Romon Gabriel

MVA
Master's Thesis
School year 2018-2019

Research internship in Optimal Transport

CREST - ENSAE
Palaiseau

Supervisors : Marco Cuturi
et Victor-Emmanuel Brunel
13/06/2019 - 06/12/2019

Contents

Introduction

This report summarizes my research internship at CREST, the topic of which was optimal transport. It is an extensive field of research, at the intersection between probability theory, analysis and geometry. The cost of entry into this theory is high, and I spent a considerable amount of time getting up to speed with the basics and browsing the literature.

This report starts with a self-contained introduction to optimal transport, so that the reader is equipped with enough knowledge to understand the rest. It is followed by an exposition of the internship's research topics: tree metrics, the algorithmic complexity of Sinkhorn's algorithm and the statistical properties of optimal transport.

Whenever a reference gave a claim without proof, or whenever I found the proof to be lacking in rigor, I chose to write my own. All the proofs in the report are my own work.

Original contributions are my implementation of the tree-Wasserstein distance, the correction of mistakes in [Altschuler et al., 2017] and the Theorem 4.2.9, which is a significant generalization of Theorem 1 in [Genevay et al., 2018].

Chapter 1

Introduction to Optimal Transport

1.1 Comparing probability measures

Comparing probability distributions is a longstanding problem of probability theory that has major implications in the fields of statistics and machine learning.

From a probabilistic standpoint, the ubiquitous notion of convergence in distribution leads naturally to the study of distances that metrize this convergence [Billingsley et al., 1971].

From a statistical standpoint, these discrepancies can be used to devise statistical tests that determine whether a sample is drawn from a known reference distribution, whether two samples are drawn from the same unknown distribution [Bhattacharya et al., 2016, Theorem 8.5 and 8.7], [Gretton et al., 2012] or whether a sample is drawn from either of two reference distributions [Lehmann and Romano, 2006, Theorem 13.1.1].

From a machine learning standpoint, these concepts are involved whenever one seeks to compare normalized histograms, which are equivalent to discrete probability distributions. In computer vision, an image may be represented by histograms of local features related to color, texture or shape. Comparing the respective histograms of two images indicates whether these images are similar [Kolouri et al., 2017]. In natural language processing, a document may be viewed as a bag-of-words, i.e. as a histogram over the vocabulary, and this provides a natural notion of distance between documents [Kusner et al., 2015]. In generative modelling, one may be interested in fitting the unknown data-generating distribution with a parametric distribution. The notion of fit depends intrinsically on the way of comparing the distributions.

Let $(\Omega, \mathcal{A})$ be a measurable space and P, Q be probability measures on that space. A classical way of comparing P and Q is via **f-divergences**.

Definition 1.1.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Standard results on slopes of convex functions imply that the limit $f'_\infty := \lim_{x\to\infty} \frac{f(x)}{x}$ exists in $\mathbb{R} \cup \{\infty\}$. By Lebesgue's decomposition theorem [Klenke, 2013, Theorem 7.33],
P writes uniquely as $P_a + P_s$ where P_a, P_s are finite non-negative measures such that $P_a \ll Q$ and $P_s \perp Q$. The f-divergence between P and Q is defined by

$$D_f(P\|Q) := \int_\Omega f\left(\frac{dP_a}{dQ}(w)\right) dQ(w) + f'_\infty P_s(\Omega)$$

Here are examples of well-known divergences. With $f : t \mapsto t \log t$ and $f : t \mapsto (t-1)^2$ one recovers respectively the **Kullback-Leibler** and the $\boldsymbol{\chi^2}$ divergences. In both cases, $f'_\infty = \infty$, so theses divergences blow up whenever P is not dominated by Q. The functions $f : t \mapsto \frac{1}{2}|t-1|$ and $f : t \mapsto (\sqrt{t}-1)^2$ define respectively the **total variation** and **squared Hellinger** divergences. For these divergences, $f'_\infty < \infty$ and $D_f(P\|Q)$ remains finite for any P, Q.

Theorem 1.1.2. 1. $D_f(P\|Q)$ is well-defined, $D_f(P\|Q) \geq 0$ and $D_f(P\|P) = 0$.
2. If f is strictly convex, $D_f(P\|Q) = 0 \implies P = Q$.

Proof. 1. Write $P = P_a + P_s$ the Lebesgue decomposition of P with respect to Q. Since f is convex, Theorem 7.9 in [Klenke, 2013] implies that $\int_\Omega f\left(\frac{dP_a}{dQ}(w)\right)^- dQ(w) < \infty$, so the integral in the definition is well-defined (although potentially equal to $+\infty$) and additionally $\int_\Omega f\left(\frac{dP_a}{dQ}(w)\right) dQ(w) \geq f\left(\int_\Omega \frac{dP_a}{dQ}(w) dQ(w)\right) = f(P_a(\Omega))$.
For $x \in \mathbb{R}$, the function $z \mapsto \frac{f(z)-f(x)}{z-x}$ is non-decreasing and bounded above by f'_∞, thus for $y > 0$ and $z = x + y$, one gets $\frac{f(x+y)-f(x)}{y} \leq f'_\infty$, that is $f(x) + f'_\infty y \geq f(x+y)$.
By the previous bound, $D(P||Q) \geq f(P_a(\Omega)) + f'_\infty P_s(\Omega)$

$$\geq f(P_a(\Omega) + P_s(\Omega)) = f(P(\Omega)) = f(1) = 0$$

The decomposition of P with respect to itself is $P = P + 0$, so that $D(P\|P) = 0$.
2. If $D_f(P\|Q) = 0$, equality is attained in Jensen's inequality. Since f is strictly convex, there is some constant C such that $\frac{dP_a}{dQ} = C$ Q-almost everywhere. By the definition of density we have $\forall A \in \mathcal{A}, P_a(A) = CQ(A)$ and since $Q(\Omega) = 1$ we get $C = P_a(\Omega)$.
Since f is strictly convex, the slope inequality is strict: $\forall y > 0, f(x) + f'_\infty y > f(x + y)$.
$D_f(P\|Q) = 0$ implies that $P_s(\Omega) = 0$, hence $P_s = 0$, $P = P_a$, $P_a(\Omega) = 1$ and $P = Q$. □

In general, f-divergences are not distances on the space of probability measures. Indeed they have no reason to be symmetric or to verify the triangle inequality. Remarkably, the total variation and Hellinger discrepancies turn out to be distances.

Remark 1.1.3. The terms "distance" and "metric" carry the same meaning in this report.

Suppose $(\Omega, \mathcal{A}) = (X, \mathcal{B}(X))$ where (X, d) is a metric space and $\mathcal{B}(X)$ denotes the Borel σ-algebra of X. The f-divergences defined above make no use of the metric d to compare probability measures on X. Therefore, one may level the criticism that f-divergences overlook geometric properties of X. This is clear when one considers two points $x, y \in X$ and the associated Dirac measures δ_x and δ_y. When $x \neq y$, the values of $D_f(\delta_x, \delta_y)$ for different f are displayed in Table 1.1.

f	Kullback-Leibler	χ^2	Total variation	Squared Hellinger
$D_f(\delta_x \| \delta_y)$	∞	∞	1	2

Table 1.1: Values of $D_f(\delta_x \| \delta_y)$ for different f, with $x \neq y$

Proof. When $x \neq y$, δ_x and δ_y have disjoint supports so δ_x is not dominated by δ_y, hence $\mathrm{KL}(\delta_x \| \delta_y) = \chi^2(\delta_x \| \delta_y) = \infty$. It is well-known [Tsybakov, 2009, Section 2.4] that the total variation can be rewritten as $\mathrm{TV}(P \| Q) = \sup_{A \in \mathcal{B}(X)} |P(A) - Q(A)|$ and the squared Hellinger as $H^2(P \| Q) = 2\left(1 - \int_X \sqrt{\frac{dP}{d\nu}(z)\frac{dQ}{d\nu}(z)} d\nu(z)\right)$ where ν is any non-negative measure that dominates both P and Q ($\nu = \frac{P+Q}{2}$ e.g.). By considering $A = \{x\}$ it is clear that $\mathrm{TV}(\delta_x \| \delta_y) = 1$. Let $\nu = \frac{1}{2}\delta_x + \frac{1}{2}\delta_y$ and note that $\delta_x \ll \nu$ and $\delta_y \ll \nu$ with respective densities $z \mapsto 2 \cdot 1_{z=x}$ and $z \mapsto 2 \cdot 1_{z=y}$, so that
$\int_X \sqrt{\frac{d\delta_x}{d\nu}(z)\frac{d\delta_y}{d\nu}(z)} d\nu(z) = \frac{1}{2}\int \sqrt{\frac{d\delta_x}{d\nu}(z)\frac{d\delta_y}{d\nu}(z)} d(\delta_x)(z) + \frac{1}{2}\int \sqrt{\frac{d\delta_x}{d\nu}(z)\frac{d\delta_y}{d\nu}(z)} d(\delta_y)(z) = 0$ and $H^2(P \| Q) = 2$. □

It is reasonable to consider instead discrepancies that compare δ_x and δ_y by relying on $d(x, y)$. This is one of the motivations behind optimal transport.

1.2 The Kantorovitch problem

Let X and Y be two Polish spaces (i.e. separable and complete metric spaces) and let $P(X), P(Y)$ denote the space of Borel probability measures on X and Y respectively. Consider $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ a lower semi-continuous **cost function** and two probability measures $\alpha \in P(X), \beta \in P(Y)$.

Definition 1.2.1. A **coupling** of (α, β) is a probability measure $\pi \in P(X \times Y)$ with marginals α and β, i.e. $\forall A \in \mathcal{B}(X), \pi(A \times Y) = \alpha(A)$ and $\forall B \in \mathcal{B}(Y), \pi(X \times B) = \beta(B)$. The set of couplings of (α, β) will be noted as $\Pi(\alpha, \beta)$.

Note that $\Pi(\alpha, \beta)$ is non-empty since it contains the trivial coupling $\alpha \otimes \beta$. **Kantorovitch's mass transportation problem** compares α and β by minimizing a linear functional over $\Pi(\alpha, \beta)$.

Definition 1.2.2. Kantorovitch's optimal transportation problem between α and β is the following:

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{X \times Y} c(x, y) d\pi(x, y) \tag{KP}$$

This problem has been studied extensively in the last 60 years and the relevant theory is developed in any of the following references [Rachev and Rüschendorf, 1998, Villani, 2003, Villani, 2008, Santambrogio, 2015].
(KP) admits the following dual formulation.

