Exposé L3

Optimal Transport and Stochastic Optimization

Mathieu Even
Ibrahim Merad

Supervised by
Gabriel Peyré

Contents

0.1 Notations

- $\mathcal{M}_+(\mathcal{X})$: The set of positive measures over $\mathcal{X}$
- $\mathcal{M}_+^1(\mathcal{X})$: The set of probability measures over $\mathcal{X}$
- $\mathbb{1}_n$: the vector of size n with all entries equal to 1
- Σ_n : the set of probability vectors of size n i.e. $\Sigma_n \coloneqq \left\{a \in \mathbb{R}_+^n : \ \sum_i a_i = 1\right\}$
- $a \otimes b$: for $a \in \Sigma_n, b \in \Sigma_m$ is the matrix whose coefficients are $(a_i b_j)_{i \in \{1,\dots,n\}, j \in \{1,\dots,m\}}$
- $u \odot v = (u_i v_i) \in \mathbb{R}^n$ the entrywise multiplication between two vectors $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$

Well known for its applications in economy, physics, and in many other fields of science, Optimal Transport of measures is an ancient problem, originally formulated by Monge in 1791 : his problem was to move parcels of land from one place to another, while minimizing the efforts and time needed. Later, in the middle of the twentieth century, Optimal Transport was developed by Kantorovitch (a mathematician who won a Nobel prize in economy) to be applied in economy, and was thus given a mathematical definition amenable for mathematical analysis and numerical computation. Kantorovitch's formulation corresponds to, given two measures over two spaces, finding a probability measure over the product space, that minimizes a global cost. This probability measure $\mathbb{P}$ found corresponds intuitively to the mass of land moved in Monge's example: $d\mathbb{P}(x, y)$ is the infinitesimal element of land moved from $[x, x + dx]$ to $[y, y + dy]$, and the cost we want to minimize is $\int c(x, y)d\mathbb{P}(x, y)$ with $c(x, y)$ being the cost in time or energy it takes to move a unit of land from x to y.

In the last two decades, the field of optimal transport has been very active in theory, numerical methods, for its applications in nearly every optimization problem, and its various bonds with differential equations, statistical physics, economy, machine learning and imaging. Hence, OT is a mathematical tool that finds application in mathematics itself (functional analysis, differential geometry), as well as in other domains (economy or imaging for instance).

The number of parameters used in practice makes optimal transport a very hard problem to solve. For instance, in image processing, it is used to interpolate two images, that can be seen as histograms and thus probability measures. However, the very high number of variables makes it hard to compute. Hence the need of effective methods to solve numerically this problem.

We will expose here numerical computational methods for solving this problem in particular cases (discrete and semi-discrete OT), after introducing generalities and theoretical bases.

These methods will be from different kinds: stochastic ones, to cope with the very high number of parameters, the use of entropic regularization to approximate solutions effectively, etc.

Chapter 1

Introduction to Optimal transport, theoretical foundations

1.1 Monge problem

The optimal assignment problem is one of the fundamental combinatorial optimization problems, which aims at finding the best assignment between two lists of points, minimizing a quantity (the cost). It is the first historical version of Optimal Transport : given a number of tasks and agents, one wants to assign exactly one task to each agent, minimizing the total cost of the assignment.

Let $(C_{i,j})_{1\leqslant i,j\leqslant n}$ be a cost matrix, $C_{i,j}$ being the cost of moving a *unit* from point i to point j.

The goal is to find $\sigma \in \mathcal{S}_n$ minimizing the total cost $\sum C_{i,\sigma(i)}$. This is the optimal assignment problem.

It can be generalized into Monge's problem, where the number of points at the arrival is different from the number of points we want to assign: the cost matrix is hence not a square matrix anymore, but $(C_{i,j})_{1\leqslant i\leqslant n,1\leqslant j\leqslant m}$. We do not seek a permutation anymore (impossible anyway if $n \neq m$, it is Kantorovitch's problem). Here is the new formulation of the problem:

Definition 1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets of points, and let α and β be two discrete probability measures over $\mathcal{X}$ and $\mathcal{Y}$ respectively i.e*

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \beta = \sum_{j=1}^{m} b_j \delta_{y_j} \qquad \sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j = 1$$

The Monge problem seeks a map $T\colon \{x_1, \dots, x_n\} \to \{y_1, \dots, y_m\}$ called a Monge map such that

$$\forall j \in \{1, \dots, m\} \qquad b_j = \sum_{i\colon T(x_i)=y_j} a_i$$

which condition will be denoted as $T_{\#}\alpha = \beta$. T must minimize a transportation cost defined by a function $c(x, y)$ defined over $\mathcal{X} \times \mathcal{Y}$

$$\min_T \left\{ \sum_{i=1}^{n} c(x_i, T(x_i))\colon T_{\#}\alpha = \beta \right\}$$

The optimal assignment problem corresponds to a Monge problem where $n = m$ and $C_{i,j} = c(x_i, y_j)$.

Observe that a Monge map may not always exist in some cases. For example, if $n = 2$, $m = 3$ and α and β are uniformly distributed then there is no map since halves cannot be summed into thirds.

Definition 2 (Push-forward operator)**.** *Given a map $T\colon \mathcal{X} \to \mathcal{Y}$, we define the pushforward operator $T_{\#}\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$. For a discrete measure $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ The push-forward measure is defined as*

$$T_{\#}\alpha = \sum_{i=1}^{n} a_i \delta_{T(x_i)}.$$

For general measures the push-forward measure $\beta = T_{\#}\alpha \in \mathcal{M}(\mathcal{Y})$ is defined such that

$$\forall h \in \mathcal{C}(\mathcal{Y}) \qquad \int_{\mathcal{Y}} h(y) d\beta(y) = \int_{\mathcal{X}} h(T(x)) d\alpha(x).$$

The push-forward operator's action upon a measure is to be interpreted as moving each mass element in the space so that we end up with a new measure according to the new distribution of mass. Notice that this operation preserves the total mass, therefore, the push-forward of a probability measure is still a probability measure.

Proposition 1. *Let α and β be be two measures over $\mathbb{R}^d$ that have densities ρ_α and ρ_β with respect to the Lebesgue measure. Let T be a smooth bijection of $\mathbb{R}^d$ such that $\beta = T_{\#}\alpha$ then we have the relation, thanks to the change of variable formula:*

$$\rho_\alpha(x) = |\det(J_T(x))| \rho_\beta(T(x)),$$

where J_T is the Jacobian matrix of T

We can now formulate Monge's problem for arbitrary measures.

Given two measures α and β over $\mathcal{X}$ and $\mathcal{Y}$ respectively and a cost function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, Monge's problem seeks a map $T\colon \mathcal{X} \to \mathcal{Y}$ such that $T_{\#}\alpha = \beta$ minimizing

$$\int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

The similarity with the previous formulation is obvious, only sums have turned into integrals. In case α and β have the same total mass and no atoms, a solution to this problem always exists since their mass would be distributed in infinitesimal amounts on their respective spaces [Brenier, 1991].

We now introduce a relaxed version of the problem where atomic mass can be split before mapping.

1.2 Kantorovich relaxation

The previous formulation of the assignement problem and Monge's problem has quite a few shortcomings. The assignment problem can only be considered for two equally sized sets of points, it's generalization, Monge's problem is slightly better but can also lead to problems that accept no solution that satisfies the mass conservation constraint. Additionally, these problems, allowing no splitting of mass, are combinatorial which makes them difficult to solve in a practical setting.

Kantorovich's relaxation gets rid of the fact that each source point can only be assigned to a single destination . Kantorovich proposes to allow the mass of a single point to be split and dispatched to potentially multiple destinations. In the discrete case, instead of a map sending each point to some definite location, the sought solution will be a coupling matrix $P \in \mathbb{R}_+^{n\times m}$ where an entry $P_{i,j}$ describes the fraction of mass moving from location x_i to y_j using previous notations.

Given two discrete measures $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, the set of admissible coupling matrices is

$$U(a,b) \coloneqq \left\{P \in \mathbb{R}_+^{n\times m} \colon P\mathbb{1}_m = a \quad \text{and} \quad P^T\mathbb{1}_n = b\right\}$$

The matrix P is a measure upon the product space which has a and b as marginals with respect to each space, this is the mass conservation constraint and is expressed as $P\mathbb{1}_m = a$ and $P^T\mathbb{1}_n = b$, indeed $P\mathbb{1}_m$ is the vector which has the sums of P's lines for entries and $P^T\mathbb{1}_n$ the sums of columns. An entry P_{ij} represents the *fraction of mass* moving from x_i to y_j.

Hence, Kantorovitch's optimal transport consists in finding $P \in U(a,b)$ minimizing the cost

$$\sum P_{ij}C_{ij} = \langle P, C\rangle \tag{1.1}$$

for a given cost matrix $C \in \mathbb{R}^{n\times m}$.

This formulation has the additional advantage of being symmetrical, indeed, if we have $P \in U(a,b)$ then also $P^T \in U(b,a)$ and if P is a minimizer for (1.1), then it is also one for the symmetrical problem.

Thus stated, this is a convex optimization problem, in fact it is a linear problem for which many solvers are available.

