Article analysis: Neural Optimal Transport

Paul Barbier and Bastien Le Chenadec

March 28, 2024

Overview

Introduction

Introduction

Well known models for generating images:

- GANs
- VAEs
- Normalizing Flows
- Diffusion models

$\rightarrow$ Explicit **Optimal Transport** (OT) is a recent approach to generative models.

Optimal Transport

Optimal Transport Problem

- μ, ν probability distributions on $\mathcal{X}$, $\mathcal{Y}$.
- Cost function: $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
- Problem: finding a **transport map** $T^* : \mathcal{X} \to \mathcal{Y}$ such that:

$$T^* \in \underset{T\#\mu=\nu}{\text{Argmin}} \int_{\mathcal{X}} c(x, T(x))d\mu(x) \tag{1}$$

- Optimal transport cost:

$$\text{Cost}(\mu, \nu) = \int_{\mathcal{X}} c(x, T^*(x))d\mu(x) \tag{2}$$

$\rightarrow$ **Deterministic** mapping T^*.

Kantorovich formulation: a relaxed OT problem

- Let $\Pi(\mu, \nu)$ the set of distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.
- Problem: find a **transport plan** $\pi^* \in \Pi(\mu, \nu)$ such that:

$$\pi^* \in \underset{\pi \in \Pi(\mu,\nu)}{\text{Argmin}} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \tag{3}$$

- Optimal transport cost:

$$\text{Cost}(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi^*(x, y) \tag{4}$$

$\rightarrow$ **Stochastic** mapping π^*.

Weak OT

- Generalization to **weak** costs :

$$C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$$

- Problem: find a **transport plan** $\pi^* \in \Pi(\mu, \nu)$ such that:

$$\pi^* \in \underset{\pi \in \Pi(\mu,\nu)}{\mathrm{Argmin}} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\pi(x) \tag{5}$$

- Optimal transport cost:

$$\mathrm{Cost}(\mu, \nu) = \int_{\mathcal{X}} C(x, \pi^*(\cdot|x)) d\pi^*(x) \tag{6}$$

$\rightarrow$ **Stochastic** mapping π^*.

Weak OT Duality

For weak OT cost C and a sufficiently regular function f,

$$f^C(x) = \inf_{\rho \in \mathcal{P}(\mathcal{Y})} \left\{ C(x, \rho) - \int_{\mathcal{Y}} f(y) d\rho(y) \right\} \tag{7}$$

Then, the dual form of the weak OT problem writes:

$$f^* \in \operatorname*{Argmax}_f \int_{\mathcal{X}} f^C(x) d\mu(x) + \int_{\mathcal{Y}} f(y) d\nu(y) \tag{8}$$

and the transport cost is equal to:

$$\mathrm{Cost}(\mu, \nu) = \int_{\mathcal{X}} f^{*C}(x) d\mu(x) + \int_{\mathcal{Y}} f^*(y) d\nu(y) \tag{9}$$

NOT

Weak dual OT reformulation

- New probability space $\mathcal{Z}$.
- Let $\rho \in \mathcal{P}(\mathcal{Z})$ with some basic assumptions.

Then, we have:

$$f^C(x) = \inf_{t:\mathcal{Z}\to\mathcal{Y}} \left\{ C(x, t\#\rho) - \int_{\mathcal{Z}} f(t(z))d\rho(z) \right\} \tag{10}$$

which leads to the maximin problem:

$$\text{Cost}(\mu, \nu) = \sup_f \inf_T \mathcal{L}(f, T) \tag{11}$$

With $\mathcal{L}(f, T) = \int_{\mathcal{Y}} f d\nu + \int_{\mathcal{X}} \left(C(x, T(x, \cdot)\#\rho) - \int_{\mathcal{Z}} f(T(x, z))d\rho(z) \right) d\mu(x)$

The trick: noise outsourcing

- Here, they introduce $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$.
- Trick known as **noise outsourcing**:

Theorem

If X and Y are random variables in suitable spaces $\mathcal{X}$ and $\mathcal{Y}$, then there exists $\eta \sim \mathcal{U}([0,1])$ with $\eta \perp\!\!\!\perp X$ and a function $h\,[0,1] \times \mathcal{X} \longrightarrow \mathcal{Y}$ such that $(X, Y) = (X, h(\eta, X))$ almost surely.