Theorem 1.2.3. Let Φ_c be the set of potentials $(\varphi, \psi) \in L^1(\alpha) \times L^1(\beta)$ such that $\varphi(x) + \psi(y) \leq c(x, y)$ for α-almost all x and β-almost all y. Then

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{X \times Y} c(x, y) d\pi(x, y) = \sup_{(\varphi,\psi) \in \Phi_c} \int \varphi(x) d\alpha(x) + \int \psi(y) d\beta(y) \tag{KP-dual}$$

Besides, the infimum in the LHS is attained and the supremum in the RHS can be restrained to potentials in $C_b(X) \times C_b(Y)$ (i.e continuous and bounded potentials, instead of just integrable).

Proof. See Theorem 1.3 in [Villani, 2003]. □

Proposition 1.2.4. When $\alpha = \delta_x$ and $\beta = \delta_y$, the only coupling of (α, β) is $\delta_x \otimes \delta_y$ and the value of (KP) is $c(x, y)$.

Proof. Let us prove that $\Pi(\delta_x, \delta_y) = \{\delta_x \otimes \delta_y\}$. Consider $\pi \in \Pi(\delta_x, \delta_y)$. Since the products of Borel sets form a pi-system, it suffices to check that

$$\forall (A, B) \in \mathcal{B}(X) \times \mathcal{B}(Y), \pi(A \times B) = \delta_x \otimes \delta_y(A \times B) = \delta_x(A)\delta_y(B)$$

When $x \notin A$ or $y \notin B$, since $\pi(A \times B) \leq \pi(A \times Y) = \delta_x(A)$ and $\pi(A \times B) \leq \pi(X \times B) = \delta_y(B)$, we have $\pi(A \times B) \leq \min(\delta_x(A), \delta_y(B)) = 0$, hence $\pi(A \times B) = 0 = \delta_x(A)\delta_y(B)$.
When $x \in A$ and $y \in B$, we have

$$\begin{aligned} 1 - \pi(A \times B) &= \pi((A \times B)^c) = \pi((A^c \times Y) \cup (X \times B^c)) \\ &\leq \pi(A^c \times Y) + \pi(X \times B^c) \\ &= \delta_x(A^c) + \delta_y(B^c) = 0 \end{aligned}$$

Hence $\pi(A \times B) = 1 = \delta_x(A)\delta_y(B)$

This proves $\pi = \delta_x \otimes \delta_y$ hence $\Pi(\delta_x, \delta_y) = \{\delta_x \otimes \delta_y\}$. Since there is only one coupling, (KP) is trivial and its minimum is $\int_{X\times Y} c(z) d(\delta_x \otimes \delta_y)(z) = c(x, y)$. □

1.3 Wasserstein distances

Suppose $X = Y$, d is a metric on X, $p \geq 1$, and consider the cost function $c(x, y) := d(x, y)^p$. In this setting, (KP) has additional interesting properties.

Definition 1.3.1. Let (X, d) be a Polish space, $p \geq 1$ and $\alpha, \beta \in P(X)$.
The p-**Wasserstein distance** between α and β is defined by

$$W_p(\alpha, \beta) := \left(\inf_{\pi \in \Pi(\alpha,\beta)} \int_{X\times X} d(x, y)^p d\pi(x, y) \right)^{1/p} \tag{p-Wasserstein}$$

Theorem 1.3.2. Let $P_p(X) = \{\mu \in P(X) | \ \forall x_0 \in X, \int_X d(x, x_0)^p d\mu(x) < \infty\}$ denote the space of probability measures with finite moment of order p.
Then W_p defines a metric on $P_p(X)$.

Proof. See Theorem 7.3 in [Villani, 2003]. □

Remark 1.3.3. W_p satisfies the axioms of a metric on $P(X)$. The only issue is that it can be equal to ∞. Restricting it to $P_p(X)$ simply ensures that it remains finite. Indeed, if

$$\begin{aligned} \alpha, \beta \in P_p(X),\ W_p^p(\alpha, \beta) &\leq \int_{X\times X} d(x, y)^p d(\alpha \otimes \beta)(x, y) \\ &\leq \int_{X\times X} 2^{p-1}(d(x, x_0)^p + d(y, x_0)^p) d(\alpha \otimes \beta)(x, y) \\ &= 2^{p-1} \left(\int_X d(x, x_0)^p d\alpha(x) + \int_X d(y, x_0)^p d\beta(y) \right) < \infty \end{aligned}$$

When $p = 1$, duality takes a very simple form.

Theorem 1.3.4. Let $\mathrm{Lip}_d = \{\varphi : X \to \mathbb{R} | \ \forall (x, y) \in X^2, |\varphi(x) - \varphi(y)| \leq d(x, y)\}$ be the set of 1-Lipschitz functions. For $\alpha, \beta \in P_1(X)$,

$$W_1(\alpha, \beta) = \sup_{\varphi \in \mathrm{Lip}_d} \int_X \varphi(x) d\alpha(x) - \int_X \varphi(x) d\beta(x)$$

Proof. See Remark 6.5 in [Villani, 2008]. □

The following proposition makes the connection between the total variation distance introduced in Section 1.1 and some Wasserstein distance.

Proposition 1.3.5. Let $d : (x, y) \mapsto 1_{x \neq y}$ be the trivial metric. Then the associated 1-Wasserstein is the total variation distance.

Proof. It is clear that for this metric, $P_1(X) = P(X)$. It is also well-known [Tsybakov, 2009, Section 2.4] that $\mathrm{TV}(\alpha \| \beta) = \sup_{A \in \mathcal{B}(X)} |\alpha(A) - \beta(A)| = \frac{1}{2} \int_X \left| \frac{d\alpha}{d\nu}(x) - \frac{d\beta}{d\nu}(x) \right| d\nu(x)$ where ν is any non-negative measure that dominates α and β ($\nu = \frac{\alpha+\beta}{2}$ e.g.).
In Theorem 1.3.4, by considering $-\varphi$ one easily sees that

$$W_1(\alpha, \beta) = \sup_{\varphi \in \mathrm{Lip}_d} \left| \int_X \varphi(x) d\alpha(x) - \int_X \varphi(x) d\beta(x) \right|$$

Let $A \in \mathcal{B}(X)$. The function $x \mapsto 1_A(x)$ is in $\mathcal{L}_d$ thus $|\alpha(A) - \beta(A)| = \left| \int 1_A d\alpha - \int 1_A d\beta \right| \leq W_1(\alpha, \beta)$. Taking the supremum over A yields $\mathrm{TV}(\alpha \| \beta) \leq W_1(\alpha, \beta)$.
For the reverse inequality, consider $\varphi \in \mathrm{Lip}_d$. The Lipschitz constraint is equivalent to $\forall x, y \in X, |\varphi(x) - \varphi(y)| \leq 1$. One can replace φ by $\varphi - c$ (this leaves $\int \varphi d\alpha - \int \varphi d\beta$ untouched and φ is still in Lip_d) and for a suitable c we can suppose WLOG that $\|\varphi\|_\infty \leq \frac{1}{2}$.
Let $\nu = \frac{\alpha+\beta}{2}$ and note that

$$\begin{aligned} \left| \int_X \varphi(x) d\alpha(x) - \int_X \varphi(x) d\beta(x) \right| &\leq \int_X |\varphi(x)| \left| \frac{d\alpha}{d\nu}(x) - \frac{d\beta}{d\nu}(x) \right| d\nu(x) \\ &\leq \frac{1}{2} \int_X \left| \frac{d\alpha}{d\nu}(x) - \frac{d\beta}{d\nu}(x) \right| d\nu(x) \\ &= \mathrm{TV}(\alpha \| \beta) \end{aligned}$$

Taking the supremum over φ finishes the proof. □

$(P_p(X), W_p)$ has interesting topological properties. Before stating these properties we need to define **weak convergence** of a sequence in $P(X)$. This is related to convergence in distribution.

Definition 1.3.6. Let $(\alpha_n)_n \in P(X)^{\mathbb{N}}$ and $\alpha \in P(X)$. $(\alpha_n)_n$ converges weakly to α (denoted $\alpha_n \Rightarrow \alpha$) if for every bounded continuous function f,

$$\lim_{n \to \infty} \int_X f(x) d\alpha_n(x) = \int_X f(x) d\alpha(x)$$

Remark 1.3.7. Let $(Y_n)_n$ be a sequence of random elements from $(\Omega, \mathcal{A}, P)$ to $(X, \mathcal{B}(X))$. Y_n converges in distribution to a random element Y if and only the sequence of image measures $(P_{Y_n})_n$ converges weakly to P_Y (formally, P_{Y_n} is the pushforward of P by Y_n).

Remark 1.3.8. The well-known Portmanteau theorem [Klenke, 2013, Theorem 13.16] provides equivalent conditions for weak convergence.

A desirable property for any metric D on $P(X)$ is that it **metrizes** weak convergence, i.e. $\alpha_n \Rightarrow \alpha \iff D(\alpha_n, \alpha) \xrightarrow[n \to \infty]{} 0$. Remarkably, the total variation and Hellinger distances define the same topology, and they do not metrize weak convergence. Returning to the example of Dirac measures, consider $X = \mathbb{R}$, $\alpha_n = \delta_{1/n}$ and $\alpha = \delta_0$. It is clear that $\alpha_n \Rightarrow \alpha$. We proved earlier that $\mathrm{TV}(\alpha_n \| \alpha) = 1$, $H^2(\alpha_n \| \alpha) = 2$ and by Proposition 1.2.4 $W_p(\alpha_n, \alpha) = d(\frac{1}{n}, 0)^p$. This suggests that Wasserstein distances have nicer properties, and this is confirmed in the next theorem.

Theorem 1.3.9. Let $(\alpha_n)_n \in P_p(X)^{\mathbb{N}}$ and $\alpha \in P_p(X)$. The following statements are equivalent:

1. $W_p(\alpha_n, \alpha) \xrightarrow{n \to \infty} 0$
2. $\alpha_n \Rightarrow \alpha$ and the moment of order p converges:
$$\forall x_0 \in X, \lim_{n\to\infty} \int_X d(x, x_0)^p d\alpha_n(x) = \int_X d(x, x_0)^p d\alpha(x)$$
3. $\alpha_n \Rightarrow \alpha$ and the following tightness condition is verified:
$$\forall x_0 \in X, \lim_{R\to\infty} \limsup_{n\to\infty} \int_{d(x,x_0)\geq R} d(x, x_0)^p d\alpha_n(x) = 0$$

Proof. See Theorem 7.12 in [Villani, 2003]. □

Remark 1.3.10. When d is bounded or X is bounded, for any $x_0 \in X$ the set $\{x \in X | \, d(x, x_0) \geq R\}$ is empty for large enough R, so that

$$\lim_{R\to\infty} \limsup_{n\to\infty} \int_{d(x,x_0)\geq R} d(x, x_0)^p d\alpha_n(x) = 0$$

In this bounded case, $P_p(X) = P(X)$ and W_p metrizes weak convergence on the whole of $P(X)$.

