10

Extensions of Optimal Transport

This chapter details several variational problems that are related to (and share the same structure of) the Kantorovich formulation of optimal transport. The goal is to extend optimal transport to more general settings: several input histograms and measures, unnormalized ones, more general classes of measures, and optimal transport between measures that focuses on local regularities (points nearby in the source measure should be mapped onto points nearby in the target measure) rather than a total transport cost, including cases where these two measures live in different metric spaces.

10.1 Multimarginal Problems

Instead of coupling two input histograms using the Kantorovich formulation (2.11), one can couple S histograms $(\mathbf{a}_s)_{s=1}^S$, where $\mathbf{a}_s \in \Sigma_{n_s}$, by solving the following multimarginal problem:

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def.}}{=} \sum_s \sum_{i_s=1}^{n_s} \mathbf{C}_{i_1,\ldots,i_S} \mathbf{P}_{i_1,\ldots,i_S}, \tag{10.1}$$

where the set of valid couplings is

$$\mathbf{U}(\mathbf{a}_s)_s = \left\{ \mathbf{P} \in \mathbb{R}^{n_1 \times \ldots \times n_S} \;:\; \forall\, s, \forall\, i_s, \sum_{\ell \neq s} \sum_{i_\ell=1}^{n_\ell} \mathbf{P}_{i_1,\ldots,i_S} = \mathbf{a}_{s,i_s} \right\}.$$

The entropic regularization scheme (4.2) naturally extends to this setting

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}_s)_s} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}),$$

and one can then apply Sinkhorn's algorithm to compute the optimal $\mathbf{P}$ in scaling form, where each entry indexed by a multi-index vector $i = (i_1, \ldots, i_S)$

$$\mathbf{P}_i = \mathbf{K}_i \prod_{s=1}^{S} \mathbf{u}_{s,i_s} \quad \text{where} \quad \mathbf{K} \overset{\text{def.}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}},$$

where $\mathbf{u}_s \in \mathbb{R}_+^{n_s}$ are (unknown) scaling vectors, which are iteratively updated, by cycling repeatedly through $s = 1, \ldots, S$,

$$\mathbf{u}_{s,i_s} \leftarrow \frac{\mathbf{a}_{s,i_s}}{\sum_{\ell \neq s} \sum_{i_\ell=1}^{n_\ell} \mathbf{K}_i \prod_{r \neq s} \mathbf{u}_{\ell,i_r}} \tag{10.2}$$

.

Remark 10.1 (General measures). The discrete multimarginal problem (10.1) is generalized to measures $(\alpha_s)_s$ on spaces $(\mathcal{X}_1, \ldots, \mathcal{X}_S)$ by computing a coupling measure

$$\min_{\pi \in \mathcal{U}(\alpha_s)_s} \int_{\mathcal{X}_1 \times \ldots \times \mathcal{X}_S} c(x_1, \ldots, x_S) \mathrm{d}\pi(x_1, \ldots, x_S), \tag{10.3}$$

where the set of couplings is

$$\mathcal{U}(\alpha_s)_s \overset{\text{def.}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}_1 \times \ldots \times \mathcal{X}_S) \;:\; \forall\, s = 1, \ldots, S, P_{s,\sharp}\pi = \alpha_s, \right\}$$

where $P_s : \mathcal{X}_1 \times \ldots \times \mathcal{X}_S \to \mathcal{X}_s$ is the projection on the sth component, $P_s(x_1, \ldots, x_S) = x_s$; see, for instance, (Gangbo and Swiech, 1998). We refer to (Pass, 2015; Pass, 2012) for a review of the main properties of the multimarginal OT problem. A typical application of multimarginal OT is to compute approximation of solutions to quantum chemistry problems, and in particular, in density functional theory (Cotar *et al.*, 2013; Gori-Giorgi *et al.*, 2009; Buttazzo *et al.*, 2012). This problem is obtained when considering the singular

Coulomb interaction cost

$$c(x_1,\dots,x_S)=\sum_{i\neq j}\frac{1}{\|x_i-x_j\|}.$$

Remark 10.2 (Multimarginal formulation of the barycenter)**.** It is possible to recast the linear program optimization (9.11) as an optimization over a single coupling over $\mathcal{X}^{S+1}$ where the last marginal is the barycenter and the other ones are the input measure $(\alpha_s)_{s=1}^S$

$$\min_{\bar\pi\in\mathcal{M}_+^1(X^{S+1})}\int_{\mathcal{X}^{S+1}}\sum_{s=1}^S\lambda_s c(x,x_s)\mathrm{d}\bar\pi(x_1,\dots,x_s,x)\qquad(10.4)$$

$$\text{subject to}\quad\forall\, s=1,\dots,S,\quad P_{s,\sharp}\bar\pi=\alpha_s.$$

This stems from the "gluing lemma," which states that given couplings $(\pi_s)_{s=1}^S$ where $\pi_s\in\mathcal{U}(\alpha_s,\alpha)$, one can construct a higher-dimensional coupling $\bar\pi\in\mathcal{M}_+^1(X^{S+1})$ with marginals π_s, *i.e.* such that $Q_{s\sharp}\bar\pi=\pi_s$, where $Q_s(x_1,\dots,x_S,x)\overset{\text{def.}}{=}(x_s,x)\in\mathcal{X}^2$. By explicitly minimizing in (10.4) with respect to the last marginal (associated to $x\in\mathcal{X}$), one obtains that solutions α of the barycenter problem (9.11) can be computed as $\alpha=A_{\lambda,\sharp}\pi$, where A_λ is the "barycentric map" defined as

$$A_\lambda:(x_1,\dots,x_S)\in\mathcal{X}^S\mapsto\operatorname*{argmin}_{x\in\mathcal{X}}\sum_s\lambda_s c(x,x_s)$$

(assuming this map is single-valued), where π is any solution of the multimarginal problem (10.3) with cost

$$c(x_1,\dots,x_S)=\sum_\ell\lambda_\ell c(x_\ell,A_\lambda(x_1,\dots,x_S)).\qquad(10.5)$$

For instance, for $c(x,y)=\|x-y\|^2$, one has, removing the constant squared terms,

$$c(x_1,\dots,x_S)=-\sum_{r\leq s}\lambda_r\lambda_s\langle x_r,\,x_s\rangle,$$

which is a problem studied in Gangbo and Swiech (1998). We refer to Agueh and Carlier (2011) for more details. This formula shows that if all the input measures are discrete $\beta_s = \sum_{i_s=1}^{n_s} \mathbf{a}_{s,i_s} \delta_{x_{s,i_s}}$, then the barycenter α is also discrete and is obtained using the formula

$$\alpha = \sum_{(i_1,\ldots,i_S)} \mathbf{P}_{(i_1,\ldots,i_S)} \delta_{A_\lambda(x_{i_1},\ldots,x_{i_S})},$$

where $\mathbf{P}$ is an optimal solution of (10.1) with cost matrix $\mathbf{C}_{i_1,\ldots,i_S} = c(x_{i_1},\ldots,x_{i_S})$ as defined in (10.5). Since $\mathbf{P}$ is a nonnegative tensor of $\prod_s n_s$ dimensions obtained as the solution of a linear program with $\sum_s n_s - S + 1$ equality constraints, an optimal solution $\mathbf{P}$ with up to $\sum_s n_s - S + 1$ nonzero values can be obtained. A barycenter α with a support of up to $\sum_s n_s - S + 1$ points can therefore be obtained. This result and other considerations in the discrete case can be found in Anderes *et al.* (2016).

