Optimal Transport Notes

Ivan Zhytkevych

October 30, 2023

Contents

Notation

(TODO: Need to get rid of some of these notations from «Computational Optimal Transport» book cause they are a bit confusing. Better to use longer but simpler ones.)

- $\mathcal{X}$ and $\mathcal{Y}$ are both the sets of measures
- $[[n]] \equiv \{1, \dots, n\}$
- $\mathbb{1}_{n,m} \equiv (a_{i,j} \in \mathbb{R} : a_{i,j} = 1)_{n \times m}$
- $\mathbb{1}_n \equiv (a_i \in \mathbb{R} : a_i = 1)_n$
- $\mathbb{I}_n$ — identity matrix of size $n \times n$
- $\operatorname{diag}(u) \equiv (a_{i,j} : a_{i,j} = u \text{ for } i = j, \quad a_{i,j} = 0 \text{ for } i \neq y)_{n \times n}$
- $\Sigma_n \equiv \{x_i : x_i \in \mathbb{R}^n_+, x_i$ is a probability vector, namely $\sum_j x_{i,j} = 1\}$ — a probability simplex with n bins
- $(\mathbf{a}, \mathbf{b}) \equiv \{(a, b) \mid a \in \Sigma_n, b \in \Sigma_m\}$ *(TODO: way to change this? not obvious that $(\mathbf{a}, \mathbf{b})$ are of size $n \times m$)*
- $(\alpha, \beta) \equiv \{(\alpha, \beta) \mid \alpha \in \mathcal{X}, \beta \in \mathcal{Y}\}$
- π is a coupling measure between α and β *(TODO: better definition?)*
- $\langle \cdot, \cdot \rangle$ — for the usual Euclidean dot-product between the vectors. For two matrices of the same size: A and B — $\langle A, B \rangle \equiv \operatorname{tr}\left(A^T B\right)$ — is the Frobenius dot-product. *(TODO: need to write more about this cause I know nothing (or forgot))*
- $f \oplus g(x, y) \equiv f(x) + g(y)$ for $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$
- for two vectors $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ define $\mathbf{f} \oplus \mathbf{g} \equiv \mathbf{f}\mathbb{1}_m^T + \mathbb{1}_n \mathbf{g}^T \in \mathbb{R}^{n \times m}$
- $\alpha \otimes \beta$ is the product measure on $\mathcal{X} \times \mathcal{Y}$.
  i.e. $\int_{\mathcal{X} \times \mathcal{Y}} g(x, y) d(\alpha \otimes \beta)(x, y) = \int_{\mathcal{X} \times \mathcal{Y}} g(x, y) d\alpha(x) d\beta(y)$ *(TODO: precise definition)*
- $\mathbf{a} \otimes \mathbf{b} \equiv \mathbf{a}\mathbf{b}^T \in \mathbb{R}^{n \times m}$
- $\mathbf{u} \odot \mathbf{v} = (\mathbf{u}_i, \mathbf{v}_i) \in \mathbb{R}^n$ for $(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^n)^2$

Figure 1: So let's wake up and begin

1 Optimal Transport for Histograms

1.1 Simplex

Probability vectors gives a probability point mass in a vector form. For each of the outcomes of the random variable corresponds one row/column in the vector.

$$\mathbf{x} = \begin{pmatrix} 0.25 & 0.5 & 0.1 & 0.15 \end{pmatrix}$$

Let $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_n)$ be a probability vector with $x_0, x_1, \ldots, x_n \geq 0$ such that

$$\sum_{i=0}^{n} x_i = 1$$

So simplex should be a set of probability vectors

$$\Sigma_n := \left\{ \mathbf{a} = (a_0, a_1, \ldots, a_n) \in \mathbb{R}_+^n : \sum_{i=1}^{n} a_i = 1, a_i \geq 0, \quad \forall i \in [[n]] \right\}$$

Definition 1.1 (Discrete measure)**.** A discrete measure with weights $\mathbf{a}$ and locations $x_1, \ldots, x_n \in \mathcal{X}$ reads

$$a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \qquad \mathbf{a}_i \geq 0 \quad \forall i \in [[n]]$$

where δ_x is the Dirac delta function, which is

$$\delta_{x_i}(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

Histogram Let ξ be a random variable (with some continuous density function $f(x)$ which is unknown).

Let $X_1, X_2, \ldots, X_n \sim \xi$ a sample.

Definition 1.2 (Histogram)**.** Piecewise constant function

$$f_n(x) = \frac{\nu_r}{n \cdot |\mathcal{I}_r|} \mathbb{1}(x \in \mathcal{I}_r), \quad r \in [[m]]$$

is called a histogram, where

- $\mathcal{I}_r$ is the division segment of the division $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_m$ of the area $\mathcal{I}$ of possible values of ξ ;
- $\nu_r = \sum_{j=1}^{n} \mathbb{1}(X_j \in \mathcal{I}_r)$ — number of elements of the sample that are in $\mathcal{I}_r$.

Remark. The histogram function for large n and small enough division of the interval is the approximation of the true density $f(x)$.

Proof. By the Law of Large Numbers: $\frac{\nu_r}{n} = \frac{\sum_{j=1}^{n} \mathbb{1}(X_j \in \mathcal{I}_r)}{n} \xrightarrow[n\to\infty]{P} \mathbf{E}\left[\mathbb{1}\left(X_1 \in \mathcal{I}_r\right)\right] = P\left(X_1 \in \mathcal{I}_r\right) = \int_{\mathcal{I}_r} f(x)dx$.