For arbitrary measures α and β, we can write this problem as the minimization on

$$\mathcal{U}(\alpha,\beta) = \left\{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\#}\pi = \alpha, P_{\mathcal{Y}\#}\pi = \beta\right\}$$

where $P_\mathcal{X}$ and $P_\mathcal{Y}$ are the projections on $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Proposition 2. *If c is a continuous function, a minimizer always exists.*

Proof. We suppose here that $\mathcal{X}$ and $\mathcal{Y}$ are compact. Let $Z \coloneqq \mathcal{X}\times\mathcal{Y}$, a compact hence, and $E = \mathcal{C}(Z)$, a normed vector space for the infinite norm (Z is compact). We know that $\mathcal{M}_+^1(Z) \subset B_{E^\star}(0,1)$ the closed unit ball of $E^\star$. According to Banach-Alaoglu's theorem, the ball is compact. Furthermore, $\mathcal{M}_+^1(Z)$ is closed. Hence,

it is compact, and the minimizer exists ($\pi \in \mathcal{M}_+^1(Z) \to \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y)$ is continuous for the weak topologies of both spaces).

□

From now on, given two measures α and β and the cost function c, we will denote the optimal cost as :

$$\mathcal{L}_c(\alpha,\beta) \coloneqq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y).$$

The Kantorovich problem can be reformulated in the formalism of random variables, indeed it is equivalent to

$$\mathcal{L}_c(\alpha,\beta) = \min_{X,Y}\{\mathbb{E}_{(X,Y)}(c(X,Y)) : X \sim \alpha, Y \sim \beta\}$$

where (X,Y) is a couple of random variables over $\mathcal{X}\times\mathcal{Y}$ and $X \sim \alpha$ means that the law of X seen as a measure over $\mathcal{X}$ corresponds to α and same for Y with β.

1.3 Metric properties of optimal transport

Optimal transport allows us to define a distance between probability distributions when the cost matrix satisfies some specific properties

Proposition 3. *Let C be a cost matrix such that $C = D^p = (D_{i,j}^p)_{i,j} \in \mathbb{R}^{n\times n}$ for some $p \geqslant 1$ where $D \in \mathbb{R}_+^{n\times n}$ is a distance over $\{1,\dots,n\}$ i.e satisfies*

(i) D is symmetrical

(ii) $D_{i,j} = 0 \iff i = j$

(iii) $\forall\, 1 \leqslant i,j,k \leqslant n\,,\ D_{i,k} \leqslant D_{i,j} + D_{j,k}$

Then

$$W_p(a,b) \coloneqq L_{D^p}(a,b)^{1/p}$$

defines a distance over Σ_n called the p-Wasserstein distance

Proof. Symmetry follows from the symmetry of Kantorovich's problem and D^p , moreover, we have $W_p(a,a) = 0$ since the optimal transport plan for this problem will be $P = \mathrm{diag}(a)$ which has zero cost since D^p has a null diagonal. All other non diagonal coefficients of D^p are strictly positive, therefore for all $a \neq b$ we have $W_p(a,b) > 0$. Proving the triangle inequality calls for some calculations, let $a,b,c \in \Sigma_n$ and P and Q be two optimal transport plans between a and b, and b and c respectively and define $\overline{b}_j \coloneqq \begin{cases} b_j & \text{if } b_j > 0 \\ 1 & \text{otherwise} \end{cases}$

We define $S \coloneqq P\,\mathrm{diag}(1/\overline{b})Q \in \mathbb{R}_+^{n\times n}$ such that $S \in U(a,c)$. Indeed, we have

$$S\mathbb{1}_n = P\,\mathrm{diag}(1/\overline{b})Q\mathbb{1}_n = P\,\mathrm{diag}(1/\overline{b})b = P(b/\overline{b}) = P\mathbb{1}_{Supp(b)} = a$$

where $\mathbb{1}_{Supp(b)}$ is the indicator vector of the support of b, this is true because $P\mathbb{1}_{Supp(b)} = P\mathbb{1}_n = a$ since $P_{i,j} = 0$ for any j such that $b_j = 0$. Conversely, we have

$$S^T\mathbb{1}_n = Q^T \operatorname{diag}(1/\bar{b})P^T\mathbb{1}_n = Q^T(b/\bar{b}) = Q^T\mathbb{1}_{Supp(b)} = c$$

for the same reasons.

Checking the triangle inequality, we have

$$\begin{aligned}
W_p(a,c) &= \left(\min_{P\in U(a,c)} \langle P, D^p\rangle\right)^{1/p} \leqslant (\langle S, D^p\rangle)^{1/p} \\
&= \left(\sum_{ik} D_{ik}^p \sum_j \frac{P_{ij}Q_{jk}}{\bar{b}_j}\right)^{1/p} \leqslant \left(\sum_{ijk} (D_{ij}+D_{jk})^p \frac{P_{ij}Q_{jk}}{\bar{b}_j}\right)^{1/p} \\
&\leqslant \left(\sum_{ijk} D_{ij}^p \frac{P_{ij}Q_{jk}}{\bar{b}_j}\right)^{1/p} + \left(\sum_{ijk} D_{jk}^p \frac{P_{ij}Q_{jk}}{\bar{b}_j}\right)^{1/p} \\
&= \left(\sum_{ij} D_{ij}^p P_{ij} \sum_k \frac{Q_{jk}}{\bar{b}_j}\right)^{1/p} + \left(\sum_{jk} D_{jk}^p Q_{jk} \sum_i \frac{P_{ij}}{\bar{b}_j}\right)^{1/p} \\
&= \left(\sum_{ij} D_{ij}^p P_{ij}\right)^{1/p} + \left(\sum_{jk} D_{jk}^p Q_{jk}\right)^{1/p} \\
&= W_p(a,b) + W_p(b,c)
\end{aligned}$$

Where the first inequality comes from the suboptimality of S, the second is the triangle inequality for D and the third is Minkowski's inequality. □

Once more, the p-Wasserstein distance's definition can be extended to the case of arbitrary measures as follows.

Proposition 4. *Assuming $\mathcal{X} = \mathcal{Y}$, let c be a cost function such that $c(x,y) = d(x,y)^p$ where d is a distance over $\mathcal{X}$ i.e satisfies*

(i) *d is symmetrical, $d(x,y) = d(y,x) \geqslant 0$*

(ii) *$d(x,y) = 0 \iff x = y$*

(iii) *$\forall\, x,y,z \in \mathcal{X}\,,\ d(x,z) \leqslant d(x,y) + d(y,z)$*

Then

$$\mathcal{W}_p(\alpha,\beta) := \mathcal{L}_{d^p}(\alpha,\beta)^{1/p}$$

defines a distance over $\mathcal{M}_+^1(\mathcal{X})$ called the p-Wasserstein distance

We do not prove this proposition, the proof can be found in [Santambrogio, 2017].

Proposition 5 (Translation invariance)**.** *Let $\mathcal{X}$ be an Euclidean space $\mathcal{X} = \mathbb{R}^d$, $d \in \mathbb{N}^\star$, for the ground cost $c(x, y) = \|x - y\|^2$, translations in the Wasserstein distance can be factored, i.e, denoting $T_\tau \colon x \mapsto x - \tau$ the translation by $\tau \in \mathbb{R}^d$, we have*

$$\mathcal{W}_2(T_{\tau\#}\alpha, T_{\tau'\#}\beta)^2 = \mathcal{W}_2(\alpha, \beta)^2 - 2\langle \tau - \tau', m_\alpha - m_\beta \rangle + \|\tau - \tau'\|^2$$

where $m_\alpha \coloneqq \int_{\mathcal{X}} x d\alpha(x) \in \mathbb{R}^d$ and $m_\beta \coloneqq \int_{\mathcal{X}} x d\beta(x) \in \mathbb{R}^d$ are the means of α and β respectively. It follows that one has the decomposition

$$\mathcal{W}_2(\alpha, \beta)^2 = \mathcal{W}_2(\overline{\alpha}, \overline{\beta})^2 + \|m_\alpha - m_\beta\|^2$$

where $\overline{\alpha}$ and $\overline{\beta}$ are the centered measures $\overline{\alpha} = T_{m_\alpha\#}\alpha$ and $\overline{\beta} = T_{m_\beta\#}\beta$

Proof. Let π be an optimal coupling for the pair (α, β), we have

$$\begin{aligned}
\mathcal{W}_2(T_{\tau\#}\alpha, T_{\tau'\#}\beta)^2 &= \int_{\mathcal{X}^2} \|(x - \tau) - (y - \tau')\|^2 \, d\pi(x, y) \\
&= \int_{\mathcal{X}^2} \|(x - y) - (\tau - \tau')\|^2 \, d\pi(x, y) \\
&= \int_{\mathcal{X}^2} \|x - y\|^2 - 2\langle \tau - \tau', x - y \rangle + \|\tau - \tau'\|^2 \, d\pi(x, y) \\
&= \mathcal{W}_2(\alpha, \beta)^2 - 2\langle \tau - \tau', \int_{\mathcal{X}^2} (x - y) d\pi(x, y) \rangle + \|\tau - \tau'\|^2 \\
&= \mathcal{W}_2(\alpha, \beta)^2 - 2\langle \tau - \tau', m_\alpha - m_\beta \rangle + \|\tau - \tau'\|^2
\end{aligned}$$

The decomposition is the particular case $\tau = m_\alpha$ and $\tau' = m_\beta$.

□

Let's remark that for instance, $\mathcal{W}_p^p(\delta_x, \delta_y) = d(x, y)$, hence $\mathcal{W}_p^p(\delta_x, \delta_y) \longrightarrow 0 \iff d(x, y) \longrightarrow 0$. This is an illustration of the fact that Wasserstein's distances are a way to quantify weak convergence.

