The Monge Gap: A Regularizer to Learn All Transport Maps

Alexander Palanevich [1]

Abstract

Optimal transport (OT) theory has been used in machine learning to study maps that can push-forward efficiently a probability measure onto another. Recent works (Makkuva et al., 2020; Korotin et al., 2019) consider maps $T = \nabla f_\theta$, where f_θ is an input convex neural network (ICNN), and fit θ with SGD using samples. A new paper proposes a radically different approach: the Monge gap (Uscidda & Cuturi, 2023), which introduces a regularizer to solve the OT problem for arbitrary costs using a simple minimization algorithm. The goal of the project is to test the proposed regularization. We provide implementations for regularizers, MLP models, training procedure. We also measure the performance of the algorithm and compare it with the results from the original paper.

Github repo: https://github.com/Sisha3342/MongeGap

1. Introduction

1.1. Optimal transport

At the core of many machine learning challenges lies the problem of learning a map $T : \mathbb{R}^d \to \mathbb{R}^d$ that is able to push-forward a probability measure μ into another ν, i.e., $T\sharp\mu = \nu$. In many applications only unmatched samples $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ from μ and $(\mathbf{y}_1, \dots, \mathbf{y}_m)$ from ν are provided, requiring a distributional approach to estimate T. In this work we focus on neural OT solvers, where T is parametrized as a neural network. That area has been largely shaped by Brenier's theorem, which states that when the cost is the squared Euclidean distance, OT maps should follow the gradients of a convex potential. Leveraging that result, Makkuva et al. (2020); Korotin et al. (2019) provided a blueprint to use input convex neural networks (ICNN) for OT estimation.

[1]Yandex School of Data Analysis, Moscow, Russia. Correspondence to: Alexander Palanevich <alexander.palanevich@gmail.com>.

1.2. Limitations of ICNNs for OT

The usage of ICNNs for OT runs into many challenges (Korotin et al., 2021): some of their parameters must be non-negative, their initialization is a subject of ongoing researches, and training requires approximating a convex conjugate with a min-max formulation. On a more fundumental level, Brenier's theorem works when the input measure is a density, while usually we deal with sample measures. One might therefore question the relevance of imposing the double requirement for a candidate map to be the gradient of a convex potential. For general costs c, these requirements are equivalent to a c-concavity constraint that is even more intractable when trying to generalize ICNNs to other costs.

1.3. The Monge gap

The Monge gap (Uscidda & Cuturi, 2023) is a new way to learn OT maps. Rather than imposing architecture constraints, authors make no assumptions on T and, instead, introduce a regularizer which quantifies whether T agrees with the theoretical properties needed for T to be an OT map. That regularizer allows a more efficient trade-off to train maps that should be OT-like, rather than exactly conforming to the OT theory. Two learning procedures are proposed: for generic costs and costs which satisfy twist condition. Authors claim that their regularized approach outperforms both ICNNs and vanilla MLPs on synthetic benchmarks (Korotin et al., 2021) and single-cell data (Bunne et al., 2021).

1.4. Contributions

The goal of the project is to implement the Monge gap regularizer and test its performance on selected datasets. We summarize our contributions as follows:

- We implement the proposed regularizers and MLP models (for both generic and structured costs).
- We present some examples of the algorithm's training/evaluating steps.
- We compare our implementation's performance with the stated one on synthetic benchmarks (Korotin et al., 2021) and single-cell data (Bunne et al., 2021).