Remark 10.3 (Relaxation of Euler equations)**.** A convex relaxation of Euler equations of incompressible fluid dynamics has been proposed by Brenier (1990, 1993, 1999, 2008) and (Ambrosio and Figalli, 2009). Similarly to the setting exposed in §7.6, it corresponds to the problem of finding a probability distribution $\bar{\pi} \in \mathcal{M}_+^1(\bar{\mathcal{X}})$ over the set $\bar{\mathcal{X}}$ of all paths $\gamma : [0,1] \to \mathcal{X}$, which describes the movement of particules in the fluid. This is a relaxed version of the initial partial differential equation model because, as in the Kantorovich formulation of OT, mass can be split. The evolution with time does not necessarily define a diffemorphism of the underlying space $\mathcal{X}$. The dynamic of the fluid is obtained by minimizing as in (7.17) the energy $\int_0^1 \|\gamma'(t)\|^2 \, \mathrm{d}t$ of each path. The difference with OT over the space of paths is the additional incompressibilty of the fluid. This incompressibilty is taken care of by imposing that the density of particules should be uniform at any time $t \in [0,1]$ (and not just imposed at initial and final times $t \in \{0,1\}$ as in classical OT). Assuming $\mathcal{X}$ is compact and denoting $\rho_{\mathcal{X}}$ the uniform distribution on $\mathcal{X}$, this reads $\bar{P}_{t,\sharp}\bar{\pi} = \rho_{\mathcal{X}}$ where $\bar{P}_t : \gamma \in \bar{\mathcal{X}} \to \gamma(t) \in \mathcal{X}$. One can discretize this problem by replacing a continuous path $(\gamma(t))_{t\in[0,1]}$

by a sequence of S points $(x_{i_1}, x_{i_2}, \ldots, x_{i_S})$ on a grid $(x_k)_{k=1}^n \subset \mathcal{X}$, and $\bar{\Pi}$ is represented by an S-way coupling $\mathbf{P} \in \mathbb{R}^{n^S} \in \mathcal{U}(\mathbf{a}_s)_s$, where the marginals are uniform $\mathbf{a}_s = n^{-1}\mathbb{1}_n$. The cost of the corresponding multimarginal problem is then

$$\mathbf{C}_{i_1,\ldots,i_S} = \sum_{s=1}^{S-1} \|x_{i_s} - x_{i_{s+1}}\|^2 + R\left\|x_{\sigma(i_1)} - x_{i_S}\right\|^2. \tag{10.6}$$

Here R is a large enough penalization constant, which is here to enforce the movement of particules between initial and final times, which is prescribed by a permutation $\sigma : [\![n]\!] \to [\![n]\!]$. This resulting multimarginal problem is implemented efficiently in conjunction with Sinkhorn iterations (10.2) using the special structure of the cost, as detailed in (Benamou *et al.*, 2015). Indeed, in place of the $O(n^S)$ cost required to compute the denominator appearing in (10.2), one can decompose it as a succession of S matrix-vector multiplications, hence with a low cost Sn^2. Note that other solvers have been proposed, for instance, using the semidiscrete framework shown in §5.2; see (Goes *et al.*, 2015; Gallouët and Mérigot, 2017).

10.2 Unbalanced Optimal Transport

A major bottleneck of optimal transport in its usual form is that it requires the two input measures (α, β) to have the same total mass. While many workarounds have been proposed (including renormalizing the input measure, or using dual norms such as detailed in § 8.2), it is only recently that satisfying unifying theories have been developed. We only sketch here a simple but important particular case.

Following Liero *et al.* (2018), to account for arbitrary positive histograms $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$, the initial Kantorovich formulation (2.11) is "relaxed" by only penalizing marginal deviation using some divergence $\mathbf{D}_\varphi$, defined in (8.3). This equivalently corresponds to minimizing an

OT distance between approximate measures

$$\mathrm{L}_{\mathbf{C}}^{\tau}(\mathbf{a},\mathbf{b}) = \min_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}} \mathrm{L}_{\mathbf{C}}(\tilde{\mathbf{a}},\tilde{\mathbf{b}}) + \tau_1 \mathbf{D}_\varphi(\mathbf{a},\tilde{\mathbf{a}}) + \tau_2 \mathbf{D}_\varphi(\mathbf{b},\tilde{\mathbf{b}}) \tag{10.7}$$

$$= \min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P}\rangle + \tau_1 \mathbf{D}_\varphi(\mathbf{P}\mathbb{1}_m|\mathbf{a}) + \tau_2 \mathbf{D}_\varphi(\mathbf{P}^\top \mathbb{1}_m|\mathbf{b}), \tag{10.8}$$

where (τ_1,τ_2) controls how much mass variations are penalized as opposed to transportation of the mass. In the limit $\tau_1=\tau_2\to+\infty$, assuming $\sum_i \mathbf{a}_i = \sum_j \mathbf{b}_j$ (the "balanced" case), one recovers the original optimal transport formulation with hard marginal constraint (2.11).

This formalism recovers many different previous works, for instance introducing for $\mathbf{D}_\varphi$ an ℓ^2 norm (Benamou, 2003) or an ℓ^1 norm as in partial transport (Figalli, 2010; Caffarelli and McCann, 2010). A case of particular importance is when using $\mathbf{D}_\varphi = \mathbf{KL}$ the Kulback–Leibler divergence, as detailed in Remark 10.5. For this cost, in the limit $\tau=\tau_1=\tau_2\to 0$, one obtains the so-called squared Hellinger distance (see also Example 8.3)

$$\mathrm{L}_{\mathbf{C}}^{\tau}(\mathbf{a},\mathbf{b}) \overset{\tau\to 0}{\longrightarrow} \mathfrak{h}^2(\mathbf{a},\mathbf{b}) = \sum_i (\sqrt{\mathbf{a}_i} - \sqrt{\mathbf{b}_i})^2.$$

Sinkhorn's iterations (4.15) can be adapted to this problem by making use of the generalized algorithm detailed in §4.6. This means that the solution has the form (4.12) and that the scalings are updated as

$$\mathbf{u} \leftarrow \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^{\frac{\tau_1}{\tau_1+\varepsilon}} \quad \text{and} \quad \mathbf{v} \leftarrow \left(\frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}}\right)^{\frac{\tau_2}{\tau_2+\varepsilon}}. \tag{10.9}$$

Remark 10.4 (Generic measure). For (α,β) two arbitrary measures, the unbalanced version (also called "log-entropic") of (2.15) reads

$$\mathcal{L}_c^{\tau}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \tau\mathcal{D}_\varphi(P_{1,\sharp}\pi|\alpha) + \tau\mathcal{D}_\varphi(P_{2,\sharp}\pi|\beta),$$

where divergences $\mathcal{D}_\varphi$ between measures are defined in (8.1). In the special case $c(x,y)=\|x-y\|^2$, $\mathcal{D}_\varphi=\mathrm{KL}$, $\mathcal{L}_c^{\tau}(\alpha,\beta)^{1/2}$ is the

Gaussian–Hellinger distance (Liero *et al.*, 2018), and it is shown to be a distance on $\mathcal{M}_+^1(\mathbb{R}^d)$.