In the unbounded case, convergence of moments or the tightness condition need to be checked to infer $W_p(\alpha_n, \alpha) \xrightarrow{n\to\infty} 0$.

1.4 The discrete case

As mentioned in the beginning of Section 1.1, machine learning practitioners often deal with normalized histograms, i.e. discrete probability measures. In this section we consider Kantorovitch's problem in the discrete case. A comprehensive reference on this topic is [Peyré et al., 2019, Chapter 2].

Let X, Y be Polish spaces and $c : X \times Y \to \mathbb{R}_+$ be a lower semi-continuous cost function. Suppose $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{y_j}$ where $n, m \in \mathbb{N}^*$, $a \in (\mathbb{R}_+^*)^n$, $b \in (\mathbb{R}_+^*)^m$ are such that $\sum_{i=1}^n a_i = \sum_{j=1}^m b_j = 1$, and $(x_i)_{i \in [\![1,n]\!]} \in X^n$, $(y_j)_{j \in [\![1,n]\!]} \in Y^m$ are support points.

Theorem 1.4.1. Let $U(a,b) = \{P \in (\mathbb{R}_+)^{n \times m} |\ P\mathbb{1}_m = a, P^T \mathbb{1}_n = b\}$ denote the convex polytope of matrices with row sums a, column sums b and non-negative entries.
In the discrete setting, (KP) has the following form:

$$\inf_{P \in U(a,b)} \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) P_{ij}$$

Proof. Given $\pi \in \Pi(\alpha, \beta)$, let us prove that π is discrete and supported on the (x_i, y_j) where $(i,j) \in [\![1,n]\!] \times [\![1,m]\!]$ (note that $\pi(\{(x_i, y_j\})$ may be 0). It suffices to prove that $\pi([\cup_{i=1}^n \{x_i\}] \times [\cup_{j=1}^m \{y_j\}] = 1$, and this is equivalent to

$$\pi(([\cup_{i=1}^n \{x_i\}]^c \times Y) \cup \big(X \times [\cup_{j=1}^m \{y_j\}]^c\big)) = 0$$

This holds since

$$\begin{aligned} \pi(([\cup_{i=1}^n \{x_i\}]^c \times Y) \cup \big(X \times [\cup_{j=1}^m \{y_j\}]^c\big)) &\leq \pi([\cup_{i=1}^n \{x_i\}]^c \times Y) + \pi(X \times [\cup_{j=1}^m \{y_j\}]^c) \\ &= \alpha([\cup_{i=1}^n \{x_i\}]^c) + \beta([\cup_{j=1}^m \{y_j\}]^c) = 0 \end{aligned}$$

Therefore there is some $P \in U(a,b)$ such that $\pi = \sum_{i=1}^n \sum_{j=1}^m P_{ij} \delta_{(x_i, y_j)}$ and the rest of the proof is easy. □

Remark 1.4.2. The discrete case boils down to a **linear programming** problem: the objective and the $n + m$ constraints are linear in P.
This problem can be solved exactly with any LP solver, but the best time complexity achieved to date is $O((n+m)nm \log^2(n+m))$ [Peyré et al., 2019, Section 3.5.3]. When $n = m$ this complexity is $O(n^3 \log^2(n))$, whereas the size of the data is n^2.

Chapter 2

Tree metrics

2.1 Introduction

The original goal of the internship was to work on extensions of the following recent paper [Le et al., 2019b], which was accepted at NIPS 2019. The paper considers the 1-Wasserstein between discrete measures on spaces where the ground metric is a **tree metric**.

Definition 2.1.1. Suppose (X, d) is a metric space and consider a tree T (i.e. a connected acyclic undirected graph) whose nodes are elements of X. To each edge $e = (z, t)$ in the tree we associate the weight $w_e = d(z, t)$. Given two nodes z and t, there is a unique path between them in the tree. $d_T(z, t)$ is defined as the sum of the weights along this path. d_T is called the tree metric associated to T.

Proposition 2.1.2. (T, d_T) is a metric space.

Definition 2.1.3. Consider (X, d) a metric space, $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$ discrete measures on X. Let T be a tree whose nodes contain the set $\{(x_i)_{i \in [\![1,n]\!]}\} \cup \{(y_j)_{j \in [\![1,m]\!]}\}$. The **tree-Wasserstein** between α and β is defined as $\mathrm{TW}(\alpha, \beta) := W_1(\alpha, \beta)$ where in the RHS α, β are taken as measures on (T, d_T).

Remark 2.1.4. In other words, we embed the original metric space (X, d) into a tree metric space (T, d_T) and we compute the 1-Wasserstein in the new space.

It turns out that $\mathrm{TW}(\alpha, \beta)$ has a nice closed form. This owes much to the simplicity of duality for 1-Wasserstein distances, as seen in Theorem 1.3.4.

Theorem 2.1.5. Let T be a tree rooted at an arbitrary $r \in X$ and let E_T denote the set of edges of T. For $e \in E_T$, let u_e denote the node closest to the root and v_e the other node (see Figure 2.1). Additionally, for any node $z \in X$, let $\Gamma(z)$ denote the subtree rooted at z. Then

$$\mathrm{TW}(\alpha, \beta) = \sum_{e \in E_T} w_e |\alpha(\Gamma(v_e)) - \beta(\Gamma(v_e))|$$

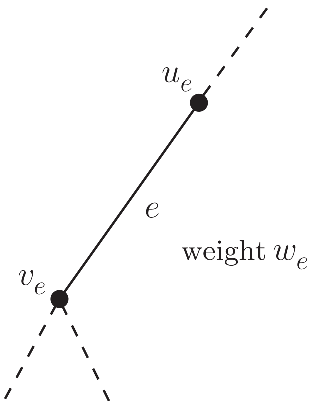

Figure 2.1: An edge $e = (u_e, v_e)$ in a rooted tree

Proof. We make use of the dual form of 1-Wasserstein. Let $f : T \mapsto \mathbb{R}$ be 1-Lipschitz for the metric d_T. By replacing f with $f - f(r)$ we can suppose WLOG that $f(r) = 0$. For every edge (u_e, v_e) define $\tilde{f}(v_e) := \frac{f(v_e) - f(u_e)}{w_e}$. Since f is 1-Lipschitz, $\tilde{f}$ takes values in $[-1, 1]$, and by construction $f(v_e) = \sum_{a \in E_T} 1_{\Gamma(v_a)}(v_e) \tilde{f}(v_a) w_a$.

Note that

$$\begin{aligned}
\int_T f(t) d\alpha(t) &= f(r)\alpha(\{r\}) + \sum_{e \in E_T} f(v_e)\alpha(\{v_e\}) \\
&= \sum_{e \in E_T} \sum_{a \in E_T} 1_{\Gamma(v_a)}(v_e) \tilde{f}(v_a) w_a \alpha(\{v_e\}) \\
&= \sum_{a \in E_T} \tilde{f}(v_a) w_a \sum_{e \in E_T} 1_{\Gamma(v_a)}(v_e) \alpha(\{v_e\}) \\
&= \sum_{a \in E_T} \tilde{f}(v_a) w_a \alpha(\Gamma(v_a))
\end{aligned}$$

Hence

$$\int_T f(t) d\alpha(t) - \int_T f(t) d\beta(t) = \sum_{e \in E_T} \tilde{f}(v_e) w_e \left[\alpha(\Gamma(v_e)) - \beta(\Gamma(v_e))\right]$$

The supremum over f is attained for f^* defined recursively by $f^*(r) = 0$ and $f^*(v_e) = w_e \operatorname{sign}(\alpha(\Gamma(v_e)) - \beta(\Gamma(v_e)) + f^*(u_e)$. This yields $\tilde{f}^*(v_e) = \operatorname{sign}(\alpha(\Gamma(v_e)) - \beta(\Gamma(v_e))$, thus

$$\mathrm{TW}(\alpha, \beta) = W_1(\alpha, \beta) = \sup_{f \in \mathrm{Lip}_{d_T}} \int_T f(t) d\alpha(t) - \int_T f(t) d\beta(t) = \sum_{e \in E_T} w_e |\alpha(\Gamma(v_e)) - \beta(\Gamma(v_e))|$$

□

Remark 2.1.6. Since the proof relies heavily on the dual form of the 1-Wasserstein, it is very likely that there is no nice closed form for the p-Wasserstein on tree metric spaces when $p > 1$.

2.2 Personal work

The beginning of the internship was dedicated to implementing the tree-Wasserstein. Note that the authors of [Le et al., 2019b] did not release their code, nor did they give pseudo-code for the tree-Wasserstein in their paper.

A computational challenge in their approach is that they consider measures $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$ but the tree T is not given *a priori*. Instead they propose methods to sample trees whose nodes are the support points of the measures. We explored and implemented other methods of sampling using the Python package `NetworkX`.

Assuming $X = \mathbb{R}^d$, $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$ we:

1. Generate a connected graph G whose nodes are the support points and whose edges are the Euclidean distances between points. The simplest and fastest way is to use the complete graph on the support points. To get a graph that is sensitive to the geometry of the support points, it is preferable to consider a k-NN graph, but it may not be connected. Even if k is reasonably large, if there are two tightly-knit and distant clusters in the support points, it will have at least two components. This can be fixed by adding random edges between the connected components.

2. Compute a directed spanning tree T on the graph. For our use in tree-Wasserstein, G or T needs to be random.

   • BFS and DFS on a connected graph both generate spanning trees. We get random spanning trees simply by randomizing the starting node of the search. An interesting question is the distribution induced by this sampling method on spanning trees of G.

   • Wilson's algorithm [Lyons and Peres, 2017, Theorem 4.1] is based on loop-erased random walks, with transition probabilities proportional to edge weights (or a function thereof). The probability of a spanning tree being sampled is the product of its edge weights.

   • Kruskal's or Prim's algorithm can be used to compute a minimum spanning tree, but these are deterministic algorithms (the minimum spanning tree is unique if all the edge weights are distinct).