Definition 3. *Let $(\alpha_k), \alpha \in \mathcal{M}_1^+(\mathcal{X})$.*

(α_k) converges weakly towards α if for all $g \in \mathcal{C}(\mathcal{X})$, we have $\int_{\mathcal{X}} g d\alpha_k \longrightarrow \int_{\mathcal{X}} g d\alpha$

This convergence can be shown to be equivalent to $\mathcal{W}_p(\alpha_k, \alpha) \longrightarrow 0$
[Villani, 2009]

Hence, Wasserstein metric is a natural way to compare two probability distributions, whether they are with continuous density, or discrete, where the second is derived from the first with small perturbations. For instance, $\mathcal{W}_1$ is widely used in practice to compare histograms (color histograms of two images, ...). If one image is the other but after transmission, we can compute the distance between these images, to evaluate the "noise".

1.4 Dual Problem

As a constrained convex minimization problem, the Kantorovich problem can be paired with a dual problem which will be a concave maximization problem. We now give this new formulation and it's relationship with the primal problem.

Proposition 6. *The dual of the previous problem is :*

$$L_C(a,b) = \max_{(f,g)\in R(a,b)} \langle f, a\rangle + \langle g, b\rangle$$

where

$$R(a,b) := \{(f,g) \in \mathbb{R}^n \times \mathbb{R}^m : \ \forall\, 1 \leqslant i \leqslant n \ \ 1 \leqslant j \leqslant m, \ f_i + g_j \leqslant C_{i,j}\}$$

Proof. Including the mass constraint through auxiliary potentials f and g we have :

$$\begin{aligned} L_C(a,b) &= \min_{P\geqslant 0} \max_{(f,g)\in\mathbb{R}^n\times\mathbb{R}^m} \langle C, P\rangle + \langle a - P\mathbb{1}_m, f\rangle + \langle b - P^T\mathbb{1}_n, g\rangle \\ &= \max_{(f,g)\in\mathbb{R}^n\times\mathbb{R}^m} \langle a, f\rangle + \langle b, g\rangle + \min_{P\geqslant 0} \langle C - f\mathbb{1}_m^T - \mathbb{1}_n g^T, P\rangle \end{aligned}$$

where the reordering of the min and max operators is justified by the existence of a solution for the linear program. We have

$$\min_{P\geqslant 0} \langle Q, P\rangle = \begin{cases} 0 & \text{if } Q \geqslant 0 \\ -\infty & \text{otherwise} \end{cases}$$

So the last term in the previous formula is an infinite penalty as soon as the constraint is violated and the original constraint over P is equivalent to that over f and g given by $R(a,b)$. □

A classical result about optimization problems gives the following property, called dual formulation of OT. (Santambrogio, 2015, Optimal Transport for applied mathematicians).

Proposition 7. *We define*

$$\mathcal{R}(\alpha,\beta) = \left\{(f,g) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{Y}} \textit{measurable} : \forall (x,y), f(x) + g(y) \leqslant c(x,y)\right\}.$$

We have

$$\mathcal{L}_C(\alpha,\beta) = \max\left\{\int_{\mathcal{X}} f d\alpha + \int_{\mathcal{Y}} g d\beta : (f,g) \in \mathcal{R}(\alpha,\beta)\right\}$$

The easier part of the proof of that result is that the right hand part of the equality is a lower bound to the left hand side.

1.5 Special cases

In most cases, computing optimal transport plans and distances requires numerical methods, however, in some cases, it can be done formally. We give a few examples of such cases before we go on.

1.5.1 Arbitrary measures in one dimension

Sometimes, optimal transport applications only require the one-dimensional case, for instance, it is the case when comparing the histograms of two grayscale images. The solution is then quite straightforward.

Definition 4. *Let α be a probability measure on $\mathbb{R}$ we define it's cumulative function as*

$$\forall x \in \mathbb{R},\ \mathcal{C}_\alpha(x) \coloneqq \int_{-\infty}^{x} d\alpha$$

It is a function $\mathcal{C}_\alpha \colon \mathbb{R} \to [0,1]$, we also define it's pseudo-inverse $\mathcal{C}_\alpha^{-1} \colon [0,1] \to \mathbb{R} \cup \{-\infty, +\infty\}$ by

$$\forall r \in [0,1]\,,\ \mathcal{C}_\alpha^{-1}(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty, +\infty\} :\ \mathcal{C}_\alpha(x) \geqslant r\}$$

For $p \geqslant 1$, for any probability measures $\alpha, \beta \in \mathcal{M}_+^1(\mathbb{R})$we have

$$\mathcal{W}_p(\alpha, \beta) = \left\| \mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1} \right\|_{L^p([0,1])}^p = \int_0^1 |\mathcal{C}_\alpha^{-1}(r) - \mathcal{C}_\beta^{-1}(r)|^p dr$$

For $p = 1$ we have more simply

$$\mathcal{W}_1(\alpha, \beta) = \|\mathcal{C}_\alpha - \mathcal{C}_\beta\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |\mathcal{C}_\alpha(x) - \mathcal{C}_\beta(x)| dx = \int_{\mathbb{R}} \left| \int_{-\infty}^{x} d(\alpha - \beta) \right| dx$$

Then, an optimal transport plan T such that $T_\# \alpha = \beta$ is given by $T = \mathcal{C}_\beta^{-1} \circ \mathcal{C}_\alpha$

Intuitively, the optimal coupling moves each point where it would give the same value for $\mathcal{C}_\beta$ as it does for $\mathcal{C}_\alpha$

1.5.2 Distance between Gaussians

The popular and widely used Gaussian distributions provide interesting instances of optimal transport. The optimal transport plan between two such distributions can be directly found according to their parameters in any dimension.

Let α and β be two Gaussian probability distributions over $\mathbb{R}^d$, $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$ and let ρ_α and ρ_β be their respective densities with respect to Lebesgue's measure, let T be the mapping

$$T \colon x \mapsto m_\beta + A(x - m_\alpha)$$

where

$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = A^T$$

such that $T_\# \rho_\alpha = \rho_\beta$

Indeed, we can verify that :

$$\begin{aligned}
\rho_\beta(T(x)) &= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\frac{1}{2}\langle T(x) - m_\beta, \Sigma_\beta^{-1}(T(x) - m_\beta)\rangle) \\
&= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\frac{1}{2}\langle x - m_\alpha, A^T\Sigma_\beta^{-1}A(x - m_\alpha) \\
&= \det(2\pi\Sigma_\beta)^{-\frac{1}{2}} \exp(-\frac{1}{2}\langle x - m_\alpha, \Sigma_\alpha^{-1}(x - m_\alpha)\rangle)
\end{aligned}$$

Because

$$A^T\Sigma_\beta^{-1}A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \underbrace{\Sigma_\alpha^{-\frac{1}{2}} \Sigma_\beta^{-1} \Sigma_\alpha^{-\frac{1}{2}}}_{\left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{-1}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = \Sigma_\alpha^{-\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = \Sigma_\alpha^{-1}$$

Moreover, since T is linear, the Jacobian of this mapping can be easily calculated

$$|\det(J_T(x))| = |\det A| = \left(\frac{\det \Sigma_\beta}{\det \Sigma_\alpha} \right)^{\frac{1}{2}}$$

since we have verified the relation given in proposition 1, we have established that $T_\# \rho_\alpha = \rho_\beta$

Further calculations can yield that T achieves optimal cost when the latter is $\|x - y\|^2$, hence the 2-Wasserstein distance between two such measures is

$$\mathcal{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2$$

where $\mathcal{B}$ is the Bures metric between two positive definite matrices defined by

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta)^2 := \operatorname{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}} \right)^{\frac{1}{2}} \right)$$

which, in the particular case of diagonal matrices $\Sigma_\alpha = \operatorname{diag}(r)$ and $\Sigma_\beta = \operatorname{diag}(s)$ for $r, s \in (\mathbb{R}_*^+)^d$ simplifies to

$$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta) = \left\| \sqrt{r} - \sqrt{s} \right\|_2$$

where $\sqrt{\cdot}$ is the entrywise square root.

Chapter 2

Discrete Optimal Transport

The section presents efficient numerical methods for OT, in the discrete case (α and β discrete measures) a and b, that we simply write as vectors of size m and n respectively. In the following paragraph, the index i is for a, and j for b.

The method consists in adding an entropic term to the quantity minimized, to make the problem a λ-convex one. But, instead of using a direct gradient descent (which can anyway only be easily implemented on the whole space, which is not the case, we minimize on $U(a,b)$), Sinkhorn's algorithm finds a minimizer of the entropicly regularized OT (with an ϵ term) with iterative computations.

Definition 5. *For $P \in U(a,b)$, $H(P) := -\sum P_{i,j}(\log(P_{i,j}) - 1)$ is the discrete entropy of a coupling matrix.*

Proposition 8. $\nabla H(P) = -(\log(P_{i,j}) - 1)_{i,j}$ *and* $\partial^2 H(P) = -\operatorname{diag}(1/P_{i,j})$.
Hence, H is 1-strongly concave.

For $\epsilon > 0$, we note

$$L_C^\epsilon(a,b) := \min_{P \in U(a,b)} (\langle P, C\rangle - \epsilon H(P)) \tag{2.1}$$

which is called the regularized problem, solving this problem yields an approximate solution for the original problem. The additional entropy term favors the solution maximizing the entropy, that is, the most diffuse coupling, which better corresponds to observable reality when modeling actual phenomena like flows of commodities or people in a market in economy.

The minimized quantity is ϵ-convex, $U(a,b)$ is closed, so a minimizer always exists. Thanks to strong convexity and the convexity of $U(a,b)$, it is unique. We note P_ϵ the minimizer.