Remark 10.5 (Wasserstein–Fisher–Rao)**.** For the particular choice of cost

$$c(x,y) = -\log\cos(\min(d(x,y)/\kappa, \pi/2)),$$

where κ is some cutoff distance, and using $\mathcal{D}_\varphi = \mathrm{KL}$, then

$$\mathrm{WFR}(\alpha,\beta) \overset{\text{def.}}{=} \mathcal{L}_c^\tau(\alpha,\beta)^{\frac{1}{2}}$$

is the so-called Wasserstein–Fisher–Rao or Hellinger–Kantorovich distance. In the special case $\mathcal{X} = \mathbb{R}^d$, this static (Kantorovich-like) formulation matches its dynamical counterparts (7.15), as proved independently by Liero *et al.* (2018) and Chizat *et al.* (2018c). This dynamical formulation is detailed in §7.4.

The barycenter problem (9.11) can be generalized to handle an unbalanced setting by replacing $\mathcal{L}_c$ with $\mathcal{L}_c^\tau$. Figure 10.1 shows the resulting interpolation, providing a good illustration of the usefulness of the relaxation parameter τ. The input measures are mixtures of two Gaussians with unequal mass. Classical OT requires the leftmost bump to be split in two and gives a nonregular interpolation. In sharp contrast, unbalanced OT allows the mass to vary during interpolation, so that the bumps are not split and local modes of the distributions are smoothly matched. Using finite values for τ (recall that OT is equivalent to $\tau = \infty$) is thus important to prevent irregular interpolations that arise because of mass splitting, which happens because of a "hard" mass conservation constraint. The resulting optimization problem can be tackled numerically using entropic regularization and the generalized Sinkhorn algorithm detailed in §4.6.

In practice, unbalanced OT techniques seem to outperform classical OT for applications (such as in imaging or machine learning) where the input data is noisy or not perfectly known. They are also crucial when the signal strength of a measure, as measured by its total mass, must be accounted for, or when normalization is not meaningful. This was the original motivation of Frogner *et al.* (2015), whose goal was to compare

sets of word labels used to describe images. Unbalanced OT and the corresponding Sinkhorn iterations have also been used for applications to the dynamics of cells in (Schiebinger *et al.*, 2017).

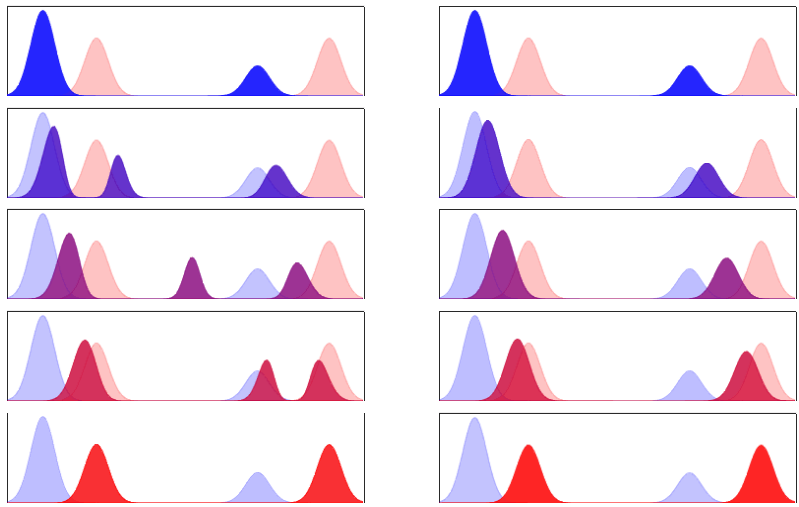

Figure 10.1: Influence of relaxation parameter τ on unbalanced barycenters. Top to bottom: the evolution of the barycenter between two input measures.

Remark 10.6 (Connection with dual norms)**.** A particularly simple setup to account for mass variation is to use dual norms, as detailed in §8.2. By choosing a compact set $B \subset \mathcal{C}(\mathcal{X})$ one obtains a norm defined on the whole space $\mathcal{M}(\mathcal{X})$ (in particular, the measures do not need to be positive). A particular instance of this setting is the flat norm (8.11), which is recovered as a special instance of unbalanced transport, when using $\mathcal{D}_\varphi(\alpha|\alpha') = \|\alpha - \alpha'\|_{\mathrm{TV}}$ to be the total variation norm (8.9); see, for instance, (Hanin, 1992; Lellmann *et al.*, 2014). We also refer to (Schmitzer and Wirth, 2017) for a general framework to define Wasserstein-1 unbalanced transport.

10.3 Problems with Extra Constraints on the Couplings

Many other OT-like problems have been proposed in the literature. They typically correspond to adding extra constraints $\mathcal{C}$ on the initial

OT problem (2.15)

$$\min_{\pi \in \mathcal{U}(\alpha,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \;:\; \pi \in \mathcal{C} \right\}. \tag{10.10}$$

Let us give two representative examples. The optimal transport with capacity constraint (Korman and McCann, 2015) corresponds to imposing that the density ρ_π (for instance, with respect to the Lebesgue measure) is upper bounded

$$\mathcal{C} = \{\pi \;:\; \rho_\pi \leq \kappa\} \tag{10.11}$$

for some $\kappa > 0$. This constraint rules out singular couplings localized on Monge maps. The martingale transport problem (see, for instance, Galichon *et al.* (2014), Dolinsky and Soner (2014), Tan and Touzi (2013), and Beiglböck *et al.* (2013)), which finds many applications in finance, imposes the so-called martingale constraint on the conditional mean of the coupling, when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$:

$$\mathcal{C} = \left\{ \pi \;:\; \forall x \in \mathbb{R}^d, \int_{\mathbb{R}^d} y \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)} \mathrm{d}\beta(y) = x \right\}. \tag{10.12}$$

This constraint imposes that the barycentric projection map (4.20) of any admissible coupling must be equal to the identity. For arbitrary (α, β), this set $\mathcal{C}$ is typically empty, but necessary and sufficient conditions exist (α and β should be in "convex order") to ensure $\mathcal{C} \neq \emptyset$ so that (α, β) satisfy a martingale constraint. This constraint can be difficult to enforce numerically when discretizing an existing problem. It also forbids the solution to concentrate on a single Monge map, and can lead to couplings concentrated on the union of several graphs (a "multi-valued" Monge map), or even more complicated support sets. Using an entropic penalization as in (4.9), one can solve approximately (10.10) using the Dykstra algorithm as explained in Benamou *et al.* (2015), which is a generalization of Sinkhorn's algorithm shown in §4.2. This requires computing the projection onto $\mathcal{C}$ for the KL divergence, which is straightforward for (10.11) but cannot be done in closed form (10.12) and thus necessitates subiterations; see (Guo and Obloj, 2017) for more details.

10.4 Sliced Wasserstein Distance and Barycenters

One can define a distance between two measures (α, β) defined on $\mathbb{R}^d$ by aggregating 1-D Wasserstein distances between their projections onto all directions of the sphere. This defines

$$\mathrm{SW}(\alpha,\beta)^2 \overset{\text{def.}}{=} \int_{\mathbf{S}^d} \mathcal{W}_2(P_{\theta,\sharp}\alpha, P_{\theta,\sharp}\beta)^2 \mathrm{d}\theta, \tag{10.13}$$

where $\mathbf{S}^d = \{\theta \in \mathbb{R}^d \;:\; \|\theta\| = 1\}$ is the d-dimensional sphere, and $P_\theta : x \in \mathbb{R}^d \to \mathbb{R}$ is the projection. This approach is detailed in (Bonneel *et al.*, 2015), following ideas from Marc Bernot. It is related to the problem of Radon inversion over measure spaces (Abraham *et al.*, 2017).

