Self-study: Variational Wasserstein Problems

Tianpei Xie

Aug. 19th., 2022

Contents

1 Previous lessons

- Recall the definition of Wasserstein distance between two measures $\alpha, \beta \in \mathcal{M}(\mathcal{X})$, where $\mathcal{M}(\mathcal{X})$ is the space of Radon measure on $\mathcal{X}$:

  Definition We suppose $\mathcal{X} = \mathcal{Y}$ and that for some $p \geq 1$, $c(x, y) = d(x, y)^p$, where d is a distance on $\mathcal{X}$, i.e.

  1. $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$;
  2. $d(x, y) = 0$ iff $x = y$;
  3. $d(x, y) \leq d(x, z) + d(z, y)$, for all $x, y, z \in \mathcal{X}$

  Then the p-***Wasserstein distance*** between $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$ on $\mathcal{X}$ is defined by $\mathcal{W}_p(\alpha, \beta) := \mathcal{L}_{d^p}(\alpha, \beta)^{\frac{1}{p}}$, where $\mathcal{X} = \mathcal{Y}$ and

  $$\begin{aligned} \mathcal{L}_{d^p}(\alpha, \beta) := \min_{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})} & \int_{\mathcal{X} \times \mathcal{Y}} d^p(x, y) d\pi(x, y) \\ \text{s.t. } & P_{\mathcal{X}\#}\pi = \alpha, \\ & P_{\mathcal{Y}\#}\pi = \beta. \end{aligned}$$

  Here $P_{\mathcal{X}\#}$ and $P_{\mathcal{Y}\#}$ are the ***push-forwards*** of the **projections** $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$.

- For two random variables X, Y, **the p-Wasserstein distance between their distributions** $\mathcal{W}_p(P_X, P_Y) := \left(\min_\pi \mathbb{E}_{(X,Y)\sim\pi}\left[d^p(X, Y)\right]\right)^{\frac{1}{p}}$.

- For discrete measures, $\alpha = \sum_{i=1}^n a_i \delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^m b_i \delta_{\boldsymbol{y}_i}$, the primal problem for optimal transport is

  $$\min_{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}} \langle \boldsymbol{P}, \boldsymbol{C} \rangle = \sum_{i,j} C_{i,j} P_{i,j} \tag{1}$$

  $$\text{s.t. } \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a} \tag{2}$$

  $$\boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b} \tag{3}$$

  $$P_{i,j} \geq 0$$

  where $\boldsymbol{C}_{n,m} := [C_{i,j}]_{i \in [1:n], j \in [1:m]}$, $C_{i,j} := c(\boldsymbol{x}_i, \boldsymbol{y}_j) \geq 0$. The feasible set is defined as

  $$U(\boldsymbol{a}, \boldsymbol{b}) := \left\{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a},\ \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b}\right\} \tag{4}$$

- the corresponding dual problem with respect to primal problem is

  $$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m} \langle \boldsymbol{\lambda}, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle \tag{5}$$

  $$\text{s.t. } \lambda_i + \mu_j \leq C_{i,j} \quad \forall\, i \in [1:n], j \in [1:m] \tag{6}$$

  where $\boldsymbol{\lambda} = [\lambda_i]_n$, $\boldsymbol{\mu} = [\mu_j]_m$ are **dual variables** (slack variables) for marginal distribution constrain $\boldsymbol{a}$ and $\boldsymbol{b}$. We denote $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} := \boldsymbol{\lambda}\mathbf{1}_m^T + \mathbf{1}_n \boldsymbol{\mu}^T \in \mathbb{R}^{n \times m}$ so that the linear constraints is $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C}$. Such dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are often referred to as "***Kantorovich potentials***." The feasible set of the dual problem is defined as

  $$R(\boldsymbol{C}) := \{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m : \boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C}\} \tag{7}$$

where $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} = \boldsymbol{\lambda}\mathbf{1}_m + \mathbf{1}_n\boldsymbol{\mu}^T$.

- By strong duality:

$$\begin{aligned}\mathcal{W}_p(\alpha,\beta)^p = \mathcal{L}_{d^p}(\alpha,\beta) &:= L_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) \\ &= \min_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}\,,\,\boldsymbol{C}\rangle \\ &= \max_{\boldsymbol{\lambda}\oplus\boldsymbol{\mu}\leq\boldsymbol{C}} \langle \boldsymbol{\lambda}\,,\,\boldsymbol{a}\rangle + \langle \boldsymbol{\mu}\,,\,\boldsymbol{b}\rangle \end{aligned} \tag{8}$$

or at optimal solution

$$\begin{aligned}\mathcal{W}_p(\alpha,\beta)^p &= \langle \boldsymbol{P}^*\,,\,\boldsymbol{C}\rangle \\ &= \langle \boldsymbol{\lambda}^*\,,\,\boldsymbol{a}\rangle + \langle \boldsymbol{\mu}^*\,,\,\boldsymbol{b}\rangle \end{aligned} \tag{9}$$

where $\boldsymbol{P}^*$ is the optimal coupling matrix for primal problem and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ are optimal Kantorovich potentials or optimal dual solutions that satisfies

$$\boldsymbol{\lambda}^* \oplus \boldsymbol{\mu}^* = \boldsymbol{C} = [\boldsymbol{D}_{i,j}^p] \tag{10}$$

That is both primal objective function and dual objective function can be used to estimate the Wasserstein distance.

- We also have the **maximum entropy optimal transport problem**

$$L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) = \min_{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}\,,\,\boldsymbol{C}\rangle - \epsilon H(\boldsymbol{P}) \tag{11}$$

where the second term is entropy

$$H(\boldsymbol{P}) := -\sum_{i,j} P_{i,j}\left(\log(P_{i,j}) - 1\right) \tag{12}$$

This problem has a unique optimal solution

$$\boldsymbol{P}^* = \operatorname{diag}(\boldsymbol{u})\,\boldsymbol{K}\operatorname{diag}(\boldsymbol{v}) \tag{13}$$

where $\boldsymbol{u} = [\exp(\lambda_i/\epsilon)] = \exp(\boldsymbol{\lambda}/\epsilon)$ and $\boldsymbol{v} = [\exp(\mu_j/\epsilon)] = \exp(\boldsymbol{\mu}/\epsilon)$.