The use of the entropy as a regularizing term is an arbitrary choice, another possibility is to use the Kullback-Leibler divergence which we will be introducing a little later.

Proposition 9. *The unique solution P_ϵ of 2.1 converges to the optimum with maximal entropy when $\epsilon \to 0$ i.e*

$$P_\epsilon \xrightarrow{\epsilon \to 0} argmin_P \{-H(P), P \in U(a,b), \langle P, C\rangle = L_C(a,b)\}.$$

Hence, $L_C^\epsilon(a,b) \to L_C(a,b)$
Also $P_\epsilon \xrightarrow{\epsilon \to \infty} a \otimes b$.

Proof. Let $(\epsilon_l)_l$ be a sequence such that $\epsilon_l \xrightarrow{l \to \infty} 0$ and $\epsilon_l > 0$. Let P_l be the solution of 2.1 for $\epsilon = \epsilon_l$. $U(a,b)$ is bounded so we can extract a subsequence from $(\epsilon_l)_l$ (which we will not relabel for simplicity) such that $P_l \to P^\star$ with $P^\star \in U(a,b)$ because $U(a,b)$ is closed.

Let P be a solution to the original problem, i.e $\langle C, P \rangle = L_C(a,b)$. By suboptimality of P_l for the original problem, we have :

$$0 \leqslant \langle C, P_l \rangle - \langle C, P \rangle$$

And by suboptimality of P for the regularized problem with $\epsilon = \epsilon_l$ we have :

$$\langle C, P_l \rangle - \epsilon_l H(P_l) \leqslant \langle C, P \rangle - \epsilon_l H(P)$$

Therefore, we have

$$0 \leqslant \langle C, P_l \rangle - \langle C, P \rangle \leqslant \epsilon_l (H(P_l) - H(P))$$

H being continuous, taking the limit for $l \to \infty$ in this expression yields $\langle C, P^\star \rangle = \langle C, P \rangle$. Moreover, dividing by ϵ_l in the last inequality and taking the limit gives $H(P) \leqslant H(P^\star) \Rightarrow -H(P^\star) \leqslant -H(P)$ so $P^\star$ is actually a solution for

$$\arg\min_P \{-H(P), P \in U(a,b), \langle P, C \rangle = L_C(a,b)\}$$

By strict convexity of H the solution $P_0^\star$ of this problem is unique, so $P^\star = P_0^\star$ and the original sequence P_l is convergent.

For $\epsilon \to \infty$, a similar method shows that the problem becomes equivalent to

$$\min_{P \in U(a,b)} -H(P)$$

the solution of which is $a \otimes b$.

□

Thus, for a small regularization, the solution converges to the optimal coupling that maximizes the entropy, whereas for a large regularization, we simply get the coupling that maximizes entropy while still being admissible.

2.1 Regularized Dual

Proposition 10. *The dual problem associated to the previously introduced regularized problem*

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

is

$$L_C^\epsilon(a,b) = \max_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} \langle f, a \rangle + \langle g, b \rangle - \epsilon \langle e^{f/\epsilon}, K e^{g/\epsilon} \rangle$$

We defer the proof because it relies on that of the next proposition.

2.2 Sinkhorn's algorithm

Definition 6. *Given a cost matrix $C \in \mathbb{R}^{n\times m}$ and a positive number $\epsilon > 0$ we define the Gibbs kernel K associated to C as*

$$K_{i,j} := \exp -C_{i,j}/\epsilon$$

Proposition 11. *The solution to the regularized problem $L_C^\epsilon(a,b) := \min_{P\in U(a,b)}(\langle P, C\rangle - \epsilon H(P))$ is unique and has the form*

$$\forall 1 \leqslant i \leqslant n \qquad 1 \leqslant j \leqslant m, \qquad P_{i,j} = u_i K_{i,j} v_j$$

where K is the Gibbs kernel associated to C and the given ϵ in the problem, for two scaling variables $(u,v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Proof. We first write the Lagrangian according to the problem's constraints ($P \in U(a,b)$) by introducing the dual variables $f \in \mathbb{R}_+^n$ and $g \in \mathbb{R}_+^m$, we have

$$\Lambda(P,f,g) = \langle P, C\rangle - \epsilon H(P) - \langle P\mathbb{1}_m - a, f\rangle - \langle P^T\mathbb{1}_n - b, g\rangle$$

When evaluating at an optimum, we have

$$\forall i,j \qquad \frac{\partial \Lambda(P,f,g)}{\partial P_{ij}} = 0$$

Effectively differentiating yields

$$\forall i,j \qquad \frac{\partial \Lambda(P,f,g)}{\partial P_{ij}} = C_{i,j} + \epsilon \log(P_{i,j}) - f_i - g_j$$

Finally, solving for an optimal P we get $P_{i,j} = e^{f_i/\epsilon}e^{-C_{i,j}/\epsilon}e^{g_j/\epsilon}$ i.e $e^{f_i/\epsilon}K_{i,j}e^{g_j/\epsilon}$ which matches the enunciated form with positive vectors u and v.

□

proof of proposition 10. We continue from the proof of proposition 11 having linked the optimal solution P to the dual potentials f and g as $P_{i,j} = e^{f_i/\epsilon}e^{-C_{i,j}/\epsilon}e^{g_j/\epsilon}$. Substituting the optimal P as a function of f and g in the previously calculated Lagrangian we get a function :

$$f, g \mapsto \langle \operatorname{diag}(e^{f/\epsilon})K \operatorname{diag}(e^{g/\epsilon}), C\rangle - \epsilon H(\operatorname{diag}(e^{f/\epsilon})K \operatorname{diag}(e^{g/\epsilon}))$$

Because $P\mathbb{1}_m - a = P^T\mathbb{1}_n - b = 0$

We also have

$$\begin{aligned} -\epsilon H(\operatorname{diag}(e^{f/\epsilon})K \operatorname{diag}(e^{g/\epsilon})) &= -\epsilon H(P) = \epsilon\langle P, \log P - \mathbb{1}_{n\times m}\rangle \\ &= \langle P, \underbrace{\epsilon\,(\log P - \mathbb{1}_{n\times m})}_{:=M}\rangle \end{aligned}$$

And for $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$ we have

$$M_{i,j} = \epsilon \log\left(e^{f_i/\epsilon} e^{-C_{i,j}/\epsilon} e^{g_j/\epsilon}\right) - \epsilon = f_i + g_j - C_{i,j} - \epsilon$$

Therefore

$$\begin{aligned} -\epsilon H(\operatorname{diag}(e^{f/\epsilon}) K \operatorname{diag}(e^{g/\epsilon})) &= \langle \operatorname{diag}(e^{f/\epsilon}) K \operatorname{diag}(e^{g/\epsilon}), f\mathbb{1}_m^T + \mathbb{1}_n g^T - C - \epsilon \mathbb{1}_{n\times m} \rangle \\ &= -\langle \operatorname{diag}(e^{f/\epsilon}) K \operatorname{diag}(e^{g/\epsilon}), C \rangle \\ &\quad + \langle f, a \rangle + \langle g, b \rangle - \epsilon \langle e^{f/\epsilon}, K e^{g/\epsilon} \rangle \end{aligned}$$

Plugging this all back into the Lagrangian, we get the desired result.

□

The obtained factorization of P's entries in proposition 11 links it to the solution of dual problem, it will also allow us to define an iterative algorithm to solve the regularized problem now.

The previous result can be rewritten in matrix form as

$$P = \operatorname{diag}(u) K \operatorname{diag}(v)$$

We can now reexpress the mass conservation constraints for u and v as

$$\operatorname{diag}(u) K v = a \quad \text{and} \quad \operatorname{diag}(v) K^T u = b$$

.

Hence, the problem now is to find u and v such that

$$u \odot Kv = a \quad \text{and} \quad v \odot K^T u = b$$

where $\odot$ is the entrywise multiplication , this problem is commonly known as the "matrix scaling problem".

An intuitive way of solving it is an iterative one. The idea is to alternatively update u and v to satisfy each constraint in the last formula, Sinkhorn's algorithm is thus defined by initializing $u^{(0)} = \mathbb{1}_m$ and $v^{(0)} = \mathbb{1}_n$ and iterating :

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

where division in meant entrywise. The initialization choice made here is arbitrary, the algorithm only needs to start from positive vectors. Note that different initializations can lead to different solutions. In fact, the sought vectors are not unique since, if u and v are two such vectors then so are λu and v/λ for any positive λ. However, it turns out that this algorithm converges.

This algorithm was introduced with a proof of convergence by [Sinkhorn, 1964] it was early used to scale a matrix to make it fit desired marginals. It was quickly adopted in the field of economics to approximate solutions for optimal transport problems and has recently received renewed attention in data sciences, machine learning and imaging.

Sinkhorn's convergence analysis is simpler using Hilbert projective metric on $\mathbb{R}^n_{+,*}$ (positive vectors).

Proposition 12 (Hilbert metric)**.** *The application defined by*

$$\forall u, v \in \mathbb{R}^n_{+,*} \times \mathbb{R}^n_{+,*}, \qquad d_{\mathcal{H}}(u, v) := \log \max_{i,j} \frac{u_i v_j}{u_j v_i}$$

is a metric on the projective cone $\mathbb{R}^n_{+,*}/\sim$ *where the relationship* $\sim$ *is defined by* $u \sim v \iff \exists r > 0,\ u = rv$*. The projective cone is a complete metric space for this distance.*

Proof. It is easy to see that the application is symmetrical and that for $u, v \in \mathbb{R}^n_{+,*} \times \mathbb{R}^n_{+,*}$ such that $u = rv$ with $r > 0$ we have $d_{\mathcal{H}}(u, v) = 0$.

