The Monge Gap: A Regularizer to Learn All Transport Maps

Théo Uscidda [1] Marco Cuturi [1 2]

Abstract

Optimal transport (OT) theory has been used in machine learning to study and characterize maps that can push-forward efficiently a probability measure onto another. Recent works have drawn inspiration from Brenier's theorem, which states that when the ground cost is the squared-Euclidean distance, the "best" map to morph a continuous measure in $\mathcal{P}(\mathbb{R}^d)$ into another must be the gradient of a convex function. To exploit that result, Makkuva et al. (2020); Korotin et al. (2020) consider maps $T = \nabla f_\theta$, where f_θ is an input convex neural network (ICNN), as defined by Amos et al. (2017), and fit θ with SGD using samples. Despite their mathematical elegance, fitting OT maps with ICNNs raises many challenges, due notably to the many constraints imposed on θ; the need to approximate the conjugate of f_θ; or the limitation that they only work for the squared-Euclidean cost. More generally, we question the relevance of using Brenier's result, which only applies to densities, to constrain the architecture of candidate maps fitted on samples. Motivated by these limitations, we propose a radically different approach to estimating OT maps: Given a cost c and a reference measure ρ, we introduce a regularizer, the Monge gap $\mathcal{M}^c_\rho(T)$ of a map T. That gap quantifies how far a map T deviates from the ideal properties we expect from a c-OT map. In practice, we drop all architecture requirements for T and simply minimize a distance (e.g., the Sinkhorn divergence) between $T\sharp\mu$ and ν, regularized by $\mathcal{M}^c_\rho(T)$. We study $\mathcal{M}^c_\rho$ and show how our simple pipeline significantly outperforms other baselines in practice.

1. Introduction

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., $T\sharp\mu = \nu$. If one were given paired samples $(\mathbf{x}_i, \mathbf{y}_i)$, the task would amount to a simple regression, easily solved by minimizing an averaged risk $c(T(\mathbf{x}_i), \mathbf{y}_i)$. In many applications, however, only unmatched samples $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from μ and $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$ from ν are provided, requiring a distributional approach to estimate T. When the input measure μ is simple and closed-form (e.g. Gaussian, or uniform), likelihood-based methods can be used, notably normalizing flows (Rezende and Mohamed, 2015), GANs (Goodfellow et al., 2014) or even diffusion models (Song et al., 2020).

Optimal Transport and the Brenier Story. When both measures are complex and can only be accessed through samples, finding a good map T poses extra challenges. This is the case, e.g., in domain adaptation (Courty et al., 2016; 2017) or in genomics (Schiebinger et al., 2019). Optimal transport (OT) theory (Santambrogio, 2015) has emerged as a prime contender for that task (Peyré and Cuturi, 2019). We focus in this work on neural OT solvers, where T is parameterized 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. (2020) provided a blueprint to use input convex neural networks (ICNN) for OT estimation, which was later exploited in various applications, notably genomics Bunne et al. (2021).

On the limitations of ICNNs for OT. While the theory motivating ICNN solvers for OT is compelling, their practical implementation runs into many challenges (Korotin et al., 2021): some of their parameters must be non-negative, initialization them, although the subject of ongoing research (Korotin et al., 2020; Bunne et al., 2022a), is still poorly understood, and training them requires approximating a convex conjugate with a min-max formulation (Amos, 2022). On a more fundamental level, the ICNN approach may not be as sound as it seems: while Brenier (1987)'s argument is valid when the input measure μ is a density, that result does not hold for sample measures. One might therefore question the relevance of imposing the double re-

[1] CREST, ENSAE [2] Apple. Correspondence to: Théo Uscidda <theo.uscidda@ensae.fr>, Marco Cuturi <cuturi@apple.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

quirement that a candidate map 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 (Rezende and Racanière, 2021; Cohen et al., 2021). We question the need for such constraints, as also done, for instance for score functions in score-based models (Saremi, 2019).

Contributions. We propose a new approach to estimate OT maps, sturdy and generic enough to work for any cost c.

