Convexity Regularizer for Neural Optimal Transport

Koffivi Gbagbe [1] **Bintang A.S.W.A.M** [1] **Nikita Gushchin (TA)** [1]

Abstract

We proposed and add a convexity regularizer in the loss of the neural optimal transport algorithm proposed in Korotin et al. (2023) to compute the optimal transport plan for strong transport costs and investigate on the stability of training as well as the quality of the inverse mapping. We evaluate the performance of our proposed algorithm on a unpaired image-to-image translation problem using a colored images of MNIST digits. Theoretical and practical implications of the results are discussed

Keywords: convexity regularizer, OT map, inverse target-to-input mapping, adversarial training, convex optimal transport.

Github repo: bin-koff

1. Introduction

Using neural networks to solve continuous optimal transport problem is a promising approach especially for unpaired style-transfer problem (see figure 1). The idea behind this method is to learn a one-to-one mapping (OT map) between the source and target data distributions. The proposed method by Korotin et al. (2023) uses adversarial training similar to Generative Adversarial Networks (GANs), which is not very stable. However, unlike GANs, this method's optimal "discriminator" must be convex, and its gradient can be used for inverse mapping from the target distribution to the source distribution. To address the issue of stability, we find necessary to insert a convexity regularizer (kind of gradient penalty in WGAN-GP) in the loss of the neural optimal transport algorithm to improve its stability during the training phase while improving the high-quality of the inverse target-to-input mapping.

General notations:

Machine Learning 2023 Course [1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Koffivi Gbagbe <koffivi.gbagbe@skoltech.ru>, Bintang A.S.W.A.M <BintangAlamSemesta.WisranAm@skoltech.ru>, Nikita Gushchin (TA) <Nikita.Gushchin@skoltech.ru>.

Final Projects of the Machine Learning 2023 Course, Skoltech, Moscow, Russian Federation, 2023.

$\mathcal{X}$ 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.

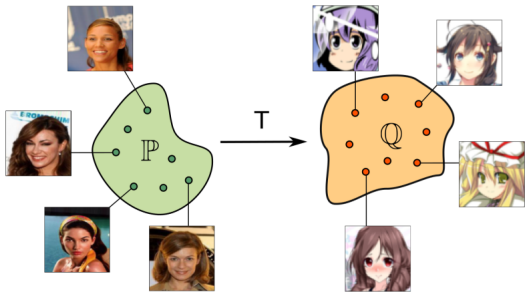

Figure 1. Example of unpaired image-to-image style translation. Inputs: two unpaired dataset sample empirically from some probability distribution function $\mathbb{P}$ and $\mathbb{Q}$; Output: a Style translation map $T : \mathcal{X} \longrightarrow \mathcal{Y}$ such that $T_{\#}\mathbb{P} = \mathbb{Q}$.

2. Preliminaries

2.1. 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

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

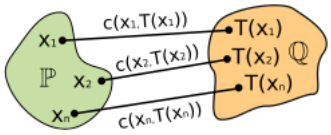

Figure 2. Monge's OT formulation.

The minimum is taken over the set of measurables 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. It is important to note that 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) therefore Kantorovitch (1958) proposed the following formulation

2.2. Strong optimal transport formulation

To overcome the mass splitting problem Kantorovitch (1958) proposed the following reformulation

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

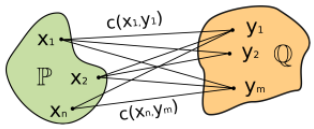

Figure 3. Strong OT formulation

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^*. Lastly, an example of strong OT cost for $\mathcal{X} = \mathcal{Y} = \mathcal{R}^D$, is the ($p$-th power of) Wasserstein-p distance $\mathcal{W}_p$, i.e. formulation (2) in which $c(x, y) = \frac{\|x-y\|^p}{2}$, $(p = 2)$

3. Related work

We overview the latest research on both OT maps and OT costs. For OT maps / OT plans, this topic is still evolving but challenging. Because there is only a few number of scalable methods have progressively been developed in response to it. According to Korotin et al. (2023), to compute the OT map, it needs two well-formulated approaches, the primal and dual formulation. Moreover, the advantage and drawback of these approach are represented in table 1

On the other hand, OT costs are basically treated as the loss function to learn generative models, the most notable example of OT costs in large-scale machine learning is Wasserstein GANs (Arjovsky et al. (2017); Gulrajani et al. (2017)).

4. Algorithm for learning the OT plan

4.1. Neural optimal transport (NOT)

Korotin et al. (2023) proposed a deep neural networks based novel algorithm to compute deterministic and stochastic OT plans which is designed for both weak and strong optimal transport costs.