- The dual problem of the **maximum entropy optimal transport problem**:

$$L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) = \max_{\boldsymbol{\lambda}\in\mathbb{R}^n,\boldsymbol{\mu}\in\mathbb{R}^m} \langle \boldsymbol{\lambda}\,,\,\boldsymbol{a}\rangle + \langle \boldsymbol{\mu}\,,\,\boldsymbol{b}\rangle - \epsilon\left\langle \exp(\boldsymbol{\lambda}/\epsilon)\,,\,\boldsymbol{K}\exp(\boldsymbol{\mu}/\epsilon)\right\rangle \tag{14}$$

where $\boldsymbol{K} = \exp(-\boldsymbol{C}/\epsilon)$ is the Gibbs distribution.

- The **probability interpretation** of orignal primal and dual Kantorovich optimal transport problem:

$$\begin{aligned}(P)\quad \mathcal{L}_c(\alpha,\beta) = \min_{(X,Y)\sim\pi} &\mathbb{E}_{(X,Y)}\left[c(X,Y)\right] \\ \text{s.t. } &X\sim\alpha, \\ &Y\sim\beta\end{aligned} \tag{15}$$

$$\begin{aligned}(D)\quad \mathcal{L}_c(\alpha,\beta) = \max_{(\lambda,\mu)\in\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} &\mathbb{E}_{X\sim\alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y\sim\beta}\left[\mu(Y)\right] \\ \text{s.t. } &\lambda(x)+\mu(y)\leq c(x,y), \quad \forall x\in\mathcal{X}, y\in\mathcal{Y},\end{aligned} \tag{16}$$

- The **probability interpretation** of primal and dual maximum entropy optimal transport problem:

$$(P) \quad \mathcal{L}^\epsilon(\alpha,\beta) := \min_{(X,Y)\sim\pi} \mathbb{E}_{(X,Y)}\left[c(X,Y)\right] + \epsilon\, I(X;Y) \tag{17}$$
$$\text{s.t. } X \sim \alpha$$
$$Y \sim \beta$$

where $I(X;Y) := \mathbb{KL}\left(\pi \parallel \alpha \otimes \beta\right)$ is the mutual information between X and Y.

$$(D) \quad \mathcal{L}^\epsilon(\alpha,\beta) := \sup_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \mathbb{E}_{X\sim\alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y\sim\beta}\left[\mu(Y)\right]$$
$$- \epsilon\mathbb{E}_{X\sim\alpha,Y\sim\beta}\left[\exp\left(\frac{-c(X,Y)+\lambda(X)+\mu(Y)}{\epsilon}\right)\right]. \tag{18}$$

- Given a cost function $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$, $f : \mathcal{X} \to \mathbb{R}$, the ***c*-transform** of f is defined as

$$f^c(y) := \inf_{x\in\mathcal{X}} c(x,y) - f(x) \tag{19}$$

The function $f^c : \mathcal{Y} \to \mathbb{R}$ is also called the ***c*-conjugate function** of f. Similarly, $g : \mathcal{Y} \to \mathbb{R}$, the ***c̄*-transform** of g is defined as

$$g^{\bar{c}}(x) := \inf_{y\in\mathcal{Y}} c(x,y) - g(y) \tag{20}$$

A function $\psi : \mathcal{X} \to \mathbb{R}$ is ***c*-concave** if there exists some function $\phi : \mathcal{Y} \to \mathbb{R}$ and cost function $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ so that ψ is the $\bar{c}$-transform of ϕ, i.e. $\psi = \phi^{\bar{c}}$. Denote ψ as c-concave($\mathcal{X}$).

A function $\phi : \mathcal{Y} \to \mathbb{R}$ is ***c̄*-concave** if there exists some function $\psi : \mathcal{X} \to \mathbb{R}$ and cost function $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ so that ϕ is the c-transform of ψ, i.e. $\phi = \psi^c$. Denote ϕ as $\bar{c}$-concave($\mathcal{Y}$)

For distance $c = d$, $f^c = f^{\bar{c}}$, thus we drop their distinctions.

Proposition 1.1 *If $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a distance, then the function $f : \mathcal{X} \to \mathbb{R}$ is **c-concave** if and only if f is **Lipschitz continuous** with Lipschitz constant less than 1 w.r.t. the distance c. We will denote by Lip_1 the set of these functions. Moreover, for every $f \in Lip_1$, i.e. $\|f\|_L \le 1$, we have the c-transform of f, $f^c = -f$. [Santambrogio, 2015]*

- Thus the dual problem (5) is equivalent to an **unconstrained optimization problem**

$$\mathcal{L}_c(\alpha,\beta) := \max_{\lambda\in\mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} \lambda d\alpha + \int_{\mathcal{Y}} \lambda^c d\beta \tag{21}$$
$$= \max_{\mu\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \mu^{\bar{c}} d\alpha + \int_{\mathcal{Y}} \mu d\beta \tag{22}$$

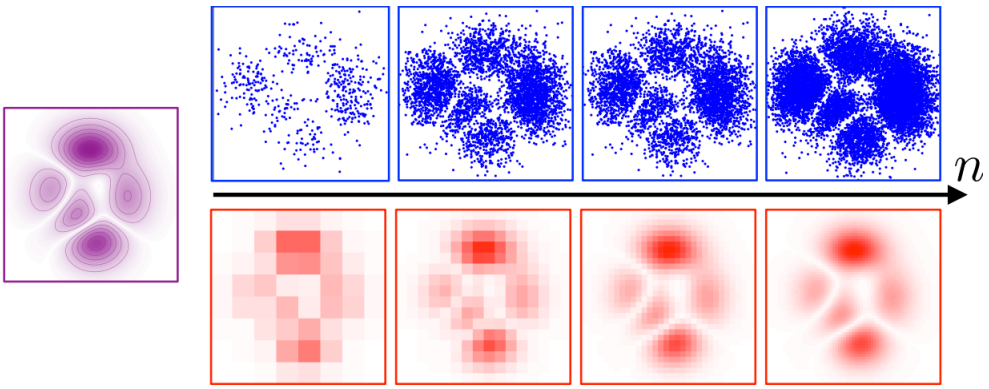

Figure 9.1: Increasing fine discretization of a continuous distribution having a density (violet, left) using a Lagrangian representation $\frac{1}{n}\sum_i \delta_{x_i}$ (blue, top) and an Eulerian representation $\sum_i \mathbf{a}_i \delta_{x_i}$ with x_i representing cells on a grid of increasing size (red, bottom). The Eulerian perspective starts from a pixelated image down to one with such fine resolution that it almost matches the original density. Weights $\mathbf{a}_i$ are directly proportional to each pixel-cell's intensity.