Lagrangian discretization and stochastic gradient descent. The advantage of this functional is that 1-D Wasserstein distances are simple to compute, as detailed in §2.6. In the specific case where $m = n$ and

$$\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \quad \text{and} \quad \beta = \frac{1}{n}\sum_{i=1}^m \delta_{y_i}, \tag{10.14}$$

this is achieved by simply sorting points

$$\mathrm{SW}(\alpha,\beta)^2 = \int_{\mathbf{S}^d} \left(\sum_{i=1}^n |\langle x_{\sigma_\theta(i)} - y_{\kappa_\theta(i)}, \theta\rangle|^2 \right) \mathrm{d}\theta,$$

where $\sigma_\theta, \kappa_\theta \in \mathrm{Perm}(n)$ are the permutation ordering in increasing order, respectively, $(\langle x_i, \theta\rangle)_i$ and $(\langle y_i, \theta\rangle)_i$.

Fixing the vector y, the function $\mathcal{E}_\beta(x) \overset{\text{def.}}{=} \mathrm{SW}(\alpha,\beta)^2$ is smooth, and one can use this function to define a mapping by gradient descent

$$x \leftarrow x - \tau \nabla \mathcal{E}_\beta(x) \quad \text{where} \tag{10.15}$$

$$\nabla \mathcal{E}_\beta(x)_i = 2 \int_{\mathbf{S}^d} \Big(\langle x_i - y_{\kappa_\theta \circ \sigma_\theta^{-1}(i)}, \theta\rangle \theta \Big) \mathrm{d}\theta$$

using a small enough step size $\tau > 0$. To make the method tractable, one can use a stochastic gradient descent (SGD), replacing this integral with a discrete sum against randomly drawn directions $\theta \in \mathbf{S}^d$ (see §5.4 for more details on SGD). The flow (10.15) can be understood as (Lagrangian implementation of) a Wasserstein gradient flow (in the sense

of §9.3) of the function $\alpha \mapsto \mathrm{SW}(\alpha,\beta)^2$. Numerically, one finds that this flow has no local minimizer and that it thus converges to $\alpha = \beta$. The usefulness of the Lagrangian solver is that, at convergence, it defines a matching (similar to a Monge map) between the two distributions. This method has been used successfully for color transfer and texture synthesis in (Rabin *et al.*, 2011) and is related to the alternate minimization approach detailed in (Pitié *et al.*, 2007).

It is simple to extend this Lagrangian scheme to compute approximate "sliced" barycenters of measures, by mimicking the Frechet definition of Wasserstein barycenters (9.11) and minimizing

$$\min_{\alpha\in\mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \,\mathrm{SW}(\alpha,\beta_s)^2, \tag{10.16}$$

given a set $(\beta_s)_{s=1}^S$ of fixed input measure. Using a Lagrangian discretization of the form (10.14) for both α and the $(\beta_s)_s$, one can perform the nonconvex minimization over the position $x = (x_i)_i$

$$\min_x \mathcal{E}(x) \overset{\text{def.}}{=} \sum_s \lambda_s \mathcal{E}_{\beta_s}(x), \quad \text{and} \quad \nabla\mathcal{E}(x) = \sum_s \lambda_s \nabla\mathcal{E}_{\beta_s}(x), \tag{10.17}$$

by gradient descent using formula (10.15) to compute $\nabla\mathcal{E}_{\beta_s}(x)$ (coupled with a random sampling of the direction θ).

Eulerian discretization and Radon transform. A related way to compute an approximated sliced barycenter, without resorting to an iterative minimization scheme, is to use the fact that (10.13) computes a distance between the Radon transforms $\mathcal{R}(\alpha)$ and $\mathcal{R}(\beta)$ where

$$\mathcal{R}(\alpha) \overset{\text{def.}}{=} (P_{\theta,\sharp}\alpha)_{\theta\in\mathbf{S}^d}.$$

A crucial point is that the Radon transform is invertible and that its inverse can be computed using a filtered backprojection formula. Given a collection of measures $\rho = (\rho_\theta)_{\theta\in\mathbf{S}^d}$, one defines the filtered backprojection operator as

$$\mathcal{R}^+(\rho) = C_d \Delta^{\frac{d-1}{2}} \mathcal{B}(\rho), \tag{10.18}$$

where $\xi = \mathcal{B}(\rho) \in \mathcal{M}(\mathbb{R}^d)$ is the measure defined through the relation

$$\forall\, g \in \mathcal{C}(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} g(x)\mathrm{d}\xi(x) = \int_{\mathbf{S}^d}\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} g(r\theta + U_\theta z)\mathrm{d}\rho_\theta(r)\mathrm{d}z\mathrm{d}\theta, \tag{10.19}$$

where U_θ is any orthogonal basis of $\theta^\perp$, and where $C_d \in \mathbb{R}$ is a normalizing constant which depends on the dimension. Here $\Delta^{\frac{d-1}{2}}$ is a fractional Laplacian, which is the high-pass filter defined over the Fourier domain as $\hat{\Delta}^{\frac{d-1}{2}}(\omega) = \|\omega\|^{d-1}$. The definition of the backprojection (10.19) adds up the contribution of all the measures $(\rho_\theta)_\theta$ by extending each one as being constant in the directions orthogonal to θ. One then has the left-inverse relation $\mathcal{R}^+ \circ \mathcal{R} = \mathbb{I}_{\mathcal{M}(\mathbb{R}^d)}$, so that $\mathcal{R}^+$ is a valid reconstruction formula.

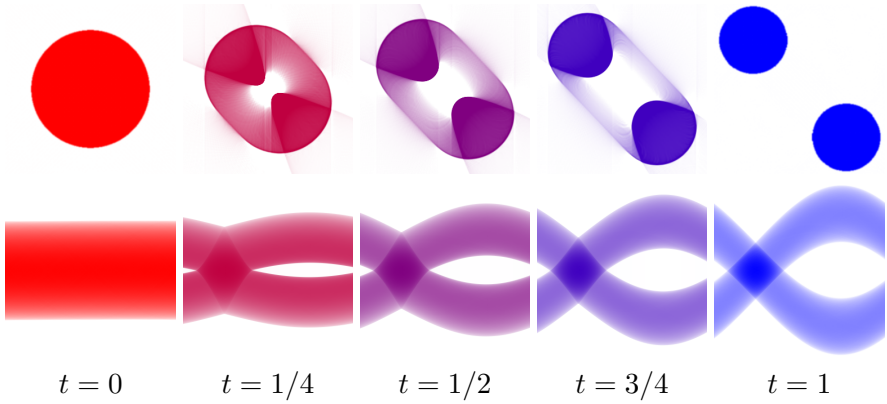

Figure 10.2: Example of sliced barycenters computation using the Radon transform (as defined in (10.20)). Top: barycenters α_t for $S = 2$ two input and weights $(\lambda_1, \lambda_2) = (1 - t, t)$. Bottom: their Radon transform $\mathcal{R}(\alpha_t)$ (the horizontal axis being the orientation angle θ).

