Convexity Regularizer for Neural Optimal Transport

Nikita Gushchin (TA)

Kofffivi F. Gbagbe

Bintang A.S.W.A.M

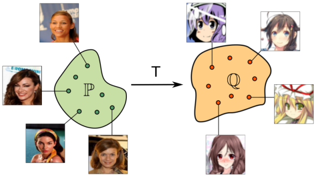
MSc Advanced Computational Science

March 24, 2023

Skoltech
Skolkovo Institute of Science and Technology

The research idea:

Let's consider the following Unpaired Image-to-Image Style Translation:



where we have:

$$\begin{cases} \textbf{Inputs: } \text{two unpaired dataset sample empirically from some p.d.f } \mathbb{P} \text{ and } \mathbb{Q} \\ \textbf{Output: } \text{a Style translation map } T : \mathcal{X} \longrightarrow \mathcal{Y} \text{ such that } T_{\#}\mathbb{P} = \mathbb{Q} \end{cases}$$

Skoltech
Skolkovo Institute of Science and Technology

The research idea (cont.):

- The method proposed in the research paper uses adversarial training (kind of GANs), which is not very stable to compute the optimal transport map T.
- From the theory , the method's optimal "discriminator" must be convex, and its gradient can be used for inverse mapping from the target distribution to the source distribution.
- Hence our goal is to add a convexity regularizer in the loss of neural optimal transport algorithm to improve its stability during transport training and the high-quality of the inverse target-to-input mapping.

General Notations

- $\mathcal{X}$ and $\mathcal{Y}$ are some polish spaces
- $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are sets of probability defined respectively on $\mathcal{X}$ and $\mathcal{Y}$
- For a measurable map $T : \mathcal{X} \longrightarrow \mathcal{Y}$ the operator $T_{\#}$ denotes the so called push-forward operator

Mathematical Background

- **Monge's Optimal Transport formulation**
  For $\mathbb{P} \in \mathcal{P}(\mathcal{X}), \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$, Monge's primal formulation of the cost is

$$\textbf{Cost}(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{T_{\#}\mathbb{P}=\mathbb{Q}} \int_{\mathcal{X}} c(x, T(x)) d\mathbb{P}(x) \tag{1}$$

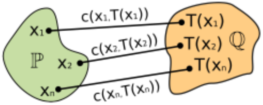


Figure: Monge's OT formulation.

Here the minimum is taken over the set of mesurables functions called transport maps $T : \mathcal{X} \longrightarrow \mathcal{Y}$ that maps the marginal distribution $\mathbb{P}$ to $\mathbb{Q}$. The optimal transport map T^* is called the OT map

Mathematical Background (cont.)

Note: The equation (1) is not symmetric and for some $\mathbb{P}, \mathbb{Q}$ it may happen that their is no T such that $T_{\#}\mathbb{P} = \mathbb{Q}$ (does not allow mass splitting)

Skoltech
Skolkovo Institute of Science and Technology

Mathematical Background (cont.)

- **Strong OT formulation**
  To overcome the mass splitting problem (Kantorovich, 1958) proposed the following reformulation

$$\textbf{Cost}(\mathbb{P},\mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}\times\mathcal{Y}} c(x, T(x))d\pi(x,y) \tag{2}$$

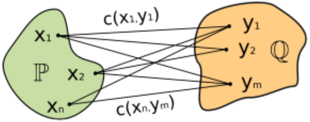


Figure: Strong OT formulation

and the minimum is taken over all the transport plan π which represent all distributions on $\mathcal{X}\times\mathcal{Y}$ whose marginals are $\mathbb{P}$ and $\mathbb{Q}$ and the optimal plan denoted by π^*