Figure 1: The comparison between Eulerian Discretization (down) and Lagrangian Discretization (top) [Peyr and Cuturi, 2019]

2 Differentiating the Wasserstein Loss

The optimal transport geometry has a unique ability, not shared with other information divergences, to leverage physical ideas (mass displacement) and geometry (a ground cost between observations or bins) to compare measures. These two facts combined make it thus very tempting to use the Wasserstein distance as a loss function.

However, the main technical challenge associated with that idea lies in approximating and **differentiating** efficiently the Wasserstein distance. We start this section by presenting methods to approximate such gradients, then follow with three important applications that can be cast as variational Wasserstein problems.

We consider a **statisical estimation problem**: given a probability distribution β arising from measurements, we need to find a model from a parameterized family of distributions $\{\alpha_\theta, \theta \in \Theta\}$ that **minimizes** the ***Wasserstein distance***, where Θ is a subset of a Euclidean space.

$$\min_{\theta\in\Theta} \mathcal{L}_c(\alpha_\theta, \beta) \tag{23}$$

where $\mathcal{L}_c(\alpha_\theta, \beta)$ is defined as the optimal value of (15). This problem in general is **non-convex**. For cases such as Gaussian measures or elliptically contoured distributions, the **Wasserstein distance based objective** function has closed form, which can be solved directly.

In most cases, however, one has to resort to a careful **discretization** of α_θ to compute a **local minimizer** for Problem (23). Two approaches can be envisioned: Eulerian or Lagrangian. Note that the discrete measure is $\alpha = \sum_i^n a_i \delta_{\boldsymbol{x}_i}$, which depends on two elements: the **position** of mass $\boldsymbol{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and the **probability measure** at each position $\boldsymbol{a} := \{a_1, \ldots, a_n\}$. Figure 1 compares the Eulerian discretization and Lagrangian discretization.

2.1 Eulerian and Lagrangian discretization

- **Eulerian Discretization** (when the **measure** of mass changes but the position of mass does not)
  A first way to discretize the problem is to suppose that both distributions $\alpha = \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i}$ and $\beta = \sum_{i=1}^{m} b_i \delta_{\boldsymbol{y}_i}$ are discrete distributions defined on fixed locations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. The parameterized measure α_θ is in that case entirely represented through the weight vector $a : \theta \to a(\theta) \in \Delta_n$, which, in practice, might be very sparse if the grid is large. This setting corresponds to the so-called class of **Eulerian discretization methods**.

  In order to obtain differentiable objective function, we use the maximum entropy OT (11). The statistical estimation problem becomes

$$\min_{\boldsymbol{\theta} \in \Theta} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}(\boldsymbol{\theta}), \boldsymbol{b}) := \mathcal{E}_E(\boldsymbol{\theta}) \tag{24}$$

  where $C_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{y}_j)$.

  Recall that the maximum entropy OT objective is **differentiable** and **convex** w.r.t. input historgrams, with gradient

$$\nabla_{(\boldsymbol{a},\boldsymbol{b})} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \begin{bmatrix} \boldsymbol{\lambda}_* \\ \boldsymbol{\mu}_* \end{bmatrix} \tag{25}$$

  where $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ are the **dual potentials** i.e. the optimal solutions of regularized dual problem (14) chosen so that $\sum_i \lambda_i = \sum_j \mu_j = 0$. The zero mean condition on $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ is important when using gradient descent to guarantee conservation of mass.

  By chain rule, we have

$$\nabla_{\boldsymbol{\theta}} \mathcal{E}_E(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{a}(\boldsymbol{\theta})^T \boldsymbol{\lambda}_* \tag{26}$$

  where $\nabla_{\boldsymbol{\theta}} \boldsymbol{a}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times d_\theta}$ is the Jacobian (differential) of the map $\boldsymbol{a}(\boldsymbol{\theta})$ and $d_\theta = \dim(\boldsymbol{\theta})$.

  The problem (24) is a convex optimization with differentiable gradient. It can be solved directly.

- **Lagrangian Discretization** (when the **position** of mass changes)
  A different approach consists in using instead fixed (typically uniform) weights and approximating an input measure α as an **empirical measure** $\alpha_{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{x}_i(\boldsymbol{\theta})}$ for a point-cloud parameterization map $\boldsymbol{X} : \boldsymbol{\theta} \to \boldsymbol{X}(\boldsymbol{\theta}) = (\boldsymbol{x}(\boldsymbol{\theta})_i)$, where $\boldsymbol{x}(\boldsymbol{\theta})_i \in \mathcal{X}$ for $i = 1, \ldots, n$. We assume here that $\mathcal{X}$ is Euclidean.

  Note that the measure changes when the **position** of mass changes, which will affect the cost function $C_{i,j}$ as well. The optimization becomes

$$\min_{\boldsymbol{\theta} \in \Theta} L^{\epsilon}_{\boldsymbol{C}(\boldsymbol{\theta})} \left(\frac{1}{n} \mathbf{1}_n, \boldsymbol{b} \right) := \mathcal{E}_L(\boldsymbol{\theta}) \tag{27}$$

  where $\boldsymbol{C}(\boldsymbol{\theta}) := [c(\boldsymbol{x}_i(\boldsymbol{\theta}), \boldsymbol{y}_j)]_{i,j}$. Note that here the cost matrix $\boldsymbol{C}(\boldsymbol{\theta})$ now depends on $\boldsymbol{\theta}$ since the support of $\alpha_{\boldsymbol{\theta}}$ changes with $\boldsymbol{\theta}$.

The objective function $L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$ can be represented as $L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \langle \boldsymbol{P}^* , \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}^*)$. We can show that the function $g : \boldsymbol{C} \to g(\boldsymbol{C}) := L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$ is a **concave** and **smooth** function. The gradient of g is

$$\nabla_{\boldsymbol{C}} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{P}^* \tag{28}$$

where $\boldsymbol{P}^* = \operatorname{diag}(\boldsymbol{u}) \exp(-\boldsymbol{C}/\epsilon) \operatorname{diag}(\boldsymbol{v})$ is the optimal primal solution (11) and it is also a function of $\boldsymbol{C}$.

