1}{p}}\end{aligned}$$

where we successively invoked (1) the subadditivity of $h$, (2) the fact that $h^{1/p}$ satisfies the triangle inequality, (3) Minkowski's inequality and (4) the computations above. Thus $C_h(\cdot,\cdot)^{1/p}$ satisfies the triangle inequality, and is a metric. □

**Theorem 1.2.10** (Weak convergence)**.** *Assume that $X = Y$, that $(x,y)\mapsto h_{x,y}(1,1)$ tends to 0 as* $\mathrm{dist}(x,y)\to 0$ *and that there exists $\alpha > 0$ such that $h_{x,x}(1,0) + h_{x,x}(0,1) < \alpha$ for all $x \in X$, then*

$$(\mu_n \rightharpoonup \mu) \Rightarrow (C_h(\mu_n,\mu)\to 0).$$

*If moreover $h$ satisfies the hypothesis of Theorem 1.2.7, then the converse implication is true and $C_h^{1/p}$ metrizes the weak convergence on $\mathscr{M}_+(X)$.*

*Proof of* $\Rightarrow$. By the continuity near the diagonal (and thus uniform continuity by compactness of $X$) of $h_{\cdot,\cdot}(1,1)$, for all $\varepsilon > 0$, there exists $\eta > 0$ such that $\mathrm{dist}(x,y) < \eta$ implies $h_{x,y}(1,1) < \varepsilon$. Similarly to the proof of the weak continuity of $C_L$ (see Theorem 1.1.26), we choose a partition $(W_i)_{i\in I}$ of $X$ of diameter smaller than $\eta$ such that $\mu(\partial W_i) = 0$ for all $i \in I$ (the mosaic lemma is valid in separable metric spaces). Since $\mu_n$ weakly converges to $\mu$, one has $\mu_n(W_i)\to\mu(W_i)$ for all $i \in I$, by the Portmanteau lemma. Let us build a pair of semi-couplings $(\gamma_1^{(n)},\gamma_2^{(n)})$ for all $n \in \mathbb{N}$ as follows. First, $\gamma_i^{(n)}|_{W_i\times W_j} = 0$ if $i \neq j$ or if $\mu_n(W_i) = \mu(W_j) = 0$. Otherwise, if $\mu_n(W_i) > \mu(W_i)$ define

$$\begin{cases}\gamma_1^{(n)}|_{W_i\times W_i} = (\mu_n|_{W_i}\otimes\mu|_{W_i})/\mu_n(W_i) + (1-\mu(W_i)/\mu_n(W_i))\,\mathrm{diag}_{\#}\mu_n\\ \gamma_2^{(n)}|_{W_i\times W_i} = (\mu_n|_{W_i}\otimes\mu|_{W_i})/\mu_n(W_i).\end{cases}$$

where $\mathrm{diag}(x) = (x,x)$ maps $X$ to the diagonal of $X^2$. If $\mu(W_i) \geq \mu_n(W_i)$ use the same definition but with the role of $\mu$ and $\mu_n$ exchanged. By construction, $\gamma_1^{(n)}$ and $\gamma_2^{(n)}$ satisfy the semi-coupling constraint and one can now control $C_h(\mu_n,\mu)$ as follows

$$C_h(\mu_n,\mu) \leq \int_{X^2} h_{x,y}(\gamma_1^{(n)},\gamma_2^{(n)}) \leq \sum_{i\in I} \varepsilon \max(\mu_n(W_i),\mu(W_i)) + \alpha|\mu_n(W_i)-\mu(W_i)|$$

and the last term tends to $\varepsilon\mu(\Omega)$ as $n\to\infty$. Since $\varepsilon>0$ is arbitrary, it follows that $C_h(\mu_n,\mu)\to 0$. □

*Proof of* $\Leftarrow$. Under the additional hypothesis of Theorem 1.2.7, we have that $\mu_n \rightharpoonup \mu$ and $\nu_n \rightharpoonup \nu$ implies $C_h(\mu_n,\nu_n)\to C(\mu,\nu)$ as a consequence of the triangular inequality. Let us choose a sequence $(\mu_n)$ that does *not* converge weakly to $\mu\in\mathscr{M}_+(\Omega)$. Either $(\mu_n(\Omega))$ is not bounded and then, clearly, $C_h(\mu_n,\mu)$ does not converge to 0. Let us make it rigorous: in this case there must exist a subsequence of minimizers such that $\gamma_1^{(n)}(S^{(n)})\to\infty$, with $S^{(n)} := \{(x,y)\in X^2\,;\, \mathrm{d}\gamma_1^{(n)}/\mathrm{d}\gamma_2^{(n)} > 2\}$. So

$$C_h(\mu_n,\mu) \geq \int_{S^{(n)}} h_{x,y}(\gamma_1^{(n)},\gamma_2^{(n)}) \geq \gamma_1^{(n)}(S^{(n)}) \inf_{(x_1,x_2)\in X^2,\alpha\in[0,\frac{1}{2}]} h_{x_1,x_2}(1,\alpha)\to\infty$$

since by the lower-semicontinuity of $h$ the infimum is strictly positive. Otherwise, if $(\mu_n(\Omega))$ is bounded, then $\mu_n$ admits a subsequence which weakly converges to $\bar\mu\neq\mu$. In this case, by lower-semicontinuity (Proposition 1.2.5), all limit points of $C_h(\mu_n,\mu)$ are greater than $C_h(\bar\mu,\mu)\neq 0$. □

We conclude this section with a useful approximation lemma that we use several times in the following of this chapter.

**Lemma 1.2.11** (Atomic approximation of semi-couplings)**.** *Let $\mu\in\mathscr{M}_+(X)$ and $\nu\in\mathscr{M}_+(Y)$. If $h$ is a continuous sublinear cost, then there exists a sequence of finite atomic measures $(\gamma_1^{(n)})$ and $(\gamma_2^{(n)})$ that weakly converge to an optimal pair of semi-couplings for $C_h(\mu,\nu)$ and such that $\lim_{n\to\infty}\int_{X\times Y} h_{x,y}(\gamma_1^{(n)},\gamma_2^{(n)}) = C_h(\mu,\nu)$.*

*Proof.* Let $(f\gamma,g\gamma)$ be an optimal pair of semi-couplings for $C_h(\mu,\nu)$, where $\gamma\in\mathscr{P}(X\times Y)$ and $f,g\in L^1(\gamma)$. Let $(B_i^{(n)},(x_i,y_i)^{(n)})_{i\in I}$ be sequence of finite pointed partitions of $X\times Y$ such that $\lim_{n\to\infty}\max_{i\in I}\mathrm{diam}\, B_i^{(n)} = 0$. We define the discrete approximations $\bar\gamma_1^{(n)} := T_\#^{(n)}(f\gamma)$ and $\bar\gamma_2^{(n)} := T_\#^{(n)}(g\gamma)$ where $T^{(n)}:\Omega^2\to\Omega^2$ maps all points in $B_i^{(n)}$ to $(x_i,y_i)^{(n)}$. Also, denote $(\pi^1_\#\bar\gamma_1^{(n)},\pi^2_\#\bar\gamma_2^{(n)}) = (\mu_n,\nu_n)$. It is clear that the discretized semi-couplings weakly converge to $(f\gamma,g\gamma)$. Moreover, for $\varepsilon>0$, there exists $n\in\mathbb{N}$ such that for all $i\in I$, by Jensen inequality, and since $h$ is continuous, uniformly on $(X\times[0,1])\times(Y\times[0,1])$,

$$h_{(x_i,y_i)^{(n)}}(\bar\gamma_1^{(n)}(B_i^{(n)}),\bar\gamma_2^{(n)}(B_i^{(n)})) \leq \varepsilon\max\{\bar\gamma_1^{(n)}(B_i^{(n)}),\bar\gamma_2^{(n)}(B_i^{(n)})\} + \int_{B_i^{(n)}} h_{x,y}(f\gamma,g\gamma)$$

By integrating on the whole domain, one has

$$\int_{X\times Y} h_{(x,y)}(\hat\gamma_1^{(n)},\hat\gamma_2^{(n)}) \leq \varepsilon\cdot(\mu(X)+\nu(Y)) + C_h(\mu,\nu)$$

and the result follows because $\varepsilon$ can be arbitrarily small. □

### 1.2.3. Optimal lift and optimal entropy-transport problems

We now make apparent the link between the semi-coupling formulation and the other formulations introduced in [103].

#### Optimal lift formulation

The optimal lift formulation[3], introduced in [103] recasts the problem of unbalanced optimal transport as a standard optimal transport problem between lifted marginals. This formulation comes from the remark that the linear map $\mathfrak{h} : \mathscr{M}_+(X \times \mathbb{R}_+) \to \mathscr{M}(X)$ defined by

$$\mathfrak{h}(\bar{\mu})(B) = \int_B \int_{\mathbb{R}} u \, \bar{\mu}(\mathrm{d}x, \mathrm{d}u) \quad \forall \text{ Borel } B \subset X \tag{1.2.5}$$

which takes the "partial expectation" w.r.t. the real variable (interpreted as "mass") is a surjection. In particular, every measure $\mu \in \mathscr{M}_+(\Omega)$ admits pre-images that we refer to as *lifts*. This remark leads to the following definition.

**Definition 1.2.12** (Optimal lift)**.** *Let $c : (X \times \mathbb{R}_+) \times (Y \times \mathbb{R}_+) \to [0,\infty]$ be a l.s.c. cost function and $(\mu,\nu) \in \mathscr{M}_+(X) \times \mathscr{M}_+(Y)$. The optimal lift formulation of unbalanced optimal transport is defined as*

$$C_c^{\mathfrak{h}}(\mu,\nu) := \inf_{\gamma} \left\{ \int c \, \mathrm{d}\gamma \,;\, (\mathfrak{h}(\pi^1_\#\gamma), \mathfrak{h}(\pi^2_\#\gamma)) = (\mu,\nu) \right\}. \tag{1.2.6}$$

*where $\gamma \in \mathscr{M}_+((X \times \mathbb{R}_+) \times (Y \times \mathbb{R}_+))$.*

Equivalently, one may use the classical optimal transport cost $C_c$ associated to $c$ and write the problem as

$$C_c^{\mathfrak{h}}(\mu,\nu) := \inf_{\substack{\bar{\mu} \in \mathscr{M}_+(X \times \mathbb{R}_+) \\ \bar{\nu} \in \mathscr{M}_+(Y \times \mathbb{R}_+)}} \{ C_c(\bar{\mu},\bar{\nu}) \,;\, (\mathfrak{h}(\bar{\mu}), \mathfrak{h}(\bar{\nu})) = (\mu,\nu) \}.$$

This definition differs from the definition in [103] in some ways: we stick to the 1-homogeneous definition of $\mathfrak{h}$, do not necessarily consider homogeneous costs, and allow general measures (not just probabilities) as couplings. This problem does not necessarily admit a minimizer under these assumptions, but it turns out that for a general (but continuous) cost function $c$, this problem boils down to a semi-coupling problem with a regularized cost. This is a new result that sheds light on the interplay between the various formulations: informally, the optimal lift formulation uses the additional degree of freedom to regularize the cost. The intuitive explanation is as follows. Consider the transfer of mass between two elements of volume $\mathrm{d}x$ and $\mathrm{d}y$. In the optimal lift formulation, the mass of $\mu(\mathrm{d}x)$ and $\nu(\mathrm{d}y)$ can be decomposed in smaller chunks. In particular, the minimization problem seeks to lower the cost using these decompositions. Taking the infimum over all decompositions corresponds to the sublinear (i.e. convex and homogeneous) regularization of the cost at the point $(x,y)$, by Lemma 1.2.14. Note that if $c$ is already assumed homogeneous, the optimal lift formulation can be equivalently restricted to probability

[3] this problem is called "homogeneous formulation" in [103] but here this name could lead to confusion as many objects in this thesis are "homogeneous".

measures. Also, there is less freedom in the semi-coupling formulation because it does not allow such decompositions. This explains why we have to make the assumption that $h$ is sublinear beforehand.

For completeness, we also prove a similar result when $c$ is already sublinear in $(u_1,u_2)$, but this result is almost included in [103] (it is just not expressed in terms of semi-couplings).

**Theorem 1.2.13** (Reduction to semi-coupling)**.** *Let $c : (X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+)\to[0,\infty]$ be a l.s.c. cost function and let $(\mu,\nu)\in\mathscr{M}_+(X)\times\mathscr{M}_+(Y)$. If $c$ is continuous then*

$$C_c^{\mathfrak{h}}(\mu,\nu)=C_h(\mu,\nu),$$

*where $h$ is the sublinear cost function such that for all $(x,y)\in X\times Y$, $h_{x,y}$ is the sublinear relaxation of $c(x,\cdot,y,\cdot)$ and $C_h$ is the associated optimal semi-coupling cost. Equality also holds if $c(x,\cdot,y,\cdot)$ is l.s.c. and sublinear for all $(x,y)\in X\times Y$.*

*Proof.* We first prove the inequality $C_h\le C_c^{\mathfrak{h}}$ which holds in general. The reverse inequality which seems intuitive from Lemma 1.2.14 would require, in general, to make a selection of minimizers whose measurability is not evident. Thus we adopt different proof strategies to prove it: it is easy when $c=h$ and it involves an approximation argument if $c$ is merely continuous.

**Step 1 ($C_h\le C_c^{\mathfrak{h}}$).** Let $\gamma\in\mathscr{P}((X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+))$ be a feasible coupling for (1.2.6) and disintegrate it w.r.t. $\pi^{x_1,x_2}:(x_1,u_1,x_2,u_2)\mapsto(x_1,x_2)$ so that it writes

$$\gamma(\mathrm{d}x_1,\mathrm{d}u_1,\mathrm{d}x_2,\mathrm{d}u_2)=\sigma_{x_1,x_2}(\mathrm{d}u_1,\mathrm{d}u_2)\bar\gamma(\mathrm{d}x_1,\mathrm{d}x_2).$$

where $(\sigma_{x_1,x_2})$ is a family of measures in $\mathscr{P}(\mathbb{R}_+^2)$ and $\bar\gamma=\pi^{x_1,x_2}_{\#}\gamma\in\mathscr{P}(X\times Y)$. We moreover define the functions in $L^1(\bar\gamma)$

$$f_i:(x_1,x_2)\mapsto\mathfrak{h}(\pi^{u_i}_{\#}(\sigma_{x_1,x_2}))\quad\text{for } i\in\{1,2\}.$$

One has, by Lemma 1.2.14 (in particular (1.2.8)),

$$\begin{aligned}\int c\mathrm{d}\gamma&=\int_{X\times Y}\left(\int_{\mathbb{R}_+^2}c(x_1,u_1,x_2,u_2)\sigma_{x_1,x_2}(\mathrm{d}u_1,\mathrm{d}u_2)\right)\mathrm{d}\bar\gamma(\mathrm{d}x_1,\mathrm{d}x_2)\\&\ge\int_{X\times Y}h_{x_1,x_2}(f_1(x_1,x_2),f_2(x_1,x_2))\bar\gamma(\mathrm{d}x_1,\mathrm{d}x_2)\ge C_h(\mu,\nu).\end{aligned}$$

where the last inequality holds because $(f_1\bar\gamma,f_2\bar\gamma)$ forms an admissible pair of semi-couplings between $(\mu,\nu)$. Taking the infimum over $\gamma$ proves the first claim.

**Step 2 ($C_h\ge C_c^{\mathfrak{h}}$ when $c=h$).** If $c$ is already sublinear w.r.t. $(u_1,u_2)$, i.e. $c=h$, let $(f_1\bar\gamma,f_2\bar\gamma)$ be an optimal semi-coupling between $(\mu,\nu)$, where $\bar\gamma\in\mathscr{P}(X\times Y)$ and $f_1,f_2\in L^1(\bar\gamma)$. We build a probability $\gamma\in\mathscr{P}((X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+))$ that satisfies $(\mathfrak{h}(\pi^{x_1,u_1}_{\#}\gamma),\mathfrak{h}(\pi^{x_2,u_2}_{\#}\gamma))=(\mu,\nu)$ by posing $T:(x,y)\mapsto(x_1,f_1(x_1,x_2),x_2,f_2(x_1,x_2))$ and $\gamma:=T_{\#}\bar\gamma$. One has

$$C_h(\mu,\nu)=\int_{X\times Y}h_{x_1,x_2}(f_1(x_1,x_2),f_2(x_1,x_2))\gamma(\mathrm{d}x_1,\mathrm{d}x_2)=\int c\,\mathrm{d}\gamma\ge C_c^{\mathfrak{h}}.$$

**Step 3 ($C_h \geq C_c^{\mathfrak{h}}$ when $c$ continuous).** Consider a sequence of finite atomic semi-couplings for $n \in \mathbb{N}$

$$\hat{\gamma}_1^{(n)} := \sum_{i \in I} f_i^{(n)} \delta_{(x_i,y_i)^{(n)}}, \qquad (\hat{\gamma}_2^{(n)}) := \sum_{i \in I} g_i^{(n)} \delta_{(x_i,y_i)^{(n)}}$$

given by Lemma 1.2.11 which is such that the sequences $(\mu_n) := (\pi^x_\# \hat{\gamma}_1^{(n)})$ and $(\nu_n) := (\pi^y_\# \gamma_2^{(n)})$ weakly converge to $\mu$ and $\nu$, and such that $\lim_{n\to\infty} \int_{X\times Y} h_{x,y}(\hat{\gamma}_1^{(n)}, \hat{\gamma}_2^{(n)}) = C_h(\mu,\nu)$. Note that this lemma applies because since $c$ is continuous, $h$ is continuous, as can be seen from the characterizations of Lemma 1.2.14.

Choose an arbitrarily small $\varepsilon > 0$. There exists an $\eta > 0$ such that for any $n \in \mathbb{N}$ and $(x_i,y_i)^{(n)}$, one can choose a measure $\sigma_i^{(n)} \in \mathscr{M}_+([\eta,\infty[^2)$ which is such that

$$\int c_{(x_i,y_i)^{(n)}}(u_1,u_2)\sigma_i^{(n)}(\mathrm{d}u_1,\mathrm{d}u_2) = h_{(x_i,y_i)^{(n)}}(f_i^{(n)}, g_i^{(n)}) + \varepsilon$$

and $(\mathfrak{h}\pi^1_\# \sigma_i^{(n)}, \mathfrak{h}\pi^2_\# \sigma_i^{(n)}) = (f_i^{(n)}, g_i^{(n)}) \in \mathbb{R}_+^2$. This is made possible by Lemma 1.2.14 and the security distance from zero is possible because we allow a small error $\varepsilon$ and $h$ is uniformly continuous. Finally, let $\gamma^{(n)} \in \mathscr{M}_+((X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+))$ be defined as

$$\gamma^{(n)} := \sum_i \delta_{(x_i,y_i)^{(n)}}(\mathrm{d}x,\mathrm{d}y) \otimes \sigma_i^{(n)}(\mathrm{d}u_1,\mathrm{d}u_2).$$

If we prove that $\gamma^{(n)}$ admits a weakly convergent subsequence $\gamma^{(n)} \rightharpoonup \gamma$, then we would have

$$\int c\,\mathrm{d}\gamma \leq \liminf_{n\to\infty} \int c\mathrm{d}\gamma^{(n)} \leq \varepsilon + \liminf_{n\to\infty} \int_{X\times Y} h_{x,y}(\hat{\gamma}_1^{(n)}, \hat{\gamma}_2^{(n)}) = \varepsilon + C_h(\mu,\nu)$$

and, combined with the fact that, by construction,

$$\mathfrak{h}(\pi^{x,u_1}_\#(\gamma^{(n)})) = \mu_n \rightharpoonup \mu \quad \text{and} \quad \mathfrak{h}(\pi^{y,u_2}_\#(\gamma^{(n)})) = \nu_n \rightharpoonup \nu$$

then we would have built a feasible lifted plan, so $C_c^{\mathfrak{h}}(\mu,\nu) \leq C_h(\mu,\nu) + \varepsilon$ and the result would follow. As a consequence, it only remains to prove the compactness of $\{\gamma^{(n)}\}_{n\in\mathbb{N}}$. The mass is uniformly upper bounded because, using the "partial expectation" constraints satisfied by $\gamma^{(n)}$, for all $n \in \mathbb{N}$,

$$\frac{1}{\eta}\mu_n(X) \geq \int_X \int_{[\eta,\infty[} \mathrm{d}(\pi^{x,u_1}_\# \gamma^{(n)}) = \gamma^{(n)}((X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+)).$$

and $(\mu_n(X))$ is bounded by some constant $M > 0$. As for the tightness, one may use a similar argument: let $\varepsilon > 0$, then posing $K_\varepsilon := (X\times[0,M/\varepsilon])\times(Y\times[0,M/\varepsilon])$ and $K_\varepsilon^c$ its complementary set in $(X\times\mathbb{R}_+)\times(Y\times\mathbb{R}_+)$, one has for all $n \in \mathbb{N}$,

$$M \geq \mu_n(X) \geq \gamma(K_\varepsilon^c)\cdot M/\varepsilon$$

and it follows $\gamma(K_\varepsilon^c) < \varepsilon$. As a consequence, by Prokorov's theorem, $\{\gamma^{(n)}\}_{n\in\mathbb{N}}$ is pre-compact, and the proof is complete. □

**Lemma 1.2.14** (Sublinear regularization)**.** *Let $f:\mathbb{R}^n\to[0,\infty]$ be a l.s.c. function and let $\tilde f$ be its sublinear regularization. One has the characterizations, for $x\in\mathbb{R}^n\setminus\{0\}$,*

$$\tilde f(x)=\inf_{(\lambda_i,x_i)_{i=1}^n\in(\mathbb{R}_+\times\mathbb{R}^n)^n}\left\{\sum_{i=1}^n\lambda_i f(x_i)\;;\;\sum_{i=1}^n\lambda_i x_i=x\right\} \tag{1.2.7}$$

$$=\inf_{\sigma\in\mathscr{M}_+(\mathbb{R}^n)}\left\{\int_{\mathbb{R}^n}f(r)\mathrm{d}\sigma(r)\;;\;(\mathfrak{h}(\pi^1_\#\sigma),\dots,\mathfrak{h}(\pi^n_\#\sigma))=x\right\} \tag{1.2.8}$$

$$=\sup_{y\in\mathbb{R}^n}\{y\cdot x\;;\;y\cdot r\le f(r),\;\forall r\in\mathbb{R}^n\}. \tag{1.2.9}$$

*Proof.* We will prove $\tilde f=(1.2.7)\ge(1.2.8)\ge(1.2.9)\ge\tilde f$. The first characterization is a corollary of Caratheodory's theorem [134, Cor. 17.1.6]. The second is clearly smaller since given an admissible point $(\lambda_i,x_i)_{i=1}^n$ for (1.2.7), one builds an admissible measure $\sum\lambda_i\delta_{x_i}$ for (1.2.8) of same value. The inequality (1.2.8) $\ge$ (1.2.9) is a convex *weak* duality result or in other words, an inequality of type $\inf\sup\ge\sup\inf$, for which one only needs the expression of the convex conjugate of $\sigma\mapsto\int f\mathrm{d}\sigma$ when $f$ l.s.c., which is classical in optimal transport theory, and of the adjoint $(\mathfrak{h}\circ\pi^i)^*(y_i)=u$ where $y_i\in\mathbb{R}$ and $u\in\mathscr{C}(\mathbb{R}^n)$ is $u(r_1,\dots,r_n)=y_ir_i$. One cannot directly claim that strong duality holds because there is no always a strictly feasible $y\in\mathbb{R}^n$ for the dual problem. Finally, denoting $T(f)$ the function defined by formula (1.2.9), one has easily $T(f)\ge T(\tilde f)$ since $f\ge\tilde f$, and $T(\tilde f)=\tilde f$ because $\tilde f^*$ is the indicator of a convex set. So $T(f)\ge\tilde f$. □

### Optimal entropy-transport problems

Another class of unbalanced optimal transport problems was introduced in [103], the *optimal entropy-transport* problems. They correspond to relaxing the marginal constraints in the Kantorovich formulation, i.e. replacing the equality constraint on the marginals of the couplings by divergence functionals that quantify how much these marginals deviate from the desired measures. For measuring this deviation, a class of functionals with good properties is given by the *$f$-divergences*, also known as *Csiszár divergences*, which are built from entropy functions. The results in this paragraph are not our contribution but will be extremely useful in other chapters of this thesis and they complete the general picture of unbalanced optimal transport. This formulation is also central for the numerical schemes developed in Chapter 3.

**Divergences between measures** The following definition is more restrictive than the one adopted in [103] but covers most of the interesting cases.

**Definition 1.2.15** (entropy function)**.** *An entropy function is a convex l.s.c. function $f:\mathbb{R}\to\mathbb{R}\cup\{\infty\}$ such that* $\operatorname{dom}f\subset\mathbb{R}_+$ *and $f(1)=0$ is a minimum.*

**Definition 1.2.16** ($f$-divergence)**.** *A $f$-divergence is an integral functional associated to the perspective of an entropy function. Given an entropy function $f$, it is defined for $(\mu,\nu)\in\mathscr{M}(X)^2$ with the Lebesgue decomposition $\mu=\sigma\nu+\mu^\perp$ as*

$$\int_X\psi_f(\mu|\nu)=\int_X f(\sigma)\,\mathrm{d}\nu+f'_\infty(1)\cdot\mu^\perp(X).$$

The convention for the last term is $\infty \cdot 0 = 0$. Notice that this divergence is worth 0 if $\mu = \nu$, increases as $\mu$ deviates from $\nu$ and is $\infty$ if either $\mu$ or $\nu$ is not nonnegative. These functionals are a particular case of integral functionals of sublinear functions and enjoy consequently the same duality, convexity and lower-semicontinuity properties (see Appendix A).

**Remark 1.2.17.** *This perspective function is a bit peculiar because the two variables of $\psi_f$ belong to $\mathbb{R}_+$. As a consequence, one can select any line in $\mathbb{R}^2$ tangent to the unit circle in the positive quadrant $\mathbb{R}^2_+$ and $\psi_f$ is the perspective function of its own restriction to this line, modulo a change of coordinate. In particular, taking the line $\mathbb{R} \times \{1\}$, one has*

$$\psi_f(\mu|\nu) = \psi_g(\nu|\mu)$$

*where $g(y) = \psi_f(1|y)$. This second description is called the reverse entropy and is used in [103] to give alternative descriptions of the optimal entropy-transport problem.*

**Optimal entropy-transport problems** We only introduce the most useful case with one entropy function.

**Definition 1.2.18** (Optimal entropy-transport)**.** *Let $c : X \times Y \to [0,\infty]$ be a l.s.c. cost function and $f$ be an entropy function. For $(\mu,\nu) \in \mathscr{M}_+(X) \times \mathscr{M}_+(Y)$, the optimal entropy transport cost is defined as*

$$C_{c,f}(\mu,\nu) := \min \left\{ \int_{X\times Y} c\mathrm{d}\gamma + \int_X \psi_f(\pi^1_\#\gamma|\mu) + \int_Y \psi_f(\pi^2_\#\gamma|\nu) \; ; \; \gamma \in \mathscr{M}_+(X\times Y) \right\}. \quad (1.2.10)$$

The following duality result is another analogue of Kantorovich duality for unbalanced optimal transport, and just like Kantorovich duality, it involves the operator $\oplus$, defined as $(\psi \oplus \psi)(x,y) = \psi(x) + \phi(y)$. Note that after a change of variables, it can be re-written under the form of Theorem 1.2.4.

**Theorem 1.2.19** (Existence, duality [103])**.** *If $(\mu(X) \cdot \mathrm{dom}\, f) \cap (\nu(Y) \cdot \mathrm{dom}\, f) \neq \emptyset$, then problem* (1.2.10) *admits a least a minimizer and moreover*

$$C_{c,f}(\mu,\nu) = \sup_{(\phi,\psi)\in\mathscr{C}(X)\times\mathscr{C}(Y)} \left\{ -\int_X f^*(-\phi)\mathrm{d}\mu - \int_X f^*(-\psi)\mathrm{d}\nu \; ; \; \phi \oplus \psi \leq c \right\}$$

*where in the case when $f$ is not superlinear, one also impose $\psi(x), \phi(y) \in \mathrm{dom}\, f^*$ for all $x \in X$ and $y \in Y$.*

*Proof.* This result is proved in a much broader setting in [103] (where also existence of dual maximizers is proved in several contexts). In the present case, this can also be seen as a consequence of Fenchel-Rockafellar duality theorem, using the duality results on integral functionals of measures (see Appendix A). □

It is insightful, as remarked in [103], to fix a pair of points $(x,y) \in X \times Y$ and solve the minimization problem "restricted" to these points. The resulting minimal value defines a sublinear cost function that can be used in a optimal lift or an optimal semi-coupling problem.

**Definition 1.2.20** (Marginal perspective [103]). *Given an entropy function $f$ and $c \in [0,\infty]$ the marginal perspective function is defined, with the convention $0 \cdot \infty = 0$, as*

$$h_c(a,b) \coloneqq \inf_{\theta \geq 0} \psi_f(\theta|a) + \psi_f(\theta|b) + \theta c.$$

*Given a cost $c : X \times Y \to [0,\infty]$ and an entropy function $f$, define the induced cost*

$$h_{x,y}(a,b) \coloneqq h_{c(x,y)}(a,b) \quad \text{for } x,y \in X \times Y \text{ and } a,b \geq 0.$$

This construction allows to obtain equivalence with the previously introduced formulations. This result is applied several times in Chapter 2 to specific examples of unbalanced optimal transport models.

**Theorem 1.2.21** (from entropy-transport to semi-coupling). *Let $h$ be the marginal perspective cost associated to a pair $(c,f)$ of cost function/entropy function. One has for all $(\mu,\nu) \in \mathscr{M}_+(X) \times \mathscr{M}_+(Y)$ and writing $\tilde{c} = h$*

$$C_{c,f}(\mu,\nu) = C^{\mathfrak{h}}_{\tilde{c}}(\mu,\nu) = C_h(\mu,\nu).$$

*Proof.* This is a particular case of [103, thm. 5.8], combined with the equivalence of Theorem 1.2.13 which is exact since the marginal perspective cost is l.s.c. sublinear. □

It should be noted that, in contrast, semi-coupling problems do not admit in general an optimal entropy-transport formulation unless they have a very specific structure.

## 1.3. From dynamic to coupling problems

In the previous sections, we have introduced and studied basic properties of two categories of unbalanced optimal transport problems: dynamic formulations, and coupling formulations. In this section, we prove that any dynamic problem is equivalent to a coupling problem. This result which the link between the two approaches developed independently in Sections 1.1 and 1.2.

We adopt the setting of Section 1.1 : $\Omega \subset \mathbb{R}^d$ is the closure of a bounded, Lipschitz, connected open domain and we recall that $\mathfrak{C}(\Omega)$ denotes the cone of $\Omega$ (Definition 1.2.6). The function $L : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a Lagrangian that satisfies Assumption 1.1.10: it is convex, continuous and admits the unique minimum $L(0,0) = 0$. We also assume the following:

$$\text{there exists } p > 0 \text{ such that } \lim_{|g|\to\infty} L(0,g)/|g|^p = 0,$$

that we simply state as "$L$ has polynomial growth".

### Minimal path cost and convexification

Any Lagrangian defines a cost function on $\Omega \times \mathbb{R}_+$ obtained by minimizing over absolutely continuous trajectories linking pairs of points.

**Definition 1.3.1** (minimal path cost)**.** *The minimal path cost associated to the Lagrangian L is, for* $(x_i,u_i) \in (\Omega \times \mathbb{R}_+)^2$,

$$c_L((x_1,u_1),(x_2,u_2)) \coloneqq \inf_{x(t),u(t)} \int_0^1 \psi_L(u(t)x'(t),u'(t) \mid u(t))\mathrm{d}t$$

*where the infimum is taken over* $(x_t,u_t)_{t\in[0,1]}$ *such that u is absolutely continuous in* $\mathbb{R}_+$ *and x is piecewise absolutely continuous in* $\Omega$*,* $(x(i),u(i)) = (x_i,u_i)$ *for* $i \in \{0,1\}$ *and with a finite number of discontinuity points for x at* $(t_i)_{i=1}^k$ *such that* $u(t_i) = 0$*.*

Allowing for discontinuities in $x$ when the mass $u$ vanishes expresses the fact that we are rather interested in absolutely continuous paths *in the cone* and intuitively, it simply means that a path that reaches the apex at some $x_a \in \Omega$ can instantaneously escape the apex from another point $x_b \in \Omega$. The definition without this property would be sufficient for us (because we will regularize $c_L$ anyways) but the construction of minimizing sequences would be more tedious. More generally, the proof of Theorem 1.3.3 only requires the definition of $c_L$ to have the following properties :

- each path $(x(t),u(t))$ in the minimizing space is such that

$$(u(t)\delta_{x(t)}, x'(t)u(t)\delta_{x(t)}, u'(t)\delta_{x(t)}) \in \mathrm{CE}_0^1(u_0\delta_{x_0}, u_1\delta_{x_1});$$

- the space of minimization contains $\mathscr{C}^1$ paths between pairs of points with positive mass;
- $c_L$ is continuous.

Since the Lagrangian is continuous, one has that $c_L$ is continuous on its domain (one can perturb a path and reach nearby points with a perturbed cost). The behavior at the apex of the cone is less clear but if we assume that $L$ has a polynomial growth, an absolutely continuous path can be constructed as in Proposition 1.1.14 and $c_L$ is continuous on $\Omega \times \mathbb{R}_+$. In general, $c_L$ does not define a sublinear cost function, but its convex regularization does so.

**Proposition 1.3.2** (convexification of $c_L$)**.** *Assume that $L$ has a polynomial growth and let $h^{(c_L)}_{x_1,x_2}$ be the convex regularization of $c_L$ for all $(x_1,x_2) \in \Omega^2$. Then $h^{(c_L)}$ is an admissible sublinear cost function, is continuous and is characterized by*

$$h^{(c_L)}_{x1,x2}(r_1,r_2) := \min_{\substack{r_1^a+r_1^b=r_1 \\ r_2^a+r_2^b=r_2}} c((x_1,r_1^a),(x_2,r_2^a)) + c((x_1,r_1^b),(x_2,r_2^b)). \tag{1.3.1}$$

*Proof.* It is clear that $h^{(c_L)}$ takes its values in $[0,\infty]$ and that it inherits 1-homogeneity from $\psi_L$, so $h_{x_1,x_2}$ is sublinear for all $(x_1,x_2)$. Also, by Lemma 1.2.14, and the fact that $c_L$ is homogeneous, one has the characterization (1.3.1) (one has to check directly that it is also valid for $(r_1,r_2) = (0,0)$). The infimum is attained because $c_L$ is continuous and the minimization set is compact. The continuity of $h^{(c_L)}$ is clear from this characterization and the continuity of $c_L$. □

This convexification of the Lagrangian cost, which does not appear in the classical optimal transport theory, is an essential step that explains differences between the cut-locus distance of minimal path associated to a Lagrangian and the minimal measure valued interpolations (see examples in Chapter 2).

### Main theorem

The following theorem, reminiscent of the Benamou-Brenier theorem [12], shows that any dynamic optimal transport problem is equivalent to a coupling problem. It is a key step to relate many apparently independent approaches for defining unbalanced optimal transport models. In Chapter 2, we apply this result on several explicit models. Note also that a similar result is obtained in [103] for the quadratic case considered in Chapter 2, with a different proof technique based on a representation formula for the dual problem.

**Theorem 1.3.3** (Dynamic to static)**.** *Let $L$ be an admissible Lagrangian of polynomial growth and $h^{(c_L)}$ be the associated sublinear cost as defined in Proposition 1.3.2. Then the dynamic and semi-coupling problems are equivalent, i.e.*

$$C_{h^{(c_L)}}(\mu,\nu) = C_L(\mu,\nu) \quad \forall (\mu,\nu) \in \mathscr{M}_+(\Omega)^2.$$

*Proof.* For ease of reading, we shall denote $h^{(c_L)}$ by simply $h$. This proof is divided into three steps: in Step 1, we show by a discrete approximation of semi-couplings that it holds $C_h \geq C_L$. In Step 2, we show that a converse inequality holds in a smooth setting, a result used in Step 3 to show that $C_h \leq C_L$, via a regularization argument.

**Step 1.** Let $(\mu,\nu)\in\mathscr{M}_+(\Omega)^2$ and let $(\hat\gamma_1^{(n)})$, let $(\hat\gamma_2^{(n)})$ be the atomic measures on $X\times Y$ given by Lemma 1.2.11 (the continuity of $h$ follows from Proposition 1.3.2): they are such that $\lim_{n\to\infty}\int_{X\times Y}h_{x,y}(\hat\gamma_1^{(n)},\hat\gamma_2^{(n)})=C_h(\mu,\nu)$. Also denote by $(\mu_n,\nu_n):=(\pi^1_\#\hat\gamma_1^{(n)},\pi^2_\#\hat\gamma_2^{(n)})$ the corresponding marginals, that satisfy $\mu_n\rightharpoonup\mu$ and $\nu_n\rightharpoonup\nu$.

By the definition of $h$, and in particular by the characterization (1.3.1), one has $C_L(u_1\delta_x,u_2\delta_y)\le h_{x,y}(u_1,u_2)$ because to any absolutely continuous path in the cone $(x(t),u(t))_{t\in[0,1]}$, corresponds an admissible dynamic interpolation between atoms of the form $\big(u(t)\delta_{x(t)}\big)_{t\in[0,1]}$ which action is $\int_0^1\psi_L(u(t)x'(t),u'(t)|u(t))\mathrm{d}t$. It follows

$$\begin{aligned}\int_{\Omega^2}h_{x,y}(\bar\gamma_1^{(n)},\bar\gamma_2^{(n)})&=\sum_{i\in I}h_{(x_i,y_i)}(f_i^{(n)},g_i^{(n)})\gamma_i^{(n)}\\&\ge\sum_{i\in I}C_L(f_i^{(n)}\gamma_i^{(n)}\delta_{x_i},g_i^{(n)}\gamma_i^{(n)}\delta_{y_i})\ge C_L(\mu_n,\nu_n).\end{aligned}$$

Since $C_L$ is weakly l.s.c., $C_L(\mu,\nu)\le\liminf C_L(\mu_n,\nu_n)$, and it follows $C_L(\mu,\nu)\le C_h(\mu,\nu)$.

**Step 2.** Let $\mu\in\mathscr{M}_+(\Omega)$ be an absolutely continuous measure of strictly positive mass, let $(v,g)\in\mathscr{C}^1([0,1]\times\Omega;\mathbb{R}^d\times\mathbb{R})$ be velocity and growth field (assume no-flux boundary conditions for $v$) and let $(\rho_t)_{t\in[0,1]}$ the unique classical solution of $\partial_t\rho_t+\nabla\cdot(v_t\rho_t)=g_t\rho_t$ such that $\rho_0=\mu$. This solution can be written as a flow $\rho_t=(Y_t)_\#(G_t\mu)$ where $Y_t$ and $G_t$ are, as in Theorem 1.1.5, the unique solutions to

$$\begin{cases}\partial_tY_t(x)=v_t(Y_t(x))\\ Y_0(x)=x\end{cases}\quad\text{and}\quad\begin{cases}\partial_tG_t(x)=g_t(Y_t(x))G_t(x)\\ G_0(x)=1\end{cases}.$$

One has

$$\begin{aligned}\int_0^1\int_\Omega\psi_L(\omega,\zeta|\rho)&=\int_0^1\int_\Omega L(v_t(x),g_t(x))\mathrm{d}(Y_t)_\#(G_t\mu)(x)\mathrm{d}t\\&=\int_\Omega\left(\int_0^1L(\partial_tY_t(x),\partial_tG_t(x)/G_t(x))G_t(x)\mathrm{d}t\right)\mathrm{d}\mu(x)\\&\overset{(1)}{\ge}\int_\Omega h_{x,Y_1(x)}(1,G_1(x))\mathrm{d}\mu(x)\overset{(2)}{\ge}C_h(\mu,\nu)\end{aligned}$$

where (1) is a consequence of $h\le c_L$ and (2) comes from the fact that $(\mathrm{id},Y_1)_\#\mu$ and $(\mathrm{id},Y_1)_\#(G_1\rho_0)$ satisfy the semi-coupling constraints of (1.2.2).

**Step 3.** We now show with the help of Step 2, that for $\mu,\nu\in\mathscr{M}_+(\Omega)$, it holds $C_h(\mu,\nu)\le C_L(\mu,\nu)$. Let $(\rho,\omega,\zeta)$ be a solution to the continuity equation between $\mu$ and $\nu$ and for $\alpha\in]0,1[$, let

$$\tilde\rho^\alpha=(1-\alpha)\rho+\alpha\,(\mathrm{d}t\otimes\mathrm{d}x)|_S,\quad\tilde\omega^\alpha=(1-\alpha)\omega,\quad\tilde\zeta^\alpha=(1-\alpha)\zeta$$

where $S\subset\mathbb{R}^d$ is a bounded set containing $\Omega+B^d(0,1)$ and $B^d(a,r)$ denotes the open ball of radius $r$ centered at point $a$ in $\mathbb{R}^d$. The new triplet $(\tilde\rho^\alpha,\tilde\omega^\alpha,\tilde\zeta^\alpha)$ is a solution to the continuity

equation between its endpoints $(\tilde\rho_0^\alpha,\tilde\rho_1^\alpha)$ which weakly converge to $(\mu,\nu)$ as $\alpha\to 0$. Since by Proposition 1.2.5, $C_h$ is weakly l.s.c. one has

$$C_h(\mu,\nu)\le \liminf_{\alpha\to 0} C_h(\tilde\rho_0^\alpha,\tilde\rho_1^\alpha)$$

and, by convexity of $\psi_L$,

$$\int_0^1\int_S \psi_L(\tilde\omega^\alpha,\tilde\zeta^\alpha|\tilde\rho^\alpha)\le \int_0^1\int_\Omega \psi_L(\omega,\zeta|\rho),$$

so is is sufficient to prove $C_h(\tilde\rho_0^\alpha,\tilde\rho_1^\alpha)\le\int_0^1\int_S\psi_L(\tilde\omega^\alpha,\tilde\zeta^\alpha|\tilde\rho^\alpha)$ for proving $C_h(\mu,\nu)\le C_L(\mu,\nu)$. In order to alleviate notations we shall denote $\tilde\rho^\delta,\tilde\omega^\delta,\tilde\zeta^\delta$ by just $\rho,\omega,\zeta$ and the new marginals $\tilde\rho_0^\delta,\tilde\rho_1^\delta$ by $\rho_0,\rho_1$ (now supported on $S\supset\Omega$) and extend the definition of $(\rho,\omega,\zeta)$ for $t<0$ by $(\rho_0\otimes \mathrm{d}t,0,0)$ and for $t>1$ by $(\rho_1\otimes \mathrm{d}t,0,0)$.

Let $r_\varepsilon\in C_c^\infty(\mathbb{R}^d;\mathbb{R}_+)$ be a mollifier of the form $r_\varepsilon(t,x)=\frac{1}{\varepsilon^d}r_1\left(\frac{x}{\varepsilon}\right)\frac{1}{\varepsilon}r_2\left(\frac{t}{\varepsilon}\right)$ where $r_1\in C_c^\infty(B^d(0,\frac12))$, $r_2\in C_c^\infty(B^1(0,\frac12))$, $r_i\ge 0$, $\int r_i=1$, $r_i$ even ($i=1,2$). Define $(\rho^\varepsilon,\omega^\varepsilon,\zeta^\varepsilon)$ the regularized triplet $(\rho*r_\varepsilon,\omega*r_\varepsilon,\zeta*_\varepsilon)$ which, by Proposition 1.1.9 solves the continuity equation on $[-\frac{\varepsilon}{2},1+\frac{\varepsilon}{2}]\times\Omega^\varepsilon$ where $\Omega^\varepsilon:=\Omega+B^d(0,\varepsilon)$ between $(\mu^\varepsilon,\nu^\varepsilon)$ that weakly converge to $\mu$ and $\nu$ as $\varepsilon\to 0$. Moreover, the vector fields $v_t^\varepsilon:=\omega_t^\varepsilon/\rho_t^\varepsilon$ and $g_t^\varepsilon:=\zeta_t^\varepsilon/\rho_t^\varepsilon$ are well defined on $S$, smooth and bounded in particular because $\rho$ is uniformly bounded from below on $S\supset\Omega^\varepsilon$ as long as $\varepsilon<1$. Besides, $v^\varepsilon$ satisfies no-flux boundary conditions on $\Omega^\varepsilon$. Let $c_\varepsilon$ be the minimal path cost related to the Lagrangian $L$ on $\Omega^\varepsilon$ and the time range $[\frac{\varepsilon}{2},1+\frac{\varepsilon}{2}]$, and let $h_{c_\varepsilon}$ be its sublinear regularization as in Proposition 1.3.2. One has by step 2 and by sublinearity of $\psi_L$ (Lemma 1.1.23)

$$C_{h^{c\varepsilon}}(\rho_0^\varepsilon,\rho_1^\varepsilon)\le\int_{-\varepsilon/2}^{1+\varepsilon/2}\int_{\Omega^\varepsilon}\psi_L(\omega^\varepsilon,\zeta^\varepsilon|\rho^\varepsilon)\le\int_0^1\int_\Omega\psi_L(\omega,\zeta|\rho)$$

Since $\Omega^\varepsilon$ shrinks as $\varepsilon\to 0$, $h_{c_\varepsilon}$ converges pointwise and increasingly to $h$ (we may define $h_\varepsilon$ and $h$ as $+\infty$ outside of $\Omega^\varepsilon$ and $\Omega$, respectively, and use characterization (1.3.1) to see this), so by Proposition 1.2.5, one has

$$C_h(\mu,\nu)\le\liminf_{\varepsilon\to 0}C_{h^{c\varepsilon}}(\rho_0^\varepsilon,\rho_1^\varepsilon)\le\int_0^1\int_\Omega\psi_L(\rho,\omega,\zeta).$$

This allows to conclude that $C_h(\mu,\nu)\le C_L(\mu,\nu)$. □

# Chapter 2.

# Unbalanced Optimal Transport Metrics

In this chapter we study specific cases of the framework developed in Chapter 1. Starting from the dynamic formulation of unbalanced optimal transport, we define a new family of metrics, denoted $\widehat{W}_p$. We prove that, just as the $W_p$ optimal transport metrics, the $\widehat{W}_p$ optimal transport-growth metrics make the space of nonnegative measure a geodesic space and the convergence in $\widehat{W}_p$ is equivalent to weak convergence.

We study in Section 2.1 the limit geodesics obtained when a *scale* parameter goes to 0 or $\infty$. In one case, we obtain the geodesics for a class of pure growth metrics, that contains in particular the Hellinger and the Total Variation metrics. In the other we obtain generalized transport geodesics, that can be built from pure $W_p$ optimal transport geodesics by rescaling and reparametrization.

In Section 2.2, we consider the quadratic case for which the *minimal path cost* is explicit. It follows explicit form of the optimal lift, semi-coupling and optimal entropy-transport formulations by the results of Section 1.3. We also show that a non-trivial variant of $\widehat{W}_2$ with weights remains a metric for a certain range of parameters (Proposition 2.2.5).

In the third Section, we consider the limit case $p = 1$ and show that it corresponds to the Bounded Lipschitz metric. We also consider more general Lagrangians which are 1-homogeneous w.r.t. the growth variable: the so-called "generalized Wasserstein distances" [129] are recovered and we prove that they are equivalent to the problem of optimal partial transport, with a dual parametrization.

The material of this chapter is partially based on the published work [44] and the submitted article [43]. The first section is, at this level of generality, entirely new.

## 2.1. Unbalanced $\widehat{W}_p$ metrics

### 2.1.1. A class of geodesic metrics

In this chapter, we consider $\Omega \subset \mathbb{R}^d$ which is a domain with our standard assumptions (see section *Notation*). Let us consider the $p$-homogeneous Lagrangian $L : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ defined by

$$L_{p,\alpha}(v,g) := \left(\tfrac{1}{\alpha}|v|\right)^p + \left(\tfrac{1}{p}|g|\right)^p. \tag{2.1.1}$$

**Definition 2.1.1** (Metric $\widehat{W}_p$)**.** *For $\mu,\nu \in \mathscr{M}_+(\Omega)$, $\alpha > 0$ a scale parameter and $p \geq 1$ a homogeneity parameter, the unbalanced $\widehat{W}_{p,\alpha}$ metric between $\mu$ and $\nu$ is defined by*

$$\widehat{W}_{p,\alpha}(\mu,\nu) := C_{L_{p,\alpha}}(\mu,\nu)^{1/p}$$

*where $C_{L_{p,\alpha}}$ is the optimal dynamic cost associated to the Lagrangian $L_{p,\alpha}$ (Definition 1.1.13). For $\alpha = 1$, we simply write $\widehat{W}_p$.*

Let us gather a few results from the previous parts (see Theorems 1.1.25 and 1.1.26 and Proposition 1.1.16).

**Theorem 2.1.2.** *The space $(\mathscr{M}_+(\Omega), \widehat{W}_{p,\alpha})$ is a complete geodesic metric space that metrizes weak convergence. Moreover, for $\beta > 0$ and $\mu,\bar{\mu},\nu,\bar{\nu} \in \mathscr{M}_+(\Omega)$, it holds $\widehat{W}_{p,\alpha}(\beta\mu,\beta\nu) = \beta^{\frac{1}{p}}\widehat{W}_{p,\alpha}(\mu,\nu)$ and*

$$\widehat{W}_{p,\alpha}(\mu+\bar{\mu},\nu+\bar{\nu}) \leq \left(\widehat{W}_{p,\alpha}(\mu,\nu)^p + \widehat{W}_{p,\alpha}(\bar{\mu},\bar{\nu})^p\right)^{\frac{1}{p}}.$$

**Proposition 2.1.3** ($\alpha$ is a scale parameter)**.** *Let $s : x \mapsto \alpha x$ be a linear rescaling of parameter $\alpha > 0$ and $s_\#$ the associated pushforward map, then*

$$s_\# : (\mathscr{M}_+(\Omega), \widehat{W}_{p,1}) \to (\mathscr{M}_+(\alpha\Omega), \widehat{W}_{p,\alpha})$$

*is an isometry.*

*Proof.* Given any solution $(\rho,\omega,\zeta)$ of the continuity equation between $\mu$ and $\nu$ on $[0,1]$, one has by Proposition 1.1.6 that $(s_\#\rho, \alpha s_\#\omega, s_\#\zeta)$ solves the continuity equation between $s_\#\mu$ and $s_\#\nu$. Moreover, thanks to Lemma 2.1.4 below, it holds

$$\int_0^1 \int_{\alpha\Omega} \psi_{L_{p,\alpha}}(\alpha s_\#\omega, s_\#\zeta | s_\#\rho) = \int_0^1 \int_\Omega \psi_{L_{p,\alpha}}(\alpha\omega,\zeta|\rho) = \int_0^1 \int_\Omega \psi_{L_{p,1}}(\omega,\zeta|\rho)$$

and the result follows. □

In the proof we used the next lemma (see also the related Lemma 1.1.23).

**Lemma 2.1.4** (Image measure and pushforward)**.** *If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is (positively) 1-homogeneous measurable and $T : X \to Y$ is a measurable bijective map of measurable inverse on measurable spaces $X,Y$, then $f(T_\#\mu) = T_\#(f(\mu))$ for all $\mu \in \mathscr{M}(X;\mathbb{R}^n)$. If $T$ is only measurable but $f$ is sublinear and $X,Y$ are Radon spaces, one has $f(T_\#\mu) \leq T_\#(f(\mu))$.*

*Proof.* Let $\lambda \in \mathscr{M}_+(X)$ be a measure that dominates $\mu$ and denote $\sigma := \mathrm{d}\mu/\mathrm{d}\lambda$. In the first case one has, for all measurable $B \subset Y$

$$T_\#\mu(B) = \mu(T^{-1}(B)) = \int_{T^{-1}(B)} \sigma \mathrm{d}\lambda = \int_B (\sigma \circ T^{-1}) \mathrm{d}(T_\#\lambda)$$

so one has $T_\#\mu = (\sigma \circ T^{-1})T_\#\lambda$. It follows

$$f(T_\#\mu)(B) = \int_B f \circ \sigma \circ T^{-1} \mathrm{d}T_\#\lambda = \int_{T^{-1}(B)} f \circ \sigma \mathrm{d}\lambda = f(\mu)(T^{-1}(B)) = (T_\# f(\mu))(B).$$

In the other case, one may apply the disintegration theorem [5, Thm. 5.3.1] to $\lambda$ w.r.t. $T_\#\lambda$ to obtain that for all Borel function $\phi : Y \to \mathbb{R}$,

$$\int_Y \phi T_\#\mu = \int_Y \int_{T^{-1}(y)} (\phi \circ T)(x)\sigma(x)\lambda_y(\mathrm{d}x)T_\#\lambda(\mathrm{d}y) = \int_Y \phi(y) \left[ \int_{T^{-1}(y)} \sigma \mathrm{d}\lambda_y \right] T_\#\lambda(\mathrm{d}y)$$

so one has $\tilde{\sigma}(y) := \int_{T^{-1}(y)} \sigma \mathrm{d}\lambda_y = \mathrm{d}(T_\#\mu)/\mathrm{d}(T_\#\lambda)$. If follows, using Jensen inequality that for all Borel function $\phi : Y \to \mathbb{R}$,

$$\int_Y \phi \mathrm{d}T_\# f(\mu) = \int_Y \phi \left[ \int_{T^{-1}(y)} f \circ \sigma \mathrm{d}\lambda_y \right] \mathrm{d}T_\#\lambda \geq \int_Y \phi(y) f(\tilde{\sigma}) \mathrm{d}T_\#\lambda = \int_Y \phi \mathrm{d}f(T_\#\mu). \qquad \square$$

This behavior w.r.t. the spatial scale can also be stated as a contraction property.

**Proposition 2.1.5** (Contraction property)**.** *Let $s : x \mapsto \alpha x$ be a linear rescaling of parameter $0 < \alpha < 1$ and $s_\#$ the associated pushforward map, then, assuming $\alpha\Omega \subset \Omega$, $s_\#$ is a contraction for $\widehat{W}_p$. If moreover a minimizer for $C_{L_p}$ between $\mu$ and $\nu$ has nonzero momentum $\omega$, then this contraction is strict.*

*Proof.* The proof is similar to Proposition 2.1.3, using additionally the fact that $L_p(v,g)$ is strictly increasing with respect to $v$. $\square$

**Remark 2.1.6** (Convention and path cost)**.** *The factor $1/p$ in the definition of the Lagrangian is motivated by the following remark (consider $\alpha = 1$). Remember that by Theorem 1.3.3, the metric $\widehat{W}_p$ admits a coupling formulation, for a cost which is derived from the minimal path cost associated to the Lagrangian (Definition 1.3.1). In this case, the cost of an absolutely continuous path $(x,u)$ in $\Omega \times \mathbb{R}_+$ is*

$$\int_0^1 \left( |\dot{x}(t)|^p + \left( \frac{\dot{u}(t)}{p\,u(t)} \right)^p \right) u(t) \mathrm{d}t.$$

*which, after the change of variable $r := u^{1/p}$, rewrites*

$$\int_0^1 \left( (r(t)|\dot{x}(t)|)^p + |\dot{r}(t)|^p \right) \mathrm{d}t.$$

*One may then interpret $(r,x)$ as polar coordinates and remark that this energy is the integral of the $p$-th norm of the velocity vector, in the local reference frame. In general, we do not know an explicit expression for the minimal path cost for this Lagrangian, except in the situation of the next lemma. The case $p = 2$ is however simpler because the 2-norm is invariant by orthonormal change of basis so—in the 1-D case—one writes the path in a fixed basis by posing $(z_1, z_2) := (r\cos x, r\sin x)$ and minimal path are just straight lines in these coordinates (see Section 2.2).*

**Lemma 2.1.7** (Pure growth cost)**.** *The minimal path cost (Definition 1.3.1) between two points in the cone $(x,a)$ and $(x,b)$ sharing the same spatial location $x \in \Omega$ associated to the Lagrangian $L_{p,\alpha}$ is $h_p(a,b) := |a^{1/p} - b^{1/p}|^p$ and is reached for the unique minimizing path $(x_t,u_t) = (x,((1-t)a^{1/p} + tb^{1/p})^p)$. The function $h_p$ is continuous and sublinear on its domain $\mathbb{R}_+^2$.*

*Proof.* It is clear that the path $(x_t,u_t)$ has a minimal cost only if $x_t = x$ for all $t \in [0,1]$. After the change of variable $r_t = u_t^{1/p}$ which preserves absolute continuity, the problem is equivalent to minimizing $\int_0^1 |\dot{r}(t)|^p \mathrm{d}t$ subject to $r_0 = a^{1/p}$ and $r_1 = b^{1/p}$. By Jensen inequality, $\int_0^1 |\dot{r}(t)|^p \mathrm{d}t \geq |\int_0^1 \dot{r}(t)\mathrm{d}t|^p$ so this problem has a unique solution, satisfying $\dot{r}_t$ constant. It corresponds to the linear interpolation $r_t = (1-t)r_0 + tr_1$ and the associated cost is $|r_1 - r_0|^p$. Given this formula, $h_p$ is clearly continuous and positively 1-homogeneous. For the subadditivity, consider $(a_1,a_2,b_1,b_2) \in \mathbb{R}_+^4$. The pure growth path energy $\int_0^1 (\dot{u}_t/u_t)^p u_t \mathrm{d}t$ is sublinear in $(u_t,\dot{u}_t)$ and in particular, taking $(u^{(i)})_t$ the two minimal paths between the couples $(a_i,b_i)$ and $(u_t)$ their sum, one has

$$\int_0^1 (\dot{u}_t/u_t)^p u_t \mathrm{d}t \leq \int_0^1 (\dot{u}_t^{(1)}/u_t^{(1)})^p u_t^{(1)} \mathrm{d}t + \int_0^1 (\dot{u}_t^{(2)}/u_t^{(2)})^p u_t^{(2)} \mathrm{d}t$$

which implies $h_p(a_1+a_2,b_1+b_2) \leq h_p(a_1,b_1) + h_p(a_2,b_2)$ and $h_p$ is sublinear. □

We deduce the following explicit geodesics which correspond to the case when the most efficient interpolation has no transport component.

**Proposition 2.1.8** (Pure growth geodesics)**.** *Let $\mu \in \mathscr{M}_+(\Omega)$, $\beta \in \mathbb{R}_+$ and $p > 1$. One has $\widehat{W}_{p,\alpha}(\mu,\beta\mu) = |1-\beta^{1/p}|\mu(\Omega)^{1/p}$ and $\rho_t = (1-t+t\beta^{1/p})^p\mu$ is the unique geodesic.*

*Proof.* We denote $L_g : u \in \mathbb{R}_+ \mapsto (u/p)^p$. For $(\rho,\omega,\zeta) \in \mathrm{CE}_0^1(\mu,\alpha\mu)$, using Lemma 2.1.4 and the fact that $L_{p,\alpha}$ only depends on the norm of its arguments, one has, recalling Lemma 2.1.7,

$$\int_0^1 \int_\Omega \psi_{L_{p,\alpha}}(\omega,\zeta|\rho) \geq \int_0^1 \psi_{L_g}(|\zeta_t|(\Omega)|\rho_t(\Omega)) \geq |1-\beta^{1/p}|^p \mu(\Omega)$$

with equality if and only if $\omega = 0$ and $g_t = \mathrm{d}\zeta_t/\mathrm{d}\rho_t$ is constant $\rho_t$ a.e. in $\Omega$ for the first inequality and $\rho_t(\Omega) = ((1-t)\mu(\Omega)^{1/p} + t\beta^{1/p}\mu(\Omega)^{1/p})^p$ for the second inequality, by Lemma 2.1.7. □

## 2.1.2. Convergence to limit models

In this section, we study the behavior of $\widehat{W}_p$ and its geodesics when the scale parameter $\alpha$ tends to 0 or $\infty$, using the notion of $\Gamma$-convergence [24].

### Some $\Gamma$-convergence results

Let us first define $\Gamma$-convergence and recall classical results that are used several times in this thesis. Since we are interested in the weak convergence, which is metrizable on $\mathscr{M}_+(\Omega)$ (the metrics $\widehat{W}_p$ are an example), we do not need to distinguish between convergence and sequential convergence.

**Definition 2.1.9.** *Let $(X,d)$ be a complete metric space and $(F_n)_{n\in\mathbb{N}}$, $F$ be functionals $X \to \mathbb{R}\cup\{\infty\}$. The sequence $(F_n)$* $\Gamma$-converges *to $F$ if, for any $x \in X$, it holds:*

$\Gamma$-liminf. *For all $(x_n)$ converging to $x$, one has $\liminf F_n(x_n) \geq F(x)$;*

$\Gamma$-limsup. *There exists $(x_n)$ converging to $x$ such that $\limsup F_n(x_n) \leq F(x)$.*

The central result motivating this notion of $\Gamma$-convergence is that from the convergence of functionals, one may deduce the convergence of minimums and of minimizers.

**Theorem 2.1.10** (Convergence of minimizers)**.** *Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of functions $X \to \mathbb{R}\cup\{\infty\}$ such that $F_n$ $\Gamma$-converges to $F$ and assume that there exists a compact $K \subset X$ such that $\inf_X F_n = \inf_K F_n$, for all $n \in \mathbb{N}$. Then $F$ admits a minimum on $X$ and $(\inf_X F_n)$ converges to $\min F$. Moreover, if $(x_n)$ is a sequence in $X$ such that*

$$\lim_n F_n(x_n) = \lim_n (\inf_X F_n)$$

*then any limit point $x$ of $(x_n)$ satisfies $F(x) = \inf_X F$.*

Before dealing with our particular case, let us examine prototypical $\Gamma$-convergence results for families of the form that we consider (these are classical results, but we will use them several times).

**Lemma 2.1.11** (Zeroth order $\Gamma$-convergence)**.** *Assume that $f,g$ are functionals $X \to \mathbb{R}_+\cup\{\infty\}$ and that $g$ is lower bounded. As $\alpha \downarrow 0$, the family of functionals $F_\alpha := f + \alpha g$ admits the $\Gamma$-limit $F : x \mapsto \inf\{\liminf f(x_n)\ ;\ x_n \to x \text{ and } x_n \in \operatorname{dom} g\}$. In particular, $F(x) = f(x)$ if $f$ is l.s.c. and $x \in \operatorname{dom} g$ and $f(x) = \infty$ if $x \notin \overline{\operatorname{dom} g}$.*

*Proof.* Let $x \in X$ and $(\alpha_n) \downarrow 0$. For the $\Gamma$-limsup, either $F(x) = \infty$ and there is nothing to prove, or one needs to show that the inf defining $F$ is reached for a sequence $(x_n)$. To see this, let $S_x \subset \mathbb{R}$ be the set of which the infimum is taken. For all $n \in \mathbb{N}$, there exists $x_n \in \operatorname{dom} g$ satisfying $f(x_n) \leq \inf S_x + 1/n$ and $d(x_n,x) < 1/n$. One has $x_n \to x$ and $\lim f(x_n) = F(x)$. Moreover, one can always duplicate some indexes to slow down the possible blow up of $g(x_n)$, and obtain a new sequence sequence $(x_{k_n})$ such that $\alpha_n g(x_{k_n}) \to 0$. So $\lim F_n(x_{k_n}) = F(x)$. For the $\Gamma$-liminf, let $(x_n)$ be a sequence converging to $x$. Either $\liminf F_n(x_n) = \infty$ and there is nothing to prove, otherwise $\liminf F_n(x_n)$ is not changed by replacing $(x_n)$ by $(\tilde{x}_n)$ the sequence where terms not in $\operatorname{dom} g$ are replaced by the next term in $\operatorname{dom} g$. Since $g$ is lower-bounded, $\liminf \alpha_n g(\tilde{x}_n) \geq 0$, so $\liminf F_n(x_n) \geq \liminf f(\tilde{x}_n) \geq F(x)$. □

We deduce in particular that limit points of minimizers of $F_\alpha$ are minimizers of $F$, but it will turn out that in the cases we are interested by, this result is not very insightful because $F$ has a huge set of minimizers. In particular, in this $\Gamma$-limit, $g$ only intervenes through its domain. For the same class of problems, finer results are obtained by considering properly rescaled functionals which have exactly the same minimizers as $F_\alpha$ for all $\alpha > 0$.

**Lemma 2.1.12** (First order Γ-convergence)**.** *Assume that $f, g$ are l.s.c. functionals $X \mapsto \mathbb{R} \cup \{\infty\}$ such that* $\inf f > -\infty$*. As $\alpha \downarrow 0$, the family of functionals defined for $\alpha > 0$ by $F_\alpha := (f - \inf f)/\alpha + g$ admits the Γ-limit*

$$F : x \mapsto \begin{cases} g(x) & \text{if } x \in \arg\min_X f \\ \infty & \text{otherwise.} \end{cases}$$

*Proof.* Let $x \in X$ and $(\alpha_n) \downarrow 0$. For the Γ-lim sup, either $F(x) = \infty$ and there is nothing to prove, or $x \in \arg\min f$. In this case, the constant sequence $x_n = x$ satisfies $\limsup F_n(x_n) = g(x) = F(x)$. For the Γ-lim inf, let $x_n \to x$. If $x \notin \arg\min f$ then $\liminf f(x_n) \geq f(x) > \inf f$ since $f$ is l.s.c. so $\liminf F_n(x_n) = \infty$. Otherwise, one has $F_n(x_n) \geq g(x_n)$ so $\liminf F_n(x_n) \geq g(x)$ because $g$ is l.s.c. □

### Application to the convergence of geodesics

Let us consider two Lagrangians defined on $\mathbb{R}^d$ and $\mathbb{R}$ respectively

$$L_v : v \mapsto |v|^p, \qquad L_g : g \mapsto \left(\frac{1}{p}|g|\right)^p$$

and the associated functionals, depending implicitly on a pair $(\mu, \nu) \in \mathscr{M}_+(\Omega)^2$ of marginals

$$f_v : (\rho, \omega, \zeta) \mapsto \iota_{\mathrm{CE}} + \int_0^1 \int_\Omega \psi_{L_v}(\omega|\rho), \qquad f_g : (\rho, \omega, \zeta) \mapsto \iota_{\mathrm{CE}} + \int_0^1 \int_\Omega \psi_{L_g}(\zeta|\rho)$$

where $\iota_{\mathrm{CE}}$ is the convex indicator of the set $\mathrm{CE}_0^1(\mu, \nu)$ of solutions $(\rho, \omega, \zeta)$ to the continuity equation between $\mu$ and $\nu$. With these notations, the functional associated to the computation of $\widehat{W}_{p,\alpha}$ is

$$F_\alpha := f_v/\alpha^p + f_g.$$

Considering the result of Lemma 2.1.12, it is natural to introduce the following variational problems, which are studied in the next sections:

$$C_{g/v}(\mu, \nu) := \min\{f_g(\rho, \omega, \zeta) \,;\, (\rho, \omega, \zeta) \in \arg\min f_v\}. \tag{2.1.2}$$

$$C_{v/g}(\mu, \nu) := \min\{f_v(\rho, \omega, \zeta) \,;\, (\rho, \omega, \zeta) \in \arg\min f_g\}. \tag{2.1.3}$$

**Lemma 2.1.13** (Equicoercivity)**.** *Let $\mu, \nu \in \mathscr{M}_+(\Omega)$ and consider the set*

$$\mathscr{A} := \{(\rho, \omega, \zeta) \in \arg\min F_\alpha \,;\, \alpha \in ]0, \infty[\}.$$

*One has, for all $(\rho, \omega, \zeta) \in \mathscr{A}$,*

$$f_v(\rho, \omega, \zeta) \leq C_{v/g}(\mu, \nu), \qquad f_g(\rho, \omega, \zeta) \leq C_{g/v}(\mu, \nu).$$

*Moreover, $\mathscr{A}$ is uniformly bounded and thus weakly pre-compact.*

*Proof.* It is simple to see that if those bounds do not hold, this contradicts the optimality of an element in $\mathscr{A}$. Since $C_{v/g}(\mu,\nu)$ and $C_{g/v}(\mu,\nu)$ are always finite (see Proposition 2.1.17 and the proof of Theorem 2.1.21), it follows that the total variations are uniformly bounded by (a simplified version of) Proposition 1.1.21. □

**Theorem 2.1.14** (Convergence of $\widehat{W}_p$ geodesics)**.** *Let $(\mu,\nu)\in\mathscr{M}_+(\Omega)^2$, $p\geq 1$ and $(\rho_t^\alpha)_{t\in[0,1]}$ be the geodesic for $\widehat{W}_{p,\alpha}$ for $\alpha\in]0,\infty[$.*

- *as $\alpha\to 0$, $\rho^\alpha$ weakly converges to the unique minimizer of* (2.1.2)*;*
- *as $\alpha\to\infty$, $\rho^\alpha$ weakly converges (up to subsequences) to minimizers of* (2.1.3)*.*

*In other words, the geodesics for $\widehat{W}_{p,\alpha}$ interpolate between pure growth geodesics and generalized optimal transport interpolations. These interpolations are studied in the next sections.*

*Proof.* For all $\alpha>0$, the dynamic minimization problem defining $C_{L_{p,\alpha}}(\mu,\nu)$ has the same minimizers as

$$\tilde{F}_\alpha := (f_v-\inf f_v)/\alpha^p + f_g.$$

By Lemma 2.1.12, this family of functionals $\Gamma$-converges (relatively to the weak topology) as $\alpha\downarrow 0$ to $F := f_g+\iota_{\operatorname{arg\,min} f_v}$. Moreover, the closure $K$ of the set of minimizers from Lemma 2.1.13 is a compact such that $\inf F_\alpha = \inf_K F_\alpha$ for all $\alpha>0$. So by Theorem 2.1.10, minimizers for $F_\alpha$ weakly converge to minimizers of $F$ which is precisely (2.1.2). For the other result, define instead $\tilde{F}_\alpha := f_v+\alpha^p(f_g-\inf f_g)$ which $\Gamma$ converges as $\alpha\uparrow\infty$ to $F:=f_v+\iota_{\operatorname{argmin} f_g}$ and the same reasonning goes through. □

In the two next sections, we study in more depth these limit geodesics. The interpretation that will emerge is that $\widehat{W}_p$ behaves as optimal transport —up to rescaling and time reparametrization— at small scales and as a pure growth model at large scales.

### 2.1.3. Pure growth metrics

In this section, we introduce and characterize a class of pure growth metrics, that can be thought of as a 1-homogenization of the $L^p$ metrics. These growth metrics contain the total variation norm and the Hellinger metric as special cases, for $p=1$ and $p=2$.