Primal formulation:	Dual formulation:
• **advantage**: Using generative models & yields complex optimization objectives with several adversarial regularizers. • **drawback**: In practice, it is quite tough to setup since they need careful selection of hyper-parameters.	• **advantage**: methods based on dual formulation have simpler optimization procedures, usually are designed for OT with quadratic cost,i.e. Wasserstein-2 distance $\mathcal{W}_2^2$ • **drawback**: generates in a version of unbalanced optimal transport that leads to sparse solutions, which either hinders the interpretability for machine learning tasks or they do not perform well in practice.

Table 1. advantage and drawback of primal and dual formulation

Algorithm 1 Neural optimal transport (NOT)

Input: distributions $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ accessible by samples mapping network T_θ; potential network $f_w : \mathbb{R}^Q \mapsto \mathbb{R}$; number of inner iterations K_T; (strong) cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \mapsto \mathbb{R}$; empirical estimator $\hat{C}(x, T(x, Z))$ for the cost
Output: learned stochastic OT map T_θ representing an OT plan between distributions $\mathbb{P}, \mathbb{Q}$
repeat
  Samples batches $Y \sim \mathbb{Q}, X \sim \mathbb{P}$; for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;
  $\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{|Z_x|} \sum_{z \in Z_x} f_w(T_\theta(x, z)) - \frac{1}{|Y|} \sum_{y \in Y} f_w(y)$;
  Update w by using $\frac{\partial \mathcal{L}_f}{\partial w}$
  for $k_T = 1$ **to** K_T **do**
    Sample $X \sim \mathbb{P}$, for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;
    $\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \Big[\hat{C}(x, T_\theta(x, Z_x)) - \frac{1}{|Z_x|} \sum_{z \in Z_x} f_w(T_\theta(x, z))$
    Update θ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$
  end for
until *converged*

For more details about Algorithm 1 see Korotin et al. (2023)

4.2. Regularized Neural optimal transport (Reg-NOT)

In order to add a regularization term to the loss of the previous algorithm we used the following steps:

- From the dual optimization problem we deduce that the potential

$$f(x) = f^*(x)\frac{|X|}{2} \tag{3}$$

- By posing $f(x) = \frac{||x||^2}{2} - \psi(x)$ and equating with equation (3) we obtain

$$\psi(x) = -f^*\frac{|X|}{2} + \frac{\|x\|^2}{2} \tag{4}$$

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

$$\psi(tx_1+(1-t)x_2)-t\psi(x_1)-(1-t)\psi(x_2) \leq 0,\ t \in [0,1] \tag{5}$$

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

$$\begin{aligned}\mathcal{L}(X_1, X_2) = \psi(tX_1 + (1-t)X_2)\\ -t\psi(X_1) - (1-t)\psi(X_2) \leq 0\end{aligned} \tag{6}$$

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

$$\mathrm{R}(X_1, X_2) = \mathbf{ReLU}(\mathcal{L}(X_1, X_2)) \tag{7}$$

- 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.

Following these steps we modified the previous algorithm as follow:

Algorithm 2 Regularized Neural optimal transport (Reg-NOT)

Input: distributions $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ accessible by samples mapping network T_θ; potential network $f_w : \mathbb{R}^Q \mapsto \mathbb{R}$; number of inner iterations K_T; (strong) cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \mapsto \mathbb{R}$; empirical estimator $\hat{C}(x, T(x, Z))$ for the cost
Output: learned stochastic OT map T_θ representing an OT plan between distributions $\mathbb{P}, \mathbb{Q}$
repeat
  Samples batches $Y \sim \mathbb{Q}, X \sim \mathbb{P}$; for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;
  Set $Y^{'} = Y$ and divide $Y^{'}$ into two samples Y_1 and Y_2
  Set $R(y_1, y_2) = ReLU(\psi(ty_1+(1-t)y_2)-t\psi(y_1)-(1-t)\psi(y_2)); t \in [0,1]$
  $\mathcal{L}_f \leftarrow \frac{1}{|X|}\sum_{x\in X}\frac{1}{|Z_x|}\sum_{z\in Z_x} f_w(T_\theta(x,z)) - \frac{1}{|Y|}\sum_{y\in Y} f_w(y) + \alpha\frac{1}{|Y_1|}\sum_{(y_1,y_2)\in Y_1\times Y_2} R(y_1, y_2)$;
  Update w by using $\frac{\partial \mathcal{L}_f}{\partial w}$
  for $k_T = 1$ **to** K_T **do**
    Sample $X \sim \mathbb{P}$, for each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;
    $\mathcal{L}_f \leftarrow \frac{1}{|X|}\sum_{x\in X}\Big[\hat{C}(x, T_\theta(x, Z_x)) - \frac{1}{|Z_x|}\sum_{z\in Z_x} f_w(T_\theta(x,z))$
    Update θ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$
  end for
until *converged*

Here α is the learning rate mutiplying the regularization term.