By chain rule, we have

$$\nabla_{\boldsymbol{\theta}} \mathcal{E}_L(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{X}(\boldsymbol{\theta})^T (\nabla F(\boldsymbol{X}(\boldsymbol{\theta}))) \tag{29}$$

where $\boldsymbol{X}(\boldsymbol{\theta}) := (\boldsymbol{x}(\boldsymbol{\theta})_i) \in \mathbb{R}^{nd \times 1}$ and $\nabla_{\boldsymbol{\theta}} \boldsymbol{X}(\boldsymbol{\theta}) \in \mathbb{R}^{nd \times d_\theta}$ is the Jacobian matrix. Here we define $F : \boldsymbol{X} \to L^{\epsilon}_{\boldsymbol{C}(\boldsymbol{X})}\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{b}\right)$. Assuming $(\mathcal{X}, \mathcal{Y})$ are convex subsets of $\mathbb{R}^d$, by chain rule we have

$$\begin{aligned} \nabla F(\boldsymbol{X}) &= \left[\nabla_{\boldsymbol{C}} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})^T \nabla_{\boldsymbol{x}_i} C \right]_i \\ &= \left[\sum_j^m P_{i,j,*} \nabla_{\boldsymbol{x}_i} c(\boldsymbol{x}_i, \boldsymbol{y}_j) \right]_{i=1}^n \in \mathcal{X}^n \end{aligned} \tag{30}$$

For instance, for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, for $c(s,t) = \|s-t\|_2^2$, one has

$$\nabla F(\boldsymbol{X}) = 2 \left[\frac{1}{n} \boldsymbol{x}_i - \sum_j^m P_{i,j,*} \boldsymbol{y}_j \right]_{i=1}^n \in \mathcal{X}^n \subset \mathbb{R}^{nd} \tag{31}$$

Note that, up to a constant, this gradient is $\mathrm{Id} - T$, where T is the **barycentric projection**.

2.2 Automatic differentiation

Note that for both Eulerian and Lagrangian discretization, the gradient in (26) and (29) requires knowledge on the exact optimal solution $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ on dual problem and $\boldsymbol{P}^*$ on primal problem This can only be achieved with acceptable precision using a very large number of **Sinkhorn iterates**.

Given the limitation on time and computational resources, rather than approximating the gradient using the value obtained at a given iterate, it is usually better to *differentiate* directly the *output of Sinkhorns algorithm*, using ***reverse mode automatic differentiation***. In this case, we use the ***Sinkhorn dual divergence*** $\mathfrak{D}^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) := \langle \boldsymbol{\lambda}_* , \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}_* , \boldsymbol{b} \rangle$, which incorporates the entropy of the regularized optimal transport, and differentiating it directly as a composition of simple maps using the inputs, either the histogram in the Eulerian case or the cost matrix in the Lagrangian cases.

Thus the objective for Eulerian and Lagrangian discretization becomes

$$\begin{aligned} (E): &\quad \mathfrak{D}^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}(\boldsymbol{\theta}), \boldsymbol{b}) \\ (L): &\quad \mathfrak{D}^{\epsilon}_{\boldsymbol{C}(\boldsymbol{\theta})}\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{b}\right) \end{aligned}$$

The cost for computing the **gradient** of functionals involving Sinkhorn dual divergences is *the same as* that of computation of the **functional** itself.

3 Wasserstein Barycenters and Clustering

3.1 Discrete measures

Given input histogram $\{\boldsymbol{b}_s\}_{s=1}^S$, where $\boldsymbol{b}_s \in \Delta_{n_s}$, and weights $\boldsymbol{w} \in \Delta_S$, a Wasserstein barycenter is computed by minimizing

$$\min_{\boldsymbol{a}\in\Delta_n} \sum_{s=1}^S w_s L_{\boldsymbol{C}_s}(\boldsymbol{a}, \boldsymbol{b}_s) \tag{32}$$

where $\boldsymbol{C}_s \in \mathbb{R}^{n\times n_s}$ is the cost matrix between $\boldsymbol{a}$ and $\boldsymbol{b}_s$. Typically, we assume $n_s = n$ and $\boldsymbol{C}_s = \boldsymbol{C} = \boldsymbol{D}^p, \forall s$ as distance matrix, and the problem becomes

$$\min_{\boldsymbol{a}\in\Delta_n} \sum_{s=1}^S w_s \mathcal{W}_p^p(\boldsymbol{a}, \boldsymbol{b}_s)$$

For discrete measures, this barycenter problem (32) is in fact a *linear program*, since one can look for the S couplings $(\boldsymbol{P}_s)_s$ between each input and the barycenter itself, which by construction must be constrained to share the same row marginal,

$$\begin{aligned} \min_{\substack{\boldsymbol{a}\in\Delta_n,\\ \{\boldsymbol{P}_s\in\mathbb{R}_+^{n\times n_s},\forall s\}}} & \sum_{s=1}^S w_s \left\langle \boldsymbol{P}_s\,,\, \boldsymbol{C}_s \right\rangle \\ \text{s.t. } & \boldsymbol{P}_s \mathbf{1}_{n_s} = \boldsymbol{a}, \forall s = 1,\dots,S \\ & \boldsymbol{P}_s^T \mathbf{1}_n = \boldsymbol{b}_s, \forall s = 1,\dots,S \end{aligned}$$

This problem is a large-scale LP and can be solved using first order methods such as subgradient descent on the dual.

3.2 Arbitrary measures

Given input measuers $\{\beta_s\}_{s=1}^S$, where $\beta_s \in \mathcal{M}_+(\mathcal{X})$, and weights $\boldsymbol{w} \in \Delta_S$, a Wasserstein barycenter for arbitrary measures is

$$\min_{\alpha\in\mathcal{M}_+(\mathcal{X})} \sum_{s=1}^S w_s \mathcal{L}_c(\alpha, \beta_s) \tag{33}$$

In the case where $\mathcal{X} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2^2$, Agueh and Carlier show that if one of the input measures has a density, then this barycenter is **unique**. Moreover, under this condition, the **mean of the barycenter** α_* is necessarily the **barycenter of the mean**, i.e.

$$\int_{\mathcal{X}} x d\alpha_* = \sum_{s=1}^S w_s \int_{\mathcal{X}} x d\beta_s$$

and the support of α_* is located in the *convex hull* of the supports of the $\{\beta_s\}_{s=1}^S$.

Let us also note that it is possible to recast (33) as a **multimarginal OT** problem.

3.3 K-means as a Wasserstein variational problem

When $\beta_s = \beta$ for all s and $\mathcal{X} = \mathbb{R}^d$, $c(x, y) = \|x - y\|_2^2$, the barycenter problem becomes

$$\min_{\alpha \in \mathcal{M}_{k,1}(\mathcal{X})} \mathcal{L}_c(\alpha, \beta)$$

where $\alpha = \sum_{i=1}^k a_i \delta_{\boldsymbol{c}_i}$ is constrained to be a **discrete measure** with a finite support of **size up to** k. This barycenter problem is equivalent to the usual **k-means** problem taking β. Indeed, one can easily show that the ***centroids*** output $\{\boldsymbol{c}_i, i = 1, \ldots, k\}$ by the k-means problem correspond to the **support** of the solution α and that its ***weights*** a_i correspond to the **fraction** of points in β *assigned to each centroid* [Canas and Rosasco, 2012].