**Definition 2.1.15** ($G_p$ metrics)**.** *Consider the function $h_p:(a,b)\mapsto|a^{1/p}-b^{1/p}|^p$ defined on $\mathbb{R}_+^2$. The pure growth $G_p$ metrics are defined for $p\geq 1$ by*

$$G_p(\mu,\nu) := \left(\int_\Omega h_p(\mu,\nu)\right)^{1/p} \text{“=”} \left(\int_\Omega |\mu^{1/p}-\nu^{1/p}|^p\right)^{1/p}.$$

The formula after the sign “=” is an abuse of notation but is easier to parse than the formula with an underlying dummy reference measure. One should keep in mind that these kind of expressions only make sense when the function being integrated is 1-homogeneous.

**Theorem 2.1.16.** *For $p \geq 1$, the space $(\mathscr{M}_+(\Omega), G_p)$ is a complete geodesic metric space, which metrizes strong convergence. A geodesic $(\rho_t)_{t\in[0,1]}$ between two measures $\mu$ and $\nu$ is given by the p-th power of the linear interpolation between the p-th roots:*

$$\rho_t = h_p((1-t)^p\mu, t^p\nu) \text{ “=” } ((1-t)\mu^{1/p} + t\nu^{1/p})^p.$$

*It is the unique geodesic if $p > 1$.*

*Proof.* The case $p = 1$ corresponds to the total variation metric restricted to the (closed) set of nonnegative measures and these results are well known so we may assume $p > 1$ in the rest of the proof. It is clear that $G_p$ is symmetric, nonnegative and that $G_p(\mu,\nu) = 0$ implies $\mu = \nu$. Moreover, for any third measure $\sigma \in \mathscr{M}_+(\Omega)$, by Minkowski's inequality (and with the abuse of notation mentioned above)

$$\begin{aligned} G_p(\mu,\nu) &= \left(\int_\Omega |\mu^{1/p} - \sigma^{1/p} + \sigma^{1/p} - \nu^{1/p}|^p\right)^{1/p} \\ &\leq \left(\int_\Omega |\mu^{1/p} - \sigma^{1/p}|^p\right)^{1/p} + \left(\int_\Omega |\nu^{1/p} - \sigma^{1/p}|^p\right)^{1/p} \\ &= G_p(\mu,\sigma) + G_p(\sigma,\nu) \end{aligned}$$

so the triangle inequality holds and $G_p$ is a metric. Consider now the equality case. Denoting $t := G_p(\mu,\sigma)/G_p(\mu,\nu)$ and assuming that $t \in ]0,1[$ is well defined, equality holds if and only if $(t-1)(\mu^{1/p} - \sigma^{1/p}) + t(\sigma^{1/p} - \nu^{1/p}) = 0$ (we identify $\mu,\nu$ to their density w.r.t. a dominating reference measure). This implies that $\sigma = \rho_t$ with the definition of $(\rho_t)$ given in the theorem so geodesics are uniquely characterized.

For the completeness property, let $\mu_n$ be a Cauchy sequence. In particular, its mass is bounded and has a subsequence that converges to $\mu \in \mathscr{M}_+(\Omega)$. Since $h_p$ is l.s.c. and sublinear by Lemma 2.1.7, general theorems on integral functionals (Appendix A) state that $G_p$ is weakly l.s.c. so it follows that for all $m \in \mathbb{N}$, $G_p(\mu_m,\mu) \leq \liminf_{n\to\infty} G_p(\mu_n,\mu_m)$. Taking the limit as $m \to 0$ and using the Cauchy property, one has $\lim_{m\to\infty} G_p(\mu_m,\mu) \leq \limsup_{m,n\to\infty} G_p(\mu_m,\mu_n) = 0$ so $(\mu_n)$ converges for the metric $G_p$.

To show that convergence in $G_p$ is equivalent to strong convergence, it is sufficient to remark that $h_p(a_n,a) \to 0$ if and only if $a_n \to a$. From this we deduce that $G_p(\mu_n,\mu) \to 0$ if and only if $\mu_n(A) \to \mu(A)$ for all Borel sets $A \subset \Omega$. □

The following proposition shows that $G_p$ corresponds indeed to the dynamic problem that is obtained in the Γ-limit of $\alpha \downarrow 0$.

**Proposition 2.1.17** (Dynamic formulation of $G_p$)**.** *One has for all $(\mu,\nu) \in \mathscr{M}_+(\Omega)$ and $p > 1$,*

$$G_p(\mu,\nu)^p = C_{g/v}(\mu,\nu).$$

*and for the unique minimizer $(\rho_t, 0, \zeta_t)$ of (2.1.2) it holds that $\rho_t$ is the geodesic between $(\mu,\nu)$ for $G_p$.*

*Proof.* Consider $\mu,\nu \in \mathscr{M}_+(\Omega)$ of positive mass (the other cases are treated in Proposition 2.1.8), $(\rho,\zeta)$ and remark that a minimizer for $C_{g/v}$ exists by standard compactness and lower-semicontinuity arguments (bounds on the mass can be obtained as in Proposition 1.1.21). Remembering Lemma 1.1.23, we known that for any $\varepsilon > 0$, it holds $\int_{-\varepsilon/2}^{1+\varepsilon/2}\int_{\Omega^\varepsilon} \psi_{L_g}(\zeta^\varepsilon|\rho^\varepsilon) \leq C_{g/v}(\mu,\nu)$ , where $\rho^\varepsilon,\zeta^\varepsilon$ are the convolutions of $\rho,\zeta$ with a mollifier supported on $]-\varepsilon/2,\varepsilon/2[\times B^d(0,\varepsilon)$.

Now remark that $\partial_t\rho^\varepsilon = \zeta^\varepsilon$ so, identifying $\rho^\varepsilon$ to its density and making the change of variable $r_t^\varepsilon = (\rho_t^\varepsilon)^{1/p}$, we get (refer to Remark 2.1.6 for more details):

$$C_{g/v}(\mu,\nu) \geq \int_{-\varepsilon/2}^{1+\varepsilon/2}\int_{\Omega^\varepsilon} |\partial r_t^\varepsilon(x)|^p \mathrm{d}x\mathrm{d}t \geq (1+\varepsilon)^{p-1} G_p(\mu^\varepsilon,\nu^\varepsilon)^p$$

by convexity and time rescaling. Passing to the limit $\varepsilon \to 0$ and using the lower-semicontinuity of $G_p$, we get $C_{g/v}(\mu,\nu) \geq G_p(\mu,\nu)^p$. For the other bound, it is direct because the geodesic for $G^p$ given in Theorem 2.1.16, when plugged as a candidate in the functional of the minimization problem (2.1.2) with $\zeta$ its weak* derivative, gives exactly $G_p(\mu,\nu)^p$. Uniqueness of geodesics for $G_p$ implies uniqueness for the minimizers of $C_{g/v}(\mu,\nu)$ too since minimizers of the latter are constant speed geodesics (it is proved just as in Proposition 1.1.19). □

**Remark 2.1.18** (Continuity equation with zero transport and Banach derivative)**.** *If $(\rho_t)_{t\in[0,1]}$ is a weakly continuous solution to the continuity equation with source with zero momentum $(\rho_t,0,\zeta_t)_{t\in[0,1]}$ then the path $(\rho_t)$ is absolutely continuous for the total variation metric and $\zeta_t$ corresponds to its weak* differential [5, Rmk. 1.1.3].*

**Remark 2.1.19** (Special cases)**.**

***Total variation*** *For $p = 1$, the distance $G_1$ is the total variation distance, which is the distance associated to the dual Banach space norm on $\mathscr{M}(\Omega)$.*

***Hellinger/Fisher-Rao*** *For $p = 2$, the distance $G_2$ is the Hellinger (a.k.a. Kakutani) distance on $\mathscr{M}_+(\Omega)$. In that case, the dynamic action of a path $(\rho_t)_{t\in[0,1]}$ is the integral in time of*

$$\int_\Omega \left(\frac{\mathrm{d}\dot\rho_t}{\mathrm{d}\rho_t}\right)^2 \mathrm{d}\rho_t = \int_\Omega (\partial_t \log\rho_t)^2 \mathrm{d}\rho_t,$$

*assuming that $\dot\rho_t \ll \rho_t$ and that $\rho_t$ is a positively lower bounded density. The action can thus be formally interpreted as a Riemannian length associated to the metric tensor $\langle\dot\mu_1,\dot\mu_2\rangle_\mu = \int \left(\frac{\mathrm{d}\dot\mu_1}{\mathrm{d}\mu}\right)\left(\frac{\mathrm{d}\dot\mu_2}{\mathrm{d}\mu}\right)\mathrm{d}\mu$. This metric tensor is known as the* Fisher-Rao *metric when restricted to tangent vectors $\mu_i$ of zero total mass. In the setting of parametric probabilities, i.e. when the densities are constrained to lie in a finite dimensional subspace of probability measures, this metric tensor, written in coordinates, is called the* Fisher information matrix.

***Limit case*** *Taking the limit $p \to \infty$ one has for all $(\mu,\nu) \in \mathscr{M}_+(\Omega)$*

$$G_\infty(\mu,\nu) \coloneqq \lim_{p\to\infty} G_p(\mu,\nu)^p = \mu_\nu^\perp(\Omega) + \nu_\mu^\perp(\Omega)$$

*where $\mu_\nu^\perp$ and $\nu_\mu^\perp$ are the singular parts appearing in the Lebesgue decomposition of $\mu$ w.r.t. $\nu$ and of $\nu$ w.r.t. $\mu$, respectively. This follows from the fact that $|1-\beta^{1/p}|^p \to 0$ if $\beta > 0$ or $1$ if $\beta = 0$. In general, $G_\infty$ is not a metric, but it boils down to the* symmetric difference pseudo-metric *when restricted to measures which are indicators of Borel sets (w.r.t. to a reference measure).*

### 2.1.4. Generalized $W_p$ optimal transport interpolation

In the opposite limit, we do not recover a metric, but a generalized notion of optimal transport interpolation with a prescribed , constant in space, growth field.

**Definition 2.1.20** (generalized $W_p$ interpolation)**.** *Let $\mu, \nu \in \mathscr{M}_+(\Omega)$. We define the generalized $W_p$ interpolation $(\rho_t)_{t\in[0,1]}$ as follows.*

- *if $\mu(\Omega)$ or $\nu(\Omega) = 0$ then $(\rho_t)$ is the $G_p$-pure growth geodesic between $\mu, \nu$;*
- *if $\mu(\Omega) = \nu(\Omega) \neq 0$ then $(\rho_t)$ is a $W_p$-optimal transport geodesic between $\mu, \nu$;*
- *otherwise, it is defined as $\rho_t = \beta_t \hat{\rho}_{\mathrm{s}(t)}$ where $(\hat{\rho}_t)_{t\in[0,1]}$ is a $W_p$-geodesic between $\mu$ and $\frac{\mu(\Omega}{\nu(\Omega}\nu$, the scalar function $\beta_t := [(1-t) + t(\nu(\Omega)/\mu(\Omega))^{1/p}]^p$ is a time dependent mass rescaling and $\mathrm{s} : [0,1] \to [0,1]$ is the time reparametrization satisfying $(s(0), s(1)) = (0,1)$ and $\mathrm{s}'(t) = \mathrm{cst} \cdot \beta_t^{1/(p-1)}$.*

This definition is reminiscent of the models studied in [105] where optimal transport interpolations with prescribed rates of growth are considered for image interpolation. The next theorem concludes the characterization of the limit geodesics, treating the case $\alpha \uparrow \infty$.

**Theorem 2.1.21.** *The minimizers to* (2.1.3) *are exactly the set of triplets $(\rho_t, \omega_t, g_t\rho_t)$ where $\rho_t$ is a generalized $W_p$ interpolation and $g_t$ a growth field which is constant in space.*

*Proof.* The case when both measures or one of the measures is zero is treated in Proposition 2.1.8. Also when $\mu(\Omega) = \nu(\Omega)$, minimizers of $f_g$ are characterized by $\zeta = 0$, so (2.1.3) boils down to a standard $W_p$-optimal transport problem. For the remaining cases, we assume $0 < \mu(\Omega), \nu(\Omega)$. Our first step is to characterize minimizers of $f_g$, the second one is to characterize the minimizers of $f_v$ in this subset.

**Step 1.** Let $(\rho, \omega, \zeta) \in \mathrm{CE}_0^1(\mu, \nu)$. One has, similarly as in Proposition 2.1.8

$$f_g(\rho, \omega, \zeta) = \int_0^1 \int_\Omega \psi_{L_g}(\zeta|\rho) \geq \int_0^1 \psi_{L_g}(\zeta(\Omega)|\rho(\Omega)) \geq |\nu(\Omega)^{1/p} - \mu(\Omega)^{1/p}|^p$$

with equality if and only if $\omega = 0$ and $g_t = \mathrm{d}\zeta_t/\mathrm{d}\rho_t$ is constant $\rho_t$ a.e. in $\Omega$ for the first one and $\rho_t(\Omega) = ((1-t)\mu(\Omega)^{1/p} + t\nu(\Omega)^{1/p})^p$ for the second one, by Lemma 2.1.7. So the minimizers of $L_g$ are characterized by $\zeta_t = g_t\rho_t$ where $g_t \in \mathbb{R}$ is determined by the initial and final total masses.

**Step 2.** Let $\beta_t := \mu(\Omega)/\rho_t(\Omega)$ and let $\mathrm{s} : [0,1] \to [0,1]$ be such that $(\mathrm{s}(0), \mathrm{s}(1)) = (0,1)$ and $\kappa \cdot (\mathrm{s}'(t))^{p-1} \cdot = \beta_t$ for some constant $\kappa > 0$. We define, on $\operatorname{argmin} f_g \cap \operatorname{dom} f_v$ (on this set the measures admit a disintegration in time w.r.t. Lebesgue), the mass rescaling

$$R : (\rho_t, \omega_t, \zeta_t)_{t\in[0,1]} \mapsto (\beta_t \cdot \rho_t, \beta_t, \cdot \omega_t)_{t\in[0,1]}.$$

It is such that $R(\rho_t, \omega_t, g_t\rho_t)$ solves the continuity equation without source between $\mu$ and $\frac{\mu(\Omega)}{\nu(\Omega)}\nu$. We also define the time reparametrization $\mathrm{t} := \mathrm{s}^{-1}$ and

$$T : (\rho_t, \zeta_t)_{t\in[0,1]} \mapsto (\rho \circ \mathrm{t}(s), \mathrm{t}'(s)\omega \circ \mathrm{t}(s))_{s\in[0,1]}$$

which preserves solutions to the continuity equation (Proposition 1.1.7). One has, for $(\rho, \omega, \zeta) := (\rho_t, \omega_t, g_t\rho_t)_{t\in[0,1]} \in \operatorname{argmin} f_g$

$$\begin{aligned}
f_v(T \circ R(\rho, \omega, \zeta)) &= \int_0^1 \left( \int_\Omega \psi_{L_v}(\mathrm{t}'(s) \cdot \beta_{\mathrm{t}(s)}\omega_{t(s)} | \beta_{\mathrm{t}(s)}\rho_{t(s)}) \right) \mathrm{d}s \\
&= \int_0^1 \mathrm{t}'(s)^p \left( \int_\Omega \psi_{L_v}(\beta_{\mathrm{t}(s)}\omega_{t(s)} | \beta_{\mathrm{t}(s)}\rho_{t(s)}) \right) \mathrm{d}s \\
&= \int_0^1 \frac{1}{\mathrm{s}'(t)^{p-1}} \left( \int_\Omega \psi_{L_v}(\beta_t\omega_t | \beta_t\rho_t, ) \right) \mathrm{d}t \\
&= \kappa \int_0^1 \int_\Omega \psi_{L_v}(\omega_t | \rho_t) \mathrm{d}t = \kappa \cdot f_v(\rho, \omega, \zeta).
\end{aligned}$$

If follows that $T \circ R$ maps $\operatorname{argmin} f_g$ to solutions of the continuity equation without source while preserving (up to a factor) the value of $f_v$. Thus $(\rho, \omega, \zeta)$ minimizes (2.1.3) if and only if $T \circ R(\rho, \omega, \zeta)$ is a $W_p$-optimal transport minimizer. Equivalently, if $(\rho, \omega)$ is a $W_p$ geodesic, then

$$R^{-1} \circ T^{-1}(\rho, \omega) = (\rho_{\mathrm{s}(t)}/\beta_t, \mathrm{s}'(t)\omega_{\mathrm{s}(t)}/\beta_t)_{t\in[0,1]}$$

is a minimizer for (2.1.3) and we recover the expression for the generalized optimal transport interpolation from Definition 2.1.20. □

## 2.2. Quadratic case

### 2.2.1. Dynamic and coupling formulations

Within the previous class of metrics, let us consider now the quadratic case $p = 2$. As the effect of the scale parameter has been studied in the previous section, we now fix $\alpha = 1$. In this case the Lagrangian is, as a function of the velocity and the growth,

$$L_2(v,g) = |v|^2 + \frac{1}{4}g^2.$$

Since the Lagrangian is superlinear, one can equivalently write the definition of $C_{L_2}$ from Definition 1.1.13 in terms of velocity and growth fields, for $\mu, \nu \in \mathscr{M}_+(\Omega)$,

$$C_{L_2}(\mu,\nu) = \min\left\{ \int_0^1 \int_\Omega \left( |v_t(x)|^2 + \frac{1}{4}g_t(x)^2 \right) \mathrm{d}\rho_t \mathrm{d}t \;;\; \partial_t \rho_t + \nabla\cdot(v_t\rho_t) = g_t\rho_t \right\}$$

where the minimum is over weakly continuous paths $(\rho_t)_{t\in[0,1]}$ such that $(\rho_0,\rho_1) = (\mu,\nu)$ and such that the continuity equation with $v_t \in L^2(\rho_t;\mathbb{R}^d)$, $g_t \in L^2(\rho_t;\mathbb{R})$ is satisfied in the weak sense. Let us specialize Definition 2.1.1 to this case.

**Definition 2.2.1.** *For $\mu,\nu \in \mathscr{M}+(\Omega)$, we define $\widehat{W}_2(\mu,\nu) := C_{L_2}(\mu,\nu)^{1/2}$. By Theorem 2.1.2, the space $(\mathscr{M}_+(\Omega),\widehat{W}_2)$ is a complete geodesic metric space where convergence is equivalent to weak convergence.*

This metric has been introduced several times in the literature during the time span of the preparation of this thesis [103, 104, 94, 40] and independently by us [44, 43]. It was attributed the name *Hellinger-Kantorovich* in [103, 104] and *Wasserstein-Fisher-Rao* in [44]. In both cases, the two terms refer to the limit models (see Section 2.1.2), or the two "metric tensors" of pure growth[1] and pure transport[2], that are inf-convoluted to generate $\widehat{W}_2$. One of the great advantages of this case is that the minimal path cost (Definition 1.3.1) has an explicit expression, which leads to an explicit optimal coupling formulation.

**Theorem 2.2.2** ($\widehat{W}_2$ as a semi-coupling)**.** *One has the alternative semi-coupling formulation and its dual*

$$\begin{aligned}\widehat{W}_2^2(\mu,\nu) &= \min_{\gamma_1,\gamma_2\in\mathscr{M}_+(\Omega^2)} \left\{ \int_{\Omega^2} h^{c_{L_2}}_{x,y}(\gamma_1,\gamma_2) \;;\; (\pi^1_\#\gamma_1, \pi^2_\#\gamma_2) = (\mu,\nu) \right\}, \\ &= \sup_{\substack{\phi,\psi\in\mathscr{C}(\Omega)\\ \phi,\psi\le 1}} \left\{ \int_\Omega \phi\,\mathrm{d}\mu + \int_\Omega \psi\,\mathrm{d}\nu \;;\; (1-\phi(x))(1-\psi(y)) \ge \cos_+^2(\mathrm{dist}(x,y)) \right\}\end{aligned}$$

*where the sublinear cost function $h^{c_{L_2}}_{x,y}$ is defined in Lemma 2.2.3 and the meaning of a sublinear function of measures is in the section* Notation*. Also, the notation* $\cos_+$ *stands for $d \mapsto \cos(\min\{d,\pi/2\})$ and the constraint in the dual problem is understood for all $(x,y) \in \Omega^2$.*

[1] The Hellinger distance is the distance induced by the Fisher-Rao metric tensor with the probability constraint removed, see Remark 2.1.19.

[2] Wasserstein or Monge-Kantorovich are two alternative names for the quadratic optimal transport distance. The latter is historically more accurate but less used.

*Proof.* This is an application of Theorem 1.3.3 on the equivalence between dynamic and coupling problems. The derivation of the minimal path cost and its convex regularization is performed in Lemma 2.2.3, while the dual to the sublinear cost is obtained by direct computations and the dual formulation follows from 1.2.4. □

**Lemma 2.2.3** (Quadratic path based cost)**.** *The minimal path cost for the Lagrangian $L_2$ is*

$$c_{L_2}((x_1,u_1),(x_2,u_2)) = u_1+u_2-2\sqrt{u_1}\sqrt{u_2}\cos(\min\{\mathrm{dist}(x_1,x_2),\pi\}).$$

*and its convex regularization is, for each $(x_1,x_2)\in\Omega^2$,*

$$h^{c_{L_2}}_{x_1,x_2}(u_1,u_2) = u_1+u_2-2\sqrt{u_1}\sqrt{u_2}\cos_+(\mathrm{dist}(x_1,x_2)).$$

*where* $\cos_+(d)=\cos(\min\{d,\pi/2\})$.

*Proof.* Consider $(x_i,u_i)\in\Omega\times\mathbb{R}_+$, for $i\in\{0,1\}$ and an absolutely continuous path $(x(t),u(t))$ joining those points. The associated cost is

$$\int_0^1 \left(|\dot{x}(t)|^2+\frac{1}{4}\left|\frac{\dot{u}(t)}{u(t)}\right|\right)u(t)\mathrm{d}t = \int_0^1 \left(r(t)^2|\dot{x}(t)|^2+|\dot{r}(t)|^2\right)\mathrm{d}t$$

after the change of variables $r(t)=\sqrt{u(t)}$, which preserves absolute continuity. First, remark that the path is minimizing only if $x(t)$ is a parametrization of a geodesic between $x_0$ and $x_1$ (otherwise, the cost could be improved by "reproducing" the same mass variations $r(t)$ on a geodesic, with a lesser cost) so we may as well write $x(t)=e_{\theta(t)}(x_0,x_1)$ where $t\mapsto e_t(x,y)$ is the unit speed parametrization of a geodesic in $\Omega$ and $\theta(t)\in\mathbb{R}$ our new unknown, that satisfies $|\dot{x}(t)|=|\dot{\theta}(t)|$. Also, it is judicious to interpret $(r,\theta)$ as polar coordinates since the integrand is equal to the norm of the velocity vector in this case. As long as $\theta(t)\in[0,2\pi[$, we may define $z(t):=r(t)\exp(i\theta(t))$ and there is a unique minimizing path in this basis which is a straight line $z(t)=(1-t)z(0)+tz(1)$ with the associated cost $|z(1)-z(0)|^2$ and $(z(0),z(1))=(\sqrt{u}_0\exp i\theta_0,\sqrt{u}_1\exp i\theta_1)$. This trajectory is not valid when $\theta_1-\theta_0>\pi$ since then, $\arg z(t)\notin[\theta_0,\theta_1]$. Otherwise, noticing that $(\theta_1-\theta_0)=\mathrm{dist}(x_0,x_1)$ (the distance dist is the geodesic distance in $\Omega$), this gives the cost

$$|\sqrt{u}_1\exp(i\,\mathrm{dist}(x_0,x_1))-\sqrt{u}_0|^2 = u_1+u_0-2\sqrt{u_0}\sqrt{u_1}\cos(\mathrm{dist}(x_0,x_1)).$$

But remember that in the definition of the minimum-path cost, "teleporting" through the apex is admissible. The shortest path doing so is defined by

$$(x(t),u(t)) = \begin{cases}(x_0,(1-2t)^2u_0) & \text{if } t\in[0,1/2]\\ (x_1,(2t-1)^2u_1) & \text{if } t\in[1/2,1]\end{cases}$$

with a corresponding cost $(\sqrt{u}_0+\sqrt{u}_1)^2$ which is the limit of the previous formula as $\mathrm{dist}(x_0,x_1)\to\pi$. For the convex regularization of $c_L$, when the argument of the cosine is smaller or equal than $\pi/2$, the function $c_{L_2}(x_1,\cdot,x_2,\cdot)$ is already sublinear (it is obviously 1-homogeneous, and $(u_1,u_2)\mapsto\sqrt{u_1u_2}$ is jointly concave). Otherwise, one may use the characterization of Proposition 1.3.1 and decompose the masses as $u_1+0$ and $0+u_2$ which gives the cost $u_1+u_2$ and that can be shown to be the unique optimal decomposition for $\mathrm{dist}(x_1,x_2)>\pi/2$. The case $\mathrm{dist}(x_1,x_2)=\pi/2$ gives the cost $u_1+u_2$ but can be obtained by any decomposition of the masses in an arbitrary number of chunks. □

### 2.2.2. Optimal entropy-transport formulations

Following the alternative formulations introduced by [103] (see Section 1.2.3) one has also an optimal lift, and an optimal entropy-transport formulations for $\widehat{W}_2$. The optimal lift formulation holds for any cost on $\Omega \times \mathbb{R}_+$ which sublinear regularization is $h^{c_{L_2}}$ and, in particular, for $c_{L_2}$ (see Theorem 1.2.13). For the optimal entropy-transport formulation, it can be obtained by a change of variables to linearize the constraints in Theorem 2.2.2.

**Theorem 2.2.4** ($\widehat{W}_2$ as an entropy-transport problem [103])**.**

$$\widehat{W}_2^2(\mu,\nu) = \min_{\gamma \in \mathscr{M}_+(\Omega^2)} \left\{ \int_{\Omega^2} c_\ell \mathrm{d}\gamma + H(\pi^1_\#\gamma|\mu) + H(\pi^2_\#\gamma|\nu) \right\},$$

$$= \sup_{\alpha,\beta \in \mathscr{C}(\Omega)} \left\{ \int_\Omega (1-e^\alpha)\mathrm{d}\mu + \int_\Omega (1-e^\beta)\mathrm{d}\nu \; ; \; \alpha(x)+\beta(y) \le c_\ell(x,y) \right\}$$

*where* $c_\ell(x,y) = -\log\cos^2_+(\mathrm{dist}(x,y))$ *and* $H(\sigma|\mu) = \int \psi_f(\sigma|\mu)$ *is the relative entropy, the* $f$*-divergence associated to the entropy* $f(s) = s\log s - s + 1$ *(see Definition 3.5.1).*

*Proof.* This theorem, as well as optimality conditions, is stated and proved in a very general setting in [103] so we will only give a sketch of a simple proof. In the dual of the semi-coupling formulation of Theorem 2.2.2, make the change of variables $(\alpha,\beta) = (-\log(1-\phi), -\log(1-\psi))$. This gives the supremum problem of the present theorem. Then one may apply Fenchel-Rockafellar duality (Theorem A.0.7) to obtain the optimal entropy transport formulation, with the qualification constraint easily checked since $c_\ell$ is nonnegative and the dual objective is continuous. One may also notice as in [103] that the marginal perspective cost (Definition 1.2.20) with respect to $c_\ell$ and the relative entropy is $h^{c_{L_2}}$. In other words, for $(x,y) \in \Omega^2$ and $a,b \ge 0$, one has by direct computations

$$h^{c_{L_2}}_{x,y}(a,b) = \min_{\theta \ge 0} \; \theta \cdot c_\ell(x,y) + \psi_f(\theta|a) + \psi_f(\theta|b).$$

We then apply Theorem 1.2.21 (taken from [103]). □

In the entropy-transport formulation, the differentiable structure of $\mathbb{R}^d$ plays no role so it can as well be defined on more general metric spaces. It is shown in [103] that if $X$ is a geodesic space, then $\widehat{W}_2$ is also geodesic. It is also shown that one may replace the cost $c_\ell$ by the quadratic cost and the optimal entropy transport problem still defines a distance, that they call the *Gaussian-Hellinger-Kantorovich* distance and the corresponding geodesic distance is $\widehat{W}_2$. In the same line, we can propose the following generalization which allows weights on the relative entropy terms.

**Proposition 2.2.5** (A variant with weights)**.** *If* $(X,d)$ *is a compact metric space, the following variant of* $\widehat{W}_2$ *defined by*

$$\widehat{W}^2_{2,\lambda}(\mu,\nu) \coloneqq \min_{\gamma \in \mathscr{M}_+(X^2)} \left\{ \int_{X^2} c_\ell \mathrm{d}\gamma + \lambda H(\pi^1_\#\gamma|\mu) + \lambda H(\pi^2_\#\gamma|\nu) \right\}$$

*defines a distance if* $0 < \lambda \le 1$ *and generally not if* $\lambda > 1$.

*Proof.* The case $\lambda = 0$ is trivial (the "distance" is equally null). If $\lambda > 0$, when dividing the objective functional by $\lambda$, one obtains the new cost $-2\log(\cos_+^{1/\lambda}(\mathrm{dist}(x,y)))$. But the function $f : d \mapsto \arccos(\cos^{\frac{1}{\lambda}}(d))$ defined on $[0,\pi/2]$ is increasing, positive, satisfies $\{0\} = f^{-1}(0)$ and for $x \in ]0,\pi/2[$ it holds

$$f'' = \frac{1}{\lambda}\frac{\cos^{\frac{1}{\lambda}-2}}{(1-\cos^{\frac{2}{\lambda}})^2}\left(1 - \frac{1}{\lambda}\sin^2 - \cos^{\frac{2}{\lambda}}\right).$$

From the convexity inequality $\mathrm{sign}[(1-X)^{\frac{1}{\lambda}} - 1 + \frac{1}{\lambda}X] = \mathrm{sign}(\frac{1}{\lambda} - 1)$ it follows that $f$ is strictly concave on $[0,\pi/2]$ if $\lambda < 1$ and strictly convex if $\lambda > 1$. Thus if $\lambda \leq 1$, $f \circ d$ still defines a distance on $X$ and consequently $\widehat{W}^2_{2,\lambda}(\mu,\nu)$ too. The strict convexity if $\lambda > 1$ implies that the distance property is generally not true in that case (it is always false if $X$ is a geodesic space of diameter greater than $\pi/2$ as we proved in [45]). □

Other useful properties of the optimal entropy-transport problems in general and of the metric $\widehat{W}_2$ in particular, are proved in Chapter 6 following our specific needs.

## 2.3. Bounded Lipschitz and optimal partial transport

The last section of this chapter focuses on the case $p = 1$ and, more generally, when the Lagrangian is 1-homogeneous w.r.t. the growth variable. We recover in particular the "generalized Wasserstein distances" introduced in [129, 130] and the optimal partial transport problem [32, 69], which we prove to be equivalent.

### 2.3.1. Recovering the Bounded Lipschitz metric

In the case $p = 1$, one has the Lagrangian

$$L_1(v,g) = |v| + |g|.$$

Since this Lagrangian is already sublinear, its perspective function does not depend on the perspective parameter. It follows that the definition of $C_{L_1}$ can be adapted from Definition 1.1.13 as

$$C_{L_1}(\mu,\nu) = \min\left\{(|\omega| + |\zeta|)([0,1]\times\Omega)\ ;\ (\rho,\omega,\zeta) \in \mathrm{CE}_0^1(\mu,\nu)\right\}$$

where $\mathrm{CE}_0^1(\mu,\nu)$ is the set of distributional solutions to the continuity equation with source (Definition 1.1.1).

**Definition 2.3.1.** *For $\mu,\nu \in \mathscr{M}_+(\Omega)$, we define $\widehat{W}_1(\mu,\nu) := C_{L_1}(\mu,\nu)$. By Theorem 2.1.2, the space $(\mathscr{M}_+(\Omega),\widehat{W}_1)$ is a complete geodesic metric space, and $\widehat{W}_1$ metrizes weak convergence.*

One also may recall Proposition 1.1.20 from Chapter 1 and give an alternative static formulation for $\widehat{W}_1$.

**Proposition 2.3.2** (Static formulation of $\widehat{W}_1$)**.** *One has alternatively, for $(\mu,\nu) \in \mathscr{M} + (\Omega)$,*

$$\begin{aligned}\widehat{W}_1(\mu,\nu) &= \min_{\substack{\omega\in\mathscr{M}_+(\Omega;\mathbb{R}^d)\\ \zeta\in\mathscr{M}_+(\Omega;\mathbb{R})}} \{(|\omega| + |\zeta|)(\Omega)\ ;\ \nu - \mu = \zeta - \nabla\cdot\omega\}\\ &= \sup_{\varphi\in\mathscr{C}^1(\Omega)} \left\{\int \varphi\,\mathrm{d}(\nu-\mu)\ ;\ \|\nabla\varphi\|_\infty \le 1,\ \|\varphi\|_\infty \le 1\right\}.\end{aligned}$$

*where the equation $\nu - \mu = \zeta - \nabla\cdot\omega$ is understood in the distributional sense, with no-flux boundary conditions on $\partial\Omega$ for $\omega$. In particular, $\widehat{W}_1$ is the well-known Bounded-Lipschitz distance on $\mathscr{M}_+(\Omega)$.*

*Proof.* It is just an application of Proposition 1.1.20 with the explicit formulae for the conjugate of $L$, $L^*(v^*,g^*) = 0$ if $|v^*| \le 1$ and $|g^*| \le 1$ and $\infty$ otherwise. □

The static problem we recover is the Bounded Lipschitz metric, a classical object that is known, for instance, to extend as a norm on the whole space of signed measures [83, 80].

### 2.3.2. Optimal partial transport

In this section, we consider Lagrangians which are 1-homogeneous w.r.t. to the growth variable and $p$-homogeneous w.r.t. the velocity, for some $p > 1$ and $\alpha > 0$:

$$L(v,g) = |v/\alpha|^p + \frac{1}{2}|g|. \tag{2.3.1}$$

This is a different Lagrangian than that of Section 2.1, but the parameters $p, \alpha$ have similar interpretations. The factor $\frac{1}{2}$ is introduced to equalize the cut locus distance to $\alpha$ whatever the value of $p$. Let us compute the minimal path cost as defined in Definition (1.3.1).

**Proposition 2.3.3** (Minimal path cost)**.** *Let $(x_0,u_0)$ and $(x_1,u_1)$ be points in $\Omega \times \mathbb{R}_+$. The minimal path cost associated to the Lagrangian* (2.3.1) *is*

$$c_L((x_0,u_0),(x_1,u_1)) = |u_1-u_0|/2 + \min\{u_0,u_1\} \cdot \min\{\mathrm{dist}(x_0,x_1)/\alpha, 1\}^p. \tag{2.3.2}$$

*This cost is already l.s.c. and sublinear in $(u_0,u_1)$ for all $(x_0,x_1) \in \Omega^2$, meaning that $h^{c_L}_{x_0,x_1} = c_L((x_0,\cdot),(x_1,\cdot))$.*

*Proof.* For any absolutely continuous path $(x(t),u(t))_{t\in[0,1]}$, its action is

$$\int_0^1 |x'(t)/\alpha|^p u(t)\mathrm{d}t + \frac{1}{2}\int_0^1 |u'(t)|\mathrm{d}t \geq \bar{u}\int_0^1 |x'(t)/\alpha|^p \mathrm{d}t + \frac{1}{2}(|u_0-\bar{u}| + |u_1-\bar{u}|)$$

where $\bar{u}$ denotes the minimum mass $\bar{u} := \min_{t\in[0,1]} u(t)$. Denoting $d := \mathrm{dist}(x_0,x_1)$ so that $\int_0^1 |x'(t)|\mathrm{d}t \geq d^p$, we may minimize the right hand side and obtain, using first order conditions, that at optimality $\bar{u} = \min\{u_0,u_1\}$ if $d < \alpha$ and $\bar{u} = 0$ if $d > \alpha$ (any $\bar{u}$ in the interval is a minimizer for $d = \alpha$). The corresponding lower bound for the path cost is the right hand side of (2.3.2). It remains to show that this cost can be approached arbitrarily close by absolutely continuous paths as in Definition (1.3.1). If $d \geq \alpha$, any path joining the apex without transport works and exactly reaches $c$. For instance, let $(x(t),u(t)) = (x_0,(1-2t)u_0)$ for $t \in [0,1/2]$ and $(x(t),u(t)) = (x_1,(1-2t)u_1)$ for $t \in [1/2,1]$. If $d < \alpha$, let $t \mapsto e_t(x,y)$ be the constant speed parametrization of a geodesic joining $(x_0,x_1)$ and assume that $u_0 \leq u_1$. Then define $(x(t),u(t)) = (e_{t/(1-\varepsilon)}(x_0,x_1),u_0)$ for $t \in [0,1-\varepsilon]$ and $(x(t),u(t)) = (x_1,u_0+(u_1-u_0)\cdot(t+\varepsilon-1)/\varepsilon)$. The cost associated to this absolutely continuous path is $u_0\cdot(d/\alpha)^p/(1-\varepsilon)^{1-p} + (u_1-u_0)/2$ which can be made arbitrarily close to $c$. □

From this explicit minimal path cost, we obtain a coupling formulation for $C_L$. We will then prove that it corresponds to the problem of optimal partial transport.

**Theorem 2.3.4** (Recovering optimal partial transport)**.** *For $(\mu,\nu) \in \mathscr{M}_+(\Omega)$, and considering the Lagrangian $L$ from* (2.3.1)*, one has*

$$C_L(\mu,\nu) = \frac{1}{2}(\mu(\Omega)+\nu(\Omega)) + \min\left\{\int_{\Omega^2}(c_{p,\alpha}(x,y)-1)\mathrm{d}\gamma(x,y)\ ;\ \gamma \in \Pi_{\leq}(\mu,\nu)\right\}$$

*where $c_{p,\alpha}(x,y) = (\mathrm{dist}(x,y)/\alpha)^p$ and $\Pi_{\leq}(\mu,\nu) = \{\gamma \in \mathscr{M}_+(\Omega^2)\ ;\ (\pi^1_\#\gamma,\pi^2_\#\gamma) \leq (\mu,\nu)\}$.*

*Proof.* Just for this proof, let us denote by $\tilde{C}$ the right hand side term. By Theorem 1.3.3, $C_L$ admits a semi-coupling formulation involving the sublinear cost from Proposition 2.3.2. Let us consider optimal semi-couplings $(f_1\gamma, f_2\gamma)$ where $\gamma \in \mathscr{P}(\Omega^2)$ and $f_1, f_2 \in L^1(\gamma)$. Let $\bar{\gamma} := \min\{f_1, f_2\}\bar{\gamma}|_{\mathscr{D}}$ where $\mathscr{D} := \{(x,y) \in \Omega^2 \,;\, \mathrm{dist}(x,y) \leq \alpha\}$. It holds $\bar{\gamma} \in \Pi_{\leq}(\mu,\nu)$ and

$$\begin{aligned} C_L(\mu,\nu) &= \int_{\Omega^2} h_{x_1,x_2}(f_1(x_1,x_2), f_2(x_1,x_2))\mathrm{d}\gamma(x_1,x_2) \\ &= \int_{\Omega} c_{p,\alpha}(x,y)\mathrm{d}\bar{\gamma}(x,y) + |\gamma_1 - \bar{\gamma}|(\Omega^2)/2 + |\gamma_2 - \bar{\gamma}|(\Omega^2)/2 \\ &= \frac{1}{2}\mu(\Omega) + \frac{1}{2}\nu(\Omega) + \int_{\Omega^2}(c_{p,\alpha}(x,y) - 1)\mathrm{d}\bar{\gamma}(x,y) \geq \tilde{C}(\mu,\nu). \end{aligned}$$

For the opposite inequality, remark that the infimum defining $\tilde{C}$ is unchanged by adding the constraint that the sub-coupling $\gamma \in \Pi_{\leq}(\mu,\nu)$ is concentrated on the set $\{(x,y) \in \Omega^2 \,;\, (c_{p,\alpha}(x,y) - 1) \leq 0\}$. For such a plan, let $\rho_1 = \mu - \pi^1_{\#}\gamma$ and $\rho_2 = \nu - \pi^2_{\#}\gamma$ and define the pair of semi-couplings

$$\gamma_i = \gamma + \mathrm{diag}_{\#}(\rho_0 \wedge \rho_1) + \mathrm{diag}_{\#}(\rho_i - \rho_0 \wedge \rho_1), \quad i \in \{0,1\}$$

where $\mathrm{diag} : x \mapsto (x,x)$ lifts $\Omega$ to the diagonal in $\Omega^2$. One has

$$\int_{\Omega^2} h_{x_1,x_2}(\gamma_1,\gamma_2) = \mu(\Omega) + \nu(\Omega) + \int_{\Omega^2}(c_{\alpha,p}(x,y) - 1)\mathrm{d}\gamma$$

and it follows $C_L(\mu,\nu) \leq \tilde{C}(\mu,\nu)$. □

The problem introduced in Theorem 2.3.4 has, after removing the constant terms and multiplying by $\alpha^p$, the same minimizers as

$$\min\left\{\int_{\Omega^2}(|y-x|^p - \alpha^p)\mathrm{d}\gamma(x,y) \,;\, \gamma \in \Pi_{\leq}(\mu,\nu)\right\}. \tag{2.3.3}$$

The latter is a formulation of the *optimal partial transport* problem, a variant of optimal transport studied in [32, 69]. It is proved in [32] that for any choice of $\alpha > 0$ corresponds the choice of $m(\alpha)$ such that $0 \leq m(\alpha) \leq \min\{\mu(\Omega), \nu(\Omega)\}$ for which the minimizers of (2.3.3) and the minimizers of

$$\min\left\{\int_{\Omega^2}|y-x|^p\mathrm{d}\gamma(x,y) \,;\, \gamma \in \Pi_{\leq}(\mu,\nu),\ \gamma(\Omega^2) = m(\alpha)\right\}$$

are the same. The variable $\alpha$ is the Lagrange multiplier for the constraint of total mass and corresponds to the maximum distance over which transport can occur (of course, this does not mean that optimal plans for (2.3.3) are obtained by simply restricting optimal transport plans to the set of points distant by $\alpha$ at most). The function $m(\alpha)$ cannot be inverted in general (think of atomic measures) but it is proved in [32, Corollary 2.11] that it can be inverted if $\mu$ or $\nu$ is absolutely continuous.

Let us finally show that $C_L$, and thus optimal partial transport (in the parametrization "maximum distance" $\alpha$), admits an optimal entropy transport formulation. A similar equivalence result was proved in [130], with slightly different definitions and for the case $p = 2$.

**Proposition 2.3.5** (Optimal entropy-transport formulation)**.** *For $\mu,\nu\in\mathscr{M}_+(\Omega)$, one has*

$$C_L(\mu,\nu)=\min_{\gamma\in\mathscr{M}_+(\Omega^2)}\left\{\int_{\Omega^2}c_{\alpha,p}(x,y)\mathrm{d}\gamma(x,y)+|\mu-\pi^1_\#\gamma|(\Omega)/2+|\nu-\pi^2_\#\gamma|(\Omega)/2\right\}.$$

This formulation is indeed an optimal entropy-transport problem since the total variation is the $f$-divergence corresponding to the entropy function $f(s)=|s-1|$ if $s\geq 0$ and $+\infty$ otherwise, i.e. $|\mu-\nu|=\int\psi_f(\nu|\mu)=\int\psi_f(\mu|\nu)$ for $\mu,\nu\in\mathscr{M}_+(\Omega)$.

*Proof.* In this proof, we denote by $\tilde{C}(\mu,\nu)$ the infimum on the right-hand side which is a minimum by Theorem 1.2.19. If we show that any minimizer also satisfies the constraint $\gamma\in\Pi_{\leq}(\mu,\nu)$ then the problem exactly rewrites as the formula of Theorem 2.3.4 and the conclusion follows. To show that the marginals of minimizers are indeed smaller than $\mu$ and $\nu$, take any $\gamma\in\mathscr{M}+(\Omega)$ and build $\bar{\gamma}$ such that $\bar{\gamma}\leq\gamma$ and $\pi^1_\#\bar{\gamma}=\mu\wedge\pi^1_\#\gamma$. By construction, one has

$$(|\mu-\pi^1_\#\gamma|-|\mu-\pi^1_\#\bar{\gamma}|)(\Omega)=|\pi^1_\#\gamma-\pi^1_\#\bar{\gamma}|(\Omega)=|\gamma-\bar{\gamma}|(\Omega),$$

and

$$(|\nu-\pi^2_\#\gamma|-|\nu-\pi^2_\#\bar{\gamma}|)(\Omega)\geq-|\pi^2_\#\gamma-\pi^2_\#\bar{\gamma}|(\Omega)=-|\gamma-\bar{\gamma}|(\Omega).$$

By denoting $F$ the functional minimized to define $\tilde{C}(\mu,\nu)$, it holds

$$F(\gamma)-F(\bar{\gamma})\geq\int_{\Omega^2}c_{p,\alpha}\mathrm{d}(\gamma-\bar{\gamma})\geq 0.$$

A similar truncation procedure for the other marginal leads to the result. □

# Part II.

# Numerical Methods

# Chapter 3.

# Scaling Algorithms for Optimal Couplings

In this chapter we introduce a class of algorithms for the numerical resolution of optimal coupling problems and study convergence with various approaches.
In Section 3.1, we make the simple yet powerful remark that many variational problems involving optimal transport including barycenters, gradient flows and their unbalanced counterparts can all be cast as a *generic problem*. We prove well-posedness and existence of minimizers for this generic formulation.
In Section 3.2, we propose to generalize the approach popularized by Cuturi for computing optimal transport distances by retaining two features (i) a strictly convex regularization and (ii) alternate maximization on the dual. We show a global convergence rate of the primal iterates in the discrete setting for a general Bregman regularization.
In Section 3.3, we derive the explicit form of the iterates for the specific case of entropic regularization which leads to a new class of simple *scaling* algorithms. We suggest a stabilization method to reach small regularization parameters without sacrificing speed and make informal comments on implementation as well as possible generalizations.
In Section 3.4, we suggest to use a method to accelerate convergence known as the *successive over-relaxation* method. This idea seems to be new in the context of optimal transport and we show that it can greatly improve convergence speed in practice. We complement this section with a detailed local convergence analysis which gives the optimal acceleration parameter and factor (these results are well known in the context of resolution of linear systems).
Finally, in Section 3.5, we consider scaling algorithms in the continuous setting. We show that the iterates for separable marginal functions are nonexpansive for the *Thompson* metric, a metric issued from non-linear Perron-Frobenius theory. In some particular cases of practical importance (computation of barycenters, gradient flows, optimal plans for $\widehat{W}_2$), we prove that the *global* convergence rate is linear for this metric.
The material of this chapter is partly based on the published article [45] and a number of new results have been added.

## 3.1. Generic formulation for scaling algorithms

In this section, we show that many instances of optimal transport like problems, including Wasserstein barycenters, gradient flows, and their unbalanced counterparts, can be phrased in a common formulation which can serve as a basis for a numerical scheme. This simple remark allows to unify the treatment of several algorithms that can be found in the literature, and to deal painlessly with their generalization to the unbalanced setting. This section gives also the opportunity to define a few variational problems where unbalanced optimal transport can be plugged in place of classical optimal transport and to state existence theorems.

The *generic formulation* we refer to is the following variational problem

$$\boxed{\min\left\{\int_{X\times Y} c(x,y)\cdot \mathrm{d}\gamma(x,y) + F_1(\pi^x_\#\gamma) + F_2(\pi^y_\#\gamma)\ ;\ \gamma\in\mathscr{M}_+(X\times Y)^n\right\}} \tag{3.1.1}$$

where $n \geq 1$ is the number of couplings involved—which can be more than one for solving barycenter or multispecies gradient flows—and

(i) $X$ and $Y$ are the spaces where the marginals are defined, which have to be finite discrete spaces for the numerical method we propose, but merely assumed compact metric in this section;

(ii) $c : X \times Y \to \bar{\mathbb{R}}^n$ is a l.s.c. cost function and "$\cdot$" the usual scalar product on $\mathbb{R}^n$;

(iii) $F_1 : \mathscr{M}_+(X)^n \to \mathbb{R}\cup\{\infty\}$ and $F_2 : \mathscr{M}_+(Y)^n \to \mathbb{R}\cup\{\infty\}$ are convex funtionnals acting on the marginals, which, for the effective numerical resolution, need to be *simple*, in a sense made precise later on.

Although it is slightly improper, we still use the word "coupling" to refer to $\gamma$ because at optimality, $\gamma$ is an optimal transport coupling between its marginals, although they are not necessarily fixed ahead.

### 3.1.1. Classical optimal transport

The classical optimal optimal transport problem can be formulated in the form (3.1.1). For $(\mu,\nu)\in\mathscr{P}(X)\times\mathscr{P}(Y)$, it is recovered by posing

$$F_1(\sigma) = \begin{cases} 0 & \text{if } \sigma=\mu \\ \infty & \text{otherwise} \end{cases}, \qquad F_2(\sigma) = \begin{cases} 0 & \text{if } \sigma=\nu \\ \infty & \text{otherwise} \end{cases}.$$

We may also use the notation $F_1 = \iota_{\{\mu\}}$ and $F_2 = \iota_{\{\nu\}}$ where "$\iota$" is the usual (convex) indicator function of a convex set. Since a singleton is weakly closed and convex, these functions are proper, convex, and weakly l.s.c.

### 3.1.2. Unbalanced optimal transport

The suitable formulation of unbalanced optimal transport that can fit in the framework of (3.1.1) is the optimal entropy-transport problem of [103], defined in Section 1.2.3. Fixing two marginals $(\mu,\nu) \in \mathscr{M}_+(X) \times \mathscr{M}_+(Y)$ and $f_1, f_2$ entropy functions as in Definition 1.2.15 ($\psi_f$ denotes their perspective function, see Appendix A), optimal entropy-transport problems are recovered by posing

$$F_1(\sigma) = \int_X \psi_{f_1}(\sigma|\mu), \qquad\qquad F_2(\sigma) = \int_Y \psi_{f_2}(\sigma|\nu).$$

We recall that by the results of Chapter 1, problems of this form are equivalent to semi-coupling problems but the latter are less suitable for the numerical approach of this chapter. Let us briefly recall for the convenience of the reader that when $f_1 = f_2 = \iota_{\{1\}}$, this problem boils down to classical optimal transport. Otherwise, optimal entropy-transport problems can be viewed as a relaxation of classical optimal transport where the marginal constraints are slackened and replaced by terms in the objective functional penalizing the deviation of the marginals of $\gamma$ from $\mu$ and $\nu$.

Important cases for this class of problems are:

- for $f(s) = s\log s - s + 1$, and $c(x,y) = -\log\cos^2(\min\{\mathrm{dist}(x,y), \pi/2)$ on $X = Y$, this problem defines (the square of) the metric $\widehat{W}_2$ considered in Chapter 2. Variants with a quadratic cost cost $c(x,y) = \mathrm{dist}(x,y)^2$ or with weighted relative entropies are also possible, see Section 2.2.2;

- for $f(s) = \alpha|s-1|/2$, the $f$-divergence corresponds to the total variation distance and one recovers the optimal partial transport problem. The parameter $\alpha$ stands for the maximum cost of transport, it is the Lagrange multiplier of the parameter "total transported mass", see Section 2.3.2;

- another choice is $f(s) = \iota_{[\alpha,\beta]}(s)$ where $0 < \alpha \leq 1 \leq \beta < \infty$. This corresponds to an optimal transport problem where, instead of imposing marginal constraints, one imposes a *density range* constraint.

### 3.1.3. Optimal transport barycenters

It is well known that the Euclidean barycenter $v_b$ of a family $(u_i)_{i=1}^n$ of points in a Euclidean space $E$ with positive weights $(w_i)_{i=1}^n$ admit the variational characterization

$$v_b \in \arg\min_{v\in E} \sum_{i=1}^n w_i|v-u_i|^2.$$

This characterization only involves metric quantities, and thus can be generalized to define barycenters on metric spaces (the name Fréchet mean or Karcher mean is sometimes used in this context): one then replace each term $|v-u_i|^2$ by $d(v,u_i)^2$ where $d$ is the metric. On the space of probability measures, the problem of Wasserstein barycenters, first studied by [1] and has found applications in image processing [133] or machine learning [154].

The Wasserstein distance (squared) is an instance of an optimal transport cost $C_c(\mu,\nu)$ computed between two probability measures $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$ with the ground cost $c(x,y) = \mathrm{dist}(x,y)^2$. A general problem of interest is thus to define the optimal transport barycenter of a finite family $(\mu_i)_i \in \mathscr{P}(X)^n$ of probability measures with positive weights $(w_i)_{i=1}^n$ as a minimizer to the problem

$$\inf\left\{ \sum_{i=1}^n w_i C_{c_i}(\mu_i,\nu) \;;\; \nu \in \mathscr{P}(Y) \right\}$$

if such a minimizer exists, where $C_{c_i}$ is the optimal transport cost associated to a ground cost $c_i : X \times Y \to \mathbb{R} \cup \{\infty\}$. Since each $C_{c_i}$ is itself defined by the variational problem (1) (in the introduction), the previous problem rewrites

$$\inf\left\{ \sum_{i=1}^n w_i \int_{X\times Y} c(x,y)\,\mathrm{d}\gamma_i(x,y) \;;\; \forall i \in \{1,\dots,n\},\, \gamma_i \in \Pi(\mu_i,\nu) \text{ for some } \nu \in \mathscr{P}(Y) \right\}$$

where $\Pi(\mu,\nu)$ denotes the set of (exact) couplings between two probability measures $\mu$ and $\nu$. This problem falls into the framework of the generic formulation by posing the vector valued cost $c(x,y) \coloneqq (w_i c_i(x,y))_{i=1}^n$ and the functions

$$F_1(\sigma) = \begin{cases} 0 & \text{if } \sigma_i = \mu_i \text{ for all } i \\ \infty & \text{otherwise} \end{cases}, \qquad F_2(\sigma) = \begin{cases} 0 & \text{if } \sigma_i = \sigma_j \text{ for all } (i,j) \\ \infty & \text{otherwise.} \end{cases}$$

### 3.1.4. Unbalanced barycenters

This definition of barycenters can be extended to the unbalanced case. Similarly as above, one fixes finite families of measures $(\mu_i)_{i=1}^n \in \mathscr{M}_+(X)^n$ and positive weights $(w_i)_{i=1}^n$ and searches for a measure $\nu \in \mathscr{M}_+(Y)$ that minimizes

$$\inf\left\{ \sum w_i C_{c_i,f_i}(\nu,\mu_i) \;;\; \nu \in \mathscr{M}_+(Y) \right\} \tag{3.1.2}$$

where now $C_{c_i,f_i}$ is an optimal entropy-transport cost associated to the ground cost $c_i$ and the entropy function $f_i$ (Definition 1.2.18). Again, such a cost admits equivalent formulations in some cases as studied in Chapter 1, but the optimal entropy-transport form is the most suitable for the numerical approach of this section. Plugging the definition of $C_{c_i,f_i}$ in the previous formula one finds that this problem can be reduced to the form (3.1.1) by posing the vector valued cost $c(x,y) \coloneqq (w_i c_i(x,y))_{i=1}^n$ and the functions

$$F_1(\sigma) = \sum_{i=1}^n w_i \psi_{f_i}(\sigma_i|\mu_i), \qquad F_2(\sigma) = \inf\left\{ \sum_{i=1}^n w_i \psi_{f_i}(\sigma_i|\nu) \;;\; \nu \in \mathscr{M}_+(Y) \right\}.$$

One may object that the unknown of interest is not so much the family of couplings $(\gamma_i)$ but the minimizer $\nu$. Fortunately, the algorithm studied later returns $\nu$ as a byproduct (see Section 5.2, the reason is that it is required to compute the proximal operator).

It should be noted that, since $\mathrm{W}_2$ and $\widehat{W}_2$ are geodesic metrics, computing the barycenters between two measures with various weights is a way to compute geodesics.

### 3.1.5. Implicit scheme for gradient flows

It is well known that a handful of evolutional partial differential equations (PDEs) are characterized as gradient flows of certain functionals in the Wasserstein space (see Chapter 6 for more details). The implicit Euler scheme, known as JKO in this context—a particular case of De Giorgi minimizing movements—is of theoretical interest since it allows to prove the existence of solutions to certain PDEs. It is also of numerical interest because it provides with a variational characterization of a time-discretized flow. Choosing a time step $\tau > 0$ and an initial condition $\mu$, the implicit scheme for the Wasserstein gradient flow of a functional $G : \mathscr{P}(X) \to \mathbb{R} \cup \{\infty\}$ consists in defining a sequence $(\mu_k^\tau)_{k\in\mathbb{N}}$ via

$$\mu_0^\tau := \mu, \qquad \mu_{k+1}^\tau \in \arg\min \left\{ G(\nu) + \frac{1}{2\tau} W_2^2(\mu_k^\tau, \nu) \; ; \; \nu \in \mathscr{P}(X) \right\},$$

where $W_2^2$ is the Wasserstein squared distance. Replacing it by an optimal transport cost $C_c$ associated to a ground cost $c : X^2 \to \mathbb{R} \cup \{\infty\}$, each time step involves to solve

$$\inf \left\{ \int_{X^2} c(x,y) \mathrm{d}\gamma(x,y) + 2\tau G(\pi^2_\# \gamma) \; ; \; \gamma \in \mathscr{P}(X^2), \pi^1_\# \gamma = \mu_k^\tau \right\}.$$

Other numerical advantages are due to the fact that implicit schemes tend to be more stable for large time steps and are applicable for non-smooth functionals. This problem can be cast in the framework of (3.1.1) by posing

$$F_1(\sigma) = \begin{cases} 0 & \text{if } \sigma = \mu_k^\tau \\ \infty & \text{otherwise} \end{cases}, \qquad F_2(\sigma) = 2\tau G(\sigma).$$

An interesting remark is that the variational problem defining an implicit step of a gradient flow can also be interpreted as a regularized inverse problem where $G$ is a regularization term and optimal transport is used as a fidelity term [72, 92]. The time variable translates then into an "amount of regularization" parameter.

### 3.1.6. Unbalanced gradient flows

The introduction of unbalanced optimal transport metrics (Chapter 2) paves the way for further applications of gradient flows in the space of nonnegative measures. Choosing an unbalanced optimal transport cost $C_{c,f}$ in the form of an optimal entropy-transport problem with cost function $c$ and entropy function $f$, the computation of one step involves the resolution of

$$\inf \left\{ \int_{X^2} c \mathrm{d}\gamma + \int_X \psi_f(\pi^1_\# \gamma | \mu_k^\tau) + \int_X \psi_f(\pi^2_\# \gamma | \nu) + 2\tau G(\nu) \; ; \; \gamma \in \mathscr{P}(X^2), \nu \in \mathscr{M}_+(X) \right\}$$

which falls in the framework of (3.1.1) by posing

$$F_1(\sigma) = \int_X \psi_f(\sigma | \mu_k^\tau), \qquad F_2(\sigma) = \inf \left\{ 2\tau G(\nu) + \int_X \psi_f(\sigma|\nu) \; ; \; \nu \in \mathscr{M}_+(X) \right\}.$$

### 3.1.7. General case of simple functionals

Before turning in the next section to the method for solving (3.1.1) numerically, let us briefly clarify what is meant by *simple* functionals in the definition of (3.1.1). Without going into to much details at this point, the method we introduce is iterative and has a complexity per iteration which is of the order of that of finding, given $\bar{\sigma} \in \mathscr{M}_+(X)$, an element in

$$\arg\min \{H(\sigma|\bar{\sigma}) + F_i(\sigma) \; ; \; \sigma \in \mathscr{M}_+(X)\} \tag{3.1.3}$$

where $H$ is the relative entropy functional (defined further in Definition 3.5.1). A *simple* functional $F_i$ is a functional such that this minimum can be computed quickly. This is the case when $X$ is discrete and $F_i$ is separable, since then (3.1.3) rewrites as a family of 1 dimensional problems that can be solved in parallel. Separability is however not strictly necessary and we study and solve in Section 5.3 a problem where $F_i$ is not separable.

In the examples reviewed above, the functions $F_i$ are sometimes defined through auxiliary variational problems. This fact is not an issue in itself for the algorithms as long as (3.1.3) can be solved efficiently. For instance, functions of the form

$$F(\sigma) = \min \left\{ \sum_{i=1}^{n} \int_X \psi_{f_i}(\sigma_i|\nu_i) + G(\nu) \; ; \; \nu \in \mathscr{M}_+(X)^I \right\} \tag{3.1.4}$$

where $G$ is convex and $(f_i)_{i=1}^n$ is a finite family of entropy functions can model a great variety of problems and in particular all the problem introduced in this section. A graphical interpretation of these problems is suggested in Figure 3.1 along with some examples.

The following proposition shows that such a function has all the required properties. It applies to all cases reviewed above in this section.

**Proposition 3.1.1** (Convexity)**.** *If $X$ is a compact metric space, $G$ is proper, weakly l.s.c. convex and lower bounded and $(f_i)_{i=1}^n$ is a family of entropy functions such that $f_i(0) > 0$ for all $i$, then* (3.1.4) *defines a proper, l.s.c. convex function and the minimum defining $F$ is attained.*

*Proof.* In order to simplify notations assume that $n = 1$ (the proof for $n > 1$ is the same) and let us denote by $J_\sigma(\nu) = \int_X \psi_f(\sigma|\nu) + G(\nu)$ the functional minimized in (3.1.4) which is jointly convex and l.s.c. Since $G$ is proper and $f(1) = 0$, $F$ is proper.

Let us now fix $\sigma \in \operatorname{dom} F$ and show that the minimum is attained: since $J_\sigma$ is weakly l.s.c. and $X$ compact, one just need to show that any minimizing sequence $(\nu_n)$ is bounded, which is guaranteed by Lemma 3.1.3.

For the convexity property, let $\sigma_a, \sigma_b \in \mathscr{M}_+(X)$ and $\theta \in ]0,1[$. Denoting $\nu_a, \nu_b$ corresponding minimizers in (3.1.4) one has by using the joint convexity of $J$:

$$\begin{aligned} F(\theta\sigma_a + (1-\theta)\sigma_b) &\leq J_{\theta\sigma_a + (1-\theta)\sigma_b}(\theta\nu_a + (1-\theta)\nu_b) \\ &\leq \theta J_{\sigma_a}(\nu_a) + (1-\theta) J_{\sigma_b}(\nu_b) = \theta F(\sigma_a) + (1-\theta) F(\sigma_b). \end{aligned}$$

Finally for the weak lower semicontinuity, let $(\sigma_n)_{n\in\mathbb{N}}$ be a sequence weakly converging to $\sigma \in \mathscr{M}_+(X)$ and let $\nu_n$ be a corresponding sequence of minimizers of (3.1.4). If $\underline{\lim} J_{\sigma_n}(\nu_n) = \infty$ there is nothing to prove, otherwise, up to a subsequence, $(J_{\sigma_n}(\nu_n))_{n\in\mathbb{N}}$ is bounded and this

implies by Lemma 3.1.3 that $(\nu_n)$ is bounded and thus converges to some $\bar{\nu} \in \mathscr{M}_+(X)$, up to a subsequence. We conclude by using the joint lower semicontinuity of $J$ in $(\sigma, \nu)$:

$$F(\sigma) \leq J_\sigma(\bar{\nu}) \leq \liminf J_{\sigma_n}(\nu_n) = \liminf F(\sigma_n). \qquad \square$$

The next theorem contains as special cases the well known existence of optimal transport plans, the existence of minimizers for the optimal entropy-transport problems, the existence of optimal transport barycenters (possibly unbalanced) and the fact that gradient flow dicrete steps (possibly unbalanced) are well defined. However, it is only stated in the simple setting of compact spaces. Just for this result, we denote $F^X$ and $F^Y$ the marginals functional in order to avoid ambiguities.

**Theorem 3.1.2** (Existence of minimizer)**.** *Assume that $X$ and $Y$ are compact metric spaces, that $F^X, F^Y$ are of the form* (3.1.4) *satisfying the assumptions of Proposition 3.1.1. Assume moreover that $F^X$ is such that $G^X$ has compact sublevel sets and $f_i^X(x) \to \infty$ as $x \to \infty$ (or the same for $F^Y$). Then a minimizer for* (3.1.1) *exists and so do the corresponding minimizers involved in the definition of $F^X$ and $F^Y$ in* (3.1.4)*.*

*Proof.* Once again, let us use the direct method of the calculus of variations. By Proposition 3.1.1 and since $c_i$ is assumed l.s.c., the minimized functional is weakly l.s.c. Assume that (3.1.1) is feasible (otherwise the claim is void). This implies by our assumption that any minimizing sequence $\gamma_n$ is bounded because the sequences $(\pi^i_\# \gamma_n)$ are bounded, by Lemma 3.1.3. $\square$

**Lemma 3.1.3.** *Assume that $f$ is an entropy function and $\mu_n, \nu_n \in \mathscr{M}_+(X)$ are two sequences of nonnegative measures on a topological space $X$ such that $\int \psi_f(\mu_n|\nu_n)$ is bounded. Then*

*(i) if $(\mu_n)$ is bounded and $f(0) > 0$ then $(\nu_n)$ is bounded;*

*(ii) if $(\nu_n)$ is bounded and $f(\infty) > 0$ then $(\mu_n)$ is bounded.*

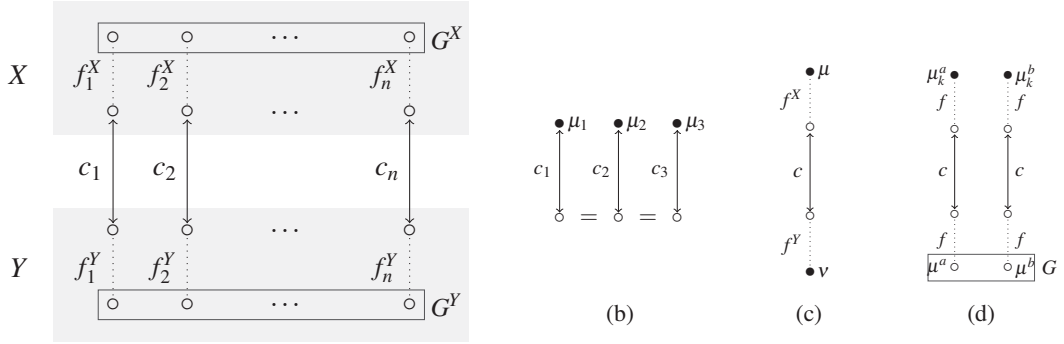
*Proof.* (i) Since $f(0) > 0$, $f(1) = 0$ and $f$ is l.s.c., there exists $r > 0$ such that $f(x) \geq \eta > 0$ for all $x \leq r$. This implies that $\nu_n(\{\mathrm{d}\mu_n/\mathrm{d}\nu_n \leq r\})$, and thus $\nu_n$, are bounded.

(ii) This is equivalent to (i) by exchanging the role of $\mu_n$ and $\nu_n$ and taking the reverse entropy of $f$, see Remark 1.2.17. $\square$

Beyond the set of problems reviewed in this section, an interesting illustration for (3.1.4) is the problem of Wasserstein propagation introduced in [153]. Given a graph with edges $\mathscr{E}$ and vertices $\mathscr{V}$, assign a probability measure $\mu_v \in \mathscr{P}(X)$ to all leaves $v \in \mathscr{V}_0$ and find a minimizer to

$$\min \left\{ \sum_{(u,v)\in\mathscr{E}} C_c(\mu_u, \mu_v) \; ; \; \mu_v \in \mathscr{P}(X) \text{ for all } v \in \mathscr{V} \setminus \mathscr{V}_0 \right\}.$$

Up to duplicating some edges (and link them with an equality constraint) until the graph admits a 2-colorization, this problem can be formulated, also in an unbalanced variant, in the framework of (3.1.1).




(a) (b) (c) (d)

Figure 3.1.: Representation of some problems of the form (3.1.1). Each circle represents a measure (fixed in black, variables in white); the terms in the variational problem are represented by arrows for transport terms, dotted lines for divergence functional terms and rectangles for the $G$ terms. (a) With $F_i$ of the form (3.1.4) (b) classical optimal transport barycenter (c) unbalanced optimal transport (d) unbalanced gradient flow of 2 species with interaction.

### 3.1.8. Bibliographical comments

Before we head into our first numerical method, a few bibliographical comments are in order: optimal transport has been, indeed, a subject of active research in optimization for years.

The Kantorovich formulation as a linear program (see introduction, eq. (1)), when restricted to discrete measures, can be directly tackled using simplex or interior point methods. For the special case of optimal linear assignment (i.e. for sums of the same number of uniform-mass Diracs) one can also use dedicated algorithms such as the Hungarian method [30] or the auction algorithm [17]. The time complexity of these algorithms is roughly cubic in the number of atoms, and hence they do not scale to very large problems. In [144] a multi-scale algorithm is developed that consistently leverages the structure of geometric transport problems to accelerate linear program solvers. In the specific case of costs that derive from a Lagrangian, first order non-smooth optimization schemes can be used to solve the dynamic formulation [12, 123] (this is an approach we study in Chapter 4). A related, but dual, method for the computation of optimal partial transport problem has been suggested in [87].

A last class of approaches deals with semi-discrete problems, when one measure has a density and the second one is discrete. This problem, introduced by Alexandrov and Pogorelov as a theoretical object, can be solved numerically with geometric tools when using the squared Euclidean cost [6]. This methods further enhanced by the use of advanced computational geometry [114, 100] can be considered the state of the art for finding optimal transport maps for the quadratic cost in 2-D or 3-D domains, but are very specific to this context and this problem. In the same framework, another approach consists in solving the Monge-Ampere equation [15].

**Entropic regularization.** In order to cope with large scale problems with arbitrary transportation costs, a recent class of approaches, initiated and revitalized by the paper of Marco

Cuturi [51], proposes to compute an approximate transport plan coupling using entropic regularization. This idea has its origins in many different fields, most notably it is connected with Schrödinger's problem in statistical physics [149, 99] and with the iterative scaling algorithm by Sinkhorn [151] a.k.a. IPFP [57]. Several follow-up articles [52, 14] have shown that the same strategy can also be used to tackle the computation of barycenters for the Wasserstein distance (as initially formulated by [1]), and for solving OT problems on geometric domains (such as regular grids or triangulated meshes) using convolution and the heat diffusion kernel [152]. Some theoretical properties of this regularization are studied in [34], including the $\Gamma$-convergence of the regularized problem toward classical transport when regularization vanishes.

**Optimization with Bregman divergences.** The success of this entropic regularization scheme is tightly linked with use of the relative entropy, a.k.a. Kullback-Leibler divergence, as a natural Bregman divergence [25] for optimization with a positivity constraint. The most simple algorithm, which is actually at the heart of Sinkhorn's iterations, is the iterative projection on affine subspaces for the relative entropy. A refined version of these iterations, which works for arbitrary convex sets (not just affine spaces) is the so-called Dykstra's algorithm [64], which can be interpreted as an iterative block-coordinates maximization on a dual problem. Dykstra's algorithm is known to converge when used in conjunction with Bregman divergences [9, 38]. Many other first order proximal methods for Bregman divergences exists. The most simple one is the proximal point algorithm [65], but most proximal splitting schemes have been extended to this setting, such as for instance ADMM [165], primal-dual schemes [39] and forward-backward [159].