Weak OT reformulation: A summary

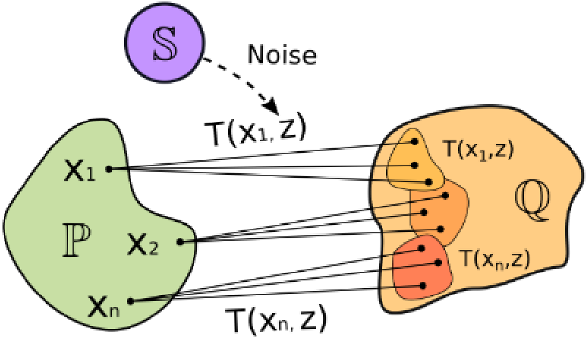

Figure: Transportation map with outsourced noise z (taken from [KSB22])

Stochastic Gradient Ascent Descent (SGAD)

Algorithm Stochastic Gradient Ascent Descent algorithm

1: **Input**: distributions μ, ν, ρ accessible by samples, mapping network $T_\theta : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^d$, potential network $f_\omega : \mathbb{R}^n \to \mathbb{R}$, number of inner iterations K_T, (weak) cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$, empirical estimator $\widehat{C}(x, T(x, Z))$ for the cost

2: **Output**: learned stochastic OT map T_θ representing an OT plan between distributions μ, ν

3: **repeat**

4: Sample batches $Y \sim \nu$, $X \sim \mu$, for each $x \in X$ sample batch $Z_x \sim \rho$

5: $\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega\big(T_\theta(x, z)\big) - \frac{1}{|Y|} \sum_{y \in Y} f_\omega(y)$

6: Update ω by using $\frac{\partial \mathcal{L}_f}{\partial \theta}$

7: **for** $k_T = 1, 2, \ldots, K_T$ **do**

8: Sample batch $X \sim \mu$, for each $x \in X$ sample batch $Z_x \sim \rho$

9: $\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \Big[\widehat{C}\big(x, T_\theta(x, Z_x)\big) - \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega\big(T_\theta(x, z)\big)\Big]$

10: Update θ by using $\frac{\partial \mathcal{L}_\theta}{\partial \theta}$

11: **until** converged

Experiments

Synthetic data

- 2D standard normal distribution for $\mathcal{X}$.
- Moon distribution for $\mathcal{Y}$.
- Weak OT cost for $\gamma = 1$:

$$C(x, \mu) = \int_{\mathcal{Y}} \frac{1}{2}\|x - y\|^2 d\mu(y) - \frac{1}{2}\mathrm{Var}(\mu) = \frac{1}{2}\|x - \int_{\mathcal{Y}} y \, d\mu(y)\|^2 \tag{12}$$

The parameters:

- simple feedforward model.
  - 2 hidden layers (100 units each).
  - ReLU activation.
- Adam optimizer.

Results: the fitted distribution

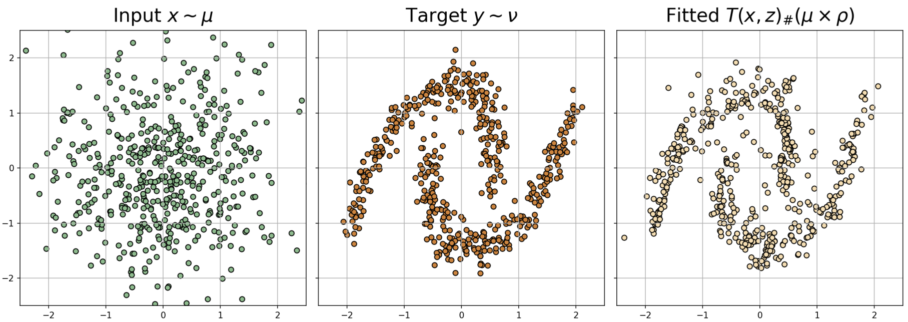

Figure: Synthetic data experiment. Left: input gaussian distribution. Middle: target distribution. Right: learned transport of the input distribution to the target distribution.