In order to compute barycenters of input densities, it makes sense to replace formula (9.11) by its equivalent using Radon transform, and thus consider independently for each θ the 1-D barycenter problem

$$\rho_\theta^\star \in \operatorname*{argmin}_{(\rho_\theta \in \mathcal{M}_+^1(\mathbb{R}))} \sum_{s=1}^{S} \lambda_s \, \mathcal{W}_2(\rho_\theta, P_{\theta,\sharp}\beta_s)^2. \tag{10.20}$$

Each 1-D barycenter problem is easily computed using the monotone rearrangement as detailed in Remark 9.6. The Radon approximation $\alpha_R \overset{\text{def.}}{=} \mathcal{R}^+(\rho^\star)$ of a sliced barycenter solving (9.11) is then obtained by the inverse Radon transform $\mathcal{R}^+$. Note that in general, α_R is not a solution to (9.11) because the Radon transform is not surjective, so that $\rho^\star$, which is obtained as a barycenter of the Radon transforms $\mathcal{R}(\beta_s)$

does not necessarily belong to the range of $\mathcal{R}$. But numerically it seems in practice to be almost the case (Bonneel *et al.*, 2015). Numerically, this Radon transform formulation is very effective for input measures and barycenters discretized on a fixed grid (*e.g.* a uniform grid for images), and $\mathcal{R}$ and well as $\mathcal{R}^+$ are computed approximately on this grid using fast algorithms (see, for instance, (Averbuch *et al.*, 2001)). Figure 10.2 illustrates this computation of barycenters (and highlights the way the Radon transforms are interpolated), while Figure 10.3 shows a comparison of the Radon barycenters (10.20) and the ones obtained by Lagrangian discretization (10.17).

Figure 10.3: Comparison of barycenters computed using Radon transform (10.20) (Eulerian discretization), Lagrangian discretization (10.17), and Wasserstein OT (computed using Sinkhorn iterations (9.18)).

Sliced Wasserstein kernels. Beside its computational simplicity, another advantage of the sliced Wasserstein distance is that it is isometric to a Euclidean distance (it is thus a "Hilbertian" metric), as detailed in Remark 2.30, and in particular formula (2.36). As highlighted in §8.3, this should be contrasted with the Wasserstein distance $\mathcal{W}_2$ on $\mathbb{R}^d$, which is not Hilbertian in dimension $d \geq 2$. It is thus possible to use this sliced distance to equip the space of distributions $\mathcal{M}_1^+(\mathbb{R}^d)$ with a reproducing kernel Hilbert space structure (as detailed in §8.3). One can, for instance, use the exponential and energy distance kernels

$$k(\alpha,\beta) = e^{-\frac{\mathrm{SW}(\alpha,\beta)^p}{2\sigma^p}} \quad \text{and} \quad k(\alpha,\beta) = -\,\mathrm{SW}(\alpha,\beta)^p$$

for $1 \leq p \leq 2$ for the exponential kernels and $0 < p < 2$ for the energy distance kernels. This means that for any collection $(\alpha_i)_i$ of input measures, the matrix $(k(\alpha_i, \alpha_j))_{i,j}$ is symmetric positive semidefinite. It is possible to use these kernels to perform a variety of machine learning tasks using the "kernel trick," for instance, in regression, classification (SVM and logistic), clustering (K-means) and dimensionality reduction (PCA) (Hofmann *et al.*, 2008). We refer to Kolouri *et al.* (2016) for details and applications.

10.5 Transporting Vectors and Matrices

Real-valued measures $\alpha \in \mathcal{M}(\mathcal{X})$ are easily generalized to vector-valued measures $\alpha \in \mathcal{M}(\mathcal{X}; \mathbb{V})$, where $\mathbb{V}$ is some vector space. For notational simplicity, we assume $\mathbb{V}$ is Euclidean and equipped with some inner product $\langle \cdot, \cdot \rangle$ (typically $\mathbb{V} = \mathbb{R}^d$ and the inner product is the canonical one). Thanks to this inner product, vector-valued measures are identified with the dual of continuous functions $g : \mathcal{X} \to \mathbb{V}$, *i.e.* for any such g, one defines its integration against the measure as

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) \in \mathbb{R}, \tag{10.21}$$

which is a linear operation on g and α. A discrete measure has the form $\alpha = \sum_i \mathbf{a}_i \delta_{x_i}$ where $(x_i, a_i) \in \mathcal{X} \times \mathbb{V}$ and the integration formula (10.21) simply reads

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) = \sum_i \langle \mathbf{a}_i, g(x_i) \rangle \in \mathbb{R}.$$

Equivalently, if $\mathbb{V} = \mathbb{R}^d$, then such an α can be viewed as a collection $(\alpha_s)_{s=1}^d$ of d "classical" real-valued measures (its coordinates), writing

$$\int_{\mathcal{X}} g(x) \mathrm{d}\alpha(x) = \sum_{s=1}^{d} \int_{\mathcal{X}} g_s(x) \mathrm{d}\alpha_s(x),$$

where $g(x) = (g_s(x))_{s=1}^d$ are the coordinates of g in the canonical basis.

Dual norms. It is nontrivial, and in fact in general impossible, to extend OT distances to such a general setting. Even coping with real-valued measures taking both positive and negative values is difficult.

The only simple option is to consider dual norms, as defined in §8.2. Indeed, formula (6.3) readily extends to $\mathcal{M}(\mathcal{X};\mathbb{V})$ by considering B to be a subset of $\mathcal{C}(\mathcal{X};\mathbb{V})$. So in particular, $\mathcal{W}_1$, the flat norm and MMD norms can be computed for vector-valued measures.

OT over cone-valued measures. It is possible to define more advanced OT distances when α is restricted to be in a subset $\mathcal{M}(\mathcal{X};\mathcal{V}) \subset \mathcal{M}(\mathcal{X};\mathbb{V})$. The set $\mathcal{V}$ should be a positively 1-homogeneous convex cone of $\mathbb{V}$

$$\mathcal{V} \overset{\text{def.}}{=} \Big\{ \lambda u \;:\; \lambda \in \mathbb{R}^+, u \in \mathcal{V}_0, \Big\}$$

where $\mathcal{V}_0$ is a compact convex set. A typical example is the set of positive measures where $\mathcal{V} = \mathbb{R}^d_+$. Dynamical convex formulations of OT over such a cone have been proposed; see (Zinsl and Matthes, 2015). This has been applied to model the distribution of chemical components. Another important example is the set of positive symmetric matrices $\mathcal{V} = \mathcal{S}^d_+ \subset \mathbb{R}^{d\times d}$. It is of course possible to use dual norms over this space, by treating matrices as vectors; see, for instance, (Ning and Georgiou, 2014). Dynamical convex formulations for OT over such a cone have been provided (Chen *et al.*, 2016b; Jiang *et al.*, 2012). Some static (Kantorovich-like) formulations also have been proposed (Ning *et al.*, 2015; Peyré *et al.*, 2017), but a mathematically sound theoretical framework is still missing. In particular, it is unclear if these static approaches define distances for vector-valued measures and if they relate to some dynamical formulation. Figure 10.4 is an example of tensor interpolation obtained using the method detailed in (Peyré *et al.*, 2017), which proposes a generalization of Sinkhorn algorithms using quantum relative entropy (10.22) to deal with tensor fields.