The algorithm we propose in this chapter extends Sinkhorn's iterations to more complex problems. Its simple structure is due to the fact that the non-linear part of the functional only involves the marginals of the couplings that are being optimized. Note that the resolution of an optimal entropy-transport formulation, in conjunction with entropic smoothing has been introduced, without proof of convergence, in [74] for application in machine learning.

## 3.2. Regularized algorithm in discrete setting

In this section, we propose a numerical method for finding approximate solutions to (3.1.1) in the discrete setting, i.e. when $X = (x_i)_{i\in I}$ and $Y = (y_j)_{j\in J}$ where $I$ and $J$ are finite sets of indices. In this setting, we may identify the set of measures and sets of functions to finite families of real numbers (indexed by $I$ or $J$). Given a cost $c = (c_{i,j})$ with $c_{i,j} \in \mathbb{R}\cup\{\infty\}$ and two convex, l.s.c., proper functions $F_1 : \mathbb{R}_+^I \to \mathbb{R}\cup\{\infty\}$ and $F_2 : \mathbb{R}_+^J \to \mathbb{R}\cup\{\infty\}$, the problem (3.1.1) rewrites as

$$\min\left\{ c\cdot\gamma + F_1\Big(\sum_{j\in J}\gamma_{i,j}\Big) + F_2\Big(\sum_{i\in I}\gamma_{i,j}\Big)\ ;\ \gamma\in\mathbb{R}_+^{I\times J}\right\}. \tag{3.2.1}$$

where $c\cdot\gamma = \sum c_{i,j}\gamma_{i,j}$. Without loss of generality and for simplicity, this section considers only one unknown plan (i.e. $n = 1$ in (3.1.1)). The case $n \in \mathbb{N}$ can be reduced to the case $n = 1$ by considering an unknown coupling $\gamma$ on the disjoint union of $n$ replica of $X \times Y$ (in Section 3.5, the case $n > 1$ reappears because, there, separability matters). We also explain in section 3.3.3 how to adapt the algorithm to this case.

### 3.2.1. Regularization

The variational problem (3.2.1) is a non-smooth and (non-strongly) convex minimization problem of typically large dimension ($\gamma$ lives on the product space $\mathbb{R}^{I\times J}$). In this section, generalizing the approach of [51] and [14], we show that simple algorithms can be derived by studying the structure of this problem and accepting to solve it only approximately. Our analysis starts with the dual of (3.2.1).

**Proposition 3.2.1** (Duality)**.** *The minimum to* (3.2.1) *is attained and equals*

$$\sup\left\{ -F_1^*(-u) - F_2^*(-v)\ ;\ (u,v)\in\mathbb{R}^I\times\mathbb{R}^J \text{ and } u_i + v_j \le c_{i,j}\ \forall (i,j)\in I\times J\right\}. \tag{3.2.2}$$

*Proof.* This is an application of Fenchel-Rockafellar duality theorem (Appendix A). The qualification constraint is satisfied because $F_1^*$ and $F_2^*$ are proper and Lemma 3.2.2 below implies that there exists a feasible couple $(u,v)$ such that $u_i + v_j < c_{i,j}$ for all $(i,j)$. □

**Lemma 3.2.2.** *Let $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ be a function with a nonempty domain included in $\mathbb{R}_+^n$. Then $f^*$ is increasing w.r.t. the usual partial ordering in $\mathbb{R}^n$ $(x \preceq y) \Leftrightarrow (x_i \le y_i\ \forall i)$. In particular, $y \in \operatorname{dom} f^*$ whenever $y \preceq x$ for some $x \in \operatorname{dom} f^*$.*

*Proof.* It follows from the definition of the conjugate and the fact that the scalar product $x \mapsto x\cdot z$ is nondecreasing for the same partial order whenever $z \in \mathbb{R}_+^n$. □

The structure of this problem is appealing: for instance one can explicitly maximize with respect to one variable, the other being fixed. The monotonicity of $-F^*(-\cdot)$ implies that, maximizing w.r.t. $u$ with $\bar{v}$ fixed and subsequently maximizing with respect to $v$ leads to a pair of so-called $c$-conjugate functions $(u,v)$ satisfying

$$u_i = \min_{j\in J}\ c_{i,j} - v_j, \qquad\qquad v_j = \min_{i\in I}\ c_{i,j} - u_i.$$

However, it is clear that this process remains stuck after these two operations, and the resulting $(u,v)$ is not a pair of maximizers (it does not even depend on $F_i$). It is a well-known fact that alternate maximization does not generally converge to a maximizer when the term in the functional that involves the two variables is not differentiable. However, since some guarantees exist when that term is smooth, we introduce a regularization in the dual problem (3.2.2) to make the coupling term smooth. This regularization can in general take the form of a Bregman divergence, which is the setting in which we will prove the general convergence result. The interest of Bregman divergences lies in Lemma 3.2.11 below, which allows to obtain a speed of convergence of the primal iterates. We specialize then to (the conjugate of) the relative entropy, which has particularly nice algebraic properties in this context.

### Bregman divergences

Bregman divergences are generally defined using functions of Legendre type.

**Definition 3.2.3** (Legendre type)**.** *A convex, l.s.c. proper function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ ($n \in \mathbb{N}^*$) is said of* Legendre type *if $h$ and $h^*$ are strictly convex on their respective domains, or on any convex set included in the domain of their subdifferential.*

This definition (equivalent to other found in the literature [134]) implies in particular that $h$ and $h^*$ are continuously differentiable on the interior of their domain—which have nonempty interior—and that $\nabla h$ and $\nabla h^*$ are inverse bijections on their domain. The Bregman divergence $D_h(x|z)$ is the evaluation at $x$ of the difference between $h$ and its linear approximation at $z$.

**Definition 3.2.4** (Bregman divergence)**.** *The Bregman divergence associated to a function of Legendre type $h$ is defined for $(x,z) \in (\mathring{\mathrm{dom}}\, h)^2$ as*

$$D_h(x|z) := h(x) - h(z) - \nabla h(z) \cdot (x - z).$$

Clearly, $D_h$ is convex w.r.t. the first argument, but it is not jointly convex in general.

**Remark 3.2.5.** *The relative entropy functional on $\mathbb{R}_+^2$ is both a $f$-divergence $\psi_h$ and a Bregman divergence $D_h$, by posing $h(x) = x\log x - x + 1$, $x \in \mathbb{R}_+$. Actually it is essentially the only such function! Indeed, for convex functions $h, f$ (with $h \in C^1(]0,\infty[)$ the fact that $D_h = \psi_f$ implies that for all $x,z > 0$ it holds*

$$h(x) - h(z) - \nabla h(z) \cdot (x - z) = z f(x/z) \Rightarrow \nabla h(x) - \nabla h(z) = \nabla f(x/z)$$

*so, modulo a constant $\nabla h = \nabla f = \alpha \log$ for $\alpha > 0$. An interesting stronger result is proved in [8]: if $\mathring{\mathrm{dom}}\, h = ]0,\infty[$ and $h$ is of class $C^3$ then $D_h$ is jointly convex if and only if $h$ is the entropy, modulo linear terms. So $f$-divergences and Bregman divergences are almost disjoint concepts, only the relative entropy endorses sometimes one role or the other. If we remove the restriction to the domain, other examples exist, such as the Hellinger divergence.*

The following properties are classical and the strict convexity estimate (iii) will be useful. We denote by $D^{*,i}$ the conjugate of the function $D$ with respect to the $i$-th variable in order to distinguish it from the joint conjugate w.r.t. all variables.

**Lemma 3.2.6** (Strict Bregman convexity)**.**

*(i) For $z \in \mathring{\mathrm{dom}}\, h$, the function $D_h(\cdot|z)$ is convex and its conjugate is*

$$D_h^{*,1}(u|z) = h^*(u + \nabla h(z)) + h(z) - z \cdot \nabla h(z),$$

*whose graph, for fixed $z$, is simply a translation of that of $h^*$.*

*(ii) By denoting $(u,v) = (\nabla h(x), \nabla h(z))$ for $u, v \in \mathring{\mathrm{dom}}\, h$, one has*

$$D_h(x|z) = h(x) + h^*(v) - v \cdot x = D_{h^*}(v|u).$$

*(iii) For $(u,\bar{u}) \in (\mathring{\mathrm{dom}}\, h^*)^2$, it holds*

$$\begin{aligned} D_h^{*,1}(u|z) - D_h^{*,1}(\bar{u}|z) - \nabla_1 D_h^{*,1}(\bar{u}|z) \cdot (u - \bar{u}) &= D_{h^*}(u + \nabla h(z) | \bar{u} + \nabla h(z)) \\ &= D_h(\bar{r}|r) \end{aligned}$$

*where $r := \nabla h^*(u + \nabla h(z))$ and $\bar{r}$ is defined likewise from $\bar{u}$.*

*Proof.* These formula are obtained by simple manipulations. For (i) combine

$$D_h^{*,1}(u|z) = u \cdot \bar{x} - h(\bar{x}) + h(z) + \nabla h(z) \cdot (\bar{x} - z) \quad \text{where} \quad \bar{x} = \nabla h^*(u + \nabla h(z))$$

with $h^*(w) = w \cdot \tilde{x} - h(\tilde{x})$ where $\tilde{x} = \nabla h^*(w)$. For (ii), observe the symmetries in

$$D_h(x|z) = h(x) + h^*(v) - v \cdot z - v \cdot (x - z) = h(x) + h^*(v) - v \cdot x = D_{h^*}(v|u).$$

For (iii), first use (i), then apply (ii). □

### Bregman regularization

Instead, of solving (3.2.1), we propose to solve a regularized problem. Let $h : \mathbb{R}^{I \times J} \to \mathbb{R} \cup \{\infty\}$ be a convex function of Legendre type of domain $\mathbb{R}_+^{I \times J}$ and $D_h$ the associated Bregman divergence (Definitions 3.2.3 and 3.2.4), let $\gamma^{(0)} \in \mathbb{R}_+^{I \times J}$ be a (discrete) reference measure and let $\varepsilon > 0$ be a small parameter. The regularized problem is

$$\boxed{\min_{\gamma \in \mathbb{R}^{I \times J}} E_\varepsilon(\gamma) \quad \text{where} \quad E_\varepsilon(\gamma) := F_1\Big(\sum_{j \in J} \gamma_{i,j}\Big) + F_2\Big(\sum_{i \in I} \gamma_{i,j}\Big) + \varepsilon D_h(\gamma|K).} \tag{3.2.3}$$

where we define the kernel (which implicitly depends on $\varepsilon$ and the cost and $\varepsilon$)

$$K := \nabla h^*\Big(\nabla h(\gamma^{(0)}) - c/\varepsilon\Big).$$

If $c$ is finite, which we assume, one can always choose $\gamma^{(0)}$ such that $K \in \mathring{\mathrm{dom}}\, h$. As detailed in the next proposition, solving (3.2.3) amounts to minimizing (3.2.1) with a regularizing term $\varepsilon D_h(\gamma|\gamma^{(0)})$ added.

**Proposition 3.2.7.** *Denoting $E(\gamma)$ the functional in* (3.2.1) *and $E_\varepsilon(\gamma)$ the functional in* (3.2.3)*, one has*

$$E(\gamma) + \varepsilon D_h(\gamma|\gamma^{(0)}) = E_\varepsilon(\gamma) + \varepsilon D_h(K|\gamma^{(0)}) + c \cdot K$$

*so that the left hand side and $E_\varepsilon(\gamma)$ only differ by a constant and minimizing one amounts to minimizing the other.*

*Proof.* Using the property $(\nabla h^*)^{-1} = \nabla h$ and the definition of $D_h$, one finds

$$\varepsilon D_h(\gamma|\gamma^{(0)}) + c \cdot \gamma = \varepsilon D_h(\gamma|K) + \varepsilon D_h(K|\gamma^{(0)}) + c \cdot K. \qquad \square$$

**Remark 3.2.8** (Interpretation)**.** *This regularization can be interpreted in several ways:*

1. *as the* first step of a proximal point algorithm. *By denoting $\gamma^{(1)}$ the solution to* (3.2.3)*, one has $\gamma^{(1)} = \operatorname{prox}^{D_h}_{E_0/\varepsilon}(\gamma^{(0)})$. It is possible to iterate this relation and define*

$$\gamma^{(\ell+1)} := \operatorname{prox}^{D_h}_{E_0/\varepsilon}(\gamma^{(\ell)}).$$

   *which is the so-called* proximal point algorithm *[65], known to converge to a solution to* (3.2.1)*. By looking at the optimality conditions, one may notice that, in the particular case of entropy regularized optimal transport, $\gamma^{(\ell)}$ can as well be obtained by performing a single proximal step with parameter $\varepsilon/\ell$, but this relation is not true in general.*

2. *as a* regularization. *In inverse problems or statistical learning tasks, when the data is incomplete, one generally introduce some prior information through a divergence function, which additionally makes the problem strictly convex. Here, $\gamma^{(0)}$ favors diffuse minimizers while solutions to the unregularized problem are typically sparse because solutions are optimal plans between their marginals. This improves the performance for statistical learning tasks [51].*

3. *as a* model in its own right. *In the specific case of optimal transport with quadratic cost with $D_h = H$ the relative entropy (defined in Section 3.3), this problem appears as a stochastic version of optimal transport. For instance, it is obtained by considering an optimal matching problem where there is (a specific model of) unknown heterogeneity in the preferences within each class [75]. It also corresponds to the* Shrödinger bridge *problem, which aims at computing the law of motion of the (indistinguishable) particules of a gaz which follow a Brownian motion, conditionally to the observation of its density at times $t = 0$ and $t = 1$ [99].*

As intended, this primal regularization corresponds to a dual problem where the term coupling $(u, v)$ is smoothed.

**Proposition 3.2.9** (Regularized dual)**.** *The minimum in* (3.2.3) *is attained at a unique point and is equal to*

$$\sup_{u\in\mathbb{R}^I, v\in\mathbb{R}^J} \bar{E}_\varepsilon(u,v) \quad \text{where} \quad \bar{E}_\varepsilon(u,v) := -F_1^*(-u) - F_2^*(-v) - \varepsilon D_h^{*,1}((u\oplus v)/\varepsilon|K)$$

*where $(u\oplus v)_{i,j} = u_i + v_j$ for $(i,j) \in I \times J$ and $D_h^{*,1}$ is the conjugate of $D_h$ w.r.t. the first variable.*

*Proof.* This is again an application of Fenchel-Rockafellar duality (Theorem A.0.7). The qualification constraint is satisfied because we see in the expression of $D_h^{*,1}$ given in Lemma 3.2.6 (i) that its domain is a translation of that of $h^*$. One then deduces from Lemma 3.2.2 that there exists a strictly feasible couple $(u,v)$. Uniqueness of the minimizer comes from the strict convexity of $h$. □

**Proposition 3.2.10** (Convergence of regularized minimizers)**.** *Assume that $h$ is superlinear (in all directions) and let $\gamma_\varepsilon$ be the minimizer to* (3.2.3) *for each $\varepsilon > 0$. Then, as $\varepsilon \to 0$, $\gamma_\varepsilon$ tends to $\gamma_0 \in \mathbb{R}_+^{I\times J}$ characterized by*

$$\{\gamma_0\} = \arg\min \left\{ D_h(\gamma|\gamma^{(0)}) \; ; \; \gamma \in \arg\min E \right\}.$$

*where $E$ is the unregularized functional in* (3.2.1)*. In the limit $\varepsilon \to \infty$, one has that $\gamma_\varepsilon$ tends to $\gamma_\infty$ that is characterized by*

$$\{\gamma_\infty\} = \arg\min \left\{ D_h(\gamma|\gamma^{(0)}) \; ; \; \gamma \in \operatorname{dom} E \right\}.$$

*Proof.* Direct proofs are possible, but we will use Γ-convergence results for conciseness. Applying Lemma 2.1.12 (first order Γ-convergence), one has by denoting $E(\gamma)$ that

$$F_\varepsilon : \gamma \mapsto \frac{1}{\varepsilon}(E(\gamma) - \inf E) + D_h(\gamma|\gamma^{(0)})$$

Γ-converges to $F := D_h(\gamma|\gamma^{(0)}) + \iota_{\arg min E}$ as $\varepsilon \downarrow 0$ and this functional has the same minimizers as $E_\varepsilon$ (see Proposition 3.2.7). Similarly, by using this time Lemma 2.1.11 (zeroth order Γ-convergence), one has that

$$\tilde{F}_\varepsilon : \gamma \mapsto \frac{1}{\varepsilon}E(\gamma) + D_h(\gamma|\gamma^{(0)})$$

Γ-converges to $\tilde{F} := D_h(\gamma|\gamma^{(0)}) + \iota_{\operatorname{dom} E}$ as $\varepsilon \uparrow \infty$, because $\operatorname{dom} E \subset \operatorname{dom} D_h(\cdot|\gamma^{(0)})$ and the latter is l.s.c. It remains to show the equicoercivity to have the convergence of minimizers. But for all $\varepsilon > 0$, minimizers $\gamma_\varepsilon$ of $F_\varepsilon$ (or equivalently $\tilde{F}_\varepsilon$) satisfy

$$D_h(\gamma_\varepsilon|\gamma^{(0)}) \le D_h(\gamma_0|\gamma^{(0)}) < \infty$$

which is enough because $D_h(\cdot|\gamma^{(0)})$ is coercive. Note that this proof is very specific to the discrete setting: in the infinite dimensional setting, the Γ-limit is in general degenerate because optimal couplings have typically infinite relative entropy. □

### 3.2.2. Convergence of dual alternate maximization

In this section we introduce the minimization algorithm and prove its convergence. Before that, let us make a small digression and derive a general primal-dual relationship for a class of minimization problems involving Bregman divergences. In this digression, we prove a bound that is crucial for obtaining the convergence of our algorithm.

### Primal-dual relationship with Bregman divergence

Consider a linear operator $A : \mathbb{R}^n \to \mathbb{R}^m$, a function of Legendre type $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and a proper l.s.c. convex function $F$ on $\mathbb{R}^m$ such that $\operatorname{dom} F \cap A(\mathring{\operatorname{dom}}\, h) \neq \emptyset$. Given $z \in \mathring{\operatorname{dom}}\, h$, we define a *minimal divergence problem* as follows

$$\min_{x \in \mathbb{R}^n} F(Ax) + D_h(x|z)\,. \tag{3.2.4}$$

This problem is feasible and strictly convex. If it is moreover coercive, it admits a unique minimizer that we denote $\bar{x}$. The regularized transport problems belong to this class of problems. The dual problem reads (minus)

$$\inf_{y \in \mathbb{R}^m} F^*(-y) + D_h^{*,1}(A^* y|z)\,. \tag{3.2.5}$$

If one assumes the qualification constraint for (3.2.5) holds, i.e. $(\mathring{\operatorname{dom}}\, h^* - \nabla h(z)) \cap A^*(\operatorname{dom} F^*) \neq \emptyset$ then strong duality holds and a pair of optimizers $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ exists and satisfies the primal-dual relationships

$$A\bar{x} \in \partial F^*(-\bar{y}) \quad \text{and} \quad A^*\bar{y} = \nabla h(\bar{x}) - \nabla h(z) \Leftrightarrow \bar{x} = \nabla h^*(A^*\bar{y} + \nabla h(z))$$

The purpose of this paragraph is the following result which gives a general method to build convergent primal iterates from dual iterates that converge to the dual maximum in value. We state it in a general abstract form, but it is an idea that has already been used in the literature [67, Thm. 1] for specific divergences.

**Lemma 3.2.11.** *Let $\bar{E}$ denote the dual functional* (3.2.5) *and assume that the constraint qualification holds. If $\bar{x}$ is the minimizer of* (3.2.4) *and $\bar{y}$ minimizes* (3.2.5) *then for all $y \in \mathbb{R}^m$*

$$D_h(\bar{x}|x_y) \leq E(y) - E(\bar{y}) \quad \textit{where} \quad x_y := \nabla h^*(A^* y + \nabla h(z)).$$

*Proof.* Writing Lemma 3.2.6 with $(u, \bar{u}) = (A^* y, A^* \bar{y})$, one has for all $y \in \mathbb{R}^m$

$$D_h^{*,1}(A^* y|z) - D_h^{*,1}(A^* \bar{y}|z) = \nabla_1 D_h^{*,1}(A^* \bar{y}|z) \cdot A^*(y - \bar{y}) + D_h(\bar{x}|x_y).$$

Moreover, by convexity, one has for any $m \in -\partial F^*(-\bar{y})$

$$F^*(-y) - F^*(-\bar{y}) \geq m \cdot (y - \bar{y}).$$

Summing the two leads to

$$E(y) - E(\bar{y}) \geq (m + A\nabla D_h^*(A^* \bar{y}|z)) \cdot (y - \bar{y}) + D_h(\bar{x}|x_y).$$

But as $\bar{y}$ minimizes (3.2.5) it holds $0 \in \partial E(\bar{y})$. This implies (also because the qualification constraint is assumed) that one can chose $m$ such that $m + A(\nabla D_h^*(A^* \bar{y}|z)) = 0$ and the result follows. □

### Main discrete convergence result

Consider alternate maximization on the dual problem, which consists in sequences $(u^{(\ell)}, v^{(\ell)})$ defined by $v^{(0)} \in \mathbb{R}^J$ given and for $\ell \in \mathbb{N}^*$,

$$u^{(\ell+1)} = \arg\max_u \bar{E}_\varepsilon(u, v^{(\ell)}) \qquad v^{(\ell+1)} = \arg\max_v \bar{E}_\varepsilon(u^{(\ell+1)}, v). \tag{3.2.6}$$

The main result of this section is the following theorem, that gives convergence of the dual objective and the primal iterates at a sublinear rate. We insist on the fact this rate is worse than what is observed in practice; it is just the only rate available to us at this level of generality. The notation $\oplus$ is defined as $(u \oplus v)_{i,j} = (u_i + v_j)$.

**Theorem 3.2.12.** *Assume that the regularized dual problem in Proposition 3.2.9 admits a maximizer and that the iterates* (3.2.6) *are well defined. Then, letting* $(u^{(\ell)}, v^{(\ell)})$ *be the sequence generated by alternate maximization and posing* $\gamma^{(\ell)} := \nabla h^*((u^{(\ell)} \oplus v^{(\ell)})/\varepsilon + \nabla h(K))$ *the corresponding primal sequence, it holds for* $\ell \geq 3$*,*

$$\varepsilon D_h(\gamma_\varepsilon | \gamma^{(\ell)}) \leq (\sup E_\varepsilon) - E_\varepsilon(u^{(\ell)}, v^{(\ell)}) \leq O(1/\ell).$$

The constant in the notation $O(\cdot)$ is given in Theorem 3.2.13. For the proof, we mainly use Lemma 3.2.11 and the following result from [10] of which we need a slightly stronger version to fit our framework.

**Theorem 3.2.13.** *Let* $F(u,v) := f(u,v) + g_1(u) + g_2(v)$ *be a convex, proper, l.s.c. function on* $\mathbb{R}^I \times \mathbb{R}^J$ *with* $f$ *continuously differentiable on* $\operatorname{dom} g_1 \times \operatorname{dom} g_2$*, of partial gradients* $\nabla_u f$ *and* $\nabla_v f$ *respectively* $L_1$ *and* $L_2$ *Lipschitz. If the set of minimizers* $A$ *of* $F$ *is nonempty, then the sequence generated by alternate minimization on* $F$ *starting from* $(v^{(0)}) \in \mathbb{R}^J$*, if well-defined, satisfies for* $\ell \geq 3$*,*

$$F(u^{(\ell)}, v^{(\ell)}) - \min F \leq \max\left\{ \left(\frac{1}{2}\right)^{\frac{\ell-2}{2}} (F(u^{(1)}, v^{(0)}) - \min F),\ \frac{8\min\{L_1, L_2\}R^2}{\ell - 2} \right\}$$

*where* $R^2 := \max_{(u,v)} \left\{ \min_{(u^*,v^*)\in A} \|(u,v) - (u^*,v^*)\|^2 \ ;\ F(u,v) \leq F(u^{(1)}, v^{(0)}) \right\}$.

*Proof.* This is [10, Thm. 3.9] where the assumption on the compactness of the set of minimizers is removed with the following remark (see the cited article for reference): [10, Lem. 3.4] holds for any minimizer $x^*$ so one may as well choose $x^*$ as the projection of $x_k$ on the set of minimizers. It follows that one may replace the definition of $R$ in [10, eq. (3.8)] by the maximum distance to the set of minimizers and the proof of [10, Lem. 3.5] goes through. □

*Proof of theorem 3.2.12.* The function $E_\varepsilon$ satisfies the hypothesis of Theorem 3.2.13 so we deduce the second inequality. The first inequality is given by Lemma 3.2.11. □

## 3.3. Discrete scaling algorithm

### 3.3.1. Entropic regularization: discrete case

In the previous section, we have shown that by adding a small regularization term, one could find an approximate minimizer of the generic formulation of coupling problems, by performing an alternate maximization algorithm on the dual problem. Yet, this procedure takes all its strength when the Bregman regularization is chosen precisely as the relative entropy. In this case, one obtains simple iterates that have a low per-iteration complexity for many cases of interest.

Additionally, the framework of entropic regularization admits a natural definition in the continuous setting where one recovers (a generalization of) the well studied minimal entropy problems. In particular and important cases, some dimension free convergence rates can be obtained. These infinite dimensional questions are treated separately in Section 3.5 and here we focus on practical implementation.

Let us instantiate the previous formula in the case of entropic regularization. Consider the separable entropy function on $\mathbb{R}^n$

$$h(x) := \sum \bar{h}(x_i) \quad \text{where} \quad \bar{h}(x_i) := \begin{cases} x_i \log x_i - x_i + 1 & \text{if } x_i > 0 \\ 1 & \text{if } x_i = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{3.3.1}$$

It is of Legendre type since it is strictly convex on the (convex) domain of its subdifferential $]0,\infty[^n$ and its convex conjugate $h^* : y \mapsto \sum(e^{y_i} - 1)$ is strictly convex on $\mathbb{R}^n$. As a side note, one has that $(\nabla h, \nabla h^*) = (\sum \log, \sum \exp)$ form a pair of inverse homeomorphisms, as expected.

**Definition 3.3.1** (Discrete relative entropy)**.** *The* relative entropy $H := D_h$ *is the Bregman divergence associated to the Legendre function defined in* (3.3.1)*. Its explicit expression for* $x, y \in \mathbb{R}^n$ *is* $H(x|y) = \sum \bar{H}(x_i|y_i)$ *where*

$$\bar{H}(x_i|y_i) := \begin{cases} x_i \log(x_i/y_i) - x_i + y_i & \text{if } x_i > 0 \text{ and } y_i > 0 \\ y_i & \text{if } x_i = 0 \text{ and } y_i \geq 0 \\ \infty & \text{otherwise,} \end{cases}$$

*and the value for* $x_i = y_i = 0$ *is obtained by l.s.c. extension.*

As explained in Remark 3.2.5, $H$ is both the Bregman divergence, and the $f$-divergence associated to $h$ (see Definition 3.5.1: for the $f$-divergence interpretation, one may see $(x_i)$ and $(y_i)$ as the list of masses of atomic measures). Now choose a reference $\gamma^{(0)} \in ]0,\infty[^{I \times J}$ and define the kernel $K_\varepsilon \in \mathbb{R}_+^{I \times J}$ by

$$(K_\varepsilon)_{i,j} = \exp(-c_{i,j}/\varepsilon)\gamma^{(0)}_{i,j}.$$

for $(i,j) \in I \times J$. The entropic regularization of the generic formulation is an instantiation of (3.2.3) that writes $\min_{\gamma \in \mathbb{R}^{I \times J}} E_\varepsilon(\gamma)$ where

$$E_\varepsilon(\gamma) = F_1(\sum_{j \in J} \gamma_{i,j}) + F_2(\sum_{i \in I} \gamma_{i,j}) + \varepsilon H(\gamma | K_\varepsilon) \tag{3.3.2}$$

and its dual is $\sup_{(u,v)\in\mathbb{R}^I\times R^J}\bar{E}_\varepsilon(u,v)$ where

$$\bar{E}_\varepsilon(u,v) = -F_1^*(-u) - F_2^*(-v) - \varepsilon\sum_{i,j}(\exp((u_i+v_j-c_{i,j})/\varepsilon)-1)\gamma_{i,j}^{(0)}. \tag{3.3.3}$$

### 3.3.2. Scaling algorithm

An alternate maximization step on $u$ given $v^{(\ell)}$ amounts to finding a maximizer $u^{(\ell+1)}$ of

$$\max_{u\in\mathbb{R}^I} -F_1^*(-u) - \varepsilon\sum_i(e^{u_i/\varepsilon}-1)\sum_j K_{i,j}e^{v_j^{(\ell)}/\varepsilon}$$

or equivalently, solve the dual (the qualification constraint are easily checked):

$$\min_{s\in\mathbb{R}^I} F_1(s) + \varepsilon H(s|(\sum_j K_{i,j}e^{v_j^{(\ell)}/\varepsilon})_i)$$

which is a proximal step with respect to the relative entropy. The primal-dual relationship are, at optimality, $s_i = e^{u_i^{(\ell+1)}/\varepsilon}(\sum_j K_{i,j}e^{v_j^{(\ell)}/\varepsilon})$. Making the change of variables $(a,b) := ((e^{u_i/\varepsilon})_i, (e^{v_j/\varepsilon})_j)$ we remark that the primal-dual relationship can be written $s_i = a^{(\ell+1)} \oslash K(b^{(\ell)})$ where $K(\cdot)$ is a matrix/vector product and $\oslash$ is a pointwise (or elementwise) division with the convention $0/0 = 0$. Thus, in these new variables, the maximization step amounts to setting

$$a^{(\ell+1)} = \mathrm{prox}^H_{F_1/\varepsilon}(K(b^{(\ell)})) \oslash K(b^{(\ell)})$$

where we denote $\mathrm{prox}^H_F(\bar{s}) = \arg\min_s\{F(s)+H(s|\bar{s})\}$. Similar computations for the other partial maximization lead to the definition of the following sequence.

**Definition 3.3.2** (Discrete scaling iterations)**.** *Given a generic problem (i.e. functions $F_1$, $F_2$ and a cost $c$) and a regularization parameter $\varepsilon > 0$, the discrete scaling iterations are the sequence $(a^{(\ell)}, b^{(\ell)})_{\ell\in\mathbb{N}^*}$ defined as $b^{(0)} = 1_J \in \mathbb{R}^J$ and*

$$a^{(\ell)} = \mathrm{prox}^H_{F_1/\varepsilon}(K(b^{(\ell-1)})) \oslash (K(b^{(\ell-1)})),$$
$$b^{(\ell+1)} = \mathrm{prox}^H_{F_2/\varepsilon}(K^T(a^{(\ell)})) \oslash (K^T(a^{(\ell)})).$$

*The corresponding primal sequence (see Theorem 3.2.13) is $(\gamma^{(\ell)})$ defined for all $\ell\in\mathbb{N}^*$ as*

$$\gamma_{i,j}^{(\ell)} = a_i^{(\ell)} K_{i,j} b_j^{(\ell)}.$$

**Remark 3.3.3.**

*(i) note that the scaling variables are of dimensions $I$ and $J$ while the dimension of the original unknown was $I\times J$. Now, the only object of dimensions $I\times J$ is $K$ and it only appears through a matrix/vector product, which is fast operation in modern software. It can even be replaced by faster operations in some cases, see Section 3.3.3;*

*(ii) the operator $s\mapsto \mathrm{prox}^H_{F_i/\varepsilon}(s)\oslash s$ has closed form or is easy to compute in many important cases that we review later in Chapter 5;*

*(iii) in the case of regularized optimal transport, the functions $F_i$ are the indicator of the marginal constraints and one recovers the standard Sinkhorn iterations a.k.a. IPFP or iterative matrix scaling. We choose to maintain the name* scaling *because the primal iterates $\gamma^{(\ell)}$ are obtained from $K$ by scaling the rows and columns with two scaling vectors.*

The following result is an instantiation of Theorem 3.2.13. It gives a very general convergence result but the provided sublinear convergence rate is much slower than what observed in practice. Linear convergence rate will be proven for specific cases in Section 3.5.

**Theorem 3.3.4** (Convergence of the scaling iterations)**.** *Let $\gamma^{(\ell)}$ be the primal iterate from Definition 3.3.2. If the dual problem* (3.3.3) *admits a maximizer then, denoting $\gamma_\varepsilon$ the minimizer of* (3.3.2) *and $\bar{E}_\varepsilon$ the dual objective functional* (3.3.3)*, it holds*

$$H(\gamma_\varepsilon|\gamma^{(\ell)}) \leq (\max \bar{E}_\varepsilon) - \bar{E}_\varepsilon(\varepsilon \log a^{(\ell)}, \varepsilon \log b^{(\ell)}) = O(1/\ell).$$

### 3.3.3. Stability and practical implementation

#### Stabilization

For small values of $\varepsilon$ the entries of the matrix $K$, $a^{(\ell)}$ and $b^{(\ell)}$ may become both very small and very large, leading to numerically imprecise values of their mutual products and overflow of the numerical range. An easy fix would be to execute the algorithm in the log domain and using stabilized versions of the operations. However, this is not entirely satisfying because then the computation of $K(b^{(\ell)})$ and $K(a^{(\ell)})$ are not simple matrix/vector product anymore but so-called "$\log\sum\exp$" operations, and the algorithm is considerably slowed down. We suggest a middle way, using a redundant parametrization of the iterates as follows:

$$a^{(\ell)} = \tilde{a}^{(\ell)} \odot \exp(\tilde{u}^{(\ell)}/\varepsilon), \qquad b^{(\ell)} = \tilde{b}^{(\ell)} \odot \exp(\tilde{v}^{(\ell)}/\varepsilon). \tag{3.3.4}$$

The idea is to keep $\tilde{a}^{(\ell)}$ and $\tilde{b}^{(\ell)}$ close to 1 and to absorb the extreme values of $a^{(\ell)}$ and $b^{(\ell)}$ into the log-domain via $\tilde{u}^{(\ell)}$ and $\tilde{v}^{(\ell)}$ from time to time.

Consider the stabilized kernels $\tilde{K}^{(\ell)}$ whose entries are given by

$$\tilde{K}^{(\ell)}_{i,j} = \exp((\tilde{u}^{(\ell)}_i + \tilde{v}^{(\ell)}_j - c_{i,j})/\varepsilon)$$

and remark that it holds, by direct computations,

$$K(b^{(\ell)}) = e^{-\tilde{u}^{(\ell)}/\varepsilon} \odot \tilde{K}^{(\ell)}(\tilde{b}^{(\ell)}), \qquad K^T(a^{(\ell)}) = e^{-\tilde{v}^{(\ell)}/\varepsilon} \odot (\tilde{K}^{(\ell)})^T(\tilde{a}^{(\ell)}).$$

The scaling iterates of Definition 3.3.2 then read

$$\tilde{a}^{(\ell+1)} = \mathrm{prox}^H_{F_1/\varepsilon}(e^{-\tilde{u}^{(\ell)}/\varepsilon} \odot s) \oslash s \quad \text{where} \quad s := \tilde{K}^{(\ell)}(\tilde{b}^{(\ell)})$$

with a similar formula for $\tilde{b}^{(\ell+1)}$. For computing these iterates, all the information we need about $F_1$ and $F_2$ can be condensed by the specification of functions $\mathrm{proxdiv}_{F_1} : \mathbb{R}^I \times \mathbb{R}^I \times \mathbb{R}_+ \to \mathbb{R}^I$ and $\mathrm{proxdiv}_{F_2} : \mathbb{R}^J \times \mathbb{R}^J \times \mathbb{R}_+ \to \mathbb{R}^J$ defined as

$$\mathrm{proxdiv}_{F_i} : (s,u,\varepsilon) \mapsto \mathrm{prox}^H_{F_i/\varepsilon}(e^{-u/\varepsilon} \odot s) \oslash s. \tag{3.3.5}$$

which have closed form, or are simple to estimate in many important cases (reviewed in Chapter 5). The variable $u$ is an offset (in logarithm domain) that appears because of the stabilization procedure, but can be set to 0 for the basic version of the algorithm. The main numerical algorithm in pseudo code is displayed in Algorithm 1 for the basic version and Algorithm 2 for the stabilized version.

**Remark 3.3.5.** *We may summarize the properties of the stabilization as follows:*

- *let aside the updates of the kernel $\tilde{K}^{(\ell)}$ which are only performed a few times, the stabilized scaling updates on $\tilde{a}^{(\ell)}$ and $\tilde{b}^{(\ell)}$ have the same complexity as the original scaling updates;*
- *the operator* proxdiv *is* guaranteed *to be stable whenever the algorithm converges. More precisely, for any $M > 1$, there exists $\ell_0 \in \mathbb{N}$ such that if an absorption step is performed at iterate $\ell_0$ by setting $(\tilde{u}^{(\ell)}, \tilde{v}^{(\ell)}) = (\varepsilon \log a^{(\ell)}, \varepsilon \log b^{(\ell)})$, then for all $\ell > \ell_0$, one has $1/M < \tilde{a}_i^{(\ell)}, \tilde{b}_j^{(\ell)} < M$ for all components $i, j$, as a consequence of the convergence of the algorithm.*

**Algorithm 1** Basic scaling algorithm. The function "proxdiv" is defined in (3.3.5).

1. initialize the kernel $K = (e^{-c_{i,j}/\varepsilon} \gamma_{i,j}^{(0)})_{i,j}$ and the iterates $(a,b) = (1_I, 1_J)$
2. while stopping criterion not satisfied repeat:
   a) $a \leftarrow \text{proxdiv}_{F_1}(K(b), 0_I, \varepsilon)$
   b) $b \leftarrow \text{proxdiv}_{F_2}(K^T(a), 0_J, \varepsilon)$
3. return $\gamma = (a_i K_{i,j} b_j)_{i,j}$

### Computing matrix multiplications

The size of the matrix $K$ is $I \times J$ so the matrix-vector multiplication step or even merely storing $K$ in memory can quickly become intractable as the sizes of $X$ and $Y$ increase. For special problems it is possible to avoid dense matrix multiplication (and storage). For instance, when $c(x,y) = |x-y|^2$ is the squared Euclidean distance on uniform Cartesian grids $(x_{ii'}), (y_{jj'})$ in $\mathbb{R}^2$ (or more generally $\mathbb{R}^d$) with step size $h > 0$, then $K$ is the separable Gaussian kernel

$$K_{ii'jj'} = \exp((|i'-i|^2 + |j'-j|^2)h^2/\varepsilon) = K^{(1)}_{ii'} K^{(2)}_{jj'}.$$

Then, multiplying by $K$ can be done by successive 1-D convolutions. For more general geometric surfaces, $K$ can also be approximated by the heat kernel [152]. These methods however cannot be directly combined with the stabilization procedure that we propose since the "stabilized kernels" $\tilde{K}$ lose that separable structure.

**Algorithm 2** Scaling algorithm with stabilization

1. initialize kernel $\tilde{K} = (e^{-c_{i,j}/\varepsilon}\gamma_{i,j}^{(0)})_{i,j}$
2. initialize the iterates $(\tilde{a},\tilde{b}) = (1_I, 1_J)$, $(u,v) = (0_I, 0_J)$
3. while stopping criterion not satisfied repeat:
   a) $\tilde{a} \leftarrow \text{proxdiv}_{F_1}(\tilde{K}(\tilde{b}), u, \varepsilon)$
   b) $\tilde{b} \leftarrow \text{proxdiv}_{F_2}(\tilde{K}^T(\tilde{a}), v, \varepsilon)$
   c) if $\max\{\|\log\tilde{a}\|_\infty, \|\log\tilde{b}\|_\infty\} >$ threshold, then absorption step:
      i. $(u,v) \leftarrow (u + \varepsilon\log\tilde{a}, v + \varepsilon\log\tilde{b})$
      ii. (optionally decrease $\varepsilon$)
      iii. $\tilde{K}_{i,j} \leftarrow \exp((u_i + v_j - c_{i,j})/\varepsilon) \cdot \gamma_{i,j}^{(0)}$
      iv. $\tilde{b} \leftarrow 1_J$
4. return $\gamma = (\tilde{a}_i \tilde{K}_{i,j} \tilde{b}_j)_{i,j}$

### Gradually decreasing $\varepsilon$ and other tricks

When running Algorithms 1 or 2 for a very small regularization $\varepsilon$, most of the entries of $K = (e^{-c_{i,j}/\varepsilon})_{i,j}$ are below machine precision at initialization. A simple workaround it to first "estimate" the dual variables $(u,v)$ by performing several iterations with higher values of $\varepsilon$. Reduction of $\varepsilon$ should be performed at step 3(c)ii in Algorithm 2. Indeed, before that step, $(\tilde{u},\tilde{v})$ are approximations of the optimal dual variable of the regularized problem so one can change $\varepsilon$ and start solving the new problem with $(u,v)$ as a clever starting point using the stabilized kernel $\tilde{K} = (e^{(\tilde{u}_i + \tilde{v}_j - c_{i,j})/\varepsilon})_{i,j}$. When $\varepsilon$ decreases, many entries of $\tilde{K}$ are below numerical precision but we may expect that this is not the case for the entries where the optimal regularized plan takes high values. Besides, this sparsity property suggests tricks for solving efficiently large problems, such as (i) to transfer the sparsity pattern of $\tilde{K}$ from high values of $\varepsilon$ to smaller ones, and (ii) when the problem has some structure (defined on a regular grid, with a regular cost $c$), to "approximate" the dual variables on coarser grids and then transfer the resulting sparsity pattern to the original, finer, grid. These tricks are useful in practice, but one should keep in mind that they yield instabilities for small $\varepsilon$ because

(i) the optimal dual potentials change with $\varepsilon$ and so does the sparsity pattern;

(ii) when interpolating the potentials $(\tilde{u},\tilde{v})$ on the finer grid, the error might be sufficient to deduce a wrong sparsity pattern, especially where the cost $c$ varies rapidly.

The heuristic we followed in our numerical experiments in Chapter 5, when mentioned, is as follows: starting from $\varepsilon = 1$, after every 100 iterations we perform an absorption step and divide

$\varepsilon$ by a factor chosen so that the final value $\varepsilon$ is reached after 10 divisions. Once the final value of $\varepsilon$ is reached, Algorithm 2 is run normally until the desired convergence criterion is met.

While our purpose when doing this gradual decreasing of $\varepsilon$ is merely to find a smart initialization of the dual variables, this idea has been proposed several times in relation to standard optimal transport problems. It has been shown that, with a sufficiently slow decrease, this asymptotically solves the unregularized optimal transport problem in [150] (but $\varepsilon$ should be of the order $1/\log(\ell)$ which is too slow to be of practical interest). Heuristically, it has been proposed to accelerate the convergence of Sinkhorn's iterations, in the same fashion as for interior point methods, but theoretical justification is an open problem [95, 145]. Numerical tricks to decrease the complexity of Algorithm 2 are developped further in [145]. See also [144, 119] for multiscale methods on the linear programming approach for optimal transport.

#### Multiple couplings

The general optimization problem (3.1.1) involved $n$ couplings, with potentially $n > 1$. For simplicity, throughout Section 3.2, we focussed on the case $n = 1$. However, the extension to $n > 1$ is rather simple. In this case, the variables $a, b, u, v$ of the algorithm lie in $\mathbb{R}^{n\times I}$ or $\mathbb{R}^{n\times J}$, the kernel $K$ is a family of $n$ matrices of size $I \times J$ , the entrywise operations (multiplication, division) are still performed entrywise and the matrix vector multiplications are performed "coupling by coupling", e.g. for $k \in \{1,\dots,n\}$ and $j \in \{1,\dots,J\}$:

$$(K(b))_{k,i} = \sum_j K_{k,i,j}\, b_{k,j}\,.$$

### 3.3.4. Generalization to pushforward operators

Scaling algorithms similar to Algorithm 1 can be formulated for solving problems of more general form than the entropic regularization of the discrete "generic problem" (3.3.2). This form has been chosen because it encompasses all the cases of interest in this thesis and allows for a unified treatment. But in general, there can be more than 2 functionals, more than 2 spaces involved and the projection operators $\pi^x_\#$ and $\pi^y_\#$ can be replaced by pushforwards of functions $t$ which are not necessarily projections (i.e. not of the form $t(x,y) = x$).

Several examples of such extensions can be found in [14] for the special case of classical optimal transport. Let us sketch this extension in the discrete setting and for the case of $n = 1$ (i.e. one unknown coupling) and give the *scaling form* of the alternate maximization on the dual and an example, without proof. It should be noted that, while our convergence proof extends verbatim to general pushforward operators (we only use the fact that they are linear), the behavior of alternate maximization for more than 2 blocs is more intricate than the case of 2 blocs (a good reference is [158]).

Let $(X^k)_{k=1}^N$ and $Z$ be finite discrete spaces of cardinalities $(I_k)_k$ and $L$, respectively and let $t^k : Z \to X_k$ be a family of $N$ surjective maps (that act on the indices, in our notations). The space $Z$ plays the role of $X \times Y$ in the previous discussions, but in this generalization the structure of a product space is lost. Given $k \in \{1,\dots,N\}$, $\gamma \in \mathbb{R}^L$ and $u \in \mathbb{R}^{I_k}$ the pushforward operator $t^k_\#$ and

its adjoint $(t^k_\#)^*$ read

$$[t^k_\# \gamma]_i = \sum_{l \in (t^k)^{-1}(i)} \gamma_l \quad \text{and} \quad [(t^k_\#)^* u]_l = u_{t^k(l)}.$$

Given a vector $K \in \mathbb{R}^L_+$ corresponding to the kernel, and $N$ convex, proper, l.s.c. functions $F_k : \mathbb{R}^{I_k} \to \mathbb{R} \cup \{\infty\}$, the regularized primal (generalizing (3.3.2)) is

$$\min_{\gamma \in \mathbb{R}^L} \sum_{k=1}^N F_k(t^k_\# \gamma) + \varepsilon H(\gamma | K) \tag{3.3.6}$$

and the dual reads, up to a constant,

$$\sup_{(u^k)_{k=1}^N \in \mathbb{R}^{\Sigma_k I_k}} -\sum_{k=1}^N F_k^*(-u^k) - \varepsilon \sum_{l=1}^L \exp\left(\tfrac{1}{\varepsilon} \textstyle\sum_{k=1}^N u^k_{t^k(l)}\right) \cdot K_l \tag{3.3.7}$$

Now for $(a^n)_{n=1}^N \in \mathbb{R}_+^{\sum_n I_n}$ and $k \in \{1,\dots,N\}$, define the operator $\mathscr{K}^k$ as

$$[\mathscr{K}^k((a^n)_{n \neq k})]_i := \sum_{l \in (t^k)^{-1}(i)} \Big(\prod_{n \neq k} a^n_{t^n(l)}\Big) \cdot K_l \tag{3.3.8}$$

These operators allow to compute the rightmost term of (3.3.7) in a "marginalized" way, using the relation

$$\langle a^k, \mathscr{K}^k((a^n)_{n \neq k}) \rangle = \sum_{l=1}^L \Big(\prod_{n=1}^N a^n_{t^n(l)}\Big) \cdot K_l$$

valid for $k \in \{1,\dots,N\}$. The key feature for obtaining this relation, and the reason why scaling algorithms generalizes naturally to linear operators which are pushforward maps, is the fact that for each $k$, the pre-images of $t^k$, namely $((t^k)^{-1}(i))_{i=1}^{I_k}$, form a partition of $Z$. We can now define the *generalized scaling* algorithm displayed in Algorithm 3. As Algorithm 1, it corresponds to alternate maximization on the dual problem written in dual form. Instead of rescaling the rows and columns of the kernel $K$, it rescales the entries of $K$ corresponding to the pre-images of the maps $t^k$.

**Algorithm 3** Generalized scaling algorithm ($\mathscr{K}$ is defined in (3.3.8))

1. initialize $(a^1,\dots,a^N) = (1_{I_1},\dots,1_{I_N})$
2. while stopping criterion not satisfied, repeat:
   a) for $k = 1,\dots,N$, do:
      i. $a^k \leftarrow \mathrm{proxdiv}(\mathscr{K}^k((a^n)_{n \neq k}), 0, \varepsilon)$
3. return the primal minimizer $(K_l \cdot \prod_{k=1}^N a^k_{t^k(l)})_{l=1}^L$

As a simple illustration, consider, in the setting of a product space $Z = X \times Y$, an extension of the generic problem where is added a function of the total mass

$$\min_{\gamma \in \mathbb{R}^{I \times J}} F_1(\sum_{j \in J} \gamma_{i,j}) + F_2(\sum_{i \in I} \gamma_{i,j}) + F_3(\sum_{i,j} \gamma_{i,j}) + \varepsilon H(\gamma | K)$$

where the pushforwards involved are the marginal projections and the total mass and $F_3 : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is a proper, l.s.c. and convex function. Applying the reasoning above, and after some rearrangement (here $Z$ is still a product space and thus it is convenient to interpret $\gamma$ as a matrix), one obtains Algorithm 4. If, for instance, $F_3$ is chosen as indicator of equality with a positive real number, this algorithm solves the optimal partial transport problem (in the "total mass" parametrization, see Section 2.3.2).

**Algorithm 4** Scaling algorithm with a function on the total mass

1. initialize kernel $K$ and variables $b = 1_J$ and $z = 1 \in \mathbb{R}$
2. while stopping criterion not satisfied, repeat:
   a) $a \leftarrow \text{proxdiv}_{F_1}(z \cdot K(b))$
   b) $b \leftarrow \text{proxdiv}_{F_2}(z \cdot K^T(a))$
   c) $z \leftarrow \text{proxdiv}_{F_3}(a^T K(b))$
3. return the primal minimizer $z \cdot (a_i K_{i,j} b_j)_{i,j}$

## 3.4. Acceleration of convergence rate

In this section, we propose to apply an acceleration method with a strong practical impact. We show that a minor modification of alternate minimization procedure with the same per-iteration complexity, known as *successive over-relaxation*, allows to improve the local convergence rate. We also perform numerical experiments that show that the global convergence rate is greatly improved in the case of Sinkhorn's iterations (which are a particular case of the scaling algorithms). This is in contrast to other local acceleration methods that typically fail to converge globally. However, proving a practically useful global convergence guarantee is left as an open question.

### 3.4.1. Successive over-relaxation

Let us consider the following modification of the alternate minimization algorithm, for minimizing a function $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

**Definition 3.4.1** (Successive over-relaxation)**.** *Let $\theta \in \mathbb{R}$ be an extrapolation parameter. The successive over-relaxation method (SOR) defines the sequence $(u^{(\ell)}, v^{(\ell)})_{\ell \in \mathbb{N}}$ defined by $(u^{(0)}, v^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^m$ and for $\ell \in \mathbb{N}^*$*

$$\begin{aligned}
\tilde{u}^{(\ell)} &= \operatorname*{argmin}_{u \in \mathbb{R}^n} F(u, v^{(\ell-1)}) \\
u^{(\ell)} &= u^{(\ell-1)} + \theta \left( \tilde{u}^{(\ell)} - u^{(\ell-1)} \right) \\
\tilde{v}^{(\ell)} &= \operatorname*{argmin}_{v \in \mathbb{R}^m} F(u^{(\ell)}, v) \\
v^{(\ell)} &= v^{(\ell-1)} + \theta \left( \tilde{v}^{(\ell)} - v^{(\ell-1)} \right).
\end{aligned}$$

So at each (half) step, one moves to a new point which is a linear combination of the previous partial coordinate and the new partial minimizer. One can readily check the following result.

**Lemma 3.4.2** (Fixed-points)**.** *If $\theta \neq 0$, the fixed points of the successive over-relaxation (Definition 3.4.1) are the same as the fixed points of the alternate minimization, recovered for $\theta = 1$.*

Relaxations of this kind are common for the acceleration of fixed points. The very algorithm of Definition 3.4.1 is presented for convex minimization in [7] and in [169], an extensive analysis of this method for solving linear systems is proposed. While this acceleration is well known for numerical linear algebra, it seems that this method has not been proposed yet to accelerate the computation of entropy regularized optimal transport. Applied to the scaling algorithms, this gives updates where the additive extrapolation transforms into a multiplicative one due to the exponential change of variables, as diplayed in Algorithm 5. We also display a specialized version for entropy regularized optimal transport in Algorithm 6, it defines a variant of Sinkhorn algorithm which is very simple to implement and that can be order of magnitude faster (see Table 3.1).

**Algorithm 5** Extrapolated scaling algorithm (acceleration parameter $\theta \in [1,2[$)

1. initialize the kernel $K$ and the iterates $(a,b) = (1_I, 1_J)$
2. while stopping criterion not satisfied repeat:
   a) $a \leftarrow a^{1-\theta} \odot \left[\mathrm{proxdiv}_{F_1}(K(b),0,\varepsilon)\right]^{\theta}$
   b) $b \leftarrow b^{1-\theta} \odot \left[\mathrm{proxdiv}_{F_2}(K^T(a),0,\varepsilon)\right]^{\theta}$
3. return $\gamma = (a_i K_{i,j} b_j)_{i,j}$

**Algorithm 6** Extrapolated Sinkhorn's algorithm (acceleration parameter $\theta \in [1,2[$)

1. initialize the kernel $K$ and the iterates $(a,b) = (1_I, 1_J)$
2. while stopping criterion not satisfied repeat:
   a) $a \leftarrow a^{1-\theta} \odot [\mu \oslash K(b)]^{\theta}$
   b) $b \leftarrow b^{1-\theta} \odot \left[\nu \oslash K^T(a)\right]^{\theta}$
3. return $\gamma = (a_i K_{i,j} b_j)_{i,j}$

### 3.4.2. Acceleration of local convergence

Let us first consider the minimization of a quadratic function, with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{n\times m}$

$$F_Q(u,v) = \frac{1}{2}u^T A u + \frac{1}{2}v^T B v + v^T C u \tag{3.4.1}$$

which is positive definite, i.e. the block matrix $H = \left(\begin{array}{c|c} A & C^T \\ \hline C & B \end{array}\right)$ is a positive definite symmetric matrix $H \succ 0$.

**Lemma 3.4.3** (Local linear convergence of alternate minimization)**.** *The alternate minimization on $F_Q$ (also known as Gauss-Seidel algorithm) starting from $u^{(0)} \in \mathbb{R}^n$ corresponds to iteratively applying the operator $T := A^{-1}C^T B^{-1}C$, i.e. $u^{(\ell)} = T^{\ell}(u^{(0)})$. The matrix $T \in \mathbb{R}^{n\times n}$ is diagonalizable in $\mathbb{R}$ and all its eigenvalues belong to $[0,1[$. In particular, alternate minimization converges linearly with the rate $\|T\|_{\infty}^{\ell} < 1$, the spectral norm of $T$.*

*Proof.* By the first order optimality conditions, one has $Au^{(\ell)} + C^T v^{(\ell-1)} = 0$ and $Bv^{(\ell)} + Cu^{(\ell)} = 0$, which gives $u^{(\ell)} = T(u^{(\ell-1)})$. Now let $\tilde{B} := C^T B^{-1} C$. Since $B^{-1} \succ 0$, one has $x^T \tilde{B} x = \|Cx\|_{B^{-1}}^2 \geq 0$ for all $x \in \mathbb{R}^n$ so $\tilde{B}$ is a symmetric, positive semi-definite matrix. Also, since $H$ is positive definite, the Schur complement $A - C^T B^{-1} C = A - \tilde{B}$ is symmetric positive definite so we deduce $\tilde{B} \prec A$. It follows that $T = A^{-1}\tilde{B} = A^{-1/2}(A^{-1/2}\tilde{B}A^{-1/2})A^{1/2}$ is diagonalizable in $\mathbb{R}$ because it is similar to $A^{-1/2}\tilde{B}A^{-1/2}$ which is symmetric. To conclude, let $\mu \in \mathbb{R}$ be an

eigenvalue of $T = A^{-1}\tilde{B}$ and $x \in \mathbb{R}^n$ such that $A^{-1}\tilde{B}x = \mu x$. Computing the scalar product with $Ax$ yields $x^T\tilde{B}x = \mu x^T Ax$. It follows, by the inequality $0 \preceq \tilde{B} \prec A$, that $0 \leq \mu < 1$. □

The spectral analysis on this quadratic case shows an acceleration of the order of the optimal methods. Such analysis is standard for the SOR method [169], but I include it in this thesis for the sake of completeness and because I was unable to find a reference for the specific case of convex minimization.

**Theorem 3.4.4.** *Denote $1-\eta$ the maximal eigenvalue of $T$ (by Lemma 3.4.3, $0 < \eta < 1$). Then, for the optimal choice of parameter $\theta = 2/(1+\sqrt{\eta})$, the asymptotical linear convergence rate of the extrapolated alternate minimization (Definition 3.4.1) on the quadratic function* (3.4.1) *is*

$$r = \frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}.$$

*This means that asymptotically, $\|(u^{(\ell)}, v^{(\ell)})\| = O(r^\ell)$.*

**Remark 3.4.5.** *For $\eta$ close to 0, an optimal solution with precision $\varepsilon > 0$ is found after an order of $-\log\varepsilon/\eta$ iterations with the alternate minimization, and $-\log\varepsilon/\sqrt{\eta}$ iterations for the accelerated version with the optimal parameter.*

*Proof.* By writing the first order optimality conditions of the partial minimization, one finds the explicit form of the updates

$$\begin{aligned} u^{(\ell+1)} &= (1-\theta)u^{(\ell)} - \theta A^{-1}C^T v^{(\ell)} \\ v^{(\ell+1)} &= (1-\theta)v^{(\ell)} - \theta B^{-1}C^T u^{(\ell+1)}. \end{aligned}$$

One iteration is equivalent to applying an endomorphism $M_\theta : \mathbb{R}^n \times \mathbb{R}^m$ to $(u^{(\ell)}, v^{(\ell)}) \in \mathbb{R}^n \times \mathbb{R}^m$: let us study the spectral properties of $M_\theta$. Let $(u,v)$ be (if it exists) a complex eigenvector of $M_\theta$ with eigenvalue $\lambda \in \mathbb{C}$. Replacing in the iteration above, we see that $(u,v)$ satisfies

$$(\lambda+\theta-1)u = -\theta A^{-1}C^T v \qquad\qquad (\lambda+\theta-1)v = -\lambda\theta B^{-1}Cu$$

and combining these two equalities, it holds in particular that $u$ is an eigenvector of $T = A^{-1}C^T B^{-1}C$ (the matrix from Lemma 3.4.3) with eigenvalue $\mu = \frac{(\lambda+\theta-1)^2}{\lambda\theta^2}$, i.e. $\lambda$ solves the equation

$$X^2 - 2(\frac{1}{2}\theta^2\mu + 1 - \theta)X + (\theta-1)^2 = 0. \tag{3.4.2}$$

of unknown $X \in \mathbb{C}$. The solutions to (3.4.2) are of the form

$$\lambda = b + \delta^{1/2} \quad \text{where} \quad \begin{cases} b = \frac{1}{2}\theta^2\mu + 1 - \theta \\ \delta = \theta^2\mu(1 - \theta + \frac{1}{4}\theta^2\mu). \end{cases}$$

and where $\delta^{1/2}$ is a (positive, negative or purely imaginary) square root of $\delta$.

A careful treatment of the several cases depending on the signs of $b$ and $\delta$ leads to the following expression, denoted $f(\theta,\mu)$, for the maximum modulus of the eigenvalue associated to $\mu$ and acceleration parameter $\theta$

$$f(\theta,\mu) = \begin{cases} \theta - 1 & \text{if } \theta^2\mu - 4(\theta-1) \leq 0 \\ \frac{1}{2}\theta^2\mu - (\theta-1) + \frac{1}{2}\sqrt{\mu\theta^2(\theta^2\mu - 4(\theta-1))} & \text{otherwise.} \end{cases}$$

This function is plotted on Figure 3.2. For a fixed $\mu$, it reaches a minimum for $\delta = 0$ and for fixed $\theta$, it is an increasing function of $\mu$, so, if $\mu_{\max}$ is the maximum eigenvalue of $T$, then $f(\theta,\mu_{\max})$ is the maximum modulus of the eigenvalues of $M_\theta$ and it is minimized for $\delta = 0$ i.e. (taking the only solution smaller than 2):

$$\bar{\theta} = \frac{2}{\mu_{\max}}(1 - \sqrt{1-\mu_{\max}}) = \frac{2}{1+\sqrt{\eta}}$$

where we have introduced $\eta = 1 - \mu_{\max}$ which is typically small. With this choice of $\theta$, we thus guarantee that all the eigenvalues of $M_\theta$ are smaller than $\theta - 1 = (1-\sqrt{\eta})/(1+\sqrt{\eta})$, so $\theta - 1$ is the spectral radius of $M_\theta$. The asymptotical rate is then given by Gelfand's formula, stating that the spectral radius equals $\lim_{k\to\infty} \|M_\theta^k\|^{1/k}$ (for any choice of consistent norm on matrices). □

For completeness, let us show that this analysis on a quadratic form corresponds indeed to the local acceleration behavior in general. For simplicity, we assume a positive Hessian at the minimum, an assumption that does not cover directly the case of Sinkhorn's iterations. An interesting direction would be to study the case of functions with separable non-smooth terms.

**Theorem 3.4.6.** *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a convex function of class $\mathscr{C}^2$ which admits a minimum where the Hessian is positive definite. Then, if it is well-defined and converges, alternate minimization starting from $u^{(0)} \in \mathbb{R}^n$ converges locally linearly at a rate $1-\eta$ with $0 < \eta < 1$. For the choice of parameter $\theta = 2/(1+\sqrt{\eta})$, the extrapolated alternate minimization converges locally at a rate $(1-\sqrt{\eta})/(1+\sqrt{\eta})$.*

*Proof.* Consider the fixed point map $\Phi_\theta : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$, defined as $(u^{(\ell+1)}, v^{(\ell+1)}) = \Phi_\theta(u^{(\ell)}, v^{(\ell)})$ and let $(\bar{u},\bar{v})$ be a minimizer of $F$. By the implicit function theorem applied on the partial gradients of $F$, there exists an open neighborhood $\mathscr{U} \times \mathscr{V} \subset \mathbb{R}^{n+m}$ of $(\bar{u},\bar{v})$ and continuously differentiable functions $G_1, G_2$ such that whenever $(u,v) \in \mathscr{U} \times \mathscr{V}$, it holds

$$\Phi_\theta(u,v) = ((1-\theta)u + \theta G_1(v), (1-\theta)v + \theta G_2((1-\theta)u + \theta G_1(v))).$$

It follows that $\Phi_\theta$ is continuously differentiable on $\mathscr{U} \times \mathscr{V}$, and its differential at the fixed point $(\bar{u},\bar{v})$ is the endomorphism $M_\theta$ from the proof of Theorem 3.4.4 associated to the quadratic Taylor approximation of $F$ around $(\bar{u},\bar{v})$. The local convergence rate follows with classical arguments and Theorem 3.4.4. □

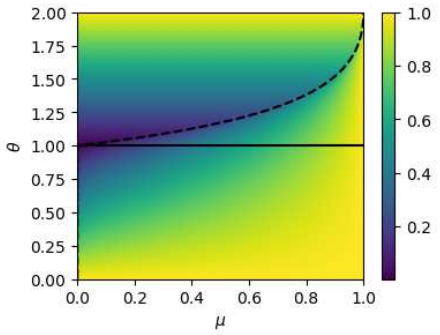



Figure 3.2.: Modulus of the transformed eigenvalues of the extrapolated algorithm $f(\theta,\mu)$ as a function of $\theta$ the extrapolation parameter and $\mu$ the original eigenvalue. (plain curve) the original eigenvalue is recovered for $\theta = 1$. (dashed curve) optimal value $\bar{\theta}$ for the acceleration parameter, as a function of $\mu_{\max}$. Below that line, the eigenvalues are real, and complex above, with modulus $\theta - 1$.

### 3.4.3. Numerical experiments on Sinkhorn's iterations

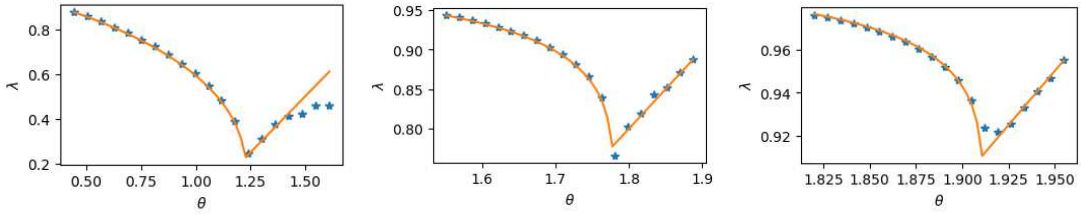
In order to compare the previous result to our specific problem of scaling iterations, we perform the following numerical experiment, focused on the specific case of standard Sinkhorn's iterations, to solve the entropy regularized optimal transport problem. Note that we observed a similar behavior for other scaling iterations.

**Linear regime** We compare the predicted local convergence behavior with practical experiments as follows. We set $I = J = 100$ and let $\mu, \nu \in \mathbb{R}^I$ be two random vectors, with entries independently uniformly distributed in $[1,2]$, and subsequently normalized to total mass 1. The entries of the cost matrix are independent, uniformly distributed on $[0,1]$. The problem we solve has thus no particular structure.

We execute the following steps: we chose $\varepsilon > 0$ and run a reference computation to compute the optimal potential $u_{\text{ref}}$. We then run standard and extrapolated Sinkhorn's iterations and measure their local linear convergence rates using a linear regression on $\|u^{(\ell)} - u_{\text{ref}}\|_\infty$ on the last third iterations[1], with a stopping criterion being $\|u^\ell - u_{\text{ref}}\|_\infty < 10^{-6}$. Note that in these experiments, we observed that the linear convergence regime started after just a few iterations (this is not true anymore for smaller values of the regularization $\varepsilon$). The theoretical and numerical convergence rates are compared on Figure 3.2. For the theoretical convergence rates, we use the formula in the proof of Theorem 3.4.4 (also displayed on Figure 3.2) which requires to measure Sinkhorn's numerical linear convergence rate.

**Global convergence** We do not have theoretical guarantees for global convergence. In practice, the algorithm may even fail to converge when the acceleration parameter $\theta$ is close to 2 outside of the local, linearly convergent, regime. In order to avoid this issue and still enjoy the

[1] the optimal $u_{\text{ref}}$ is unique only up to a translation, so one should recenter $u^{(\ell)}$ and $u_{\text{ref}}$ before computing their distance, for instance by subtracting the mean.




(a) $\varepsilon = 0.1$ and $\mu = 0.6$ (b) $\varepsilon = 0.01$ and $\mu = 0.984$ (c) $\varepsilon = 0.001$ and $\mu = 0.9978$

Figure 3.3.: Comparison of the theoretical local linear convergence rates of the extrapolated scaling iterations (orange lines) and the numerical results (blue crosses). The three plots correspond to the resolution of the same optimal transport problem with various regularization parameters $\varepsilon$ and $\mu$ corresponds to the experimental linear convergence rate of Sinkhorn's iterations.

same per-iteration complexity, we propose a heuristic for the evolution of $\theta$ that turns out quite robust and very efficient. At each iteration, of index $\ell$, set

$$\theta^{(\ell)} = \min\{\theta_{\text{opt}}, 1 + \frac{\ell}{\kappa + \ell}\}$$

where $\theta_{\text{opt}}$ is ideally computed from the linear convergence rate of Sinkhorn's iterations on the same problem with the formula of Theorem 3.4.4, or in practice, estimated from the resolution of similar problems and $\kappa \geq 1$ a parameter. The informal idea behind this heuristic is to attribute a weight $(1-\lambda)^\kappa$ to each eigenvalue in $[0, \lambda_{\max}]$ of the linearized iteration and to see which one has the strongest magnitude at iteration $\ell$, i.e. $\lambda^{(\ell)} = \arg\max \lambda^\ell (1-\lambda)^\kappa$. Our heuristic corresponds to setting $\theta^{(\ell)} = 1 + \lambda^{(\ell)}$ (as is the proof of Theorem 3.4.4). In practice, it performs well outside of the linear convergence regime too.

We display on Table 3.1 the number of iterations before convergence of the extrapolated iterations in comparison to the standard Sinkhorn's iterations, for solving the same random OT problem as in the previous paragraph, with $\kappa = 1$. For highly regularized problems, the running time is divided by 2 to 10. For small regularization parameters, the gain is of orders of magnitude.

| $\varepsilon$ | # iter Sinkhorn | # iter extrapolated | ratio | $\theta_{\text{opt}}$ |
|---|---|---|---|---|
| $10^{-1}$ | 16 | 8 | **2.0** | 1.25 |
| $10^{-2}$ | 316 | 46 | **6.9** | 1.70 |
| $10^{-3}$ | $7 \times 10^3$ | 189 | **37.4** | 1.93 |
| $10^{-4}$ | $4.1 \times 10^4$ | 709 | **57.5** | 1.96 |
| $10^{-5}$ | $8.2 \times 10^4$ | 985 | **82.8** | 1.97 |

Table 3.1.: Comparison of the number of iterations before convergence for Sinkhorn's iterations and the extrapolated version for several values of $\varepsilon$ and with the optimal acceleration parameter $\theta_{\text{opt}}$ computed from the linear convergence rate of Sinkhorn's iterations. Stopping criterion is $\|u^{(\ell)} - u_{\text{opt}}\|_\infty < 2 \times 10^{-6}$ (except for the last line where the tolerance is $10^{-8}$). The iterates are initialized with $v^{(0)} = 0$, except for the 2 last lines, where we use the optimal solution of the previous line in order to avoid numerical overflow. Complexity per iteration is the same for both algorithms so the ratio of number of iterations is also the running time ratio.

## 3.5. Scaling algorithms in continuous setting

In the context of standard optimal transport, the entropic regularization has been well studied in the continuous setting [2]. It corresponds to the Schrödinger bridge problem [99], and to the infinite dimensional extension of the so-called *DAD problem* [21, 118]. We refer to [86] for an up to date review on the DAD problem and its generalization to the scaling of linear operators in the finite dimensional setting. As regards the present thesis, a journey in the continuous setting is worthwhile since we obtain, for specific cases of the scaling algorithm, linear and dimension free convergence rates of the primal iterates.