We can also conclude that for some point $x_r \in \mathcal{I}_r$

$$\int_{\mathcal{I}_r} f(x)dx = |\mathcal{I}_r| \cdot f(x_r)$$

is true because of the theorem about the mean and that the function $f(x)$ is continuous.

Then pick $n \to \infty$ and the division infinitely small, which gives us:

$$\frac{\nu_r}{n \cdot |\mathcal{I}_r|} \approx f(x_r)$$

□

2 Monge Problem

The first problem that may come into a mind is about transporting some mass from point x into y. The two densities (for x and y) are f and g respectively. So we would like to find such a map T that is optimal. The problem is:

$$\min \int |x - T(x)| \, f(x)dx$$

Generalizing, we can consider other costs $c(x, y)$:

$$\min \int c(x, T(x))f(x)dx$$

But we want to work with measures μ and ν and get mass balance $\mu(\mathbb{R}^n) = \nu(\mathbb{R}^n)$.

The thing to conserve mass may be written like this:

$$\mu(T^{-1}(A)) = \nu(A) \;\; \forall A \subset \mathcal{Y}$$

And then we can rewrite the Monge formulation of OT:

$$\min \left\{ \int_{\mathbb{R}^n} c(x, T(x))d\mu(x) \mid u(T^{-1}(A)) = \nu(A) \;\; \forall A \subset \mathcal{Y} \right\}$$

Discrete measures:

$$\alpha = \sum_{i=1}^{n} = \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$$

Seek for a map that associates to each point x_i a single point y_i and which must push the mass of α toward the mass of β:

$$T : \{x_1, \dots, x_n\} \to \{y_1, \dots, y_m\}$$

$$\forall j \in [[m]], \;\; \mathbf{b}_j = \sum_{i:T(x_i)=y_j} a_i$$

compactly

$$T_{\#}\alpha = \beta$$

This map should minimize the transportation cost which is the sum of each single point transportation:

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \;:\; T_{\#}\alpha = \beta \right\}$$

2.1 Push-forward operator

(TODO:)

2.2 Kantorovich Relaxation

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

where $\mathbb{1}_n = \left(a_i = 1, i = \overline{1,n}\right)$.

Kantorovich optimal transport reads:

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle := \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}$$

For discrete measures of the form

we store matrix C as all the pairwise consts between points in the supports of α, β

$$\mathbf{C}_{i,j} \equiv c(x_i, y_j)$$

defining

$$\mathcal{L}_c(\alpha, \beta) \equiv L_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$$

That means that the this formulation of optimal transport between discrete measures is the same as the problem between their associated probability weight vectors $\mathbf{a}, \mathbf{b}$ except that the cost matrix $\mathbf{C}$ depends on the support of α, β.

2.3 Wasserstein distance

Proposition 2.0.1. Suppose that $n = m$ and that for some $p \geq 1$

$$\mathbf{C} = \mathbf{D}^p = (\mathbf{D}_{i,j}^p)_{i,j} \in \mathbb{R}^{n \times n}$$

where $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is a distance on $[[n]]$, i.e.

1. $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ is symmetric
2. $D_{i,j} = 0 \Leftrightarrow i = j$
3. $\forall (i, j, k) \in [[n]]^3,\ \mathbf{D}_{i,k} \leq \mathbf{D}_{i,j} + D_{j,k}$

Then

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

defines the p-*Wasserstein distance* on Σ_n, i.e. W_p is symmetric, positive, $W_p(\mathbf{a}, \mathbf{b}) = 0$ if and only of $\mathbf{a} = \mathbf{b}$, and it satisfies the triangle inequality.

2.4 Entropic Regularization of Optimal Transport

2.4.1 Entropic Regularization

Definition 2.1 (Disrete entropy)**.** The discrete entropy of a coupling matrix is defined as

$$\mathbf{H}(\mathbf{P}) \equiv -\sum_{i,j} \mathbf{P}_{i,j} \left(\log\left(\mathbf{P}_{i,j}\right) - 1\right)$$

As you may see from the definition, the same function works for vectors, as we use the sum.

Remark. The function $\mathbf{H}$ is 1-strongly concave, because its Hessian is $\partial^2 \mathbf{H}(P) = -\operatorname{diag}\left(\dfrac{1}{\mathbf{P}_{i,j}}\right)$ and $\mathbf{P}_{i,j} \leq 1$.

The idea behind the use of entropy is to use it as a regularizing function to obtain approximate solutions to the original transport problem:

$$L_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b}) \equiv \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$$

3 Optimal Transport for Probability Measures

3.1 Probability Measures

The applied object in OT is a measure (probability measure). Let's give some definitions and explain them

Definition 3.1 (Measure)**.** A function $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, \infty)$ is called a **measure** if:

1. $\mu(\varnothing) = 0$
2. Countable additivity: $\mu\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$

(TODO: redo this definition)

Definition 3.2 (Probability measure)**.** A function $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ is called a **probability measure** if:

1. $\mu(\varnothing) = 0$
2. Countable additivity: $\mu\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$

3.1.1 General measures

(TODO: Radon measures $\mathcal{M}(\mathcal{X})$)