3. Prepare T for the computation of d_{TW}. Let T' be the tree obtained by reversing the edges in T and compute a topological sort of T'. This yields an ordered list of nodes ℓ. By construction, the leaves of T are the first elements of ℓ. For each node in ℓ, add its mass under α and β to the corresponding masses of its parent in T (which is its child in T'). When this process ends, by construction, the masses at each `node` are $(\alpha(\Gamma(\texttt{node})), \beta(\Gamma(\texttt{node})))$, the masses under α and β of the subtree rooted at `node`, which is exactly what we need to compute TW.
   As a sanity check, one can verify that the masses affected to the root of T are $(1, 1)$.

4. Compute d_{TW}. This is done by simply looping over the edges of T'.

We provide cost analyses for both types of graphs in Table 2.1 and 2.2. Let $N = n+m$ be the number of nodes of the graph. The number of edges in the complete graph is $\frac{N(N-1)}{2} = \Theta(N^2)$ and $\frac{kN}{2} = \Theta(kN)$ for the k-NN graph.

Step		Time complexity	Space complexity
Generation		$\Theta(N^2 d)$	$\Theta(N^2)$
Spanning tree	BFS/DFS	$\Theta(N^2)$	$O(N)$
	Wilson	?	$O(N)$
Preparation		$\Theta(N)$	$O(N)$
TW computation		$\Theta(N)$	$\Theta(1)$

Table 2.1: Complexity estimates for a complete graph

Step		Time complexity	Space complexity
Generation		?	$\Theta(kN)$
Spanning tree	BFS/DFS	$\Theta(kN)$	$O(N)$
	Wilson	?	$O(N)$
Preparation		$\Theta(N)$	$O(N)$
TW computation		$\Theta(N)$	$\Theta(1)$

Table 2.2: Complexity estimates for a k-NN graph

Once the tree is constructed, the time complexity for computing $\text{TW}(\alpha, \beta)$ is $\Theta(n+m)$. This is much better than the $O((n+m)nm\log^2(n+m))$ of LP solvers.

The goal of the internship was to adapt the notion of sliced Wasserstein barycenters developed in [Bonneel et al., 2015] to the tree-Wasserstein. As explained above, trees depend on the measures α and β because the support points must be contained in their nodes. On the contrary, sliced Wasserstein is based on sampling random lines in $\mathbb{R}^d$ and projecting the support points on those lines (so the lines are independent of α and β).

Because of this significant difference, I was not able to come up with a notion of tree-sliced Wasserstein barycenters. Such a notion was recently proposed in [Le et al., 2019a], but I have not had time to look into it.

Given a ground metric d, two discrete measures α, β and a tree T, the authors of [Le et al., 2019b] do not study the link between $\mathrm{TW}(\alpha, \beta)$ and $W_1(\alpha, \beta)$. If the tree T is sufficiently well-chosen, we expect that $\mathrm{TW}(\alpha, \beta)$ is not too far from the true 1-Wasserstein.

The identity mapping $\mathrm{id} : (X, d) \to (T, d_T)$ is a **metric embedding**. We define the expansion of id as $\sup_{x \neq y \in X} \frac{d_T(x,y)}{d(x,y)}$, its contraction as $\sup_{x \neq y \in X} \frac{d(x,y)}{d_T(x,y)}$ and its distortion as the product of the expansion and the contraction. Equivalently, the distortion of id is the minimum $\lambda \geq 1$ such that

$$\exists \mu > 0, \forall (x, y) \in X, \mu d_T(x, y) \leq d(x, y) \leq \lambda \mu d_T(x, y)$$

A desirable tree is a T such that id has low distortion. It is not always possible to construct such a tree: in [Rabinovich and Raz, 1998] it is proven that any embedding of the n-point cycle C_n into a tree metric has distortion at least $\frac{n}{3} - 1$. **Probabilistic embeddings** are a possible workaround [Bartal, 2019]. The paper [Fakcharoenphol et al., 2004] provides a way of sampling trees with distortion $O(\log n)$ (sometimes referred to as FRT trees because of the author's initials).

When browsing the literature on tree metrics, I noticed that the paper [Leeb, 2018] makes the connection between FRT trees and the approximation of $W_1(\alpha, \beta)$ by $\mathrm{TW}(\alpha, \beta)$. Numerical experiments carried out on the MNIST dataset show that both quantities are not far from each other (see Figure 2.2)

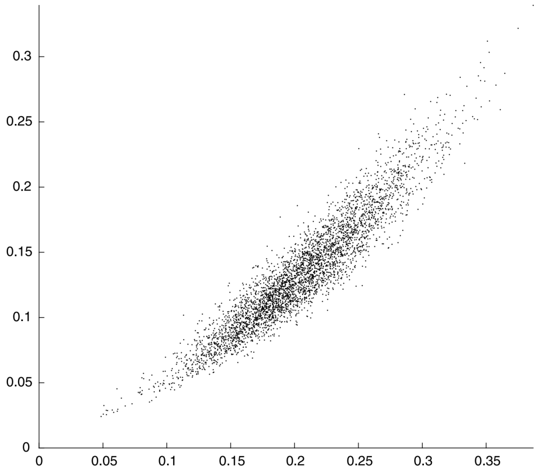

Figure 2.2: True W_1 versus approximate W_1, using 25 trees, from [Leeb, 2018]

Chapter 3

Regularized Optimal Transport

3.1 Introduction

In the discrete case, regularizing Kantorovitch's problem with an entropic term enables the use of an elementary iterative scheme that converges quickly to the minimum. This regularization and the iterative method were introduced in [Cuturi, 2013].

In the general case, the regularized Kantorovitch's problem is defined using the Kullback-Leibler divergence (already introduced in Section 1.1). Let $(\Omega, \mathcal{A})$ be a measurable space and P, Q be probability measures on that space. We define

$$\mathrm{KL}(P\|Q) := \begin{cases} \int_\Omega \log\left(\frac{dP}{dQ}(w)\right) dP(w) & \text{if } P \ll Q \\ \infty & \text{otherwise} \end{cases}$$

If $P \ll Q$, this rewrites as $\int_\Omega \log\left(\frac{dP}{dQ}(w)\right) \frac{dP}{dQ}(w) dQ(w)$ and this is consistent with the definition via f-divergences (with $f : x \mapsto x \log x$).

Definition 3.1.1. Let X,Y be Polish spaces, $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ a lower semi-continuous cost function and two probability measures $\alpha \in P(X), \beta \in P(Y)$. The regularized Kantorovitch's problem is the following.
Given $\varepsilon > 0$ a regularization parameter, we consider

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{X\times Y} c(x,y) d\pi(x,y) + \varepsilon \, \mathrm{KL}(\pi\|\alpha \otimes \beta) \qquad \text{(Regularized KP)}$$

Remark 3.1.2. The term $\mathrm{KL}(\pi\|\alpha \otimes \beta)$ in the objective forces the transportation plans π to be absolutely continuous with respect to $\alpha \otimes \beta$. In general a coupling $\pi \in \Pi(\alpha, \beta)$ need not be dominated by $\alpha \otimes \beta$ (consider π the uniform distribution on the diagonal of the unit rectangle $[0,1] \times [0,1]$ and α, β having $\mathcal{U}([0,1])$ distribution).

The regularized problem also admits a dual formulation. A version of the dual problem was first stated in [Genevay et al., 2016], without proof. A proof was added in [Genevay, 2019, Proposition 4] but it is purely formal and not rigorous. The recent paper [Clason et al., 2019] provides an in-depth study of duality in the setting where $X = \mathbb{R}^d$ and the regularization is modified:

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{X \times Y} c(x,y) d\pi(x,y) + \varepsilon \, \mathrm{KL}(\pi \| \lambda_d \otimes \lambda_d)$$

The authors make heavy use of functional analysis and I plan on spending time on this paper in the coming weeks. In [Genevay et al., 2016, Genevay et al., 2018, Genevay, 2019], duality is stated as follows:

Proposition 3.1.3. Let X, Y be arbitrary metric spaces, and let $\mathcal{C}(X)$, $\mathcal{C}(Y)$ denote the spaces of continuous functions with real values. The convex dual of (Regularized KP) is

$$\sup_{\substack{\varphi \in \mathcal{C}(X) \\ \psi \in \mathcal{C}(Y)}} \int_X \varphi(x) d\alpha(x) + \int_Y \psi(y) d\beta(y) - \varepsilon \int_{X \times Y} e^{\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}} d(\alpha \otimes \beta)(x,y) + \varepsilon \tag{Dual Reg. KP 1}$$

In the recent paper [Mena and Weed, 2019], it is stated as follows:

Proposition 3.1.4. When $X = Y = \mathbb{R}^d$ and $c : (x,y) \mapsto \|x-y\|_2^2$, the dual of (Regularized KP) is

$$\sup_{\substack{\varphi \in L^1(\alpha) \\ \psi \in L^1(\beta)}} \int_{\mathbb{R}^d} \varphi(x) d\alpha(x) + \int_{\mathbb{R}^d} \psi(y) d\beta(y) - \varepsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{\frac{\varphi(x)+\psi(y)-\|x-y\|_2^2}{\varepsilon}} d(\alpha \otimes \beta)(x,y) + \varepsilon \tag{Dual Reg. KP 2}$$

The space in which the dual is optimized certainly depends on the regularity of X, Y, α, β. By the best of our knowledge, there is no result on this dependence in the literature. I plan on studying this issue in the coming weeks, as this is important for the work carried out in Chapter 3.