### 3.5.1. Continuous entropic regularization

#### Regularization and $\Gamma$-convergence

Assume that $(X, \mathrm{d}x)$ and $(Y, \mathrm{d}y)$ are two Polish spaces endowed with reference probability measures and consider minimizing the entropic regularization of the generic formulation (3.1.1):

$$E_\varepsilon(\gamma) := \int_{X\times Y} c \cdot \mathrm{d}\gamma + F_1(\pi^x_\# \gamma) + F_2(\pi^y_\# \gamma) + \varepsilon H(\gamma | \mathrm{d}x \otimes \mathrm{d}y). \tag{3.5.1}$$

where $\mathrm{d}x \otimes \mathrm{d}y$ is the product measure on $X \times Y$, $\gamma \in \mathscr{M}_+(X \times Y)^n$ and $H$ is defined as follows. Note that in this section we come back to the setting of $n$ couplings as in Section 3.1 because it allows to treat the case of barycenters (balanced and unbalanced) while preserving the pointwise separability of the functions $F_1$ and $F_2$. It is also interesting to see how the non-expansiveness property (shown below) extends to the case $n > 1$ in a non-trivial way.

**Definition 3.5.1** (Relative entropy)**.** *The relative entropy (a.k.a. Kullback-Leibler divergence) is defined as the $f$-divergence associated to $f(s) = s\log(s) - s + 1$ and has the explicit expression, for $\mu, \nu \in \mathscr{M} + (X)$*

$$H(\mu|\nu) = \int \psi_f(\mu|\nu) = \begin{cases} \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}) \mathrm{d}\nu - \int \mu + \int \nu & \text{if } \mu \ll \nu \\ \infty & \text{otherwise.} \end{cases}$$

*This definition is consistent with the discrete Definition 3.3.1 and we use therefore the same notation. For vector valued measures, $H$ is the sum of the relative entropies on each component (with possibly positive weights).*

The following proposition gives conditions for the minimizer of the regularized problem to converge to a true regularizer, under an abstract assumption (this is a basic result and the assumption is for instance too strong to deal with unbalanced optimal transport with entropies with linear growth).

**Assumption 3.5.2** (Density of finite entropy plans)**.** *Any feasible plan for $E_0$ can be approximated (weakly) by a sequence of plans with finite entropy and same marginals.*

[2] by continuous, we mean spaces which are not necessarily discrete. The setting in this section is, after the first paragraph, that of measurable spaces.

**Proposition 3.5.3** (Minimizers and their convergence)**.** *Assume that $c$ is a (family of) l.s.c., lower bounded cost and that $F_1$, $F_2$ are convex, l.s.c. functionals. If $E_\varepsilon$ is feasible for some $\varepsilon > 0$, then it is feasible and admits a unique minimizer $\gamma_\varepsilon$ for all $\varepsilon > 0$. If furthermore $c$ is continuous and Assumption 3.5.2 is satisfied, then $E_\varepsilon$ $\Gamma$-converges to $E_0$. In particular, if $F_1$ or $F_2$ have compact sublevel sets, then $\gamma_\varepsilon$ converges weakly to a minimizer of $E_0$ as $\varepsilon \downarrow 0$.*

*Proof.* It is clear that a feasible coupling for some $\varepsilon > 0$ is feasible for all $\varepsilon > 0$. Moreover, thanks to our assumptions on $c$, $E_\varepsilon$ is weakly l.s.c. and also the sublevel sets of $H(\cdot|\mathrm{d}x \otimes \mathrm{d}y)$ are compact [103, prop. 2.10], so there exists a minimizer for all $\varepsilon > 0$ by the direct method of the calculus of variations. Moreover, the relative entropy is strictly convex w.r.t. its first argument so the minimizer is unique. The second result is a result of $\Gamma$-convergence of $E_\varepsilon$ to $E_0$ (see Section 2.1.2). For any $\gamma_0 \in \operatorname{dom} E_0$, the sequence given by Assumption 3.5.2 is, up to a reindexing to slow down the possible blow up of the entropy, a recovery sequence, in particular because since $c$ is continuous, $\int c \mathrm{d}\gamma$ is weakly continuous. This proves the $\Gamma$-lim sup. The $\Gamma$-lim inf, is obvious because the relative entropy is nonnegative and the other terms are weakly l.s.c. □

**Remark 3.5.4.**

*(i) Theorem 3.1.2 gives some conditions for $F_1$ or $F_2$ to have compact sublevel sets.*

*(ii) Assumption 3.5.2 holds when $X = Y$ is a bounded subset of $\mathbb{R}^d$, $\mathrm{d}x = \mathrm{d}y$ is the (rescaled) Lebesgue measure and the domain of $F_1$ and $F_2$ is included in the set of measures with finite entropy, as shown in [34] (where the $\Gamma$-convergence of entropic regularization of standard OT is studied). It also holds in the discrete case if $\gamma^{(0)}$ has full support, when one recovers Proposition 3.2.10.*

*(iii) the previous result does not inform on which minimizer of $E_0$ is chosen in the limit. This is an intricate question that is studied in [59] for standard OT with distance cost on $\mathbb{R}$.*

### Choice of reference measure

The previous result shows that the choice of $\mathrm{d}x$ and $\mathrm{d}y$ is critical for obtaining the weak convergence to a solution of (3.1.1) when $\varepsilon \to 0$. A necessary condition is that the support of $\mathrm{d}x \otimes \mathrm{d}y$ should contain the support of an optimizer of the unregularized problem. For an optimal entropy-transport problems between two measures $(\mu, \nu)$, in contrast to standard OT, the choice $\mathrm{d}x = \mu/\mu(X)$ and $\mathrm{d}y = \nu/\nu(Y)$ is not always suitable. If one of the entropy function is not superlinear, the marginals of the minimizer are not necessarily dominated by $\mu$ or $\nu$. In this case, the support of the reference measure $\mathrm{d}x\mathrm{d}y$ should contain the sets

$$\{(x, \arg\min_y c(x,y)); x \in X\}, \quad \text{and} \quad \{(\arg\min_x c(x,y), y); y \in Y\}.$$

where the singular part of minimizers is optimally located. In our numerical experiments, $X$ and $Y$ are discrete spaces and the choice of $\mathrm{d}x$ and $\mathrm{d}y$ goes in hand with the choice of a discretization grid on which we find approximate solutions to the original problem.

### Variational problem on densities and duality

From the very definition of the entropy $H$ (Definition 3.5.1), any feasible $\gamma$ of the entropy regularized problem (3.5.1) admits an $L^1$ density with respect to the reference measure $\mathrm{d}x \otimes \mathrm{d}y$. Accordingly, it is natural to reformulate the problem as a variational entropy regularized problem on measurable functions:

$$\min \left\{ F_1(\pi^x_\# r) + F_2(\pi^y_\# r) + \varepsilon H(r|K) \; ; \; r \in L^1(X \times Y)^n \right\} \tag{$P_\varepsilon$}$$

where $F_1(s) \leftrightarrow F_1(s\mathrm{d}x)$, $F_2(s) \leftrightarrow F_2(s\mathrm{d}y)$ and $K \in L^\infty_+(X \times Y)^n$ is defined componentwise by

$$K_k(x,y) := \exp\left(-c_k(x,y)/\varepsilon\right) \qquad k \in \{1,\dots,n\} \tag{3.5.2}$$

with the convention $\exp(-\infty) = 0$, the projection operator acts on each component of $r$ and $H$ is the sum of the relative entropy on each component (possibly with positive weights). Throughout this section, we only make the following general assumptions on the objects involved in $(P_\varepsilon)$:

**Assumption 3.5.5.**

*(i)* $(X, \mathrm{d}x)$ *and* $(Y, \mathrm{d}y)$ *are probability spaces (i.e. measured spaces with unit total mass). The product space* $X \times Y$ *is equipped with the product measure* $\mathrm{d}x\mathrm{d}y := \mathrm{d}x \otimes \mathrm{d}y$*;*

*(ii)* $F_1 : L^1(X)^n \to \mathbb{R} \cup \{\infty\}$ *and* $F_2 : L^1(Y)^n \to \mathbb{R} \cup \{\infty\}$ *are weakly l.s.c., convex and proper functionals;*

*(iii)* $K \in L^\infty_+(X \times Y)^n$ *and* $\varepsilon > 0$*.*

We begin with a general duality result, similar to duality results that can be found in the literature on entropy minimization [20].

**Proposition 3.5.6** (Duality)**.** *The dual problem of* $(P_\varepsilon)$ *is*

$$\sup \left\{ -F_1^*(-u) - F_2^*(-v) - \varepsilon \langle e^{(u \oplus v)/\varepsilon} - 1, K \rangle \; ; \; (u,v) \in L^\infty(X)^n \times L^\infty(Y)^n \right\} \tag{$D_\varepsilon$}$$

*where* $u \oplus v : (x,y) \mapsto u(x) + v(y)$ *and one has* $\min(P_\varepsilon) = \sup(D_\varepsilon)$*. If* $(P_\varepsilon)$ *is feasible then it admits a unique minimum* $r = (r_k)_k \in L^1(X \times Y)^n$ *and* $(u,v)$ *maximize* $(D_\varepsilon)$ *if and only if*

$$\begin{cases} -u \in \partial F_1(\pi^x_\# r) \\ -v \in \partial F_2(\pi^y_\# r) \end{cases} \qquad \text{and} \quad r_k(x,y) = e^{\frac{u_k(x)}{\varepsilon}} K_k(x,y) e^{\frac{v_k(x)}{\varepsilon}} \text{ for } k = 1,\dots,n. \tag{3.5.3}$$

*Proof.* In this proof, the spaces $L^\infty$ and $(L^\infty)^*$ are endowed with the strong and the weak* topology respectively. As $L^1(X \times Y)^n$ can be identified with a subset of the topological dual of $L^\infty(X \times Y)^n$, the function $H(\cdot|K)$ can be extended on $(L^\infty(X \times Y)^n)^*$ as $G(r) = H(r|K)$ if $r \in L^1(X \times Y)^n$ and $\infty$ otherwise. Its convex conjugate $G^* : L^\infty(X \times Y)^n \to \mathbb{R}$ is

$$G^*(w) = \sum_{k=1}^n \int_{X \times Y} (e^{w_k(x,y)} - 1) K_k(x,y) \mathrm{d}x\mathrm{d}y = \langle e^w - 1, K \rangle .$$

and is everywhere continuous for the strong topology [135, Theorem 4] (this property relies on the finiteness of d$x$d$y$ and the boundedness of $K$). The linear operator $A : L^\infty(X)^n \times L^\infty(Y)^n \to L^\infty(X\times Y)^n$ defined by $A(u,v) : (x,y) \mapsto u(x)+v(y)$ is continuous and its adjoint is defined on $(L^\infty(X\times Y)^n)^*$ (identified with a subset of $\mathscr{M}(X\times Y)^n$) by $A^*(r) = (\pi^x_\# r, \pi^y_\# r)$. Since $F_1$ and $F_2$ are convex, l.s.c., proper and since $G^*$ is everywhere continuous on $L^\infty(X\times Y)^n$, strong duality and the existence of a minimizer for $(P_\varepsilon)$ is given by Fenchel-Rockafellar theorem (see Appendix A). More explicitly, this theorem states that

$$\sup_{(u,v)\in L^\infty(X)^n\times L^\infty(Y)^n} -F_1^*(-u) - F_2^*(-v) - \varepsilon G^*(A(u,v)/\varepsilon)$$

and

$$\min_{r\in(L^\infty(X\times Y)^n)^*} F_1(P^X_\# r) + F_2(P^Y_\# r) + \varepsilon G(r)$$

are equal, the latter being exactly $(P_\varepsilon)$ since $G$ is infinite outside of $L^1(X\times Y)^n$. It states also that if $(u,v)$ maximizes $(D_\varepsilon)$, then any minimizer of $(P_\varepsilon)$ satisfies $r \in \partial G^*(A(u,v)/\varepsilon)$ and the expression for the subdifferential of $G^*$ is a particular case of Lemma 3.5.12 below. Finally, uniqueness of the minimizer for $(P_\varepsilon)$ comes from the strict convexity of $G$. □

### 3.5.2. Continuous scaling iterations

In this section, we recover the continuous scaling iterations from the alternate maximization on the dual problem. The derivation are similar (and contain as a particular case) those of Section 3.3, but more care is needed in the proofs. Let $\mathscr{K}$ and $\mathscr{K}^T$ be the linear operators defined, for $a : X \to [0,\infty[^n$ and $b : Y \to [0,\infty[^n$ measurable, for $k = 1,\dots,n$ as

$$[\mathscr{K} b]_k(x) := \int_Y K_k(x,y) b_k(y) \mathrm{d}y \quad \text{and} \quad [\mathscr{K}^T a]_k(y) := \int_X K_k(x,y) a_k(x) \mathrm{d}x. \tag{3.5.4}$$

Given $v^{(0)} \in L^\infty(Y)^n$, the alternate maximization procedure applied to the dual problem $(D_\varepsilon)$ defines, for $\ell \in \mathbb{N}^*$, the iterates

$$\begin{cases} u^{(\ell+1)} = \arg\max_{u\in L^\infty(X)^n} -F_1^*(-u) - \varepsilon\langle e^{u/\varepsilon}, \mathscr{K} e^{v^{(\ell)}/\varepsilon}\rangle_X, \\ v^{(\ell+1)} = \arg\max_{v\in L^\infty(Y)^n} -F_2^*(-v) - \varepsilon\langle e^{v/\varepsilon}, \mathscr{K}^T e^{u^{(\ell+1)}/\varepsilon}\rangle_Y, \end{cases} \tag{3.5.5}$$

where we used the fact that, by Fubini-Tonelli, one has

$$\langle e^{(u\oplus v)/\varepsilon}, K\rangle_{X\times Y} = \langle e^{\frac{u}{\varepsilon}}, \mathscr{K} e^{\frac{v}{\varepsilon}}\rangle_X = \langle e^{\frac{v}{\varepsilon}}, \mathscr{K}^T e^{\frac{u}{\varepsilon}}\rangle_Y. \tag{3.5.6}$$

Conditions which ensure the existence of these iterates are postponed to Theorem 3.5.14. For the moment, remark that by strict convexity, they are uniquely defined when they exist. The continuous scaling iterations are defined as follows.

**Definition 3.5.7** (Continuous scaling iterations)**.** *The continuous scaling iterations is the sequence $(a^{(\ell)}, b^{(\ell)})_{\ell\in\mathbb{N}^*}$ defined by $b^{(0)} \in L^\infty_+(Y)$ and, for $\ell \in \mathbb{N}^*$,*

$$a^{(\ell+1)} := \frac{\mathrm{prox}^H_{F_1/\varepsilon}(\mathscr{K} b^{(\ell)})}{\mathscr{K} b^{(\ell)}}, \quad b^{(\ell+1)} := \frac{\mathrm{prox}^H_{F_2/\varepsilon}(\mathscr{K}^T a^{(\ell+1)})}{\mathscr{K}^T a^{(\ell+1)}} \tag{3.5.7}$$

In this definition, the division is performed pointwise with the convention $0/0 = 0$ (this has to be thought of as computing a relative density). The proximal operator for the relative entropy $H$ is defined for $F_1$ (and similarly for $F_2$) as

$$\mathrm{prox}^H_{F_1/\varepsilon}(z) := \arg\min \{F_1(s) + \varepsilon H(s|z) \;;\; s : X \to \mathbb{R}^n \text{ measurable}\}.$$

The following proposition shows that, as long as they are well defined, these iterates are related to the alternate dual maximization iterates by a change of variables.

**Proposition 3.5.8.** *Define* $a^{(0)} := \exp(v^{(0)}/\varepsilon)$*, let* $(a^{(\ell)}, b^{(\ell)})$ *be the scaling iterates* (3.5.7) *and* $(u^{(\ell)}, v^{(\ell)})$ *the alternate dual maximization iterates defined in* (3.5.5)*. If for all* $\ell \in \mathbb{N}$*, either* $\log a^{(\ell)} \in L^\infty(X)^n$ *and* $\log b^{(\ell)} \in L^\infty(Y)^n$*, or* $u^{(\ell)} \in L^\infty(X)^n$ *and* $v^{(\ell)} \in L^\infty(Y)^n$ *then*

$$(a^{(\ell)}, b^{(\ell)}) = (e^{u^{(\ell)}/\varepsilon}, e^{v^{(\ell)}/\varepsilon}).$$

The proof of this proposition uses the following lemma.

**Lemma 3.5.9.** *Let* $(T, \mathrm{d}t)$ *be a measured space and* $v \in L^1_+(T)$*. For any* $u : T \to \mathbb{R}$ *measurable, if* $H(u|v) < \infty$ *then* $u \in L^1_+(T)$*.*

*Proof.* Without loss of generality, one can assume $v$ positive since for $\mathrm{d}t$-a.e. $t$, if $v(t) = 0$ then $u(t) = 0$. The subgradient inequality at $\exp(1)$ gives, for all $s \in \mathbb{R}$, $s \leq h(s) + e^1 - 1$, where $h(s) = s\log(s) - s + 1$. Consequently,

$$\int_T u\mathrm{d}t = \int_T (u(t)/v(t))v(t)\mathrm{d}t \leq \int_T \big(h(u(t)/v(t)) + e^1 - 1\big)\, v(t)\mathrm{d}t < \infty. \qquad \square$$

*Proof of Proposition 3.5.8.* Suppose that $v^{(\ell)} \in L^\infty(Y)^n$ and that $b^{(\ell)} = e^{v^{(\ell)}/\varepsilon}$. One has $\mathscr{K} b^{(\ell)} \in L^1(X)^n$ and from Lemma 3.5.9, one can compute $\mathrm{prox}^H_{F_1/\varepsilon}(\mathscr{K} b^{(\ell)})$ in $L^1(X)^n$. Fenchel-Rockafellar duality gives (see the proof of Proposition 3.5.6 for a more detailed application in a similar setting):

$$\sup_{u \in L^\infty(X)^n} -F_1^*(-u) - \varepsilon\langle e^{\frac{u}{\varepsilon}}, \mathscr{K} e^{\frac{v^{(\ell)}}{\varepsilon}}\rangle = \min_{s \in L^1(X)^n} F_1(s) + \varepsilon H(s|\mathscr{K} e^{\frac{v^{(\ell)}}{\varepsilon}})$$

and the optimality conditions state that $u^\star$ maximizes the problem on the right if and only if the minimizer $s^\star := \mathrm{prox}^H_{F_1/\varepsilon}(\mathscr{K} e^{\frac{v^{(\ell)}}{\varepsilon}})$ of the problem on the left belongs to the subdifferential of $u \mapsto \langle e^{\frac{u}{\varepsilon}}, \mathscr{K} e^{\frac{v^{(\ell)}}{\varepsilon}}\rangle$ at the point $u^\star$. That is (as a consequence of Lemma 3.5.12 below) if and only if for $\mathrm{d}x$ almost every $x \in X$,

$$s^\star(x) = e^{u^\star(x)/\varepsilon} \cdot (\mathscr{K} e^{v^{(\ell)}/\varepsilon})(x)$$

Thus if $\varepsilon \log a^{(\ell+1)}$ belongs to $L^\infty(X)^n$ or if $u^\star \in L^\infty(X)^n$ exists, then $u^\star = \varepsilon \log a^{(\ell+1)}$. The result follows by induction. $\square$

### 3.5.3. Existence of iterates for integral functionals

Our next step is to give conditions on $F_1$ and $F_2$ that guarantee the existence of the scaling iterates (3.5.7) and an equivalence with alternate maximization on the dual (3.5.5). The main condition involves the notions of normal integrands and integral functionals.

**Definition 3.5.10** (Normal integrands and integral functionals [139])**.** *A function $f : X \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called a* normal integrand *if its epigraphical mapping $X \ni x \mapsto \mathrm{epi} f(x,\cdot)$ is closed-valued and measurable. A convex* integral functional *is a function $F : L^1(X)^n \to \mathbb{R} \cup \{\infty\}$ of the form*

$$F(s) = I_f(s) := \int_X f(x, s(x))\mathrm{d}x$$

*where $f$ is a normal integrand and $f(x,\cdot)$ is convex for all $x \in X$. In this Section, we say that $F$ is an* admissible *integral functional if moreover for all $x \in X$, $f(x,\cdot)$ takes nonnegative values, has a domain which is a subset of $[0,\infty[^n$ and if there exists $s \in L^1(X)^n$ such that $I_f(s) < \infty$.*

**Example 3.5.11.**

*(i) for finite dimensional problems (when $X$ and $Y$ have a finite number of points), integral functionals are simply sums of pointwise l.s.c. functions;*

*(ii) the $f$-divergence between two densities $(u,v) \in L^1_+(X)^2$, which is defined as $\int_X \psi_f(u(x)|v(x))\mathrm{d}x$ for an entropy function $f$ as in Definition 1.2.15 is an admissible integral functional, as a function on $(u,v)$ jointly as well as a function of $u$ or $v$ separately [139, Prop. 14.45].*

*(iii) consider a function of the general form* (3.1.4) *between densities, i.e. for $u \in L^1(X)^n$*

$$F(u) = \inf_{v \in L^1(X)^n} \left\{ \sum \int_X \psi_f(u_i(x)|v_i(x))\mathrm{d}x + \int_X g(v(x))\mathrm{d}x \right\}.$$

*where $f$ an entropy function and $g : \mathbb{R}^n \to \mathbb{R}_+ \cup \{\infty\}$ convex, proper, l.s.c. Then minimization can be performed pointwise in $v$ and $F$ is a convex integral functional ([139, Prop. 14.47] and Proposition 3.1.1).*

In general, this concept allows to deal conveniently with measurability issues: conjugation (as well as subdifferentiation) can be performed pointwise.

**Lemma 3.5.12.** *If $F$ is an admissible integral functional associated to the convex normal integrand $f$, then $F$ is convex and weakly l.s.c., $f^*$ is also a normal convex integrand, $F^* = I_{f^*}$, where conjugation is computed w.r.t. the second variable.*

*Proof.* This property can be found in [138] under the assumption of existence of a feasible point $s^\star \in L^1(X)^n$ for $I_f$ and a feasible point $u^\star \in L^\infty(X)^n$ for $I_{f^*}$. Our admissibility criterion requires the existence of $s^\star$ and one has

$$I_{f^*}(0) = \int_X f^*(x,0)\mathrm{d}x = -\int_X \inf_{s \in \mathbb{R}^n} f(x,s)\mathrm{d}x < \infty$$

since $\inf_s f(x,s) \in [0, f(x,s^\star(x))]$. □

It follows also from general results on normal integrand that the operator $\mathrm{prox}^H$ is separable.

**Proposition 3.5.13.** *Let $s : X \to \mathbb{R}^n$ be measurable. If $F = I_f$ is an admissible integral functional then for almost all $x \in X$,*

$$\left(\mathrm{prox}^H_{F/\varepsilon}(s)\right)(x) = \mathrm{prox}^H_{f(x,\cdot)/\varepsilon}(s(x))\,.$$

Remark that in this proposition, we used the same notation $H$ for the relative entropy between two functions and between two vectors. In both cases, it is the sum—discrete or continuous—of the pointwise relative entropy.

*Proof.* The problem which defines the proximal operator is that of minimizing $I_f(z) + H(z|s) := I_g(z)$ over measurable functions $z : X \to \mathbb{R}^n$ with

$$g : (x,z) \in X \times \mathbb{R}^n \mapsto f(x,z) + H(z|s(x))\,.$$

The function $(x,z) \mapsto H(z|s(x))$ is a convex normal integrand by [139, Prop. 14.30 and 14.45c]. Thus $g$ is itself a normal convex integrand, as the sum of normal convex integrands [139, Prop. 14.44]. Then a minimization interchange result [139, Thm. 14.60]] states that minimizing $I_g$ is the same as minimizing $g$ pointwise. □

By Lemma 3.5.12, if $F_1$ and $F_2$ are admissible integral functionals then $F_1^*$ and $F_2^*$ are also integral functionals. So the alternating optimization on $(D_\varepsilon)$ can be relaxed to the space of measurable functions and still be well defined:

$$\begin{cases} u^{(\ell+1)} = \underset{u:X\to\mathbb{R}^n}{\mathrm{argmax}} -I_{f_1^*}(-u) - \varepsilon\langle e^{\frac{u}{\varepsilon}}, \mathscr{K} e^{\frac{v^{(\ell)}}{\varepsilon}}\rangle_X \\ v^{(\ell+1)} = \underset{v:Y\to\mathbb{R}^n}{\mathrm{argmax}} -I_{f_2^*}(-v) - \varepsilon\langle e^{\frac{v}{\varepsilon}}, \mathscr{K}^T e^{\frac{u^{(\ell+1)}}{\varepsilon}}\rangle_Y\,. \end{cases} \tag{3.5.8}$$

The following theorem gives existence, uniqueness of this iterates and a precise relation with the scaling iterates (3.5.7).

**Theorem 3.5.14.** *Let $F_1$ and $F_2$ be admissible integral functionals associated to the normal integrands $f_1$ and $f_2$ as in Definition* (3.5.10)*. Assume that for all $x \in X$ and $y \in Y$, there exists points $s_1$ and $s_2$ with strictly positive coordinates such that $f_1(x,s_1) < \infty$ and $f_2(y,s_2) < \infty$ and that $K$ takes positive values. Define $a^{(0)} = 1$ and let $(a^{(\ell)}, b^{(\ell)})$ be the scaling iterates* (3.5.7)*. Then, with initialization $v^{(0)} = 0$, the iterates $(u^{(\ell)}, v^{(\ell)})$ in* (3.5.8) *are well defined, unique, and for all $\ell \in \mathbb{N}$ one has $(a^{(\ell)}, b^{(\ell)}) = (e^{\frac{u^{(\ell)}}{\varepsilon}}, e^{\frac{v^{(\ell)}}{\varepsilon}})$.*

*Proof.* Suppose that $v^{(\ell)} : Y \to \mathbb{R}^n$ is a well-defined measurable function and that $b^{(\ell)} = e^{v^{(\ell)}/\varepsilon}$. As $K$ is positive, $\mathscr{K} b^{(\ell)}$ is positive d$x$ a.e. Let $a^{(\ell+1)}$ be computed with (3.5.7). Thanks to Proposition 3.5.13, the proximal operator can be decomposed as pointwise optimization problems and our assumptions allows to apply Fenchel-Rockafellar duality (Appendix A) in the case where both problems reach their optima

$$\max_{u\in\mathbb{R}^n} -f_1^*(x,-u) - \varepsilon\langle e^{u/\varepsilon}, \mathscr{K} b^{(\ell)}\rangle = \min_{s\in\mathbb{R}^n} f_1(x,s) + \varepsilon H(s|\mathscr{K} b^{(\ell)}(x))$$

with the relation between optimizers: $e^{u/\varepsilon} = s/\mathscr{K} b^{(\ell)}$. The minimized function is a strictly convex normal integrand because $H(\cdot|\mathscr{K} b^{(\ell)}(x))$ is a normal integrand and sum of normal integrands are normal [139, Prop. 14.44]. It follows that the function of pointwise minimizers $x \mapsto s(x)$ is uniquely well-defined and measurable by [139, Thm. 14.37], so the function of pointwise maximizers $x \mapsto u(x)$ is also measurable. This shows that $a^{(\ell+1)} = e^{u^{(\ell+1)}/\varepsilon}$ and one concludes by induction. □

### 3.5.4. Global linear convergence in continuous setting

After these technical preliminaries, we are ready to state a nonexpansiveness result and a convergence result that hold in the infinite dimensional case. We start by making sure that fixed points of the scaling iterations are minimizers.

**Proposition 3.5.15** (Fixed point)**.** *Under the assumptions of Theorem 3.5.14, if the scaling iterations* (3.5.7) *admit a fixed point* $(a,b)$ *such that* $\log a \in L^\infty(X)^n$ *and* $\log b \in L^\infty(Y)^n$ *then* $(\varepsilon \log a, \varepsilon \log b)$ *is a solution of* $(D_\varepsilon)$ *and the function r defined for each* $k = 1,\dots,n$ *by* $r_k(x,y) = a_k(x)K_k(x,y)b_k(y)$ *is the unique solution of* $(P_\varepsilon)$*.*

*Proof.* As a consequence of Proposition 3.5.13, we can write the optimality condition of a fixed point of (3.5.8) for almost every $x \in X$ as

$$u_k(x) = \varepsilon \log\left(\frac{a_k(x)\mathscr{K} b_k(x)}{\mathscr{K} b_k(x)}\right)$$

for some $u(x) \in -\partial f_1(x, a(x)\mathscr{K} b(x))$. Thus $-\varepsilon\log(a) \in \partial I_{f_1}(\pi^x_\# r)$ because $(\pi^x_\# r)(x) = a(x)\mathscr{K} b(x)$ for a.e. $x$. Similar derivations for $b$ show that the couple $(\varepsilon\log a, \varepsilon \log b)$ and $r$ satisfies the optimality conditions (3.5.3). □

#### Metrics from non-linear Perron-Frobenius theory

The specific form of the scaling iterations make them suitable for tools from non-linear Perron-Frobenius theory [98] which studies the convergence and fixed points of iterated maps satisfying some homogeneity or positivity properties. It is well known that Sinkhorn's iterations—a special case of the scaling iterations (3.5.7)—converges linearly for the so-called *Hilbert projective metric* [73, 21]. This metric is defined as follows (we use Birkhoff's definition, specialize it to the cone $L^\infty_+(X)^n$ and denote $\preceq$ the associated partial order).

**Definition 3.5.16** (Hilbert's projective metric)**.** *Let* $\sim$ *be the equivalence relationship on* $L^\infty_+(X)^n$ *defined as* $u \sim v$ *if and only if there exists* $\alpha,\beta > 0$ *such that* $u \preceq \alpha v \preceq \beta u$*. For* $x,y \in L^\infty_+(X)^n \setminus \{0\}$ *define the maximum and minimum ratios as*

$$M(u/v) := \inf\{\alpha > 0\,;\, u \preceq \alpha v\}, \qquad m(u/v) := \sup\{\beta > 0\,;\, \beta v \preceq u\}.$$

*The Hilbert's projective metric is defined as*

$$d_H(u,v) := \begin{cases} \log(M(u/v)/m(u/v)) & \text{if } u \sim v \text{ and } u,v \neq 0,\\ 0 & \text{if } u = v = 0,\\ \infty & \text{otherwise.}\end{cases}$$

This defines an extended metric on the spaces of rays of $L^\infty_+(X)^n$. It has the property that maps which are order preserving and $p$-homogeneous are $p$-Lipschitz and linear maps which have a *finite projective diameter* are also Lipschitz contractions [31]. It follows in particular that when the operators $\mathscr{K}$ and $\mathscr{K}^T$ have finite projective diameter (which is true if $c$ is bounded) then Sinkhorn's iterations are contractive. For the general case of scaling iterations, there is a variant of the Hilbert metric whose properties are slightly more convenient, called the *Thompson metric* [155].

**Definition 3.5.17** (Thompson's metric)**.** *Using the notations of Definition 3.5.16, the Thompson's metric is defined as*

$$d_T(u,v) := \begin{cases} \max\{\log M(u/v), \log M(v/u)\} & \text{if } u \sim v \text{ and } u,v \neq 0 \\ 0 & \text{if } u = v = 0 \\ \infty & \text{otherwise.} \end{cases}$$

Each part of $L^\infty_+(X)^n$ for the equivalence relationship "$\sim$" endowed with the metric $d_T$ is a complete metric space (for instance the set $\left\{s \in L^\infty_+(X)^n \,;\, \log s \in L^\infty(X)^n\right\}$ is a part of $L^\infty_+(X)^n$). Moreover, an order preserving mapping $T : L^\infty_+(X)^n \to L^\infty_+(Y)^n$ is non-expansive under Thompson's metric if and only if $f$ is subhomogeneous, i.e., $\lambda T(u) \preceq T(\lambda u)$ for all $u \in L^\infty_+(X)^n$ and $0 \leq \lambda \leq 1$.

### Contraction properties of scaling iterates

In this section, we avoid subtleties by working with strictly positive functions. While not all positive functions belong to the same part, if a part intersects $L^\infty_{++}(X)$ then it is entirely included in $L^\infty_{++}(X)$. Let us first isolate cases that we will refer to frequently.

**Assumption 3.5.18.** *The function $F$ is an admissible integral functional and* $\operatorname{prox}^H_F(s)$ *preserves strict positivity.*

For $\operatorname{prox}^H_F$ to preserve (strict) positivity, it is necessary and sufficient that almost everywhere, $f(x,\cdot)$ has a positive feasible point, where $F = I_f$. Indeed, in that case, the qualification constraint is satisfied for the proximal problem and Fenchel-Rockafellar theorem guarantees that the entropy is subdifferentiable at the minimum.

**Assumption 3.5.19.** *The function $F$ is the relative entropy with respect to a family of densities $p \in L^1_{++}(X)^n$, i.e. for $s \in L^\infty_{++}(X)^n$ one has $F_1(s) = \lambda \sum_{k=1}^n H(s_k|p_k)$ with $\lambda > 0$ (in particular, it satisfies Assumption 3.5.18.*

The next lemma shows why it is relevant to introduce the Thompson metric in the context of scaling iterations.

**Lemma 3.5.20** (Contraction properties)**.** *If $F$ satisfies Assumption 3.5.18, then the operator $s \mapsto \operatorname{prox}^H_F(s)/s$ is nonexpansive for the Thompson's metric on $L^\infty_{++}(X)^n$. If $F$ satisfies, Assumption 3.5.19, then this operator is Lipschitz contractive of contracting ratio $(1+\lambda^{-1})^{-1}$ on $L^\infty_{++}(X)^n$.*

*Proof.* According to Proposition 3.5.13, the operator $\mathrm{prox}_F^H$ can be computed pointwisely so we can work directly with the operator $T : \mathbb{R}_+^n \to \mathbb{R}_+^n$ defined as $T(a) \coloneqq P(a) \oslash a$ where $P(a) \coloneqq \mathrm{prox}_f^H(a)$. Let $a, b \succ 0$ and let us first show that $P$ is order preserving. Combining the optimality conditions of the proximal step, we get

$$\log(T(a)) \in -\partial f(P(a)), \qquad \log(T(b)) \in -\partial f(P(b)). \tag{3.5.9}$$

where the logarithm acts componentwise. But the subdifferential of $f$ is a monotone set so it follows,

$$(\log(T(a)) - \log(T(b))) \cdot (P(a) - P(b)) \leq 0.$$

Rearranging the terms yields

$$(\log a - \log b) \cdot (P(a) - P(b)) \geq (\log P(a) - \log P(b)) \cdot (P(a) - P(b))$$

where the last term is nonnegative because log is the gradient of the entropy and, as such, is monotone. It follows that $P$ is order preserving and, coming back to (3.5.9) again, that $T$ is order *reversing*. For subhomogeneity, let $0 < \lambda \leq 1$ and notice that, since $\lambda a \preceq a$, it holds

$$T(\lambda a) = P(\lambda a) \oslash (\lambda a) \preceq P(a) \oslash (\lambda a) = \lambda^{-1} T(a).$$

Combining these properties, it follows that the operator $a \mapsto 1 \oslash T(a)$ is order preserving and subhomogeneous so it is nonexpansive for $d_T$ and since $d_T$ is invariant by inversion, so is $T$. For the case $F = \lambda \sum_{k=1}^n H(s_k|p_k)$, the optimality conditions give the explicit formula $[T(u)]_k = (u_k/p_k)^{\lambda/(\lambda+1)}$ and one can directly check the claim from the definition of $d_T$. □

Since the operator from Lemma 3.5.20 is at the core of the scaling iterations, the next nonexpansiveness result follows easily. We need an assumption on $\mathscr{K}$ which is always satisfied, for instance, if $\mathscr{K}$ is derived from a finite valued transport cost.

**Theorem 3.5.21** (Scaling iterations are nonexpansive)**.** *If $F_1$ and $F_2$ satisfy Assumption 3.5.18 and $\mathscr{K}$ preserves strict positivity, then the scaling iterations are nonexpansive for the Thompson metric on $L^\infty_{++}(X)$ and $L^\infty_{++}(Y)$. They are $(1+\frac{\varepsilon}{\lambda})^{-1}$-Lipschitz contractive if $F_1$ or $F_2$ satisfies Assumption 3.5.19.*

*Proof.* Nonexpansiveness follows from Lemma 3.5.20 and the fact that $\mathscr{K}$ is an order preserving linear operator, and so is nonnexpansive for the Thompson metric. Under Assumption 3.5.19, it holds, writing $\alpha = (1+\varepsilon/\lambda)^{-1}$,

$$d_T(a^{(\ell+1)}, a^{(\ell)}) \leq \alpha\, d_T(\mathscr{K} b^{(\ell)}, \mathscr{K} b^{(\ell-1)}) \leq \alpha\, d_T(b^{(\ell)}, b^{(\ell-1)}) \leq \alpha\, d_T(a^{(\ell)}, a^{(\ell-1)}). \qquad \square$$

We deduce in particular a convergence result when a part of $L^\infty_{++}$ is stable under the action of the scaling iterates (for instance, when their logarithms remain in $L^\infty$). This condition is easy to check in the discrete setting because then, parts are characterized by the pattern of zeros. This corollary contains thus, in the discrete case, the linear convergence for the computation of optimal entropy-transport problems and unbalanced barycenters with the relative entropy divergence, or gradient flows for the metric $\widehat{W}_2$.

**Corollary 3.5.22.** *If $F_1$ satisfies Assumption 3.5.18 and $F_2$ satisfies Assumption 3.5.19, and moreover $d_T(b^{(\ell_0+1)}, b^{(\ell_0)})$ is finite for some $\ell_0 \in \mathbb{N}$, then the scaling iterates $(a^{(\ell)}, b^{(\ell)})$ starting from $b^{(0)} = 1$ converge to $(a,b) \in L^\infty_{++}(X)^n \times L^\infty_{++}(Y)^n$ at a linear rate. For instance, for $(a^{(\ell)})_\ell$ and $\ell_0 = 0$, it holds:*

$$d_T(a^{(\ell)}, a) \leq (1 + \varepsilon/\lambda)^{-\ell} d_T(a^{(0)}, a) < \infty.$$

*In particular, if $(u,v) \coloneqq (\varepsilon \log a, \varepsilon \log b) \in L^\infty(X)^n \times L^\infty(Y)^n$, then $(u,v)$ maximizes the dual problem $(D_\varepsilon)$.*

*Proof.* The additional assumption guarantees that the iterates $(b^{(\ell)})_\ell$ and $(a^{(\ell)})_\ell$ are in the same part of $L_+(X)$ or $L_+(Y)$ for $\ell \geq \ell_0$ which are complete metric spaces with the metric $d_T$. The conclusion follows by Theorem 3.5.21 and Banach fixed point theorem. The optimality of $(u,v)$ is given by Proposition 3.5.15. □

# Chapter 4.

# Proximal Splitting for Dynamic Formulations

In this short chapter, we propose a numerical method for solving the dynamic formulations of unbalanced optimal transport introduced in Chapter 1. We adopt and adapt the approach described in [123] (that deals with the balanced case) and give more details on certain aspects.
In Section 4.1, we describe the splitting algorithm and the staggered grid discretization. In Section 4.2 we derive the method to compute quickly the projection maps in the spectral domain, considering various boundary conditions, and the discretized proximal maps for various action functionals. We conclude in Section 4.3 with numerical computations for various unbalanced optimal transport models, on 1-D and 2-D domains.
This chapter is an augmented version of a section of the published article [44].

## 4.1. Discretization and optimization algorithm

The general problem we propose to solve is of the following:

$$\min \int_0^T \int_\Omega f(\rho_t(x), \omega_t(x), \zeta_t(x)) \mathrm{d}x\mathrm{d}t \tag{4.1.1}$$

where $(\rho, \omega, \zeta)$ are densities on $[0,T] \times \Omega$ that satisfy the continuity equation between two densities $(\mu, \nu)$ on $\Omega \subset \mathbb{R}^d$, i.e.

$$\partial_t \rho + \nabla \cdot \omega = \zeta, \qquad (\rho_0, \rho_T) = (\mu, \nu).$$

and $f : \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is a simple proper, l.s.c. and jointly convex function. The approach that we describe is an extension of [123]. See also the related [85] where other extensions are considered, such as Riemannian costs and the introduction of additional physical constraints.

### 4.1.1. Discretization

For clarity of the presentation, we consider a domain $\Omega = [0,L]$ which is a line segment: the extension to bounded domains in higher dimension is just a matter of notations (if the domain is not a hypercube, one may enclose it in a hypercube and extend $f_1$ as being $\infty$ if $\rho \neq 0$ outside of $\Omega$). Likewise, the extension to functions $f$ with a dependency in $(t,x)$ is just a matter of notations.

The time and space domain $[0,T] \times [0,L]$ is discretized into $N_t \times N_x$ rectangular cells denoted $W_{i,j}$, the center of which form the so-called *centered grid*, denoted $G_c$. The grids steps are denoted $(h_t, h_x) = (T/N_t, L/N_x)$. We consider moreover *staggered grids* which describe the boundaries of the cells: the samples of the time-staggered grid $G_t$ are located at the center of the time boundaries and the space-staggered grid $G_x$ is defined similarly. These grids $(G_c, G_t, G_x)$ are shown on Figure 4.1 and defined in mathematical terms as

$$\begin{aligned}
G_c &= \{((i-1/2)h_t, (j-1/2)h_x) : 1 \leq i \leq N_t, 1 \leq j \leq N_x\}, \\
G_x &= \{((i-1/2)h_t, (j-1)h_x) : 1 \leq i \leq N_t, 1 \leq j \leq N_x + 1\}, \\
G_t &= \{((i-1)h_t, (j-1/2)h_x) : 1 \leq i \leq N_t + 1, 1 \leq j \leq N_x\}.
\end{aligned}$$

The unknown we consider are pairs of centered variables $(\hat{\rho}, \hat{\omega}, \hat{\zeta}) \in V_c \times V_c \times V_c$ and staggered variables $(\rho, \omega, \zeta) \in V_t \times V_x \times V_c$, where $V_c, V_t$ and $V_x$ denote the real valued functions supported on $G_c, G_t$ and $G_x$, respectively.

#### Continuity equation

The continuity equation constraint is enforced on the staggered variables by asserting the mass conservation in each cell. Interpreting $\rho_{i,j}$ as the averaged density on the left time boundary of the cell and $\omega_{i,j}$ the averaged momentum on the bottom space boundary, and $\zeta_{i,j}$ as the averaged source over the whole cell, this amounts to the conservation of mass equation

$$h_x(\rho_{i+1,j} - \rho_{i,j}) + h_t(\omega_{i,j+1} - \omega_{i,j}) = h_x h_t \zeta_{i,j}.$$

With the operator $A : V_t \times V_x \times V_c \to V_c$ defined as

$$(A(\rho,\omega,\zeta))_{ij} \coloneqq \frac{1}{h_t}(\rho_{i+1,j}-\rho_{i,j}) + \frac{1}{h_x}(\omega_{i,j+1}-\tilde{\omega}_{i,j}) - \zeta_{i,j}, \tag{4.1.2}$$

the continuity equation constraint is $A(\rho,\omega,\zeta)=0$. Notice that imposing boundary conditions amounts to replacing the staggered variables on a boundary by fixed values, but it is convenient to still include the fixed values in the indexation. Then, the operator $A$—which is linear when no boundary constraints are enforced—becomes an affine operator.

Solutions to this discrete continuity equation, can be associated to a continuous solution with densities $(\tilde{\rho},\tilde{\omega},\tilde{\zeta})$ defined on $[0,T]\times[0,L]$ through piecewise constant or affine interpolation. These continuous solutions are defined on each cell $W_{i,j}$ by: $\tilde{\rho}(t,x)$ is constant in space and linearly interpolate in time between $\rho_{i,j}$ and $\rho_{i+1,j}$, $\tilde{\omega}(t,x)$ is constant in time and linearly interpolate in space between $\omega_{i,j}$ and $\omega_{i,j+1}$, and $\tilde{\zeta}(t,x)$ is constant in space and time equal to $\zeta_{i,j}$.

### Functional discretization and interpolation constraint

Ideally, one would exactly solve the dynamic problem among piecewise constant or affine densities $(\tilde{\rho},\tilde{\omega},\tilde{\zeta})$ as described above, but the integral $\int_0^T\int_0^L f(\tilde{\rho},\tilde{\omega},\tilde{\zeta})\mathrm{d}x\mathrm{d}t$ is generally hard to evaluate and not convenient to minimize. Instead, we replace, on each cell $W_{i,j}$, the integral of $f$ by $f$ applied to the averaged densities. Notice that by Jensen inequality, it holds

$$f\Big(\fint_{W_{i,j}}\tilde{\rho},\fint_{W_{i,j}}\tilde{\omega},\fint_{W_{i,j}}\tilde{\zeta}\Big) \le \fint_{W_{i,j}} f_1(\tilde{\rho},\tilde{\omega},\tilde{\zeta})$$

with equality if $\tilde{\rho},\tilde{\omega}$ are constant on the cell (here $\fint$ denote the average, or normalized integral).

We then consider an auxiliary triplet of variables, the centered variables $(\hat{\rho},\hat{\omega},\hat{\zeta}) \in V_c \times V_c \times V_c$ which are related to the staggered variables by the interpolation constraint: $\hat{\rho}=Q_t(\rho)$, $\hat{\omega}=Q_x(\omega)$, $\hat{\zeta}=\zeta$ where $Q_t : V_t \to V_c$ and $Q_x : V_x \to V_c$ are interpolation operators

$$Q_t(\rho)_{i,j} = \frac{1}{2}(\rho_{i,j}+\rho_{i+1,j}), \qquad Q_x(\omega)_{i,j} = \frac{1}{2}(\omega_{i,j}+\omega_{i,j+1}) \tag{4.1.3}$$

which are linear if no boundary conditions are given, affine otherwise. This constraints states exactly that the centered variables should be the cell averages of the continuous densities $(\tilde{\rho},\tilde{\omega},\tilde{\zeta})$

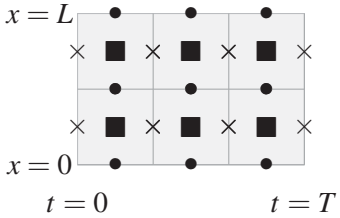



Figure 4.1.: (■) centered $G_c$, (×) staggered in time $G_t$ and (•) staggered in space $G_x$ grids. The grey rectangle is the space-time domain and here $N_t=3$, $N_x=2$.

built by interpolation from the discrete staggered variables $(\rho, \omega, \zeta)$ as explained above. The action functional is then approximated by (we drop the $h_t h_x$ factor which do not change the minimizers):

$$F(\hat{\rho}, \hat{\omega}, \hat{\zeta}) := \sum_{i,j} f(\hat{\rho}_{i,j}, \hat{\omega}_{i,j}, \hat{\zeta}_{i,j})$$

### Discrete minimization problem

This discretization scheme leads to replacing (4.1.1) by finite dimensional convex variational problem

$$\min \left\{ F(\hat{\rho}, \hat{\omega}, \hat{\zeta}) + \iota_{\{0\}}(A(\rho, \omega, \zeta)) + \iota_{\{(\hat{\rho}, \hat{\omega}, \hat{\zeta})\}}(Q_t(\rho), Q_x(\omega), \zeta) \right\} \tag{4.1.4}$$

where the variables are $(\rho, \omega, \zeta) \in V_t \times V_x \times V_c$ and $(\hat{\rho}, \hat{\omega}, \hat{\zeta}) \in V_c \times V_c \times V_c$. The source of inaccuracies with respect to the continuous problem (4.1.1) are (i) restriction of the to a certain finite dimensional class of densities which are piecewise constant or affine and (ii) averaging on each cell to compute the action.

## 4.1.2. Operator splitting algorithm

The structure of the problem, as a sum of non differentiable "simple" convex functions, makes it amenable to operator splitting algorithms. There are several algorithms known to solve such problems with comparable convergence properties (see [123] for a benchmark for solving the dynamic formulation of optimal transport). In this chapter, we focus on an implementation of the Douglas Rachford algorithm [48, 124] (equivalent to the *alternating direction method of multipliers*). This algorithm allows to solve problems of the form

$$\min_{x \in \mathbb{R}^N} g_1(x) + g_2(x)$$

when $g_1$ and $g_2$ are proper, l.s.c. convex functions that are "simple" in the sense that their (squared-Euclidean) proximal mapping

$$\mathrm{prox}_{g_i}(\bar{x}) = \arg\min_{x \in \mathbb{R}^N} g_i(x) + \frac{1}{2}\|x - \bar{x}\|^2$$

is easy to compute numerically. This algorithm builds a sequence $(z^{(\ell)})_{\ell \in \mathbb{N}}$ as follows. Let $z^{(0)} \in \mathbb{R}^N$ and define for $\ell \geq 0$:

$$\begin{aligned} x^{(\ell)} &= \mathrm{prox}_{\gamma g_1}(z^{(\ell)}), \\ y^{(\ell)} &= \mathrm{prox}_{\gamma g_2}(2x^{(\ell)} - z^{(\ell)}), \\ z^{(\ell+1)} &= z^{(\ell)} + 2\alpha(y^{(\ell)} - x^{(\ell)}). \end{aligned}$$

where $0 < \alpha < 1$ and $\gamma > 0$ are parameters (corresponding, respectively, to an operator averaging and a step size). If there exists $x$ such that $0 \in \partial g_1 + \partial g_2$, then $z^{(\ell)}$ converges to a minimizer of $g_1 + g_2$.

Coming back to our problem (4.1.4), one may for instance pose

$$g_1(\rho,\omega,\zeta,\hat{\rho},\hat{\omega},\hat{\zeta}) = F(\hat{\rho},\hat{\omega},\hat{\zeta}) + \iota_{\{0\}}(A(\rho,\omega,\zeta))$$
$$g_2(\rho,\omega,\zeta,\hat{\rho},\hat{\omega},\hat{\zeta}) = \iota_{\{(\hat{\rho},\hat{\omega},\hat{\zeta})\}}(Q_t(\rho),Q_x(\omega),\zeta),$$

which is the splitting that we have implemented. The remaining task before practical implementation is to compute the associated proximal maps.

## 4.2. Computing proximal maps

### 4.2.1. Projection maps

In case the function $g$ is the indicator of a set, the proximal operator $\mathrm{prox}_{\gamma g}$ becomes the projection operator on that set. The algorithm involves thus to compute the projections on:

- the continuity constraint, i.e. $\{(\rho,\omega,\zeta) \in V_t \times V_s \times V_c \,;\, A(\rho,\omega,\zeta) = 0\}$ ;
- the interpolation constraints, i.e. $\{(\rho,\hat{\rho}) \in V_t \times V_c \,;\, \hat{\rho} = Q_t(\rho)\}$ (and similarly for $(\omega,\hat{\omega})$).

One also has the projection on the equality constraint $\hat{\zeta} = \zeta$ which is trivially obtained as $((\hat{\zeta}+\zeta)/2,(\hat{\zeta}+\zeta)/2)$.

#### General projection formula

Let us start with the general, classical, projection formula on affine subspaces.

**Proposition 4.2.1.** *Given $\bar{x} \in \mathbb{R}^N$, a matrix $A : \mathbb{R}^{N\times M}$ and $b \in \mathrm{im}A$, the solution to*

$$x \in \arg\min \left\{ |x-\bar{x}|^2 \,;\, Ax = b \right\}$$

*is unique and given by $x = \bar{x} - A^*\lambda$ where $\lambda = (AA^*)^{-1}(A\bar{x}-b)$ is a Lagrange multiplier (not unique if $AA^*$ is not invertible).*

*Proof.* The dual problem reads $\min_\lambda \frac{1}{2}|A^*\lambda|^2 - \langle \lambda, A\bar{x}-b\rangle$ and the qualification constraint is satisfied. Any $\lambda$ satisfying $(AA^*)\lambda = A\bar{x}-b$ is optimal and the primal dual relationship yields $x = \bar{x} - A^*\lambda$ at optimality. □

With a similar proof, we obtain the following proposition.

**Proposition 4.2.2.** *Given $\bar{x} \in \mathbb{R}^N$, $\bar{y} \in \mathbb{R}^M$ and a matrix $Q \in \mathbb{R}^{N\times M}$, the solution to*

$$x \in \arg\min\{ |x-\bar{x}|^2 + |y-\bar{y}|^2 \,;\, Qx = y\}$$

*is unique and given by $(x,y) = (\bar{x} - Q^*\lambda, \bar{y} + \lambda)$ where $\lambda = (QQ^* + I)^{-1}(A\bar{x}-\bar{y})$.*

### Solving in frequency domain

According to these formula, the computation of projections involves finding a pre-image of a symmetric matrix $S \in \mathbb{R}^{N\times N}$, i.e. given $u \in \mathbb{R}^N$ find $\lambda \in \mathbb{R}^N$ such that $S\lambda = u$. If there exists an extension operator $u \in \mathbb{R}^N \mapsto u^{(e)} \in \mathbb{R}^{\mathbb{N}}$ and a time invariant linear filter $L : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ such that for some $M \geq N$,

$$u^{(e)} \text{ is } M\text{-periodic} \quad \text{and} \quad u_i^{(e)} = u_i, \quad (Su)_i = (Lu^{(e)})_i \quad \text{for all } i \in \{1,\dots,N\}. \tag{4.2.1}$$

then this can be solved efficiently in $O(N\log N)$ operations by pointwise multiplication in the frequency domain, because the discrete Fourier basis $\{n \mapsto e^{2i\pi(k+1)n/N}\}_{k=1}^N$ diagonalizes $L$ [107]. In other words, one looks for a periodic extension of $u$ where $Su$ appears as the result of a convolution, which is then easy to invert in the Fourier domain.

**Proposition 4.2.3** (Inversion in spectral domain)**.** *Let $\lambda, u \in \mathbb{R}^n$ be such that*

$$w_k \hat{\lambda}_k^{(e)} = \hat{u}_k^{(e)} \quad \textit{for all } k \in \{1,\dots,m\}$$

*where $(w_1,\dots,w_m)$ are the transfer coefficients of $L$, and $\hat{\cdot}$ denote the discrete Fourier transform. Then it holds $L\lambda^{(e)} = u^{(e)}$ and thus $S\lambda = u$.*

*Proof.* The first claim is classical in linear filter theory [107]. For the second, it holds $(S\lambda)_i = (L\lambda^{(e)})_i = u_i^{(e)} = u_i$, for $i \in \{1,\dots,n\}$. □

The choice of the periodic extension depends on the boundary conditions and on the filter: let us discuss our cases of interest.

### Continuity constraint

The operator $A$ defined in (4.1.2), its adjoint $A^*$ and $AA^*$ can be written as

$$A = \begin{pmatrix} D_t \mid D_x \mid -I \end{pmatrix}, \qquad A^* = \left(\begin{array}{c} D_t^* \\ \hline D_x^* \\ \hline I \end{array}\right) \quad \text{and} \quad AA^* = D_tD_t^* + D_xD_x^* + I.$$

where $D_t$ and $D_x$ are finite-difference operator with respect to the axis $t$ and $x$, respectively, with some specific boundary conditions. Let us focus on a single difference operator $D$ to precisely study the boundary conditions.

**Fixed boundary constraints** For fixed boundary constraints, one has[1]

$$D = \frac{1}{h}\begin{pmatrix} 0 & 1 & & \\ & -1 & 1 & \\ & & -1 & 0 \end{pmatrix} \quad \text{and} \quad DD^* = \frac{1}{h^2}\begin{pmatrix} 1 & -1 & \\ -1 & 2 & -1 \\ & -1 & 1 \end{pmatrix}.$$

The corresponding filter is (minus) the Laplacian $L(u)_i = -u_{i-1} + 2u_i - u_{i+1}$ and the extension is of period $M = 2N$ and satisfies $u_{1-i}^{(e)} = u_i$ (even extension w.r.t. $\frac{1}{2}$ and $N + \frac{1}{2}$). In this case, $\hat{u}^{(e)}$ might be directly obtained with the DCT-II transform (inverted by the DCT-III divided by $m$).

[1] it is implicit in our notation that the middle row is repeated diagonally as many times as necessary.

**Proposition 4.2.4.** *The transfer coefficients of $DD^*$ are $w_1 = 0$ and $w_k = \frac{1}{h^2}(2 - 2\cos(\pi(k-1)/N))$ for $k \in \{2,\dots,N\}$.*

In the higher dimensional case, one simpler filters independently along each dimension. Note that $w_1 = 0$ implies that the linear filter is sometimes not invertible: this correspond to cases where one imposes inconsistent boundary conditions (e.g. a continuity equation without source but marginals with unequal masses).

**Free boundary constraints** For free boundary conditions, one has

$$\bar{D} = \frac{1}{h}\begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & -1 & 1 \end{pmatrix} \quad \text{and} \quad \bar{D}\bar{D}^* = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{pmatrix}.$$

Again, $\bar{D}\bar{D}^*$ corresponds to (minus) the Laplacian on an infinite grid with the extension being as follows (and corresponds to DST-I transform): $M = 2(N+1)$ and

$$u_0^{(e)} = 0 \quad \text{and} \quad u_{-i}^{(e)} = u_i^{(e)}.$$

**Proposition 4.2.5.** *The transfer coefficients of $\bar{D}\bar{D}^*$ are obtained, for $k \in \{1,\dots,N\}$, as $w_k = \frac{1}{h^2}(2 - 2\cos(k\pi/(N+1)))$.*

**Projection on continuity constraint with source term** With the source term, as can be seen from the expression of $AA^*$, we simply add 1 to the transfer coefficients. Given the previous results, it is easy to combine various filters along each axis: one successively compute the transform associated to the boundary constraint along each axis and sum all the transfer coefficients. To fix ideas, the projection of $(\rho_0, \omega_0, \zeta_0)$ on the continuity equation with fixed boundary constraints is obtained as follows:

1. replace the boundary values by the boundary constraints;
2. define $u = A(\rho_0, \omega_0, \zeta_0) \in V_c$ (since the fixed values are part of $(\rho_0, \omega_0)$, $b$ does not appear explicitly);
3. define $\hat{u} \in V_c$ by applying the DCT-II transform along axis $t$ and $x$;
4. define $\hat{\lambda} \in V_c$ by solving, in the frequency domain
$$\left[\frac{1}{h^2}(2 - 2\cos(\pi(k-1)/N)) + 1\right]\hat{\lambda}_k = \hat{u}_k.$$
(the source term adds 1 to the transfer coefficients so they never vanish).
5. invert the transform to obtain $\lambda$ and the projection is $(\rho, \omega, \zeta) = (\rho_0, \omega_0) - A^*\lambda$.


(a) Time boundaries (b) Projected $\rho$ (c) Projected $\omega$ (d) Projected $\zeta$

Figure 4.2.: Projection of 0 on the continuity equation with source constraint with time boundary conditions shown on (a) and no flux boundary conditions on $\omega$. For (b)-(d) time is horizontal and space vertical as in Figure 4.1. For $\rho$ the colormap ranges from 0 (blue) to 1 (yellow) and for $\omega, \zeta$ it ranges form $-1$ (blue) to 0 (yellow) to 1 (red).

### Interpolation constraint

The projection formula that is relevant for the interpolation constraint is in Proposition 4.2.2 with the operator $Q$ defined in (4.1.3).

**Fixed boundary conditions** For fixed boundary conditions, one has

$$Q = \frac{1}{2}\begin{pmatrix} 0 & 1 & & \\ & 1 & 1 & \\ & & 1 & 0 \end{pmatrix} \quad \text{and} \quad QQ^* = \frac{1}{4}\begin{pmatrix} 1 & 1 & \\ 1 & 2 & 1 \\ & 1 & 1 \end{pmatrix}.$$

The matrix $(QQ^* + I)$ corresponds to the filter $u_i \mapsto \frac{1}{4}(u_{i-1} + 6u_i + u_{i+1})$ if one assumes *odd* extension with respect to the indices $\frac{1}{2}$ and $N + \frac{1}{2}$. This corresponds to DST-II, invertible with DST-III (and division by $2N$).

**Proposition 4.2.6.** *The transfer coefficients of* $(QQ^* + I)$ *are* $w_N = 1$ *and* $w_k = \frac{3}{2} + \frac{1}{2}\cos(k\pi/N)$ *for* $k \in \{1, \ldots, N-1\}$.

**Free boundary conditions** In this case, one has,

$$\bar{Q} = \frac{1}{2}\begin{pmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \end{pmatrix} \quad \text{and} \quad \bar{Q}\bar{Q}^* = \frac{1}{4}\begin{pmatrix} 2 & 1 & \\ 1 & 2 & 1 \\ & 1 & 2 \end{pmatrix}.$$

The matrix $(\bar{Q}\bar{Q}^* + I)$ corresponds to the filter $u_i \mapsto \frac{1}{4}(u_{i-1} + 6u_i + u_{i+1})$ if one assumes *odd* extension with respect to the indices 0 and $N + 1$. This corresponds to DST-I, invertible with itself (and division by $2(N+1)$).

**Proposition 4.2.7.** *The transfer coefficients of* $(\bar{Q}\bar{Q}^* + I)$ *are* $w_k = \frac{3}{2} + \frac{1}{2}\cos(k\pi/(N+1))$ *for* $k \in \{1, \ldots, N\}$.

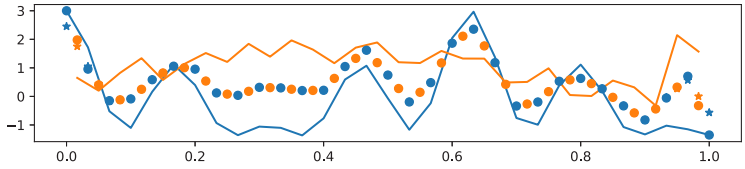



Figure 4.3.: Projection on the interpolation constraints in 1d computed numerically. Blue stands for staggered and orange for centered values. (lines) initial values. (•) projection with fixed boundary values (∗) projections with free boundary values.

### 4.2.2. Computing proximal maps

It remains to give explicit formula for $\mathrm{prox}_{\gamma F}$. Since $F$ is separable, one has for $(\hat{\rho},\hat{\omega},\hat{\xi}) \in V_c \times V_c^d \times V_c$

$$\mathrm{prox}_{\gamma F}(\hat{\rho},\hat{\omega},\hat{\xi}) = \left(\mathrm{prox}_{\gamma f}(\hat{\rho}_{i,j},\hat{\omega}_{i,j},\hat{\xi}_{i,j})\right)_{i,j}$$

so knowing how to compute $\mathrm{prox}_{\gamma f}(x,y,z)$ for $(x,y,z) \in \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$ is sufficient for computing $\mathrm{prox}_{\gamma F}$. Moreover, if $f$ is of the form $f_1(x,y)+f_2(z)$ or $f_1(y)+f_2(x,z)$ then one may again compute the proximal maps separately. Some interesting cases are

(i) given a differentiable, superlinear and convex Lagrangian $L:\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$, the function $f(x,y,z)=\psi_L(y,z|x)$. This corresponds to the general class of problems introduced in Chapter 1. The case where the Lagrangian is $L(v,g)=|v|^2+g^2/4$ corresponds to the key case of the $\widehat{W}_2$ metric.

(ii) the functions $f(x,y,z)=f_1(x,y)+|z|^p$ for $p\geq 1$ where $f_1$ sublinear penalize the source term independently. The case $p=1$, which corresponds to optimal partial transport (see Chapter 2). The case $p=2$ has been proposed by [106].

(iii) the functions $f(x,y,z)=|y|+f_2(x,z)$ correspond to a class of problems studied in [147] where minimal flow methods are proposed.

Let us derive the proximal maps of these "building block" functions.

#### Power functions

For power functions, the proximal map is easy to compute using the optimality conditions.

#### Perspective of Lagrangians

In the models introduced in Chapter 1 we consider a function $f(x,y,z)=\psi_L(y,z|x)$ which is the perspective function (see Appendix A) of a convex, continuous Lagrangian $L:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}$.

| $f(y)$ | $\mathrm{prox}_{\gamma f}(y)$ |
|---|---|
| $\lvert y\rvert$ | $\mathrm{sign}(y)(\lvert y\rvert-\gamma)_+$ |
| $\lvert y\rvert^2$ | $y/(1+2\gamma)$ |
| $\lvert y\rvert^p,\ p\geq 1$ | $\mathrm{sign}(y)u$, where $u\geq 0$ and $u+p\gamma u^{p-1}=\lvert y\rvert$ |

Table 4.1.: Proximal operator associated to convex power functions.

**General case** The following result is proved in [47].

**Proposition 4.2.8** (Proximal of a perspective)**.** *Let $L$ be a convex, superlinear, and differentiable Lagrangian on $\mathbb{R}^N$, then*

$$\mathrm{prox}_{\gamma\psi_L}(x,y)=\begin{cases}(0,0) & \text{if } x+\gamma L^*(y/\gamma)\leq 0\\ (x+\gamma L^*(u),y-\gamma u) & \text{otherwise,}\end{cases}$$

*where $u$ is uniquely characterized by $y=\gamma u+(x+\gamma L^*(u))\nabla L^*(u)$.*

**Quadratic Lagrangian** The most important case is $L(v)=|v|^2/2$. Specializing Proposition 4.2.8 to this case leads to the formula

$$\mathrm{prox}_{\gamma\psi_L}(x,y)=\begin{cases}(0,0) & \text{if } x+|y|^2/(2\gamma)\leq 0,\\ (x,0) & \text{if } y=0 \text{ and } x\geq 0,\\ \left(x+\frac{\gamma}{2}\frac{|y|^2}{\alpha^2},y(1-\frac{\gamma}{\alpha})\right) & \text{otherwise,}\end{cases}$$

where $\alpha$ is the unique real solution to the third order polynomial equation in $X$

$$X^3-X^2(\gamma+x)-\gamma|y|^2/2=0.$$

We may then use the fact that we know that this root is the unique real one to reduce the amount of computations for finding the root. After the change of variable $x^*=x+\frac{\gamma}{2}\frac{|y|^2}{\alpha^2}$, one has the equivalent characterization which was proposed in [123]

$$\mathrm{prox}_{\gamma\psi_L}(x,y)=\begin{cases}(x^*,\frac{x^*y}{x^*+\gamma}) & \text{if } x^*>0\\ (0,0) & \text{otherwise,}\end{cases}$$

where $x^*$ is the largest real root of the third order polynomial equation in $X$

$$(X-x)(X+\gamma)^2-\gamma|y|^2/2.$$

**Quadratic Lagrangian with weights** Note that for computing $\hat{W}_2$ interpolations, one generally has a quadratic Lagrangian with a weight

$$f(x,y,z) = \psi_L(\omega|\rho) + \alpha^2 \psi_L(\zeta|\rho).$$

We suggest to take advantage of the rescaling property proved in Proposition 2.1.3 to deal with this case without effort using the previous formula: first pushforward $(\mu,\nu)$ with the spatial rescaling $x \mapsto x/\alpha$, solve the problem between these new marginals with $\alpha = 1$ (which is simpler) and pushforward the solution (with a rescaled momentum) $(\rho, \alpha\omega, \zeta)$ with the inverse map $x \mapsto \alpha x$ to obtain the solution to the original problem. Numerically, applying this pushforward is as simple as changing the lengths of the domain and multiplying the density by a scalar. This trick applies more generally to power Lagrangians.

**Power Lagrangians** For other Lagrangians of the form $L(v) = |v|^p$ with $p > 1$ (the case $p = 1$ is covered by the paragraph on power functions), one has a similar characterization, posing $q = p/(p-1)$ the conjugate exponent:

$$\mathrm{prox}_{\gamma\psi_L}(x,y) = \begin{cases} (0,0) & \text{if } x + |y|^q/(q\gamma^{q-1}) \leq 0, \\ (x,0) & \text{if } y = 0 \text{ and } x \geq 0, \\ \left(x + \frac{\gamma}{q}\frac{|y|^q}{\alpha^q}, y(1-\frac{\gamma}{\alpha})\right) & \text{otherwise,} \end{cases}$$

where $\alpha$ is the unique real solution to

$$X^{2q-1} - \gamma X^{2q-2} - (xX^q + \gamma|y|^q/q)|y|^{q-2} = 0.$$

a good approximation of which can be found by Newton's method in a few iterations using the fact that $0 < \alpha < \gamma$.

## 4.3. Numerical results

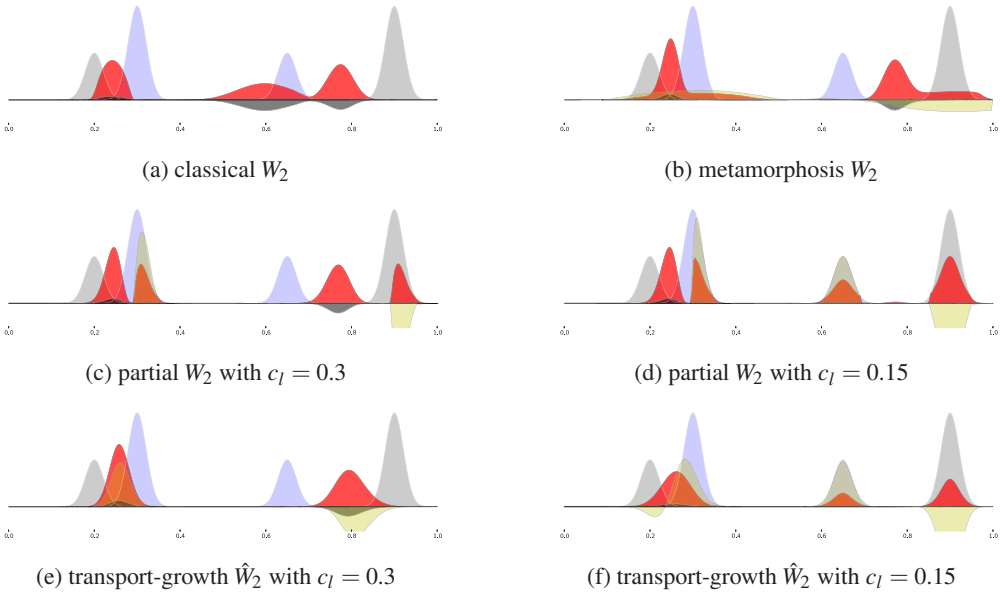
**Transport of Gaussian bumps.** We first consider the interpolation between two measures $(\mu,\nu)$ which have the same mass on $\Omega = [0,1]$ with Neumann boundary conditions. The result is shown on Figure 4.4. The initial and the final measures are both composed of two Gaussian densities of mass 1 and 2 centered at, from left to right, $x = 0.2$, 0.3, 0.65 and 0.9. The problem is discretized into $N_t \times N_x = 11 \times 256$ samples. We observe that the behavior of geodesics is highly dependent on the choice of action and of "cut-locus" parameter $c_l$. The action functionals considered are, with $L(v) = |v|^2$ and $\psi_L$ its perspective function (see Appendix A):

**classical** $W_2$: $f(\rho,\omega,\zeta) = \psi_L(\omega|\rho) + \iota_{\{0\}}(\zeta)$;

**metamorphosis** $W_2$: $f(\rho,\omega,\zeta) = \psi_L(\omega|\rho) + |\zeta|^2$;

**partial** $W_2$: $f(\rho,\omega,\zeta) = \psi_L(\omega|\rho) + (c_l/2)|\zeta|$;

**transport growth** $\hat{W}_2$: $f(\rho,\omega,\zeta) = \psi_L(\omega|\rho) + (8c_l/\pi)\psi_L(\zeta|\rho)$.



(a) classical $W_2$ (b) metamorphosis $W_2$

(c) partial $W_2$ with $c_l = 0.3$ (d) partial $W_2$ with $c_l = 0.15$

(e) transport-growth $\hat{W}_2$ with $c_l = 0.3$ (f) transport-growth $\hat{W}_2$ with $c_l = 0.15$

Figure 4.4.: Interpolations between two densities $(\mu, \nu)$ of equal mass on the segment $[0,1]$. (gray) $\mu$, (blue) $\nu$, (red) $\rho_{t=1/2}$, (black) $\omega_{t=1/2}$, (yellow) $\zeta_{t=1/2}$. We indicate by $c_l$ the theoretical maximum distance a Dirac can travel (if relevant).