The two main important results of the research paper

- **Corollary:**

$$\textbf{Cost}(\mathbb{P}, \mathbb{Q}) = \underset{f}{\text{Sup}}\underset{T}{\text{Inf}}\mathcal{L}(f, T),$$

where $\mathcal{L}(f, T) = \int_{\mathcal{Y}} f(y)d\mathbb{Q}(y) + \int_{\mathcal{X}} \left(c(x, T(x, \dot{)}_{\#}\mathbb{S}) - \int_{\mathcal{Z}} f(T(x, z))d\mathbb{S} \right) \mathbb{P}(x)d$

- **Lemma:**
For any f^* that maximizes the dual form of the weak OT cost between $\mathbb{P}$ and $\mathbb{Q}$ and for any stochastic map T^* which realizes some optimal transport plan π^*, it holds that

$$T^* \in \underset{T}{\text{argInf}}\mathcal{L}(f^*, T).$$

Skoltech
Skolkovo Institute of Science and Technology

Algorithm to Learn the Optimal Transport Plans

Algorithm 1: Neural optimal transport (NOT)

Input :distributions $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ accessible by samples; mapping network $T_\theta : \mathbb{R}^P \times \mathbb{R}^S \to \mathbb{R}^Q$; potential network $f_\omega : \mathbb{R}^Q \to \mathbb{R}$; number of inner iterations K_T; (weak) cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$; empirical estimator $\widehat{C}\big(x, T(x, Z)\big)$ for the cost;

Output :learned stochastic OT map T_θ representing an OT plan between distributions $\mathbb{P}, \mathbb{Q}$;

repeat

  Sample batches $Y \sim \mathbb{Q}$, $X \sim \mathbb{P}$; for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;

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

  Update ω by using $\frac{\partial \mathcal{L}_f}{\partial \omega}$;

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

    Sample batch $X \sim \mathbb{P}$; for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;

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

    Update θ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$;

until *not converged*;

Skoltech
Skolkovo Institute of Science and Technology

Convexity Regularization (theoretical)

- From the theory we have the potential $f(x) = f^*(x)\frac{|X|}{2}$
- By posing $f(x) = \frac{||x||^2}{2} - \psi(x)$ and equating with the above formula we obtain

$$\boxed{\psi(x) = -f^*\frac{|X|}{2} + \frac{||x||^2}{2}}$$

- We desire ψ to be a convex function which satisfies the Jensen's inequality

$$\psi(tx_1 + (1-t)x_2) - t\psi(x_1) - (1-t)\psi(x_2) \leq 0, \;\; t \in [0,1] \tag{3}$$

- Then, we define and penalize the loss function to hold Jensen's inequality

$$\mathcal{L}(X_1, X_2) = \psi(tX_1 + (1-t)X_2) - t\psi(X_1) - (1-t)\psi(X_2) \leq 0 \tag{4}$$

Convexity Regularization (cont.)

- To make loss positive we may apply the **ReLU** function; hence let define

$$\mathrm{R}(X_1, X_2) = \textbf{ReLU}(\mathcal{L}(X_1, X_2))$$

- Finally we uses the mean of $\mathrm{R}(X_1, X_2)$ as a regularization term which can be multiplied by a given learning rate, in this case we are using two learning rate, 0.01 and 0.001

Sample of our results

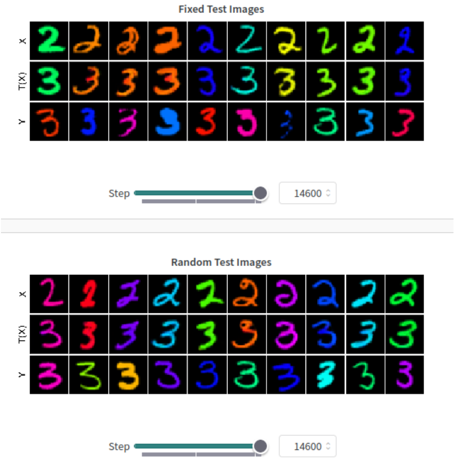


Figure: Regularized Model in Training with a learning rate of 0.01

Sample of our results(cont.)

Figure: Regularized Model in Training with a learning rate of 0.001