2. Preliminaries

2.1. Monge and Kantorovich problems

We consider throughout this work a compact subset $\Omega \subset \mathbb{R}^d$, a continuous cost function $c : \Omega \times \Omega \to \mathbb{R}$ and two probability distributions μ, ν, which are absolutely continuous w.r.t. the Lebesgue measure. The Monge problem consists of finding a map (among all maps $T : \Omega \to \Omega$ that push-forward μ onto ν) which minimizes the averaged displacement cost:

$$W_c(\mu, \nu) := \inf_{T\sharp\mu=\nu} \int_\Omega c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\mu(\mathbf{x}). \quad (1)$$

We call any solution to (1) a c-OT map between μ and ν. Solving this problem is difficult: the constraint set is not convex and can even be empty. Instead of transport maps, the Kantorovich formulation of OT seeks for couplings $\pi \in \Pi(\mu, \nu)$, i.e., probability measures on $\Omega \times \Omega$ that have μ and ν as respective marginals:

$$W_c(\mu, \nu) := \min_{\pi\in\Pi(\mathbf{x},\mathbf{y})} \iint_{\Omega\times\Omega} c(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}, \mathbf{y}). \quad (2)$$

An optimal coupling π^* for the problem (2) always exists.

2.2. Primal-dual relashionship

For any $\varphi : \Omega \to \mathbb{R}$, writing $\varphi^c(\mathbf{y}) = \inf_{\mathbf{x}} c(\mathbf{x}, \mathbf{y}) - \varphi(\mathbf{x})$ as its c-transform, one can derive the Kantorovich dual:

$$W_c(\mu, \nu) := \min_{\varphi:\Omega\to\mathbb{R}} \int_\Omega \varphi\mathrm{d}\mu + \int_\Omega \varphi^c\mathrm{d}\nu. \quad (3)$$

Let φ^* be the optimal potential. When $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$ with $h : \Omega \to \mathbb{R}$ beeing strictly convex and h^* beeing its convex conjugate, the optimal map reads (Santambrogio, 2015):

$$T^* : \mathbf{x} \mapsto \mathbf{x} - \nabla h^* \circ \nabla\varphi^*(\mathbf{x}). \quad (4)$$

When $h = \frac{1}{p} \|\cdot\|_p^p$ with $p \geq 1$ and q s.t. $\frac{1}{p} + \frac{1}{q} = 1$, we have $h^* = \frac{1}{q} \|\cdot\|_q^q$. In particular, when $h = \|\cdot\|_2^2$ one recovers the Brenier Theorem: $T^* = \mathrm{Id} - \nabla\varphi^*$.

2.3. Entropic regularization

When both μ and ν are instantiated as samples, the Kantorovich problem's (2) objective can be smoothed out using entropic regularization (Cuturi, 2013). For $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, \hat{\nu}_n = \frac{1}{m}\sum_{j=1}^m \delta_{\mathbf{y}_j}$ and $\epsilon > 0$, we form $\mathbf{C} = [c(\mathbf{x}_i, \mathbf{y}_j)]_{ij}$ and set:

$$W_{c,\epsilon} := \min_{\mathbf{P}} \langle \mathbf{P}, \mathbf{C}\rangle - \epsilon H(\mathbf{P}), \quad (5)$$

where $\mathbf{P} \in \mathbb{R}_+^{n\times m}, \mathbf{P}\mathbf{1}_m = \frac{1}{n}\mathbf{1}_n, \mathbf{P}^T\mathbf{1}_\mathbf{n} = \frac{1}{m}\mathbf{1}_m$ and $H(\mathbf{P}) = -\sum_{i,j} \mathbf{P}_{ij} \log(\mathbf{P}_{ij})$. In addition to resulting in better computational and statistical performance (Chizat et al., 2020), entropic regularization also results in a strongly convex problem with a unique solution. Besides, one can also define the Sinkhorn divergence:

$$S_{c,\epsilon} := W_{c,\epsilon}(\mu, \nu) - \frac{1}{2}(W_{c,\epsilon}(\mu, \mu) + W_{c,\epsilon}(\nu, \nu)). \quad (6)$$

Under some assumptions on c (Feydy et al., 2019), this is a valid discrepancy measure between probability distributions. Notably, the quadratic cost ℓ_2^2 satisfies theses assumptions.