3.2 The discrete case

Proposition 3.2.1. In the discrete case where $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{y_j}$ with $n, m \in \mathbb{N}^*$, $a \in (\mathbb{R}_+^*)^n$, $b \in (\mathbb{R}_+^*)^m$, the regularized problem writes as

$$\inf_{P \in U(a,b)} \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) P_{ij} + \varepsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right)$$

Proof. The proof is similar to that of Theorem 1.4.1. It suffices to notice that $\dfrac{d\pi}{d\alpha \otimes \beta}(x,y) = \displaystyle\sum_{i=1}^n \sum_{j=1}^m \frac{P_{ij}}{a_i b_j} 1_{x=x_i} 1_{y=y_j}$. □

Theorem 3.2.2. The dual formulation for the discrete regularized problem is

$$\sup_{\substack{\lambda \in \mathbb{R}^n \\ \mu \in \mathbb{R}^m}} \sum_{i=1}^n \lambda_i a_i + \sum_{j=1}^m \mu_j b_j - \varepsilon \sum_{i=1}^n \sum_{j=1}^m a_i b_j \exp\left(\frac{\lambda_i + \mu_j - c(x_i, y_j)}{\varepsilon}\right) + \varepsilon$$

Proof. Let g denote the objective function in Proposition 3.2.1. g is strictly convex with domain $\operatorname{dom} g = (\mathbb{R}_+)^{n \times m}$, g is continuous on $\operatorname{dom} g$, and coercive. The constraint set is compact, thus the infimum is actually a minimum, and there is a unique solution.
Since there are only two equality constraints that are linear in P, strong duality holds. Let $L(P, \lambda, \mu) = \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) P_{ij} + \varepsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right) - \lambda^T (P \mathbb{1}_m - a) - \mu^T (P^T \mathbb{1}_n - b)$ be the Lagrange function. For fixed λ, μ, L is convex in P and $P_{ij}^* = a_i b_j \exp\left(\frac{\lambda_i + \mu_j - c(x_i, y_j)}{\varepsilon} - 1\right)$ is a critical point, so the dual function is
$g(\lambda, \mu) = L(P^*, \lambda, \mu) = \sum_{i=1}^n \lambda_i a_i + \sum_{j=1}^m \mu_j b_j - \varepsilon \sum_{i=1}^n \sum_{j=1}^m a_i b_j \exp\left(\frac{\lambda_i + \mu_j - c(x_i, y_j)}{\varepsilon} - 1\right)$
With the change of variable $\lambda \leftarrow \lambda - \frac{\varepsilon}{2} \mathbb{1}_n$, $\mu \leftarrow \mu - \frac{\varepsilon}{2} \mathbb{1}_m$, the dual problem is equivalent to

$$\sup_{\substack{\lambda \in \mathbb{R}^n \\ \mu \in \mathbb{R}^m}} \sum_{i=1}^n \lambda_i a_i + \sum_{j=1}^m \mu_j b_j - \varepsilon \sum_{i=1}^n \sum_{j=1}^m a_i b_j \exp\left(\frac{\lambda_i + \mu_j - c(x_i, y_j)}{\varepsilon}\right) + \varepsilon$$

□

Since the problem is convex and the constraints are linear, according to [Boyd, 2004, Section 5.2.3] the supremum in the dual is attained for some (λ^*, μ^*). [Boyd, 2004, Section 5.5.2] then implies that the optimal P^* satisfies

$$P_{ij}^* = a_i b_j \exp\left(\frac{\lambda_i^* + \mu_j^* - c(x_i, y_j)}{\varepsilon} - 1\right) = (a_i e^{\lambda_i^*/\varepsilon - 1/2}) e^{-c(x_i, y_j)/\varepsilon} (b_j e^{\mu_j^*/\varepsilon - 1/2})$$

Letting $K \in (\mathbb{R}_+^*)^{n \times m}$ be the matrix with entries $e^{-c(x_i, y_j)/\varepsilon}$, we infer that there exists $u \in (\mathbb{R}_+^*)^n$ and $v \in (\mathbb{R}_+^*)^m$ such that $P^* = \operatorname{diag}(u) K \operatorname{diag}(v)$. By Sinkhorn's theorem [Idel, 2016], such a $P^* \in U(a,b)$ is unique, so if we find u and v such that $\operatorname{diag}(u) K \operatorname{diag}(v) \in U(a,b)$, we can conclude that $\operatorname{diag}(u) K \operatorname{diag}(v)$ is optimal.

Therefore, the problem boils down to finding $u \in (\mathbb{R}_+^*)^n$ and $v \in (\mathbb{R}_+^*)^m$ such that $\operatorname{diag}(u) K \operatorname{diag}(v) \in U(a,b)$. This is known as a **matrix scaling** problem and there is a wealth of literature on this topic. The constraint rewrites as $u \odot (Kv) = a$ and $v \odot (K^T u) = b$, and this motivates the iterative Sinkhorn's algorithm: let $v^{(0)} = \mathbb{1}_m$ and

$$u^{(\ell+1)} = \frac{a}{K v^{(\ell)}} \text{ and } v^{(\ell+1)} = \frac{b}{K^T u^{(\ell+1)}}$$

where the quotients are entrywise.

3.3 Algorithmic complexity of the Sinkhorn

The algorithmic complexity of Sinkhorn's algorithm is a natural object of study: given $\gamma > 0$, how many iterations are sufficient to get a $\hat{P} \in U(a,b)$ such that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j) \hat{P}_{ij} \leq \inf_{P \in U(a,b)} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j) P_{ij} + \gamma$$

Note that we seek to approximate the optimal value of the un-regularized problem, and Sinkhorn's algorithm is based on the regularized version of the problem. Of course, the regularization parameter ε is chosen as a function of γ.

The algorithmic complexity is still an active topic of research. In [Altschuler et al., 2017] it was shown that it can be done in $O\left(\frac{n^2 \log n}{\gamma^3}\right)$ time. In [Dvurechensky et al., 2018] it was reduced to $O\left(\frac{n^2 \log n}{\gamma^2}\right)$.

While reading [Altschuler et al., 2017] I noticed that something was wrong in the proof of Lemma 3. Their proof yields an upper bound $s + \log\left(\frac{1}{\ell}\right)$, instead of the claimed $\log\left(\frac{s}{\ell}\right)$. In a few days, I managed to devise a complicated proof for an upper bound in $\log\left(1 + \frac{2s}{\ell}\right)$. After some time, I noticed there was also an issue in Algorithm 3. In the case $k = 1$,

$$A^{(1)} := \mathbf{D}\left(\exp\left(x^1\right)\right) A \, \mathbf{D}\left(\exp\left(y^1\right)\right) = \mathbf{D}\left(\frac{r_i}{r_i\left(\frac{A}{\|A\|_1}\right)}\right) A = \|A\|_1 \mathbf{D}\left(\frac{r_i}{r_i(A)}\right) A$$

Hence $\|A^{(1)}\|_1 = \|A\|_1$ and Lemma 2 fails for $k = 2$, since its proof relies critically on $\|A^{(1)}\|_1 = 1$. I noted that it could be fixed by defining $y^0 := -\log \|A\|_1$. Incidentally, this also fixes the issue in Lemma 3.

I sent an email to the authors and I got a reply from Altschuler who recognized the issues and proposed to set $A^{(k)} = \mathbf{D}\left(\exp\left(x^1\right)\right) A^{(0)}\, \mathbf{D}\left(\exp\left(y^1\right)\right)$ in line 11 of Algorithm 3, and to replace A_{ij} with $A_{ij}^{(0)}$ in the definition of the potential function.

I also spent time reading the details of [Dvurechensky et al., 2018] and thinking of a way to obtain sharper bounds.

Chapter 4

Statistical properties of Optimal Transport

4.1 Introduction

In this chapter we consider measures on $\mathbb{R}^d$. In practice, the measure α is only known through the empirical measure $\hat{\alpha}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ where $(X_1, \ldots, X_n)$ are i.i.d samples $\sim \alpha^{\otimes n}$. A straightforward consequence of Glivenko-Cantelli theorem is that almost surely $\hat{\alpha}_n \Rightarrow \alpha$. Wasserstein distances can be used to measure the speed of this convergence. The rate at which $W_1(\alpha, \hat{\alpha}_n)$ goes to 0 was first studied in [Dudley, 1969] and the properties of $W_p(\alpha, \hat{\alpha}_n)$ remain an active topic of research [Weed et al., 2019].

The paper [Genevay et al., 2018] studies the statistical properties of regularized optimal transport with an arbitrary cost function under the strong assumption that the measures have bounded supports. The recent paper [Mena and Weed, 2019] generalizes the previous one to sub-gaussian measures, but the cost is restrained to the squared Euclidean norm (in other words, they consider the 2-Wasserstein distance).

We have worked on a generalization of Theorem 1 in [Genevay et al., 2018] to measures with unbounded supports. We present our work in the next section.

4.2 Extension of Theorem 1 in [Genevay et al., 2018]

We suppose that the cost function is $c(x, y) := \|x - y\|_2^p$ where $p \geq 1$. In other words, we consider p-Wasserstein distances. We let $W_{p,\varepsilon}^p(\alpha, \beta)$ denote the minimum of Regularized KP with regularization parameter ε. Unlike [Mena and Weed, 2019] we do not assume that the measures are sub-gaussian, we only suppose that they have finite moments of sufficiently large order. The theorem we consider is not covered or generalized in [Mena and Weed, 2019].

We prove the following theorem.

Theorem. Let α, β be probability measures on $\mathbb{R}^d$ and $p \geq 1$. Suppose α has a moment of order $\max(d+1, p)$ and β has a moment of order p. Then, as $\varepsilon \to 0$, the following inequality on the p-Wasserstein holds:

$$0 \leqslant W_{p,\varepsilon}^p(\alpha, \beta) - W_p^p(\alpha, \beta) \lesssim \varepsilon \log\left(\tfrac{1}{\varepsilon}\right)$$

where the sign $\lesssim$ hides constants depending on d, p and the moments of α and β.

1. Some facts

The bound $0 \leqslant W_{p,\varepsilon}^p(\alpha, \beta) - W_p^p(\alpha, \beta) \leqslant \big(C(\pi^\Delta) - C(\pi_0)\big) + \varepsilon H(\pi^\Delta)$ is still valid for measures with unbounded supports.

It is more convenient for later computations to suppose that the cube Q_0^Δ is centered at the origin. Therefore, we deviate from [Genevay et al., 2018] and define for $i \in \mathbb{Z}^d$,

$$Q_i^\Delta := \left[\Delta\left(i_1 - \frac{1}{2}\right), \Delta\left(i_1 + \frac{1}{2}\right)\right) \times \ldots \times \left[\Delta\left(i_d - \frac{1}{2}\right), \Delta\left(i_d + \frac{1}{2}\right)\right)$$

We define the block approximation π^Δ as in [Genevay et al., 2018] (see the definition there).