5. Experiments and Results

In this section, we will represent our experiment on toy dataset, i.e. color-MNIST, here X-colored images of MNIST digit 2, Y-colored images of MNIST digit 3. Aside from that, style transfer for such data is to map digit 2 to the digit 3 of the same color. We are also eventually succeed to running three cases, the first case is purely without regularization,whereas the second case is using regularization with $lr = 0.01$. On the flip side, it requires to introduce loss function (denoted as f-loss) and Frechet Inception Distance(FID). Loss function is basically Loss functions measure how far an estimated value is from its true value, in whih it maps decisions to their associated costs. Whereas, FID is a metric for evaluating the quality of generated images and specifically developed to evaluate the performance of generative adversarial networks. A lower FID indicates higher-quality images; conversely, a higher FID correlates lower-quality image and the relationship might be linear.

5.1. Source-to-Target Mapping of Neural Optimal Transport (NOT)

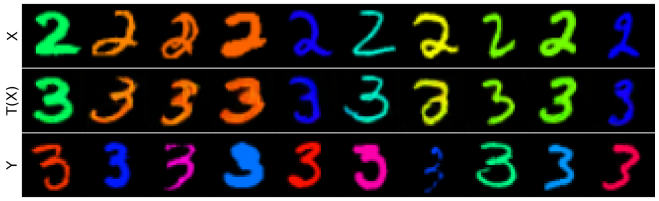

Figure 4. Fixed Test Images without Regularization

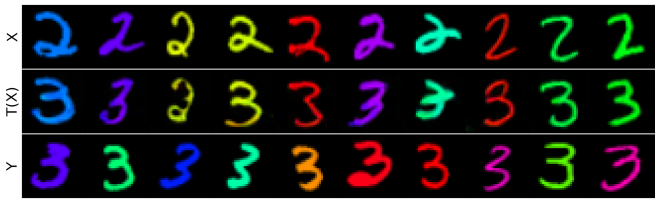

Figure 5. Random Test Images without Regularization

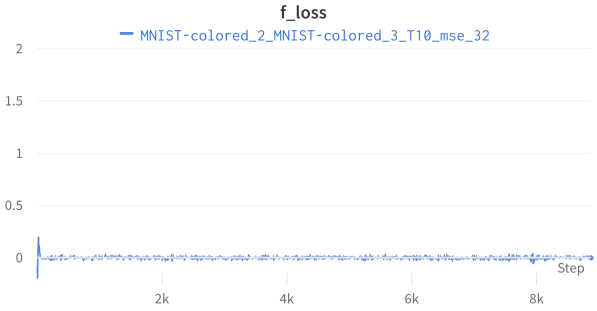

Figure 6. Loss-function without regularization calculated until 8800 iterations

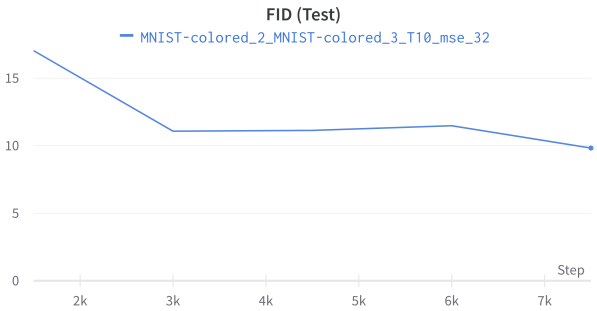

Figure 7. FID score without regularization

5.2. Source-to-Target Mapping of Regularized Neural Optimal Transport (Reg-NOT) with a learning rate = 0.01

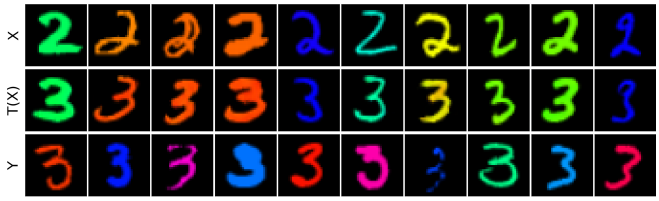

Figure 8. Fixed Test Images using Regularization

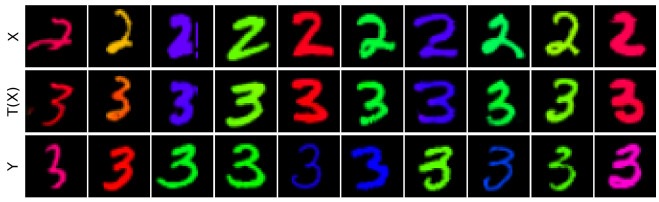

Figure 9. Random Test Images using Regularization

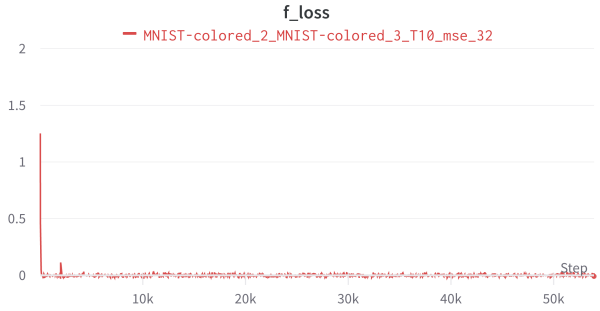

Figure 10. Loss-function using regularization calculated until 53900 iterations

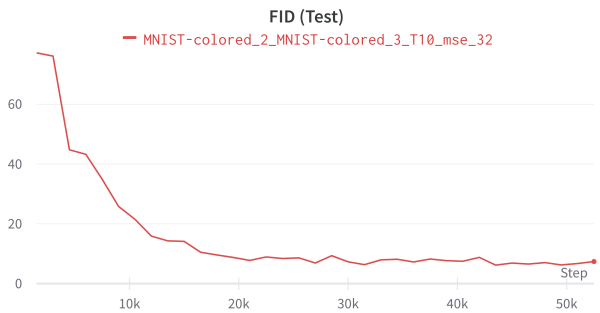

Figure 11. FID score using regularization

We have been running our experiments using GPU from Google-Colab with additional assistance from optimizing