3. Related works

3.1. Neural OT map estimation

As recalled in the introduction, duality theory can guide the choice of neural OT architectures. This motivates using ICNNs for squared-Euclidean costs. Within the original paper, some of the ICCN approaches (Bunne et al., 2022; Fan et al., 2020) were used to compare their peformance with the Monge gap regularization (see Section 3.2).

3.2. The Monge gap

The Monge gap (Uscidda & Cuturi, 2023) is a regularizer to estimate optimal transport maps with any ground cost c. As the goal of the project is to implement and test this particular algorithm, we cover some of its properties in details. The full list of properties and their proofs can be found in the original paper.

3.2.1. Definition and properties

Given a cost c and a reference measure ρ, the Monge gap of a vector field $T : \Omega \to \Omega$ is defined as:

$$\mathcal{M}_\rho^c(T) := \int_\Omega c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\rho(\mathbf{x}) - W_c(\rho, T\sharp\rho). \quad (7)$$

The regularizer has the following properties:

- For any vector field T, $\mathcal{M}_\rho^c(T) \geq 0$.
- T is a c-OT map between ρ and $T\sharp\rho \leftrightarrow \mathcal{M}_\rho^c(T) = 0$.
- Let μ, ν be probability measures on Ω, $\mathrm{Spt}(\mu) \subset \mathrm{Spt}(\rho)$ and $T\sharp\mu = \nu$. Then $\mathcal{M}_\rho^c(T) = 0$ implies that T is a c-OT map between μ and ν.

3.2.2. Estimation

Given emperical measures $\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, T\sharp\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^n \delta_{T(\mathbf{x}_i)}$, we can simply consider the plug-in estimator $\mathcal{M}_{\hat{\rho}_n}^c(T)$. Under mild assumptions that $T\sharp\hat{\rho}_n \to T\sharp\rho$ in law, it almost surely holds:

$$\lim_{n\to+\infty} \mathcal{M}_{\hat{\rho}_n}^c(T) = \mathcal{M}_\rho^c(T). \quad (8)$$

Evaluating the Monge gap $\mathcal{M}_\rho^c(T)$ requires solving an OT problem. To alleviate computational issues, authors use an

entropic regularization, as introduced in 5:

$$\mathcal{M}^c_{\hat{\rho}_n,\epsilon}(T) := \frac{1}{n}\sum_{i=1}^{n} c(\mathbf{x}_i, T(\mathbf{x}_i)) - W_{c,\epsilon}(\hat{\rho}_n, T\sharp\hat{\rho}_n). \quad (9)$$

The estimator in (9) retains many of the appealing properties of the unregularized Monge gap:

- Choosing $\epsilon = 0$, one recovers $\mathcal{M}^c_{\hat{\rho}_n,0}(T) = \mathcal{M}^c_{\hat{\rho}_n}(T)$.
- For $\epsilon > 0$, one has $\mathcal{M}^c_{\hat{\rho}_n,\epsilon}(T) > 0$.

We must note that after introducing an entropic regularization, we no longer have $\mathcal{M}^c_{\hat{\rho}_n,\epsilon}(T) = 0$ when T is optimal.

3.2.3. Learning with the Monge gap

Let μ, ν be the source and target measures and T_θ be a parametrized family of maps. The OT problem 1 balances two goals: (i) ensure $T_\theta\sharp\mu = \nu$, while (ii) minimizing the averaged cost of this displacement. The Monge gap can handle (ii), while any fitting loss defined through a divergence Δ would work. This translates to the following optimization problem:

$$\min_\theta \mathcal{L}(\theta) := \Delta(T_\theta\sharp\mu, \nu) + \lambda_{\text{MG}}\mathcal{M}^c_\rho(T). \quad (10)$$

3.2.4. Handling costs with structure

For costs $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$ with h beeing strictly convex, the minimization problem (10) can be refined. Using 4, $\nabla\varphi^*(\mathbf{x})$ can be directly parametrized as F_θ:

$$T_\theta : \mathbf{x} \mapsto \mathbf{x} - \nabla h^* \circ F_\theta(\mathbf{x}). \quad (11)$$