Proposition 4.2.1. π^Δ is dominated by $\alpha \otimes \beta$ and

$$\frac{d\pi^\Delta}{d\alpha \otimes \beta}(z) = \sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} 1_{Q_{ij}^\Delta}(z)$$

Proof. Let $C \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ be such that $\alpha \otimes \beta(C) = 0$. This implies $\forall i, j,\ \alpha \otimes \beta(C \cap Q_{ij}^\Delta) = 0$. Then

$$\pi^\Delta(C) = \sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} \, \alpha|_{Q_i^\Delta} \otimes \beta|_{Q_j^\Delta} \left(C \cap Q_{ij}^\Delta\right)$$

and it suffices to prove $\alpha|_{Q_i^\Delta} \otimes \beta|_{Q_j^\Delta}(C \cap Q_{ij}^\Delta) = \alpha \otimes \beta(C \cap Q_{ij}^\Delta)$. This is clearly true when $C = A \times B$. Since the products of Borel sets are a π-system, the equality holds on $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$. Let us compute the density: for $C \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$,

$$\begin{aligned}
\pi^\Delta(C) &= \sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} \, \alpha|_{Q_i^\Delta} \otimes \beta|_{Q_j^\Delta}(C \cap Q_{ij}^\Delta) \\
&= \sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} \alpha \otimes \beta(C \cap Q_{ij}^\Delta) \\
&= \int 1_C(z) \left(\sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} 1_{Q_{ij}^\Delta}(z) \right) d\alpha \otimes \beta(z)
\end{aligned}$$

Hence $\dfrac{d\pi^\Delta}{d\alpha \otimes \beta}(z) = \displaystyle\sum_{ij} \frac{\pi_0(Q_{ij}^\Delta)}{\alpha_i^\Delta \beta_j^\Delta} 1_{Q_{ij}^\Delta}(z)$. □

2. Upper bound on $C(\pi^\Delta) - C(\pi_0)$

Proposition 4.2.2.

$$C(\pi^\Delta) - C(\pi_0) \leq \sum_{ij} \pi_0(Q_{ij}^\Delta) \left(\sup_{(x,y) \in Q_{ij}^\Delta} \|\nabla c(x,y)\|_2 \right) \operatorname{diam}(Q_{ij}^\Delta)$$

Proof.

$$\begin{aligned}
C(\pi^\Delta) &= \int c(x,y) d\pi^\Delta(x,y) \\
&= \sum_{ij} \int_{Q_{ij}^\Delta} c(x,y) d\pi^\Delta(x,y) \\
&\leq \sum_{ij} \pi^\Delta(Q_{ij}^\Delta) \sup_{(x,y) \in Q_{ij}^\Delta} c(x,y) \\
&= \sum_{ij} \pi_0(Q_{ij}^\Delta) \sup_{(x,y) \in Q_{ij}^\Delta} c(x,y)
\end{aligned}$$

and similarly $C(\pi_0) \geq \sum_{ij} \pi_0(Q_{ij}^\Delta) \inf_{(x,y) \in Q_{ij}^\Delta} c(x,y)$.
Note that

$$\begin{aligned}
\left| \sup_{(x,y) \in Q_{ij}^\Delta} c(x,y) - \inf_{(x,y) \in Q_{ij}^\Delta} c(x,y) \right| &= \left| \max_{(x,y) \in Q_{ij}^\Delta} c(x,y) - \min_{(x,y) \in Q_{ij}^\Delta} c(x,y) \right| \\
&\leq \left(\sup_{(x,y) \in Q_{ij}^\Delta} \|\nabla c(x,y)\|_2 \right) \operatorname{diam}(Q_{ij}^\Delta)
\end{aligned}$$

hence the claim. □

Proposition 4.2.3.

$$\sup_{(x,y)\in Q_{ij}^{\Delta}} \|\nabla c(x,y)\|_2 \leq \begin{cases} \sqrt{2}p\Delta^{p-1}(d^{(p-1)/2}+\|i-j\|_2^{p-1})2^{p-2} & \text{if } p\geq 2 \\ \sqrt{2}p\Delta^{p-1}(d^{(p-1)/2}+\|i-j\|_2^{p-1}) & \text{if } p<2 \end{cases}$$

Proof. For $(h_1,h_2)\in\mathbb{R}^d\times\mathbb{R}^d$,

$$\begin{aligned} c((x,y)+(h_1,h_2)) &= (\|x-y+h_1-h_2\|_2^2)^{p/2} \\ &= c(x,y)+p\|x-y\|_2^{p-2}\langle\begin{pmatrix} h_1 \\ h_2 \end{pmatrix},\begin{pmatrix} x-y \\ -(x-y) \end{pmatrix}\rangle + o(\|h\|) \end{aligned}$$

Hence $\nabla c(x,y) = p\|x-y\|_2^{p-2}\begin{pmatrix} x-y \\ -(x-y) \end{pmatrix}$ and

$$\|\nabla c(x,y)\|_2 = p\|x-y\|_2^{p-2}\sqrt{2}\|x-y\|_2 = \sqrt{2}p\|x-y\|_2^{p-1}$$

Let $C=\left[-\frac{\Delta}{2},\frac{\Delta}{2}\right]^d$ be the cube centered at 0 with side Δ and $(i,j)\in\mathbb{Z}^d\times\mathbb{Z}^d$ be fixed. Consider $x\in Q_i^{\Delta}$ and $y\in Q_j^{\Delta}$. Note that

$$\begin{aligned} \|x-y\|_2 &= \|(x-\Delta i)-(y-\Delta j)+\Delta(i-j)\|_2 \\ &\leq \|(x-\Delta i)-(y-\Delta j)\|_2+\Delta\|i-j\|_2 \\ &\leq \operatorname{diam} C+\Delta\|i-j\|_2 \\ &\leq \sqrt{d}\Delta+\Delta\|i-j\|_2 \end{aligned}$$

Thus for $(x,y)\in Q_{ij}^{\Delta}$ we have $\|\nabla c(x,y)\|_2 \leq \sqrt{2}p[\sqrt{d}\Delta+\Delta\|i-j\|_2]^{p-1}$

$$\leq \begin{cases} \sqrt{2}p\Delta^{p-1}(d^{(p-1)/2}+\|i-j\|_2^{p-1})2^{p-2} & \text{if } p\geq 2 \\ \sqrt{2}p\Delta^{p-1}(d^{(p-1)/2}+\|i-j\|_2^{p-1}) & \text{if } p<2 \end{cases}$$

□

It remains to deal with $\operatorname{diam}(Q_{ij}^{\Delta})$. As a subset of $\mathbb{R}^{2d}$, Q_{ij}^{Δ} is a cube with side Δ, so we have

$$\operatorname{diam}(Q_{ij}^{\Delta}) \leq \sqrt{2d}\Delta$$

Combining Proposition 4.2.2 and 4.2.3, we get the following proposition.

Proposition 4.2.4.

$$C(\pi^\Delta) - C(\pi_0) \leq \begin{cases} \sum_{ij} \pi_0(Q_{ij}^\Delta) \left(2^{p-3/2} p\Delta^{p-1}[d^{(p-1)/2} + \|i-j\|_2^{p-1}]\right) \left(\sqrt{2d}\Delta\right) & \text{if } p \geq 2 \\ \sum_{ij} \pi_0(Q_{ij}^\Delta) \left(\sqrt{2} p\Delta^{p-1}[d^{(p-1)/2} + \|i-j\|_2^{p-1}]\right) \left(\sqrt{2d}\Delta\right) & \text{if } p < 2 \end{cases}$$

$$= \begin{cases} 2^{p-1} p d^{p/2}\Delta^p + 2^{p-1} p\sqrt{d}\Delta^p \left[\sum_{ij} \pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right] & \text{if } p \geq 2 \\ 2pd^{p/2}\Delta^p + 2p\sqrt{d}\Delta^p \left[\sum_{ij} \pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right] & \text{if } p < 2 \end{cases} \tag{1}$$

The next proposition bounds $\Delta^{p-1}\left[\sum_{ij} \pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right]$.

Proposition 4.2.5.

$$\Delta^{p-1}\left[\sum_{ij} \pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right] \leq \begin{cases} 2^{2p-3}\left(\int \|x\|_2^{p-1} d\alpha(x) + \int \|y\|_2^{p-1} d\beta(y)\right) & \text{if } p \geq 2 \\ 2^{p-1}\left(\int \|x\|_2^{p-1} d\alpha(x) + \int \|y\|_2^{p-1} d\beta(y)\right) & \text{if } p < 2 \end{cases}$$

Proof. Starting with the bound $\|i-j\|_2^{p-1} \leq (\|i\|_2^{p-1} + \|j\|_2^{p-1})2^{p-2}$ if $p \geq 2$ and $\|i-j\|_2^{p-1} \leq (\|i\|_2^{p-1} + \|j\|_2^{p-1})$ otherwise, we get

$$\Delta^{p-1}\left[\sum_{ij} \pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right] \leq \begin{cases} 2^{p-2}\Delta^{p-1}\sum_i \left(\alpha(Q_i^\Delta) + \beta(Q_i^\Delta)\right)\|i\|_2^{p-1} & \text{if } p \geq 2 \\ \Delta^{p-1}\sum_i \left(\alpha(Q_i^\Delta) + \beta(Q_i^\Delta)\right)\|i\|_2^{p-1} & \text{if } p < 2 \end{cases} \tag{2}$$

and it suffices to bound $\Delta^{p-1}\sum_i \alpha(Q_i^\Delta)\|i\|_2^{p-1}$.

Let $i \in \mathbb{Z}^d$ be fixed and $x \in Q_i^\Delta$. Then for every k, $|x_k - \Delta i_k| \leq \frac{\Delta}{2}$, and the triangle inequality yields $\Delta\left(|i_k| - \frac{1}{2}\right) \leq |x_k|$. If $|i_k| \geq 1$, we have $-\frac{|i_k|}{2} \leq -\frac{1}{2}$ and $\frac{\Delta}{2}|i_k| \leq |x_k|$. Note that this trivially holds when $i_k = 0$, so that

$$x \in Q_i^\Delta \implies \frac{\Delta}{2}\|i\|_2 \leq \|x\|_2$$

Hence

$$\int \|x\|_2^{p-1} d\alpha(x) = \sum_i \int_{Q_i^\Delta} \|x\|_2^{p-1} d\alpha(x) \geq \frac{\Delta^{p-1}}{2^{p-1}} \sum_i \alpha(Q_i^\Delta)\|i\|_2^{p-1}$$

that is

$$\Delta^{p-1}\sum_i \alpha(Q_i^\Delta)\|i\|_2^{p-1} \leq 2^{p-1}\int \|x\|_2^{p-1} d\alpha(x)$$

Plugging this back in (2),

$$\Delta^{p-1}\left[\sum_{ij}\pi_0(Q_{ij}^\Delta)\|i-j\|_2^{p-1}\right] \leq \begin{cases} 2^{2p-3}\left(\int \|x\|_2^{p-1}d\alpha(x)+\int \|y\|_2^{p-1}d\beta(y)\right) & \text{if } p\geq 2 \\ 2^{p-1}\left(\int \|x\|_2^{p-1}d\alpha(x)+\int \|y\|_2^{p-1}d\beta(y)\right) & \text{if } p<2 \end{cases}$$

