Neural Optimal Transport

Petr Mokrov

Deep Generative Models course at MIPT

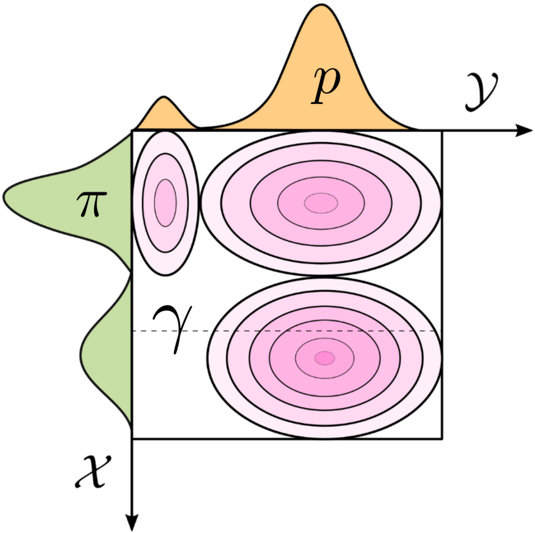
2023

1-Wasserstein distance

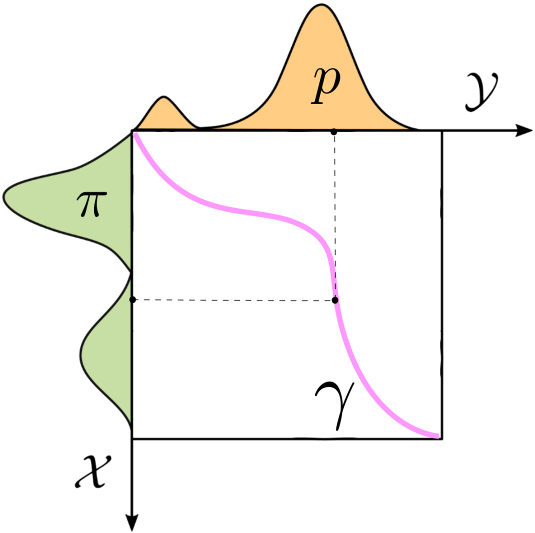
$$W_c(\pi, p) = \inf_{\gamma \in \prod(\pi,p)} \mathbb{E}_{(x,y)\sim\gamma} c(x,y) = \inf_{\gamma \in \prod(\pi,p)} \int c(x,y)\gamma(x,y)dxdy$$

- $\prod(\pi, p)$ – the set of all joint distributions $\gamma(x, y)$ with marginals π and p ($\int \gamma(x, y)dx = p(y)$, $\int \gamma(x, y)dy = \pi(x)$)
- $\gamma(x, y)$ – transportation plan (the amount of "dirt" that should be transported from point x to point y).
- $\gamma(x, y)$ – the amount, $c(x, y)$ – cost (not necessary to be metric).

1-Wasserstein distance (general case)



1-Wasserstein distance (deterministic OT plan case)



Wasserstein distance and Kantorovich Duality

Theorem 1 (Kantorovich Duality)

Let X and Y be Polish spaces, π an p are probability measures on X and Y, $c(x, y) : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ be a lower semi-continious cost function. Let

$$J(\phi, f) = \int_X \phi(x)\mathrm{d}\pi(x) + \int_Y f(y)\mathrm{d}p(y)$$

Then

$$\inf_{\gamma \in \prod(\pi, p)} \mathbb{E}_{(x,y)\sim\gamma} c(x, y) = \sup_{\phi(x)+f(y)\le c(x,y)} J(\phi, f)$$

Filippo Santambrogio: Optimal transport for applied mathematicians.

Kantorovich Duality. Alternative formulation

Definition

$c(x, y) : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ be a lower semi-continious cost function. Let $f : Y \to \mathbb{R} \cup \{-\infty\}$. Then

$$f^c(x) \stackrel{\text{def}}{=} \inf_{y \in Y} (c(x, y) - f(y))$$

is called **c-transform** of f.

One can write the Kantorovich Duality in the following form:

Theorem 1* (Kantorovich Duality)

In conditions of **Theorem 1**:

$$\inf_{\gamma \in \prod(\pi, p)} \mathbb{E}_{(x,y)\sim\gamma} c(x, y) = \sup_{f(y)} J(f^c, f) = \sup_{f(y)} \left(\mathbb{E}_{x\sim\pi} f^c(x) + \mathbb{E}_{y\sim p} f(y) \right)$$

Filippo Santambrogio: Optimal transport for applied mathematicians.