- Rather than imposing architecture choices to mimic OT maps, we make no assumption on T and, instead, introduce a *regularizer* which quantifies whether T agrees with the theoretical properties needed for T to be an OT map.
- The *Monge gap* regularizer $\mathcal{M}_\rho^c$ uses a *reference* measure ρ (that need not be necesseraly equal to μ), and is the difference between the expectation of $c(X, T(X))$, $X \sim \rho$, and the c-Wasserstein distance between ρ and $T\sharp\rho$.
- We show that the Monge gap characterizes the optimality of a map T between μ and ν. More formally, when $T\sharp\mu = \nu$ and the support $\mathrm{Spt}(\mu) \subset \mathrm{Spt}(\rho)$, we show that $\mathcal{M}_\rho^c(T) = 0$ iff T is an optimal map.
- We show that $\mathcal{M}_\rho^c$ is convex when $c(\cdot,\cdot) = \|\cdot - \cdot\|_2^2$, a property which is *still* valid when using a Sinkhorn finite-sample estimator for the 2-Wasserstein distance.
- We propose two learning procedures to estimate Monge maps using the Monge gap: (i) for general costs c, we simply add the Monge Gap of a vector field T to a fitting loss measuring the difference between $T\sharp\mu$ and the true target distribution ν and (ii) when the cost satisfies the twist condition, we take advantage of the structure induced by such costs on the optimal map, and propose instead to directly parameterize the gradient of the potential.
- We provide ample evidence on toy data, synthetic benchmarks (Korotin et al., 2021) and single-cell data that our regularized approach outperforms both ICNNs and vanilla MLPs, but also works for other more exotic costs.

2. Background on optimal transport

Monge and Kantorovich formulation. 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, $\nu \in \mathcal{P}(\Omega)$. The notation $\mu \in \mathcal{P}(\Omega)$, $\mu \ll \mathcal{L}_d$ means that μ is absolutely continuous w.r.t. the Lebesgue measure. The Monge problem consists of finding, among all map $T : \Omega \to \Omega$ that push-forward μ onto ν, that which minimizes the averaged displacement cost:

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

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 (1942) formulation of OT seeks for couplings $\pi \in \Pi(\mu,\nu)$, i.e., probability measures supported on $\Omega \times \Omega$ that have μ and ν as respective marginals:

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

An optimal coupling $\pi^\star$ always exists. When Problem (1) is feasible, both formulations coincide and $\pi^\star = (\mathrm{Id}, T^\star)\sharp\mu$.

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

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

Taking an optimal dual potential $\varphi^\star$ (which always exists under our assumptions on Ω and c) and an optimal coupling $\pi^\star$, the complementary slackness reads: $\forall(\mathbf{x}_0,\mathbf{y}_0) \in \mathrm{Spt}(\pi^\star)$, $\varphi^\star(\mathbf{x}_0) + \varphi^{\star,c}(\mathbf{y}_0) = c(\mathbf{x}_0,\mathbf{y}_0)$. Assume that $\varphi^\star$ is differentiable at $\mathbf{x}_0$, which is true under mild assumptions, and that c is sub-differentiable w.r.t. the first variable. Exploiting the definition of $\varphi^{\star,c}$:

$$(\mathbf{x}_0,\mathbf{y}_0) \in \mathrm{Spt}(\pi^*) \Leftrightarrow \nabla\varphi^\star(\mathbf{x}_0) \in \partial_1 c(\mathbf{x}_0,\mathbf{y}_0)\,. \tag{4}$$

Translation Invariant Costs. In particular, when $c(\mathbf{x},\mathbf{y}) = h(\mathbf{x}-\mathbf{y})$ with $h : \Omega \to \mathbb{R}$ strictly convex, differentiable everywhere, then its gradient map can be inverted everywhere. Indeed, one has $(\nabla h)^{-1}(\mathbf{x}) = \nabla h^*(\mathbf{x})$, with h^* the convex conjugate of h (Santambrogio, 2015, Box 1.12). Then, in that specific case, the optimal map reads:

$$T^\star : \mathbf{x} \mapsto \mathbf{x} - \nabla h^* \circ \nabla\varphi^\star(\mathbf{x})\,. \tag{5}$$

In particular, when $h = \frac{1}{2}\|\cdot\|_2^2$ one recovers the Brenier (1987) Theorem: $T^\star = \mathrm{Id} - \varphi^\star = \nabla f^\star$ where $f^\star := \frac{1}{2}\|\cdot\|_2^2 - \varphi^\star$ can be shown to be convex.