We notice that, if for some i, j we have $\frac{u_i v_j}{u_j v_i} < 1$ then $\frac{u_j v_i}{u_i v_j} > 1$ which ensures that the application is positive to begin with. It also implies that, if $d_{\mathcal{H}}(u, v) = 0$ then $\forall i, j\ \frac{u_i v_j}{u_j v_i} = 1 \Rightarrow \frac{u_i}{v_i} = \frac{u_j}{v_j} = r \Rightarrow u = rv$.

Finally, for $u, v, w \in \left(\mathbb{R}^n_{+,*}\right)^3$ there exist integers k and l such that

$$d_{\mathcal{H}}(u, w) = \log \frac{u_k w_l}{u_l w_k} = \log \left(\frac{u_k v_l}{u_l v_k} \times \frac{v_k w_l}{v_l w_k} \right) = \log \frac{u_k v_l}{u_l v_k} + \log \frac{v_k w_l}{v_l w_k} \leqslant d_{\mathcal{H}}(u, v) + d_{\mathcal{H}}(v, w)$$

which establishes the triangle inequality. □

The following theorem will play a critical role in proving the convergence of Sinkhorn's algorithm

Theorem 1. *Let* $M \in \mathbb{R}^{n \times m}_{+,*}$ *be a positive matrix, then for* $u, v \in \mathbb{R}^n_{+,*} \times \mathbb{R}^n_{+,*}$ *we have*

$$d_{\mathcal{H}}(Mu, Mv) \leqslant \lambda(M) d_{\mathcal{H}}(u, v) \text{ where } \begin{cases} \lambda(M) := \frac{\sqrt{\eta(M)}-1}{\sqrt{\eta(M)}+1} < 1 \\ \eta(M) := \max_{i,j,k,l} \frac{M_{i,k} M_{j,l}}{M_{j,k} M_{i,l}} \end{cases}$$

i.e M *is a strict contraction on the cone of positive vectors.*

This fundamental theorem was proved by [Birkhoff, 1957].

Theorem 2 (Convergence of Sinkhorn's algorithm)**.** *Regarding Sinkhorn's iteration, we have* $(u^{(l)}, v^{(l)}) \to (u^\star, v^\star)$ *with rates of convergence (measured through the Hilbert metric)*

$$d_{\mathcal{H}}(u^{(l)}, u^\star) = O(\lambda(K)^{2l}), \qquad d_{\mathcal{H}}(v^{(l)}, v^\star) = O(\lambda(K)^{2l})$$

And denoting $P^{(l)} := \operatorname{diag}(u^{(l)}) K \operatorname{diag}(v^{(l)})$*, we also have*

$$d_{\mathcal{H}}(u^{(l)}, u^\star) \leqslant \frac{d_{\mathcal{H}}(P^{(l)} \mathbb{1}_m, a)}{1 - \lambda(K)} \qquad d_{\mathcal{H}}(v^{(l)}, v^\star) \leqslant \frac{d_{\mathcal{H}}({P^{(l)}}^T \mathbb{1}_n, b)}{1 - \lambda(K)}$$

Proof. We notice that, for any vectors $u, v \in \mathbb{R}^n_{+,*} \times \mathbb{R}^n_{+,*}$ we have

$$d_{\mathcal{H}}(u, v) = d_{\mathcal{H}}(u/v, \mathbb{1}_n) = d_{\mathcal{H}}(\mathbb{1}_n/u, \mathbb{1}_n/v)$$

In fact, for any $w \in \mathbb{R}^n_{+,*}$ we have $d_{\mathcal{H}}(u,v) = d_{\mathcal{H}}(w \odot u, w \odot v)$. Thus, we have thanks to Theorem 1

$$\begin{aligned} d_{\mathcal{H}}(u^{(l+1)}, u^\star) &= d_{\mathcal{H}}\left(\frac{a}{Kv^{(l)}}, \frac{a}{Kv^\star}\right) \\ &= d_{\mathcal{H}}\left(Kv^{(l)}, Kv^\star\right) \\ &\leqslant \lambda(K) d_{\mathcal{H}}\left(v^{(l)}, v^\star\right) \end{aligned}$$

Similarly, we have

$$\begin{aligned} d_{\mathcal{H}}(v^{(l)}, v^\star) &= d_{\mathcal{H}}\left(\frac{b}{K^T u^{(l)}}, \frac{b}{K^T u^\star}\right) \\ &= d_{\mathcal{H}}\left(K^T u^{(l)}, K^T u^\star\right) \\ &\leqslant \lambda(K^T) d_{\mathcal{H}}\left(u^{(l)}, u^\star\right) \end{aligned}$$

and since $\lambda(K^T) = \lambda(K)$ we end up with

$$d_{\mathcal{H}}(u^{(l+1)}, u^\star) \leqslant \lambda(K)^2 d_{\mathcal{H}}\left(u^{(l)}, u^\star\right) \text{ and } d_{\mathcal{H}}(v^{(l)}, v^\star) \leqslant \lambda(K)^2 d_{\mathcal{H}}\left(v^{(l-1)}, v^\star\right)$$

By iterating these inequalities, we establish the theorem's first claim. Moreover, using the triangle inequality we have

$$\begin{aligned} d_{\mathcal{H}}(u^{(l)}, u^\star) &\leqslant d_{\mathcal{H}}\left(u^{(l+1)}, u^{(l)}\right) + d_{\mathcal{H}}\left(u^{(l+1)}, u^\star\right) \\ &\leqslant d_{\mathcal{H}}\left(\frac{a}{Kv^{(l)}}, u^{(l)}\right) + \lambda(K) d_{\mathcal{H}}\left(u^{(l)}, u^\star\right) \\ &= d_{\mathcal{H}}\left(a, u^{(l)} \odot \left(Kv^{(l)}\right)\right) + \lambda(K) d_{\mathcal{H}}\left(u^{(l)}, u^\star\right) \end{aligned}$$

and since $u^{(l)} \odot \left(Kv^{(l)}\right) = P^{(l)}\mathbb{1}_n$ rearranging the inequality gives the second part of the theorem (the second inequality is obtained similarly)

□

Although the convergence is geometrical, the multiplication of matrices is here very costly. An idea is to use the semi-dual problem and a stochastic averaged method, which will allow us to tackle the semidiscrete problem later on.

2.3 Illustration Of Sinkhorn's Method

In this section, we illustrate the previous method with examples taken from our codes.

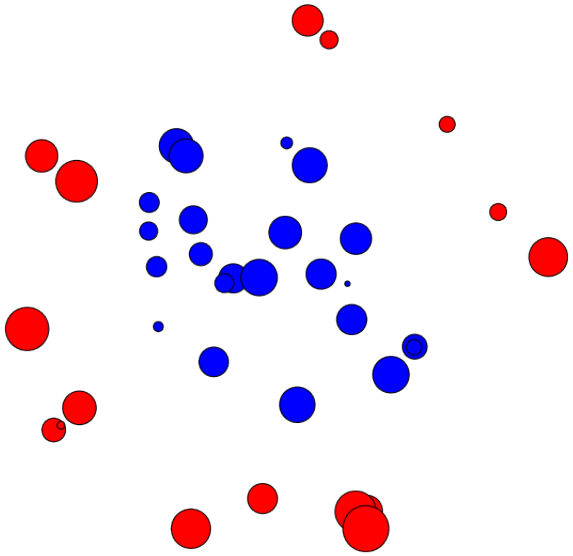

Figure 2.1: Two discrete mass distributions

The figure above shows two discrete mass distributions, let a be the blue and b the red, we want to solve the optimal transportation problem with these two measures.

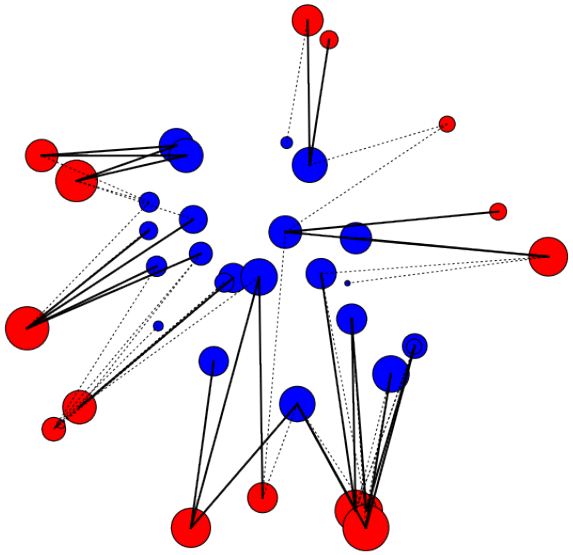

Figure 2.2: The mapping of the optimal coupling computed through Sinkhorn

Each segment is placed where the mass transported between its extremities is not negligible, dashed segments represent smaller mass movement.

Chapter 3

Semi-discrete Optimal Transport

In this, chapter, we consider the optimal transport problem between a discrete measure and an arbitrary one, especially one having a density with respect to the Lebesgue measure. This leads to interesting geometrical interpretations in small dimensions and falls within the scope of application of stochastic optimization algorithms in higher dimensions. We first need to introduce the notion of c-transform which is crucial for what follows.

3.1 c-transform and $\bar{c}$-transform, semi-dual problem

We define the indicator function for a constraint given as a set $\mathcal{C}$ as

$$\iota_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases}$$

this will be used, in constrained optimization problems, to incorporate the constraint into the function being optimized as an infinite penalization as soon as the constraint is violated.