□

Combining (1) and Proposition 4.2.5, we get

$$C(\pi^\Delta)-C(\pi_0) \leq \begin{cases} 2^{p-1}pd^{p/2}\Delta^p + 2^{3p-4}p\sqrt{d}\Delta\left(\int \|x\|_2^{p-1}d\alpha(x)+\int \|y\|_2^{p-1}d\beta(y)\right) & \text{if } p\geq 2 \\ 2pd^{p/2}\Delta^p + 2^{p}p\sqrt{d}\Delta\left(\int \|x\|_2^{p-1}d\alpha(x)+\int \|y\|_2^{p-1}d\beta(y)\right) & \text{if } p<2 \end{cases} \tag{3}$$

3. Upper bound on $H(\pi^\Delta)$

Let us bound $H(\pi^\Delta)$ from above. As in [Genevay et al., 2018], using the density of π^Δ w.r.t $\alpha\otimes\beta$ we get

$$H(\pi^\Delta) = \sum_{ij}\log\left(\frac{\pi_0(Q_{ij}^\Delta)}{\alpha(Q_i^\Delta)\beta(Q_j^\Delta)}\right)\pi_0(Q_{ij}^\Delta)$$

Since π_0 is a coupling, we have $\pi_0(Q_{ij}^\Delta)\leq \min(\alpha(Q_i^\Delta),\beta(Q_j^\Delta))$, thus

$$\frac{\pi_0(Q_{ij}^\Delta)}{\alpha(Q_i^\Delta)\beta(Q_j^\Delta)} \leq \min\left(\frac{1}{\alpha(Q_i^\Delta)},\frac{1}{\beta(Q_j^\Delta)}\right)$$

and

$$\sum_{ij}\log\left(\frac{\pi_0(Q_{ij}^\Delta)}{\alpha(Q_i^\Delta)\beta(Q_j^\Delta)}\right)\pi_0(Q_{ij}^\Delta) \leq \min\left(\sum_i \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right),\sum_j \beta(Q_j^\Delta)\log\left(\frac{1}{\beta(Q_j^\Delta)}\right)\right)$$

It suffices to bound the quantity $H^\Delta(\alpha) := \sum_i \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)$ (note the minus sign compared to [Genevay et al., 2018]).
Consider some $M>0$ (that will be chosen later) and write

$$H^\Delta(\alpha) = \sum_{\substack{i\\ \|i\|_\infty\leq M}}\alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) + \sum_{\substack{i\\ \|i\|_\infty> M}}\alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) \tag{4}$$

Proposition 4.2.6. If $M \geq 1$,

$$\sum_{\substack{i \\ \|i\|_\infty \leq M}} \alpha(Q_i^\Delta) \log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) \leq d \log(2\lfloor M \rfloor + 1)$$

Proof. Let α_M^Δ be the distribution with density w.r.t $\lambda_{\mathbb{R}^d}$ given by

$$f_{\alpha,\Delta,M}(x) = \frac{1}{\sum\limits_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)} \sum_{\substack{i \\ \|i\|_\infty \leq M}} \frac{\alpha(Q_i^\Delta)}{\Delta^d} 1_{Q_i^\Delta}(x)$$

For $i \in \mathbb{Z}^d$ such that $\|i\|_\infty \leq M$ and $x \in Q_i^\Delta$, we have for all k, $|x_k - \Delta i_k| \leq \frac{\Delta}{2}$, hence $|x_k| \leq \Delta(M + \frac{1}{2})$. As a consequence, α_M^Δ is supported on a cube of side $\Delta(2M+1)$.

The differential entropy of α_M^Δ w.r.t $\lambda_{\mathbb{R}^d}$ is

$$\begin{aligned} H_{\lambda_{\mathbb{R}^d}}(\alpha_M^\Delta) &= \frac{1}{\sum\limits_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)} \sum_{\substack{i \\ \|i\|_\infty \leq M}} \int_{Q_i^\Delta} \frac{\alpha(Q_i^\Delta)}{\Delta^d} \log\left(\frac{\Delta^d}{\alpha(Q_i^\Delta)} \sum_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)\right) d\lambda_{\mathbb{R}^d}(x) \\ &= \frac{1}{\sum\limits_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)} \left[\sum_{\substack{i \\ \|i\|_\infty \leq M}} \alpha(Q_i^\Delta) \log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)\right] - d \log\left(\frac{1}{\Delta}\right) + \log\left(\sum_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)\right) \end{aligned}$$

Hence

$$\sum_{\substack{i \\ \|i\|_\infty \leq M}} \alpha(Q_i^\Delta) \log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) = \sum_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta) \left[H_{\lambda_{\mathbb{R}^d}}(\alpha_M^\Delta) + d \log\left(\frac{1}{\Delta}\right) - \log\left(\sum_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)\right)\right]$$

It is not clear how to bound $\log\left(\sum\limits_{\substack{j \\ \|j\|_\infty \leq M}} \alpha(Q_j^\Delta)\right)$ from below, so this path is likely a dead end.

We make use of Jensen's inequality instead. Note that

$$\begin{aligned}\sum_{\substack{i\\ \|i\|_\infty\le M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) &= \sum_{\substack{j\\ \|j\|_\infty\le M}} \alpha(Q_j^\Delta)\cdot \sum_{\substack{i\\ \|i\|_\infty\le M}} \frac{\alpha(Q_i^\Delta)}{\sum\limits_{\substack{j\\ \|j\|_\infty\le M}}\alpha(Q_j^\Delta)}\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)\\ &\le \sum_{\substack{j\\ \|j\|_\infty\le M}} \alpha(Q_j^\Delta)\cdot\log\left(\frac{|\{i\in\mathbb{Z}^d,\ \|i\|_\infty\le M\}|}{\sum\limits_{\substack{j\\ \|j\|_\infty\le M}}\alpha(Q_j^\Delta)}\right)\\ &= \sum_{\substack{j\\ \|j\|_\infty\le M}} \alpha(Q_j^\Delta)\cdot\log\left(\frac{(2\lfloor M\rfloor+1)^d}{\sum\limits_{\substack{j\\ \|j\|_\infty\le M}}\alpha(Q_j^\Delta)}\right)\end{aligned}$$

The function $[0,a]\to\mathbb{R}, x\mapsto x\log\left(\frac{a}{x}\right)$ where $a>0$ is increasing over $[0,\frac{a}{e}]$ and decreasing over $[\frac{a}{e},a]$. If $1\le\frac{a}{e}$, the function restricted to $[0,1]$ attains its maximum at 1. The inequality $\frac{1}{2}\left(\exp(\frac{1}{d})-1\right)\le\frac{1}{2}\left(\exp(1)-1\right)<0.9$ implies that as long as $M\ge 1$,

$$\sum_{\substack{i\\ \|i\|_\infty\le M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)\le\log((2\lfloor M\rfloor+1)^d)=d\log(2\lfloor M\rfloor+1)$$

□

Remark 4.2.7. It is remarkable that this bound is independent of α.

In (4) it remains to bound

$$\sum_{\substack{i\\ \|i\|_\infty> M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)$$

Proposition 4.2.8. Suppose α has a moment of order q for q sufficiently large. Let X be a random vector with distribution α and $M\ge\dfrac{2}{\Delta}\left(eE(\|X\|_2^q)\right)^{1/q}$. Then

$$\begin{aligned}&\sum_{\substack{i\\ \|i\|_\infty> M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)\\ &\le\frac{E(\|X\|_2^q)\,2^q d3^{d-1}}{\Delta^q M^{q-d}}\left(\frac{q}{q-d}\log(M+1)+\frac{q}{2}\log d+\frac{q}{(q-d)^2}+\log\left(\frac{\Delta^q}{2^qE(\|X\|_2^q)}\right)\right)\end{aligned}$$

Proof. Note that

$$\begin{aligned}\alpha(Q_i^\Delta) &= P(X \in Q_i^\Delta)\\ &\le P(\|X\|_2 \ge \frac{\Delta}{2}\|i\|_2)\\ &\le \frac{E(\|X\|_2^q)\ 2^q}{\Delta^q}\frac{1}{\|i\|_2^q}\end{aligned}$$

The function $[0,1] \to \mathbb{R}, x \mapsto x\log\left(\frac{1}{x}\right)$ is increasing over $[0,\frac{1}{e}]$. Assuming $M \ge \dfrac{2}{\Delta}\left(eE(\|X\|_2^q)\right)^{1/q}$, we get

$$\sum_{\substack{i\\ \|i\|_\infty > M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right) \le \frac{E(\|X\|_2^q)\ 2^q}{\Delta^q}\left(q\sum_{\substack{i\\ \|i\|_\infty > M}}\frac{\log(\|i\|_2)}{\|i\|_2^q} + \log\left(\frac{\Delta^q}{2^qE(\|X\|_2^q)}\right)\sum_{\substack{i\\ \|i\|_\infty > M}}\frac{1}{\|i\|_2^q}\right) \tag{5}$$

Let us find an upper bound on $\displaystyle\sum_{\substack{i\\ \|i\|_\infty > M}}\frac{1}{\|i\|_2^q}$. The other sum will be dealt with similarly.