For the classical $W_2$ interpolation on Figure 4.4a, conservation of mass enforces the bumps to split, yielding a somehow unnatural interpolation. The effect of a non-homogeneous functional is visible on Figure 4.4b. The non-homogeneity of the action requires to choose of a reference measure (here the discrete Lebesgue measure) and spread densities. Consider now the partial $W_2$ interpolations on Figure 4.4c. The domain $\Omega = [0,1]$ is split in an active set where mass is transported and an inactive set where the source is nonzero. Note that interpolations in that case are not unique on the inactive set and the result depends on the initialization of the optimization algorithm. Finally, Figures 4.4e and 4.4f display the interpolation at $t = 1/2$ for the $\hat{W}_2$ with different values of cut-loci. In the first case, we obtain a geodesic consisting of two traveling bumps which inflate or deflate while in the second case, only one bump travels as $c_l$ is now smaller than the distance between the two bumps on the right.

**Synthetic 2D experiments.** We now interpolate between two densities $(\mu, \nu)$ of equal mass on the domain $\Omega = [0,1]^2$. The initial density $\mu$ is the indicator of the ring centered at $(\frac{1}{2}, \frac{1}{2})$, internal diameter 0.5 and external diameter 0.7. The final density $\nu$ is obtained from $\mu$ by a random smooth deformation. The domain is discretized in $N_t \times N_{x_1} \times N_{x_2} = 12 \times 64 \times 64$ samples. We compare on Figure 4.5 the interpolations for the Hellinger metric (defined in Remark 2.1.19 and 3 different actions (as defined in the previous paragraph): classical $W_2$, partial $W_2$ (with $c_l = 0.2$) and $\hat{W}_2$ (with $c_l = 0.4$). Notice that the two first rows thus show the limit geodesics for $\hat{W}_2$ when $c_l$ tends to 0 or $\infty$ (see Section 2.1.2).

Figure 4.6 helps to better understand how the mass moves during interpolations. On the

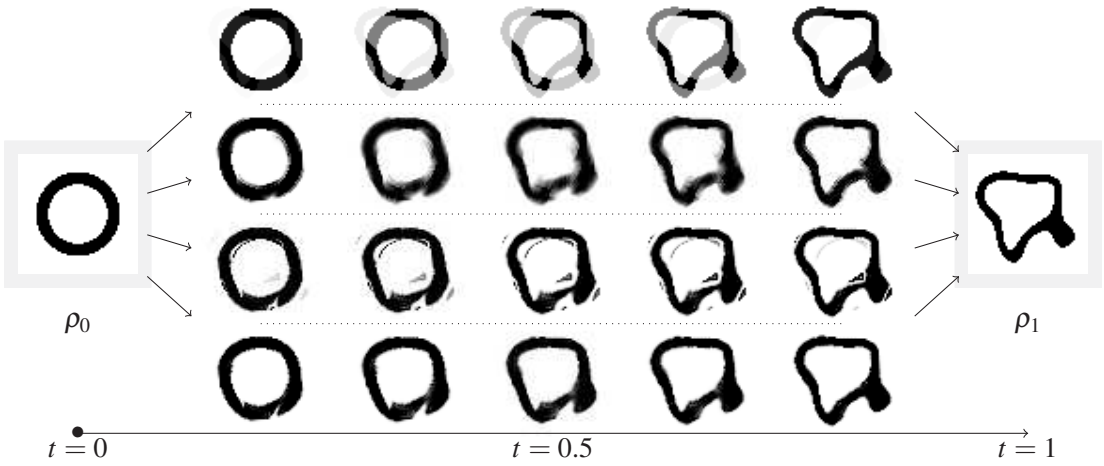


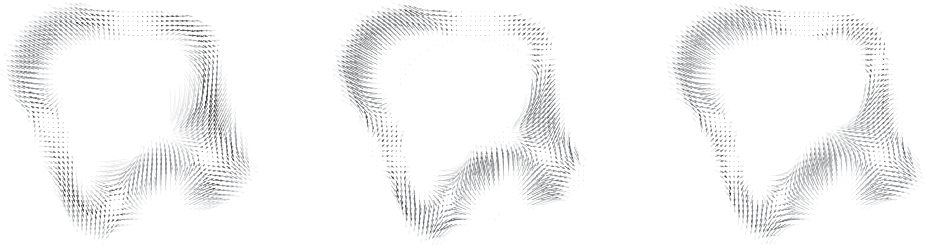

Figure 4.5.: Interpolation between $\mu$ and $\nu$ for four metrics: Hellinger (1st row), $W_2$ (2nd row), partial $W_2$ (3rd row), $\hat{W}_2$ (4th row).

top row, the velocity field shows that the classical $W_2$ interpolation transports a lot of mass to the bottom left protuberance in order to balance mass. On the contrary, actions allowing local variations of mass attenuate the component of the velocity field which is tangential to the ring. Finally, by looking at the source maps in the bottom row, we see the inactive sets of partial optimal transport with well defined boundaries and how this strongly differs to the smooth source related to $\hat{W}_2$ interpolation.

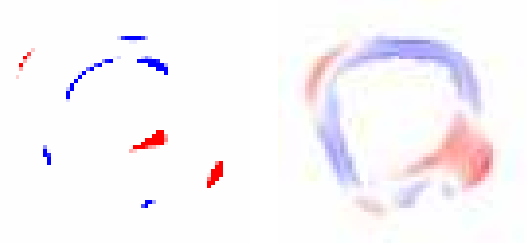
Figure 4.7 displays another experiment intended to again illustrate the behavior of geodesics. The measures $\mu$ and $\nu$ have same total mass, the domain $\Omega = [0,1]^2$ is discretized into $N_t \times N_{x_1} \times N_{x_2} = 12 \times 44 \times 44$ cells. We show the interpolation for classical $W_2$, partial $W_2$ (with $c_l = 0.5$) and for the transport-growth $\hat{W}_2$ (with $c_l = 0.5$).

**Biological images interpolation.** Interpolating between shapes with varying masses was our original motivation for introducing the dynamic formulation of unbalanced optimal transport. This is the object of our last numerical experiment. The initial and final densities $\mu, \nu$ represent a segmented brain taken at different times. This example is rather challenging for image matching algorithms: matter is creased, folded and grows unevenly. The spatial domain is $\Omega = [0, 0.82] \times [0,1]$ and the time/space domain is discretized into $12 \times 89 \times 73$ cells. We display the interpolation for classical $W_2$ and for the transport-growth metric $\hat{W}_2$ with $c_l = 0.2$. Figure 4.8 displays the interpolations and Figure 4.9 displays the velocity and rate of growth fields at time $t = 1/2$.

Most of the tissue growth is located at the bottom of the domain. Consequently, the velocity field associated to the $W_2$ geodesic is dominated by top to bottom components as observed on Figure 4.9a. To the contrary, the interpolation for $\hat{W}_2$ locally adapts the rate of growth, and the velocity field is arguably more consistent to the underlying true evolution. However, several artifacts inherent to optimal transport models are retained: some matter is teared off the brain lining and brought somewhere else to fill a need of mass. This behavior is clearly observed on



(a) classical $W_2$ (b) partial $W_2$ (c) transport-growth $\hat{W}_2$



(d) partial $W_2$ (e) transport-growth $\hat{W}_2$

Figure 4.6.: First row: velocity field $v_t = \omega_t / \rho_t$ at time $t = 1/2$. The higher the density $\rho_t$, the darker the arrow. Second row: source $\zeta_t$ at time $t = 1/2$. Blue stands for negative density and red for positive.

Figure 4.8 near the bulges that appear on the right and left sides of the brain. As a final remark, it is proved in [94, 103] that the optimality conditions for $\widehat{W}_2$ imply that the velocity field is the gradient of the rate of growth: this fact is observed numerically by comparing 4.9b and Figure 4.9c.

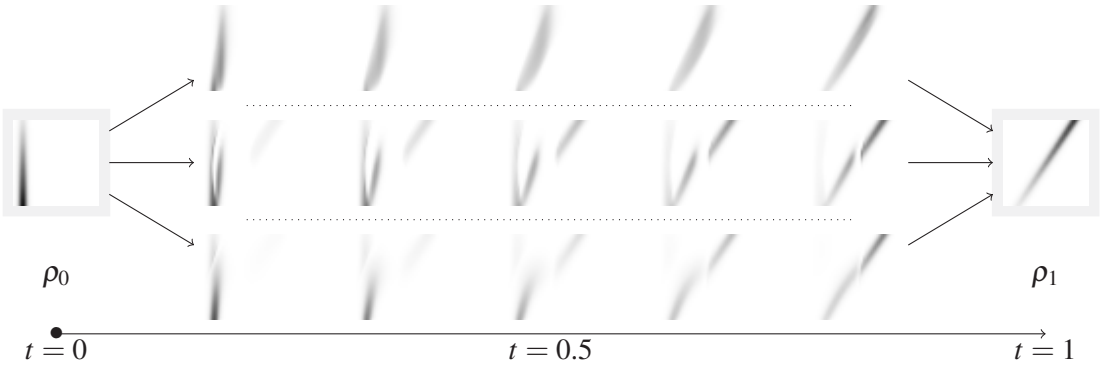



Figure 4.7.: Geodesics between $\mu$ and $\nu$ for various actions. (1st row) classical $W_2$, (2nd row) partial $W_2$ and (3rd row) transport-growth $\hat{W}_2$ . Notice that the mass $\mu$ is concentrated at the bottom left that of $\nu$ is concentrated at top right.

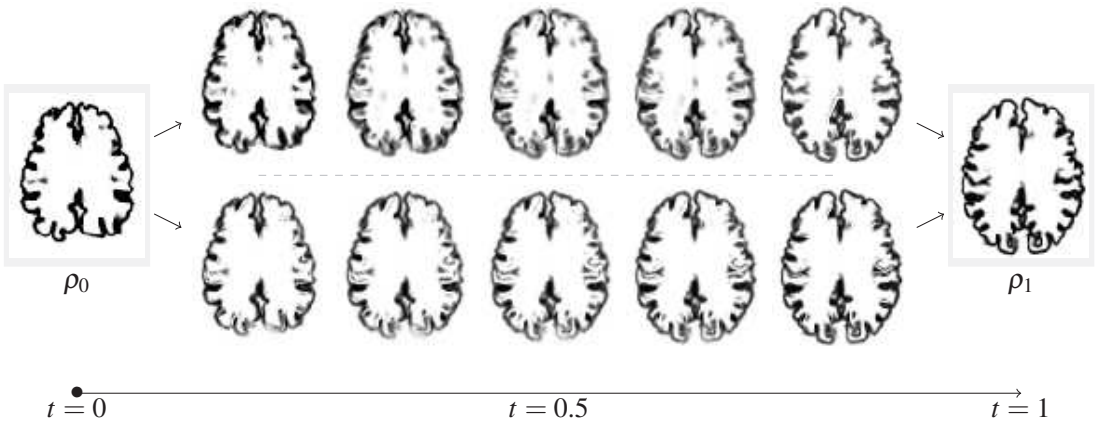
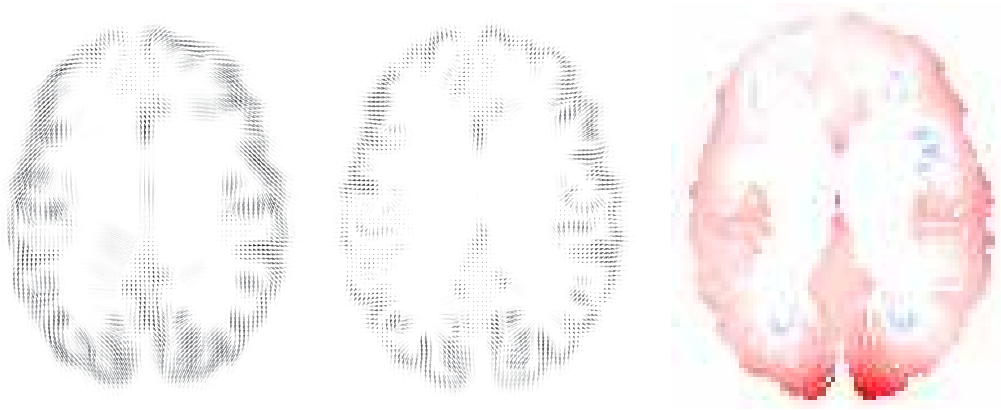



Figure 4.8.: Interpolation between $\mu$ and $\nu$ for two metrics. (top row) classical $W_2$ between the rescaled densities and (bottom row) transport-growth $\hat{W}_2$ with $c_l = 0.2$.



(a) $v_{0.5}$ for $W_2$ (b) $v_{0.5}$ for $\mathrm{WF}_\kappa$ (c) $g_{0.5}$ for $\mathrm{WF}_\kappa$

Figure 4.9.: (a) and (b): velocity field $v_t = \omega_t / \rho_t$ of the interpolation at time $t = 1/2$ (the higher the density of $\rho_{0.5}$, the darker the arrow). (c) rate of growth $g_t = \zeta_t / \rho_t$ at time $t = 1/2$. Blue stands for negative density and red for positive and $g_t$ is set to zero when $\rho_t \approx 0$.

# Part III.

# Applications

# Chapter 5.

# Illustrations for Scaling Algorithms

This chapter is the numerical and applied counterpart of Chapter 3. We review several problems that fit in the framework of generic formulations, derive the explicit form of the scaling iterates and display illustrations.
In Section 5.1, we compute optimal entropy-transport couplings for 4 choices of $f$-divergences: exact marginal constraints, relative entropy, total variation, and range constraint. We display 1-D and 2-D experiments as well as a color transfer experiment where the unbalanced framework comes naturally.
In Section 5.2, we derive the explicit form of the scaling algorithm for various unbalanced optimal transport barycenter problems. This is applied to the computation of geodesics for $\widehat{W}_2$, to barycenters in 1-D as well a as a comparison between balanced and unbalanced barycenters in 2-D.
In Section 5.3, we explain how to use scaling algorithms for computing time discretized gradient flows w.r.t. the optimal transport of unbalanced optimal transport. This is applied to the challenging problem of computing Wasserstein gradient flow of the functional "total variation of the gradient".
The content of this chapter was published in [45], except the gradient flow experiment which is new.

## Foreword: discrete densities

In this chapter, we consider the discretization of problems which have a continuous nature. In this situation, it is natural to work with dimensionless quantities which do not overly depend on the discretization: this suggests to deal with densities instead of measures. This leads to a change of variable which induces a minor modification of Algorithm 1. Consider $(X, \mathrm{d}x)$ and $(Y, \mathrm{d}y)$ the discretization of two domains, endowed with their discretized reference measures. We define the reference measure on the product space $\gamma^{(0)} = \mathrm{d}x \otimes \mathrm{d}y$ and we parametrize the *density* of the unknown coupling as $(a_i K_{i,j} b_j)$. This gives the scaling algorithm in "density" variables displayed in Algorithm 7.

In this variant, we do not use the "proxdiv" operator of the functions that act on measures $F_i$, but that of the functions that act on densities $\bar{F}_i$, with the correspondence

$$\bar{F}_1(s) = F_1(s\mathrm{d}x) \text{ for } s \in \mathbb{R}^X \quad \text{and} \quad \bar{F}_2(s) = F_2(s\mathrm{d}y) \text{ for } s \in \mathbb{R}^Y$$

In the next paragraphs, we give explicit formula for the operators

$$\mathrm{prox}^H_{\bar{F}/\varepsilon} : \bar{s} \in (\mathbb{R}^I)^n \mapsto \arg\min \left\{ \bar{F}(s) + \varepsilon H(s|\bar{s}) \; ; \; s \in (\mathbb{R}^I)^n \right\}$$

and the associated "proxdiv" operators (3.3.5). We recall that the operator "proxdiv" is the building block of the scaling iterations and is defined, for a marginal function $F$, a regularization parameter $\varepsilon$ and variables $(s,u) \in (\mathbb{R}^I_+)^n \times (\mathbb{R}^I)^n$ as

$$\mathrm{proxdiv}_F(s,u,\varepsilon) := \mathrm{prox}^H_{F/\varepsilon}(e^{-u/\varepsilon} \odot s) \oslash s.$$

The variable $u$ is an offset that is only used in the stabilized version of the algorithm and the standard version is recovered by setting $u = 0$.

**Discretization adopted** In all the numerical experiments, we consider a domain $[0,1]^d$ for $d = 1, 2$ or 3, that is divided into cubic cells, the discrete space $X = Y$ is the grids of cell centers and the cost matrix is the evaluation of the cost function between pairs of cell centers.

**Algorithm 7** Scaling algorithm in discrete density variables

1. define the discretized reference measures $(\mathrm{d}x, \mathrm{d}y) \in \mathbb{R}^I \times \mathbb{R}^J$
2. initialize the kernel $K = (e^{-c_{i,j}/\varepsilon})_{i,j}$ and the iterates $(a,b) = (1_I, 1_J)$
3. while stopping criterion not satisfied repeat:
   a) $a \leftarrow \mathrm{proxdiv}_{\bar{F}_1}(K(b \odot \mathrm{d}y), 0_I, \varepsilon)$
   b) $b \leftarrow \mathrm{proxdiv}_{\bar{F}_2}(K^T(a \odot \mathrm{d}x), 0_J, \varepsilon)$
4. return $\gamma = (a_i K_{i,j} b_j)_{i,j}$, the density of the optimal plan w.r.t. $\mathrm{d}x \otimes \mathrm{d}y$

# 5.1. Unbalanced optimal transport

## 5.1.1. Derivation of the algorithm

Let $p \in \mathbb{R}_+^I$ be the discretized density of a measure $\mu$. Recalling Sections 3.1.1 and 3.1.2, the functions $\bar{F}_1$ and $\bar{F}_2$ involved in the definition of optimal transport and unbalanced optimal transport are of the form

$$\bar{F}(s) := \sum_{i=1}^{I} \psi_f(s_i|p_i) \tag{5.1.1}$$

where $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is an entropy function (Definition 1.2.15) and $\psi_f$ is the associated perspective function (see Appendix A). The first order optimality conditions for the operator $\mathrm{prox}_F^H$ are given in the next proposition.

**Proposition 5.1.1.** *Let $\bar{s} \in \mathbb{R}_+^I$ and consider $\bar{F}$ defined in* (5.1.1) *for some $p \in \mathbb{R}_+^I$. If it holds $0 \in \mathrm{dom}\, f$ or $\bar{s}_i = 0 \Rightarrow p_i = 0$, then $\mathrm{prox}_{F/\varepsilon}^H(\bar{s}) = (s_i)_i$ where for each $i \in \{1,\dots,I\}$ one has*

$$\begin{cases} s_i = 0 & \text{if } \bar{s}_i = 0, \\ s_i = \bar{s}_i \exp(-f'_\infty/\varepsilon) & \text{if } p_i = 0 \text{ and } \bar{s}_i > 0, \\ 0 \in \varepsilon \log(s_i/\bar{s}_i) + \partial f(s_i/p_i) & \text{otherwise.} \end{cases}$$

This formula allows to explicitly compute the "proxdiv" operators of the examples introduced in Section 3.1, as listed in Table 5.1. The names used in the leftmost column stand for the following cases:

OT: standard optimal transport, i.e. $f(s) = \iota_{\{1\}}(s)$;

$\mathrm{H}_\lambda$: relative entropy with weight $\lambda > 0$, i.e. $f(s) = \lambda(s\log s - s + 1)$;

$\mathrm{TV}_\lambda$: total variation with weight $\lambda > 0$, i.e. $f(s) = |s-1| + \iota_{\mathbb{R}_+}(s)$;

$\mathrm{RG}_{[\alpha,\beta]}$: range constraint with $0 \le \alpha \le \beta \le \infty$, i.e. $f(s) = \iota_{[\alpha,\beta]}(s)$.

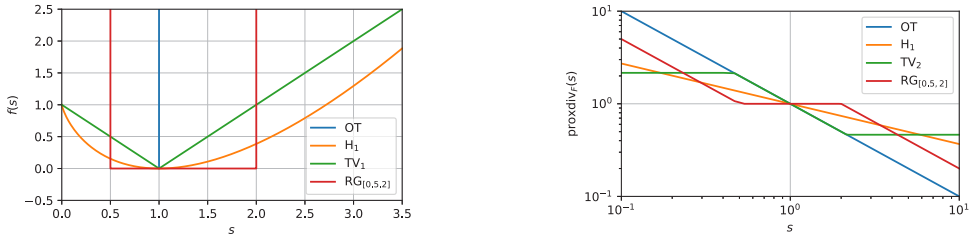
These entropy functions as well as the associated proxdiv operators are displayed on Figure 5.1. Remark that the slopes of proxdiv in log domain are smaller than 1: this corresponds to the non-expansiveness property proved in Lemma 3.5.20. In Table 5.1 the first line corresponds to standard Sinkhorn's iterations which are also recovered in the second and third line by letting $\lambda \to +\infty$ and by setting $\alpha = \beta = 1$ in the fourth line.

## 5.1.2. Numerical example in 1-D.

Consider the domain $[0,1]$ discretized into $I = 1000$ cells and consider the marginals displayed on Figures 5.2a-5.2b of discrete densities $p, q \in \mathbb{R}_+^I$. We run Algorithm 2 for marginal functions $F_i$ of the form (5.1.1) with the choices of entropy functions listed above and for the cost functions $c_{\mathrm{quad}}(x,y) = |y-x|^2$ or $c_{\ell,\alpha}(x,y) = -\log\cos_+^2(|y-x|\cdot 2/(\alpha\pi))$ which corresponds to the static cost in the definition of $\widehat{W}_2$ with maximum distance of transport $\alpha$.

| | $\mathrm{prox}^H_{F/\varepsilon}(s)$ | $\mathrm{proxdiv}_F(s,u,\varepsilon)$ |
|---|---|---|
| OT | $p$ | $p/s$ |
| $\mathrm{H}_\lambda$ | $s^{\frac{\varepsilon}{\varepsilon+\lambda}} \cdot p^{\frac{\lambda}{\varepsilon+\lambda}}$ | $(p/s)^{\frac{\lambda}{\lambda+\varepsilon}} \cdot e^{-u/(\lambda+\varepsilon)}$ |
| $\mathrm{TV}_\lambda$ | $\min\left\{s\cdot e^{\frac{\lambda}{\varepsilon}}, \max\left\{s\cdot e^{-\frac{\lambda}{\varepsilon}}, p\right\}\right\}$ | $\min\left\{e^{\frac{\lambda-u}{\varepsilon}}, \max\left\{e^{-\frac{\lambda+u}{\varepsilon}}, p/s\right\}\right\}$ |
| $\mathrm{RG}_{[\alpha,\beta]}$ | $\min\{\beta\, p, \max\{\alpha\, p, s\}\}$ | $\min\left\{\beta\, p/s, \max\left\{\alpha\, p/s, e^{-u/\varepsilon}\right\}\right\}$ |

Table 5.1.: Some divergence functionals and the associated prox and proxdiv operators for $s, p \in \mathbb{R}^I_+$ and $u \in \mathbb{R}^I$ (all operators act pointwise).




(a) Graphs of the entropy functions $f$ (more precisely: boundaries of the epigraphs).

(b) Operator $\mathrm{proxdiv}_F$ for $p = 1$, $u = 0$ and varying $s \in \mathbb{R}_+$ (in log scale)

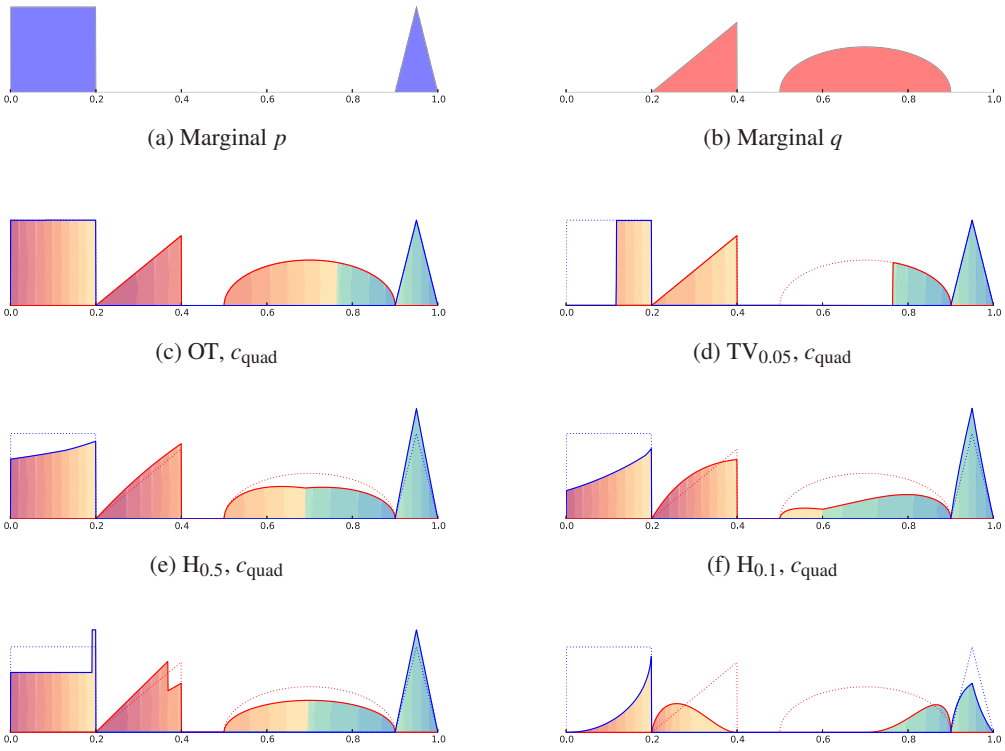
Figure 5.1.: $f$-divergences and proxdiv operators for the examples of Table 5.1

On Figure 5.2, we display the marginals $(\sum_j \gamma_{i,j} w_j)_i$ and $(\sum_i \gamma_{i,j} w_i)_j$ of the density $(\gamma_{i,j})$ output by the algorithm, with a color scheme that illustrates how mass is transported. Figure 5.3 displays the entries of $\gamma$ which are greater than $10^{-5}$. This can be interpreted as the approximate support of the optimal discrete plan. The stabilization procedure of Algorithm 2 allows to choose a small parameter $\varepsilon = 10^{-6}$ and to return a quasi-deterministic plan. Remark that for the TV case (orange support), the straight segments correspond to a unit density on the diagonal: this is the plan with minimal entropy among all the optimal plans as expected from Proposition 3.5.3.

Figure 5.4 displays the primal-dual gap $(P_\varepsilon) - (D_\varepsilon)$ as a function of the iteration number $\ell$ when running Algorithm 1 with $\varepsilon = 0.01$. More precisely, we display

$$\tilde{P}_\varepsilon(r^{(\ell)}) - \tilde{D}_\varepsilon(u^{(\ell)}, v^{(\ell)})$$

where $\tilde{P}_\varepsilon$ and $\tilde{D}_\varepsilon$ are defined from the primal (3.3.2) and dual (3.3.3) functionals by replacing the indicator of constraints by the exponential of the distance to the set, and where $r^{(\ell)}$, $(u^{(\ell)}, v^{(\ell)})$ are the primal and dual iterates as in Theorem 3.3.4. The marginals are the same as shown on Figures 5.2a-5.2b, the cost is $c_{\mathrm{quad}}$ and the number of discretization points is $I = 200$.



(a) Marginal $p$

(b) Marginal $q$

(c) OT, $c_{\text{quad}}$

(d) $\text{TV}_{0.05}$, $c_{\text{quad}}$

(e) $\text{H}_{0.5}$, $c_{\text{quad}}$

(f) $\text{H}_{0.1}$, $c_{\text{quad}}$

(g) $\text{RG}_{[0.7,1.2]}$, $c_{\text{quad}}$

(h) $\text{H}_1$, $c_{\ell,0.2}$ (corresponds to $\widehat{W}_{2,\alpha}$ with $\alpha = 0.4/\pi$)

Figure 5.2.: (a)-(b) Input marginals. (c)-(h) Marginals of the optimal plan $\gamma$ displayed together for the functions $F$ and cost functions as introduced in the text body (specified in the caption). The color shows the location of the same subset of mass before and after transportation.

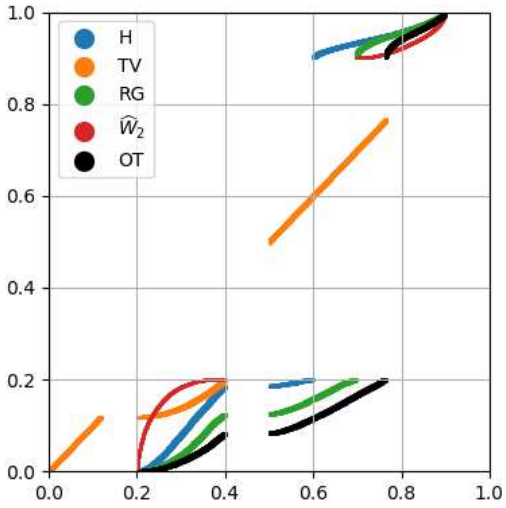



Figure 5.3.: Colored areas are such that $\gamma_{i,j} > 10^{-8}$ for all the examples displayed on Figure 5.2 (the legend $\widehat{W}_2$ corresponds to (h) and (e) is not represented). The first marginal is horizontal, the second vertical. Orange and black superimposed at the top.

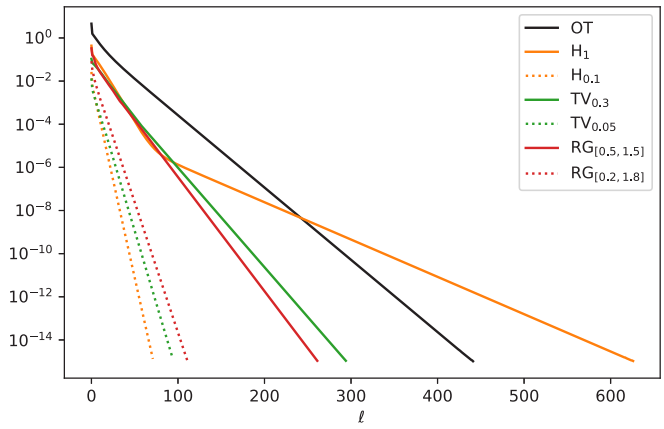
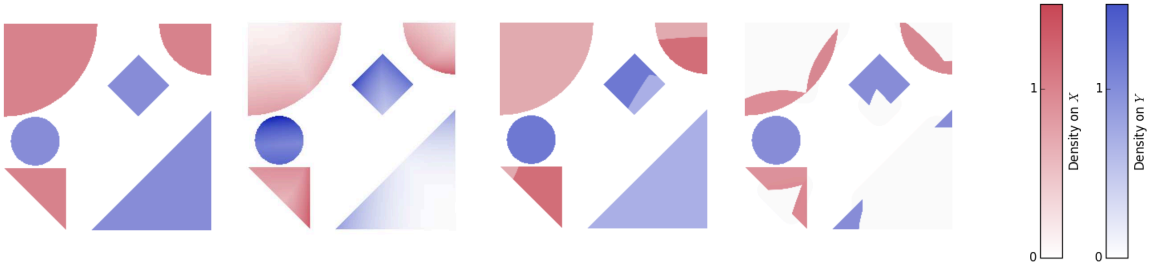



Figure 5.4.: Primal dual gap as a function of the iterations for Algorithm 1 applied to unbalanced optimal transport problems (indicator of sets replaced by exponential functions). Note that convergence can be accelerated (see Section 3.4).




(a) $p$ and $q$ (b) $\mathrm{H}_{0.1}$ (c) $\mathrm{RG}_{[0.7,1.2]}$ (d) $\mathrm{TV}_{0.05}$

Figure 5.5.: Marginals of the optimal plan $\gamma$ for several choices of $f$-divergences.

(a) OT (b) $\mathrm{H}_{0.1}$ (c) $\mathrm{RG}_{[0.7,1.2]}$ (d) $\mathrm{TV}_{0.05}$

Figure 5.6.: Representation of the transport plans for the experiments of Figure 5.5.

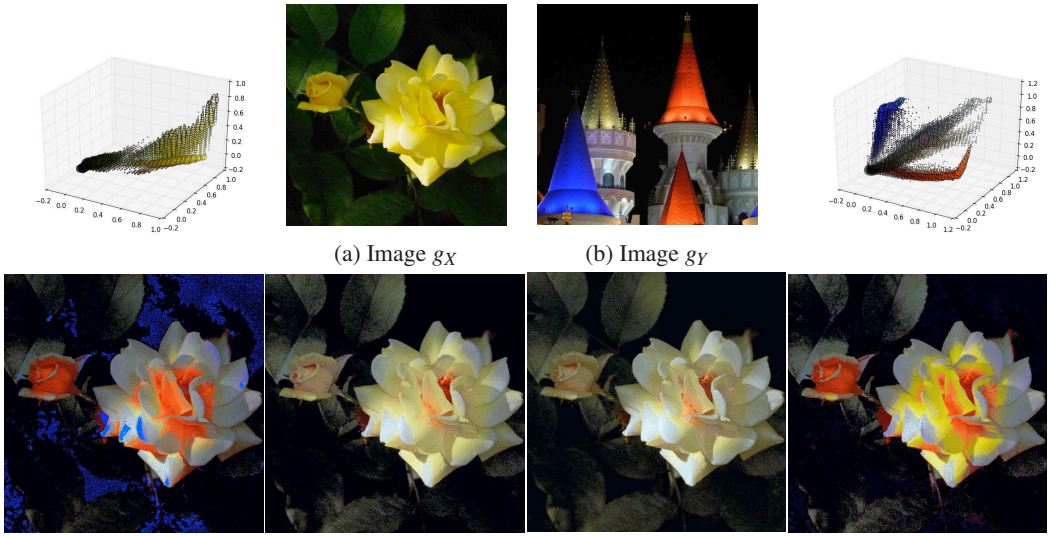
### 5.1.3. Numerical examples in 2-D.

Consider the domain $[0,1]^2$ discretized into $I = 200 \times 200$ cells, the marginals $p$, $q$ displayed together on Figure 5.5a and the quadratic cost $c_{\mathrm{quad}}(x,y) = |y-x|^2$. We run Algorithm 1 for several choices of $f$-divergences and use the separability of the matrix $K$ to decrease the complexity of each iteration (see Section 3.3.3). The regularization parameter $\varepsilon$ has been fixed to $10^{-4}$. Figure 5.5 shows the marginals of the optimal coupling $\gamma$ and Figure 5.6 illustrates the resulting transport plan: points with the same color correspond to the same mass particle before and after transport.

### 5.1.4. Color transfer

Color transfer is a classical task in image processing where the goal is to impose the color histogram of one image onto another image. Optimal transport between histograms has proven useful for problems of this sort such as contrast adjustment [54] and color transfer via 1-D optimal transport [131]. Indeed, optimal transport produces a correspondence between histograms which minimizes the total amount of color distortion, where the notion of distortion is specified by the cost function. It thus maintains visual consistency.

In our experiments we represent colors in the three-dimensional "CIE-Lab" space (one coordinate for luminance and two for chrominance), resized to fit into a cuboid $[0,1]^3$, discretized into



(a) Image $g_X$ (b) Image $g_Y$

(c) OT (d) $H_{0.03}$ (e) $RG_{[0,5]}$ (f) $TV_{0.2}$

Figure 5.7.: A challenging color transfer experiment where the colors of the image $g_Y$ are transferred to the image $g_X$ (histograms of colors are schematically displayed next to them). In all cases $F_1$ is a hard marginal constraint (of type "OT") and $F_2$ is the divergence with respect to $q$ specified in the caption. Note that in (f) some colors "stay on place" due to the fact that the entropy function used is not superlinear.

$64 \times 32 \times 32$ regular cells and we choose the quadratic cost $c_2(x,y) = |x-y|^2$. The anisotropic discretization of the domain accounts for the fact that the eye is more sensitive to variations in luminance than variations in chrominance.

Let $\Omega \subset \mathbb{R}^2$ be the image domain. An image is described by a function $g : \Omega \to X$ and its color histogram is the pushforward of the Lebesgue measure on $\Omega$ by $g$. Let $g_X : \Omega \to X$ and $g_Y : \Omega \to Y (= X)$ be two images and $p$, $q$ be the discretized densities of the associated color histograms. We run Algorithm 1 using the separability of $K$ (see Section 3.3.3) to obtain an (unbalanced) optimal transport plan $\gamma$. An approximate transport map $T : X \to Y$ is then computed according to the map of pointwise conditional expectations, know as the barycentric projection map. This transformation is common in applications when one needs to produce a map from a diffuse plan [152]. The modified image is finally obtained as $T \circ g_X$.

On Figure 5.7, we display a color transfer between very dissimilar images, computed with the parameter $\varepsilon = 0.002$. This intentionally challenging example is insightful as it exhibits the strong effect of the choice of the divergence. Unbalanced optimal transport tend to adapt the amount of each color of target histogram so as to match the modes of the initial histogram and yields meaningful results.

## 5.2. Unbalanced optimal transport barycenters

The computation of entropy regularized Wasserstein barycenters has been considered in [52, 133] (see also [35] for other approaches). With the scaling algorithms, it is possible to compute balanced as well as unbalanced barycenters, within the same framework.

### 5.2.1. Derivation of the algorithm

The coupling formulation of (unbalanced) optimal transport barycenters has been introduced in Sections 3.1.3 and 5.1. In the present section, we consider the numerical resolution of the discretization of problems of the form

$$\min_{\nu} \sum \alpha_k C_{f,c}(\mu^{(k)}, \nu) \tag{5.2.1}$$

where $C_{f,c}$ is an optimal entropy transport cost defined with the entropy function $f$ and the transport cost $c$, $(\mu^{(k)})_k$ is a family of measures and $\nu$ the unknown barycenter. We are going to generalize this slightly and consider possibility different entropies $f_1$ and $f_2$ for each marginal.

Let us chose a family of $n$ discretized densities $(p^k)_{k=1}^n \in (\mathbb{R}^I)^n$ and $(\alpha^{(k)}) \in \mathbb{R}_+^n$ a family of positive weights. In order to obtain closed forms expressions, *for this section only*, we consider an entropic regularization with a different weight on each component

$$H^{(\alpha)}(r|s) := \sum_k \alpha H(r^{(k)}|s^{(k)}).$$

This does not impact the theoretical properties of the regularization, but the definition of the kernel $K$ has to be adapted as, for each component $K^{(k)} := e^{-c/(\alpha^{(k)}\varepsilon)}$, for Proposition 3.2.7 to still hold. The problem of solving (5.2.1) corresponds to defining

$$\bar{F}_1(s) = \sum_{k=1}^n \alpha^{(k)} \sum_{i=1}^I \psi_{f_1}(s_i^{(k)}|p_i^{(k)}), \qquad \bar{F}_2(s) = \inf_{r \in \mathbb{R}_+^I} \sum_{k=1}^n \alpha^{(k)} \sum_{i=1}^I \psi_{f_1}(s_i^{(k)}|r_i).$$

Computing the proximal operator for $F_1$ with respect to the *weighted* entropy $H^{(\alpha)}$ can be done component wise

$$\mathrm{prox}_{F_1}^{H^{(\alpha)}}(s) = (\mathrm{prox}^H_{\psi_{f_1}(\cdot|p^{(1)})}(r^{(1)}), \dots, \mathrm{prox}^H_{\psi_{f_1}(\cdot|p^{(n)})}(r^{(n)}) \tag{5.2.2}$$

(remark that the weights $\alpha$ do not appear anymore as they cancel) and this case has been treated in Section 5.1. Let us turn our attention to $F_2$. Computing the prox and proxdiv operators requires to solve, for each index $i \in \{1, \dots, I\}$, a problem of the form

$$\min_{(\tilde{s},r) \in \mathbb{R}^n \times \mathbb{R}} \sum_{k=1}^n \alpha_k \left( \varepsilon H(\tilde{s}_k|s_k) + \psi_f(\tilde{s}_k|r) \right). \tag{5.2.3}$$

where $f$ is an entropy function. If $\psi_f$ is smooth, first order optimality conditions for (5.2.3) are simple to obtain. The next proposition deals with the general case where more care is needed.

**Proposition 5.2.1.** *Let $(s_i)_{i=1}^n \in \mathbb{R}_+^n$ be such that there exists feasible points $(\tilde{s}^0, r^0)$ for* (5.2.3) *such that $(s_i > 0) \Rightarrow (\tilde{s}_i^0 > 0)$. Then $(\tilde{s}, r)$ solves* (5.2.3) *if and only if*

- $(s_i = 0) \Leftrightarrow (\tilde{s}_i = 0)$ *and*
- *there exists $b \in \mathbb{R}^n$ such that $\sum_{k=1}^n \alpha_k b_k = 0$ and, for all $k \in \{1, \dots, n\}$:*

$$\begin{cases} (\varepsilon \log(s_k/\tilde{s}_k), b_k) \in \partial \psi_f(\tilde{s}_k | r) & \text{if } s_k > 0, \\ b_k \in \partial_2 \psi_f(0, r) & \text{otherwise.} \end{cases}$$

*Proof.* First assume that $s_i > 0$ for all $i \in \{1, \dots, n\}$. The positivity assumption on the feasible point implies that the sum of the subdifferential is the subdifferential of the sum by continuity of the relative entropy on the positive orthant. In this case, a minimizer necessarily satisfies $\tilde{s}_i > 0$ for all $i$. Consequently, the subdifferential of the function in (5.2.3) at $((\tilde{s}_i)_n, r)$ is the set of vectors in $\mathbb{R}^{n+1}$ of the form

$$\begin{pmatrix} \vdots \\ \varepsilon \alpha_i \log(\tilde{s}_i / s_i) + \alpha_i a_i \\ \vdots \\ \lambda \sum_k \alpha_k b_k \end{pmatrix}$$

with $(a_i, b_i) \in \partial \psi_f(\tilde{s}_i | r)$. Writing the first order optimality condition yields the second condition of the proposition. Now for all $i$ such that $s_i$ is null, set $\tilde{s}_i = 0$ (this is required for feasibility) and repeat the reasoning above by withdrawing the variables $\tilde{s}_i$ which have been fixed. □

Once the optimal barycenter density $r \in \mathbb{R}_+^I$ is found with the help of Proposition 5.2.1, determining the optimal values for $\tilde{s}$ can be done componentwise as in (5.2.2) with the help of Proposition 5.1.1. In Table 5.2 we provide formulae for the optimal $r$ for choices of entropy function. For the subsequent computation of $\tilde{s}_k$ (and the proxdiv step) we refer to Table 5.1, where one must replace the marginal density $p$ by the barycenter density $r$.

*Proof of the formulas in Table 5.2.* For $(s_i)_{i=1}^n \in \mathbb{R}^n \geq 0$, we derive the expression for $r$ given in Table 5.2, by applying Proposition 5.2.1.

**Case** OT. This case boils down to solving $\min_r \sum \alpha^{(k)} H(r | s_k)$ and first order optimality conditions allow to conclude.

**Case** $\mathrm{H}_\lambda$. First remark that the assumption of Proposition 5.2.1 is satisfied and that $r = 0$ if and only if for all $k$, $s_k = 0$ (otherwise, the joint subdifferential is empty). As $f$ is smooth, its joint subdifferential is the singleton $\partial H(\tilde{s} | r) = \{(\log(\tilde{s}/r), 1 - \tilde{s}/r)\}$ if $\tilde{s}, r > 0$. Also, since $H(0|r) = r + \iota_{[0,\infty[}(r)$, one has $\partial_2 H(0, r) = \{1\}$ if $r > 0$. Thus, optimality conditions in Proposition 5.2.1 yields the system

$$\begin{cases} \log \frac{\tilde{s}_k}{r} = \frac{\varepsilon}{\lambda} \log \frac{s_k}{\tilde{s}_k} & \text{if } s_k > 0, \\ \tilde{s}_k = 0 & \text{if } s_k = 0, \\ \sum \alpha_k (1 - \frac{\tilde{s}_k}{r}) = 0 \,. \end{cases}$$

| $f$ | Formula for $r$ as a function of $s \in \mathbb{R}^n$ |
|---|---|
| OT | $r = \left(\prod_k s_k^{\alpha_k}\right)^{\frac{1}{\sum_k \alpha_k}}$ |
| $\mathrm{H}_\lambda$ | $r = \left(\frac{\sum_k \alpha_k s_k^{\frac{\varepsilon}{\varepsilon+\lambda}}}{\sum_k \alpha_k}\right)^{\frac{\varepsilon+\lambda}{\varepsilon}}$ |
| $\mathrm{TV}_\lambda$ | if $\sum_{k\notin I_+} \alpha_k \geq \sum_{k\in I_+} \alpha_k$ then $r=0$ otherwise solve: $\sum_{k\notin I_+} \alpha_k + \sum_{k\in I_+} \alpha_k \max\left(-1, \min\left(1, \frac{\varepsilon}{\lambda}\log\frac{r}{s_i}\right)\right) = 0$ |
| $\mathrm{RG}_{[\beta_1,\beta_2]}$ | if $s_k = 0$ for some $k$ then $r=0$ otherwise solve: $\sum_k \alpha_k \left[\beta_2 \min\left(\log\frac{\beta_2 r}{s_k}, 0\right) + \beta_1 \max\left(\log\frac{\beta_1 r}{s_k}, 0\right)\right] = 0$ |

Table 5.2.: Expression for the minimizer $r$ of (5.2.3) as a function of $s \in \mathbb{R}^n_+$ where $I_+ = \{k\,;\, s_k > 0\}$. For the implicit equations of cases TV and RG, an exact solution can be given quickly because computing $\log r$ consists in finding the root of a piecewise linear non-decreasing function (with at most $2n$ pieces), which is guaranteed to change its sign.

**Case $\mathrm{TV}_\lambda$.** As seen in Appendix A, one has $\psi_f(\tilde{s}|r) = \sup_{(a,b)\in B} a\cdot\tilde{s} + b\cdot r$ with $B = \{(a,b) \in \mathbb{R}^2\,;\, a \leq 1, b \leq 1, a+b \leq 0\}$. The set of points in $B$ at which this supremum is attained is the set $\partial\psi_f(\tilde{s}|r)$. With the notations of Proposition 5.2.1, one has with $a_k = \frac{\varepsilon}{\lambda}\log\frac{s_k}{\tilde{s}_k}$,

$$
\begin{array}{ll}
(1)\ \tilde{s}_k > r > 0 \Leftrightarrow -b_k = a_k = 1 & (2)\ r > \tilde{s}_k > 0 \Leftrightarrow -b_k = a_k = -1 \\
(3)\ \tilde{s}_k = r > 0 \Leftrightarrow -b_k = a_k \in [-1,1] & (4)\ r > \tilde{s}_k = 0 \Leftrightarrow b_k = 1 \\
(5)\ \tilde{s}_k > r = 0 \Leftrightarrow a_k = 1 \text{ and } b_k \leq -1 & (6)\ \tilde{s}_k = r = 0 \Leftrightarrow b_k \leq 1\,.
\end{array}
$$

Let us first deal with the case $r = 0$ (cases (5) and (6)). By condition $\sum \alpha_k b_k = 0$ in Proposition 5.2.1, it is possible if and only if $\sum_{k\notin I_+} \alpha_k \geq \sum_{k\in I_+} \alpha_k$. Now assume that $r > 0$. If $\tilde{s}_k > 0$ (cases (1), (2) and (3)) then $b_k$ can be expressed as $\max\{-1, \min\{1, \frac{\varepsilon}{\lambda}\log\frac{r}{s_k}\}\}$. Otherwise, $b_k = 1$. The implicit expression given for $r$ is thus just the condition $\sum \alpha_k b_k = 0$.

**Case $\mathrm{RG}_{[\beta_1,\beta_2]}$.** In this case, $\psi_f$ is the support function of $B = \{(a,b) \in \mathbb{R}^2\,;\, \text{for } i \in \{1,2\}, b \leq -\beta_i \cdot a\}$. With the notations of Proposition 5.2.1, one has with $a_k = \frac{\varepsilon}{\lambda}\log\frac{s_k}{\tilde{s}_k}$,

$$
\begin{array}{ll}
(1)\ 0 < \beta_1 r < \tilde{s}_k < \beta_2 h \Leftrightarrow a_k = b_k = 0 & (2)\ 0 < \beta_1 r = \tilde{s}_k \Leftrightarrow b_k = -\beta_1 a_k \\
(3)\ 0 < \beta_2 r = \tilde{s}_k \Leftrightarrow b_k = -\beta_2 a_k & (4)\ 0 = r = \tilde{s}_k \Leftrightarrow (b_k, a_k) \in B\,.
\end{array}
$$

If $s_k = 0$ for some $k \in \{1,\ldots,n\}$ then $r = 0$ (this is the only feasible point). Otherwise, $r > 0$ and the condition $\sum \alpha_k b_k = 0$ gives the implicit equation. □

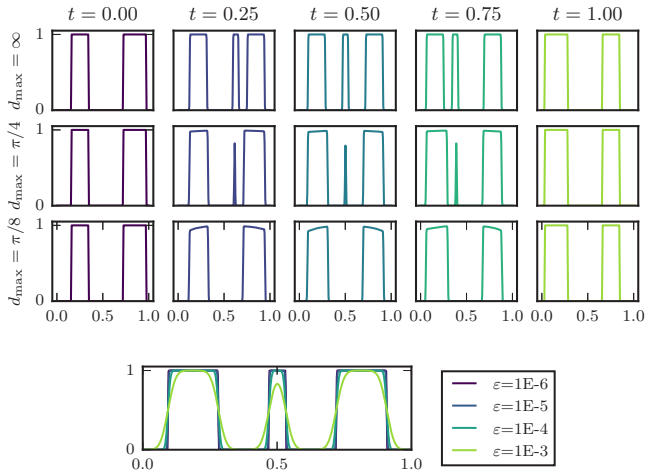


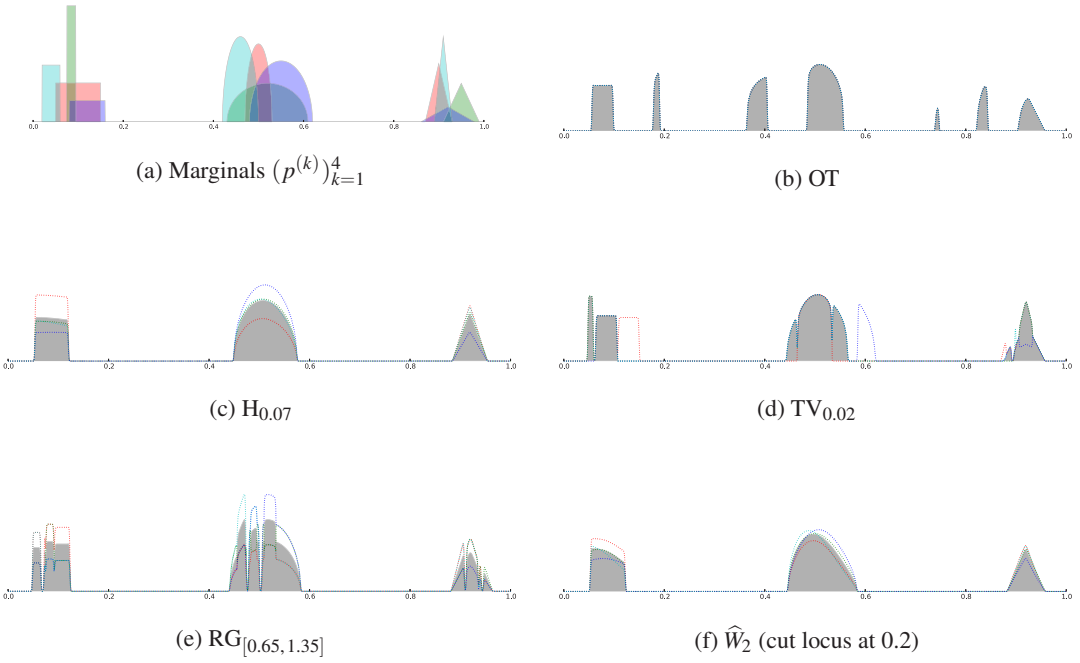

Figure 5.8.: *Top:* Geodesics in $W_2$ distance ($d_{\max} = \infty$) and $\widehat{W}_2$ distance ($d_{\max}$ gives the cut-locus distance) as weighted (unbalanced) barycenters between endpoints. For $d_{\max} = \infty$ the mass difference between left and right blocks must be compensated by transport. As $d_{\max}$ is reduced, the mass difference is increasingly compensated by growth and shrinkage. In all experiments $\varepsilon = 10^{-6}$. *Bottom:* Midpoint of $W_2$ geodesic for various regularization $\varepsilon$.

### 5.2.2. Numerical experiments

Geodesics for the Wasserstein $\mathrm{W}_2$ metric and unbalanced optimal transport $\widehat{W}_2$ (Section 2.2) metrics can be computed as weighted barycenters between their endpoints. In Figure 5.8 geodesics for the $\widehat{W}_2$ distance on $X = [0,1]$ for different cut-loci $d_{\max}$ and the $W_2$ distance (corresponding to $d_{\max} = \infty$) are compared and the influence of entropy regularization is illustrated. The interval $[0,1]$ was discretized into 512 cells.

In Figure 5.9, we display Fréchet means-like experiments for a family of 4 given marginals where $X$ is the segment $[0,1]$ discretized as above. The discretized densities of the marginals $(p^{(k)})_{k=1}^4$ consist each of the sum of three bumps (centered near the points $x = 0.1$, $x = 0.5$ and $x = 0.9$). These computations where performed with Algorithm 2 with $\varepsilon = 10^{-5}$. Unbalanced variants (Figures 5.9c-5.9f) tend to preserve more the structure in three bumps in contrast to classical optimal transport (Figure 5.9b).

Figure 5.10 displays barycenters with the quadratic cost $c_{\mathrm{quad}}$ and the marginal penalization $\mathrm{OT}/\mathrm{OT}$ (classical optimal transport) and $\mathrm{H}/\mathrm{H}$ (this corresponds to the "Gaussian-Hellinger-Kantorovich" metric defined in Section 2.2) respectively. The densities considered on $[0,3]^2$ are discretized into $200 \times 200$ samples. Computations where performed using Algorithm 2 with



(a) Marginals $(p^{(k)})_{k=1}^4$ (b) OT

(c) $\mathrm{H}_{0.07}$ (d) $\mathrm{TV}_{0.02}$

(e) $\mathrm{RG}_{[0.65,1.35]}$ (f) $\widehat{W}_2$ (cut locus at 0.2)

Figure 5.9.: Barycenter-like problems on $X = [0,1]$. Except for (f), the first entropy $f_1$ is $\iota_{\{1\}}$ which corresponds to equality constraints w.r.t. the densities $(p^{(k)})_{k=1}^4$, the cost is $c_{\mathrm{quad}}(x,y) = |y-x|^2$, the weights are $(1,1,1,1)/4$ and the function $f_2$ is of the type specified in the legend (using the same notations as in the previous section). Figure (f) represents the Fréchet mean for the $\widehat{W}_2$ metric. The dotted lines represent the marginal of the optimal plans w.r.t. the second factor.

$\varepsilon = 10^{-6}$. The barycenter coefficients are the following:

$$
\begin{array}{ccccccccc}
 & & & & (0,1,0) & & & & \\
 & & & (1,3,0)/4 & & (0,3,1)/4 & & & \\
 & & (2,2,0)/4 & & (1,2,1)/4 & & (0,2,2)/4 & & \\
 & (3,1,0)/4 & & (2,1,1)/4 & & (1,1,2)/4 & & (0,1,3)/4 & \\
(1,0,0) & & (3,0,1)/4 & & (2,0,2)/4 & & (1,0,3)/4 & & (0,0,1)
\end{array}
$$

The input densities have been specifically designed to showcase the difference between the balanced and unbalanced models. The input densities have a similar global structure but mass is unevenly distributed globally. This leads optimal transport to scatter mass over the domain as it has to compromise between many locations for the superfluous mass, while the unbalanced model simply adjusts the mass locally.

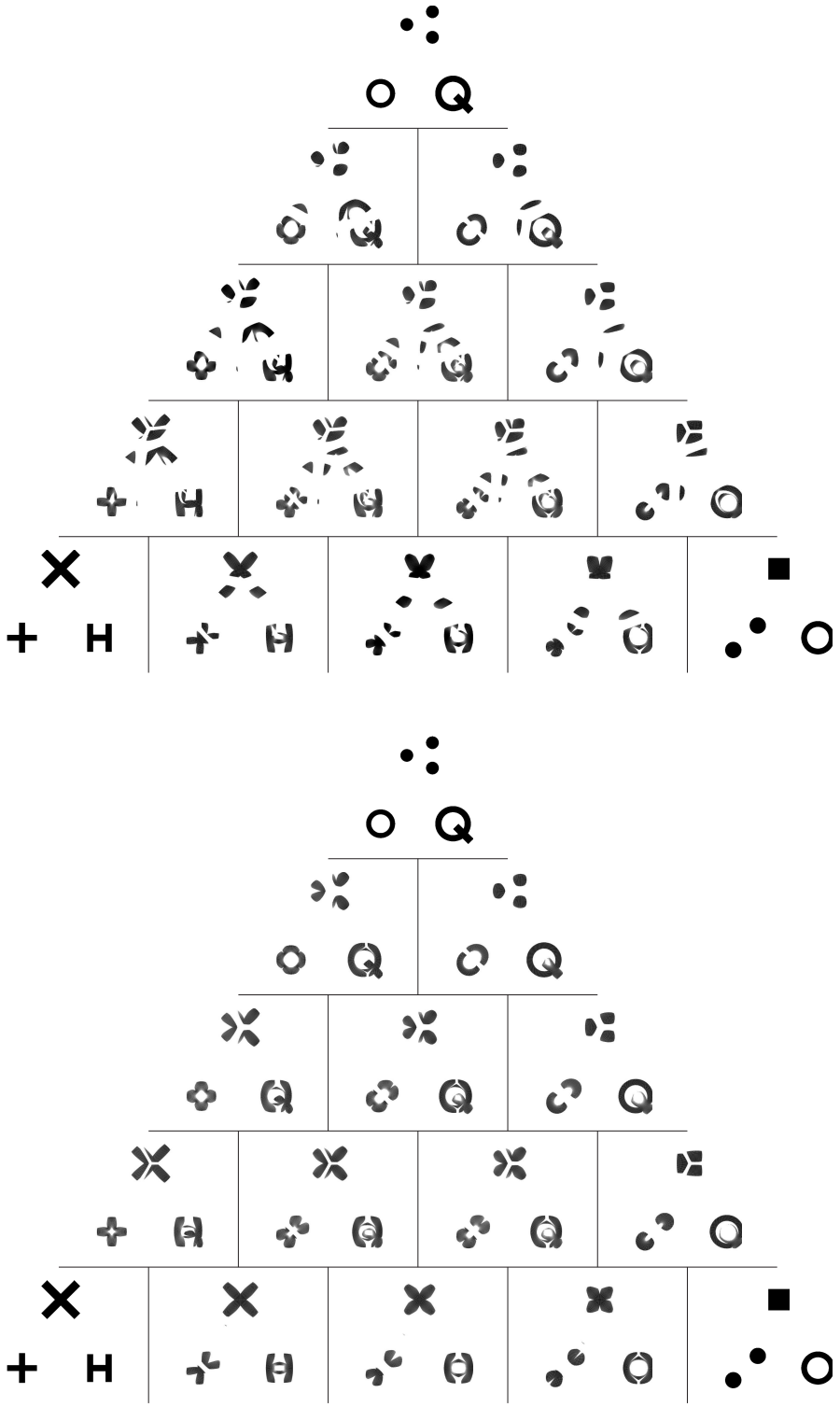


Figure 5.10.: *Top*: barycenters with quadratic cost and exact marginal constraints (Wasserstein barycenters). *Bottom:* barycenters with quadratic cost and entropies $f(s) = s\log s - s + 1$ (Gaussian-Hellinger-Kantorovich barycenters). The three input densities are at the corners and the colormap is linear from 0 (white) to 1 (black).

## 5.3. Time discretized gradient flows

The basic framework of gradient flows has been briefly laid out in Section 3.1. A large variety of dedicated numerical schemes has been proposed for spatial discretization and solving of these time-discrete flows, such as for instance finite differences [27], finite volumes [37] and Lagrangian schemes [13]. A bottleneck of these schemes is the high computational complexity due to the resolution of an OT-like problem at each time step. The use of entropic regularization has been proposed recently in [127] and studied theoretically in [34] for Wasserstein gradient flow.

In this section, we consider the numerical resolution of time discretized gradient flows for $W_2$ or $\widehat{W}_2$. The framework of scaling algorithm is versatile, simple, extends to the case of unbalanced gradient flow and can be stabilized and accelerated to reach higher precisions, faster. We derive the "proxdiv" operator for a few cases and showcase the method on the computation of the Wasserstein gradient flow of the functional "total variation of the gradient". In the next chapter, we separately and in more details the numerical behavior for a $\widehat{W}_2$ gradient flow.

### Computing "proxdiv" maps for separable gradient flows

**$W_2$ gradient flows** The JKO scheme for time discretized Wasserstein $W_2$ gradient flows of a functional $G$ involves a function of the form

$$\bar{F}(s) = 2\tau\bar{G}(s)$$

where $\bar{G}$ is the functional expressed in terms of density.

**Example 5.3.1** ("proxdiv" for $W_2$ discretized gradient flow of $G$)**.**

- *If $G$ is the relative entropy w.r.t. a reference measure, i.e. $\bar{G}(s) = H(s|p)$ for $p \in \mathbb{R}^I_+$ then one has for each index $i$*

$$[\mathrm{proxdiv}_F(s,u,\varepsilon)]_i = (p_i/s_i)^{2\tau/(2\tau+\varepsilon)} \cdot e^{-u_i/(\varepsilon+2\tau)}$$

- *some "splitting schemes" for crowd motion [110, 112] or dentritic growth [111] involve the Wasserstein projection on a set of bounded densities. This can be solved with scaling algorithms by posing $G$ the convex indicator of a maximum density constraint, i.e. $G(s) = 0$ if $s_i \leq p_i$ for all $i \in \{1,\dots,I\}$, $\infty$ otherwise, then*

$$[\mathrm{proxdiv}_F(s,u,\varepsilon)]_i = \min\{e^{-u_i/\varepsilon}, p_i/s_i\}$$

**$\widehat{W}_2$ gradient flows** In the next chapter, we consider in details a gradient flow for the unbalanced metric $\widehat{W}_2$ introduced in Chapter 2.2. In this case, one has a marginal function of the form

$$\bar{F}(s) = \inf_{p \in \mathbb{R}^I_+} H(s|p) + 2\tau\bar{G}(p)$$

If $G$ is a sum of pointwise functionals $G(p) = \sum_{i=1}^{I} g_i(p_i)$ then the operator $\mathrm{prox}_{F/\varepsilon}^{H}$ is given by the pointwise minimization problem

$$\inf_{(\tilde{s},p)\in\mathbb{R}^2} \varepsilon H(\tilde{s}|s) + H(\tilde{s}|p) + 2\tau g_i(p)$$

for which the first order optimality conditions read

$$\begin{cases} 0 = \varepsilon \log(\tilde{s}/s) + \log(\tilde{s}/p) \\ (\tilde{s}/p - 1)/(2\tau) \in \partial g_i(p) \end{cases}$$

and $\tilde{s} = 0$ if $s = 0$ or $p = 0$. We leave the study of a specific example for the next chapter.

### Wasserstein gradient flow of "total variation of the gradient"

We conclude with the challenging case of the Wasserstein gradient flow of the functional $\rho \mapsto \|\nabla\rho\|_1$, the total variation of the gradient. This flow has been studied in [36] where is it shown to converge (in some cases) to a fourth order and non-linear evolution PDE that reads

$$\partial_t \rho_t + \nabla\cdot(\rho_t \nabla p_t) = 0 \quad \text{with} \quad p_t = \nabla\cdot(\nabla\rho_t/|\nabla\rho_t|)$$

Dedicated numerical schemes have been studied in [63, 29, 16] to solve this PDE and it can be seen there that it is challenging to obtain efficient and precise numerical schemes already in the 1-D case.

For a discretized 1-D domain of grid step $h$, the marginal function is

$$\bar{F}(s) = \frac{2\tau}{h}\sum_{i=1}^{I-1} |s_{i+1} - s_i|.$$

It is not separable, so we suggest to split it in two parts

$$\bar{F}_e = \frac{2\tau}{h}\sum_{i\ \text{even}} g(s_{i+1}, s_i) \qquad \bar{F}_o = \frac{2\tau}{h}\sum_{i\ \text{odd}} g(s_{i+1}, s_i) \tag{5.3.1}$$

where $g(a,b) = |a-b|$. The operator "proxdiv" of $F_e$ and $F_o$ can be deduced from that for $g$ which has the following expression when evaluated at $(a,b),(u,v)$:

$$\begin{cases} (0,0) & \text{if } a = b = 0 \\ (e^{-(u+\tau/h)/\varepsilon}, e^{-(v-\tau/h)/\varepsilon}) & \text{if } b = 0 \text{ or } a/b > e^{(u-v+2\tau/h)/\varepsilon} \\ (e^{-(u-\tau/h)/\varepsilon}, e^{-(v+\tau/h)/\varepsilon}) & \text{if } b = 0 \text{ or } a/b < e^{(u-v-2\tau/h)/\varepsilon} \\ (\sqrt{b/a})e^{-(u+v)/(2\varepsilon)}, (\sqrt{a/b})e^{-(u+v)/(2\varepsilon)}) & \text{otherwise.} \end{cases}$$

Since we do not have an explicit expression for $\mathrm{proxdiv}_{\bar{F}}$, we need to depart from the exact setting of scaling algorithms. We propose to use Algorithm 3 with three "proxdiv" operators (in practice it is more efficient to apply the "proxdiv" associated to the fixed marginal constraint less often). We display on Figure 5.11 the flow computed using this method, starting from a piecewise constant density. It is known that in that case, the evolution remains piecewise constant, as observed in [36].

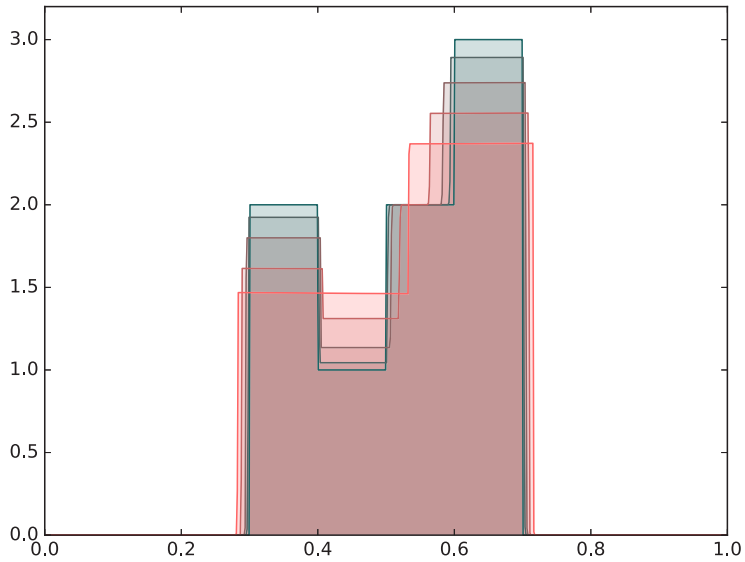



Figure 5.11.: Time discretized Wasserstein gradient flow of the functional $G(s) = \|\nabla s\|_1$ starting from piecewise constant density. The time steps shown are $t = 10^{-3} \times (0; 2.5; 7.3; 17; 37)$ (from green to red) and the time resolution of the computed flow is twice higher. $\varepsilon = 2 \times 10^{-6}$.

# Chapter 6.

# Gradient Flow of Hele-Shaw Type

In this chapter, we show that the solutions of an evolution PDE of Hele-Shaw type, studied in mathematical biology as a tumor growth model, can be recovered as the gradient flow of a degenerate functional for the unbalanced metric $\widehat{W}_2^2$.

We prove that the key ingredients of the theory of gradient flows in metric spaces apply. On the one hand, any *minimizing movement* is a solution of the PDE: this implies the existence of solutions without regularity assumptions on the initial density. On the other, when the domain is convex, every solution of the PDE is a *EVI* solution of gradient flow, which implies uniqueness.

In Section 6.1, we introduce the PDE of interest, motivate its study and state the main result as well as an informal justification. The whole of Section 6.2 is devoted to the proofs of existence and uniqueness.

In Section 6.3, starting from the time discretized minimizing movement scheme, we proceed with the spatial discretization and derive the scaling algorithm for solving the discrete step. We show that the numerical scheme is consistent throughout, in that is recovers a solution to the PDE when all parameters tend to their limit successively.

In Section 6.4, we compare the numerical results with explicit spherical solutions in order to assess the precision of the numerical method as well as the convergence of the scheme. We conclude in Section 6.5 with numerical illustrations in 1-D and 2-D.

This chapter is a joint work with Simone Di Marino.

## 6.1. Background and main results

### 6.1.1. A Hele-Shaw growth model as a gradient flow

Modeling tumor growth is a longstanding activity in applied mathematics that has become a valuable tool for understanding cancer development. At the macroscopic and continuous level, there are two main categories of models: the cell density models – which describe the tumor as a density of cells which evolve in time – and the free boundary models – which describe the evolution of the domain conquered by the tumor by specifying the geometric motion of its boundary. Perthame et al. [126, 113] have exhibited a connection between these two approaches: by taking the incompressible limit of a standard density model of growth/diffusion, one recovers a free boundary model of Hele-Shaw type.