One can show that approximating $\mathcal{L} \approx \mathcal{L}^\epsilon$ using entropic regularization results in smoothed out assignments that appear in **soft-clustering** variants of k-means, such as *mixtures of Gaussians*.

3.4 Infinite mixtures and non-parametric Bayesian prior

It is possible to generalize (33) to a possibly **infinite collection of measures**. This problem is described by considering a probability distribution M over the space $\mathcal{M}_+^1(\mathcal{X})$ of probability distributions, i.e. $M \in \mathcal{M}_+^1(\mathcal{M}_+^1(\mathcal{X}))$. A **barycenter** is then a solution of

$$\min_{\alpha \in \mathcal{M}_+(\mathcal{X})} \mathbb{E}_{\beta \sim M} \left[\mathcal{L}_c(\alpha, \beta)\right] := \int_{\mathcal{M}_+^1(\mathcal{X})} \mathcal{L}_c(\alpha, \beta) dM(\beta). \tag{34}$$

where β is a *random measure* distributed according to M. M is a distribution over distributions, i.e. a ***non-parametric Bayesian*** **prior**. Then we can see that the problem (33) is a special case when $M = \sum_s w_s \delta_{\beta_s}$ is a discrete measure with finite support on space of measures.

Drawing S measures uniformly randomly would results in $\{\hat{\beta}_s\}_{s=1}^S$ with weight $w_i = \frac{1}{S}$. Define the unique solution of corresponding barycenter problem as $\hat{\beta}_S = \arg\min_\alpha \sum_{s=1}^S \mathcal{L}_c(\alpha, \hat{\beta}_s)/S$. It is shown that the barycenters $\hat{\beta}_S$ has consistency, i.e.

$$\mathcal{L}_c(\hat{\beta}_S, \alpha_*) \overset{S \to \infty}{\to} 0$$

where α_* is the optimal solution of infinite mixture barycenter problem (34). The convergence is in expectation or with high probability. This can be interpreted as a **law of large numbers** over the **Wasserstein space**.

3.5 Special cases

3.5.1 Barycenter of Gaussian measures is Gaussian

When all the measures are Gaussian $\beta_s = \mathcal{N}(\boldsymbol{m}_s, \boldsymbol{\Sigma}_s)$ for all s, and $c(x, y) = \|x - y\|_2^2$, the **barycenter** (33) has a closed form solution and itself is a **Gaussian measure** $\alpha_* = \mathcal{N}(\boldsymbol{m}_*, \boldsymbol{\Sigma}_*)$, where the **mean of barycenter** is the **barycenter of means** under Euclidean distance c

$$\boldsymbol{m}_* = \sum_{s=1}^S w_s \boldsymbol{m}_s$$

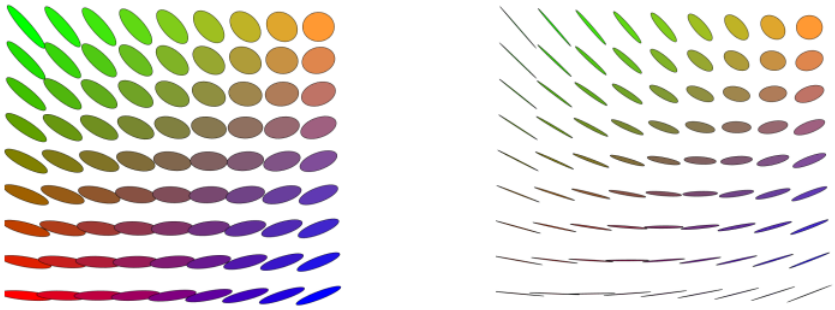

Figure 9.2: Barycenters between four Gaussian distributions in 2-D. Each Gaussian is displayed using an ellipse aligned with the principal axes of the covariance, and with elongations proportional to the corresponding eigenvalues.

Figure 2: The smooth interpolation of Gaussian measures.

and the **covariance** matrix of **barycenter** is the **barycenter of covariance matrices** under ***Bures metric***:

$$\mathbf{\Sigma}_* = \arg\min_{\mathbf{\Sigma}\in\mathcal{S}_+} \sum_{s=1}^{S} w_s \mathcal{B}\left(\mathbf{\Sigma}, \mathbf{\Sigma}_s\right)^2$$

This is a convex optimization problem and Bures metric is defined as

$$\mathcal{B}\left(\mathbf{\Sigma}_\alpha, \mathbf{\Sigma}_\beta\right)^2 = \mathrm{tr}\left(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2\left(\mathbf{\Sigma}_\alpha^{\frac{1}{2}}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}}\right).$$

$\mathbf{\Sigma}_*$ is the fix-point solution of the following map

$$\mathbf{\Sigma}_* = \Phi(\mathbf{\Sigma}_*)$$
$$\text{where } \Phi(\mathbf{\Sigma}_*) = \sum_{s=1}^{S} w_s \left(\mathbf{\Sigma}_*^{\frac{1}{2}}\mathbf{\Sigma}_s\mathbf{\Sigma}_*^{\frac{1}{2}}\right)^{\frac{1}{2}}$$

3.5.2 1-dimensional distributions on $\mathbb{R}$

For 1-D distributions, the $\mathcal{W}_p$ barycenter can be computed almost in closed form using the fact that the transport is the monotone rearrangement. For empirical distribution $\beta_s := \frac{1}{n}\sum_{i=1}^{n}\delta_{y_{i,s}}$, where the points are assumed to be ***ordered*** $y_{1,s} \le y_{2,s} \le \ldots \le y_{n,s}$. The barycenter α_w is also an empirical measure on n points

$$\alpha_w = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{i,w}}$$

where the mass location $x_{i,w}$ is the result of barycenter map A_w

$$x_{i,w} = A_w(y_{i,s}) := \arg\min_{x\in\mathbb{R}} \sum_{s=1}^{S} w_s \left\|x - y_{i,s}\right\|_p^p; \quad \forall\, i = 1, \ldots, S.$$

For instance, for $p = 2$, one has $x_{i,w} = \sum_{s=1}^{S} w_s y_{i,s}$ for all i.