OT over positive matrices. A related but quite different setting is to replace discrete measures, *i.e.* histograms $\mathbf{a} \in \Sigma_n$, by positive matrices with unit trace $A \in \mathcal{S}^+_n$ such that $\mathrm{tr}(A) = 1$. The rationale is that the eigenvalues $\lambda(A) \in \Sigma_n$ of A play the role of a histogram, but one also has to take care of the rotations of the eigenvectors, so that this problem is more complicated.

One can extend several divergences introduced in §8.1 to this setting. For instance, the Bures metric (2.42) is a generalization of the Hellinger

distance (defined in Remark 8.3), since they are equal on positive diagonal matrices. One can also extend the Kullback–Leibler divergence (4.6) (see also Remark 8.1), which is generalized to positive matrices as

$$\mathbf{KL}(A|B) \overset{\text{def.}}{=} \operatorname{tr}\left(P\log(P) - P\log(Q) - P + Q,\right) \tag{10.22}$$

where $\log(\cdot)$ is the matrix logarithm. This matrix $\mathbf{KL}$ is convex with both of its arguments.

It is possible to solve convex dynamic formulations to define OT distances between such matrices (Carlen and Maas, 2014; Chen *et al.*, 2016b; Chen *et al.*, 2017). There also exists an equivalent of Sinkhorn's algorithm, which is due to Gurvits (2004) and has been extensively studied in (Georgiou and Pavon, 2015); see also the review paper (Idel, 2016). It is known to converge only in some cases but seems empirically to always work.

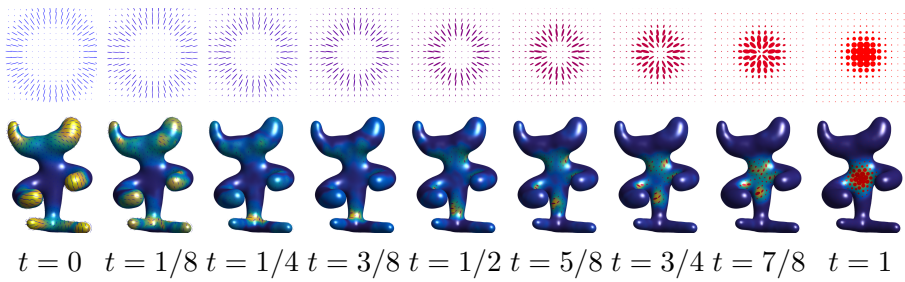

Figure 10.4: Interpolations between two input fields of positive semidefinite matrices (displayed at times $t \in \{0, 1\}$ using ellipses) on some domain (here, a 2-D planar square and a surface mesh), using the method detailed in Peyré *et al.* (2017). Unlike linear interpolation schemes, this OT-like method transports the "mass" of the tensors (size of the ellipses) as well as their anisotropy and orientation.

10.6 Gromov–Wasserstein Distances

For some applications such as shape matching, an important weakness of optimal transport distances lies in the fact that they are not invariant to important families of invariances, such as rescaling, translation or rotations. Although some nonconvex variants of OT to handle such global transformations have been proposed (Cohen and Guibas, 1999;

Pele and Taskar, 2013) and recently applied to problems such as cross-lingual word embeddings alignments (Grave *et al.*, 2019; Alvarez-Melis *et al.*, 2019; Grave *et al.*, 2019), these methods require specifying first a subset of invariances, possibly between different metric spaces, to be relevant. We describe in this section a more general and very natural extension of OT that can deal with measures defined on different spaces without requiring the definition of a family of invariances.

10.6.1 Hausdorff Distance

The Hausdorff distance between two sets $A, B \subset \mathcal{Z}$ for some metric $d_{\mathcal{Z}}$ is

$$\mathcal{H}_{\mathcal{Z}}(A,B) \overset{\text{def.}}{=} \max\Big(\sup_{a\in A}\inf_{b\in B} d_{\mathcal{Z}}(a,b), \sup_{b\in B}\inf_{a\in A} d_{\mathcal{Z}}(a,b)\Big),$$

see Figure 10.5. This defines a distance between compact sets $\mathcal{K}(\mathcal{Z})$ of $\mathcal{Z}$, and if $\mathcal{Z}$ is compact, then $(\mathcal{K}(\mathcal{Z}), \mathcal{H}_{\mathcal{Z}})$ is itself compact; see (Burago *et al.*, 2001).

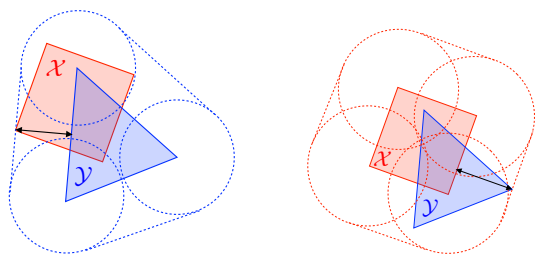

Figure 10.5: Computation of the Hausdorff distance in $\mathbb{R}^2$.

Following Mémoli (2011), one remarks that this distance between sets (A, B) can be defined similarly to the Wasserstein distance between measures (which should be somehow understood as "weighted" sets). One replaces the measures couplings (2.14) by sets couplings

$$\mathcal{R}(A,B) \overset{\text{def.}}{=} \left\{ R \in \mathcal{X}\times\mathcal{Y} \;:\; \begin{array}{l} \forall\, a\in A, \exists b\in B, (a,b)\in R \\ \forall\, b\in B, \exists a\in A, (a,b)\in R \end{array} \right\}.$$

With respect to Kantorovich problem (2.15), one should replace integration (since one does not have access to measures) by maximization,

and one has

$$\mathcal{H}_{\mathcal{Z}}(A,B) = \inf_{R \in \mathcal{R}(A,B)} \sup_{(a,b)\in R} d(a,b). \tag{10.23}$$

Note that the support of a measure coupling $\pi \in \mathcal{U}(\alpha,\beta)$ is a set coupling between the supports, *i.e.* $\mathrm{Supp}(\pi) \in \mathcal{R}(\mathrm{Supp}(\alpha), \mathrm{Supp}(\beta))$. The Hausdorff distance is thus connected to the ∞-Wasserstein distance (see Remark 2.20) and one has $\mathcal{H}(A,B) \leq \mathcal{W}_\infty(\alpha,\beta)$ for any measure (α,β) whose supports are (A,B).

10.6.2 Gromov–Hausdorff distance

The Gromov–Hausdorff (GH) distance (Gromov, 2001) (see also (Edwards, 1975)) is a way to measure the distance between two metric spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ by quantifying how far they are from being isometric to each other, see Figure 10.6. It is defined as the minimum Hausdorff distance between every possible isometric embedding of the two spaces in a third one,

$$\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) \overset{\text{def.}}{=} \inf_{\mathcal{Z},f,g} \left\{ \mathcal{H}_{\mathcal{Z}}(f(\mathcal{X}), g(\mathcal{Y})) \;:\; \begin{array}{c} f : \mathcal{X} \overset{\text{isom}}{\longrightarrow} \mathcal{Z} \\ g : \mathcal{Y} \overset{\text{isom}}{\longrightarrow} \mathcal{Z} \end{array} \right\}.$$

Here, the constraint is that f must be an isometric embedding, meaning that $d_{\mathcal{Z}}(f(x), f(x')) = d_{\mathcal{X}}(x,x')$ for any $(x,x') \in \mathcal{X}^2$ (similarly for g). One can show that $\mathcal{GH}$ defines a distance between compact metric spaces up to isometries, so that in particular $\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) = 0$ if and only if there exists an isometry $h : \mathcal{X} \to \mathcal{Y}$, *i.e.* h is bijective and $d_{\mathcal{Y}}(h(x), h(x')) = d_{\mathcal{X}}(x,x')$ for any $(x,x') \in \mathcal{X}^2$.