Entropic regularization. When both μ and ν are instantiated as samples, as usual in a machine learning context, the Kantorovich (1942) Problem (2) translates to a linear program, whose objective can be smoothed out using entropic regularization (Cuturi, 2013). For empirical measures $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, $\hat{\nu}_n = \frac{1}{n}\sum_{j=1}^n \delta_{\mathbf{y}_j}$ and $\varepsilon > 0$, we form $\mathbf{C} = [c(\mathbf{x}_i,\mathbf{y}_j)]_{ij}$ and set:

$$W_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) := \min_{\mathbf{P}\in U_n} \langle \mathbf{P},\mathbf{C}\rangle - \varepsilon H(\mathbf{P})\,, \tag{6}$$

where $U_n = \{\mathbf{P} \in \mathbb{R}_+^{n\times m}, \mathbf{P}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n, \mathbf{P}^T\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n\}$ is the Birkhoff polytope and $H(\mathbf{P}) = -\sum_{i,j=1}^n \mathbf{P}_{ij}\log(\mathbf{P}_{ij})$ the entropy. As ε goes to 0, one recovers the classical OT problem, namely $W_{c,0} = W_c$. In addition to resulting in better computational and statistical performance (Genevay et al., 2018; Mena and Niles-Weed, 2019; Chizat et al., 2020), entropic regularization also results in a strongly convex problem, with a unique solution, making $W_{c,\varepsilon}$ differentiable everywhere in its inputs via (Danskin, 1967)'s theorem. Besides, one can define the Sinkhorn divergence $S_{c,\varepsilon}(\mu,\nu) := W_{c,\varepsilon}(\mu,\nu) - \frac{1}{2}\left(W_{c,\varepsilon}(\mu,\mu) + W_{c,\varepsilon}(\nu,\nu)\right)$ (Ramdas et al., 2017; Feydy et al., 2019; Salimans et al., 2018; Genevay et al., 2019) which is, under some assumptions on c (see Feydy et al. (2019, Theorem 1)), a valid non-negative discrepancy measure between probability distributions. The quadratic cost satisfies theses assumptions and we we note $W_{\ell_2^2,\varepsilon}$ and $S_{\ell_2^2,\varepsilon}$ in that case.

3. The Monge Gap

We introduce in this section the Monge gap, a regularizer to estimate optimal transport maps with any ground cost c.

Definition 3.1 (The Monge Gap)**.** Given a cost c and a reference measure $\rho \in \mathcal{P}$, the Monge gap of a measurable vector field $T : \Omega \to \Omega$ is defined as:

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

By definition of Eq. (1), the Monge problem between ρ and $T\sharp\rho$ is feasible for any measure, notably discrete, since there exists at least one map, T itself, that satisfies the push-forward constraint. With this in mind, because the Monge gap is simply the optimality gap of the Monge problem, one can deduce immediately the following properties:

- For any vector field T, $\mathcal{M}^c_\rho(T) \geq 0$.
- T is a c-OT map between ρ and $T\sharp\rho \Leftrightarrow \mathcal{M}^c_\rho(T) = 0$.

Intuitively, the Monge gap $\mathcal{M}^c_\rho$ measures the gap between the cost incurred when moving from ρ to $T\sharp\rho$ using T, to the optimal one (not necessarily T) realized by a c-OT map $T^\star$. See Figure 1 for a simple illustration.

3.1. Estimation from Samples.

In practice, we estimate the Monge gap using i.i.d. samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from ρ. Given the empirical measures $\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ and $T\sharp\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^n \delta_{T(\mathbf{x}_i)}$, we simply consider the plug-in estimator $\mathcal{M}^c_{\hat{\rho}_n}(T)$, which is a consistent estimator of $\mathcal{M}^c_\rho(T)$. The proof of the following Prop. 3.2 can be found in Appendix A.1.