Since F_θ intends to parametrize a conservative vector field, authors follow recent papers that propose a regularization penalizing a lack of conservativity (Chao et al., 2022). This regularizer penalizes the assymetry of $\text{Jac}_\mathbf{x}F$ for $\mathbf{x} \sim \rho$:

$$\mathcal{C}_\rho(F) = \mathbb{E}_{\mathbf{x}\sim\rho}\left\|\text{Jac}_\mathbf{x}F - \text{Jac}^T_\mathbf{x}F\right\|_2^2. \quad (12)$$

Computing the full Jacobian might be too costly, so Hutchinson (1990) trace estimator is considered to use Jacobian vector products and vector Jacobian products instead. With this in mind, the emperical counterpart for (12) translates to:

$$\mathcal{C}_{\hat{\rho}_n}(F) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|\text{Jac}_{\mathbf{x}_i}F\mathbf{v}_j - \text{Jac}^T_{\mathbf{x}_i}F\mathbf{v}_j\right\|_2^2, \quad (13)$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \rho$ and $\mathbf{v}_1, \ldots, \mathbf{v}_m \sim \mathcal{N}(0, \text{I}_d)$. The final optimization problem used for structured costs can be formulated as follows:

$$\begin{aligned}\min_\theta \mathcal{L}(\theta) := \Delta((\text{I}_d - \nabla h^* \circ F_\theta)\sharp\mu, \nu) \\ + \lambda_{\text{MG}}\mathcal{M}^c_\rho(\text{I}_d - \nabla h^* \circ F_\theta) + \lambda_{\text{cons}}\mathcal{C}_\rho(F_\theta).\end{aligned} \quad (14)$$

4. Algorithms and models

We fit an OT map $T_\theta = \text{I}_d - F_\theta$ with an emperical counterpart of optimization problem (14). Map F_θ is parametrized as a MLP with 3-4 hidden layers and a residual connection, all carefully initialized in a specific manner.

4.1. Initialization schemes overview

We use the same initialization schemes as in the original paper (Uscidda & Cuturi, 2023), see Figure 1. The idea is to fix a residual connection as an affine map $T_{A,b}$, while fitting hidden layers initialized with Glorot and Bengio (2010) technique. This setup makes our model behave like the chosen $T_{A,b}$ from the beginning.

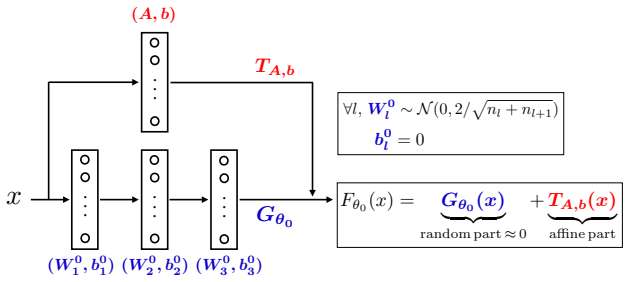

Figure 1. Initialization scheme to match affine maps applied to a 3 hidden layers MLP.

4.2. Identity initialization

The idea of the identity initialization is to make our mapping behave closely to the identity mapping. As we use the parametrization $T_\theta = \text{I}_d - F_\theta$ with F_θ beeing a MLP, we set $T_{A,b} = 0$ (not using the residual connection at all), so we have $T_{\theta_0} = \text{I}_d - F_{\theta_0} \approx \text{I}_d$.