More precisely, they consider a mono-phasic density of cells $\rho(x,t)$ (with $x \in \mathbb{R}^d$ the space variable and $t \geq 0$ the time) whose motion is driven by a scalar pressure field $p(x,t)$ through Darcy's law and which grows according to the rate of growth which is modeled as a function of the pressure $\Phi(p(x,t))$ where $\Phi$ is continuously decreasing and null for $p$ greater than a so-called "homeostatic" pressure. The equation of evolution for $\rho$ is then

$$\begin{cases} \partial_t \rho - \nabla \cdot (\rho \nabla p) = \Phi(p)\rho\,, \text{ for } t > 0 \\ \qquad\qquad p = \rho^m, \quad \text{with } m \geq 1\,, \\ \qquad \rho(0,\cdot) = \rho_0 \in L^1_+(\Omega) \end{cases} \tag{6.1.1}$$

where the relation $p = \rho^m$ accounts for a slow-diffusive motion. For suitable initial conditions, they show that when $m$ tends to infinity – the so-called *stiff* or *incompressible* or *hard congestion* limit – the sequence of solutions $(\rho^m, p^m)$ of (6.1.1) tends to a limit $(\rho^\infty, p^\infty)$ satisfying a system of the form (6.1.1) where the relation between $\rho$ and $p$ is replaced by the "Hele-Shaw graph" constraint $p(1-\rho) = 0$.

We propose to study directly this stiff limit system from a novel mathematical viewpoint, focusing on the case of a rate of growth depending linearly on the pressure $\Phi(p) = 4(\lambda - p)_+$, with a homeostatic pressure $\lambda > 0$. In a nutshell, we show that the stiff limit system

$$\begin{cases} \partial_t \rho - \nabla \cdot (\rho \nabla p) = 4(\lambda - p)_+ \rho \\ p(1-\rho) = 0 \\ 0 \leq \rho \leq 1 \\ \rho(0,\cdot) = \rho_0 \end{cases} \tag{6.1.2}$$

characterizes gradient flows of the functional $G : \mathscr{M}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$ defined as, with $\mathscr{L}^d$ the Lebesgue measure on $\mathbb{R}^d$,

$$G(\rho) = \begin{cases} -\lambda \rho(\Omega) & \text{if } \rho \ll \mathscr{L}^d \text{ and } \frac{d\rho}{d\mathscr{L}^d} \leq 1, \\ +\infty & \text{otherwise} \end{cases} \tag{6.1.3}$$

in the space of nonnegative measures $\mathscr{M}_+(\Omega)$ endowed with the metric $\widehat{W}_2$ (see Chapter 2.2). This approach has the following advantages:

- on a qualitative level, it gives a simple interpretation of the Hele-Shaw tumor growth model. Namely, (6.1.2) describes *the most efficient way for a tumor to gain mass under a maximum density constraint*, where efficiency translates as small displacement and small rate of growth.

- on a theoretical level, we show existence of solutions to (6.1.2) with minimal regularity assumptions on the initial condition (weaker than in [113]) and uniqueness on compact convex domains. Besides, we believe the main interest is to give another application of the theory of gradient flows in non-Hilbertian metric spaces, beyond Wasserstein spaces.

- on a numerical level, relying on recent advances on algorithms for unbalanced optimal transport problems [45], the gradient flow approach allows for a simple numerical scheme for computing solutions to (6.1.2).

### 6.1.2. Main result

In order to make precise statements, let us define what is meant by *gradient flow*. In $\mathbb{R}^d$, the gradient flow of a function $G:\mathbb{R}^d\to\{+\infty\}$ is a continuous curve $x:\mathbb{R}_+\to\mathbb{R}^d$ which is solution to the Cauchy problem

$$\begin{cases} \frac{d}{dt}x(t) = -\nabla G(x(t)), \text{ for } t>0 \\ x(0) = x_0 \in \mathbb{R}^d. \end{cases} \tag{6.1.4}$$

However, in a (non-Riemannian) metric space $(X,d)$, the gradient $\nabla G$ of a functional $G:X\to\mathbb{R}\cup\{+\infty\}$ is not clearly defined anymore. Yet, several extensions of the notion of gradient flow exist, relying on the variational structure of (6.1.4), see [5] for a general theory. One approach is that of *minimizing movements* introduced by De Giorgi, that originates from the discretization in time of (6.1.4) through the implicit Euler scheme : starting from $\tilde{x}_0^\tau = x_0 \in X$ define a sequence of points $(\tilde{x}_k^\tau)_{k\in\mathbb{N}}$ as follows

$$\tilde{x}_{k+1} \in \operatorname*{argmin}_{x\in X} \left\{ G(x) + \tfrac{1}{2\tau} d(x,\tilde{x}_k^\tau)^2 \right\}. \tag{6.1.5}$$

By suitably interpolating this discrete sequence, and making the time step $\tau$ tend to 0, we recover a curve which, in a Euclidean setting, is a solution to (6.1.4). This leads to the following definition.

**Definition 6.1.1** (Uniform minimizing movements)**.** *Let $(X,d)$ be a metric space, $G:X\to\mathbb{R}\cup\{+\infty\}$ be a functional and $x_0\in X$. A curve $x:\mathbb{R}_+\to X$ is a uniform minimizing movement if it is the pointwise limit as of a sequence of curves $x^{\tau_i}$ defined as $x^{\tau_i}(t)=\tilde{x}_k^{\tau_i}$ for $t\in[k\tau_i,(k+1)\tau_i[$ for some sequence generated by* (6.1.5)*, with $\tau_i\to 0$.*

When the metric space is the space of probability measures endowed with an optimal transport metric $(\mathscr{P}(\Omega),W_2)$, this time discretization is known as the *JKO scheme*. It is named after the authors of the seminal paper [90] where it is used to recover (in particular) the heat equation by taking the uniform minimizing movement of the entropy functional. A more precise and more restrictive notion of gradient flow is given by the *evolutional variational inequality* (EVI).

**Definition 6.1.2** (EVI gradient flow)**.** *An absolutely continuous curve $(x(t))_{t\in[0,T]}$ in a metric space $(X,d)$ is said to be an* $\mathrm{EVI}_\alpha$ *(for $\alpha\in\mathbb{R}$) solution of gradient flow of $G:X\to\mathbb{R}\cup\{\infty\}$ if for all $y\in\mathrm{dom}\,G$ and a.e. $t\in]0,T[$ it holds*

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

The main result of this chapter, proved in Section 6.2, makes a link between the tumor growth model (6.1.2), the metric $\widehat{W}_2$ and the functional (6.1.3). Note that in this chapter, we assume a definition of the distance $\widehat{W}_2$ that is based on the Euclidean metric on $\mathbb{R}^d$ and not the geodesic distance of $\Omega$ (they differ when $\Omega$ is not convex).

**Theorem 6.1.3.** *Let $\Omega$ be an open bounded $H^1$-extension domain of $\mathbb{R}^d$, let $\rho_0\in L^1_+(\Omega)$ such that $\rho_0\leq 1$ and $T>0$. Then any minimizing movement is a solution of* (6.1.2) *on $[0,T]$ starting from $\rho_0$ with some $p\in L^2([0,T],H^1(\Omega))$. Moreover if $\Omega$ is convex we have that every solution of* (6.1.2) *is an* $\mathrm{EVI}_{(-2\lambda)}$ *solution of gradient flow of $G$ in the metric space $(\mathscr{M}_+(\Omega),\widehat{W}_2)$; in particular in this case we have uniqueness for* (6.1.2)*.*

*Proof.* The existence result is in Proposition 6.2.14 and the EVI characterization, with uniqueness, in Proposition 6.2.17. □

**Remark 6.1.4.** *The concept of solutions to the system* (6.1.2) *is understood in the weak sense as in Proposition 1.1.3, i.e. we say that the family of triplets $(\rho_t,v_t,g_t)_{t\geq0}$ is a solution to*

$$\partial_t\rho_t-\nabla\cdot(\rho_t v_t)=g_t\rho_t$$

*if for all $\phi\in C^\infty_c(\bar\Omega)$, the function $t\mapsto\int_\Omega\phi(x)d\rho_t(x)$ is well defined, absolutely continuous on $[0,+\infty[$ and for a.e. $t\geq0$ we have*

$$\frac{d}{dt}\int_\Omega\phi d\rho_t=\int_\Omega(\nabla\phi\cdot v_t+\phi g_t)d\rho_t\,.$$

*This property implies that $t\mapsto\rho_t$ is weakly continuous, that the PDE is satisfied in the distributional sense, and imposes no-flux boundary conditions for $v_t$. Equation* (6.1.2) *is a specialization of this equation with $v_t=-\nabla p_t$ and $g_t=4(\lambda-p_t)_+$.*

### 6.1.3. Short informal derivation

Before proving the result rigorously, let us present an informal discussion, inspired by [143], in order to grasp the intuition behind the result. Stuying the optimality conditions in the dynamic formulation of $\widehat{W}_2$ given in Theorem 1.1.18, one sees that the velocity and the growth fields are derived from a dual potential (proofs of this fact can be found in [94, 103]) as $(v_t,g_t)=(\nabla\phi_t,4\phi_t)$ and one has

$$\widehat{W}_2^2(\mu,\nu)=\inf_\phi\left\{\int_0^1\int_\Omega(|\nabla\phi_t|^2+4|\phi_t|^2)\mathrm{d}\rho_t\mathrm{d}t\;;\;\partial_t\rho_t=-\nabla\cdot(\nabla\phi_t\rho_t)\right\}$$

where $(\rho_t)_{t\in[0,1]}$ is a path that interpolates between $\mu$ and $\nu$. This suggest to interpret $\widehat{W}_2$ as a Riemannian metric with tangent vectors at a point $\rho \in \mathscr{M}_+(\Omega)$ of the form $\partial_t \rho = -\nabla\cdot(\nabla\phi\rho) + 4\phi\rho$ and the metric tensor

$$\langle \partial_t\rho_1, \partial_t\rho_2\rangle_\rho = \int_\Omega (\nabla\phi_1\cdot\nabla\phi_2 + 4\phi_1\cdot\phi_2)\mathrm{d}\rho.$$

Now consider a smooth functional by $F : \mathscr{M}_+(\Omega) \to \mathbb{R}$ and denote $F'$ the unique function such that $\frac{d}{d\varepsilon}F(\rho+\varepsilon\chi)|_{\varepsilon=0} = \int_\Omega F'(\rho)d\chi$ for all admissible perturbations $\chi \in \mathscr{M}(\Omega)$. Its gradient at a point $\rho$ satisfies, for a tangent vector $\partial_t\rho = -\nabla\cdot(\rho\nabla\phi) + 4\phi\rho$, by integration by part,

$$\langle \mathrm{grad}_\rho F, \partial_t\rho\rangle_\rho = \int_\Omega F'(\rho)\partial_t\rho = \int_\Omega (\nabla F'(\rho)\cdot\nabla\phi + 4F'(\rho)\cdot\phi)\rho$$

which shows that, by identification, one has

$$\mathrm{grad}_\rho F = -\nabla\cdot(\rho\nabla F'(\rho)) + 4\rho F'(\rho).$$

In particular, taking the functional $F_m(\rho) = -\lambda\rho(\Omega) + \frac{1}{m+1}\rho^{m+1}$ ($\rho$ is identified with its Lebesgue density) the associated gradient flow is the diffusion-reaction system (6.1.1) because $F_m'(\rho) = -\lambda + \rho^m$. The functional $G$ introduced in (6.1.3) can be understood as the "stiff" limit as $m \to \infty$ of the sequence of functionals $F_m$. Theorem 6.1.3 expresses, with different tools, that the gradient flow structure is preserved in the limit $m \to \infty$ where one recovers the hard congestion model (6.1.2).

### 6.1.4. Related works

In the context of Wasserstein gradient flows — or so called JKO flows, free boundary models have already been modeled in [121, 79] where a thin plate model of Hele-Shaw type is recovered by minimizing the interace energy. More recently, crowd motions have been modeled with these tools in [110, 112] in a series of work pioneering the study of Wasserstein gradient flows with a hard congestion constraint.

The success of the gradient flow approach in the field of PDEs has naturally led to generalizations of optimal transport metrics in order to deal with a wider class of evolution PDEs, such as the heat flow with Dirichlet boundary conditions [70], and diffusion-reaction systems [101]. The specific metric $\widehat{W}_2$, has recently been used to study population dynamics [93], and gradient flows structure for more generic smooth functionals have been explored in [77] where the authors consider splitting strategy, i.e. they deal with the transport and the growth term in a alternative, desynchronized manner.

Our work was pursued simultaneously and independently of [76] where this very class of tumor growth model are studied using tools from optimal transport. These two works use different approaches and are complementary: our focus is on the stiff models (6.1.1) and we directly study the incompressible system with specific tools while [76] focuses primary on the diffusive models (6.1.1), and recover the stiff system by taking a double limit. Their approach is thus not directly based on a gradient flows, but is more flexible and allows to deal with nutrient systems.

## 6.2. Proof of the main theorem

### 6.2.1. Entropy-transport problems

In this section we consider the optimal entropy-transport problems (as in Section 1.2.3) associated to cost functions $c : \Omega^2 \to \mathbb{R} \cup \{\infty\}$ and the entropy function $f(s) = s\log s - s + 1$, defined as

$$C_{c,f}(\mu_1,\mu_2) := \inf_{\gamma \in \mathscr{M}_+(\Omega\times\Omega)} \left\{ \int c(x,y)\,\mathrm{d}\gamma(x,y) + H(\pi^1_\#\gamma|\mu_1) + H(\pi^2_\#\gamma|\mu_2) \right\} \tag{6.2.1}$$

where $H$ is the relative entropy functional (defined e.g. in Definition 3.5.1). *In this whole chapter, the entropy function $f$ is the same*, but the cost varies. The main role is played by the cost

$$c_\ell(x,y) = -\log\cos^2(\min\{|y-x|, \pi/2\})$$

for which one recovers the definition of $\widehat{W}_2$ in Theorem 2.2.4 (with the important distinction that we consider the Euclidean distance, so imagine we have enclosed $\Omega$ in a bigger, convex set on which we define $\widehat{W}_2$). A family of Lipschitz costs $c_n$ approximating $c_\ell$ is also used. These costs are constructed from the following approximation argument for the function $g_\ell : [0,\infty) \to [0,\infty]$ defined by

$$g_\ell(t) := -\log\cos^2(\min\{t, \pi/2\}). \tag{6.2.2}$$

**Lemma 6.2.1.** *The function $g_\ell$ is a convex and satisfies $g_\ell'^2 = 4(e^{g_\ell} - 1)$ in $[0,\pi/2)$. It can be approximated by an increasing sequence of strictly convex Lipschitz functions $g_n : [0,\infty) \to [0,\infty)$ such that*

*(i) $0 \le g_n \le g_m \le g_\ell$ for every $n \le m$ and $g_n(t) \uparrow g(t)$ pointwise for every $t$; moreover $g_n(t) = g(t)$ for $t \in [0,1]$.*

*(ii) For all $n$ we have that $g_n'^2 \le 4(e^{g_n} - 1)$ in $[0,\infty)$ and $e^{g_n(t)} - 1 \ge t^2$.*

*Proof.* It is easy to compute the first and second derivative of $g_\ell$ to deduce that we have also $g_\ell''(t) = 2e^{g_\ell} \ge 2$. Now we consider and increasing sequence of points $t_n \le \pi/2$ such that $t_n > 1$ and $t_n \uparrow \pi/2$; then we define functions $g_n$ such that $g_n(0) = 0$, $g_n'(0) = 0$ and

$$g_n''(t) = \begin{cases} 2e^{g_\ell(t)} & \text{if } t \le t_n \\ e^{-t} & \text{otherwise.} \end{cases} \tag{6.2.3}$$

Since $g_n'' > 0$ uniformly on bounded sets we have that $g_n$ is strictly convex; moreover it is Lipschitz since

$$\begin{aligned} g_n'(t) = g_n'(t) - g_n'(0) &= \int_0^t g_n''(s)\,ds \\ &\le \int_0^\infty g_n''(s)\,ds = \int_0^{t_n} g_\ell''(s)\,ds + \int_{t_n}^\infty e^{-s}\,ds \\ &\le g_\ell'(t_n) + 1. \end{aligned}$$

Furthermore clearly since $t_n$ is increasing we have that $g_n''$ is an increasing sequence of functions, in fact $g_\ell''(t) > e^{-t}$. Moreover it is clear that $g_n(t) = g_\ell(t)$ for $t \in [0,t_n]$ (and in particular in $[0,1]$), and so we have $g_n \uparrow g_\ell$ in $[0,\pi/2)$ but, since $g_n$ are increasing functions in $t$, we conclude also that $g_n(t) \to +\infty$ for every $t \geq \pi/2$.

As for (ii) we denote $F(t) = g_n'(t)^2$ and $G(t) = 4(e^{g_n(t)} - 1)$. First we notice that $F(t) = G(t)$ for $t \in [0,t_n]$, since here $g_n$ agrees with $g_\ell$, that satisfies the differential equation; then, for $t > t_n$, we can apply the Cauchy's mean value theorem to $F$ and $G$, that are both strictly increasing and differentiable in $(t_n,\infty)$. In particular there exists $t_n < s < t$ such that

$$\frac{F(t)-F(t_n)}{G(t)-G(t_n)} = \frac{F'(s)}{G'(s)} = \frac{2g_n'(s)g_n''(s)}{4g_n'(s)e^{g_n(s)}} = \frac{g_n''(s)}{2e^{g_n(s)}} \leq e^{-s} \leq 1;$$

knowing that $F(t_n) = G(t_n)$ and that $G(t) > G(t_n)$ we get immediately that $F(t) \leq G(t)$.

For the second inequality we will use that $e^t - 1 \geq t$ and $e^t - 1 \geq t^2/2$. We choose $t_n$ big enough such that $g_\ell(t_n)/t_n \geq \sqrt{2}$: this is always possible since $g_\ell(t)/t \to \infty$ as $t \uparrow \pi/2$. Then from (6.2.3) we have $g_n''(t) \geq 2$ for $t \leq t_n$ and in particular $g_n(t) \geq t^2$ in that region and so we get

$$e^{g_n(t)} - 1 \geq e^{t^2} - 1 \geq t^2 \qquad \forall t \leq t_n,$$

while if $t \geq t_n$ by convexity we have $g_n(t) \geq \frac{g_n(t_n)}{t_n} t \geq \sqrt{2}t$ and so

$$e^{g_n(t)} - 1 \geq e^{\sqrt{2}t} - 1 \geq \frac{1}{2}(\sqrt{2}t)^2 = t^2 \qquad \forall t \geq t_n. \qquad \square$$

Let us clarify and introduce the notations used in this chapter.

- one has $c_\ell(x,y) = g_\ell(|y-x|)$ and the associated optimal entropy-transport cost is $C_\ell = C_{c_\ell,f} = \widehat{W}_2^2$;
- for the approximating costs we denote $c_n(x,y) = g_n(|y-x|)$ and $C_n = C_{c_n,f}$;

We begin proving two lemmas.

**Lemma 6.2.2.** *Let us consider $\mu_1,\mu_2$ two measures in $\Omega$ and a Borel cost $c \geq 0$. It holds*

$$C_{c,f}(\mu_1,\mu_2) \geq \left(\sqrt{\mu_1(\Omega)} - \sqrt{\mu_2(\Omega)}\right)^2.$$

*Proof.* In the sequel, we denote $\mu_i(\Omega) = m_i$. It is clear that we can suppose $c = 0$ (in fact $C_{c',f} \leq C_{c,f}$ whenever $c' \leq c$) and write our problem as

$$\begin{aligned} C_{0,f}(\mu_1,\mu_2) &= \min_{\gamma \in \mathscr{M}_+(\Omega\times\Omega)} \left\{H(\pi^1_\#\gamma|\mu_1) + H(\pi^2_\#\gamma|\mu_2)\right\} \\ &= \min_{M\geq 0} \min_{\gamma \in \mathscr{M}_+(\Omega\times\Omega)} \left\{H(\pi^1_\#\gamma|\mu_1) + H(\pi^2_\#\gamma|\mu_2) \ : \ m(\gamma) = M\right\}. \end{aligned}$$

We can restrict ourselves to the case $\pi^i_\#\gamma \ll \mu_i$, where we have $\pi^i_\#\gamma = \sigma_i\mu_i$ and using Jensen inequality applied to $f(t) := t\log t - t + 1$ it holds

$$\begin{aligned} H(\pi^i_\#\gamma_i|\mu_i) &= m_i \int_{\mathbb{R}^d} f(\sigma_i)\, d\frac{\mu_i}{m_i} \geq m_i f\left(\int_{\mathbb{R}^d} \sigma_i d\frac{\mu_i}{m_i}\right) \\ &= m_i f\left(\frac{M}{m_i}\right) = M\log\left(M/m_i\right) - M + m_i, \end{aligned}$$

with equality if we choose $\gamma = \frac{M}{m_1}\mu_1 \otimes \frac{M}{m_2}\mu_2$. In particular, we have

$$C_{0,f}(\mu_1,\mu_2) = \min_{M\geq 0}\left\{ M\log\left(\frac{M^2}{m_1 m_2}\right) + m_1 + m_2 - 2M \right\},$$

the minimizer is $M = \sqrt{m_1 m_2}$, so $C_{c,f}(\mu_1,\mu_2) \geq C_{0,f}(\mu_1,\mu_2) = (\sqrt{m_1}-\sqrt{m_2})^2$. □

The following characterization is proved in [103, Thm. 4.14] and Corollaries.

**Theorem 6.2.3.** *Let us consider an L-Lipschitz cost $c \geq 0$. Then we have*

$$C_{c,f}(\mu_1,\mu_2) = \max_{\alpha,\beta\in \mathrm{Lip}_L(\Omega)}\left\{ \int (1-e^{-\alpha})\,\mathrm{d}\mu_1 + \int (1-e^{-\beta})\,\mathrm{d}\mu_2 \ :\ \alpha+\beta\leq c \right\}.$$

*Here and in the following, the constraint has to be understood as $\alpha(x)+\beta(y)\leq c(x,y)$, for all $(x,y)\in\Omega^2$. Moreover, if $\gamma$ denotes a minimizer in the primal problem and $\alpha,\beta$ maximizers in the dual, we have the following compatibility conditions:*

*(i)* $\pi^1_\# \gamma = e^{-\alpha}\mu_1$;

*(ii)* $\pi^2_\# \gamma = e^{-\beta}\mu_2$;

*(iii)* $\alpha(x)+\beta(y) = c(x,y)$ *for $\gamma$-a.e. $(x,y)$. In particular if $\mu_1$ is absolutely continuous with respect to the Lebesgue measure one has $\nabla\alpha(x) = \partial_x c(x,y)$ for $\gamma$-a.e. $(x,y)$.*

*(iv)* $C_{c,f}(\mu_1,\mu_2) = \mu_1(\mathbb{R}^d)+\mu_2(\mathbb{R}^d) - 2\gamma(\mathbb{R}^d\times\mathbb{R}^d)$. *In particular we have that $\pi^1_\#\gamma$ and $\pi^2_\#\gamma$ are unique and $\alpha$ and $\beta$ are uniquely defined in the support of $\mu_1$ and $\mu_2$, respectively.*

Some stability properties follow, both in term of the measures and of the costs.

**Proposition 6.2.4.** *Let us consider an L-Lipschitz cost $c\geq 0$. Then if $\mu_{n,i} \rightharpoonup \mu_i$ for $i=1,2$ and all the measures are supported on a bounded domain $\Omega$ then, denoting by $\alpha_n,\beta_n$ the maximizers in the dual problem, we have that $\alpha_n\to\alpha$ and $\beta_n\to\beta$ locally uniformly where $\beta$ and $\alpha$ are maximizers in the dual problem for $\mu_1$ and $\mu_2$. Moreover $C_{c,f}(\mu_{n,1},\mu_{n,2})\to C_{c,f}(\mu_1,\mu_2)$.*

*Proof.* First we show that $\lim_{n\to\infty} C_{c,f}(\mu_{n,1},\mu_{n,2}) = C_{c,f}(\mu_1,\mu_2)$. Let us consider $\{\gamma_n\}$ the set of optimal plans in (6.2.1) which forms a precompact set because the associated marginals do [103, Prop. 2.10]. Thus, from a subsequence of indices that achieves $\liminf_{n\to\infty} C_{c,f}(\mu_{n,1},\mu_{n,2})$, one can again extract a subsequence for which the optimal plans weakly converge, to an *a priori* suboptimal plan. Using the joint semicontinuity of the entropy and the continuity of the cost we deduce

$$\liminf_{n\to\infty} C_{c,f}(\mu_{n,1},\mu_{n,2}) \geq C_{c,f}(\mu_1,\mu_2).$$

Similarly, the sequence of optimal dual variables $\alpha_n,\beta_n$ form a precompact set since it is a sequence of bounded $L$-Lipschitz functions (see [103, Lem. 4.9]). Any weak cluster point

$\alpha_0, \beta_0$ satisfies $\alpha_0 + \beta_0 \le c$ and in particular, taking again a subsequence of indices achieving the lim sup and $\alpha_0, \beta_0$ a cluster point of it, we have

$$\begin{aligned}\limsup_{n\to\infty} C_{c,f}(\mu_{n,1},\mu_{n,2}) &= \limsup_{n\to\infty} \int_\Omega (1-e^{-\alpha_n})\,\mathrm{d}\mu_{n,1} + \int_\Omega (1-e^{-\beta_n})\,\mathrm{d}\mu_{n,2}\\ &= \int_\Omega (1-e^{-\alpha_0})\,\mathrm{d}\mu_1 + \int_\Omega (1-e^{-\beta_0})\,\mathrm{d}\mu_2 \le C_{c,f}(\mu_1,\mu_2).\end{aligned}$$

Therefore, the limit of the costs is the cost of the limits and the inequalities are equalities. We deduce that every weak limit of $\{\gamma_n\}$ is an optimal plan, and also that $\alpha_0$ and $\beta_0$ are the unique maximizers for the dual problem of $\mu_1$ and $\mu_2$, proving the claim. □

**Proposition 6.2.5.** *Let us consider a increasing sequence of lower semi-continuous costs $c_n(x,y)$ and let us denote by $c(x,y) = \lim_n c_n(x,y)$, $C_n := C_{c_n,f}$ and $C := C_{c,f}$. Then for every $\mu_1, \mu_2 \in \mathscr{M}(\mathbb{R}^d)$ we have $C_n(\mu_1,\mu_2) \uparrow C(\mu_1,\mu_2)$. Moreover*

*(i) any weak limit of optimal plans $\gamma_n$ for $C_n(\mu_1,\mu_2)$ is optimal for $C(\mu_1,\mu_2)$;*

*(ii) if $\phi_n, \psi_n$ are optimal potentials for $C_n(\mu_1,\mu_2)$, we have $\phi_n \to \phi$ in $L^1(\mu_1)$ and $\psi_n \to \psi$ in $L^1(\mu_2)$, where $\phi$ and $\psi$ are optimal potentials for $C$;*

*(iii) in the case $c = c_\ell$ and $c_n = g_n(|x-y|)$ (as in Lemma 6.2.1) we have also that $(\phi_n, \nabla\phi_n) \to (\phi, \nabla\phi)$ in $L^2(\mu_1)$ and similarly for $\psi_n$.*

*Proof.* As in the previous proof we take $\gamma_n$ as minimizers for the primal problem of $C_n(\mu_1,\mu_2)$. They form a pre compact set and so up to subsequences they converge to $\gamma$, which is *a priori* a suboptimal plan for $C(\mu_1,\mu_2)$. Let us fix $m > 0$ and then we know that for any $n \ge m$ we have $c_n \ge c_m$ and so

$$C_n(\mu_1,\mu_2) \ge H(\pi^1_\#\gamma_n|\mu_1) + H(\pi^2_\#\gamma_n|\mu_2) + \int c_m\,\mathrm{d}\gamma_n.$$

Now, using the semicontinuity of the entropy and the semicontinuity of $c_m$ we get

$$\liminf_{n\to\infty} C_n(\mu_1,\mu_2) \ge H(\pi^1_\#\gamma|\mu_1) + H(\pi^2_\#\gamma|\mu_2) + \int c_m\,\mathrm{d}\gamma.$$

Taking the supremum in $m$ and then the definition of $C$ we get

$$\liminf_{n\to\infty} C_n(\mu_1,\mu_2) \ge H(\pi^1_\#\gamma|\mu_1) + H(\pi^2_\#\gamma|\mu_2) + \int c\,\mathrm{d}\gamma \ge C(\mu_1,\mu_2). \tag{6.2.4}$$

Noticing that $C(\mu_1,\mu_2) \ge C_n(\mu_1,\mu_2)$ we can conclude. In particular $\gamma$ is optimal since we have equality in all inequalities of (6.2.4).

In order to prove (ii) notice that, since in every inequality we had equality, in particular we have $H(\pi^1_\#\gamma_n|\mu_1) \to H(\pi^1_\#\gamma|\mu_1)$. Since $\gamma_n \rightharpoonup \gamma$ and $\pi^1_\#\gamma_n = (1-\phi_n)\mu_1$ and $\pi^1_\#\gamma = (1-\phi)\mu_1$ we conclude by Lemma 6.2.6.

In the case we are in the hypotheses of (iii), we have that

$$\gamma_n = (\mathrm{id}, h_n(\nabla\phi))_\#(1-\phi_n)\mu_1 \rightharpoonup (\mathrm{id}, h(\nabla\phi))_\#(1-\phi)\mu_1,$$

where $h_n$ is converges pointwise to $h$. Then by Lemma 6.2.6 below, we deduce that $\nabla\phi_n \to \nabla\phi$ in measure with respect to $\mu_1$. Using Proposition 6.2.7 we have

$$\int (4\phi_n^2 + |\nabla\phi_n|^2)\,\mathrm{d}\mu_1 \le 4C_n(\mu_1,\mu_2) \le 4C(\mu_1,\mu_2) = \int (4\phi^2 + |\nabla\phi|^2)\,\mathrm{d}\mu_1,$$

where the last inequality can be proven as the first part of Proposition 6.2.7 in the case $c = c_\ell$. Now, let $v_n = (2\phi_n, \nabla\phi_n)$ and $v = (2\phi, \nabla\phi)$. Since $\limsup_n \int |v_n|^2\,\mathrm{d}\mu_1 \le \int |v|^2\,\mathrm{d}\mu_1$, we have $v_n \rightharpoonup w$ in $L^2(\mu_1)$, up to subsequences: however since $v_n \to v$ in measure we conclude $v_n \rightharpoonup v$ in $L^2(\mu_1)$ but then using $\|v\|_2^2 \ge \lim_{n\to\infty} \|v_n\|_2^2$ we finally conclude $v_n \to v$ in $L^2(\mu_1)$. □

**Lemma 6.2.6.** *Let $(X,d,\bar\mu)$ be a metric measure space with finite measure. Let $\mu_n = f_n\bar\mu$ and $\mu = f\bar\mu$ where $f, f_n$ are densities. Let us suppose that $\mu_n \rightharpoonup \mu$ and $H(\mu_n|\bar\mu) \to H(\mu|\bar\mu)$. Moreover let us assume there exists maps $T_n : X \to \mathbb{R}^d$ such that $(\mathrm{id}, T_n)_\#\mu_n \rightharpoonup (\mathrm{id}, T)_\#\mu$, with $T$ bounded. Then we have, up to a subsequence,*

*(i) $f_n \to f$ in $L^1(\bar\mu)$;*

*(ii) $T_n(x) \to T(x)$ for $\mu$-a.e. $x \in X$.*

*Proof.* The first point is a well known consequence of the strict convexity of $t\log t$ (see for example [163, Theorem 3]) and the fact that since $f_n$ are uniformly $\mu$ integrable we have $f_n \rightharpoonup f$ in $L^1(\bar\mu)$. For the second point it is sufficient to notice that thanks to the first point we have $(\mathrm{id}, T_n)_\#\mu \rightharpoonup (\mathrm{id}, T)_\#\mu$ and then we can apply [5, Lemma 5.4.1] and then pass to a subsequence. □

The following estimate allows to capture the infinitesimal behavior of the entropy-transport metrics.

**Proposition 6.2.7.** *Let $\mu_1 \in \mathscr{M}_+(\Omega)$ be an absolutely continuous measure and let $\phi := 1 - e^{-\alpha}$ where $\alpha$ is the potential relative to $\mu_1$ in the minimization problem $C_n(\mu_1,\mu_2)$. Then we have that*

$$\int_{\mathbb{R}^d} (|\nabla\phi|^2 + 4\phi^2)\,d\mu_1 \le 4C_n(\mu_1,\mu_2);$$

*moreover for every $f \in \mathscr{C}^2_c(\mathbb{R}^d)$ we have*

$$\left| \int_{\mathbb{R}^d} f\,d(\mu_2 - \mu_1) + \frac{1}{2}\int_{\mathbb{R}^d} \nabla f\cdot\nabla\phi\,d\mu_1 + 2\int_{\mathbb{R}^d} f\phi\,d\mu_1 \right| \le 5\|f\|_{\mathscr{C}^2} C_n(\mu_1,\mu_2),$$

*where $\|f\|_{\mathscr{C}_2} = \|f\|_\infty + \|\nabla f\|_\infty + \|D^2 f\|_\infty$.*

*Proof.* In the sequel we will work always $\gamma$-a.e., where $\gamma$ is the optimal plan for $C_n(\mu_1,\mu_2)$. We have $\alpha(x) + \beta(y) = c(x,y) = g_n(|y-x|)$ and $|\nabla\alpha(x)| = g_n'(|y-x|)$. We first find an upper bound for the gradient term, using an inequality from Lemma 6.2.1:

$$\begin{aligned}
\int_{\mathbb{R}^d} |\nabla\phi|^2\,\mathrm{d}\mu_1 &= \int_{\mathbb{R}^d} |\nabla\alpha|^2 e^{-2\alpha}\,\mathrm{d}\mu_1 = \int_{\mathbb{R}^{2d}} |g_n'(|y-x|)|^2 e^{-\alpha}\,\mathrm{d}\gamma \\
&\le \int_{\mathbb{R}^{2d}} 4(e^c - 1)e^{-\alpha}\,\mathrm{d}\gamma = \int_{\mathbb{R}^{2d}} 4e^{\alpha+\beta}e^{-\alpha}\,\mathrm{d}\gamma - 4\int_{\mathbb{R}^{2d}} e^{-\alpha}\,\mathrm{d}\gamma \\
&= 4\int e^{\beta}\,\mathrm{d}\gamma - 4\int e^{-\alpha}\,\mathrm{d}\gamma = 4\,\mu_2(\mathbb{R}^d) - 4\int e^{-2\alpha}\,\mathrm{d}\mu_1\,.
\end{aligned}$$

Adding the term $4\int|\phi|^2 d\mu_1$ allows to prove the first inequality:

$$\begin{aligned}\int_{\mathbb{R}^d}(|\nabla\phi|^2+4|\phi|^2)\,\mathrm{d}\mu_1 &\le 4\mu_2(\mathbb{R}^d)+4\int\left[(1-e^{-\alpha})^2-e^{-2\alpha}\right]d\mu_1\\ &=4\left(\mu_2(\mathbb{R}^d)+\mu_1(\mathbb{R}^d)-2(e^{-\alpha}\mu_1)(\mathbb{R}^d)\right)\\ &=4C_n(\mu_1,\mu_2)\,.\end{aligned}$$

As a byproduct, we have also shown:

$$\int_{\mathbb{R}^{2d}}(e^c-1)e^{-\alpha}\,d\gamma\le C_n(\mu_1,\mu_2). \tag{6.2.5}$$

For the second part we will split the estimate in several parts:

$$\int_{\mathbb{R}^d}f\,\mathrm{d}\mu_2-\int_{\mathbb{R}^d}f\,\mathrm{d}\mu_1+\frac{1}{2}\int_{\mathbb{R}^d}\nabla f\cdot\nabla\phi\,\mathrm{d}\mu_1+2\int_{\mathbb{R}^d}f\phi\,\mathrm{d}\mu_1=$$

$$=\int_{\mathbb{R}^{2d}}\left(f(y)e^{\beta}-f(x)e^{\alpha}+\frac{1}{2}\nabla f(x)\cdot\nabla\alpha+2f(x)(1-e^{-\alpha})e^{\alpha}\right)\mathrm{d}\gamma=$$

$$\begin{aligned}&\int_{\mathbb{R}^{2d}}\Big(f(y)-f(x)-\nabla f(x)\cdot(y-x)\Big)e^{-\alpha}\,\mathrm{d}\gamma\\ &\qquad+\int_{\mathbb{R}^{2d}}\nabla f(x)\cdot\left(\frac{\nabla\alpha(x)}{2}+(y-x)e^{-\alpha}\right)\mathrm{d}\gamma\\ &\qquad+\int_{\mathbb{R}^{2d}}f(y)(e^{\beta}-e^{-\alpha})\,d\gamma+\int_{\mathbb{R}^{2d}}f(x)(1-e^{-\alpha})^2e^{\alpha}\,\mathrm{d}\gamma\\ &\qquad=(I)+(II)+(III)+(IV)\end{aligned}$$

Now for (I) we use the Lagrange formula for the remainder in Taylor expansion and then, using (6.2.5) and Lemma 6.2.1 (ii), namely $e^c-1\ge|x-y|^2$, we get

$$\begin{aligned}(I)&\le\frac{1}{2}\|D^2f\|_\infty\int_{\mathbb{R}^{2d}}|y-x|^2e^{-\alpha}\,\mathrm{d}\gamma\\ &\le\frac{1}{2}\|D^2f\|_\infty\int_{\mathbb{R}^{2d}}(e^c-1)e^{-\alpha}\,\mathrm{d}\gamma\\ &\le\frac{1}{2}\|D^2f\|_\infty C_n(\mu_1,\mu_2).\end{aligned}$$

For the second term:

$$\begin{aligned}(II)&=\int_{\mathbb{R}^{2d}}\nabla f\cdot\left(\frac{\nabla\alpha(x)}{2}+(y-x)\right)e^{-\alpha}\mathrm{d}\gamma+\frac{1}{2}\int_{\mathbb{R}^{2d}}\nabla f\cdot\nabla\alpha(x)\cdot(1-e^{-\alpha})\,\mathrm{d}\gamma\\ &\le\|\nabla f\|_\infty\left(\int_{\mathbb{R}^{2d}}\left|\frac{\nabla\alpha(x)}{2}+(y-x)\right|e^{-\alpha}\,\mathrm{d}\gamma+\frac{1}{2}\int_{\mathbb{R}^{2d}}|\nabla\alpha|\cdot|1-e^{-\alpha}|\,\mathrm{d}\gamma\right)\\ &=(IIa)+(IIb)\end{aligned}$$

Now we have $\nabla\alpha(x) = \frac{(x-y)}{|x-y|}g_n'(|x-y|)$ and in particular we have

$$2\left|\frac{\nabla\alpha(x)}{2} + (y-x)\right| = |g_n'(z) - 2z|,$$

where $z = |x-y|$ and we can verify that $|g_n'(z) - 2z| \le 4(e^{g_n(z)} - 1)$ independently of $n$. In fact if $g_n'(z) \ge 1$ or $z \ge 1$ this is obvious since we have

$$|g_n'(z) - 2z| \le \max\{g_n'(z), 2z\} \le \max\{g_n'(z), 2z\}^2 \le 4(e^{g_n} - 1),$$

where in the last inequality we used Lemma 6.2.1 (ii). In the case $z \le 1$ instead we have that $g_n'(z) = g'(z) = 2\tan(z)$ (by Lemma 6.2.1 (i)), and so, calling $t = \tan(z)$, and using that $|\arctan''(t)| \le 1$ for every $t$ we have

$$|g_n'(z) - 2z| = 2|\tan(z) - z| = 2|t - \arctan(t)| \le t^2 = \frac{1}{4}g_n'(z)^2 \le e^{g_n(z)} - 1.$$

In particular we obtain

$$\begin{aligned}(IIa) &= \frac{\|\nabla f\|_\infty}{2}\int_{\mathbb{R}^{2d}} |g_n'(z) - 2z|\, e^{-\alpha}\mathrm{d}\gamma \\ &\le \|\nabla f\|_\infty \int_{\mathbb{R}^{2d}} 4(e^c - 1)\, e^{-\alpha}\mathrm{d}\gamma \le 4\|\nabla f\|_\infty \cdot C_n(\mu_1, \mu_2).\end{aligned}$$

Then we use the inequality $ab \le \frac{a^2}{4c} + cb^2$ with $a = |\nabla\alpha|$, $b = |1 - e^{-\alpha}|$ and $c = e^\alpha$ in order to get

$$\begin{aligned}(IIb) &= \frac{\|\nabla f\|_\infty}{2}\int_{\mathbb{R}^d} |\nabla\alpha| \cdot |1 - e^{-\alpha}| e^{-\alpha}\,\mathrm{d}\mu_1 \\ &\le \frac{\|\nabla f\|_\infty}{2}\int_{\mathbb{R}^d} \big(\frac{1}{4}|\nabla\alpha|^2 e^{-2\alpha} + (1 - e^{-\alpha})^2\big)\,\mathrm{d}\mu_1 \\ &= \frac{\|\nabla f\|_\infty}{8}\int_{\mathbb{R}^d} (|\nabla\phi|^2 + 4\phi^2)\,\mathrm{d}\mu_1 \le \frac{\|\nabla f\|_\infty}{2} C_n(\mu_1, \mu_2).\end{aligned}$$

Then we have

$$\begin{aligned}(III) &\le \int_{\mathbb{R}^{2d}} |f|(e^{\beta+\alpha} - 1)e^{-\alpha}\,\mathrm{d}\gamma \\ &\le \|f\|_\infty \int_{\mathbb{R}^{2d}} (e^c - 1)e^{-\alpha}\,\mathrm{d}\gamma \le \|f\|_\infty C_n(\mu_1, \mu_2)\end{aligned}$$

and in the end we conclude with

$$(IV) \le \int_{\mathbb{R}^d} |f||1 - e^{-\alpha}|^2\,\mathrm{d}\mu_1 \le \|f\|_\infty \int_{\mathbb{R}^d} |\phi|^2\,\mathrm{d}\mu_1 \le \|f\|_\infty C_n(\mu_1, \mu_2). \qquad \square$$

### 6.2.2. One step of the scheme

For a measure $\mu \in \mathscr{M}_+(\Omega)$ which is absolutely continuous of density bounded by 1 and a cost function $c$, consider the problem

$$\mathrm{prox}^c_{\tau G}(\mu) := \mathrm{argmin}\left\{ -\lambda \int_\Omega \rho + \frac{C_{c,f}(\mu,\rho)}{2\tau} \;:\; \rho \in \mathscr{M}_+(\Omega), \rho \le 1 \right\} \tag{6.2.6}$$

which corresponds to an implicit Euler step as introduced in (6.1.5): notice however that for a general cost $c$, the optimal entropy-transport cost $C_{c,f}$ is not always the square of a distance. We first show that this *proximal operator* is well defined.

**Proposition 6.2.8** (Existence and uniqueness)**.** *If $2\tau\lambda < 1$ and $c : \Omega^2 \to [0,\infty]$ is a strictly convex, l.s.c., increasing function of the distance, then $\mathrm{prox}^c_{\tau G}$ is a well defined map on $\{\mu \in L^1_+(\Omega) \;:\; \mu \le 1\}$, that is, the minimization problem admits a unique minimizer. Denoting $\rho := \mathrm{prox}^c_{\tau G}(\mu)$, it holds*

*(i)* $\sqrt{\rho(\Omega)} \le \frac{1+2\lambda\tau}{1-2\lambda\tau}\sqrt{\mu(\Omega)}$;

*(ii)* $C_{c,f}(\mu,\rho) \le 2\tau\lambda(\rho(\Omega)-\mu(\Omega)) \le \frac{(4\lambda\tau)^2}{(1-2\lambda\tau)^2}\mu(\Omega)$.

*Proof.* The definition of the proximal operator requires to solve a problem of the form

$$\min_{\gamma \in \mathscr{M}_+(\Omega\times\Omega)} I(\gamma) \quad \text{where} \quad I(\gamma) := \int_{\Omega\times\Omega} c\,d\gamma + F_1(\pi^1_\#\gamma) + F_2(\pi^2_\#\gamma)$$

where $F_1(\sigma) = H(\sigma|\mu)$ and $F_2(\sigma) = \inf_{\rho\le 1}\{H(\sigma|\rho) - 2\tau\lambda\rho(\Omega)\}$ are both convex functions of the marginals of $\gamma$ (note that the optimal $\rho$ in the definition of $F_2$ is explicit using the pointwise first order optimality conditions : $\rho(x) = \min\{1, \sigma(x)/(1-2\tau\lambda)\}$, for a.e. $x \in \Omega$. In order to prove the existence of a minimizer, one could apply Theorem 3.1.2 directly, but let us give a proof which do not assume compactness of $\Omega$ (although this is assumed), since we need the mass estimates anyways.

Remark that $\gamma = \mu\otimes\mu$ is feasible and $I$ is weakly l.s.c. so we only have to show that the closed sublevel set $S = \{\gamma \in \mathscr{M}_+(\Omega^2) \;:\; I(\gamma) \le I(\mu\otimes\mu)\}$ is tight, and thus compact, in order to prove the existence of a minimizer. Let us consider $\gamma \in S$ and $\rho(x) = \min\{1, \pi^2_\#(x)/(1-2\tau\lambda)\}$: then we have

$$-\lambda\mu(\Omega) \ge I(\mu\otimes\mu) \ge I(\gamma) \ge -\lambda\rho(\Omega) + \frac{C_{c,f}(\mu,\rho)}{2\tau}.$$

Then, using Lemma 6.2.2 we obtain

$$-\lambda\mu(\Omega) \ge -\lambda\rho(\Omega) + \frac{1}{2\tau}\left(\sqrt{\mu(\Omega)} - \sqrt{\rho(\Omega)}\right)^2;$$

by rearranging the term, it follows $\sqrt{\rho(\Omega)} \le \frac{1+2\lambda\tau}{1-2\lambda\tau}\sqrt{\mu(\Omega)}$, so we have a bounded mass as long as $2\lambda\tau < 1$. Thanks to the positivity of $H$, this implies that $F_2(\pi^2_\#\gamma)$ is lower bounded for

$\gamma \in S$ and thus both $F_1(\pi^1_\# \gamma)$ and $\int c\,\mathrm{d}\gamma$ are upper bounded (since nonnegative). Incidentally, we obtained also (ii), an estimate for the dissipated energy

$$C_{c,f}(\mu,\rho) \le 2\tau\lambda\,(\rho(\Omega)-\mu(\Omega)) \le \frac{(4\lambda\tau)^2}{(1-2\lambda\tau)^2}\mu(\Omega). \tag{6.2.7}$$

The upper bound on $F_1(\pi^1_\# \gamma)$ and the superlinear growth at infinity of the entropy implies that $S$ is bounded and $\{\gamma_1 := \pi^1_\#\gamma \,:\, \gamma \in S\}$ is tight (see [103, Prop. 2.10]). Let $\varepsilon > 0$ and $K_1$ be a compact set such that $\gamma_1(\Omega \setminus K_1) < \varepsilon/2$ for all $\gamma \in S$. The assumptions on $c$ guarantee that the set $K_\lambda := \{(x,y) \in K_1 \times \Omega \,:\, c(x,y) \le \lambda\}$ is compact for $\lambda \in \mathbb{R}$, and by the Markov inequality, $\int_{K_1\times\Omega} c\,\mathrm{d}\gamma \ge \lambda\,\gamma((K_1\times\Omega)\setminus K_\lambda)$. Consequently, for $\lambda$ big enough, it holds for all $\gamma \in S$:

$$\gamma(\Omega^2 \setminus K_\lambda) = \gamma_1(\Omega\setminus K_1) + \gamma((K_1\times\Omega)\setminus K_\lambda) \le \varepsilon/2 + \varepsilon/2 \le \varepsilon$$

which proves the tightness of $S$ and shows the existence of a minimizer.

For uniqueness, observe that if $\gamma$ is a minimizer, then it is a deterministic coupling. Indeed, $\gamma$ is an optimal coupling for the cost $c$ between its marginals, which are absolutely continuous. But $c$ satisfies the *twist condition* which guarantees that any optimal plan is actually a map, because $c$ is a strictly convex function of the distance.

Now take two minimizers $\gamma^a$ and $\gamma^b$ and define $\tilde\gamma = \frac{1}{2}\gamma^a + \frac{1}{2}\gamma^b$ which is also a minimizer, by convexity. Note that $\tilde\gamma$ must be a deterministic coupling too, which is possible only if the maps associated to $\gamma^a$ and $\gamma^b$ agree almost everywhere on $(\mathrm{spt}(\pi^1_\#\gamma^a) \cap \mathrm{spt}(\pi^1_\#\gamma^b)) \times \Omega$. Finally, since all the terms in the functional are convex, it must hold $F_1(\pi^1_\#\tilde\gamma) = \frac{1}{2}F_1(\pi^1_\#\gamma^a) + \frac{1}{2}F_1(\pi^1_\#\gamma^b)$. But $F_1$ is strictly convex so $\pi^1_\#\gamma^a = \pi^1_\#\gamma^b$ and thus $\gamma^a = \gamma^b$. This, of course, implies the uniqueness of $\rho$ which is explicitly determined from the optimal $\gamma$. □

We now use the dual formulation in order to get information on the minimizer.

**Proposition 6.2.9.** *Let us consider $\rho = \mathrm{prox}^c_{\tau G}(\mu)$. Then there exists a Lipschitz optimal potential $\phi$ relative to $\rho$ for the problem $C_{c,f}(\rho,\mu)$ such that $\phi \le 2\tau\lambda$ and*

$$\rho(x) = \begin{cases} 1 & \text{if } \phi < 2\tau\lambda, \\ [0,1] & \text{if } \phi = 2\tau\lambda. \end{cases}$$

*Proof.* In the problem (6.2.6), let us consider a competitor $\bar\rho \le 1$ and define $\rho_\varepsilon = \rho + \varepsilon(\bar\rho - \rho)$. Since $\rho_\varepsilon$ is still admissible as a competitor we have that

$$-\lambda\int_\Omega \rho + \frac{C_{c,f}(\mu,\rho)}{2\tau} \le -\lambda\int_\Omega \rho_\varepsilon + \frac{C_{c,f}(\mu,\rho_\varepsilon)}{2\tau}.$$

We can now use the fact that, if $\phi_\varepsilon$ and $\psi_\varepsilon$ are the maximizing potentials in the dual formulation for $C_{c,f}(\mu,\rho_\varepsilon)$, we have

$$C_{c,f}(\mu,\rho_\varepsilon) = \int \phi_\varepsilon\,\mathrm{d}\rho_\varepsilon + \int \psi_\varepsilon\,\mathrm{d}\mu \qquad \text{and} \qquad C_{c,f}(\mu,\rho) \ge \int \phi_\varepsilon\,\mathrm{d}\rho + \int \psi_\varepsilon\,\mathrm{d}\mu,$$

because $\phi_\varepsilon, \psi_\varepsilon$ are admissible potentials also for $\mu$ and $\rho$. In particular we deduce that

$$-\lambda\int_\Omega \rho + \frac{1}{2\tau}\int \phi_\varepsilon\,\mathrm{d}\rho \le -\lambda\int_\Omega \rho_\varepsilon + \frac{1}{2\tau}\int \phi_\varepsilon\,\mathrm{d}\rho_\varepsilon;$$

$$0 \le -\lambda \int_\Omega \varepsilon \mathrm{d}(\bar\rho - \rho) + \varepsilon \frac{1}{2\tau} \int \phi_\varepsilon \, \mathrm{d}(\bar\rho - \rho).$$

Dividing this inequality by $\varepsilon$ and then let $\varepsilon \to 0$, using that $\phi_\varepsilon \to \phi_0$ locally uniformly by Proposition 6.2.4, we get

$$\int_\Omega (\phi_0 - 2\lambda\tau)\, \mathrm{d}\rho \le \int_\Omega (\phi_0 - 2\lambda\tau)\, \mathrm{d}\bar\rho \qquad \forall 0 \le \bar\rho \le 1,$$

where $\phi_0$ is an optimal (Lipschitz) potential relative to $\rho$. This readily implies

$$\rho(x) = \begin{cases} 1 & \text{if } \phi_0 < 2\tau\lambda \\ [0,1] & \text{if } \phi_0 = 2\tau\lambda \\ 0 & \text{if } \phi_0 > 2\tau\lambda. \end{cases}.$$

Now it is sufficient to take $\phi = \inf\{2\tau\lambda, \phi_0\}$ and we have that $\phi$ is still an admissible potential since $(1-\phi)(1-\psi) \ge (1-\phi_0)(1-\psi) \ge e^{-c}$ and moreover we have $\int \phi \, \mathrm{d}\rho = \int \phi_0 \, \mathrm{d}\rho$ and so $\phi$ also is optimal. □

**Lemma 6.2.10** (Stability of prox)**.** *Let $(c_n)_{n\in\mathbb{N}}$ be an increasing sequence of Lipschitz cost functions, each satisfying the hypotheses of Proposition 6.2.8 and let $c$ be the limit cost. Then $\rho_n = \mathrm{prox}^{c_n}_{\tau G}(\mu)$ converges weakly to $\rho = \mathrm{prox}^{c}_{\tau G}(\mu)$.*

*Proof.* By Proposition 6.2.8 we know that $\{\rho_n\}$ have equi-bounded mass and in particular, up to subsequences, $\rho_n \rightharpoonup \bar\rho$ which is such that $\bar\rho \le 1$. Fix $m \in \mathbb{N}$ and $n \ge m$; by the minimality of $\rho_n$ we know that for every $\nu$ we have

$$\begin{aligned} \frac{C_{c_m,f}(\rho_n,\mu)}{2\tau} - \lambda \int_\Omega \rho_n &\le \frac{C_{c_n,f}(\rho_n,\mu)}{2\tau} - \lambda \int_\Omega \rho_n \\ &\le \frac{C_{c_n,f}(\nu,\mu)}{2\tau} - \lambda \int_\Omega \nu \le \frac{C_{c,f}(\nu,\mu)}{2\tau} - \lambda \int_\Omega \nu. \end{aligned}$$

Taking now the limit as $n \to \infty$, using the continuity of $C_{c_m,f}$ (Proposition 6.2.4), we get

$$-\lambda \int_\Omega \bar\rho + \frac{C_{c_m,f}(\bar\rho,\mu)}{2\tau} \le -\lambda \int_\Omega \nu + \frac{C_{c,f}(\nu,\mu)}{2\tau}.$$

Finally, we can take the limit $m \to \infty$ and use that $C_{c_m,f} \uparrow C_{c,f}$ (Proposition 6.2.5) in order to get

$$-\lambda \int_\Omega \bar\rho + \frac{C_{c,f}(\bar\rho,\mu)}{2\tau} \le -\lambda \int_\Omega \nu + \frac{C_{c,f}(\nu,\mu)}{2\tau},$$

that is, $\bar\rho$ is a minimizer for the limit problem and so by the uniqueness $\bar\rho = \rho$. □

**Lemma 6.2.11.** *Let us consider $\rho = \mathrm{prox}^{c_\ell}_{\tau G}(\mu) = \mathrm{prox}^{\widehat{W}_2}_{\tau G}(\mu)$. If $\Omega$ is a regular domain[1] then there exists $p \in H^1(\Omega)$ that verifies $p \ge 0$, $p(1-\rho) = 0$ and such that*

$$\int_\Omega (|\nabla p|^2 + 4(p-\lambda)^2)\, d\rho \le \frac{\widehat{W}_2^2(\rho,\mu)}{\tau^2}$$

[1] we need that $\Omega$ is an $H^1$ extension domain, that is, there exists $C > 0$ such that for every $f \in H^1(\Omega)$ there exists $\tilde f$ with $\tilde f|_\Omega = f$ and $\|\tilde f\|_{H^1(\mathbb{R}^d)} \le C\|f\|_{H^1(\Omega)}$.

*and such that for all* $f \in \mathscr{C}^2(\Omega)$,

$$\left| \int_\Omega f\, d(\mu-\rho) - \tau \int_\Omega (\nabla p \cdot \nabla f + 4(p-\lambda) f)\, d\rho \right| \leq \mathrm{cst} \cdot \|f\|_{\mathscr{C}^2} \widehat{W}_2^2(\rho,\mu).$$

*Proof.* We first use the approximated problem $\rho_n = \mathrm{prox}_{\tau G}^{c_n}(\mu)$. By Lemma 6.2.10 we know that $\rho_n \rightharpoonup \rho$. Using Proposition 6.2.9 we know there exist optimal potentials $\phi_n$ such that, calling $p_n = \frac{1}{2\tau}(2\tau\lambda - \phi_n)$, we have $p_n \in H^1(\Omega)$, $p_n \geq 0$ and $p_n(1-\rho_n) = 0$. Moreover, thanks to Proposition 6.2.7 we also know that

$$\tau^2 \int_\Omega \left(|\nabla p_n|^2 + 4(\lambda - p_n)^2\right) \mathrm{d}x = \frac{1}{4} \int \left(|\nabla \phi_n|^2 + 4|\phi_n|^2\right) \mathrm{d}\rho_n \leq C_n(\rho_n, \mu) \tag{6.2.8}$$

$$\left| \int_\Omega f\, d\mu - \int_\Omega f\, d\rho_n + \tau \int_\Omega (-\nabla p_n \cdot \nabla f + 4(\lambda - p_n) f)\, \mathrm{d}x \right| \leq 5\|f\|_{C^2} C_n(\rho_n, \mu) \tag{6.2.9}$$

In particular, using Equations (6.2.7) and (6.2.8), we get that $p_n$ is equibounded in $H^1(\Omega)$. Thanks to the hypothesis on $\Omega$, there exist a sequence $\tilde{p}_n$ equibounded in $H^1(\mathbb{R}^d)$ such that $\tilde{p}_n|_\Omega = p_n$; in particular there is a subsequence of $\tilde{p}_n$ that is weakly converging in $H^1(\mathbb{R}^d)$ and strongly in $L^2$ to some $p \in H^1$, $p \geq 0$. Since we have $\rho_n \rightharpoonup \rho$ in duality with $\mathscr{C}$ and so also in duality with $L^1$, thanks to the $L^\infty$ bound, we get that $\int_\Omega p_n(1-\rho_n) \to \int_\Omega p(1-\rho)$ and so we have $\int_\Omega p(1-\rho)\, \mathrm{d}x = 0$ that implies $p(1-\rho) = 0$ almost everywhere in $\Omega$, since $\rho \leq 1$ and $p \geq 0$. Now we can pass to the limit both Equation (6.2.8) and (6.2.9) getting the conclusion. □

### 6.2.3. Convergence of minimizing movement

We consider an initial density $\rho_0 \in L^1_+(\Omega)$ and define the discrete gradient flow scheme as introduced in (6.1.5) which depends on a *time step* $\tau > 0$

$$\begin{cases} \rho_0^\tau = \rho_0 \in L^1_+(\Omega) \\ \rho_{n+1}^\tau = \mathrm{prox}_{\tau G}^{\widehat{W}_2}(\rho_n) \quad \text{for } n \geq 1, \end{cases} \tag{6.2.10}$$

define $p_{n+1}^\tau$ as the pressure relative to the couple $\rho_n^\tau, \rho_{n+1}^\tau$ (provided by Lemma 6.2.11) and extend all these quantities in a piecewise constant fashion as in Definition 6.1.1 in order to obtain a family of time dependant curves $(\rho^\tau, p^\tau)$:

$$\begin{cases} \rho^\tau(t) = \rho_{n+1}^\tau \\ p^\tau(t) = p_{n+1}^\tau \end{cases} \quad \text{for } t \in ]\tau n, \tau(n+1)]. \tag{6.2.11}$$

The next lemmas exhibit the regularity in time of $\rho^\tau$, which improves as $\tau$ diminishes.

**Lemma 6.2.12.** *There exists a constant* $\mathrm{cst} > 0$ *such that for any* $\tau > 0$*, the sequence of minimizers satisfies*

$$\sum_k \widehat{W}_2^2(\rho_k^\tau, \rho_{k+1}^\tau) \leq \mathrm{cst} \cdot \tau.$$

*Proof.* By optimality, $\rho^\tau_{k+1}$ satisfies $\widehat{W}_2^2(\rho^\tau_k,\rho^\tau_{k+1}) \le 2\tau\left(G(\rho^\tau_k) - G(\rho^\tau_{k+1})\right)$. By summing over $k$, one obtains a telescopic sum which is upper bounded by $2\tau(G(\rho_0) - \inf G)$ and $\inf G = -\lambda|\Omega|$ is finite because $\Omega$ has a finite Lebesgue measure $|\Omega|$. □

The consequence of this bound is a Hölder property, a standard result for gradient flows.

**Lemma 6.2.13** (Discrete Hölder property)**.** *Let $\rho_0 \le 1$ and $T > 0$. There exists a constant* $\mathrm{cst} > 0$ *such that for all $\tau > 0$ and $s,t \ge 0$, it holds*

$$\widehat{W}_2(\rho^\tau_t,\rho^\tau_s) \le \mathrm{cst}\cdot(\tau + |t-s|)^{1/2}.$$

*In particular, if $(\tau_n)_{n\in\mathbb{N}}$ converges to 0, then, up to a subsequence, $(\rho^{\tau_n})$ weakly converges to a $(1/2)$-Hölder curve $(\rho_t)_{t\in[0,T]}$.*

*Proof.* The first result is direct if $s$ and $t$ are in the same time interval $]\tau k, \tau(k+1)]$ so we suppose that $s - t \ge \tau$ and let $k,l$ be such that $t \in ]\tau(k-1),\tau k]$ and $s \in ]\tau(l-1),\tau l]$. By the triangle inequality and the Cauchy-Schwarz inequality, one has

$$\widehat{W}_2(\rho^\tau(t),\rho^\tau(s)) \le \sum_{i=k}^{l-1} \widehat{W}_2(\rho^\tau_i,\rho^\tau_{i+1}) \le \left(\sum_{i=k}^{l-1} \widehat{W}_2(\rho^\tau_i,\rho^\tau_{i+1})^2\right)^{\frac12} (l-k)^{\frac12}.$$

By using Lemma 6.2.12 and the fact that $|l-k| \le 1 + |s-t|/\tau$, the first claim follows.

As for the second claim, let us adapt the proof of Arzelà-Ascoli theorem to this discontinuous setting. Let $(q_i)_{i\in\mathbb{N}}$ be a enumeration of $\mathbb{Q}\cap[0,T]$ and let $\tau_k \to 0$. Since $\{\rho^{\tau_k}(q_1)\}$ is bounded in $L^\infty \cap L^1(\Omega)$, it is weakly pre-compact and thus there is a subsequence $\tau_{k^{(1)}}$ such that $\rho^{\tau_{k^{(1)}}}(q_1)$ converges. By induction, for any $i \in \mathbb{N}$, one can extract a subsequence $\tau_{k^{(i)}}$ from $\tau_{k^{(i-1)}}$ so that $\rho^{\tau_{k^{(i)}}}(q_i)$ converges.

Now, we form the diagonal subsequence $(\rho^m)_{m\in\mathbb{N}}$ whose $m$-th term is the $m$-th term in the $m$-th subsequence $\rho^{\tau_{k^{(m)}}}$. By construction, for every rational $q_i \in [0,T]$, $\rho^m(q_i)$ converges. Moreover, for every $t \in [0,T]$ and for $q \in [0,T]\cap\mathbb{Q}$

$$\widehat{W}_2(\rho_n(t),\rho_m(t)) \le C(\tau_{k^{(n)}} + |t-q|)^{\frac12} + \widehat{W}_2(\rho_n(q),\rho_m(q)) + C(\tau_{k^{(m)}} + |t-q|)^{\frac12}$$

by the triangle inequality and the discrete Hölder property. As $q$ cana be arbitrarily close to $t$, one sees that $(\rho_m(t))_{m\in\mathbb{N}}$ is Cauchy and thus converges. Let us denote $\rho$ the limit. For $0 \le s \le t \le T$, and $m \in \mathbb{N}$, it holds

$$\widehat{W}_2(\rho(t),\rho(s)) \le \widehat{W}_2(\rho(t),\rho_m(t)) + \widehat{W}_2(\rho_m(t),\rho_m(s)) + \widehat{W}_2(\rho_m(s),\rho(s)).$$

In the right-hand side, the middle term is upper bounded by $C(\tau_{k^{(m)}} + |s-t|)^{\frac12} \to C|s-t|^{\frac12}$ and the other terms tend to 0 so by taking the limit $m \to \infty$, one obtains the $(1/2)$-Hölder property. □

Collecting all the estimates established so far, we obtain an existence result.

**Proposition 6.2.14** (Existence of solutions)**.** *The family $(\rho^\tau, p^\tau)$ defined in* (6.2.11) *admits weak cluster points $(\rho,p)$ as $\tau \downarrow 0$ which are solutions to the evolution PDE* (6.1.2) *on $[0,T]$, for all $T > 0$.*

*Proof.* Define the sequences of momentum $E_n^\tau = -\rho_n^\tau \nabla p_n^\tau$ and source term $D_n^\tau = 4\rho_n^\tau(\lambda - p_n^\tau)$ and extend these quantities in a piecewise constant fashion as in (6.2.11) on the time interval $[0,T]$ for some $T > 0$. Gathering the results, let us first show that there exists a constant $C = C(T,\rho_0)$ such that

(i) $\rho_t^\tau p_t^\tau = p_t^\tau$ and $p_t^\tau \geq 0$;

(ii) $\int_0^T \int_\Omega (|\nabla p^\tau|^2 + |p^\tau|^2)\mathrm{d}x\mathrm{d}t \leq \int_0^T \int_\Omega (|E_t^\tau|^2 + |D_t^\tau|^2)\,\mathrm{d}x\,\mathrm{d}t \leq C$;

(iii) $\left|\int_\Omega \phi\, d\rho_t^\tau - \int_\Omega \phi\, d\rho_s^\tau - \int_s^t \int_\Omega (E_r^\tau \cdot \nabla\phi + D_r^\tau \phi)\,\mathrm{d}x\,\mathrm{d}r\right| \leq C\|\phi\|_{\mathscr{C}^2} \max\{\tau, \sqrt{\tau}\}$, for all $\phi \in \mathscr{C}^2(\Omega)$;

(iv) $\int_0^T \int_\Omega |\nabla p^\tau| \mathrm{d}x\mathrm{d}t < C$.

Property (i) is a direct from Lemma 6.2.11 and the definition of the curves $p^\tau$ and $\rho^\tau$. One then proves (ii) and (iv) by using Lemma 6.2.11 and property (i). Indeed, one has

$$\int_\Omega (|E_n^\tau|^2 + |D_n^\tau|^2)\mathrm{d}x \leq \int_\Omega (|\nabla p_n^\tau|^2 + 4(\lambda - p_n)^2)\mathrm{d}\rho_n \leq \frac{1}{\tau^2}\widehat{W}_2^2(\rho_{n-1}^\tau, \rho_n^\tau).$$

Integrating now the interpolated quantities it follows, by Lemma 6.2.12,

$$\int_0^T \int_\Omega (|E_t^\tau|^2 + |D_t^\tau|^2)\mathrm{d}x\mathrm{d}t \leq \sum_{n=1}^{\lceil \frac{T}{\tau} \rceil} \tau \int_\Omega (|E_n^\tau|^2 + |D_n^\tau|^2)\mathrm{d}x \leq \frac{1}{\tau}\sum_n \widehat{W}_2^2(\rho_{n-1}^\tau, \rho_n^\tau) \leq C.$$

Property (iii) is obtained from Lemma 6.2.11 in a similar way. Indeed, for all $\phi \in \mathscr{C}^2(\Omega)$, by denoting $I_a^b = \int_a^b \int_\Omega (E_r^\tau \nabla\phi + D_r^\tau \phi)\mathrm{d}x\mathrm{d}r$,

$$\begin{aligned}\left|\int_\Omega \phi(\rho_t^\tau - \rho_s^\tau)\mathrm{d}x - I_s^t\right| &= \left|I_s^{k\tau} - I_t^{l\tau} + \right. \\ &\qquad \left. \sum_{i=k}^{l-1}\left(\int_\Omega \phi(\rho_{i+1}^\tau - \rho_i^\tau)\mathrm{d}x - \tau\int_\Omega (E_{i+1}^\tau \nabla\phi + D_{i+1}^\tau\phi)\mathrm{d}x\right)\right| \\ &\leq |I_s^{k\tau}| + |I_{kl}^t| + C\|\phi\|_{\mathscr{C}^2}\sum_{i=k}^{l-1}\widehat{W}_2^2(\rho_i^\tau, \rho_{i+1}^\tau)\end{aligned}$$

where $k = \lceil \frac{s}{\tau} \rceil$ and $l = \lceil \frac{t}{\tau} \rceil$. By Lemma 6.2.12, the last term is bounded by $C\tau$ and by Lemma 6.2.11, $|I_s^{k\tau}|$ and $|I_t^{l\tau}|$ are controlled by $C\sqrt{\tau}$ :

$$|I_s^{k\tau}| \leq \tau\left|\int_\Omega (E_k^\tau \nabla\phi + D_k^\tau\phi)\mathrm{d}x\right| \leq \|\phi\|_{H^1(\Omega)}\widehat{W}_2(\rho_{k-1}, \rho_k) \leq C\|\phi\|_{H^1(\Omega)}\sqrt{\tau}.$$

So property (iii) is shown.

Let us now take a sequence $\tau_k \to 0$ and pass those relations to the limit. Recall that from the discrete Hölder property (Lemma 6.2.13), up to a subsequence, $(\rho^{\tau_k})$ admits a weakly continuous limit $(\rho_t)_{t\in[0,T]}$. Moreover, thanks to the $L^2$-norm bound (ii) we have, up to a subsequence $(E^{\tau_k}, D^{\tau_k}) \rightharpoonup (E,D)$. In particular, looking at relation (iii), we obtain, for all $\phi \in \mathscr{C}^2(\Omega)$,

$$\int_\Omega \phi d\rho_t - \int_\Omega \phi d\rho_s = \int_s^t \int_\Omega (E_r \cdot \nabla\phi + D_r\phi)\mathrm{d}x\mathrm{d}r.$$

which means that $(\rho, E, D)$ is a weak solution of $\partial_t \rho_t + \nabla \cdot E_t = D_t$.