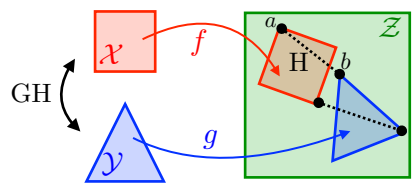

Figure 10.6: The GH approach to compare two metric spaces.

Similarly to (10.23) and as explained in (Mémoli, 2011), it is possible

to rewrite equivalently the GH distance using couplings as follows:

$$\mathcal{GH}(d_{\mathcal{X}}, d_{\mathcal{Y}}) = \frac{1}{2} \inf_{R \in \mathcal{R}(\mathcal{X},\mathcal{Y})} \sup_{((x,y),(x',y')) \in R^2} |d_{\mathcal{X}}(x,x') - d_{\mathcal{X}}(y,y')|.$$

For discrete spaces $\mathcal{X} = (x_i)_{i=1}^n, \mathcal{Y} = (y_j)_{j=1}^m$ represented using a distance matrix $\mathbf{D} = (d_{\mathcal{X}}(x_i, x_{i'}))_{i,i'} \in \mathbb{R}^{n \times n}$, $\mathbf{D}' = (d_{\mathcal{Y}}(y_j, y_{j'}))_{j,j'} \in \mathbb{R}^{m \times m}$, one can rewrite this optimization using binary matrices $\mathbf{R} \in \{0,1\}^{n \times m}$ indicating the support of the set couplings R as follows:

$$\mathrm{GH}(\mathbf{D}, \mathbf{D}') = \frac{1}{2} \inf_{\mathbf{R}\mathbb{1}>0, \mathbf{R}^\top \mathbb{1}>0} \max_{(i,i',j,j')} \mathbf{R}_{i,j}\mathbf{R}_{j,j'}|\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|. \quad (10.24)$$

The initial motivation of the GH distance is to define and study limits of metric spaces, as illustrated in Figure 10.7, and we refer to (Burago *et al.*, 2001) for details. There is an explicit description of the geodesics for the GH distance (Chowdhury and Mémoli, 2016), which is very similar to the one in Gromov–Wasserstein spaces, detailed in Remark 10.8.

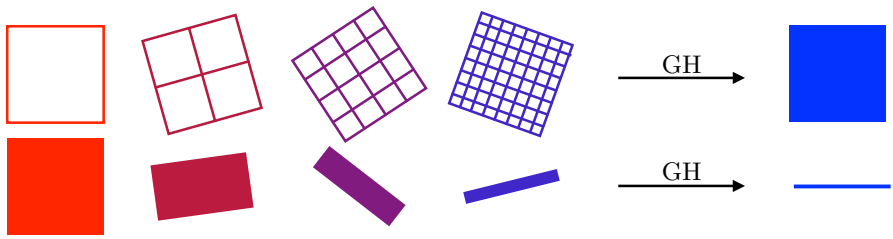

Figure 10.7: GH limit of sequences of metric spaces.

The underlying optimization problem (10.24) is highly nonconvex, and computing the global minimum is untractable. It has been approached numerically using approximation schemes and has found applications in vision and graphics for shape matching (Mémoli and Sapiro, 2005; Bronstein *et al.*, 2006).

It is often desirable to "smooth" the definition of the Hausdorff distance by replacing the maximization by an integration. This in turn necessitates the introduction of measures, and it is one of the motivations for the definition of the GW distance in the next section.

10.6.3 Gromov–Wasserstein Distance

Optimal transport needs a ground cost $\mathbf{C}$ to compare histograms $(\mathbf{a}, \mathbf{b})$ and thus cannot be used if the bins of those histograms are not defined on the same underlying space, or if one cannot preregister these spaces to define a ground cost between any pair of bins in the first and second histograms, respectively. To address this limitation, one can instead only assume a weaker assumption, namely that two matrices $\mathbf{D} \in \mathbb{R}^{n \times n}$ and $\mathbf{D}' \in \mathbb{R}^{m \times m}$ quantify similarity relationships between the points on which the histograms are defined. A typical scenario is when these matrices are (power of) distance matrices. The GW problem reads

$$\mathrm{GW}((\mathbf{a}, \mathbf{D}), (\mathbf{b}, \mathbf{D}'))^2 \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}) \tag{10.25}$$

$$\text{where} \quad \mathcal{E}_{\mathbf{D}, \mathbf{D}'}(\mathbf{P}) \overset{\text{def.}}{=} \sum_{i,j,i',j'} |\mathbf{D}_{i,i'} - \mathbf{D}'_{j,j'}|^2 \mathbf{P}_{i,j} \mathbf{P}_{i',j'},$$

see Figure 10.8. This problem is similar to the GH problem (10.24) when replacing maximization by a sum and set couplings by measure couplings. This is a nonconvex problem, which can be recast as a quadratic assignment problem (LMB07) and is in full generality NP-hard to solve for arbitrary inputs. It is in fact equivalent to a graph matching problem (Lyzinski *et al.*, 2016) for a particular cost.

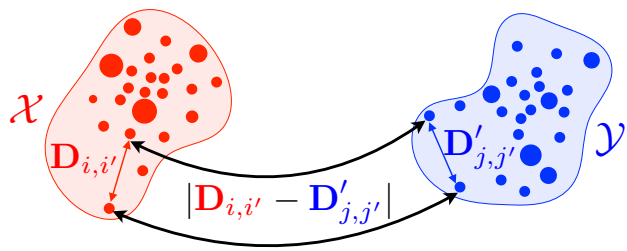

Figure 10.8: The GW approach to comparing two metric measure spaces.

One can show that GW satisfies the triangular inequality, and in fact it defines a distance between metric spaces equipped with a probability distribution, here assumed to be discrete in definition (10.25), up to isometries preserving the measures. This distance was introduced and studied in detail by Mémoli (2011). An in-depth mathematical

exposition (in particular, its geodesic structure and gradient flows) is given in (Sturm, 2012). See also (Schmitzer and Schnörr, 2013a) for applications in computer vision. This distance is also tightly connected with the GH distance (Gromov, 2001) between metric spaces, which have been used for shape matching (Mémoli, 2007; Bronstein *et al.*, 2010).