the browser to prevent Google-Colab from getting disconnected, and it consumes around seven hours for performing task without regularization whereas GPU-Tesla P1000-PCIE-12GB consumes around one and half day for performing regularization task with $lr = 0.01$. The loss function of both without regularization and with regularization had irregularly risen and fall while the trend had asymptotically been undergoing to zero-value. On the other side, in the aspect of rate of convergence, FID score using regularization is more excellent compared to FID score without regularization. Thus, the images quality during regularization (as shown in figure.8 and figure.9) are well-performed compared to the images quality without regularization (as shown in figure.4 and figure.5)

5.3. Target-to-Source Mapping with the Discriminator of Reg-NOT

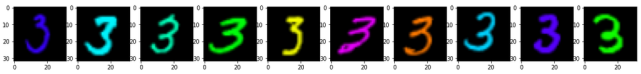

Figure 12. Digits considered for the inverse mapping

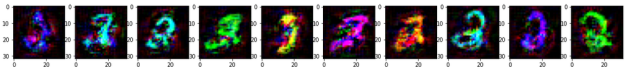

Figure 13. inverses of the digits 3 of figure 16

From the above picture we can see that despite the fact that our proposed algorithm converges it still have some difficulties to realise the inverse mapping task. However we can see that at least the colors seems being conserved on only problem is the kind of blurring due to some kind of noise.

5.4. Comparing NOT and Reg-NOT on inverse mapping problem

Note: The models used are saved at the 7499 step

Using Reg-NOT

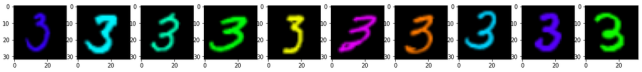

Figure 14. Digits considered for the inverse mapping

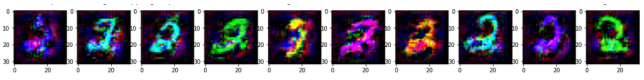

Figure 15. inverses of the digits 3 of figure 16

Using NOT

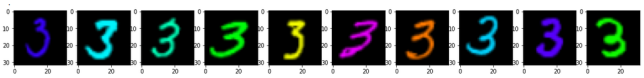

Figure 16. Digits considered for the inverse mapping

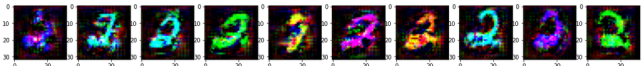

Figure 17. inverses of the digits 3 of figure 16

From both experiments we can see that none of the two algorithms successfully mapped the digits three ”3” correctly to the digits two ”2”.

6. Conclusion and Future Work

In particular, we can retain from our experiment that by adding a convexity regularizer to the Neural Optimal Transport algorithm, the algorithm is still converged and can learn successfully for the source-to-target OT map. However, it remains challenging to successfully tackle with the target-to-source OT map and will be our upcoming task to handle this sub-problem.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Kantorovitch, L. On the translocation of masses. *Management Science*, 5:1–4, 1958.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport, 2023.