In order to conclude it remains to prove that $D = 4(\lambda - p)\rho$ and $E = -\rho \nabla p$ for some admissible pressure field $p$. As $p^\tau$ is a bounded family in the Hilbert space $L^2([0,1], H^1(\Omega))$, there exist weak limits $p$ when $\tau \to 0$. The property $p \geq 0$ is obvious but the Hele-Shaw complementary relation $p(1-\rho) = 0$ is more subtle. We obtain it by combining the spatial regularity of $p^\tau$ with the time regularity of $\rho^\tau$ as was done for the Wasserstein case in [110]. Using the complementary relation $p^\tau(1-\rho^\tau) = 0$, one has for all $0 < a < b < T$:

$$0 = \frac{1}{b-a} \int_a^b \int_\Omega p_t^\tau(x)(1-\rho_a^\tau(x))\mathrm{d}x\mathrm{d}t + \frac{1}{b-a} \int_a^b \int_\Omega p_t^\tau(x)(\rho_a^\tau(x) - \rho_t^\tau(x))\mathrm{d}x\mathrm{d}t\,.$$

Denoting $p_{[a,b]} := \int_a^b p_t \mathrm{d}t$, the first term converges to $\int_\Omega p_{[a,b]}(x)(1-\rho_a(x))\mathrm{d}x$ because $(p^\tau_{[a,b]})_\tau$ converges to $p_{[a,b]}$ — weakly in $H^1(\Omega)$ and thus strongly in $L^2_{loc}(\Omega)$ since $\Omega$ bounded — and $(\rho_a^\tau)_\tau$ converges weakly to $\rho_a$ in duality with $L^1(\Omega)$. Additionally, for every Lebesgue point $a$ of $t \mapsto p_t$ (seen as a map in the separable Hilbert space $L^2(\Omega)$) we have

$$\int_\Omega p_{[a,b]}(x)(1-\rho_a(x))\mathrm{d}x\mathrm{d}t \xrightarrow[b\to a]{} \int_\Omega p_a(x)(1-\rho_a(x))\mathrm{d}x\,.$$

For the second term, we use Lemma 6.2.15 (stated below) and obtain

$$\begin{aligned}
\int_a^b \int_\Omega p_t^\tau(x)(\rho_a^\tau(x) - \rho_t^\tau(x))\mathrm{d}x\mathrm{d}t &\leq 2\int_a^b \|p_t^\tau\|_{H^1(\mathbb{R}^d)} \widehat{W}_2(\rho_a^\tau, \rho_t^\tau)\mathrm{d}t \\
&\leq C\sqrt{\tau + (b-a)} \left(\int_a^b \|p_t^\tau\|^2_{H^1(\mathbb{R}^d)}\mathrm{d}t\right)^{\frac{1}{2}} \left(\int_a^b \mathrm{d}t\right)^{\frac{1}{2}} \\
&\leq C(b-a)\sqrt{1 + \tau/(b-a)} \left(\int_a^b \|p_t^\tau\|^2_{H^1(\mathbb{R}^d)}\mathrm{d}t\right)^{\frac{1}{2}}.
\end{aligned}$$

Notice that since the geodesics used in Lemma 6.2.15 may exit the domain $\Omega$ we have to use the $H^1$ norm of $p^{\tau(t,\cdot)}$ on the whole $\mathbb{R}^d$, in the sense that we extend it, and thanks to the regularity of $\Omega$ we have $\|p^\tau(t,\cdot)\|_{H^1(\mathbb{R}^d)} \leq C\|p^\tau(t,\cdot)\|_{H^1(\Omega)}$. In this way the functions $t \mapsto \|p^\tau(t,\cdot)\|^2_{H^1(\mathbb{R}^d)}$ are $\tau$-uniformly bounded in $L^1([0,1])$ and so admit a weak cluster point $\sigma \in \mathscr{M}_+([0,T])$ as $\tau \to 0$. Thus, for a.e. $a \in [0,T]$,

$$\lim_{\tau\to 0} \frac{1}{b-a} \int_a^b \int_\Omega p_t^\tau(x)(\rho_a^\tau(x) - \rho_t^\tau(x))\mathrm{d}x\mathrm{d}t \leq C\sqrt{\sigma([a,b])} \xrightarrow[b\to a]{} 0\,.$$

As a consequence, for a.e. $a$, $\int_\Omega p_a(x)(1-\rho_a(x))\mathrm{d}x = 0$, and since $p \geq 0$ and $\rho \leq 1$, this implies $p(1-\rho) = 0$ a.e.

We are finally ready to recover the expressions for $E$ and $D$ by writing these quantities as linear functions of $p$ and $\rho$ which are preserved under weak convergence and then plugging the nonlinearities back using $p(1-\rho) = 0$. For $D^\tau \rightharpoonup D$ one has

$$D^\tau = 4(\lambda - p^\tau)\rho^\tau = 4(\lambda\rho^\tau - p^\tau) \underset{\tau\to 0}{\rightharpoonup} 4(\lambda\rho - p) = 4(\lambda - p)\rho = D,$$

while for $E^\tau \rightharpoonup E$ one has

$$E^\tau = -\rho^\tau \nabla p^\tau = -\nabla p^\tau \underset{\tau\to 0}{\rightharpoonup} -\nabla p = -\rho\nabla p = E\,. \qquad \square$$

In the proof, we used the following lemma which is well-known for the case of Wasserstein distances, and illustrates a link between $\widehat{W}_2$ and $H^{-1}$ norms. Its proof is a simple adaptation of the Wasserstein case, given the geodesic convexity result of Theorem 6.2.16. Notice that, as in the Wasserstein case, this lemma can be generalized to the case where $L^p$ bounds on the measures imply a comparison between $\widehat{W}_2$ and the $W^{-1,q}$ norm, where $\frac{1}{p}+\frac{2}{q}=1$.

**Lemma 6.2.15.** *Let $(\mu,\nu)\in\mathscr{M}_+(\mathbb{R}^d)$ be absolutely continuous measures with density bounded by a constant $C$. Then, for all $\phi\in H^1(\mathbb{R}^d)$, it holds*

$$\int_{\mathbb{R}^d}\phi d(\mu-\nu)\leq 2\sqrt{C}\|\phi\|_{H^1(\mathbb{R}^d)}\widehat{W}_2(\mu,\nu)\,.$$

*Proof.* Consider a minimizing geodesic $(\rho_t)_{t\in[0,1]}$ between $\mu$ and $\nu$ for the distance $\widehat{W}_2$ and $(v,\alpha)\in L^2([0,1],L^2(\rho_t))$ the associated velocity and growth fields. These quantities satisfy the constant speed property $\|v_t\|^2_{L^2(\rho_t)}+\|\alpha_t\|^2/4=\widehat{W}_2^2(\mu,\nu)$ for a.e. $t\in[0,1]$ (see Proposition 1.1.19 or [94, 104] for the $\mathbb{R}^d$ setting). Moreover, by Theorem 6.2.16, $L^\infty$ bounds are preserved along geodesics. Let us take $\phi\in H^1(\mathbb{R}^d)$ and notice that by approximation we can suppose that its support is bounded; then it holds

$$\begin{aligned}\int_{\mathbb{R}^d}\phi\mathrm{d}(\mu-\nu)&=\int_0^1\frac{d}{dt}\left(\int_{\mathbb{R}^d}\phi\rho_t\right)\mathrm{d}t=\int_0^1\int_{\mathbb{R}^d}(\nabla\phi\cdot v_t+\phi\alpha_t)\,\rho_t\mathrm{d}x\mathrm{d}t\\&\leq\sqrt{\int_0^1\int_{\mathbb{R}^d}(|\nabla\phi|^2+4|\phi|^2)\rho_t\mathrm{d}x\mathrm{d}t}\sqrt{\int_0^1\int_{\mathbb{R}^d}(|v_t|^2+\frac{1}{4}|\alpha_t|^2)\rho_t\mathrm{d}x\mathrm{d}t}\\&\leq 2\sqrt{C}\|\phi\|_{H^1(\mathbb{R}^d)}\widehat{W}_2(\mu,\nu)\,.\end{aligned}$$

□

This lemma relies on an announced result of geodesic convexity for $\widehat{W}_2$ [102]. We also heavily rely on this result for proving uniqueness next.

**Theorem 6.2.16.** *Let us consider $\mu_t$ be a geodesic of absolutely continuous measures connecting the two absolutely continuous measures $\mu_0$ and $\mu_1$. Then, for every $m>1$ we have that $t\mapsto\int\left(\frac{d\mu_t}{d\mathscr{L}^d}\right)^m dx$ is convex. In particular if $\mu_1,\mu_0\leq M\mathscr{L}^d$, we have $\mu_t\leq M\mathscr{L}^d$ too.*

### 6.2.4. Proof of uniqueness

**Proposition 6.2.17** (Uniqueness)**.** *If $\Omega$ is convex, every solution of* (6.1.2) *is an* $\mathrm{EVI}_{(-2\lambda)}$ *solution of gradient flow of $G$ in the metric space $(\mathscr{M}_+(\Omega),\widehat{W}_2)$ and we have uniqueness for* (6.1.2)*.*

*Proof.* We follow the same lines as [60], using the announced geodesic convexity results from Theorem 6.2.16. Let us consider two solutions $(\rho_t^1,p_t^1)$ and $(\rho_t^2,p_t^2)$ and assume that we can prove that we have (distributionally)

$$\begin{aligned}\frac{d}{dt}\widehat{W}_2^2(\rho_t^1,\rho_t^2)=&\int_\Omega\left(-\nabla\phi_t\cdot\nabla p_t^1+4\phi_t(\lambda-p_t^1)\right)\mathrm{d}\rho_t^1\\&+\int_\Omega\left(-\nabla\psi_t\cdot\nabla p_t^2+4\psi_t(\lambda-p_t^2)\right)\mathrm{d}\rho_t^2,\quad(6.2.12)\end{aligned}$$

where $\phi_t, \psi_t$ is a couple of optimal potentials for $\rho_t^1, \rho_t^2$. Then using Lemma 6.6.1 we get

$$\frac{d}{dt}\widehat{W}_2^2(\rho_t^1,\rho_t^2) \le 4\lambda \int \phi_t \,\mathrm{d}\rho_t^1 + 4\lambda \int \psi_t \,\mathrm{d}\rho_t^2 = 4\lambda \widehat{W}_2^2(\rho_t^1,\rho_t^2),$$

and so by Grönwall's lemma it follows $\widehat{W}_2^2(\rho_t^1,\rho_t^2) \le e^{2\lambda t}\widehat{W}_2^2(\rho_t^1,\rho_t^2)$. So we are left to prove (6.2.12) in the distributional sense. Notice that (6.2.12) is true if we can prove that for every $0 < s < r < T$ we have

$$\widehat{W}_2^2(\rho_r^1,\rho_r^2) - \widehat{W}_2^2(\rho_s^1,\rho_s^2) = \int_s^r \left( \int_\Omega (\nabla\phi_t \cdot v_t^1 + \phi_t r_t^1)\mathrm{d}\rho_t^1 + \int_\Omega (\nabla\psi_t \cdot v_t^2 + \psi_t r_t^2)\mathrm{d}\rho_t^2 \right) \mathrm{d}t$$

where we can suppose $\partial_t \rho_t^1 + \nabla\cdot(v_t^1\rho_t^1) = r_t^1\rho_t^1$ with $\iint (|v_t^1|^2 + (r_t^1)^2)\mathrm{d}\rho_t^1\mathrm{d}t < \infty$ and similarly for $\rho_t^2$. Let us fix $n$ and consider $C_n(\rho_t^1,\rho_t^2)$ and a couple of optimal potentials $\phi_n, \psi_n$. In particular for every $s$ we have

$$C_n(\rho_s^1,\rho_s^2) \ge \int \phi_n \,\mathrm{d}\rho_s^2 + \int \psi_n \mathrm{d}\rho_s^2,$$

with equality for $s = t$. Now with a slight modification of [60, Lemma 2.3], we can prove that there exists a full measure set where we can differentiate both sides an the derivatives are equal. In particular, using that $t \mapsto C_n(\rho_t^1,\rho_t^2)$ is absolutely continuous, we get

$$C_n(\rho_r^1,\rho_r^2) - C_n(\rho_s^1,\rho_s^2) = \int_s^r \left( \int_\Omega (\nabla\phi_{n,t} \cdot v_t^1 + \phi_{n,t} r_t^1)\mathrm{d}\rho_t^1 + \int_\Omega (\nabla\psi_{n,t} \cdot v_t^2 + \psi_{n,t} r_t^2)\mathrm{d}\rho_t^2 \right) \mathrm{d}t$$

and then letting $n \to \infty$ we conclude, using that $(\phi_{n,t}, \nabla\phi_{n,t}) \to (\phi_t, \nabla\phi_t)$ in $L^2(\rho_t^1)$ thanks to Proposition 6.2.5 (iii).

The EVI characterization is easily deduced from those previous computations. Taking a solution $(\rho_t, p_t)_{t\in[0,T]}$ and any $\mu \in \mathscr{M}_+(\Omega)$ such that $\mu \le 1$, we have, by denoting $(\phi_t, \psi_t)$ the optimal potentials for $(\rho_t, \mu)$,

$$\frac{1}{2}\widehat{W}_2^2(\rho_t,\mu) = \int_\Omega (\nabla\phi_t \cdot v_t + \phi_t r_t)\mathrm{d}\rho_t \le 2\lambda \int_\Omega \phi_t \mathrm{d}\rho_t$$

by Lemma 6.6.1 and we conclude using Theorem 6.2.3 (i) and (iv), which proves that one has

$$2\lambda \int_\Omega \phi_t \mathrm{d}\rho_t \le G(\mu) - G(\rho_t) + \lambda \widehat{W}_2^2(\rho_t,\mu). \qquad \square$$

## 6.3. Numerical scheme

The characterization of the tumor growth model (6.1.2) as a gradient flow suggests a constructive method for computing solutions through the time discretized scheme (6.1.5). In this section we build on the numerical method established in Chapter 3 to numerically compute solutions.

First, let us recall that the resolution of one step of the scheme involves, for a given time step $\tau > 0$ and previous step $\mu \in L^1_+(\Omega)$, such that $\mu \leq \mathscr{L}^d$ to compute

$$\nu \in \operatorname{argmin} \left\{ 2\tau G(\nu) + \widehat{W}_2^2(\nu,\mu)^2 \right\} \tag{6.3.1}$$

According to Proposition 6.2.8, by using the optimal entropy-transport problem and exchanging the infima, this problem can be written in terms of one variable $\gamma$ which stands for the unknown coupling

$$\min_{\gamma \in \mathscr{M}_+(\Omega^2)} \left\{ \int_{\Omega^2} c(x,y)d\gamma + H(\pi^1_\# \gamma | \mu) + \inf_{\nu \in \mathscr{M}_+(\Omega)} \left\{ H(\pi^2_\# \gamma | \nu) + 2\tau G(\nu) \right\} \right\}, \tag{6.3.2}$$

which admits a unique minimizer $\gamma^*$ and the optimal $\nu^*$ can be recovered from $\gamma^*$ through the first order pointwise optimality conditions as

$$\nu^* = \min\{1, \pi^2_\# \gamma^* / (1 - 2\tau\lambda)\}.$$

The subject addressed in this Section is the numerical resolution of (6.3.2). Since $\lambda$ is redundant with $\tau$, we fix $\lambda = 1$ for the rest of this chapter.

### 6.3.1. Spatial discretization

Let $\mathscr{W} = (W_i, x_i)_{i=1}^N$ be a pointed partition of a compact domain $\Omega \subset \mathbb{R}^d$ where $x_i$ is a point which belongs to the interior of the set $W_i$ for all $i$ (in our experiments, $W_i$ will always be a $d$-dimensional cube and $x_i$ its center). We denote by $\operatorname{diam} \mathscr{W}$ the quantity $\max_i \operatorname{diam} W_i$. An atomic approximation of $\mu \in \mathscr{M}_+(\Omega)$ is given by the measure

$$\mu_{\mathscr{W}} = \sum_{i=1}^N \alpha_i w_i \delta_{x_i}$$

where $w_i \coloneqq \mathscr{L}^d(W_i)$ is the (positive) Lebesgue measure of $W_i$, $\alpha_i \coloneqq \mu(W_i)/w_i$ are the locally averaged densities and $\delta_{x_i}$ is the Dirac measure of mass 1 concentrated at the point $x_i$. This is a proper approximation since for a sequence of partitions $(\mathscr{W}_k)_{k\in\mathbb{N}}$ such that $\operatorname{diam} \mathscr{W}_k \to 0$ then $\mu_{\mathscr{W}_k}$ converges weakly to $\mu$ (indeed, $\mu_{\mathscr{W}_k}$ is the pushforward of $\mu$ by the map $W_{k,i} \ni x \mapsto x_{k,i}$ which converges uniformly to the identity map as $k \to \infty$).

Now assume that we are given a vector $\alpha \in \mathbb{R}^N_+$. For a discrete coupling $\gamma \in \mathbb{R}^{N\times N}$ seen as a square matrix, let $J$ be the convex functional defined as

$$J(\gamma) \coloneqq \langle c, \gamma \rangle + F_1(\gamma w) + F_2(\gamma^T w) \tag{6.3.3}$$

where $\gamma w$, $\gamma^T w$ are matrix/vector products, $\langle c, \gamma \rangle := \sum_{i,j} c(x_i, y_j)\gamma_{i,j} w_i w_j$ and

$$F_1 : \mathbb{R}^N \ni \beta \mapsto H(\beta|\alpha)$$
$$F_2 : \mathbb{R}^N \ni \beta \mapsto \min_{s\in[0,1]^N} \left\{ H(\beta|s) - 2\lambda\tau \sum_i s_i w_i \right\}$$

and, for $\alpha, \beta \in \mathbb{R}^N_+$, the discrete relative entropy (Definition 3.3.1). With these definitions, solving the finite dimensional convex optimization problem

$$\gamma^* \in \arg\min_{\gamma\in\mathbb{R}^{N\times N}_+} J(\gamma) \tag{6.3.4}$$

is nothing but solving a discrete approximation of (6.3.2) where the maximum density constraint is not with respect to the Lebesgue measure anymore, but with respect to its discretized version. This is formalized in the following simple proposition.

**Proposition 6.3.1.** *Let $\mathscr{W}$ be a partition of $\Omega$ as above and let $\gamma^*$ be obtained through* (6.3.4). *Then the measure $\nu := \sum_i \beta_i w_i \delta_{w_i}$ where $\beta = \min\{1, ((\gamma^*)^T w)/(1-2\lambda\tau)\}$ does not depend on the choice of $\gamma^*$ and is a minimizer of*

$$\min_{\nu\in\mathscr{M}_+(\Omega)} \widehat{W}_2^2(\mu_{\mathscr{W}}, \nu) - 2\tau\nu(\Omega) + \iota_C(\nu)$$

*where $\iota_C$ is the convex indicator of the set $C$ of measures which are upper bounded by the discretized Lebesgue measure $\sum_i w_i \delta_{x_i}$.*

*Proof.* This result essentially follows by construction. Let us denote by (P) the minimization problem in the proposition: (P) can be written as a minimization problem over couplings $\gamma \in \mathscr{M}_+(\Omega\times\Omega)$ as in (6.3.2). But in this case, any feasible $\gamma$ is discrete because both marginals must be discrete in order to have finite relative entropies. Thus (P) reduces to the finite dimensional problem (6.3.4) and the expression for $\beta$ is obtained by first order conditions. Finally, (P) is strictly convex as a function of $\nu$, hence the uniqueness. □

The following proposition guarantees that the discrete measure $\nu_k$ built in Proposition 6.3.1 properly approximates the continuous solution.

**Proposition 6.3.2** (Consistency of discretization)**.** *Let $(\mathscr{W}_k)_{k\in\mathbb{N}}$ be a sequence of partitions of $\Omega$ such that* $\operatorname{diam}\mathscr{W}_k \to 0$ *and for all $k$ compute $\nu_k$ as in Proposition 6.3.1. Then the sequence $(\nu_k)$ converges weakly to the continuous minimizer of* (6.3.1).

*Proof.* As a sequence of bounded measures on a compact domain, $(\nu_k)$ admits weak cluster points. Let $\bar{\nu}$ be one of them. The fact that for all $k$, $\nu_k$ is upper bounded by the discretized Lebesgue measure $\sum_i w_i\delta_{x_i}$ implies that $\bar{\nu}$ is upper bounded by the Lebesgue measure in $\mathbb{R}^d$, since the discretized Lebesgue measure weakly converges to the Lebesgue measure. Now let $\sigma \in \mathscr{M}_+(\Omega)$ be any measure of density bounded by 1. By Proposition 6.3.1, one has for all $k \in \mathbb{N}$,

$$\widehat{W}_2^2(\mu_{\mathscr{W}_k}, \nu_k) - 2\tau\nu_k(\Omega) \le \widehat{W}_2^2(\mu_{\mathscr{W}_k}, \sigma_{\mathscr{W}_k}) - 2\tau\sigma_{\mathscr{W}_k}(\Omega)\,.$$

Since the distance $\widehat{W}_2$ and the total mass are continuous functions under weak convergence one obtains, in the limit $k \to \infty$,

$$\widehat{W}_2^2(\mu, \bar{\nu}) - 2\tau\bar{\nu}(\Omega) \leq \widehat{W}_2^2(\mu, \sigma) - 2\tau\sigma(\Omega)$$

which proves that $\bar{\nu}$ minimizes (6.3.1). By Proposition 6.2.8, this minimizer is unique. □

### 6.3.2. Scaling algorithm

After this thorough discretization, we can apply scaling algorithms (see Algorithm 1) to solve the entropic regularized discrete problem, with a regularization parameter $\varepsilon > 0$, see Chapter 3. The explicit form of the "proxdiv operator (3.3.5) is given in the next proposition.

**Proposition 6.3.3.** *One has* $\mathrm{proxdiv}_{F_1}(s,u,\varepsilon) = (\alpha \oslash s)^{\frac{1}{1+\varepsilon}} \odot e^{\frac{-u}{1+\varepsilon}}$ *and*

$$\mathrm{proxdiv}_{F_2}(s,u,\varepsilon) = \begin{cases} \left((1-2\tau\lambda)e^u\right)^{\frac{-1}{\varepsilon}} & \text{if } s \leq (1-2\tau\lambda)^{\frac{1+\varepsilon}{\varepsilon}} e^{\frac{u}{\varepsilon}} \\ \left(s \odot e^u\right)^{\frac{-1}{1+\varepsilon}} & \text{otherwise.} \end{cases}$$

The linear convergence rate of Theorem 3.5.21 applies in this case because $F_1$ is a relative entropy functional. The next proposition is a direct consequence of it (in the statement, we ignore points $x_k$ of zero mass $\mu(\mathscr{W}_k) = 0$ since the corresponding entries for $a^{(\ell)}$ are always 0).

**Proposition 6.3.4.** *The sequence of dual variables* $v^{(\ell)} := \varepsilon \log b^{(\ell)}$ *converges linearly in* $\ell_\infty$*-norm to the optimal (regularized) dual variable* $v_\varepsilon$*, more precisely*

$$\|v^{(\ell)} - v_\varepsilon\|_\infty \leq \frac{\|v^{(1)} - v_\varepsilon\|_\infty}{(1+\varepsilon)^\ell}.$$

*and the same holds true for* $u^{(\ell)} := \varepsilon \log a^{(\ell)}$*. Moreover, the matrix* $\gamma^{(\ell)} := (a_i^{(\ell)} K_{i,j} b_j^{(\ell)})$ *converges to the minimizer of the entropy regularized problem.*

*Proof.* The "proxdiv" operator (the one from Proposition 6.3.3 and the one from Table 5.1, line 2) preserve positivity as can be seen from their explicit expression. So Corollary 3.5.22 applies. □

Finally, recall that by Proposition 6.3.1, given the optimal regularized coupling $\gamma_\varepsilon$ one recovers the regularized discrete new step through

$$(\nu_\varepsilon)_i = \min\left\{1, \frac{\sum_j (\gamma_\varepsilon)_{ij} w_j}{1-2\lambda\tau}\right\}.$$

In what follows, we take $p_\varepsilon := (2\tau - 1 + e^{-\nu_\varepsilon})/(2\tau)$ as the expression for the regularized pressure since in the regularized version of (6.3.1), it is the term in the subgradient of the upper bound constraint at optimality. However, we do not attempt to establish a rigorous convergence result of $p_\varepsilon$ to the true pressure field.The algorithm for the computation of one step is recalled in Algorithm 8 for the reader's convenience. Note also that for very small values of the parameter $\varepsilon > 0$, it is advised to use the acceleration method described in Section 3.4 which allows to converge orders of magnitude faster in practice.

**Algorithm 8** Compute one step of the flow

1. input $\rho^{(0)}$ the initial density, $w$ the discretized Lebesgue measure and $\tau$ the time step and $\lambda$ the growth "strength" in the model;
2. for time step $k = 1, \dots, n$ do
   a) compute $\gamma$ from $\rho^{(k-1)}$ using Algorithm 1 (or its stabilized, accelerated version);
   b) define $\rho^{(k)} = (\min\{1, (\sum_j \gamma_{i,j} w_j)/(1-2\lambda\tau)\})_i$ the new density
   c) define $p^{(k)} = (2\tau - 1 + e^{v_i})/(2\tau)_i$ the new pressure

### Some remarks on convergence of the scheme

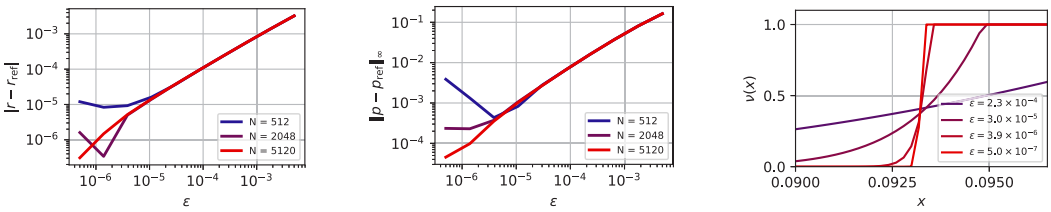
Gathering results from the previous Sections, we have proved that the scheme solves the tumor growth model (6.1.2) when

- the number of iterations $\ell \to \infty$ (Prop. 6.3.4);
- the entropic regularization $\varepsilon \to 0$ (Prop. 3.2.10);
- the spatial discretization $\operatorname{diam} \mathscr{W} \to 0$ (Prop. 6.3.2);
- the time step $\tau \to 0$ (Prop. 6.2.14).

successively, in this order. In practice, one has to fix a value for these parameters. We did not provide explicit error bounds for all these approximations, but it is worth highlighting that a bad choice leads to a bad output. As already known for Wasserstein gradient flows [109, Remark 4], there is for instance a *locking* effect when the discretization $\operatorname{diam} \mathscr{W}$ is too coarse compared to the time step $\tau$. In this case, the cost of moving mass from one discretization cell to another is indeed big compared to the gain it results in the functional.

Let us perform some numerical experiments for one step of the flow (the study of the effect of $\tau$ is postponed to the next section). We fix a time step $\tau = 0.005$, a domain $\Omega = [0,1]$, an initial density $\rho_0$ which is the indicator of the segment $[0.1, 0.9]$ and use a uniform spatial discretization $(W_i, x_i) = (\left[\frac{i-1}{N}, \frac{i}{N}\right[, \frac{2i-1}{2N})$ for $i \in \{1, \dots, N\}$. The scaling iterations are stopped as soon as $|\log a^{(\ell+1)} - \log a^{(\ell)}|_\infty < 10^{-6}$.

We first run a reference computation, with very fine parameters ($N = 8192$ and $\varepsilon = 2 \times 10^{-7}$) and then compare this with what is obtained with degraded parameters, as shown on Figure 6.1. On Figure 6.1(a)-(b), the error on the radius of the new step $r = (\sum_i \beta_i)/2N$ and the $\ell_\infty$ error on the pressure (with respect to the reference computation) are displayed. Since the initial density is the indicator of a segment, the new step is expected to be also the indicator of a segment, see Section 6.4. On Figure 6.1(c), we display the left frontier of the new density and observe how it is smoothed when $\varepsilon$ increases (the horizontal scale is strongly zoomed).




(a) Error on radius (b) Error on pressure (c) Smoothing effect

Figure 6.1.: Effect of discretization and entropic regularization

## 6.4. Test case: spherical solutions

In this section, we show that when the initial condition $\rho_0$ has unit density on a sphere and vanishes outside, then the solution of (6.1.2) are explicit, using Bessel functions. Knowing this exact solutions allows to assess the quality of the numerical algorithm.

### 6.4.1. Explicit solution

Let us construct the explicit solution for $\rho_0(x) = \chi_{B(0,r)}$. For $\alpha > -1$, we define the modified Bessel function of the first kind:

$$I_\alpha(x) := \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+\alpha+1)} \left(\frac{x}{2}\right)^{2m+\alpha}$$

and

$$H_\alpha(x) := x^{-\alpha} I_\alpha(x) \quad \text{and} \quad K_\alpha(x) := x^\alpha I_\alpha(x)\,.$$

The following mini-lemma states properties of these functions that are relevant here.

**Lemma 6.4.1.** *With the definitions as above, we have the following properties:*

*(i)* $y = I_\alpha(x)$ *satisfies the equation* $x^2y'' + xy' - (x^2 - \alpha^2)y = 0$*;*

*(ii)* $y = H_\alpha(\beta x)$ *satisfies the equation* $x^2y'' + (2\alpha+1)xy' - \beta^2x^2y = 0$ *and, up to constants is the unique locally bounded at* $0$*;*

*(iii)* $H'_\alpha(x) = xH_{\alpha+1}(x)$ *and* $K'_{\alpha+1}(x) = xK_\alpha(x)$.

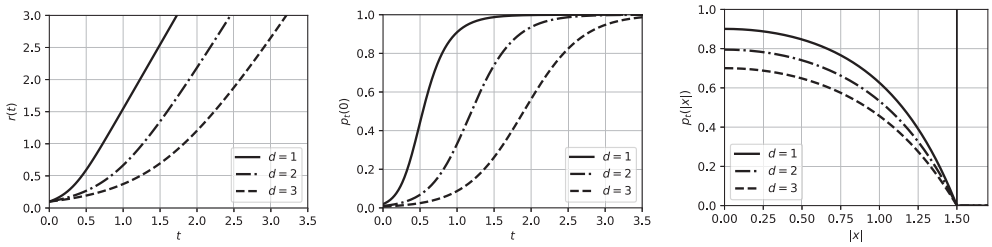
*Proof.* The proof that $I_\alpha$ satisfies the equation is trivial and can be done coefficient by coefficient since everything is converging absolutely. Moreover it is known that all the other independent functions explode like $\log(x)$ near zero. Also the equation for $H_\alpha$ is easy to derive and it is clear that it is the unique regular one. As for (ii) it can be deduced straightly deriving their formulas (using $\Gamma(x+1) = x\Gamma(x)$ where required) from the definition of $I_\alpha$ □

With the help of these Bessel functions, one can give the explicit solution when the initial condition is the indicator of a ball. Some properties of these solutions are displayed on Figure 6.2.

**Proposition 6.4.2.** *Consider the initial condition* $\rho_0 = \chi_{B_{r_0}}$ *for some* $r_0 > 0$*. Then there is a unique solution for* (6.1.2) *which is the indicator of an expanding ball* $\rho_t = \chi_{B_{r(t)}}$*, where the radius evolves as*

$$r(t) = \frac{1}{2} K_{d/2}^{-1}\left[e^{4\lambda t} K_{d/2}(2r_0)\right].$$

*Moreover, the pressure is radial and given by* $p_t(x) = \lambda\left(1 - \frac{H_{d/2-1}(2|x|)}{H_{d/2-1}(2r(t))}\right)_+$.




(a) Radius vs time

(b) Pressure at origin vs time

(c) Pressure vs distance from origin (fixed radius).

Figure 6.2.: Some properties of the spherical solutions, computed with the explicit formulae of Proposition 6.4.2 in dimensions 1, 2 and 3. In all cases, $\lambda = 1$, for (a)-(b) the initial condition is $r_0 = 0.1$ and for (c) the density radius is $r = 1.5$.

*Proof.* Taking $\beta > 0$, let us solve the evolution for the equation

$$\partial \rho_t - \nabla \cdot (\nabla p_t \rho_t) = \beta^2 (\lambda - p_t)\rho_t \quad \text{and} \quad p_t(1-\rho_t) = 0,$$

where the statement of the proposition corresponds to $\beta = 2$. In this case we suppose everything is radial, in particular we guess that $\rho_t = \chi_{B_{r(t)}}$ and also the pressure is radial. The pressure $p_t$ will depend only on $r(t)$ and it is the only function that satisfies

$$\begin{cases} -\Delta p = \beta^2(\lambda - p) & \text{in } B_{r(t)} \\ p = 0 & \text{on } \partial B_{r(t)}. \end{cases}$$

Again, by symmetry we can suppose that $p_t(x) = f_t(|x|)$ where $f_t : \mathbb{R}_+ \to \mathbb{R}$ satisfies, with the expression of the Laplacian in spherical coordinates,

$$\begin{cases} f_t''(s) + \frac{d-1}{s} f_t'(s) - \beta^2 (f_t(s) - \lambda) = 0 & \text{if } s \in [0, r(t)[, \\ f_t(s) = 0 & \text{if } s \geq r(t). \end{cases}$$

When we impose that $f_t'(0) = 0$, if we consider $g = f - \lambda$ then we can see that the solution, assuming its smoothness and using Lemma 6.4.1, is

$$g_t(s) = C_t H_\alpha(\beta s),$$

for some $C_t$ and $\alpha = \frac{d}{2} - 1$. Then the condition $f_t(r(t)) = 0$ implies that $g_t(r(t)) = -\lambda$ and this fixes $C_t = -\lambda / H_\alpha(\beta r(t))$. Now we have that $r'(t) = \frac{\partial p_t}{\partial n} = |f_t'(r(t))|$ and in particular we get:

$$\begin{aligned} r' = |g_t'(r)| &= -C_t \beta H_\alpha'(\beta r) = \lambda \beta \frac{\beta r H_{\alpha+1}(\beta r)}{H_\alpha(\beta r)} \\ &= \lambda \beta \frac{I_{\alpha+1}(\beta r)}{I_\alpha(\beta r)} = \lambda \beta \frac{K_{\alpha+1}(\beta r)}{\beta r K_\alpha(\beta r)} = \lambda \beta^2 \frac{K_{\alpha+1}(\beta r)}{\beta K_{\alpha+1}'(\beta r)}, \end{aligned}$$

and so we deduce that

$$\frac{d}{dt}\log(K_{\frac{d}{2}}(\beta r(t)) = \lambda\beta^2$$

and thus

$$K_{\frac{d}{2}}(\beta r(t)) = e^{\lambda\beta^2 t}K_{\frac{d}{2}}(\beta r(0)).$$

Since for $\alpha > 0$, $K_\alpha$ is strictly increasing and defines a bijection on $[0,+\infty[$, we have a well defined solution for (6.1.2). Uniqueness follows by Theorem 6.1.3 or [126] (since the initial density is of bounded variation). □

### 6.4.2. Numerical results and comparison

We now use the explicit spherical solution to assess the convergence of the numerical scheme when $\tau$ tends to zero. We fix an initial condition $\rho_0$ which is the indicator of a ball of radius 0.4, we fix a final time $t_f = 0.05$, and we observe the convergence towards the true solution of the continuous PDE (6.1.2) when more and more intermediate time steps are taken (ie. as $\tau$ decreases). We perform the experiments in the 1-D and the 2-D cases and the results are displayed on Figure 6.3 with the following formulae:

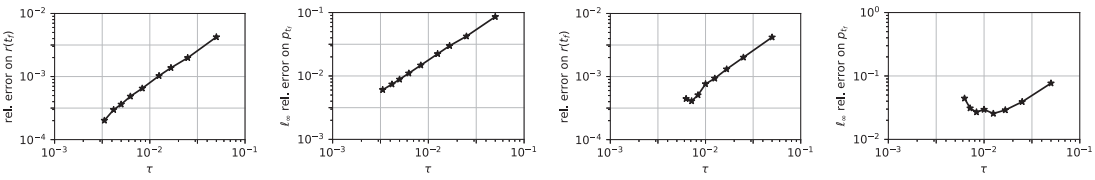
$$\text{rel. error on radius} : \frac{|r_{\text{num}} - r_{\text{th}}|}{r_{\text{th}}} \quad \text{and} \quad \text{rel. error on pressure} : \frac{\|p_{\text{num}} - p_{\text{th}}\|_\infty}{\|p_{\text{th}}\|_\infty}$$

where the subscripts "th" and "num" refer to the theoretical and numerical computations. The theoretical pressure is compared to the numerical one on the points of the grid.

**Dimension 1.** In the 1-D case, $\Omega = [0,1]$ is uniformly discretized into $N = 4096$ cells $(W_i, x_i) = ([(i-1)/N, i/N], (2i-1)/(2N))$ and $\varepsilon = 10^{-6}$. The results are displayed on Figure 6.3(a)-(b), where the numerical radius is computed through $r_{\text{num}} = (\sum_i \beta_i)/2N$.

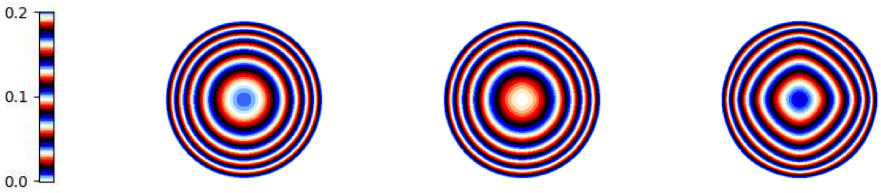
**Dimension 2.** In the 2-D case, $\Omega = [0,1]^2$ is uniformly discretized into $N^2$ sets $W_{i,j} = [(i-1)/N, i/N] \times [(j-1)/N, j/N]$ and $x_{i,j} = ((2i-1)/(2N), (2j-1)/(2N))$ with $N = 256$ and $\varepsilon = 2 \times 10^{-5}$. Compared to the 1-D case, those parameters are less fine so that the computation can run in a few hours. The numerical radius is computed as $r_{\text{num}} = \sqrt{\sum_i \beta_i/(\pi N^2)}$.

**Comments** We clearly observe the rate of convergence in $O(\tau)$ of the discretized scheme to the true solution. However, the *locking* effect (mentioned in Section 6.3.2) starts being non-negligible for small values of $\tau$ in 2-d. This effect is even more visible on the 2-d pressure because the discretization is coarser. The pressure variables at $t = t_f$ for 2 different values of $\tau$ are displayed on Figure 6.4: the solution is more sensitive to the non-isotropy of the mesh when the time step $\tau$ is small. The use of random meshes could be useful to reduce this effect.




(a) Density (1d) (b) Pressure (1d) (c) Density (2d) (d) Pressure (2d)

Figure 6.3.: For a fixed initial radius $r_0 = 0.4$ and final time $t_f = 0.05$, we assess the convergence of the scheme as the time discretization $\tau$ tends to 0 by comparing the computed $\rho_{t_f}$ and $p_{t_f}$ with the theoretical ones (see text body). Experiments performed on the interval $[0,1]$ discretized into 1024 samples with $\varepsilon = 10^{-6}$ and on the square $[0,1]^2$ discretized into $256^2$ samples and $\varepsilon = 5.10^{-6}$.




(a) theoretical (exact) (b) $\tau = 2.5 \times 10^{-2}$ (c) $\tau = 6.25 \times 10^{-3}$

Figure 6.4.: Pressure field at $t = t_f$, theoretical and numerical values for an initial density which is a ball in $[0,1]^2$. The pressure is a decreasing function of the distance to the center: here the colormap puts emphasis on the level sets so that the anisotropy due to the discretization is apparent.

## 6.5. Illustrations

We conclude this chapter with a series of flows computed numerically.

### 6.5.1. On a 1-D domain

We consider a measure $\rho_0$ of density bounded by 1 on the domain $[0,1]$ discretized into 1024 cells (Figure 6.5-(a), darkest shade of blue) and compute the evolution of the flow with parameters $\tau = 10^{-2}$ and $\varepsilon = 10^{-6}$. The density at every fourth step is shown with colors ranging from blue to yellow as time increases. Each density is displayed behind the previous ones, without loss of information since density is non-decreasing with time. The numerical pressure is displayed on Figure 6.5-(b).

**Splitting scheme** We also compare this evolution with a splitting scheme, inspired by [77], that allows for a greater freedom in the choice of the function $\Phi$ that relates the pressure to the rate of growth in (6.1.1). This scheme alternates implicit steps w.r.t. the Wasserstein metric and the Hellinger metric and is as follows. Let $\rho_0 \in \mathscr{M}_+(\Omega)$ be such that $\rho_0 \leq 1$ and define, for $n \in \mathbb{N}$,

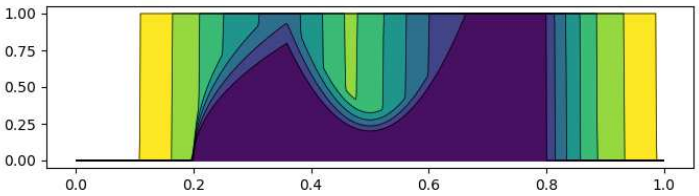
$$\begin{cases} \rho^\tau_{2n+1} = P^{W_2}(\rho^\tau_{2n}) \\ \rho^\tau_{2n+2}(x) = \rho^\tau_{2n+1}(x) \,/\, (1 - \tau\Phi(p^\tau_{2n+1}(x))) \text{ for all } x \in \Omega, \end{cases}$$

where $P^{W_2}$ is the projection on the set of densities bounded by 1 for the Wasserstein distance and $p_n$ is the pressure corresponding the projection of $\rho_n$ (as in [53]). The degenerate functional we consider is outside the domain of validity of the results of [77] and we do not know how to prove the true convergence of this scheme for non linear $\Phi$. It is thus introduced here merely for informal comparison with the case $\Phi$ linear.

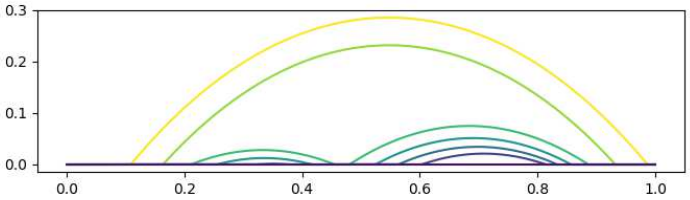
It is rather simple to adapt Algorithm 1 to compute Wasserstein projections on the set of measures of density bounded by 1.On Figure 6.6, we display such flows for rates of growth of the form $\Phi(p) = 4(1-p)^\kappa$, for three different values of $\kappa$. For these computations, the segment $[0,1]$ is divided into 1024 cells, $\tau = 10^{-2}$, $\varepsilon = 10^{-6}$ and we display the density after the projection step, at the same times than on Figure 6.5. With $\kappa = 1$, we should recover the same evolution than on Figure 6.5. As $\kappa$ increases, the rate of growth is smaller when the pressure is positive, as can be observed on Figures 6.6-(a-c).

### 6.5.2. On a 2-D non-convex domain

Our last illustration is performed on the square $[0,1]^2$ discretized into $256^2$ samples. The initial density $\rho_0$ is the indicator of a set and the parameters are $\tau = 0.015$ and $\varepsilon = 5.10^{-6}$. The first row of Figure 6.7 shows the flow at every 10th step (equivalent to a time interval of 0.15). Except at its frontier (because of discretization), the density remains the indicator of a set at all time. The bottom row of Figure 6.7 displays the pressure field, with a colormap that puts emphasis on the level sets. Notice how its level sets are orthogonal to the boundaries.

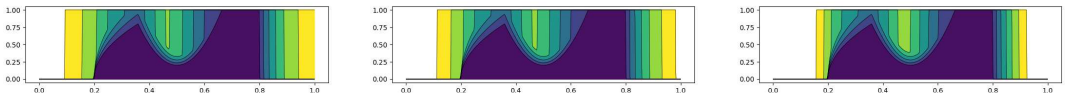



(a) Density




(b) Pressure

Figure 6.5.: A flow on $[0,1]$ and the associated pressure. Time shown are $t = 0, 0.04, 0.08, \ldots, 0.24$. Colors range from blue to yellow as time increases



(a) $\kappa = 0.01$ (b) $\kappa = 1$ (c) $\kappa = 10$

Figure 6.6.: Flow on the line with the implicit splitting scheme, with rate of growth of the form $\Phi(p) = 4(1-p)^{\kappa}$.

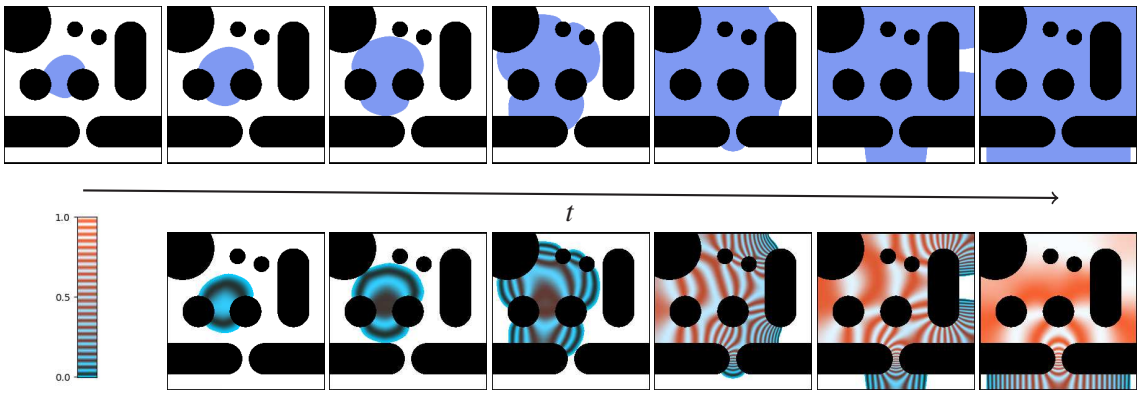



Figure 6.7.: Evolution on a non convex 2-d domain (obstacles in black). (top row) evolution of the density, the colormap is linear from white to blue as the density goes from 0 to 1. (bottom row) pressure represented with a striped colormap to make the level sets apparent. White area corresponds to $p = \rho = 0$.

## 6.6. Appendix of Chapter 6

This appendix gathers technical results that are used in the proof of uniqueness.

**Lemma 6.6.1.** *Let $\mu_0, \mu_1$ be two absolutely continuous measures on $\Omega$ convex such that $\mu_0, \mu_1 \leq 1$ and let us consider $p \in H^1(\Omega)$, such that $p \geq 0$ and $p(1-\mu_0) = 0$. Then, if we consider $\phi$ an optimal potential between $\mu_0$ and $\mu_1$, we have*

$$\int_{\mathbb{R}^d} (2p\phi + \frac{1}{2}\nabla p \cdot \nabla \phi)\, \mathrm{d}\mu_0 \geq 0.$$

*Proof.* Let us consider $\mu_t$ the geodesic between $\mu_0$ and $\mu_1$. We know that $\mu_t \leq 1$ by Theorem 6.2.16 and $\mu_t$ will be supported on $\Omega$ as well. In particular we have $\int_\Omega p\, \mathrm{d}\mu_t \leq \int_\Omega p = \int_\Omega p \mathrm{d}\mu_0$. But then using Lemma 6.6.3 it is easy to conclude

$$\int_{\mathbb{R}^d} (2p\phi + \frac{1}{2}\nabla p \cdot \nabla \phi)\, \mathrm{d}\mu_0 = -\frac{d}{dt}\bigg|_{t=0} \int p(x)\, \mathrm{d}\mu_t \geq 0. \qquad \square$$

**Theorem 6.6.2.** *Let us consider two absolutely continuous measures $\mu_0$ and $\mu_1$. Then, given $\phi$ an optimal potential for $\widehat{W}_2^2(\mu_0, \mu_1)$ relative to $\mu_0$, we consider the quantities*

$$\begin{cases} \alpha_t(x) & = (1-\phi(x)t)^2 + \frac{t^2|\nabla\phi(x)|^2}{4} \\ X_t(x) & = x - \mathrm{arctan}\big(\frac{t|\nabla\phi(x)|}{2-2t\phi(x)}\big)\frac{\nabla\phi(x)}{|\nabla\phi(x)|}. \end{cases}$$

*Then we have that $\mu_t = (X_t)_\#(\alpha_t \mu_0)$ is the geodesic for $\widehat{W}_2^2$ between $\mu_0$ and $\mu_1$.*

*Proof.* Let us consider $Y_t(x,r) = (r\sqrt{\alpha_t(x)}, X_t(x))$: then $Y_t$ are geodesics in the cone $\mathfrak{C}(\Omega)$. By the cone construction in [103] to have that if $\mu_0, \mu_1$ are two measures on $\mathbb{R}^d$ and $\phi$ is an optimal potential for $\widehat{W}_2(\mu_0, \mu_1)$ we have that $\phi(x)r^2$ is an optimal potential for $\nu_0(x,r) = \mu_0(x)f(r), \nu_1 = (Y_1)_\#(\nu_0)$ for any $f$ such that $\int f(r)r^2 \mathrm{d}r = 1$. Notice that $\mathfrak{h}\nu_0 = \mu_0$, $\mathfrak{h}\nu_1 = \mu_1$ and moreover $W_2(\nu_0, \nu_1) = \widehat{W}_2(\mu_0, \mu_1)$. In particular since $\nu_t = (Y_t)_\#\nu_0$ is a geodesic for $W_2$ on the cone, we will have that $\mu_t = \mathfrak{h}\nu_t$ is the geodesic for $\widehat{W}_2$. $\square$

**Lemma 6.6.3.** *Let $\mu_0, \mu_1$ be two absolutely continuous measures on $\mathbb{R}^d$ such that $\mu_0, \mu_1 \leq 1$ and let us consider $f \in H^1(\mathbb{R}^d)$. Then, if we consider $\mu_t$ the geodesic for $\widehat{W}_2$ between $\mu_0$ and $\mu_1$, we have*

$$\frac{d}{dt}\bigg|_{t=0} \int_{\mathbb{R}^d} f\, \mathrm{d}\mu_t = -\int_{\mathbb{R}^d} (2f\phi + \frac{1}{2}\nabla f \cdot \nabla \phi)\, \mathrm{d}\mu_0$$

*Proof.* Using Theorem 6.6.2 we can write explicitly

$$\int f\, \mathrm{d}\mu_t = \int f(X_t(x))\alpha_t(x)\, \mathrm{d}\mu_t.$$

Now we can use that $\frac{d}{dt}X_t = -\nabla\phi(x)/2\alpha_t$ in order to get

$$\frac{d}{dt}\bigg|_{t=0} \int f\, \mathrm{d}\mu_t = -\frac{1}{2}\int_{\mathbb{R}^d} \nabla f(X_t)\cdot\nabla\phi\, \mathrm{d}\mu_0 + \int_{\mathbb{R}^d} f(X_t)\Big(\frac{t|\nabla\phi|^2}{2} - 2(1-t\phi)\phi\Big)\, \mathrm{d}\mu_0.$$

While this calculation is clear when $f \in \mathscr{C}_c^\infty$ in order to make sense for $f \in H^1$ we have to consider the finite difference and integrate this inequality:

$$\begin{aligned}\frac{\int f(x)\,\mathrm{d}\mu_t - \int f(x)\,\mathrm{d}\mu_0}{t} &= \frac{1}{t}\int_{\mathbb{R}^d}\int_0^t \frac{d}{dt} f(X_t(x))\alpha_t(x)|_{t=s}\,\mathrm{d}s\,\mathrm{d}\mu_0 \\ &= \frac{1}{t}\int_{\mathbb{R}^d}\int_0^t \left[-\frac{1}{2}\nabla f(X_s)\nabla\phi(x)\right. \\ &\qquad \left. + f(X_s)\Big(\frac{s|\nabla\phi|^2}{2} - 2(1-s\phi)\phi\Big)\right]\mathrm{d}s\,\mathrm{d}\mu_0 \\ &= \int_{\mathbb{R}^d}\left[-\frac{1}{2}\mathscr{A}_t(\nabla f)\cdot\nabla\phi + \mathscr{C}_t(f)|\nabla\phi| - 2\mathscr{B}_t(f)\phi\right]\mathrm{d}\mu_0,\end{aligned}$$

where we denoted by $\mathscr{A}_t, \mathscr{B}_t, \mathscr{C}_t$ three linear operator which we will show that are acting continuously from $L^2(\mathbb{R}^d)$ to $L^2(\mu_0)$, thus proving the formula for $f \in H^1(\mathbb{R}^d)$. Explicitly we have

$$\mathscr{A}_t(g)(x) = \frac{1}{t}\int_0^t g(X_s(x))\,\mathrm{d}s \qquad \mathscr{B}_t(g)(x) = \frac{1}{t}\int_0^t g(X_s(x))(1-s\phi(x))\,\mathrm{d}s$$

$$\mathscr{C}_t(g)(x) = \frac{1}{t}\int_0^t g(X_s(x))\frac{s|\nabla\phi(x)|}{2}\,\mathrm{d}t.$$

Notice that for $0 < s \le t < 1$ we have always $1 - s\phi \ge 1 - t$. Now using Jensen and then Fubini we get:

$$\begin{aligned}\int |\mathscr{A}_t(g)(x)|^2\,\mathrm{d}\mu_0 &\le \frac{1}{t}\int_0^t \int g(X_s)^2\,\mathrm{d}\mu_0\,\mathrm{d}s \\ &\le \frac{1}{(1-t)^2}\cdot\frac{1}{t}\int_0^t\int g(X_s)^2\alpha_s\,\mathrm{d}\mu_0\,\mathrm{d}s \\ &= \frac{1}{(1-t)^2}\cdot\frac{1}{t}\int_0^t\int g(x)^2\,\mathrm{d}\mu_s\,\mathrm{d}s\end{aligned}$$

$$\begin{aligned}\int |\mathscr{B}_t(g)(x)|^2 + |\mathscr{C}_t(g)(x)|^2\,\mathrm{d}\mu_0 &\le \frac{1}{t}\int_0^t\int g(X_s)^2\left((1-s\phi(x))^2 + \frac{s^2|\nabla\phi(x)|^2}{4}\right)\mathrm{d}\mu_0\mathrm{d}s \\ &= \frac{1}{t}\int_0^t\int g(X_s)^2\alpha_s\,\mathrm{d}\mu_0\,\mathrm{d}s \\ &= \frac{1}{t}\int_0^t\int g(x)^2\,\mathrm{d}\mu_s\,\mathrm{d}s.\end{aligned}$$

We can thus conclude thanks to the fact that $\mu_s \le 1$, by Theorem 6.2.16. In particular we have $\|\mathscr{A}_t\| \le 1/(1-t)$ and $\|\mathscr{B}_t\|, \|\mathscr{C}_t\| \le 1$. Now we only need to show that $\mathscr{A}_t g \to g$, $\mathscr{B}_t g \to g$, $\mathscr{C}_t g \to 0$, where all these convergences are to be considered strongly in $L^2$. Thanks to the fact that these operators are bounded it is sufficient to show that this is true for a dense set of functions. But for $g \in \mathscr{C}_c^\infty$ for $s < 1/2$ we have $|g(X_s) - g(x)| \le L \cdot arctan(s|\nabla\phi(x)|)$ where $L$

is the Lipschitz constant of $g$. Then we have (using that $\phi, \nabla\phi \in L^2(\mu_0)$)

$$\int |\mathscr{A}_t(g)(x) - g(x)|^2 \, \mathrm{d}\mu_0 \leq L^2 \int \mathrm{arctan}(t|\nabla\phi|)^2 \, \mathrm{d}\mu_0 \to 0,$$
$$\int |\mathscr{B}_t(g)(x) - \mathscr{A}_t(g)(X)|^2 \, \mathrm{d}\mu_0 \leq \frac{\|g\|_\infty}{4} \int t^2\phi^2 \, \mathrm{d}\mu_0 \to 0,$$
$$\int |\mathscr{C}_t(g)(x)|^2 \, \mathrm{d}\mu_0 \leq \frac{\|g\|_\infty}{16} \int t^2|\nabla\phi|^2 \, \mathrm{d}\mu_0 \to 0.$$

In particular we proved that, for every $f \in H^1(\mathbb{R}^d)$ we have

$$\frac{d}{dt}\int_{\mathbb{R}^d} f \, \mathrm{d}\mu_t|_{t=0} = -\int_{\mathbb{R}^d} (\frac{1}{2}\nabla f \cdot \nabla\phi + 2f\phi) \, \mathrm{d}\mu_0. \qquad \square$$

# Chapter 7.

# Algorithmic Approach for Matrices

In this chapter, we consider an extension of the entropy-transport problem to allow optimal transport of measures whose values are positive semidefinite (PSD) matrices. The fidelity between the input PSD-valued measures is captured using the Von-Neumann quantum entropy. We propose a quantum-entropic regularization of the resulting convex optimization problem, which can be solved efficiently using an iterative scaling algorithm. The crux of this approach lies in the special algebraic properties of the quantum relative entropy, that allow to painlessly derive scaling algorithms even in this non-commutative setting. We detail a simple adaptation of the model to add a "fixed trace" constraint, which can be better suited for some applications. We also extend this formulation and the algorithm to compute barycenters of a collection of input tensor fields.
This chapter is a partial reproduction with minor modifications of [128], where additional applications to procedural noise generation, anisotropic meshing, diffusion tensor imaging and spectral texture synthesis can be found.

## 7.1. Motivation and previous work

In this chapter, we consider the optimal transport between measures that take values in the space of positive semi-definite (PSD) matrices. Our aim is to define a model that is more faithful to the geometry of PSD matrices than simply using an input dependent cost [156, 71], and that can still be solved efficiently.

Our approach is algorithmic: we consider the state of the art numerical methods for solving optimal transport and derive a model that can be solved in a similar way. This leads to a model, that turns out to be almost as easy to solve than a standard entropy regularized optimal transport (for PSD matrices of small size) but is able to take into account the additive structure of the cone of PSD matrices. In the rest of this section, we review motivations for defining optimal transport between fields of PSD matrices and previous work.

### Tensor field processing

Tensor-valued data are ubiquitous in various areas of imaging science, computer graphics and vision. In medical imaging, diffusion tensor imaging (DTI) [164] directly maps observed data to fields of tensors, and specific processing methods have been developed for this class of data (see e.g. [62, 58]). Tensor fields are also at the heart of anisotropic diffusion techniques in image processing [166], anisotropic meshing [4, 55, 23], and anisotropic texture generation [96]; they also find application in line drawing [160] and data visualization [84].

### OT and Sinkhorn on pairs of tensors

Although this is not the topic of this paper, we note that several notions of optimal transport have been defined between two tensors *without* any spatial displacement. Gurvits introduced in [81] a Sinkhorn-like algorithm to couple two tensors by an endomorphism that preserves positivity. This algorithm, however, is only known to converge in the case where the two involved tensors are the identity matrices; see [78] for a detailed analysis. In contrast to this "static" formulation that seeks for a joint coupling between the tensors, a geodesic dynamic formulation is proposed in [33]; see also [42, 41] for a related approach.

### OT on tensor fields

The simplest way to define optimal transport-like distances between arbitrary vector-valued measures is to use dual norms [116], which correspond to generalizations of $W_1$ optimal transport for which transport cost equals ground distance. The corresponding metrics, however, have degenerate behavior in interpolation and barycenter problems (much like the $L^1$ norm on functions) and only use the linear structure of matrices rather than their multiplicative structure. More satisfying notions of optimal transport have recently been proposed in a dynamical (geodesic) way [88].A static formulation of a tensor-valued optimal transport is proposed in [117], but it differs significantly from ours. It is initially motivated using a lifting that squares the number of variables, but a particular choice of cost reduces the computation to the optimization of a pair of couplings, just like a semi-coupling problem (Chapter 1).

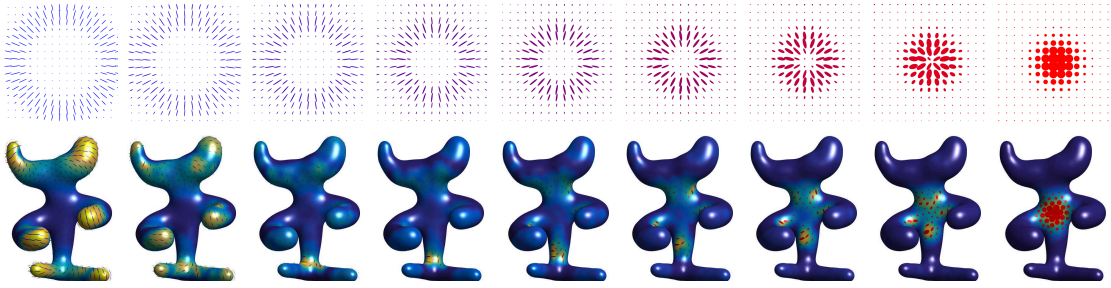



Figure 7.1.: Given two input fields of PSD matrices (displayed at times $t \in \{0,1\}$ using ellipses) on some domain (here, a 2-D planar square and a surface mesh), our Quantum Optimal Transport (Q-OT) method defines a continuous interpolating path for $t \in [0,1]$. Unlike linear interpolation schemes, Q-OT transports the "mass" of the tensors (size of the ellipses) as well as their anisotropy and orientation. This interpolation, and its extension to finding the barycenter of several input fields, is computed using a fast extension of the well-known Sinkhorn algorithm.

### 7.1.1. Notation specific to this chapter

In the following, we denote $S^d \subset \mathbb{R}^{d\times d}$ the space of symmetric matrices, $S^d_+$ the closed convex cone of positive semidefinite matrices, and $S^d_{++}$ the open cone of positive definite matrices. We denote $\exp : S^d \to S^d_{++}$ the matrix exponential, which is defined as $\exp(P) = U \operatorname{diag}_s(e^{\sigma_s})U^\top$ where $P = U \operatorname{diag}_s(\sigma_s)U^\top$ is an eigendecomposition of $P$. We denote $\log : S^d_{++} \to S^d$ the matrix logarithm $\log(P) = U \operatorname{diag}_s(\log \sigma_s)U^\top$, which is the inverse of exp on $S^d_{++}$. We adopt some conventions in order to deal conveniently with singular matrices. We extend $(P,Q) \mapsto P\log(Q)$ by lower semicontinuity on $(S^d_+)^2$, i.e. writing $Q = U \operatorname{diag}_s(\sigma_s)U^\top$ and $\tilde{P} = U^\top P U$,

$$P\log Q := \begin{cases} P\log Q & \text{if } \ker Q = \emptyset, \\ U[\tilde{P}\operatorname{diag}_s(\log\sigma_s)]U^\top & \text{if } \ker Q \subset \ker P, \\ +\infty & \text{otherwise,} \end{cases}$$

with the convention $0\log 0 = 0$ when computing the matrix product in square brackets. Moreover, for $(P,(Q_i)_{i\in I}) \in S^d \times (S^d_+)^I$, the matrix $\exp(P + \sum_i \log Q_i)$ is by convention the matrix in $S^d_+$ which kernel is $\sum_i \ker Q_i$, and is unambiguously defined on the orthogonal of this space.

A tensor-valued measure $\mu$ defined on some space $X$ is a vector-valued measure, where the "mass" $\mu(A) \in S^d_+$ associated to a measurable set $A \subset X$ is a PSD matrix. In this chapter, in order to derive computational schemes, we focus on discrete measures. Such a measure $\mu$ is a sum of Dirac masses $\mu = \sum_{i\in I} \mu_i \delta_{x_i}$ where $(x_i)_i \subset X$, and $(\mu_i)_i \in S^d_+$ is a collection of PSD matrices. In this case, $\mu(A) = \sum_{x_i\in A} \mu_i$. Figure 7.2 shows graphically some examples of tensor-valued measures; we use this type of visualization through the chapter. In the following, since the sampling points $(x_i)_i$ are assumed to be fixed and clear from the context, to ease readability, we do not make the distinction between the measure $\mu$ and the collection of matrices $(\mu_i)_i$. This is an abuse of notation, but it is always clear from context whether we are referring to a measure or a collection of matrices.

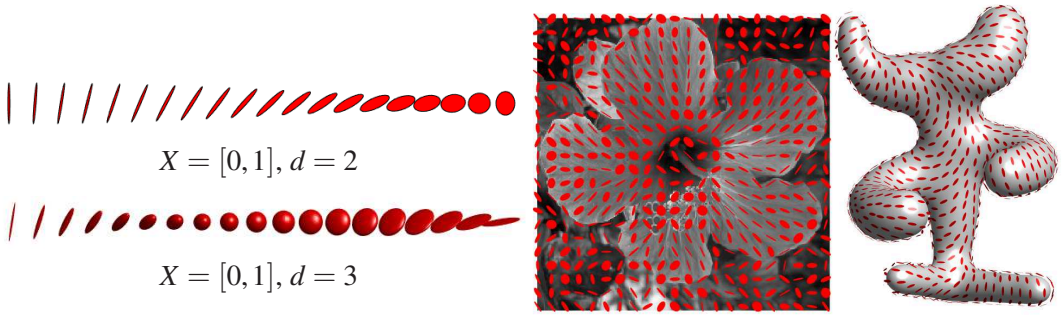



Figure 7.2.: Displays of various types of tensor-valued measures $\mu$. The principal directions of an ellipse at some $x_i \in X$ are the eigenvectors of $\mu_i \in S_+^d$, while the principal widths are given by its eigenvalues.

The quantum entropy (also called von Neumann entropy) of a tensor-valued measure is (minus)

$$H_q(\mu) := \sum_i H_q(\mu_i) \quad \text{where} \tag{7.1.1}$$

$$\forall P \in S^d, \quad H_q(P) := \operatorname{tr}(P\log(P) - P + \mathrm{Id}_{d\times d}) + \iota_{S_+^d}(P), \tag{7.1.2}$$

where $\iota_C$ is the indicator function of a closed convex set $C$. Note that $H_q$ is a convex function. The quantum relative entropy (a.k.a quantum Kullback-Leibler divergence) is the Bregman divergence associated to $H_q$. For a collection of PSD matrices $\mu = (\mu_i)_i, \xi = (\xi_i)_i$ in $S_+^d$ corresponding to measures defined on the same grid, it is defined as

$$H_q(\mu|\xi) := \sum_i H_q(\mu_i|\xi_i), \tag{7.1.3}$$

where for all $(P,Q) \in S_+^d \times S_+^d$, we denote

$$H_q(P|Q) := \operatorname{tr}(P\log(P) - P\log(Q) - P + Q) + \iota_{S_{++}^d}(P)$$

which is convex with respect to both arguments. The inner product between collections of matrices $\mu = (\mu_i)_i, \xi = (\xi_i)_i$ is

$$\langle \mu, \xi \rangle := \sum_i \langle \mu_i, \xi_i \rangle := \sum_i \operatorname{tr}(\mu_i \xi_i^\top).$$

Given a collection of matrices $\gamma = (\gamma_{i,j})_{i\in I, j\in J}$ the marginalization operators read

$$\gamma \mathbb{1}_J := \left(\sum_j \gamma_{i,j}\right)_i \quad \text{and} \quad \gamma^\top \mathbb{1}_I := \left(\sum_i \gamma_{i,j}\right)_j.$$

## 7.2. Quantum entropy-transport problem

We consider two atomic measures

$$\mu = \sum_{i \in I} \mu_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j \in J} \nu_j \delta_{y_j} \tag{7.2.1}$$

where $(x_i)_i \subset X$ and $(y_j)_j \subset Y$, and $(\mu_i)_i \in S_+^d$ and $(\nu_j)_j \in S_+^d$ are collections of PSD matrices.

### 7.2.1. Transportation of matrices

We define a coupling $\gamma = \sum_{i,j} \gamma_{i,j} \delta_{(x_i, y_j)}$ as a measure over the product $X \times Y$ that encodes the transport of mass between $\mu$ and $\nu$. In the matrix case, $\gamma_{i,j} \in S_+^d$ is now a PSD matrix, describing the exchange between $\mu_i$ and $\nu_j$. Exact (balanced) transport would mean that the marginals $(\gamma \mathbb{1}_J, \gamma^\top \mathbb{1}_I)$ must be equal to the input measures $(\mu, \nu)$. But as remarked by Ning et al. [117], in contrast to the scalar case, in the matrix case (dimension $d > 1$), this constraint is in general too strong, and there might exist no coupling satisfying these marginal constraints. Following [103], we propose to use a relaxed formulation where the discrepancy between the marginals $(\gamma \mathbb{1}_J, \gamma^\top \mathbb{1}_I)$ and the input measures $(\mu, \nu)$ is quantified according to some divergence between measures.

In the scalar case, the most natural divergence is the relative entropy (see Chapter 2). We propose here to use its quantum counterpart (7.1.3) via the following convex program

$$C_q(\mu, \nu) \coloneqq \min_{\gamma} \langle \gamma, c \rangle + \rho_1 H_q(\gamma \mathbb{1}_J | \mu) + \rho_2 H_q(\gamma^\top \mathbb{1}_I | \nu) \tag{7.2.2}$$

subject to the constraint $\forall (i,j), \gamma_{i,j} \in S_+^d$. Here $\rho_1, \rho_2 > 0$ are constants balancing the transport cost versus the cost of modifying the matrices.