In the first chapter, proposition 7 we introduced the dual problem for arbitrary measures α, β :

$$\mathcal{L}_C(\alpha, \beta) = \max \left\{ \int_X f d\alpha + \int_Y g d\beta : (f, g) \in \mathcal{R}(\alpha, \beta) \right\}$$

which we can reformulate thanks to the indicator function as

$$\max_{(f,g)} \mathcal{E}(f, g) \coloneqq \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y) - \iota_{\mathcal{R}(\alpha,\beta)}(f, g)$$

where $\mathcal{R}(\alpha, \beta) = \left\{ (f, g) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{Y}} \text{ measurable } : \forall (x, y), f(x) + g(y) \leqslant c(x, y) \right\}$

The idea of c-transform is to define new functions by minimizing over f and g alternatively as follows

$$\forall y \in \mathcal{Y}, \qquad f^c(y) \coloneqq \inf_{x \in \mathcal{X}} c(x, y) - f(x)$$

$$\forall x \in \mathcal{X}, \qquad g^{\overline{c}}(x) \coloneqq \inf_{y \in \mathcal{Y}} c(x,y) - g(y)$$

where $\overline{c}(y,x) \coloneqq c(x,y)$. So that

$$f^c \in \operatorname*{argmax}_g \mathcal{E}(f,g) \qquad \text{and} \qquad g^{\overline{c}} \in \operatorname*{argmax}_f \mathcal{E}(f,g)$$

i.e fixing g, $g^{\overline{c}}$ is optimal for the problem, and vice versa the other way around.

Given two function (f,g) within the constraint set, replacing them by $(f^c, g^{\overline{c}})$ improves the solution, but we have

$$f^{c\overline{c}c} = f^c \qquad \text{and} \qquad g^{\overline{c}c\overline{c}} = g^{\overline{c}}$$

which means that this improvement does not continue by repeating the procedure. However, the c-tranform allows us to reformulate the problem, once more, as follows

$$\begin{aligned}\mathcal{L}_c(\alpha,\beta) &= \max_{f \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} f^c(y) d\beta(y) \\ &= \max_{g \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} g^{\overline{c}}(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y)\end{aligned}$$

which is certainly more convenient since it is an optimization on a single potential now. It is the semidual problem.

3.2 Semi-discrete problem

In the context of semi-discrete OT, where β is discrete i.e $\beta = \sum_j b_j \delta_{y_j}$, with $b \in \Sigma^m$ it's distribution vector, in this case, the dual problem reads :

$$\mathcal{L}_c(\alpha,\beta) = \max \left\{ \int_X f d\alpha + \langle g, b \rangle : (f,g) \in \mathcal{R}(\alpha,\beta) \right\}$$

We can use the definition of $\overline{c}$-transform on g only needing to consider the support of β :

$$\forall x \in \mathcal{X}, \;\; g^{\overline{c}}(x) \coloneqq \min_{1 \leqslant j \leqslant m} c(x, y_j) - g_j$$

We have, as seen previously :

$$\mathcal{L}_c(\alpha,\beta) = \max_{g \in \mathbb{R}^m} E(g)$$

where

$$E(g) = \sum_j g_j b_j + \int_{\mathcal{X}} g^{\overline{c}}(x) d\alpha(x)$$

This is thus an optimization in finite dimension, named the semi-dual problem.

Remark that in this case, $\forall x \in \mathcal{X}, g^{\bar{c}}(x) = \min_j(c(x, y_j) - g_j)$, hence, if c is continuous, the $j \in \{1, ..., m\}$ minimizing $c(x, y_j) - g_j$ is locally constant.

This leads to the definition of the Laguerre cells :

$$\mathbb{L}_g(y_j) \coloneqq \{x \in \mathcal{X} : \forall k \neq j,\ c(x, y_j) - g_j \leqslant c(x, y_k) - g_k\}$$

which induces a disjoint decomposition of $\mathcal{X}$, enabling us to rewrite

$$E(g) = \sum_j \int_{\mathbb{L}_g(y_j)} (c(x, y_j) - g_j)\, d\alpha(x) + \langle g, b \rangle$$

Computing the gradient of this function we have

$$\forall 1 \leqslant j \leqslant m, \quad \nabla E(g)_j = b_j - \int_{\mathbb{L}_g(y_j)} d\alpha(x)$$

For an optimal g, the gradient is zero and every $x \in \mathbb{L}_g(y_j)$ is mapped into y_j agreeing with the previous remark.

The Laguerre cells (also called power diagrams) represent the regions into which the points of the support of the discrete distribution are mapped. The mass conservation constraint imposes that each region encloses as much mass for α as that of the point from β it is mapped into.

3.3 Entropic Semi-discrete Formulation

Proposition 13. *The* min *operator for vectors $z \in \mathbb{R}^n$ can be smoothely approximated by :*

$$\min_\epsilon z = -\epsilon \log \sum_i e^{-z_i/\epsilon}$$

for $\epsilon > 0$, and we have $\min_\epsilon z \xrightarrow{\epsilon \to 0} \min z$

Proof. Let $i_{min} = \arg\min_i z_i$ (pick a random index that accomplishes the minimum if not unique) and let $A = \{i \mid z_i = \min z\}$ and $k = \#A$ we suppose $k \neq n$ (the result is otherwise clear), we write :

$$\begin{aligned}
\min_\epsilon z = -\epsilon \log \sum_i e^{-z_i/\epsilon} &= -\epsilon \log \left(e^{-\min z/\epsilon} \left(k + \sum_{i \notin A} e^{(\min z - z_i)/\epsilon} \right) \right) \\
&= \min z - \epsilon \log \left(k \left(1 + \frac{1}{k} \sum_{i \notin A} e^{(\min z - z_i)/\epsilon} \right) \right) \\
&= \min z - \underbrace{\epsilon \log(k)}_{\xrightarrow{\epsilon \to 0} 0} - \underbrace{\epsilon \log \left(1 + \frac{1}{k} \sum_{i \notin A} e^{(\min z - z_i)/\epsilon} \right)}_{\sim_{\epsilon \to 0} \underbrace{\epsilon \frac{1}{k} \sum_{i \notin A} e^{(\min z - z_i)/\epsilon}}_{\xrightarrow{\epsilon \to 0} 0}}
\end{aligned}$$

Because $\min z - z_i < 0$ for $i \notin A$

□

Proposition 14. *Let f be a continuous coercive convex function of $\mathbb{R}^n = \mathcal{X}$, then* $\inf f$ *can be smoothly approximated by :*

$$\inf f = \lim_{\epsilon \to 0} -\epsilon \log \left(\int_{\mathcal{X}} e^{-f(x)/\epsilon} dx \right)$$

Proof. We give a proof for the case $n = 1$, the other cases can be treated similarly.

$$\begin{aligned}
-\epsilon \log \left(\int_{\mathbb{R}} e^{-f(x)/\epsilon} dx \right) &= -\epsilon \log \left(e^{-\inf f/\epsilon} \int_{\mathbb{R}} e^{(\inf f - f(x))/\epsilon} dx \right) \\
&= \inf f \underbrace{-\epsilon \log \left(\int_{\mathbb{R}} e^{(\inf f - f(x))/\epsilon} dx \right)}_{:= \xi^\epsilon}
\end{aligned}$$

also

$$\xi^\epsilon = \lim_{\eta \to 0} -\epsilon \log \sum_{k \in \mathbb{Z}} \eta e^{f_k^\eta/\epsilon}$$

Where we defined $f_k^\eta = \sup_{x \in [k\eta, (k+1)\eta[} \inf f - f(x)$
For some k_0 we have $f_{k_0}^\eta = 0$ therefore

$$\begin{aligned}
\xi^\epsilon &= \lim_{\eta \to 0} -\epsilon \log \left(\eta \left(1 + \sum_{k \neq k_0} e^{f_k^\eta/\epsilon} \right) \right) \\
&\sim_{\epsilon \to 0} \lim_{\eta \to 0} -\epsilon \log(\eta) - \epsilon \sum_{k \neq k_0} e^{f_k^\eta/\epsilon}
\end{aligned}$$

Remembering that, thanks to the function's properties, the sum on the right is finite and tends to 0 as ϵ tends to 0, by carefully making ϵ tend to 0 faster than η, we get that $\xi^\epsilon \to 0$ and the result follows.

□

The dual of the entropic problem between two arbitrary measures is:

$$\mathcal{L}_c^\epsilon(\alpha,\beta) := \max_{(f,g)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x)d\alpha(x) + \int_{\mathcal{Y}} g(y)d\beta(y) - \epsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{f(x)+g(y)-c(x,y)}{\epsilon}} d\alpha(x)d\beta(y)$$

As previously, we can effectuate a c-transform on one of the two variables, which can be smoothed, minimizing explicitly while fixing one of the two variables:

$$\forall y \in \mathcal{Y}, \qquad f^{c,\epsilon}(y) := -\epsilon \log\left(\int_{\mathcal{X}} e^{\frac{f(x)-c(x,y)}{\epsilon}} d\alpha(x)\right)$$

$$\forall x \in \mathcal{X}, \qquad g^{\bar{c},\epsilon}(x) := -\epsilon \log\left(\int_{\mathcal{Y}} e^{\frac{g(y)-c(x,y)}{\epsilon}} d\beta(y)\right)$$

In the semi-discrete case described previously, we have a simplified expression of g^c:

$$\forall x \in \mathcal{X}, \qquad g^{\bar{c},\epsilon}(x) := -\epsilon \log\left(\sum_j e^{\frac{g_j-c(x,y_j)}{\epsilon}} b_j\right)$$

Hence, the quantity we want to minimize is:

$$E^\epsilon(g) := \int_{\mathcal{X}} g^{\bar{c},\epsilon}(x)d\alpha(x) - \langle g, b\rangle$$

Which gradient can be computed as follows :

$$\forall 1 \leqslant j \leqslant n, \qquad \frac{\partial E^\epsilon(g)}{\partial g_j} = -\int_{\mathcal{X}} \chi_j^\epsilon(x)d\alpha(x) + b_j$$

where

$$\chi_j^\epsilon(x) = \frac{e^{\frac{g_j-c(x,y_j)}{\epsilon}}}{\sum_k e^{\frac{g_k-c(x,y_k)}{\epsilon}}}$$

is the smoothed indicator of the previously defined Laguerre cells.