In the general case, one needs to use the cumulative functions. Recall that the inverse of c.d.f, the generalized quantile function $F_\alpha^{-1} : [0,1] \to \mathbb{R} \cup \{-\infty\}$ is defined as

$$F_\alpha^{-1}(r) := \min_x \{x \in \mathbb{R} \cup \{-\infty\} : \ F_\alpha(x) \geq r\}.$$

The $\mathcal{W}_p$ distance is computed as

$$\mathcal{W}_p(\alpha, \beta)^p = \int_0^1 \left\| F_\alpha^{-1}(r) - F_\beta^{-1}(r) \right\|_p^p dr$$

So the **quantile** function of the *barycenter* is computed via the barycenter map

$$F_\alpha^{-1}(r) = A_w(F_{\beta_s}^{-1}(r))_{s=1}^S = \text{argmin}_g \sum_s^S w_s \left\| g - F_{\beta_s}^{-1} \right\|_{L_p}^p$$

which can be used, for instance, to compute barycenters between discrete measures supported on less than n points in $O(n \log(n))$ operations, using a simple sorting procedure.

3.5.3 Affine push-forward measures

Consider that all $\beta_s = T_{r_s,u_s,\#}\alpha_0$ are push-forward measures from some base measure α_0. The map $T_{r,u} : x \to r\,x + u$ is **affine**, i.e. via scaling and translation. It can be shown that the barycenter α_* of $\{\beta_s\}$ is a **push-forward measure** from α_0

$$\begin{aligned} \alpha_* &= T_{r_*,u_*,\#}\alpha_0 \\ \text{where } r_* &= \left(\sum_s w_s / r_s \right)^{-1} \quad \text{i.e. the harmonic mean} \\ u_* &= \sum_s w_s u_s \quad \text{i.e. the arithmetic mean} \end{aligned}$$

3.6 Entropic approximation

One can use entropic smoothing and approximate the solution of (33) using

$$\min_{\alpha \in \mathcal{M}_+(\mathcal{X})} \sum_{s=1}^{S} w_s \mathcal{L}_c^\epsilon(\alpha, \beta_s) \tag{35}$$

for some $\epsilon > 0$. This is a **smooth convex** minimization problem, which can be tackled using gradient descent.

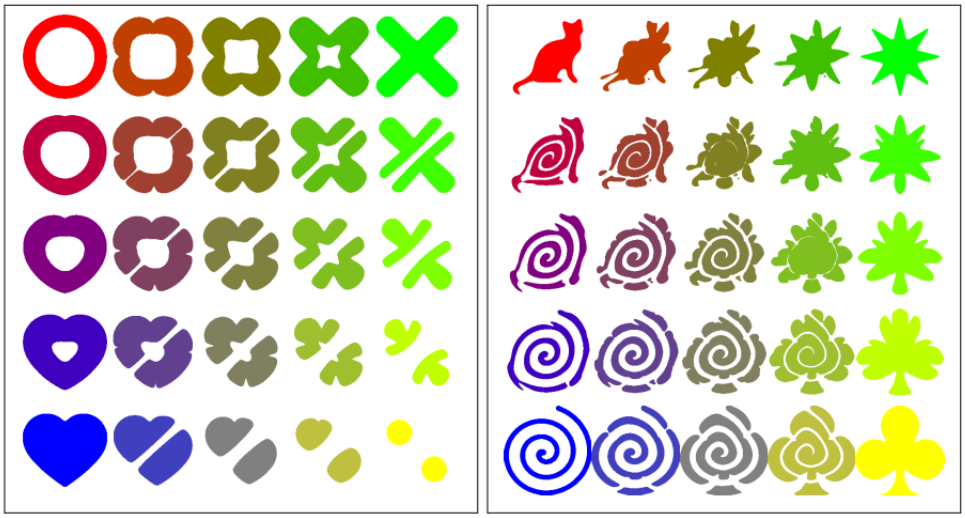

Figure 9.3: Barycenters between four input 2-D shapes using entropic regularization (9.15). To display a binary shape, the displayed images shows a thresholded density. The weights $(\lambda_s)_s$ are bilinear with respect to the four corners of the square.

Figure 3: The barycenter shape obtained via entropic regularization [Peyr and Cuturi, 2019]

3.6.1 Primal solution and generalized Sinkhorn algorithm

The problem (35) can be reformulated using KL divergence

$$\begin{aligned} \min_{\boldsymbol{P}_s \in \mathbb{R}_+^{n \times n_s}, \forall s} & \sum_{s=1}^{S} w_s \mathbb{KL}\left(\boldsymbol{P}_s \,\|\, \boldsymbol{K}_s\right) \\ \text{s.t. } & \boldsymbol{P}_s^T \mathbf{1}_n = \boldsymbol{b}_s, \quad s = 1, \ldots, S \\ & \boldsymbol{P}_1 \mathbf{1}_{n_1} = \ldots = \boldsymbol{P}_s \mathbf{1}_{n_s} \end{aligned} \tag{36}$$

The optimal solution is

$$\boldsymbol{P}_s = \operatorname{diag}\left(\boldsymbol{u}_s\right) \boldsymbol{K}_s \operatorname{diag}\left(\boldsymbol{v}_s\right), \quad s = 1, \ldots, S \tag{37}$$

and the scalings are sequentially updated via **generalized Sinkhorn algorithm**

$$\boldsymbol{v}_{s,t+1} \leftarrow \frac{\boldsymbol{b}}{\boldsymbol{K}_s^T \boldsymbol{u}_{s,t}} := \boldsymbol{b} \oslash \boldsymbol{K}^T \boldsymbol{u}_{s,t} \tag{38}$$

$$\boldsymbol{u}_{s,t+1} \leftarrow \frac{\boldsymbol{a}_{t+1}}{\boldsymbol{K}_s \boldsymbol{v}_{s,t+1}} := \boldsymbol{a}_{t+1} \oslash \boldsymbol{K} \boldsymbol{v}_{s,t+1} \tag{39}$$

$$\text{where } \boldsymbol{a}_{t+1} = \prod_s \left(\boldsymbol{K} \boldsymbol{v}_{s,t+1}\right)^{w_s} \quad \text{i.e. the geometric mean}$$

3.6.2 Dual problem

The optimal $(\boldsymbol{u}_{s,*}, \boldsymbol{v}_{s,*})$ appearing in (37) can be written as $(\boldsymbol{u}_{s,*}, \boldsymbol{v}_{s,*}) = (\exp(\boldsymbol{\lambda}_{s,*}/\epsilon), \exp(\boldsymbol{\mu}_{s,*}/\epsilon))$, where $(\boldsymbol{\lambda}_{s,*}, \boldsymbol{\mu}_{s,*})$ are the solutions of the following dual program

$$\begin{aligned} \max_{(\boldsymbol{\lambda}_s, \boldsymbol{\mu}_s), \forall s} & \sum_{s=1}^{S} w_s \left\{ \langle \boldsymbol{\mu}_s \,,\, \boldsymbol{b}_s \rangle - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}_s/\epsilon\right) \,,\, \boldsymbol{K}_s \exp\left(\boldsymbol{\mu}_s/\epsilon\right) \right\rangle \right\} \\ \text{s.t.} & \sum_{s=1}^{S} w_s \boldsymbol{\lambda}_s = \boldsymbol{0} \end{aligned} \tag{40}$$

As with the original case, the generalized Sinkhorn algorithm (38) and (39) can be obtained by alternating minimization with respect to $\boldsymbol{\mu}_s$ and $\boldsymbol{\lambda}_s$.