The matrix $c_{i,j} \in \mathbb{R}^{d \times d}$ measures the cost of displacing an amount of (matrix) mass $\gamma_{i,j}$ between $x_i$ and $y_j$ as $\operatorname{tr}(\gamma_{i,j} c_{i,j})$. A typical cost, assuming $X = Y$ is a metric space endowed with a distance dist, is

$$c_{i,j} = \operatorname{dist}(x_i, y_j)^\alpha \operatorname{Id}_{d \times d},$$

for some $\alpha > 0$. In this case, one should interpret the trace as the global mass of a tensor, and the total transportation cost is simply

$$\langle \gamma, c \rangle = \sum_{i,j} \operatorname{dist}(x_i, y_j)^\alpha \operatorname{tr}(\gamma_{i,j}).$$

**Remark 7.2.1** (Classical OT)**.** *In the scalar case $d = 1$,* (7.2.2) *recovers exactly an optimal entropy-transport problem (see Chapter 1). For isotropic tensors, i.e., all $\mu_i$ and $\nu_j$ are scalar multiples of the identity $\operatorname{Id}_{d \times d}$, the computation also collapses to the scalar case (the $\gamma_{i,j}$ are also isotropic). More generally, if all the $(\mu_i, \nu_j)_{i,j}$ commute, they diagonalize in the same orthogonal basis, and* (7.2.2) *reduces to defining $d$ independent optimal entropy-transport problems along each eigendirection.*

**Remark 7.2.2** (Cost between Dirac masses). *When $\mu = P\delta_x$ and $\nu = Q\delta_x$ are two Dirac masses at the same location $x$ and associated tensors $(P,Q) \in (S_+^d)^2$, one obtains the following "distance" between tensors (assuming $\rho_1 = \rho_2 = 1$ for simplicity)*

$$W_q(P,Q) := \sqrt{C_q(P\delta_x, Q\delta_x)} = \operatorname{tr}\left(P+Q-2\mathfrak{M}(P,Q)\right)^{\frac{1}{2}} \tag{7.2.3}$$

*where $\mathfrak{M}(P,Q) := \exp(\log(P)/2 + \log(Q)/2)$. When $(P,Q)$ commute, we obtain $W_q(P,Q) = \|\sqrt{P} - \sqrt{Q}\|_2$ which is a distance. In the general case, we do not know whether $W_q$ is a distance (basic numerical tests do not exclude this property).*

**Remark 7.2.3** (Quantum transport on curved geometries). *If $(\mu,\nu)$ are defined on a non-Euclidean space $Y = X$, like a smooth manifold, then formulation* (7.2.2) *should be handled with care, since it assumes all the tensors $(\mu_i,\nu_j)_{i,j}$ are defined in some common basis. For smooth manifolds, the simplest workaround is to assume that these tensors are defined with respect to carefully selected orthogonal bases of the tangent planes, so that the field of bases is itself smooth. Unless the manifold is parallelizable, in particular if it has a trivial topology, it is not possible to obtain a globally smooth orthonormal basis; in general, any such field necessarily has a few singular points. In the following, we compute smoothly-varying orthogonal bases of the tangent planes (away from singular points) following the method of Crane et al. [50]. In this setting, the cost is usually chosen to be $c_{i,j} = \operatorname{dist}(x_i,x_j)^\alpha \mathrm{Id}_{d\times d}$ where* dist *is the geodesic distance on $X$.*

**Remark 7.2.4** (Measure lifting). *As mentioned in the previous chapter, an alternative to compute optimal transport between tensor fields would be to rather lift the input measure $\mu$ to a measures $\bar{\mu} := \sum_{i\in I} \delta_{(\mu_i,x_i)}$ defined over the space $X \times S_+^d$ (and similarly for the lifting $\bar{\nu}$ of $\nu$) and then use traditional scalar optimal transport over this lifted space (using a ground cost taking into account both space and matrix variations). This naive approach do not take into account the additive structure of the PSD matrices, and results in very different interpolations. For example, a sum of two nearby Diracs on $X = \mathbb{R}$*

$$\mu = P\delta_0 + Q\delta_s \quad \text{where} \quad P := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad Q := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

*is treated by our method as being very close to $\mathrm{Id}_{2\times 2}\delta_0$ (which is the correct behaviour of a* measure*), whereas it would be lifted to $\bar{\mu} = \delta_{(0,P)} + \delta_{(s,Q)}$ over $\mathbb{R}\times S_2^+$, which is in contrast very far from $\delta_{(0,\mathrm{Id}_{2\times 2})}$. Yet, the discussion in the previous chapter makes it clear that $C_q$ cannot be continuous under weak convergence, a default shared by all static formulations for vector valued measures.*

### 7.2.2. Quantum transport interpolation

Given two input measures $(\mu,\nu)$, we denote by $\gamma$ a solution of (7.2.2) or, in practice, its regularized version (see (7.3.1) below). The coupling $\gamma$ defines a (fuzzy) correspondence between the tensor fields. A typical use of this correspondence is to compute a continuous interpolation between these fields. Section 7.3.4 shows some numerical illustrations of this interpolation.

Note also that Section 7.4 proposes a generalization of this idea to compute an interpolation (barycenter) between more than two input fields.

Mimicking the definition of the optimal transport interpolation (the so-called McCann displacement interpolation; see for instance [142]), we propose to use $\gamma$ to define a path $t \in [0,1] \mapsto \mu_t$ interpolating between $(\mu,\nu)$. For simplicity, we assume the cost has the form $c_{i,j} = \mathrm{dist}(x_i,y_j)^\alpha \mathrm{Id}_{d\times d}$ for some ground metric dist on $X = Y$. We also suppose we can compute efficiently the interpolation between two points $(x_i,y_j) \in X^2$ as

$$x_{i,j}^t \coloneqq \arg\min_{x\in X} (1-t)\,\mathrm{dist}^2(x_i,x) + t\,\mathrm{dist}^2(y_j,x).$$

For instance, over Euclidean spaces, $g_t$ is simply a linear interpolation, and over more general manifold, it is a geodesic segment. We also denote

$$\bar{\mu}_i \coloneqq \mu_i \Big(\sum_j \gamma_{i,j}\Big)^{-1} \quad \text{and} \quad \bar{\nu}_j \coloneqq \nu_j \Big(\sum_i \gamma_{i,j}\Big)^{-1}$$

the adjustment factors which account for the imperfect match of the marginal associated to a solution of (7.3.1); the adjusted coupling is

$$\gamma_{i,j}^t \coloneqq [(1-t)\bar{\mu}_i + t\bar{\nu}_j]\gamma_{i,j}.$$

Finally, the interpolating measure is defined as

$$\forall t \in [0,1], \quad \mu_t \coloneqq \sum_{i,j} \gamma_{i,j}^t \delta_{x_{i,j}^t}. \tag{7.2.4}$$

One easily verifies that this measure indeed interpolates the two input measures, i.e. $(\mu_{t=0},\mu_{t=1}) = (\mu,\nu)$. This formula (7.2.4) generates the interpolation by creating an atom $\gamma_{i,j}^t \delta_{x_{i,j}^t}$ for each coupling entry $\gamma_{i,j}$, and this atom travels between $\mu_i\delta_{x_i}$ (at $t=0$) and $\nu_j\delta_{y_j}$ (at $t=1$).

**Remark 7.2.5** (Computational cost)**.** *We observed numerically that, similarly to the scalar case, the optimal coupling $\gamma$ is sparse, meaning that only of the order of $O(|I|)$ non-zero terms are involved in the interpolating measure* (7.2.4)*. Note that the entropic regularization algorithm detailed in Section 7.3 destroys this exact sparsity, but we found numerically that thresholding to zero the small entries of $\gamma$ generates accurate approximations.*

## 7.3. Quantum Sinkhorn

The convex program (7.2.2) defining quantum OT is computationally challenging because it can be very large scale (problem size is $|I| \times |J|$) for imaging applications, and it involves matrix exponential and logarithm. In this section, leveraging recent advances in computational OT initiated by Cuturi [51], we propose to use a similar entropy regularized strategy (see also section 7.1), but this time with the quantum entropy (7.1.1).

### 7.3.1. Quantum entropic regularization

We define an entropic regularized version of (7.2.2)

$$C_{q,\varepsilon}(\mu,\nu) := \min_{\gamma} \langle \gamma, c\rangle + \rho_1 H_q(\gamma \mathbb{1}_J|\mu) + \rho_2 H_q(\gamma^\top \mathbb{1}_I|\nu) + \varepsilon H_q(\gamma). \tag{7.3.1}$$

Note that when $\varepsilon = 0$, one recovers the original problem (7.2.2). This is a strongly convex program, with a unique solution. The crux of this approach, just as in the scalar case (see Chapter 3), is that its convex dual has a particularly simple structure, which is amenable to a simple alternating maximization strategy.

**Proposition 7.3.1.** *The dual problem associated to* (7.3.1) *reads*

$$C_{q,\varepsilon}(\mu,\nu) = \max_{u,v} -\mathrm{tr}\Big[\rho_1 \sum_i (e^{u_i+\log(\mu_i)} - \mu_i) + \rho_2 \sum_j (e^{v_j+\log(\nu_j)} - \nu_j) + \varepsilon \sum_{i,j} e^{\mathscr{K}(u,v)_{i,j}}\Big], \tag{7.3.2}$$

*where $u = (u_i)_{i\in I}, v = (v_j)_{j\in J}$ are collections of arbitrary symmetric (not necessarily in $S^d_+$) matrices $u_i, v_j \in S^d$ and where we define*

$$\mathscr{K}(u,v)_{i,j} := -\frac{c_{i,j} + \rho_1 u_i + \rho_2 v_j}{\varepsilon}. \tag{7.3.3}$$

*Furthermore, the following primal-dual relationships hold at optimality:*

$$\forall (i,j), \quad \gamma_{i,j} = \exp\big(\mathscr{K}(u,v)_{i,j}\big). \tag{7.3.4}$$

*Proof.* First note that $\gamma = 0$ is always feasible. Applying Fenchel–Rockafellar duality (Appendix A) to (7.3.1) leads to the dual program

$$\max_{u,v} -\varepsilon H_q{}^*(\mathscr{K}_0(u,v)|\xi) - \rho_1 H_q{}^*(u|\mu) - \rho_2 H_q{}^*(v|\nu) \tag{7.3.5}$$

where here $H_q{}^*(\cdot|\mu)$ corresponds to the Legendre transform with respect to the first argument of the quantum relative entropy, $\mathscr{K}_0(u,v)_{i,j} := -\frac{\rho_1 u_i + \rho_2 v_j}{\varepsilon}$ and, for all $i, j$, $\xi_{i,j} := \exp(-c_{i,j}/\varepsilon)$. The Legendre formula

$$H_q{}^*(u|\mu) = \textstyle\sum_i \mathrm{tr}(\exp(u_i + \log(\mu_i)) - \mu_i)$$

shows that the qualification constraint is satisfied and leads to the desired result. □

### 7.3.2. Quantum Sinkhorn algorithm

The following proposition states that the maximization with respect to either $u$ or $v$ leads to two fixed-point equations. These fixed points are conveniently written using the log-sum-exp operator,

$$\mathrm{LSE}_j(K) \coloneqq \Big(\log \sum_j \exp(K_{i,j})\Big)_i, \tag{7.3.6}$$

where the sum on $j$ is replaced by a sum on $i$ for $\mathrm{LSE}_i$.

**Proposition 7.3.2.** *For $v$ fixed (resp. $u$ fixed), the minimizer $u$ (resp. $v$) of* (7.3.2) *satisfies*

$$\forall i, \quad u_i = \mathrm{LSE}_j(\mathscr{K}(u,v))_i - \log(\mu_i), \tag{7.3.7}$$
$$\forall j, \quad v_j = \mathrm{LSE}_i(\mathscr{K}(u,v))_j - \log(\nu_j), \tag{7.3.8}$$

*where $\mathscr{K}(u,v)$ is defined in* (7.3.3)*.*

*Proof.* Writing the first order condition of (7.3.2) with respect to each $u_i$ leads to

$$\rho_1 e^{u_i + \log(\mu_i)} - \rho_1 \sum_j e^{\mathscr{K}(u,v)_{i,j}} = 0$$

which gives the desired expression. A similar expression holds for the first order conditions with respect to $v_j$. □

A simple fixed point algorithm is then obtained by replacing the explicit alternating minimization with respect to $u$ and $v$ with just one step of fixed point iteration (7.3.7) and (7.3.8). To make the resulting fixed point contractive and ensure linear convergence, one introduces parameters $(\tau_1, \tau_2)$, that are akin to the over-relaxation parameter of the acceleration method detailed in Section 3.4.

The quantum Sinkhorn algorithm is detailed in Algorithm 9. It alternates between the updates of $u$ and $v$, using relaxed fixed point iterations associated to (7.3.7) and (7.3.8). We use the following $\tau$-relaxed assignment notation

$$a \overset{\tau}{\leftarrow} b \quad \text{means that} \quad a \leftarrow (1-\tau)a + \tau b. \tag{7.3.9}$$

The algorithm outputs the scaled kernel $\gamma_{i,j} = \exp(K_{i,j})$.

**Remark 7.3.3** (Choice of $\tau_k$)**.** *In the scalar case, i.e. $d = 1$ (and also for isotropic input tensors), when using $\tau_k = \frac{\varepsilon}{\rho_k + \varepsilon}$ for $k = 1,2$, one retrieves exactly the scaling iterations for unbalanced transport as described in Chapter 5, and each update of $u$ (resp. $v$) exactly solves the fixed point* (7.3.7) *(resp.* (7.3.8)*). Moreover, it is simple to check that these iterates are contractant whenever*

$$\tau_k \in \,]0, \tfrac{2\varepsilon}{\varepsilon + \rho_k}[ \quad \text{for } k = 1,2.$$

*and this property has been observed experimentally for higher dimensions $d = 2,3$. Using higher values for $\tau_k$ is akin to the acceleration method of Section 3.4 and generally improves the (linear) convergence rate. Figure 7.3 displays a typical example of convergence, and exemplifies the usefulness of using large values of $\tau_k$, which leads to a speed-up of a factor 6 with respect to the usual Sinkhorn's choice $\tau_k = \frac{\varepsilon}{\varepsilon + \rho_k}$.*

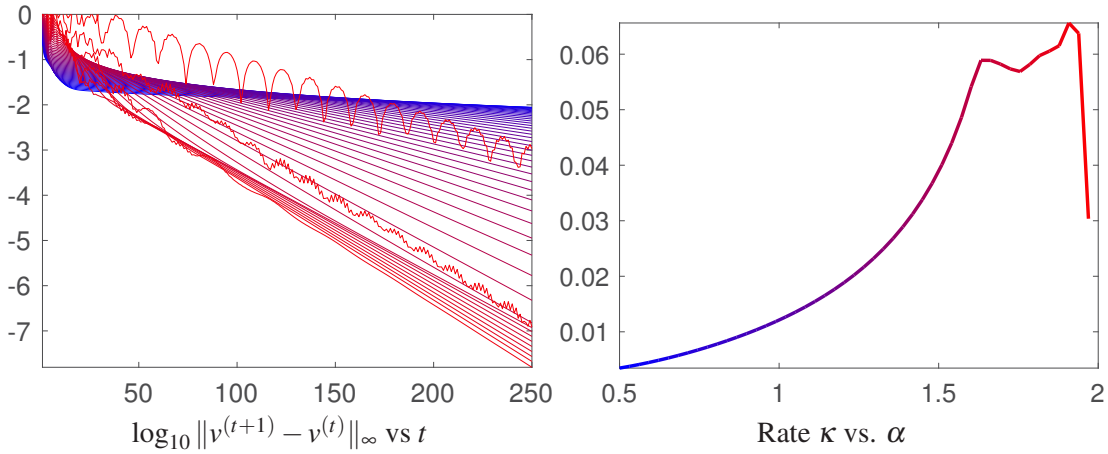



Figure 7.3.: Display of convergence of Algorithm 9 for the example displayed on the first row of Figure 7.1. Denoting $v^{(t)}$ the value of the variable $v$ at iteration $t$, the left plot shows the fixed point residual error for increasing values of $\tau_1 = \tau_2 = \frac{\alpha\varepsilon}{\varepsilon+\rho}$ with $\alpha \in [0.5, 2]$ (blue to red). The algorithm exhibits a linear convergence rate, $\log_{10}\|v^{(t+1)} - v^{(t)}\|_\infty \sim -\kappa t$ for some $\kappa > 0$, and the right plot displays $\kappa$ as a function of $\alpha$ (this plot should be compared to Figure 3.3).

**Remark 7.3.4** (Stability)**.** *In contrast to the usual implementation of Sinkhorn's algorithm, which is numerically unstable for small $\varepsilon$ because it requires to compute $e^{u/\varepsilon}$ and $e^{v/\varepsilon}$, the proposed iterations using the LSE operator are stable. The algorithm can thus be run for arbitrary small $\varepsilon$, although the linear speed of convergence is of course impacted.*

**Remark 7.3.5** (log and exp computations)**.** *A major computational workload of the Q-Sinkhorn Algorithm 9 is the repetitive computation of matrix exp and log. For $d \in \{2,3\}$ it is possible to use closed-form expressions to diagonalize the tensors, so that the overall complexity is comparable with the usual scalar case $d = 1$. While the illustrations we display only consist of low-dimensional settings, high dimensional problems are of interest, typically for machine learning applications. In these cases, one has to resort to iterative procedures, such as rapidly converging squaring schemes [2, 3].*

**Remark 7.3.6** (Computational complexity)**.** *For low-dimensional problems, the Q-Sinkhorn Algorithm 9 scales to grid sizes of roughly 5k points (with machine-precision solutions computed in a few minutes on a standard laptop). For large scale grids, even storing the full coupling $\gamma$ becomes prohibitive. We however observed numerically that, similarly to the usual scalar case, the optimal $\gamma$ solving* (7.3.1) *is highly sparse (up to machine precision for small enough $\varepsilon$). We thus found that the use of the multi-scale refinement strategy introduced in [145] is able to make the Q-Sinkhorn scale to high resolution grids.*

**Remark 7.3.7** (Gurvits' non-commutative Sinkhorn)**.** *Let us insist on the fact that the proposed Q-Sinkhorn Algorithm 9 is unrelated to Gurvits' Sinkhorn algorithm [81]. While Gurvits' iterations compute a coupling between a pair of input tensors, our method rather couples two fields*

**Algorithm 9** Quantum-Sinkhorn iterations to compute the optimal coupling $\gamma$ of the regularized transport problem (7.3.1). The operator $\mathscr{K}$ is defined in (7.3.3).

1. $\forall k = 1,2, \quad \tau_k \in ]0, \frac{2\varepsilon}{\varepsilon+\rho_k}[$
2. $\forall (i,j) \in I \times J, \quad (u_i, v_j) \leftarrow (0_{d\times d}, 0_{d\times d})$
3. for $s = 1,2,3,\ldots,$ do
4. a) $K \leftarrow \mathscr{K}(u,v)$
   b) $\forall i \in I, \quad u_i \overset{\tau_1}{\leftarrow} \mathrm{LSE}_j(K_{i,j}) - \log(\mu_i)$
   c) $K \leftarrow \mathscr{K}(u,v)$
   d) $\forall j \in J, \quad v_j \overset{\tau_2}{\leftarrow} \mathrm{LSE}_i(K_{i,j}) - \log(\nu_j)$
5. return $(\gamma_{i,j} = \exp(K_{i,j}))_{i,j}$

*of tensors (viewed as tensor-valued measures). Our usage of the wording "quantum" refers to the notion of quantum entropy* (7.1.1) *and is not inspired by quantum physics.*

### 7.3.3. Trace constrained extension

The quantum optimal transport problem (7.2.2) does not impose that the marginals of the coupling $\gamma$ match exactly the inputs $(\mu,\nu)$. It is only in the limit $(\rho_1,\rho_2) \to (+\infty,+\infty)$ that an exact match is obtained, but as explained in Section 7.2.1, this might leads to an empty constraint set.

To address this potential issue, we propose to rather only impose the *trace* of the marginals to match the trace of the input measures, in order to guarantee conservation of mass (as measured by the trace of the tensors). We thus propose to solve the entropy regularized problem (7.3.1) with the extra constraint

$$\forall i \in I, \quad \sum_j \mathrm{tr}(\gamma_{i,j}) = \mathrm{tr}(\mu_i) \quad \text{and} \quad \forall j \in J, \quad \sum_i \mathrm{tr}(\gamma_{i,j}) = \mathrm{tr}(\nu_j).$$

The feasibility conditions for this problem are more strict: one should first require $\sum \mathrm{tr}(\mu_i) = \sum \mathrm{tr}(\nu_i)$ but even this is not sufficient since a finite relative entropy implies $\ker \mu_i \subset \ker \sum_j \gamma_{i,j}$ (and similarly for the other marginal).

The two extra trace constraints introduce two dual Lagrange multipliers $(\alpha,\beta) \in \mathbb{R}^I \times \mathbb{R}^J$ and the optimal coupling relation (7.3.4) is replaced by

$$\forall (i,j), \quad \gamma_{i,j} = \exp\big(\mathscr{K}(u,v,\alpha,\beta)_{i,j}\big) \tag{7.3.10}$$

$$\text{where} \quad \mathscr{K}(u,v,\alpha,\beta)_{i,j} := -\frac{c_{i,j} + \rho_1 u_i + \rho_2 v_j + \alpha_i + \beta_j}{\varepsilon}.$$

Q-Sinkhorn algorithm (Algorithm 9) is extended to handle these two extra variables $(\alpha,\beta)$ by simply adding two steps to update these variables

$$\begin{aligned}\forall i\in I,\quad \alpha_i \leftarrow \alpha_i + \varepsilon\,\mathrm{LSTE}_j(K)_i \quad \text{where} \quad K \coloneqq \mathscr{K}(u,v,\alpha,\beta),\\ \forall j\in J,\quad \beta_j \leftarrow \beta_j + \varepsilon\,\mathrm{LSTE}_i(K)_j \quad \text{where} \quad K \coloneqq \mathscr{K}(u,v,\alpha,\beta).\end{aligned}$$

where we introduced the log-sum-trace-exp operator

$$\mathrm{LSTE}_j(K) \coloneqq \Big(\log\sum_j \mathrm{tr}(\exp(K_{i,j}))\Big)_i$$

(and similarly for $\mathrm{LSTE}_i$). Note that in this expression, the exp is matrix-valued, whereas the log is real-valued.

### 7.3.4. Numerical illustrations

Figures 7.1, 7.4 and 7.5 illustrate on synthetic examples of input tensor fields $(\mu,\nu)$ the Q-OT interpolation method. We recall that it is obtained in two steps:

(i) One first computes the optimal $\gamma$ solving (7.3.1) using Q-Sinkhorn iterations (Algorithm 9).

(ii) Then, for $t\in[0,1]$, one computes $\mu_t$ using this optimal $\gamma$ with formula (7.2.4).

Figure 7.4 shows examples of interpolations on a 1-D domain $X=Y=[0,1]$ with tensors of dimension $d=2$ and $d=3$, and a ground cost $c_{i,j}=|x_i-y_j|^2\mathrm{Id}_{d\times d}$. It compares the optimal transport interpolation, which achieves a "mass displacement," to the usual linear interpolation $(1-t)\mu+t\nu$, which only performs a pointwise interpolation of the tensors.

Figure 7.5 shows the effect of taking into account the anisotropy of tensors into the definition of optimal transport. In the case of isotropic tensors (see Remark 7.2.1), the method reduces to the (unbalanced) scalar optimal transport, and in 1-D it corresponds to the monotone rearrangement [142]. In contrast, the quantum optimal transport of anisotropic tensors is forced to reverse the ordering of the transport map in order for tensors with similar orientations to be matched together. This example illustrates that the behaviour of our tensor interpolation is radically different from only applying classical scalar-valued optimal transport to the trace of the tensor (which would result in the same coupling as the one obtained with isotropic tensors, Figure 7.5, left).

Figure 7.1 shows larger scale examples. The first row corresponds to $X=Y=[0,1]^2$ and $d=2$, with cost $c_{i,j}=\|x_i-y_j\|^2\mathrm{Id}_{2\times 2}$, which is a typical setup for image processing. The second row corresponds to $X=Y$ being a triangulated mesh of a surface, with a cost proportional to the squared geodesic distance $c_{i,j}=\mathrm{dist}(x_i,y_j)^2\mathrm{Id}_{2\times 2}$.

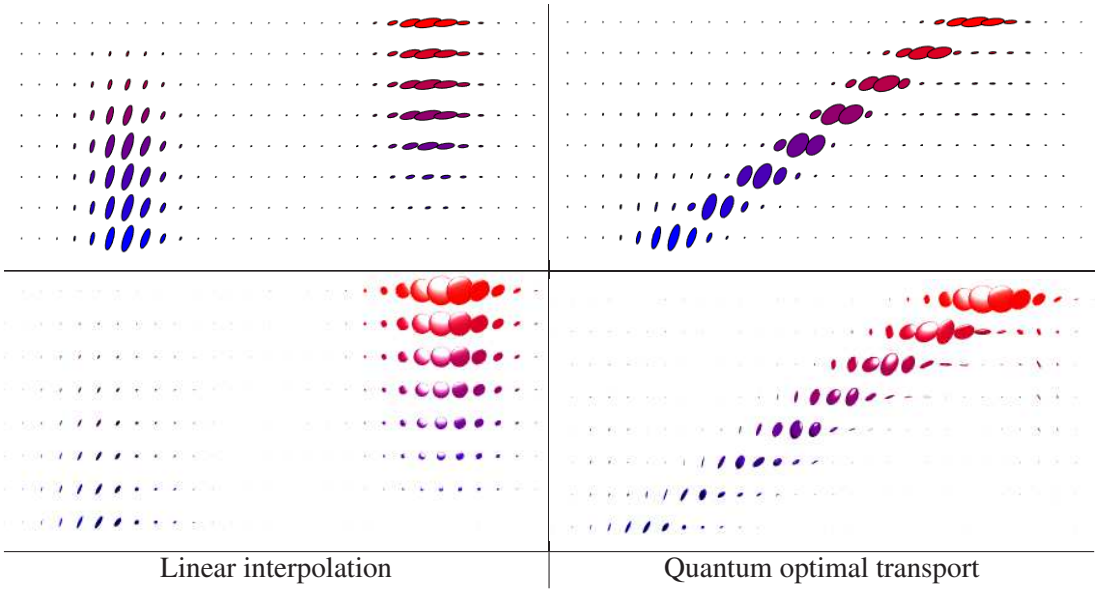



Figure 7.4.: Comparison of linear and quantum-optimal transport interpolation (using formula (7.2.4)). Each row shows a tensor field $\mu_t$ (top $d = 2$, bottom $d = 3$) along a linear segment from $t = 0$ to $t = 1$ ($t$ axis is vertical).

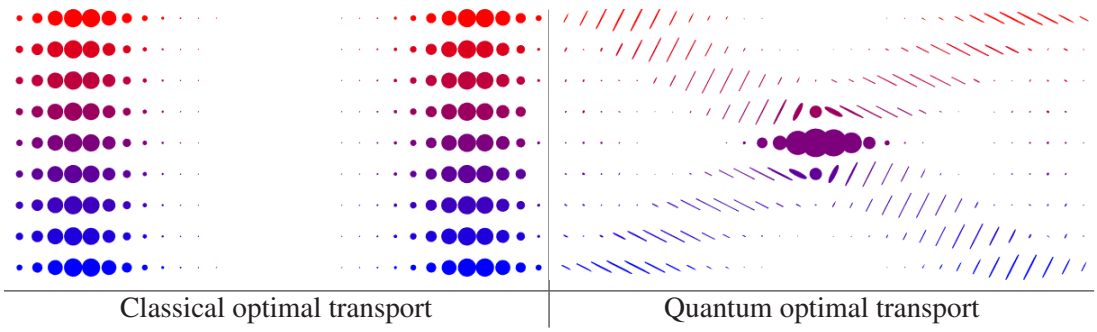



Figure 7.5.: Comparison of classical optimal transport (i.e. between isotropic tensors) and quantum optimal transport (between anisotropic tensors) interpolation (using formula (7.2.4)), using the same display as Figure 7.4.

## 7.4. Quantum barycenters

Following Agueh and Carlier [1], we now propose a generalization of the quantum optimal transport problem, where, instead of coupling only two input measures, one tries to couple an arbitrary set of inputs, and compute their Fréchet mean.

### 7.4.1. Barycenter optimization problem

Given some input measures $(\mu^\ell)_\ell$, the quantum barycenter problem reads

$$\min_\nu \sum_\ell w_\ell C_{q,\varepsilon}(\mu^\ell, \nu), \tag{7.4.1}$$

where $(w_\ell)_\ell$ is a set of positive weights normalized so that $\sum_\ell w_\ell = 1$. In the following, for simplicity, we set

$$\rho_1 = \rho \quad \text{and} \quad \rho_2 = +\infty$$

in the definition (7.2.2) of $W_\varepsilon$. Note that the choice $\rho_2 = +\infty$ corresponds to imposing the exact marginal constraint $\gamma^\top \mathbb{1}_J = \nu$.

**Remark 7.4.1** (Barycenters between single Dirac masses)**.** *If all the input measures are concentrated on single Diracs $\mu^\ell = P_\ell \delta_{x_\ell}$, then the single Dirac barycenter (unregularized, i.e., $\varepsilon = 0$) for a cost $\mathrm{dist}(x,y)^\alpha \mathrm{Id}_{d\times d}$ is $P\delta_{x^\star}$ where $x^\star \in X$ is the usual barycenter for the distance* dist*, solving*

$$x^\star \in \operatorname*{argmin}_x \mathscr{E}(x) = \sum_\ell w_\ell \,\mathrm{dist}^\alpha(x_\ell, x)$$

*and the barycentric matrix is*

$$P = e^{-\frac{\mathscr{E}(x^\star)}{\rho}} \exp\Big(\sum_\ell w_\ell \log(P_\ell)\Big). \tag{7.4.2}$$

*Figure 7.6 illustrates the effect of a pointwise interpolation (i.e. at the same location $x_\ell$ for all $\ell$) between tensors.*

Problem (7.4.1) is convex, and similarly to (7.3.2), it can be rewritten in dual form.

**Proposition 7.4.2.** *The optimal $\nu$ solving* (7.4.1) *is the solution of*

$$\max_{(u^\ell, v^\ell)} \min_\nu -\sum_\ell w_\ell \,\mathrm{tr}\Big[\rho \sum_i e^{u_i^\ell + \log(\mu_i^\ell)} \quad + \quad \sum_j \nu_j v_j^\ell \quad + \quad \varepsilon \sum_{i,j} e^{\mathscr{K}(u^\ell, v^\ell)_{i,j}}\Big], \tag{7.4.3}$$

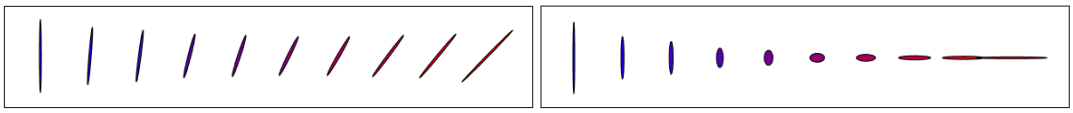


Figure 7.6.: Two examples of pointwise (without transportation) interpolations, using formula (7.4.2). Here $P_1$ and $P_2$ are represented using the blue/red ellipses on the left/right, and weights are $(w_1, w_2) = (1-t, t)$ for $t \in [0,1]$ from left to right.

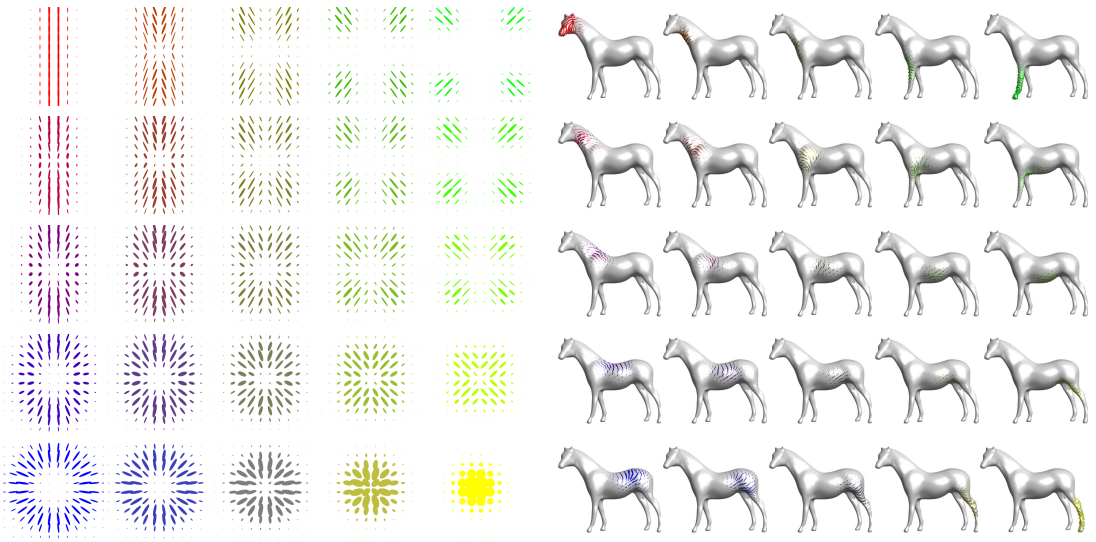


Figure 7.7.: $5 \times 5$ barycenters of four input measures (displayed in the four corners). The weighs $w \in \mathbb{R}^4$ correspond to bilinear interpolation weights (7.4.8) inside the square.

*where here we define* $\mathscr{K}$ *as*

$$\mathscr{K}(u,v)_{i,j} := -\frac{c_{i,j} + \rho u_i + v_j}{\varepsilon}. \tag{7.4.4}$$

### 7.4.2. Quantum barycenter Sinkhorn

Similarly to Proposition 7.3.2, the dual solutions of (7.4.3) satisfy a set of coupled fixed point equations.

**Proposition 7.4.3.** *Optimal* $(u^\ell, v^\ell)_\ell$ *for* (7.4.3) *satisfy*

$$\forall (i,\ell), \quad \mathrm{LSE}_j(\mathscr{K}(u^\ell, v^\ell)_{i,j}) - \log(\mu_i^\ell), = u_i^\ell \tag{7.4.5}$$

$$\forall (j,\ell), \quad \mathrm{LSE}_i(\mathscr{K}(u^\ell, v^\ell)_{i,j}) = \log(\nu_j) \tag{7.4.6}$$

$$\sum_\ell w_\ell v^\ell = 0. \tag{7.4.7}$$

*Proof.* The proof of (7.4.5) and (7.4.6) is the same as the one of Proposition 7.3.2. Minimization of (7.4.3) on $\nu$ leads to (7.4.7). □

The extension of the Q-Sinkhorn algorithm to solve the barycenter problem (10) is detailed in Algorithm 10. It alternates between the updates of $u$, $\nu$ and $v$, using the relaxed version of the fixed point equations (7.4.5), (7.4.6) and (7.4.7). The notation $\overset{\tau}{\leftarrow}$ refers to a relaxed assignment as defined in (7.3.9).

**Remark 7.4.4** (Choice of $\tau$)**.** *Remark 7.3.3 also applies for this Sinkhorn-like scheme, and setting* $(\tau_1, \tau_2) = (\frac{\varepsilon}{\rho+\varepsilon}, 1)$ *leads, in the scalar case* $d = 1$*, to the algorithm for barycenters in*

**Algorithm 10** Quantum-Barycenter iterations to compute the barycenter measure $\nu$ solving (7.4.1). The operator $\mathscr{K}$ is defined in (7.4.4).

1. Choose $\tau_1 \in ]0, \frac{2\varepsilon}{\varepsilon+\rho}[$, $\tau_2 \in ]0,2[$.
2. $\forall (i,j) \in I \times J, \quad (u_i, v_j) \leftarrow (0_{d\times d}, 0_{d\times d})$
3. for $s = 1,2,3,\ldots$ do
   a) for $\ell = 1,\ldots,L$ do
      i. $K^\ell \leftarrow \mathscr{K}(u^\ell, v^\ell)$
      ii. $\forall i \in I, \quad u_i^\ell \overset{\tau_1}{\leftarrow} \mathrm{LSE}_j(K_{i,j}^\ell) - \log(\mu_i^\ell)$,
      iii. $K^\ell \leftarrow \mathscr{K}(u^\ell, v^\ell)$
   b) $\forall j \in J, \quad \log(\nu_j) \leftarrow \sum_\ell w_\ell (\mathrm{LSE}_i(K_{i,j}^\ell) + v_j^\ell/\varepsilon)$.
   c) for $\ell = 1,\ldots,L$
      i. $\forall j \in J, \quad v_j^\ell \overset{\tau_2}{\leftarrow} \varepsilon \mathrm{LSE}_i(K_{i,j}^\ell) + v_j^\ell - \varepsilon \log(\nu_j)$.
4. return $\nu$

*Chapter 5. We found experimentally that this choice leads to contracting (and hence linearly converging) iterations, and that higher values of $\tau$ usually accelerate the convergence rate (as expected from Section 3.4).*

**Remark 7.4.5** (Scalar and isotropic cases)**.** *Note that in the scalar case $d = 1$ and for isotropic input tensors (multiples of the identity), one retrieves the provably convergent unbalanced barycenter algorithm of Chapter 5.*

### 7.4.3. Numerical illustrations

Figure 7.7 shows examples of barycenters $\nu$ solving (7.4.1) between four input measures $(\mu^\ell)_{\ell=1}^4$. The horizontal/vertical axes of the figures are indexed by $(t_1,t_2) \in [0,1]^2$ (on a $5\times 5$ grid) and parameterize the weights $(w_\ell)_{\ell=1}^4$ appearing in (7.4.1) as

$$(w_1,w_2,w_3,w_4) \coloneqq ((1-t_1)(1-t_2),(1-t_1)t_2,t_1(1-t_2),t_1,t_2). \tag{7.4.8}$$

The left part of Figure 7.7 corresponds to measures on $X = Y = [0,1]^2$ with $d = 2$ and ground cost $c_{i,j} = \|x_i - x_j\|^2 \mathrm{Id}_{2\times 2}$. The right part of Figure 7.7 corresponds to measures on $X = Y$ being a surface mesh with $d = 2$ (the tensors are defined on the tangent planes) and a ground cost is $c_{i,j} = \mathrm{dist}(x_i,x_j)^2 \mathrm{Id}_{2\times 2}$ where $d_X$ is the geodesic distance on the mesh.

# Conclusion

There was a burning need in applications of optimal transport to extend the theory to deal with unnormalized measures. The purpose of this thesis was to study these extensions in a systematic way, both from the theoretical and the numerical point of views.

In the first part, we have developed a consistent framework for unbalanced optimal transport. Various formulations are possible, based on different directions of generalization of optimal transport. We have explored them and shown how to switch from one to another. We have also studied generalizations of the $p$-Wasserstein distances to the space of nonnegative measures, that interpolate between pure growth and pure transport distances. Our approach allows to define new models, such as the $\widehat{W}_2$ distance, and also recovers known extensions such as optimal partial transport.

In the second part, we have developed numerical methods to deal with these extensions. We have shown that the approach based on entropy regularization and Sinkhorn's algorithm extends naturally to a class of *scaling algorithms* in this new setting. We have studied their convergence properties with various techniques (convex analysis, local analysis, Thompson metric). We also proposed an extension of the existing numerical methods to deal with the dynamic formulations.

In the third part, we have dealt with more specific applications. We have illustrated the behavior of unbalanced optimal transport and shown its relevance in some tasks such as shapes processing or color transfer. We have studied in depth an evolution PDE for tumor growth which is shown to characterize some gradient flows for the metric $\widehat{W}_2$. Finally, we have seen that scaling algorithms can be extended to compute optimal transport-like interpolations and barycenters of measures that take PSD matrix values.

These contributions open the way for a better handling of mass variations in applications of optimal transport. They also give insight for other extensions such as the case of vector valued measures. This is an interesting direction of future research with tools already developed for signal processing [71, 156] or PDEs [170]. In this thesis, we have considered the case of SDP matrices with algorithmic efficiency in mind. A salient open question is how to combine the geometry of the domain and that of the codomain of these measures in an optimal transport-like metric with reasonable properties such as existence of geodesics and continuity under weak convergence.

# Appendix A.

# Convex Functions

## Main concepts

**Dual descriptions** The description of convex sets as intersections of half-spaces is the principle underlying the theory of convex duality. This relationship is clear if $V$ is a topological vector space with a Hausdorff and locally convex topology, since then

- an arbitrary intersection of closed half-spaces is a closed convex set;
- any closed convex set is the intersection of the closed half-spaces containing it.

The first statement is a direct consequence of the stability of convexity and closedness by arbitrary intersections, while the second statement is a corollary of the Hahn-Banach theorem, which uses the strong axiom of choice at this level of generality (the material of this appendix is adapted from [66, 137] that should be consulted for more details).

With appropriate definitions, a similar relationship can be established between convex functions and families of affine forms. A function $f : V \to \mathbb{R} \cup \{\infty\}$ is said

- *convex* if its epigraph $\mathrm{epi} f := \{(u,\alpha) \in V \times \overline{\mathbb{R}}\,;\, f(u) \leq \alpha\}$ is a convex set, or equivalently if for all $(u_0,u_1) \in V^2$ and $\theta \in ]0,1[$, it holds

$$f(\theta u_0 + (1-\theta)u_1) \leq \theta f(u_0) + (1-\theta) f(u_1),$$

- *proper* if $\mathrm{epi} f$ is not empty or equivalently if $f$ is not identically $+\infty$,
- *lower-semicontinuous* if $\mathrm{epi} f$ is closed or equivalently (when the topology is first countable), if $f(u) \leq \liminf_{u_n \to u} f(u_n)$ for all $u \in V$.

**Definition A.0.1.** *The set of functions $f : V \to \overline{\mathbb{R}}$ which are convex, proper and lower-semicontinuous is denoted by $\Gamma_0(V)$ in this appendix.*

The next result is at the basis of convex duality (see [66, prop. 3.1] with other definitions and pathological cases removed).

**Proposition A.0.2.** *The set $\Gamma_0(V)$ is precisely the set of proper functions which are pointwise supremum of a nonempty family of continuous affine functions.*

**Conjugacy** If follows that the knowledge of the continuous affine minorants of $f \in \Gamma_0(V)$

$$\{(u^*, \alpha) \in V^* \times \mathbb{R} \; ; \; \langle u, u^* \rangle - \alpha \leq f(u), \; \forall u \in V\} \tag{A.0.1}$$

where $V^*$ is the *topological dual* of $V$—the set of continuous linear forms on $V$—is enough to characterize $f$. One can even restrict ourselves to the maximal affine minorant, that is only taking the supremum over all $\alpha$ satisfying the inequality in (A.0.1). This lead us to the definition of the *conjugate function* of $f$.

**Definition A.0.3.** *For a function $f : V \to \mathbb{R} \cup \{\infty\}$, its conjugate is the function $f^* : V^* \to \mathbb{R} \cup \{\infty\}$ defined for all $u^* \in V^*$ as*

$$f^*(u^*) \coloneqq \sup_{u \in V} \{\langle u, u^* \rangle - f(u)\}.$$

The dissymmetry between the sets $V$ and $V^*$ is only apparent and can be removed with suitable choices topologies. Two Hausdorff, locally convex, topological vector spaces are said *topologically paired* if all continuous linear functionals on one space can be identified with the elements of the other, and vice-versa[1]. This pairing comes with a bilinear form $\langle \cdot, \cdot \rangle : V \times V^* \to \mathbb{R}$. In this framework, it follows almost by construction and by Proposition A.0.2 that

- if $f$ is proper then $f^* \in \Gamma_0(V^*)$;
- if $f \in \Gamma_0(V)$ then $(f^*)^* = f$: there is a bijection between $\Gamma_0(V)$ and $\Gamma_0(V^*)$;

**Subdifferentiability** Finally, one can wonder what are the *exact* affine minorants of $f$ at some point $u \in V$. These are the couples $(u^*, \alpha)$ that reach equality in (A.0.1) or, in geometrical terms, the hyperplans that touch the epigraph of $f$ at $u$.

**Definition A.0.4.** *The slope of an affine minorant which is exact at a point $u \in V$ is called a subgradient of $f$ at $u$. The set of sugradients is called the subdifferential of $f$ at $u$ and is denoted $\partial f(u)$.*

**Proposition A.0.5.** *If $f : V \to \mathbb{R} \cup \{\infty\}$ and $u \in V$, one has $u^* \in \partial f(u)$ if and only if*

- *$\langle v - u, u^* \rangle + f(u) \leq f(v)$ for all $v \in V$, or equivalently*
- *$f(u) + f^*(u^*) = \langle u, u^* \rangle$.*

It follows from the last characterization and from the definition of $f^*$ that $\partial f(u) = \{u^* \in V^* \; ; \; f^*(u^*) \leq \langle u, u^* \rangle - f(u)\}$ so $\partial f(u)$ is convex and closed. Also, if $f \in \Gamma_0$ one deduces $u^* \in \partial f(u) \Leftrightarrow u \in \partial f^*(u^*)$. As for the first characterization, it implies a fundamental result in convex variational analysis:

$$f \text{ reaches a global minimum at } u \in V \quad \Leftrightarrow \quad 0 \in \partial f(u).$$

[1] A Euclidean space is paired with itself. A Banach space (equipped with the strong or weak topology) and its topological dual equipped with the weak* topology is another example.

## Examples of duality

**Adjoint of a linear map** While the adjoint of a linear map may be defined under rather weak assumptions, only the simple definition for linear operators which are continuous and of full domain is used in this thesis.

**Definition A.0.6.** *Let $(U,U^*)$ and $(V,V^*)$ be two pairs of topologically paired spaces and $A : U \to V$ be a continuous linear operator. For all $v^* \in V^*$, the application $u \to \langle Au, v^* \rangle$ is a continuous linear form on $U$ and thus admits a representer in $U^*$ that we denote $A^*v^*$. This uniquely defines a linear operator $A^* : V^* \to U^*$ called the* adjoint *of $A$.*

For instance, the continuity equation (Definition 1.1.1) and its variants are affine constraints w.r.t. a linear map that is only defined implicitly through its adjoint map. This fact is convenient when deriving duality results.

**Fenchel-Rockafellar** A duality result that is intensively used in this thesis is the Fenchel-Rockafellar theorem. See [137, Thm. 19-20] for a rigorous proofs and a more general statement. By convention, when an infimum is attained, we replace the sign "inf" by "min".

**Theorem A.0.7** (Fenchel-Rockafellar)**.** *Let $(U,U^*)$ and $(V,V^*)$ be two couples of topologically paired spaces. Let $f \in \Gamma_0(U)$, $g \in \Gamma_0(V)$ and $A : U \to V$ be a continuous linear operator of adjoint $A^* : V^* \to U^*$. The so-called* qualification constraint *is the property: there exists $x \in \operatorname{dom} f$ such that $g$ is continuous at $Ax$. If the qualification constraint holds, then*

$$\sup_{u \in U} -f(-u) - g(Au) = \min_{v^* \in V^*} f^*(A^*v^*) + g^*(v^*)$$

*and the* min *is attained. Moreover, if the minimum is finite, $(u, v^*) \in U \times V^*$ is a couple of optimizers if and only if $Au \in \partial g^*(v^*)$ and $A^*v^* \in \partial f(-u)$.*

*Elements of proof.* The following is a mnemonic way to recover the result rather than the proof since the difficult part of min/sup inversion is skipped. For any $y \in E$, it holds

$$\begin{aligned}
\sup_{u \in U} -f(-u) - g(Au) &= \sup_{u \in U} \{-f(-u) - \sup_{v^* \in V^*} \langle Au, v^* \rangle - g^*(v^*)\} \\
&= \sup_{u \in U} \inf_{v^* \in V^*} \{g^*(v^*) + \langle -u, A^*v^* \rangle - f(-u)\} \\
&= \inf_{v^* \in V^*} \{g^*(v^*) + \sup_{u \in U} \langle u, A^*v^* \rangle - f(u)\} \\
&= \inf_{v^* \in V^*} \{g^*(v^*) + f^*(A^*v^*)\}.
\end{aligned}$$

The subdifferential inclusions can be recovered by looking at the optimality conditions in the second line. □

## Sublinear and perspective functions

### Sublinear functions

**Definition A.0.8.** *Let $f : V \to \mathbb{R} \cup \{\infty\}$ be a function on a linear space. It is said :*

- subadditive *if for all* $(x,y) \in V^2$*, it holds* $f(x+y) \leq f(x)+f(y)$.
- positively $p$-homogeneous *if for all* $x \in V$ *and* $\lambda \in [0,+\infty[$*, it holds*

$$f(\lambda x) = \lambda^p f(x).$$

In general we will just say "$p$-homogeneous", or even just "homogeneous" if $p = 1$. Let us introduce the class of sublinear functions, that are natural in variational problems involving measures.

**Definition/Proposition A.0.9.** *A function* $f : V \to \mathbb{R} \cup \{\infty\}$ *that satisfies any two of the following properties satisfies all three and is called a* sublinear *function.*

*(i)* $f$ *is subadditive and* $f(0) \leq 0$*;*

*(ii)* $f$ *is convex;*

*(iii)* $f$ *is positively 1-homogeneous.*

*Proof.*
- (i)&(ii)⇒(iii). For $\lambda \in ]0,1[$ and $x \in V$, convexity and $f(0) \leq 0$ imply $f(\lambda x) \leq \lambda f(x)$ and $f((1-\lambda)x) \leq (1-\lambda)f(x)$. The subadditivity in turns gives

$$f(x) \leq f(\lambda x) + f((1-\lambda)x) \leq f(x)$$

so all inequalities are equalities and $f(\lambda x) = \lambda f(x)$. If $\lambda > 1$, simply exchange the role of $x$ and $\lambda x$ above and pose $\tilde{\lambda} = 1/\lambda$. Finally, for $\lambda = 0$, remark that for a subadditive function, $f(0) \leq 2f(0)$ so $f(0) = 0$. Note that if $f(0) \leq 0$ was not specified in (i) then a counter-example in a normed space is $f(x) = 1 + \|x\|$ which is subadditive, convex, but not homogeneous.

- (i)&(iii)⇒(ii). Let $(x,y) \in V^2$ and $\theta \in ]0,1[$, it holds

$$f(\theta x + (1-\theta)y) \leq f(\theta x) + f((1-\theta)y) = \theta f(x) + (1-\theta)f(y)\,.$$

Note that the property $f(0) \leq 0$ is redundant here.

- (ii)&(iii)⇒(i). For $(x,y) \in V^2$ one has

$$f(x+y) = 2f(x/2+y/2) \leq 2(f(x)/2 + f(y)/2) = f(x)+f(y)\,. \qquad \square$$

**Proposition A.0.10.** *If* $f : V \to \mathbb{R} \cup \{\infty\}$ *is sublinear then* $f^* = \iota_{\mathscr{A}}$ *for some convex, closed, nonempty set* $\mathscr{A}$*. Reciprocally, if* $\mathscr{A}$ *is a nonempty set then* $\iota^*_{\mathscr{A}}$ *is sublinear.*

*Proof.* If $f$ is sublinear, then $f^*(v^*) = \sup_{v \in V}\{\langle v, v^*\rangle - f(v)\}$ is nonnegative because $f(0) = 0$. Moreover, if $f^*(v^*) > 0$ then there exists $(v,\alpha) \in V \times ]0,\infty[$ such that $\langle v, v^*\rangle - f(v) \geq \alpha$. It follows

$$f^*(v^*) \geq \sup_{\lambda \geq 0}\{\langle \lambda v, v^*\rangle - f(\lambda v)\} \geq \sup_{\lambda \geq 0} \lambda\alpha = \infty.$$

Thus $f^*$ is the indicator of a set, whose properties can be deduced from the fact that $f^* \in \Gamma_0(V^*)$. For the second claim, one only has to check that $\iota^*_{\mathscr{A}}$ is positively 1-homogeneous, which is easy. $\square$

### Perspective functions

Due to their homogeneity properties, sublinear functions are entirely determined by specifying their values on a set that intersect all rays (half-lines of origin 0). If this set is taken as $\{1\} \times V \subset \mathbb{R} \times V$ and the values specified by a function $f : V \to \mathbb{R} \cup \{\infty\}$, the associated sublinear function is called the *perspective* of $f$. Detailed analytical properties of these functions can be found in [46].

**Definition A.0.11.** *Let $f : V \to \mathbb{R} \cup \{\infty\}$ be function on a linear space. Its perspective function $\psi : \mathbb{R} \times V \to \mathbb{R} \cup \{\infty\}$ is defined as*

$$\psi_f(t,x) := \begin{cases} t\, f\left(\frac{x}{t}\right) & \text{if } t > 0, \\ f'_\infty(x) & \text{if } t = 0, \\ \infty & \text{if } t < 0. \end{cases}$$

In the previous definition, the values of $\psi$ for $t = 0$ are defined in a way to preserve lower-semicontinuity.

**Definition A.0.12.** *The function $f'_\infty := \psi_f(0,x)$ is called the* recession *or* horizon *function of $f$. Taking any $z \in \operatorname{dom} f$, it is defined for $x \in V$ as*

$$f'_\infty(x) := \iota^*_{\operatorname{dom} f^*}(x).$$

Alternative expressions for the recession function are, if $f \in \Gamma_0(V)$:

$$f'_\infty(x) = \sup\{f(x+y) - f(y)\ ;\ y \in \operatorname{dom} f\} = \lim_{\lambda \to \infty} \left(f(\lambda x + y) - f(y)\right)/\lambda.$$

**Proposition A.0.13.** *One has $\psi^*_f = \iota\{(s,y) \in \mathbb{R} \times E^* : s + f^*(y) \leq 0\}$. Also, $\psi_f \in \Gamma_0(\mathbb{R} \times V)$ if and only if $f \in \Gamma_0(V)$. By noting $u = x/t$, if $V = \mathbb{R}^n$ and $f$ is differentiable, then it holds for $t > 0$:*

$$\nabla \psi_f(t,x) = (f(u) - u \cdot \nabla f(u), \nabla f(u))$$

*and with $H_f$ the Hessian of $f$ (if $f$ is twice differentiable):*

$$H_{\psi_f}(t,x) = \frac{1}{t} \begin{pmatrix} u \cdot H_f(u) u & -(H_f(u)u)^* \\ -H_f(u)u & H_f(u) \end{pmatrix}$$

*Proof.* If $f \in \Gamma_0(V)$, the expression suggested for $\psi^*_f$ defines a function in $\Gamma_0(\mathbb{R} \times V^*)$ and one find by direct computation that $(\psi^*_f)^* = \psi_f$. The differential formula also come from direct computations. □

## Duality for sublinear functions of measures

The following is a rephrasing of [136, Thm 6], with simplified assumptions thanks to [22, Lem. A.2]. We use the notation introduced the section *Notation* for the sublinear function of a measure.

**Theorem A.0.14.** *Let $X$ be compact metric space and $f : X \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ a l.s.c. function such that for all $x \in X$, $f_x(\cdot)$ is sublinear and proper. Then $I_f : \mathscr{M}(X;\mathbb{R}^n) \to \mathbb{R} \cup \{\infty\}$ and $I_{f^*} : \mathscr{C}(X;\mathbb{R}^n) \to \mathbb{R} \cup \{\infty\}$ defined as*

$$I_f(\mu) \coloneqq \int_X f_x(\mu) \quad \textit{and} \quad I_{f^*}(\phi) \coloneqq \begin{cases} 0 & \textit{if } \phi(x) \in \operatorname{dom} f(x,\cdot)^*,\ \forall x \in \Omega, \\ \infty & \textit{otherwise.} \end{cases}$$

*form a pair of convex, proper, l.s.c. conjugates functions, where the topology considered are the strong topology for $\mathscr{C}(X;\mathbb{R}^n)$ and the weak topology for $\mathscr{M}(X;\mathbb{R}^n)$.*

# Bibliography


[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Awad H Al-Mohy and Nicholas J Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM J. Sci. Comput.*, 31(3):970–989, 2009.

[3] Awad H Al-Mohy and Nicholas J Higham. Improved inverse scaling and squaring algorithms for the matrix logarithm. *SIAM J. Sci. Comput.*, 34(4):C153–C169, 2012.

[4] Pierre Alliez, David Cohen-Steiner, Olivier Devillers, Bruno Lévy, and Mathieu Desbrun. Anisotropic polygonal remeshing. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 485–493. ACM, 2003.

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

[6] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[7] Alfred Auslender. Optimisation: méthodes numériques. 1976.

[8] Heinz H Bauschke and Jonathan M Borwein. Joint and separate convexity of the Bregman distance. *Studies in Computational Mathematics*, 8:23–36, 2001.

[9] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[10] Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

[11] Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

[12] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[13] Jean-David Benamou and Guillaume Carlier. Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1):1–26, 2015.

[14] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[15] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampere equation. *Journal of Computational Physics*, 260:107–126, 2014.

[16] Martin Benning, Luca Calatroni, Bertram Düring, and Carola-Bibiane Schönlieb. A primal-dual approach for a total variation Wasserstein flow. In *Geometric Science of Information*, pages 413–421. Springer, 2013.

[17] Dimitri P Bertsekas and David A Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96, 1989.

[18] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[19] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *ACM Transactions on Graphics (TOG)*, volume 30, page 158. ACM, 2011.

[20] Jonathan M Borwein and Adrian S Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

[21] Jonathan M Borwein, Adrian Stephen Lewis, and Roger D Nussbaum. Entropy minimization, DAD problems, and doubly stochastic kernels. *Journal of Functional Analysis*, 123(2):264–307, 1994.

[22] Guy Bouchitté and Michel Valadier. Integral representation of convex functionals on a space of measures. *Journal of functional analysis*, 80(2):398–420, 1988.

[23] Sébastien Bougleux, Gabriel Peyré, and Laurent D Cohen. Image compression with anisotropic geodesic triangulations. In *Proc. of ICCV'09*, pages 2343–2348, 2009.

[24] Andrea Braides. *Gamma-convergence for Beginners*, volume 22. Clarendon Press, 2002.

[25] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[26] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[27] Martin Burger, José Antonio Carrillo de la Plata, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinetic and Related Models*, 3(1):59–83, 2010.

[28] Martin Burger, Marzena Franek, and Carola-Bibiane Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209–253, 2012.

[29] Martin Burger, Marzena Franek, and Carola-Bibiane Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209–253, 2012.

[30] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[31] Peter J Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

[32] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[33] Eric A Carlen and Jan Maas. An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy. *Communications in Mathematical Physics*, 331(3):887–926, 2014.

[34] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[35] Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

[36] Guillaume Carlier and Clarice Poon. On the total variation Wasserstein gradient flow and the TV-JKO scheme. *arXiv preprint arXiv:1703.00243*, 2017.

[37] José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

[38] Yair Censor and Simeon Reich. The dykstra algorithm with Bregman projections. *Communications in Applied Analysis*, 2(3):407–420, 1998.

[39] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[40] L Chayes and HK Lei. Optimal transport for non-conservative systems. *ArXiv e-prints*, 2014.

[41] Yongxin Chen, Wilfrid Gangbo, Tryphon T Georgiou, and Allen Tannenbaum. On the matrix Monge-Kantorovich problem. *Preprint arXiv:1701.02826*, 2017.

[42] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: A quantum mechanical approach. *Preprint arXiv:1610.03041*, 2016.

[43] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[44] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, pages 1–44, 2016.

[45] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *To appear in Mathematics of Computation*, 2016.

[46] Patrick L Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, pages 1–18, 2016.

[47] Patrick L Combettes and Christian L Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, 2016.

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

[49] Nicolas Courty, Remi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. 2017 (Submited).

[50] Keenan Crane, Mathieu Desbrun, and Peter Schröder. Trivial connections on discrete surfaces. In *Computer Graphics Forum*, volume 29, pages 1525–1533. Wiley Online Library, 2010.

[51] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 2292–2300, 2013.

[52] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[53] Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. BV estimates in optimal transportation and applications. *Archive for Rational Mechanics and Analysis*, 219(2):829–860, 2016.

[54] Julie Delon. Movie and video scale-time equalization application to flicker reduction. *IEEE Transactions on Image Processing*, 15(1):241–248, 2006.

[55] Laurent Demaret, Nira Dyn, and Armin Iske. Image compression by linear splines over adaptive triangulations. *Signal Processing*, 86(7):1604–1616, 2006.

[56] Françoise Demengel and Roger Temam. Convex-functions of a measure and applications. *Indiana University Mathematics Journal*, 33(5):673–709, 1984.

[57] W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[58] Rachid Deriche, D Tschumpelé, Christophe Lenglet, and Mikaël Rousson. *Variational Approaches to the Estimation, Regularization, and Segmentation of Diffusion Tensor Images*, pages 517–530. Springer US, Boston, MA, 2006.

[59] Simone Di Marino and Jean Louet. The entropic regularization of the Monge problem on the real line. *arXiv preprint arXiv:1703.10457*, 2017.

[60] Simone Di Marino and Alpár Richárd Mészáros. Uniqueness issues for evolution equations with density constraints. *Mathematical Models and Methods in Applied Sciences*, 26(09):1761–1783, 2016.

[61] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

[62] Ian L Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123, 2009.

[63] Bertram Düring and Carola-Bibiane Schönlieb. A high-contrast fourth-order PDE from imaging: numerical solution by ADI splitting. *Multi-scale and High-Contrast Partial Differential Equations, H. Ammari et al.(eds.)*, pages 93–103, 2012.

[64] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[65] Jonathan Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[66] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[67] Jalal M Fadili and Gabriel Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2011.

[68] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[69] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[70] Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[71] Jan Henrik Fitschen, Friederike Laus, and Bernhard Schmitzer. Optimal transport for manifold-valued images. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 460–472. Springer, 2017.

[72] Marzena Magdalene Franek. *Variational Methods using Transport Metrics and Applications*. PhD thesis, 2011.

[73] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[74] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[75] A. Galichon and B. Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. Technical report, Preprint SSRN-1487307, 2009.

[76] Thomas Gallouët, Maxime Laborde, and Leonard Monsaingeon. An unbalanced optimal transport splitting scheme for general advection-reaction-diffusion problems. *arXiv preprint arXiv:1704.04541*, 2017.

[77] Thomas Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

[78] Tryphon T Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3):033301, 2015.

[79] Lorenzo Giacomelli and Felix Otto. Variatonal formulation for the lubrication approximation of the Hele-Shaw flow. *Calculus of Variations and Partial Differential Equations*, 13(3):377–403, 2001.

[80] Kevin Guittet. *Extended Kantorovich norms: a tool for optimization*. PhD thesis, INRIA, 2002.

[81] Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448–484, 2004.

[82] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

[83] Leonid G Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

[84] Ingrid Hotz, Louis Feng, Hans Hagen, Bernd Hamann, Kenneth Joy, and Boris Jeremic. Physically based methods for tensor field visualization. pages 123–130. IEEE Computer Society, 2004.

[85] Romain Hug, Emmanuel Maitre, and Nicolas Papadakis. Multi-physics optimal transportation and image interpolation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1671–1692, 2015.

[86] Martin Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

[87] Noureddine Igbida and Van Thanh Nguyen. Optimal partial transport problem with Lagrangian costs. 2017.

[88] Xianhua Jiang, Lipeng Ning, and Tryphon T Georgiou. Distances and Riemannian metrics for multivariate spectral densities. *IEEE Transactions on Automatic Control*, 57(7):1723–1735, July 2012.

[89] Chloé Jimenez. Dynamic formulation of optimal transport problems. *Journal of Convex Analysis*, 15(3):593, 2008.

[90] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[91] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk SSSR*, volume 37, pages 199–201, 1942.

[92] Johan Karlsson and Axel Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *arXiv preprint arXiv:1612.02273*, 2016.

[93] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A fitness-driven cross-diffusion system from population dynamics as a gradient flow. *Journal of Differential Equations*, 261(5):2784–2808, 2016.

[94] Stanislav Kondratyev, Léonard Monsaingeon, Dmitry Vorotnikov, et al. A new optimal transport distance on the space of finite radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

[95] JJ Kosowsky and Alan L Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural networks*, 7(3):477–490, 1994.

[96] Ares Lagae, Sylvain Lefebvre, and Philip Dutré. Improving Gabor noise. *IEEE Transactions on Visualization and Computer Graphics*, 17(8):1096–1107, 2011.

[97] Jan Lellmann, Dirk A Lorenz, Carola Schönlieb, and Tuomo Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[98] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

[99] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. A*, 34(4):1533–1574, 2014.

[100] Bruno Levy. A numerical algorithm for $L^2$ semi-discrete optimal transport in 3d. *M2AN, to appear*, 2015.

[101] Matthias Liero and Alexander Mielke. Gradient structures and geodesic convexity for reaction–diffusion systems. *Phil. Trans. R. Soc. A*, 371(2005):20120346, 2013.

[102] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. On geodesic $\lambda$-convexity with respect to the Hellinger-Kantorovich distance. In preparation.

[103] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *arXiv preprint arXiv:1508.07941*, 2015.

[104] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: The Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

[105] Damiano Lombardi and Emmanuel Maitre. Eulerian models and algorithms for unbalanced optimal transport. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1717–1744, 2015.

[106] Jan Maas, Martin Rumpf, Carola Schönlieb, and Stefan Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1745–1769, 2015.

[107] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[108] Stefania Maniglia. Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *Journal de mathématiques pures et appliquées*, 87(6):601–626, 2007.

[109] Bertrand Maury and A Preux. 13 pressureless euler equations with maximal density constraint: a time-splitting scheme. *Topological Optimization and Optimal Transport: In the Applied Sciences*, 17:333, 2017.

[110] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Mathematical Models and Methods in Applied Sciences*, 20(10):1787–1821, 2010.

[111] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. Congestion-driven dendritic growth. *Discrete Contin. Dyn. Syst*, 34(4):1575–1604, 2014.

[112] Bertrand Maury, Aude Roudneff-Chupin, Filippo Santambrogio, and Juliette Venel. Handling congestion in crowd motion modeling. *Networks and Heterogeneous Media*, 6(3, September 2011):485–519, 2011.

[113] Antoine Mellet, Benoît Perthame, and Fernando Quiros. A hele-shaw problem for tumor growth. *Journal of Functional Analysis*, 2017.

[114] Quentin Mérigot. A multiscale approach to optimal transport. *Comput. Graph. Forum*, 30(5):1583–1592, 2011.

[115] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[116] Lipeng Ning and Tryphon T Georgiou. Metrics for matrix-valued measures via test functions. In *53rd IEEE Conference on Decision and Control*, pages 2642–2647. IEEE, 2014.

[117] Lipeng Ning, Tryphon T Georgiou, and Allen Tannenbaum. On matrix-valued Monge–Kantorovich optimal mass transport. *IEEE transactions on automatic control*, 60(2):373–382, 2015.

[118] Roger D Nussbaum. Entropy minimization, Hilbert's projective metric, and scaling integral kernels. *Journal of functional analysis*, 115(1):45–99, 1993.

[119] Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

[120] Felix Otto. *Double degenerate diffusion equations as steepest descent*. Sonderforschungsbereich 256, 1996.

[121] Felix Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: A mean-field theory. *Archive for Rational Mechanics and Analysis*, 141(1):63–103, 1998.

[122] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[123] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[124] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[125] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Computer vision, 2009 IEEE 12th international conference on*, pages 460–467. IEEE, 2009.

[126] Benoît Perthame, Fernando Quirós, and Juan Luis Vázquez. The Hele-Shaw asymptotics for mechanical models of tumor growth. *Archive for Rational Mechanics and Analysis*, 212(1):93–127, 2014.

[127] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[128] Gabriel Peyré, Lénaïc Chizat, François-Xavier Vialard, and Justin Solomon. Quantum optimal transport for tensor field processing. *to appear in European Journal of Applied Mathematics*, 2017.

[129] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[130] Benedetto Piccoli and Francesco Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

[131] François Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123–137, 2007.

[132] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *Proc. SSVM'15*, 2015.

[133] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[134] R Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[135] R Tyrrell Rockafellar. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

[136] R Tyrrell Rockafellar. Integrals which are convex functionals. ii. *Pacific Journal of Mathematics*, 39(2):439–469, 1971.

[137] R Tyrrell Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

[138] R Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear operators and the calculus of variations*, pages 157–207. Springer, 1976.

[139] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[140] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[141] Martin Rumpf and Benedikt Wirth. Variational methods in shape analysis. In *Handbook of Mathematical Methods in Imaging*, pages 1363–1401. Springer, 2011.

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

[143] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[144] Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *J Math Imaging Vis*, 56(2):238–259, 2016.

[145] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

[146] Bernhard Schmitzer and Christoph Schnörr. Object segmentation by shape matching with Wasserstein modes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 123–136. Springer, 2013.

[147] Bernhard Schmitzer and Benedikt Wirth. Dynamic models of Wasserstein-1-type unbalanced transport. *arXiv preprint arXiv:1705.04535*, 2017.

[148] Bernhard Schmitzer and Benedikt Wirth. A framework for Wasserstein-1-type metrics. *arXiv preprint arXiv:1701.01945*, 2017.

[149] Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

[150] Meisam Sharify, Stéphane Gaubert, and Laura Grigori. Solution of the optimal assignment problem by diagonal scaling algorithms. *arXiv preprint arXiv:1104.3830*, 2011.

[151] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[152] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[153] Justin Solomon, Raif Rustamov, Guibas Leonidas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 306–314, 2014.

[154] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

[155] Anthony C Thompson. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society*, 14(3):438–443, 1963.

[156] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation Lp̂ distance for signal analysis. *Journal of Mathematical Imaging and Vision*, pages 1–24, 2016.

[157] Alain Trouvé and Laurent Younes. Metamorphoses through lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.

[158] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[159] Quang Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.

[160] Amir Vaxman, Marcel Campen, Olga Diamanti, Daniele Panozzo, David Bommes, Klaus Hildebrandt, and Mirela Ben-Chen. Directional field synthesis, design, and processing. *Comput. Graph. Forum*, 35(2):545–572, 2016.

[161] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[162] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[163] Augusto Visintin. Strong convergence results related to strict convexity. *Communications in Partial Differential Equations*, 9(5):439–466, 1984.

[164] Brian A Wandell. Clarifying human white matter. *Annual Review of Neuroscience*, 2016.

[165] Huahua Wang and Arindam Banerjee. Bregman alternating direction method of multipliers. In *Advances in Neural Information Processing Systems*, pages 2816–2824, 2014.

[166] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.

[167] Benedikt Wirth, Leah Bar, Martin Rumpf, and Guillermo Sapiro. Geodesics in shape space via variational time discretization. In *EMMCVPR*, volume 9, pages 288–302. Springer, 2009.

[168] Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer Science & Business Media, 2010.

[169] David M Young. *Iterative solution of large linear systems*. Elsevier, 2014.

[170] Jonathan Zinsl and Daniel Matthes. Transport distances and geodesic convexity for systems of degenerate diffusion equations. *Calculus of Variations and Partial Differential Equations*, 54(4):3397–3438, 2015.

[171] Étienne Ghys. Gaspard monge. http://images.math.cnrs.fr/Gaspard-Monge,1094.html, 2012. Accessed: 2017-08-30.

## Résumé

L'objet de cette thèse est d'étendre le cadre théorique et les méthodes numériques du transport optimal à des objets plus généraux que des mesures de probabilité. En premier lieu, nous définissons des modèles de transport optimal entre mesures positives suivant deux approches, interpolation et couplage de mesures, dont nous montrons l'équivalence. De ces modèles découle une généralisation des métriques de Wasserstein. Dans une seconde partie, nous développons des méthodes numériques pour résoudre les deux formulations et étudions en particulier une nouvelle famille d'algorithmes de "scaling", s'appliquant à une grande variété de problèmes. La troisième partie contient des illustrations ainsi que l'étude théorique et numérique, d'un flot de gradient de type Hele-Shaw dans l'espace des mesures. Pour les mesures à valeurs matricielles, nous proposons aussi un modèle de transport optimal qui permet un bon arbitrage entre fidélité géométrique et efficacité algorithmique.



## Abstract

This thesis generalizes optimal transport beyond the classical ``balanced" setting of probability distributions. We define unbalanced optimal transport models between nonnegative measures, based either on the notion of interpolation or the notion of coupling of measures. We show relationships between these approaches. One of the outcomes of this framework is a generalization of the p-Wasserstein metrics. Secondly, we build numerical methods to solve interpolation and coupling-based models. We study, in particular, a new family of scaling algorithms that generalize Sinkhorn's algorithm. The third part deals with applications. It contains a theoretical and numerical study of a Hele-Shaw type gradient flow in the space of nonnegative measures. It also adresses the case of measures taking values in the cone of positive semi-definite matrices, for which we introduce a model that achieves a balance between geometrical accuracy and algorithmic efficiency.



## Mots Clés

Transport optimal, analyse convexe, optimisation, mesures positives, géométrie de l'information, algorithme de Sinkhorn, traitement d'image, flots de gradient, convergence faible, traitement de champs de tenseurs, barycentres, entropie relative, espace métrique géodésique

## Keywords

Optimal transport, convex analysis, optimization, nonnegative measures, information geometry, Sinkhorn's algorithm, image processing, gradient flows, weak convergence, tensor field processing, barycentres, relative entropy, geodesic metric space