Neural Optimal Transport (NOT) Objective

We substitute $f^c(x)$ as $\inf_{y \in Y}(c(x,y) - f(y))$ and interchange [1] $\inf_{y \in Y}$ and integral sign $\int_X$ in the expression $\mathbb{E}_{x \sim \pi} f^c(x)$:

$$\mathbb{E}_{x \sim \pi} f^c(x) = \int_X f^c(x) \mathrm{d}\pi(x) = \int_X \inf_{y \in Y} \big(c(x,y) - f(y)\big) \mathrm{d}\pi(x)$$

$$= \inf_{T: X \to Y} \int_X \big(c(x, T(x)) - f(T(x))\big) \mathrm{d}\pi(x)$$

NOT objective: Kantorovich Duality as max min problem:

$$\sup_{f(y)} \inf_{T: X \to Y} \left\{ \int_X \big(c(x, T(x)) - f(T(x))\big) \mathrm{d}\pi(x) + \int_Y f(y) \mathrm{d}p(y) \right\}$$

[1] See *R.T. Rockafellar*, Integral functionals, normal integrands and measurable selections.

NOT: How to recover optimal transport plan?

Proposition: Optimal maps solve the $\max\min$ **problem**

Let there exist a **deterministic** OT plan γ^*, given by:

$$\mathrm{d}\gamma^*(x,y) = \mathrm{d}\pi(x)\delta[y = \tilde{T}^*(x)].$$

Then for every maximizer f^* of **NOT objective** it holds:

$$\tilde{T}^* \in \operatorname*{arg\,min}_{T:X\to Y} \left\{ \int_X \big(c(x, T(x)) - f^*(T(x))\big)\mathrm{d}\pi(x) + \int_Y f^*(y)\mathrm{d}p(y) \right\},$$

i.e. $(f^*, \tilde{T}^*)$ is the *saddle* point of **NOT objective**.

Korotin et. al.: Neural Optimal Transport https://arxiv.org/pdf/2201.12220.pdf

NOT: How to recover optimal transport plan?

Question 1

Does every saddle point (f^*, T^*) of **NOT objective** recover OT plan?

Answer: **No**, even if there exist a **deterministic** OT plan. The maps T^*, which solve **NOT objective**, but don't recover OT plan, are called *fake* solutions.

Question 2

How to guarantee, that every saddle point (f^*, T^*) of **NOT objective** recover OT plan?

Answer: We need **Weak optimal transport**.

Korotin et. al.: Neural Optimal Transport https://arxiv.org/pdf/2201.12220.pdf

Weak optimal transport

$$\mathrm{Cost}_C(\pi, p) = \inf_{\gamma \in \prod(\pi, p)} \int_X C(x, \pi(\cdot|x)) \underbrace{\mathrm{d}\gamma_x(x)}_{\mathrm{d}\pi(x)}$$

- $\prod(\pi, p)$ – the set of all joint distributions $\gamma(x, y)$ with marginals π and p ($\int \gamma(x, y)dx = p(y)$, $\int \gamma(x, y)dy = \pi(x)$)
- $C : X \times \mathcal{P}(Y) \to \mathbb{R}$ – **weak** transportation cost.
- $\gamma_x \in \mathcal{P}(X)$ – projection of γ_x to X (first marginal distribution), i.e. $\gamma_x = \pi$.

Question. Why does weak OT generalize strong OT, i.e., $W_c(\pi, p)$?

Examples of weak OT costs

ε-**Entropic OT,** $\varepsilon > 0$**:**

$$C(x, \gamma(\cdot|x)) = \int_Y c(x, y)\mathrm{d}\gamma(y|x) - \varepsilon H(\gamma(\cdot|x))$$

ε-**weak quadratic cost:**

$$C(x, \gamma(\cdot|x)) = \int_Y \frac{1}{2}\|x - y\|_2^2 \mathrm{d}\gamma(y|x) - \frac{\varepsilon}{2}\mathrm{Var}(\gamma(\cdot|x))$$

ε-**weak kernel cost**[2]**:**

$$C(x, \gamma(\cdot|x)) = \int_Y \frac{1}{2}\|x - y\|_2 \mathrm{d}\gamma(y|x) - \frac{\varepsilon}{4}\int_Y \int_Y \|y - y'\|_2 \mathrm{d}\gamma(y|x)\mathrm{d}\gamma(y'|x)$$

[2] *Korotin et. al.*: Kernel Neural Optimal Transport
https://arxiv.org/pdf/2205.15269.pdf

Weak NOT objective

The **NOT objective** is straightforwardly generalized for weak case:

$$\sup_{f} \inf_{\gamma \in \prod(\mathbb{P})} \Bigg\{ \int_X C\big(x, \gamma(\cdot|x)\big) \mathrm{d}\pi(x) \\ - \int_X \bigg[\int_Y f(y)\, \mathrm{d}\gamma\,(y|x) \bigg] \mathrm{d}\pi(x) + \int_Y f(y) \mathrm{d}p(y) \Bigg\}$$

Stochastic map $T : X \times Z \to Y, Z \sim \mathbb{S}$ reformulation:

$$\sup_{f} \inf_{T} \Bigg\{ \int_X C\big(x, T_x \sharp \mathbb{S}\big) \mathrm{d}\pi(x) \\ - \int_X \int_Z f(T_x(z)) \mathrm{d}\mathbb{S}(z) \mathrm{d}\pi(x) + \int_Y f(y) \mathrm{d}p(y) \Bigg\}$$