Having brought the problem to the form of a function optimization, we can use stochastic algorithms (see appendix) to solve it. The Laguerre cells start out as a Voronoi tessellation with respect to the discrete measure's support and whose boundaries are progressively moved at each iteration until they match the solution of the semi-discrete problem.

3.4 Illustration

For now, let the spaces $\mathcal{X}$ and $\mathcal{Y}$ be $\mathbb{R}^2$, and the cost function the euclidean distance. β is a discrete measure (which support is in $[0,1]^2$, to simplify), and α is the probability measure which density is the sum of two Gaussian functions, centered respectively in $(0.2, 0.2)$ and $(0.8, 0.8)$ with variance $\sigma^2 = 0.1$.

The energy E we want to minimize is (in the dual problem, or the entropic dual one) in the form $E(x) \coloneqq \mathbb{E}_Y(f(x,Y))$, which corresponds to what is described in Appendix 1 . Hence, gradient descent methods can be used, which need the computation of ∇E, done in the two previous sections.

Let's take the $\varepsilon \geqslant 0$ for the following example, the problem being the smoothed one if $\varepsilon > 0$, and the semi dual one if $\varepsilon = 0$.

In either case,

$$\forall 1 \leqslant j \leqslant n, \qquad \frac{\partial E^\epsilon(g)}{\partial g_j} = -\int_{\mathcal{X}} \chi_j^\epsilon(x) d\alpha(x) + b_j$$

where

$$\chi_j^\epsilon(x) = \frac{e^{\frac{g_j - c(x,y_j)}{\epsilon}}}{\sum_k e^{\frac{g_k - c(x,y_k)}{\epsilon}}},$$

an expression still valid for $\varepsilon = 0$.

The computation of this gradient can thus be done with a method that is similar to Monte-Carlo ones.

The initialization is done with a null vector g, giving us what is called the Voronoi's cells of the points defining the support of β (let's name them (y_j). Indeed, the Laguerre cells (Voronoi cells in this particular case) map each point of $\mathbb{R}^2$ to the closest y_j This can be illustrated as follows:

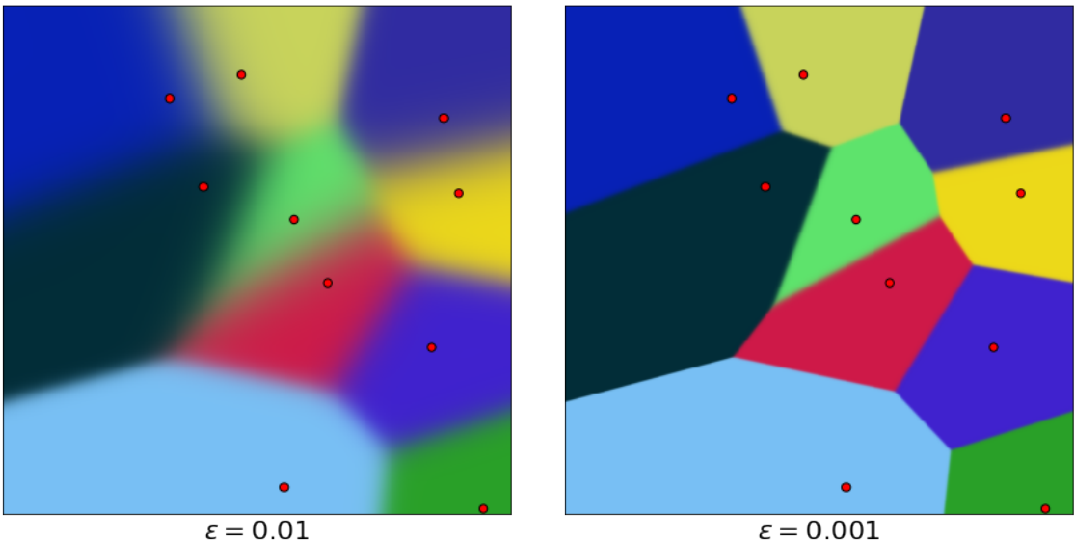

Figure 3.1: Voronoi cells

It is here clear that the entropic regularization is the smoothed semi-dual problem, becoming less and less smooth for small ε, and for $\varepsilon \longrightarrow 0$ it becomes the semi-dual problem.

With the two gaussians, instead of having Voronoi configuration, each cell is modified so as to compensate for the accumulation of mass at the center of each Gaussian.

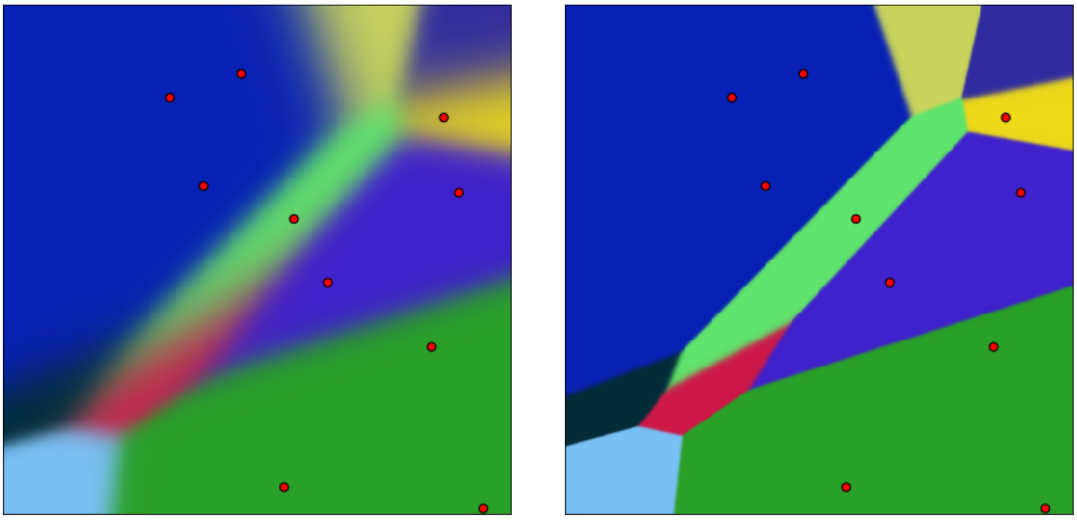

Figure 3.2: Laguerre cells, corresponding to Voronoi cells above

Note that the cells color is not changed from Figure 3.1 to 3.2 for $\epsilon = 0.01$, so that the evolution of each cell can be shown : mass tends to concentrate at the center of each Gaussian.

We give another example where β is not uniformly distributed and where α is uniform over $[0, 1]^2$ so as to show how the cells adapt by shrinking or expand to contain a mass corresponding to that of the dirac that is mapped into them.

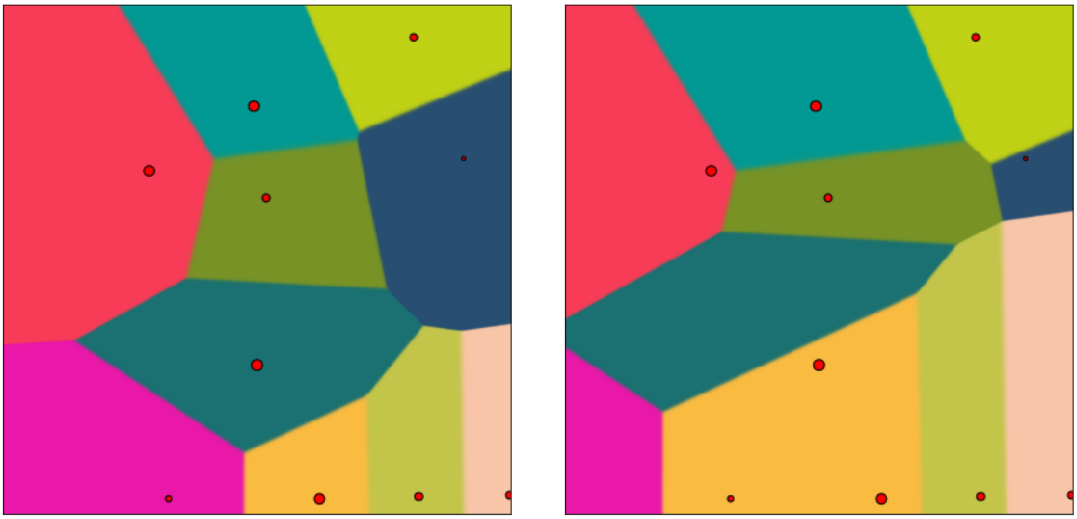

Figure 3.3: Voronoi and Laguerre cells for a non uniform β and uniform α

We can estimate the amount of mass in each cell by it's size because we have chosen α to be uniform. One can see that the cells are in different sizes and this gives a clue which dirac is being mapped into them.