3.7 Wasserstein propagation

It is possible to generalize the barycenter problem (32), where one looks for distributions $(\boldsymbol{b}_u)_{\mathcal{U}}$ at some given set $\mathcal{U}$ of nodes in a **graph** $\mathcal{G} = (\mathcal{U} \cup \mathcal{V}, \mathcal{E})$ given a set of fixed input distributions $(\boldsymbol{b}_v)_{\mathcal{V}}$ on the *complementary* set $\mathcal{V}$ of the nodes. The unknown are determined by minimizing the overall transportation distance between **all pairs of nodes** $(u, v) \in \mathcal{E}$ forming edges in the graph

$$\min_{\{\boldsymbol{b}_u \in \Delta_{n_u}\}_{\mathcal{U}}} \sum_{(u,v)\in\mathcal{E}} L_{\boldsymbol{C}_{u,v}}(\boldsymbol{b}_u, \boldsymbol{b}_v)$$

where the cost matrices $\boldsymbol{C}_{u,v} \in \mathbb{R}^{n_u \times n_v}$ need to be specified by the user.

The barycenter problem (32) is a special case of this problem where the considered graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{V}, \mathcal{E})$ is **star-shaped**, where $\mathcal{U}$ is a single vertex connected to all the other vertices $\mathcal{V}$ (the weight w_s associated to $\boldsymbol{b}_s$ can be absorbed in the cost matrix). Introducing explicitly a coupling $\boldsymbol{P}_{u,v} \in U(\boldsymbol{b}_u, \boldsymbol{b}_v)$ for each edge $(u, v) \in \mathcal{E}$, and using entropy regularization, one can rewrite this problem similarly as in (36), and one extends Sinkhorn iterations (38) and (39).

4 Minimum Kantorovich Estimators

4.1 Challenges for Maximum Likelihood Estimation

Given empirical measures $\beta_n = \sum_{i=1}^{n} \frac{1}{n} \delta_{\boldsymbol{x}_i}$, $\boldsymbol{x}_i \subset \mathcal{X}$, i.i.d generated from some unknown distribution β_*, the goal of model estimation is to fit a parametric model $\theta \to \alpha_\theta \in \mathcal{M}(\mathcal{X})$ to the observed empirical input measure β_n. The standard ***Maximum Likelihood Estimation (MLE)*** minimize the *negative log-likelihood function*

$$\min_{\theta} \mathcal{L}_{MLE}(\alpha_\theta, \beta_n)$$

where α_θ has density ρ_θ with respect to Lebesgue measure and $\mathcal{L}_{MLE}(\alpha_\theta, \beta_n) = -\sum_i \log \rho_\theta(\boldsymbol{x}_i)$. We know that the MLE convergences to KL divergence $\mathcal{L}_{MLE}(\alpha, \beta_n) \to \mathbb{KL}(\alpha \,\|\, \beta_*)$ where β_* is the underlying distribution of data.

MLE has **challenges** to estimate α_θ that is ***singular***, i.e. does not have density function ρ_θ w.r.t. Lebesgue measure, or when the underlying distribution β_* is ***singular***. In both cases, the KL divergence is not well defined since the support of α_θ and β may not overlap. Another issue is when ρ_θ is complicated and hard to compute.

One typical case is when the task is to learn a genertive model α_θ lies on a ***low-dimensional sub-manifold***. Define the transform map $h_\theta : \mathcal{Z} \to \mathcal{X}$ so that

$$\alpha_\theta = h_{\theta,\#}\zeta, \quad \zeta \in \mathcal{M}(\mathcal{Z})$$

where $\mathcal{Z}$ is a low-dimensional subspace. Usually, α_θ does not have density w.r.t. Lebesgue measure on $\mathcal{X}$ since its support does not cover the entire space $\mathcal{X}$. Furthermore, computing this density is usually intractable, while generating i.i.d. samples from α_θ is achieved by computing $\boldsymbol{x}_i = h_\theta(\boldsymbol{z}_i)$, where $(\boldsymbol{z}_i)_i$ are i.i.d. samples from ζ.

For instance, in low-rank covariance esitmation, $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$, where $\boldsymbol{A} \in \mathbb{R}^{d\times r}$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_z)$, $\boldsymbol{\Sigma}_x = \boldsymbol{A}\boldsymbol{\Sigma}_z\boldsymbol{A}^T$ has rank $r \leq d$. This is the degenerate case for covariance estimation.

4.2 Learning generative models with Minimum Kantorovich Estimators

We can use the Wasserstein distance as objective function in place of MLE objective. In particular, for Lipschitz function f , we can write $\mathcal{W}_1$ in its dual form as

$$\mathcal{W}_1(\alpha,\beta) = \mathcal{L}_c(\alpha,\beta) := \sup_{f\in \mathrm{Lip}_1} \left\{ \int_{\mathcal{X}} f(x) d\alpha_\theta(x) - \int_{\mathcal{X}} f(x) d\beta(x) \right\} \tag{41}$$

Note that since $\alpha_\theta = h_{\theta,\#}\zeta$, we have $f(x)d\alpha_\theta(x) = f(x)d(h_{\theta,\#}\zeta)(x) = f(h_\theta(z))d\zeta(z)$. The estimation problem becomes

$$\min_\theta \mathcal{L}_c(\alpha_\theta, \beta_n) = \mathcal{L}_c(h_{\theta,\#}\zeta, \beta_n) \tag{42}$$

$$\Rightarrow \min_\theta \inf_{\pi\in U(\zeta,\beta_n)} \mathbb{E}_\pi \left[c(h_\theta(Z), Y)\right] \quad \text{(primal)} \tag{43}$$

$$\Rightarrow \min_\theta \sup_{f\in \mathrm{Lip}_1} \left\{\mathbb{E}_{Z\sim\zeta}\left[f(h_\theta(Z))\right] - \mathbb{E}_{X\sim\beta_n}\left[f(X)\right]\right\} \quad \text{(dual)} \tag{44}$$

The optimal transportation distances are weaker than MLE since they are not invariant to important families of invariances, such as rescaling, translation or rotations.