4.3. Gaussian initialization

This strategy makes our mapping look like a linear transform between two Gaussian distributions. Using sample measures $\hat{\mu}_n, \hat{\nu}_n$, we can estimate emperical means and covariances for both source/target distributions and make the following mapping:

$$T_\mathcal{N} : \mathbf{x} \mapsto A(\mathbf{x} - m_{\hat{\mu}_n}) + m_{\hat{\nu}_n}, \quad (15)$$

where $A = \Sigma^{-1/2}_{\hat{\mu}_n}\left(\Sigma^{1/2}_{\hat{\mu}_n}\Sigma_{\hat{\nu}_n}\Sigma^{1/2}_{\hat{\mu}_n}\right)^{1/2}\Sigma^{-1/2}_{\hat{\mu}_n}$. Then, we initialize $T_{A,b} = \text{I}_d - T_\mathcal{N}$, so $T_{\theta_0} = \text{I}_d - F_{\theta_0} \approx T_\mathcal{N}$.

5. Experiments and results

Throughout the work, we fix reference measure $\rho = \mu$, fitting loss $\Delta = W_{\ell_2^2,\epsilon}$ and cost $c = \frac{1}{2}\|\cdot\|_2^2$. Whenever we run the Sinkhorn algorithm (Cuturi, 2013) to solve (5), we set $\epsilon = 0.01 \cdot \text{mean}(\mathbf{C})$. The only exception is when we compute the Sinkhorn divergence $S_{\ell_2^2,\epsilon}$ during evaluation stages, for which we use $\epsilon = 0.1$. For the Sinkhorn algorithm we set 300 iterations to converge. The hidden layers sizes are set according to the sizes of ICNNs used in the original paper, so the absolute difference between the parameters number is not greater than 1% (exact numbers can be seen in the project's repository). The number of Hutchinson vectors m for regularizer (13) is set to the 20% of the dimension d (number of features for datasets objects).

5.1. Metrics

To study the performance of the proposed regularization, we use (i) the Sinkhorn divergence between the target and the fitted target measures $S_{\ell_2^2,\epsilon}(\nu, \hat{T}\sharp\mu)$ (6) and, when OT T^* is known, (ii) the $\mathcal{L}_2$ unexplained variance percentage:

$$\mathcal{L}_2^{\text{UV}}(\hat{T}) := 100 \cdot \frac{\mathbb{E}_\mu \left\| \hat{T}(X) - T^*(X) \right\|_2^2}{\text{Var}_\nu(X)}. \qquad (16)$$

Both of the mentioned metrics ideally should be close to zero. For (16), there exists a trivial baseline model $\hat{T}(X) = \mathbb{E}\nu$ with $\mathcal{L}_2^{\text{UV}} = 100$.

5.2. High dimensional benchmark pairs

We use the Korotin et al. (2021) benchmark, providing pairs of Gaussians mixtures μ, ν in dimension $d \in \{2, 4, \dots, 256\}$, for which the optimal map for the squared Euclidean cost is known. We compare the Monge gap regularization with identity and Gaussian initializations (see Subsection 4.1) against a baseline model $\mathbb{E}\nu$. We report both $\mathcal{L}_2^{\text{UV}}$, $S_{\ell_2^2,\epsilon}(\nu, \hat{T}\sharp\mu)$ metrics and explain the differences/similarities in the obrained results.

5.2.1. Experiment setup

We set $\lambda_{\text{MG}} = 1$ and $\lambda_{\text{cons}} = 0.01$ for $d \leq 64$, and $\lambda_{\text{MG}} = 10$ and $\lambda_{\text{cons}} = 0.1$ for $d \geq 128$. We train the MLPs for $n_{\text{iters}} = 10^5$ iterations with a batch size $B = 1024$ and the Adam optimizer (Kingma & Ba, 2014). For $d \leq 64$ we use a learning rate $\eta = 0.01$, along with a polynomial schedule of power $p = 1.5$ to decrease it to 10^{-5}. For $d \geq 128$ we change the initial learning rate to $\eta = 0.001$ but keep the same polynomial schedule. When using the Gaussian initializer scheme, we instantiate it on 4096 samples.