Results: a closer look at the map

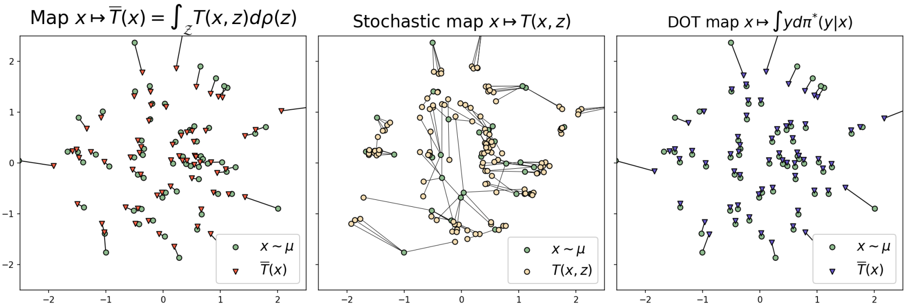

Figure: Synthetic data experiment. Left: average of the learned transport for different points. Middle: learned transport for a batch of points. Right: optimal transport plan obtained with the POT library.

Large-scale dataset

- CelebA dataset with 200k celebrity face images.
- CartoonSet with 100k avatar 2D generated images.

Figure: Images samples from CartoonSet100K

Large-scale experiment: the parameters

- Resnet to parametrise f.
- U-Net to parametrise the map T.
- image size: 128×128.
- Batch size: 128.
- Adam optimizer.
- $\mathrm{lr} = 1 \times 10^{-4}$.
- $k_T = 10$.

Results: it works!

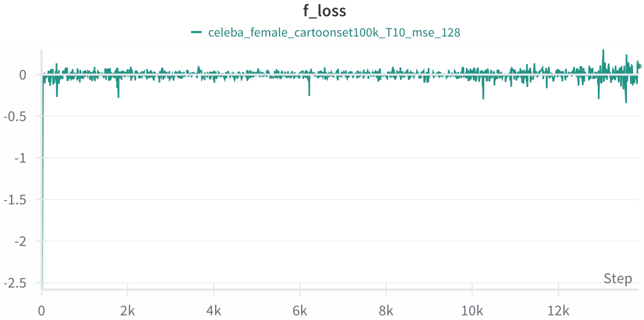

Figure: $\mathcal{L}_f$ during training

Results: some generated avatars

Figure: Iteration 1

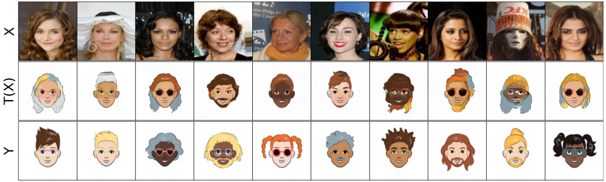

Figure: Iteration 5000

Results: an issue

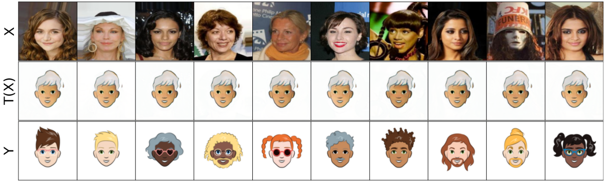

Figure: Iteration 13800

- Mode collapse!

Conclusion

Conclusion

- An algorithm to learn optimal transport maps.
- New approach to incorporate OT in ML.
- With good results on large-scale datasets!
- Good perspectives for generative models.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev, *Neural optimal transport*, arXiv (Cornell University) (2022).