Chapter 4

Stochastic Optimization for Discrete Optimal Transport

In chapter 2, we saw how the discrete optimal transport problem could be solved thanks to Sinkhorn's algorithm. Remember that each iteration of the said algorithm requires matrix vector product which can be very costly, especially in large scale problems where dimensions are high. We now introduce a stochastic approach to the same problem as presented in [Aude et al., 2016].

Definition 7. *Let P and K be two coupling matrices, we define the Kullback-Leibler divergence between them as*

$$\mathrm{KL}(P|K) \coloneqq \sum_{i,j} P_{i,j} \left(\log \left(\frac{P_{i,j}}{K_{i,j}} \right) - 1 \right)$$

The Kullback-Leibler divergence intuitively measures the difference between the two distributions, however, one should note that it is not symmetric and can not be thought of as a distance.

We give a new formulation of the regularized dual that uses the Kullback-Leibler divergence instead of the entropy

$$L_C^\epsilon(a,b) := \min_{P \in U(a,b)} \langle P, C \rangle + \epsilon \,\mathrm{KL}(P|a \otimes b) \tag{4.1}$$

This new formulation is equivalent to the previous one, the Kullback-Leibler divergence quantifies the difference between two couplings and $a \otimes b$ is the coupling that maximizes entropy, so this formulation penalizes distance from maximum entropy solutions as well and leads to the same solution as the formulation using entropy.

Using the same tricks as in the previous chapter, we define the constraint set :

$$U_c \coloneqq \{(f,g) \in \mathbb{R}^n \times \mathbb{R}^m : \ \forall\, 1 \leqslant i \leqslant n \ \ 1 \leqslant j \leqslant m, \ f_i + g_j \leqslant C_{i,j}\}$$

It follows from this that we have an indicator function ι_{U_c} for this constraint set and we can define a smoothe approximation for it

$$\iota_{U_c}^\epsilon(f,g) \coloneqq \begin{cases} \iota_{U_c}(f,g) & \text{if } \epsilon = 0 \\ \epsilon \sum_{ij} \exp \frac{f_i + g_j - C_{ij}}{\epsilon} & \text{if } \epsilon > 0 \end{cases}$$

For any $g \in \mathbb{R}^m$, we define it's c-transform and it's smoothed approximation

$$\forall 1 \leqslant 1 \leqslant n \qquad g_i^{c,\epsilon} := \begin{cases} \min_j C_{ij} - g_j & \text{if } \epsilon = 0 \\ -\epsilon \sum_j \exp \frac{g_j - C_{ij}}{\epsilon} & \text{if } \epsilon > 0 \end{cases}$$

We can now give the Dual and semi-dual formulation of the previous problem, the dual reads

$$\max_{(f,g) \in U_c} \langle f, a \rangle + \langle g, b \rangle - \iota_{U_c}^{\epsilon}(f, g)$$

and we use the c-tranform to obtain the semi-dual formulation, replacing f by $g^{c,\epsilon}$

$$\max_{g \in \mathbb{R}^m} H_\epsilon(g) := \langle g^{c,\epsilon}, a \rangle + \langle g, b \rangle - \epsilon$$

Now H_ϵ can be expressed as an expectation $H_\epsilon(g) = \mathbb{E}_I(h_\epsilon(I, g))$ where I is a random variable on $\{1, \dots, n\}$ distributed according to a and h_ϵ is defined by $h_\epsilon(i, v) := \langle g, b \rangle + g_i^{c,\epsilon} - \epsilon$, which is justified since a is a probability distribution.

Now that the problem has taken such a shape, a stochastic optimization is possible to perform gradient ascent on the semi-dual problem. Indeed, the gradient of h_ϵ can be computed as

$$\nabla_g h_\epsilon(i, g)_k = b_k - \frac{\exp \frac{g_k - C_{ik}}{\epsilon}}{\sum_j \exp \frac{g_j - C_{ij}}{\epsilon}}$$

Solving this problem yields g from which we recover a solution to the dual problem though $f = g^{c,\epsilon}$ and from there the coupling matrix P can be found as $P_{ij} = a_i \exp\left(\frac{f_i + g_j - C_{ij}}{\epsilon}\right) b_j$.

Bibliography

G. Aude, M. Cuturi, G. Peyré, and F. Bach. Stochastic Optimization for Large-scale Optimal Transport. *ArXiv e-prints*, May 2016.

Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957. ISSN 00029947. URL http://www.jstor.org/stable/1992971.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *ArXiv e-prints*, June 2016.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: 10.1002/cpa.3160440402. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160440402.

G. Peyré and M. Cuturi. Computational Optimal Transport. *ArXiv e-prints*, March 2018.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 9783319208282. URL https://books.google.fr/books?id=UOHHCgAAQBAJ.

Filippo Santambrogio. Euclidean, metric, and wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, Apr 2017. ISSN 1664-3615. doi: 10.1007/s13373-017-0101-1. URL https://doi.org/10.1007/s13373-017-0101-1.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35 (2):876–879, 06 1964. doi: 10.1214/aoms/1177703591. URL https://doi.org/10.1214/aoms/1177703591.

Cedric Villani. *Optimal Transport Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin Heidelberg, 2009.

Appendix A

Stochastic Optimization Algorithms

We consider a function $\mathcal{E}\colon \mathbb{R}^p \to \mathbb{R}$ for some $p \in \mathbb{N}$ that is expressed as

$$\mathcal{E}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

which is a very common form in various optimization contexts like machine learning for example, the functions f_i would often represent distances to a large set of samples.

To find a minimum for this function (assuming appropriate hypotheses), one can consider the usual gradient descent iteration which would read

$$w_{k+1} = w_k - \tau_k \nabla \mathcal{E}(w_k)$$

for an appropriate choice of the sequence of step sizes (τ_k), the gradient would be computed as

$$\nabla \mathcal{E}(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)$$

This algorithm is also called *Batch Gradient Descent* where the whole gradient is computed at each step, however, as n is often very large, each step is very costly. The idea of stochastic gradient descent is to alleviate this burden by only computing a single term of the previous sum at each iteration.

A.1 Stochastic Gradient Descent

Let $\mathcal{E}$ be as previously defined, the SGD iteration is defined as

$$w_{k+1} = w_k - \tau_k \nabla f_{i(k)}(w_k)$$

where $i(k)$ is uniformly drawn at random from $\{1, \dots, n\}$ at each step. The validity of this method can be justified by the fact that uniformly drawing the index $i(k)$ makes this estimation of the gradient unbiased i.e $\mathbb{E}_I \nabla f_I(x) = \nabla \mathcal{E}(x)$ for I a random variable uniformly distributed on $\{1, \dots, n\}$. Moreover, this can be seen as exploiting the redundancy of the many samples over each other (when applicable). This can be very beneficial to complexity for a large n.

The choice of the step sizes τ_k is important to ensure convergence, it must tend to 0 so as to cancel the noise induced by the random sampling while remaining non negligible to allow progress. A typical rate that satisfies both conditions is to have $\tau_l \sim_\infty l^{-1}$, so, given an initial step size τ_0 and the number of iterations l_0 serving as a "warmup" phase, one can choose :

$$\tau_l := \frac{\tau_0}{1 + l/l_0}$$

Note that, beyond improving the complexity, the stochastic approach allows us to optimize functions that are formulated as expectations, i.e of the form:

$$\mathcal{E}(x) := \mathbb{E}_Y(f(x, Y))$$

where Y is a random variable. Thanks to Monte Carlo methods, we are also enabled to optimize functions that take the form of an integral

$$\mathcal{E}(x) := \int_{\mathcal{Y}} f(x, y) d\mu(y)$$

as long as μ is a distribution we can sample from.

A.2 Stochastic Gradient Descent with Averaging

It is possible to improve SGD's convergence rate by outputting the average of the iterates, that is, consider the iteration over auxiliary variables

$$\tilde{w}_{k+1} = \tilde{w}_k - \tau_k \nabla \mathcal{E}(\tilde{w}_k)$$

and output the average

$$w_k := \frac{1}{k} \sum_{i=1}^{k} \tilde{w}_i$$

it is also possible to avoiding storing all previous iterates by computing a running average at each iteration as

$$w_{k+1} = \frac{1}{k} \tilde{w}_k + \frac{k-1}{k} w_k$$

In this method, the step size can more advantageously be chosen to have a rate of $l^{-1/2}$ like

$$\tau_l := \frac{\tau_0}{1 + \sqrt{l/l_0}}$$

A.3 Stochastic Averaged Gradient Descent

Assuming sufficient memory ressources, it is possible to improve the stochastic method further by memorizing all previously computed gradients.

The gradients are stored in $(G_i)_{i \in \{1,\dots,n\}}$ (simply 0 if not yet calculated), where n is the total size of the dataset (or number of terms defining the optimized function) which requires a memory space in $O(np)$. This allows us to have a better

approximation g of the actual gradient that is enhanced along the iteration. The algorithm reads

$$
\begin{aligned}
&h \leftarrow \nabla f_{i(k)}(w_k) \\
&g \leftarrow g - G_{i(k)} + h \\
&G_{i(k)} \leftarrow h \\
&w_{k+1} \leftarrow w_k - \tau g
\end{aligned}
$$

observe that, this time, the step size is fixed as in BGD, it must be chosen to be of the order of $1/L$ where L is the optimized function's Lipschitz constant. This algorithm improves over the two previous stochastic methods and has the same convergence rate as BGD. This improvement comes from exploiting the fact that n is finite, thus making it unviable for optimizing expectations.