5.2.2. Results

The obtained results are presented in Figure 2. Compared to the metrics presented by Uscidda & Cuturi (2023), we see a quite close behaviour for the $\mathcal{L}_2^{\text{UV}}$ metric, while the Sinkhorn divergence $S_{\ell_2^2,\epsilon}(\nu, \hat{T}\sharp\mu)$ has the same dynamics but worse absolute values compared to the ones presented in the paper. These differences might be explained by the uncertainty in MLPs hidden sizes: the exact strategy for increasing or decreasing layers sizes is not provided, only the final difference of 1% between MLPs and ICNNs parameters number is stated. Also there might be differences in the Sinkhorn algorithm implementations used (e.g. number of iterations to converge).

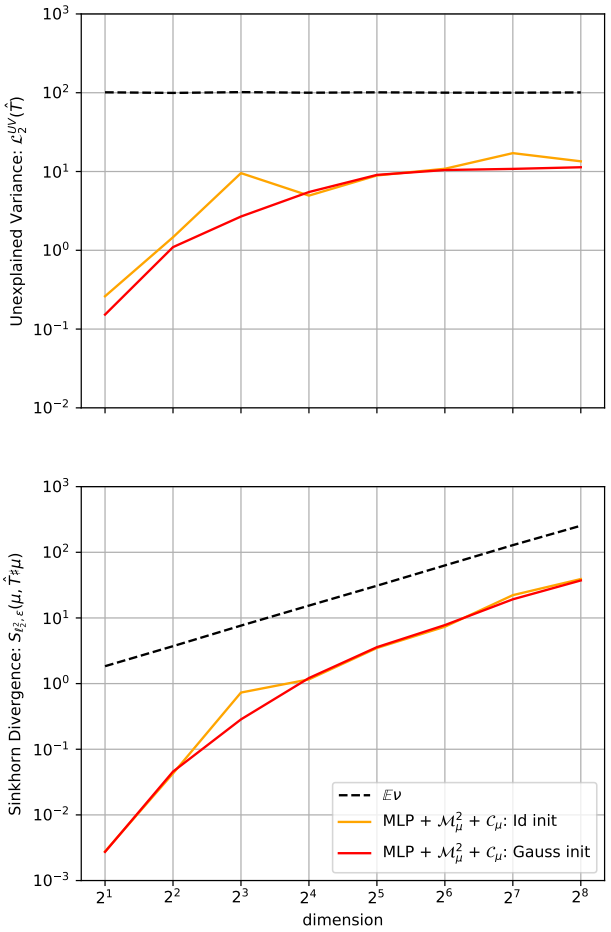

Figure 2. Performances of the Monge gap-based models and baselines on estimating the OT map between pairs of Gaussian mixtures in dimension $d \in \{2, 4, \dots, 256\}$ of the Korotin et al. (2021) benchmark.

The source code for training/evaluating process, together with logs and saved checkpoints can be found in the project's repository:

https://github.com/Sisha3342/MongeGap

5.3. Single-cell genomics

Predicting the response of cells to a perturbation is a central question in biology. In this context, feature descriptions of control and treated cells can be treated as probability measures μ and ν, and perturbation fitted as a transport map T. We predict responses of cells populations to cancer treatments using the proteomic dataset used in (Bunne et al., 2021). The dataset used can be accessed here. For simplicity we consider 4i profiling technology only. We use 10 treatments and 48 features for experiments (the full list of drugs/features selected can be found in the project's source code). Preparations regarding the dataset (e.g. features choise, train/test splitting strategy) were inspired by the following repository (Bunne et al., 2021):

https://github.com/bunnech/cellot.