Write

$$\sum_{\substack{i\\ \|i\|_\infty > M}}\frac{1}{\|i\|_2^q} = \sum_{\substack{k\in\mathbb{N}\\ k>M}}\sum_{\substack{i\\ \|i\|_\infty = k}}\frac{1}{\|i\|_2^q}$$

Since $k > M \iff k > \lfloor M \rfloor$ and $\dfrac{1}{\|i\|_2} \le \dfrac{1}{\|i\|_\infty}$,

$$\sum_{\substack{i\\ \|i\|_\infty > M}}\frac{1}{\|i\|_2^q} \le \sum_{k>\lfloor M\rfloor}\sum_{\substack{i\\ \|i\|_\infty = k}}\frac{1}{\|i\|_\infty^q} = \sum_{k>\lfloor M\rfloor}\frac{|\{i\in\mathbb{Z}^d,\ \|i\|_\infty = k\}|}{k^q}$$

Note that

$$\begin{aligned}|\{i\in\mathbb{Z}^d,\ \|i\|_\infty = k\}| &= |\{i\in\mathbb{Z}^d,\ \forall j, |i_j| \le k \text{ and } \exists j_0, |i_{j_0}| = k\}|\\ &= |\bigcup_{j_0=1}^{d}\{i\in\mathbb{Z}^d,\ \forall j, |i_j| \le k \text{ and } |i_{j_0}| = k\}|\\ &\le \sum_{j_0=1}^{d}|\{i\in\mathbb{Z}^d,\ \forall j, |i_j| \le k \text{ and } |i_{j_0}| = k\}|\\ &= d|\{i\in\mathbb{Z}^d,\ \forall j, |i_j| \le k \text{ and } |i_1| = k\}|\\ &= d(2k+1)^{d-1}\end{aligned}$$

This yields

$$\begin{aligned}\sum_{\substack{i\\ \|i\|_\infty > M}} \frac{1}{\|i\|_2^q} &\leq \sum_{k>\lfloor M\rfloor} d2^{d-1}\frac{1}{k^{q-d+1}}\left(1+\frac{1}{2k}\right)^{d-1}\\ &\leq d2^{d-1}\left(1+\frac{1}{2M}\right)^{d-1}\sum_{k>\lfloor M\rfloor}\frac{1}{k^{q-d+1}}\\ &\leq d2^{d-1}\left(1+\frac{1}{2M}\right)^{d-1}\int_{\lfloor M\rfloor+1}^{\infty}\frac{1}{t^{q-d+1}}dt\\ &\leq d2^{d-1}\left(1+\frac{1}{2M}\right)^{d-1}\frac{1}{M^{q-d}}\end{aligned}$$

Regarding $\displaystyle\sum_{\substack{i\\ \|i\|_\infty > M}} \frac{\log(\|i\|_2)}{\|i\|_2^q}$, using the inequality $\|i\|_2 \leq \sqrt{d}\|i\|_\infty$,

$$\sum_{\substack{i\\ \|i\|_\infty > M}} \frac{\log(\|i\|_2)}{\|i\|_2^q} \leq \frac{\log d}{2}\sum_{k>\lfloor M\rfloor}\sum_{\substack{i\\ \|i\|_\infty = k}}\frac{1}{\|i\|_\infty^q} + \sum_{k>\lfloor M\rfloor}\sum_{\substack{i\\ \|i\|_\infty = k}}\frac{\log(\|i\|_\infty)}{\|i\|_\infty^q}$$

and it suffices to bound the summand on the far right. The function $x \mapsto \frac{\log(x)}{x^a}$ is decreasing for $x \geq e^{1/a}$.
Since $M \geq 1$ and q is an integer, we have $q-d+1 \geq 2$, hence $\lfloor M\rfloor + 1 \geq 2 \geq e^{1/(q-d+1)}$ thus

$$\begin{aligned}\sum_{k>\lfloor M\rfloor}\sum_{\substack{i\\ \|i\|_\infty = k}}\frac{\log(\|i\|_\infty)}{\|i\|_\infty^q} &\leq d2^{d-1}\left(1+\frac{1}{2M}\right)^{d-1}\int_{\lfloor M\rfloor+1}^{\infty}\frac{\log t}{t^{q-d+1}}dt\\ &\leq d2^{d-1}\left(1+\frac{1}{2M}\right)^{d-1}\frac{(q-d)\log(M+1)+1}{(q-d)^2M^{q-d}}\end{aligned}$$

Plugging everything back into (5), we get

$$\begin{aligned}&\sum_{\substack{i\\ \|i\|_\infty > M}} \alpha(Q_i^\Delta)\log\left(\frac{1}{\alpha(Q_i^\Delta)}\right)\\ &\leq \frac{E(\|X\|_2^q)\;2^q d2^{d-1}}{\Delta^q M^{q-d}}\left(1+\frac{1}{2M}\right)^{d-1}\left(\frac{q}{2}\log d + \frac{q}{q-d}\log(M+1) + \frac{q}{(q-d)^2} + \log\left(\frac{\Delta^q}{2^qE(\|X\|_2^q)}\right)\right)\\ &\leq \frac{E(\|X\|_2^q)\;2^q d3^{d-1}}{\Delta^q M^{q-d}}\left(\frac{q}{q-d}\log(M+1) + \frac{q}{2}\log d + \frac{q}{(q-d)^2} + \log\left(\frac{\Delta^q}{2^qE(\|X\|_2^q)}\right)\right)\end{aligned}$$

□

Plugging everything back into (4),

$$\begin{aligned}H^\Delta(\alpha) &\le d\log(2\lfloor M\rfloor + 1) + \frac{E(\|X\|_2^q)\ 2^q d 3^{d-1}}{\Delta^q M^{q-d}}\left(\frac{q}{q-d}\log(M+1) + \frac{q}{2}\log d + \frac{q}{(q-d)^2} + \log\left(\frac{\Delta^q}{2^q E(\|X\|_2^q)}\right)\right)\\ &\le d\log(3M) + \frac{E(\|X\|_2^q)\ 2^q d 3^{d-1}}{\Delta^q M^{q-d}}\left(\frac{q}{q-d}\log(2M) + \frac{q}{2}\log d + \frac{q}{(q-d)^2} + \log\left(\frac{\Delta^q}{2^q E(\|X\|_2^q)}\right)\right)\end{aligned} \tag{6}$$

It remains to pick an M that minimizes the RHS. The optimal M cannot be found computationally, however it is not difficult to find a reasonable choice (up to multiplicative constants). Remember that we assumed $M \ge 1$ and $M \ge \dfrac{2}{\Delta}\left(eE(\|X\|_2^q)\right)^{1/q}$. Because of the second assumption, M is expected to be quite large, so that (6) is approximately

$$H^\Delta(\alpha) \lesssim d\log M + \tfrac{C\log M}{\Delta^q M^{q-d}}$$

It is minimized in M when both summands have the same order, i.e. when $M \propto \frac{1}{\Delta^{\frac{q}{q-d}}}$. Plugging this choice of M back in (6) yields for $q \in (d, d + \frac{d}{C_1})$

$$H^\Delta(\alpha) \lesssim d\log\left(\tfrac{1}{\Delta^{\frac{q}{q-d}}}\right) + C_1 q\log\Delta = \tfrac{q(d-C_1(q-d))}{q-d}\log\left(\tfrac{1}{\Delta}\right)$$

Together with (3), this yields

$$0 \leqslant W_{p,\varepsilon}^p(\alpha,\beta) - W_p^p(\alpha,\beta) \lesssim C_2\Delta^p + C_3\Delta + \varepsilon\tfrac{q(d-C_1(q-d))}{q-d}\log\left(\tfrac{1}{\Delta}\right)$$

As ε goes to 0, the optimal Δ is expected to go to 0 as well, so that $C_2\Delta^p$ is negligible and the bound turns into

$$0 \leqslant W_{p,\varepsilon}^p(\alpha,\beta) - W_p^p(\alpha,\beta) \lesssim C_4\Delta + \varepsilon\tfrac{q(d-C_1(q-d))}{q-d}\log\left(\tfrac{1}{\Delta}\right)$$

This is similar to what [Genevay et al., 2018] obtains, and we conclude similarly with the following theorem.

Theorem 4.2.9. Let α, β be probability measures on $\mathbb{R}^d$ and $p \ge 1$. Suppose α has a moment of order $\max(d+1, p)$ and β has a moment of order p. Then, as $\varepsilon \to 0$, the following inequality on the p-Wasserstein holds:

$$0 \leqslant W_{p,\varepsilon}^p(\alpha,\beta) - W_p^p(\alpha,\beta) \lesssim \varepsilon\log\left(\tfrac{1}{\varepsilon}\right)$$

where the sign $\lesssim$ hides constants depending on d, p and the moments of α and β.

4.3 Next steps

We are currently investigating whether we can generalize [Mena and Weed, 2019] to the p-Wasserstein and to the case of measures with finite moments. As stated in Section 3.1, it is important to determine rigorously the space to which dual potentials belong.

Bibliography

[Altschuler et al., 2017] Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974.

[Bartal, 2019] Bartal, Y. (2019). Lecture notes on metric embedding theory and its algorithmic applications. https://moodle2.cs.huji.ac.il/nu18/pluginfile.php/327451/mod_resource/content/1/METAP19_Lecture_10.pdf.

[Bhattacharya et al., 2016] Bhattacharya, R., Lin, L., and Patrangenaru, V. (2016). *A course in mathematical statistics and large sample theory*. Springer.

[Billingsley et al., 1971] Billingsley, P., Dudley, R., et al. (1971). Convergence of probability measures. *Bulletin of the American Mathematical Society*, 77(1):25–27.

[Bonneel et al., 2015] Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45.

[Boyd, 2004] Boyd, S. (2004). *Convex optimization*. Cambridge university press.

[Clason et al., 2019] Clason, C., Lorenz, D. A., Mahler, H., and Wirth, B. (2019). Entropic regularization of continuous optimal transport problems. *arXiv preprint arXiv:1906.01333*.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

[Dudley, 1969] Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

[Dvurechensky et al., 2018] Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*.

[Fakcharoenphol et al., 2004] Fakcharoenphol, J., Rao, S., and Talwar, K. (2004). A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497.

[Genevay, 2019] Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis.

[Genevay et al., 2018] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2018). Sample complexity of Sinkhorn divergences. *arXiv preprint arXiv:1810.02733*.

[Genevay et al., 2016] Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448.

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

[Idel, 2016] Idel, M. (2016). A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*.

[Klenke, 2013] Klenke, A. (2013). *Probability theory: a comprehensive course*. Springer Science & Business Media.

[Kolouri et al., 2017] Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

[Le et al., 2019a] Le, T., Huynh, V., Ho, N., Phung, D., and Yamada, M. (2019a). On scalable variant of wasserstein barycenter. *arXiv preprint arXiv:1910.04483*.

[Le et al., 2019b] Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019b). Tree-sliced approximation of wasserstein distances. *arXiv preprint arXiv:1902.00342*.

[Leeb, 2018] Leeb, W. (2018). Approximating snowflake metrics by trees. *Applied and Computational Harmonic Analysis*, 45(2):405–424.

[Lehmann and Romano, 2006] Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[Lyons and Peres, 2017] Lyons, R. and Peres, Y. (2017). *Probability on trees and networks*, volume 42. Cambridge University Press.

[Mena and Weed, 2019] Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *arXiv preprint arXiv:1905.11882*.

[Peyré et al., 2019] Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

[Rabinovich and Raz, 1998] Rabinovich, Y. and Raz, R. (1998). Lower bounds on the distortion of embedding finite metric spaces in graphs. *Discrete & Computational Geometry*, 19(1):79–94.

[Rachev and Rüschendorf, 1998] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media.

[Santambrogio, 2015] Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63.

[Tsybakov, 2009] Tsybakov, A. B. (2009). Introduction to nonparametric estimation.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Weed et al., 2019] Weed, J., Bach, F., et al. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648.