Remark 10.7 (Gromov–Wasserstein distance)**.** The general setting corresponds to computing couplings between metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \alpha_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \alpha_{\mathcal{Y}})$, where $(d_{\mathcal{X}}, d_{\mathcal{Y}})$ are distances, while $\alpha_{\mathcal{X}}$ and $\alpha_{\mathcal{Y}}$ are measures on their respective spaces. One defines

$$\mathcal{GW}((\alpha_{\mathcal{X}}, d_{\mathcal{X}}), (\alpha_{\mathcal{Y}}, d_{\mathcal{Y}}))^2 \stackrel{\text{def.}}{=} \min_{\pi \in \mathcal{U}(\alpha_{\mathcal{X}}, \alpha_{\mathcal{Y}})} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^2 \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y'). \tag{10.26}$$

$\mathcal{GW}$ defines a distance between metric measure spaces up to isometries, where one says that $(\mathcal{X}, \alpha_{\mathcal{X}}, d_{\mathcal{X}})$ and $(\mathcal{Y}, \alpha_{\mathcal{Y}}, d_{\mathcal{Y}})$ are isometric if there exists a bijection $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\varphi_\sharp \alpha_{\mathcal{X}} = \alpha_{\mathcal{Y}}$ and $d_{\mathcal{Y}}(\varphi(x), \varphi(x')) = d_{\mathcal{X}}(x, x')$.

Remark 10.8 (Gromov–Wasserstein geodesics)**.** The space of metric spaces (up to isometries) endowed with this $\mathcal{GW}$ distance (10.26) has a geodesic structure. Sturm (2012) shows that the geodesic between $(\mathcal{X}_0, d_{\mathcal{X}_0}, \alpha_0)$ and $(\mathcal{X}_1, d_{\mathcal{X}_1}, \alpha_1)$ can be chosen to be $t \in [0,1] \mapsto (\mathcal{X}_0 \times \mathcal{X}_1, d_t, \pi^\star)$, where $\pi^\star$ is a solution of (10.26) and for all $((x_0, x_1), (x_0', x_1')) \in (\mathcal{X}_0 \times \mathcal{X}_1)^2$,

$$d_t((x_0, x_1), (x_0', x_1')) \stackrel{\text{def.}}{=} (1-t) d_{\mathcal{X}_0}(x_0, x_0') + t d_{\mathcal{X}_1}(x_1, x_1').$$

This formula allows one to define and analyze gradient flows which minimize functionals involving metric spaces; see Sturm (2012). It is, however, difficult to handle numerically, because it involves computations over the product space $\mathcal{X}_0 \times \mathcal{X}_1$. A heuristic approach is used in (Peyré *et al.*, 2016) to define geodesics and barycenters of metric measure spaces while imposing the cardinality of the

involved spaces and making use of the entropic smoothing (10.27) detailed below.

10.6.4 Entropic Regularization

To approximate the computation of GW, and to help convergence of minimization schemes to better minima, one can consider the entropic regularized variant

$$\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}) - \varepsilon \mathbf{H}(\mathbf{P}). \tag{10.27}$$

As proposed initially in (Gold and Rangarajan, 1996; Rangarajan *et al.*, 1999), and later revisited in (Solomon *et al.*, 2016a) for applications in graphics, one can use iteratively Sinkhorn's algorithm to progressively compute a stationary point of (10.27). Indeed, successive linearizations of the objective function lead to consider the succession of updates

$$\mathbf{P}^{(\ell+1)} \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C}^{(\ell)} \rangle - \varepsilon H(\mathbf{P}) \quad \text{where} \tag{10.28}$$

$$\mathbf{C}^{(\ell)} \overset{\text{def.}}{=} \nabla \mathcal{E}_{\mathbf{D},\mathbf{D}'}(\mathbf{P}^{(\ell)}) = -\mathbf{D}\mathbf{P}^{(\ell)}\mathbf{D}',$$

which can be interpreted as a mirror-descent scheme (Solomon *et al.*, 2016a). Each update can thus be solved using Sinkhorn iterations (4.15) with cost $\mathbf{C}^{(\ell)}$. Figure 10.9 displays the evolution of the algorithm. Figure 10.10 illustrates the use of this entropic GW to compute soft maps between domains.

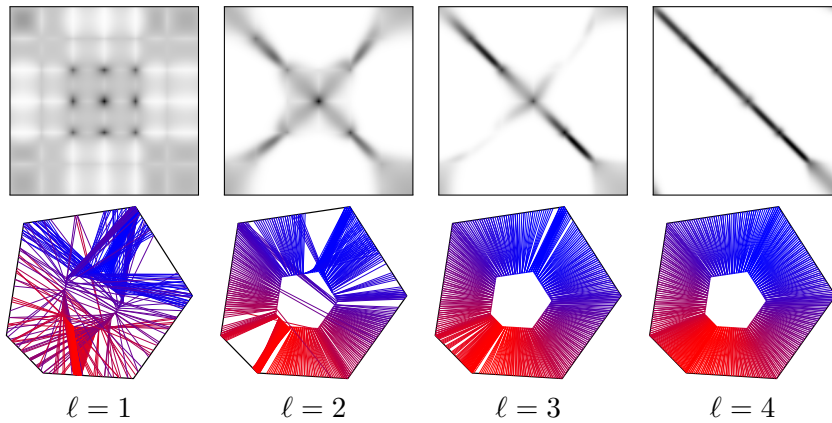

Figure 10.9: Iterations of the entropic GW algorithm (10.28) between two shapes $(x_i)_i$ and $(y_j)_j$ in $\mathbb{R}^2$, initialized with $\mathbf{P}^{(0)} = \mathbf{a} \otimes \mathbf{b}$. The distance matrices are $\mathbf{D}_{i,i'} = \|x_i - x_{i'}\|$ and $\mathbf{D}'_{j,j'} = \|y_j - y_{j'}\|$. Top row: coupling $\mathbf{P}^{(\ell)}$ displayed as a 2-D image. Bottom row: matching induced by $\mathbf{P}^{(\ell)}$ (each point x_i is connected to the three y_j with the three largest values among $\{\mathbf{P}^{(\ell)}_{i,j}\}_j$). The shapes have the same size, but for display purposes, the inner shape $(x_i)_i$ has been reduced.

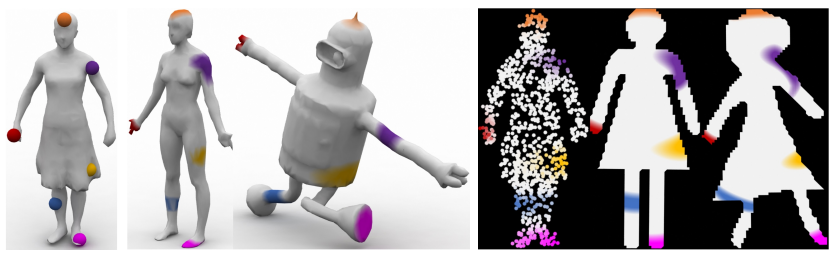

Figure 10.10: Example of fuzzy correspondences computed by solving GW problem (10.27) with Sinkhorn iterations (10.28). Extracted from (Solomon *et al.*, 2016a).

Acknowledgements

We would like to thank the many colleagues, collaborators and students who have helped us at various stages when preparing this survey. Some of their inputs have shaped this work, and we would like to thank in particular Jean-David Benamou, Yann Brenier, Guillaume Carlier, Vincent Duval and the entire MOKAPLAN team at Inria; Francis Bach, Espen Bernton, Mathieu Blondel, Nicolas Courty, Rémi Flamary, Alexandre Gramfort, Young-Heon Kim, Daniel Matthes, Philippe Rigollet, Filippo Santambrogio, Justin Solomon, Jonathan Weed; as well as the feedback by our current and former students on these subjects, in particular Gwendoline de Bie, Lénaïc Chizat, Aude Genevay, Hicham Janati, Théo Lacombe, Boris Muzellec, Francois-Pierre Paty, Vivien Seguy.