For $\mathcal{E}(\theta) := \mathcal{L}_c(\alpha_\theta, \beta_n)$, we can compute the subgradient function based on dual (41) when the **dual potential** function f is available,

$$\nabla_\theta \mathcal{E}(\theta) = \int_{\mathcal{Z}} \left[\nabla_\theta h_\theta(z)\right]^T \nabla f(h_\theta(z)) d\zeta(z) \tag{45}$$

Here $\nabla f(x)$ is gradient w.r.t. $x = h_\theta(z)$ and $\nabla_\theta h_\theta(z)$ is the differential (with respect to θ) of h_θ.

This formula is hard to use numerically, first because it requires first computing a continuous function f, which is a solution to a semi-discrete problem (16). For OT loss, this can be achieved using stochastic optimization, but this is hardly applicable in high dimension. Another option is to impose a *parametric* form for this **potential**, for instance expansion in an *RKHS* or a ***deep-network approximation***. (e.g. the ***Wasserstein GAN*** [Arjovsky et al., 2017]). A last issue is

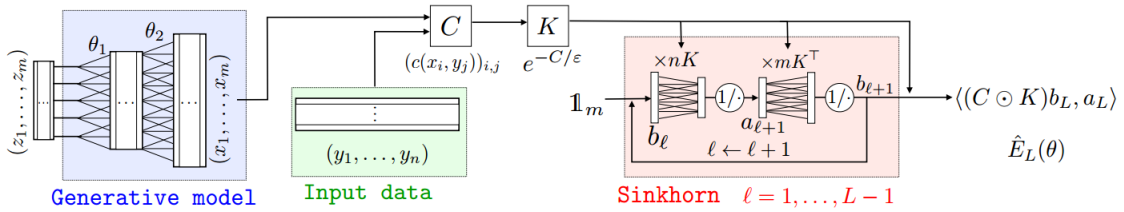

Figure 1: For a given fixed set of samples $(z_1, \ldots, z_m)$, and input data $(y_1, \ldots, y_n)$, flow diagram for the computation of Sinkhorn loss function $\theta \mapsto \hat{E}_\varepsilon^{(L)}(\theta)$. This function is the one on which automatic differentiation is applied to perform parameter learning. The display shows a simple 2-layer neural network $g_\theta : z \mapsto x$, but this applies to any generative model.

Figure 4: Learning generative model with Sinkhorn divergence [Genevay et al., 2018]

that it is **unstable** numerically because it requires the computation of the gradient $\nabla f(x)$ of the dual potential f.

We can also solve the primal problem directly, which is a semi-discrete problem. Note that $\beta_n = \sum_{i=1}^n \frac{1}{n} \delta_{\boldsymbol{x}_i}$, $\boldsymbol{x}_i \subset \mathcal{X}$, thus the objective becomes

$$\min_{\{\pi_i \in \mathcal{M}_+(\mathcal{Z})\}_i^n,} \sum_i^n \int_{\mathcal{Z}} c(h_\theta(z), \boldsymbol{x}_i) d\pi_i(z) \tag{46}$$
$$\text{where} \sum_i d\pi_i(z) = \frac{1}{n}$$
$$\sum_i \pi_i(z) = \zeta$$

Based on (46), we can obtain gradient

$$\nabla_\theta \mathcal{E}(\theta) = \sum_i^n \int_{\mathcal{Z}} [\nabla_\theta h_\theta(z)]^T \nabla_1 c(h_\theta(z), \boldsymbol{x}_i) d\pi_i(z), \tag{47}$$

where $\nabla_1 c(x, y))$ is the gradient of c w.r.t. first argument x. Note that as opposed to (45), this formula does not involve computing the gradient of the potentials being solutions of the dual OT problem.

The class of estimators obtained using $\mathcal{L} = \mathcal{L}_c$, often called **minimum Kantorovich estimators**, was initially introduced in [Bassetti et al., 2006], also [Canas and Rosasco, 2012]; It has been used in the context of generative models by [Montavon et al., 2016] to train restricted Boltzmann machines and in [Bernton et al., 2017] in conjunction with approximate Bayesian computations. Approximations of these computations using Deep Network are used to train deep generative models for both ***GAN*** [Arjovsky et al., 2017] and ***Variational Auto-Encoder (VAE)*** [Tolstikhin et al., 2018]; see also [Genevay et al., 2017, 2018, Salimans et al., 2018]. Note that the use of **Sinkhorn divergences** for parametric model fitting is used routinely for **shape matching** and **registration**, see [Gold et al., 1998, Rangarajan and Chui, 2000, Myronenko and Song, 2010, Feydy et al., 2017].

OT can also be applied to **metric learning** to learn cost function c [Wang and Guibas, 2012, Cuturi and Avis, 2014, Zen et al., 2014, Huang et al., 2016]; or to **domain adaptation** [Flamary et al., 2016] and transfer learning [Pan and Yang, 2009].

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*, 1(8):9, 2017.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 25, 2012.

Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer, 2017.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Gan and vae from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

Steven Gold, Anand Rangarajan, Chien-Ping Lu, Suguna Pappu, and Eric Mjolsness. New algorithms for 2d and 3d point matching: pose estimation and correspondence. *Pattern recognition*, 31(8):1019–1031, 1998.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. *Advances in neural information processing systems*, 29, 2016.

Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 29, 2016.

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Anand Rangarajan and Haili Chui. Applications of optimizing neural networks in medical image registration. In *Artificial Neural Networks in Medicine and Biology*, pages 99–104. Springer, 2000.

Tim Salimans, Dimitris Metaxas, Han Zhang, and Alec Radford. Improving gans using optimal transport. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.

I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview. net, 2018.

Fan Wang and Leonidas J Guibas. Supervised earth movers distance learning and its computer vision applications. In *European Conference on Computer Vision*, pages 442–455. Springer, 2012.

Gloria Zen, Elisa Ricci, and Nicu Sebe. Simultaneous ground metric learning and matrix factorization with earth mover's distance. In *2014 22nd International Conference on Pattern Recognition*, pages 3690–3695. IEEE, 2014.