5.3.1. Experiment setup

We study the performance of the regularized model with Gaussian initialization, and compare it with a vanilla MLP (i.e. without regularizers). We perform a 60%-40% train-test split on both conrol and treated cells, and evaluate the models on the 40% of unseen control and treated cells. When using regularization, we set $\lambda_{\text{MG}} = 1$ and $\lambda_{\text{cons}} = 0.01$. The models are trained for $n_{\text{iters}} = 10^4$ iterations with a batch size $B = 512$ and the ADAM optimizer using a learning rate $\eta = 0.001$, along with a polynomial schedule of power $p = 1.5$ to decrease it to $\eta = 10^{-5}$. When using regularizations, the Gaussian scheme is instantiated from the half of the training set.

5.3.2. Results

The obtained results are presented in Figure 3. Each point on the plot corresponds to its own treatment. The x axis corresponds to the Sinkhorn divergence $S_{\ell_2^2,\epsilon}$ estimated for the vanilla MLP, while the y axis stands for the Sinkhorn divergence for the regularized MLP. Compared to the original results (Uscidda & Cuturi, 2023), the performance of our implementation is worse. It still holds that the metric difference between the vanilla MLP and the regularized MLP is quite insignificant, while the absolute numbers for the Sinkhorn divergence are higher. The reasons for such differences might be the same as mentioned in 5.2.2.

6. Conclusion

During the project work we studied the new regularization approach to solve the OT problem (Uscidda & Cuturi, 2023). We implemented the proposed regularizers/models and assessed their performance on selected tasks: synthetic benchmarks (Korotin et al., 2021) and single-cell genomics (Bunne et al., 2021). For the $\mathcal{L}_2^{\text{UV}}$ metric we obtained close results, while for the $S_{\ell_2^2,\epsilon}$ metric our implementation performs notably worse than the original one. For futher work, we should consider using the alternative Sinkhorn algorithm implementations (e.g. the same approach but with more iterations to converge), make more experiments with hidden layers sizes and study all the 34 treatments for the single-cell dataset, as well as other 9 treatments for the scRNA technology. The results of the proposed experiments will help us to understand the differences in implementation's performance better.

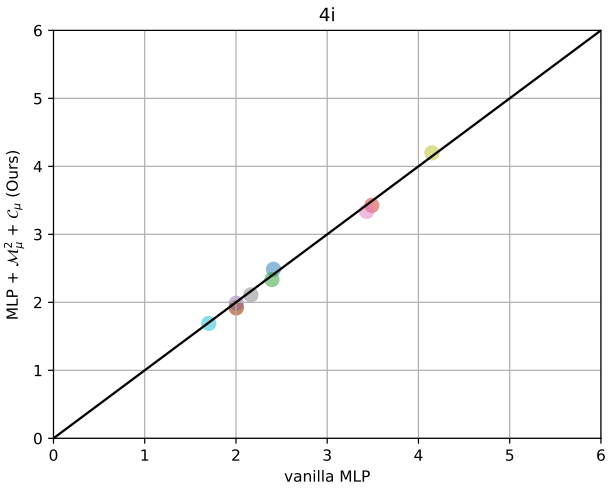

Figure 3. Fitting of a transport map $\hat{T}$ to predict the responses of cells populations to cancer treatments on 4i dataset.

References

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021. doi: 10.1101/2021.12.15.472775. URL https://www.biorxiv.org/content/early/2021/12/15/2021.12.15.472775.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Chao, C.-H., Sun, W.-F., Cheng, B.-W., and Lee, C.-Y. Quasi-conservative score-based generative models. *arXiv preprint arXiv:2209.12753*, 2022.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26, 06 2013.

Fan, J., Taghvaei, A., and Chen, Y. Scalable computations of wasserstein barycenter via input convex neural networks. *arXiv preprint arXiv:2007.04462*, 2020.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Hutchinson, M. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866. URL https://doi.org/10.1080/03610919008812866.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Korotin, A., Li, L., Genevay, A., Solomon, J. M., Filippov, A., and Burnaev, E. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Uscidda, T. and Cuturi, M. The monge gap: A regularizer to learn all transport maps. *arXiv preprint arXiv:2302.04953*, 2023.