Proposition 3.2. *Almost surely, $\mathcal{M}^c_{\hat{\rho}_n}(T) \to \mathcal{M}^c_\rho(T)$.*

Evaluating the Monge gap $\mathcal{M}^c_{\hat{\rho}_n}(T)$ requires solving an OT problem. To alleviate computational issues, we use an entropic regularization $\varepsilon \geq 0$, as introduced in Eq. (6):

$$\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T) := \tfrac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_i)) - W_{c,\varepsilon}(\hat{\rho}_n, T\sharp\hat{\rho}_n)\,. \quad (8)$$

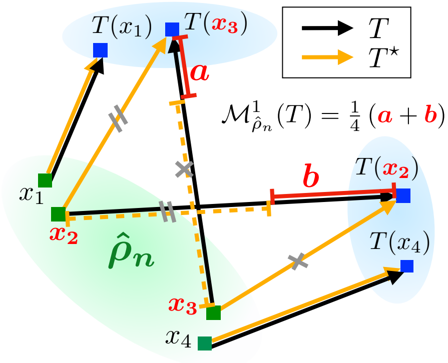


Figure 1. Sketch of the Monge Gap $\mathcal{M}^1_{\hat{\rho}_n}(T)$ instantiated with the euclidean cost $c(\cdot,\cdot) = \|\cdot - \cdot\|_2$, where $\hat{\rho}_n$ is a discrete measure supported on four points. Because the OT map $T^\star$ between $\hat{\rho}_n$ and $T\sharp\hat{\rho}_n$ does not coincide with T (notably on points x_2, x_3), the Monge gap here is positive, and equal to differences in lengths that amount to $(a+b)/4$ in the plot.

The estimator in Eq. (8), while being far more effective to compute, retains many of the appealing properties of the unregularized Monge gap:

- Choosing $\varepsilon = 0$, one recovers $\mathcal{M}^c_{\hat{\rho}_n,0}(T) = \mathcal{M}^c_{\hat{\rho}_n}(T)$.
- For $\varepsilon > 0$, one has $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T) > 0$ (see Appendix A.3).

Moreover, entropic regularization makes this estimator is differentiable everywhere in its inputs. See § C.1 for more insights on the Monge gap gradient.

3.2. Relation to Cyclical Monotonicity.

To gain intuition about what $\mathcal{M}^c_{\hat{\rho}_n}$ quantifies, we introduce the notion of cyclical monotonicity. Recall that a set $\Gamma \subset \Omega \times \Omega$ is c-CM if for any $n \in \mathbb{N}$, any set $\{\mathbf{x}_1, ..., \mathbf{x}_n\} \times \{\mathbf{y}_1, ..., \mathbf{y}_n\} \subset \Gamma$ and permutation $\sigma \in \mathcal{S}_n$ one has:

$$\sum_{i=1}^n c(\mathbf{x}_i, \mathbf{y}_i) \leq \sum_{i=1}^n c(\mathbf{x}_i, \mathbf{y}_{\sigma(i)})\,.$$

Setting $\mathbf{y}_i := T(\mathbf{x}_i)$, the Monge gap estimator using permutations (Peyré and Cuturi, 2019, Proposition 2.1) is:

$$\mathcal{M}^c_{\hat{\rho}_n}(T) = \frac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_i)) - \min_{\sigma\in\mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_{\sigma(i)})\,,$$

can therefore be interpreted as a quantification of the violation of the cyclical monotonicity of the set $\Gamma := \mathrm{Spt}\left((\mathrm{Id}, T)\sharp\rho\right)$, measured on sampled points $\{(\mathbf{x}_1, T(\mathbf{x}_1)), ..., ..., (\mathbf{x}_n, T(\mathbf{x}_n))\} \subset \Gamma$. Under the assumptions made on c and Ω, the cyclical monotonicity of that set is equivalent to the optimality of T, see (Santambrogio, 2015, Theorem 1.38, Theorem 1.49).

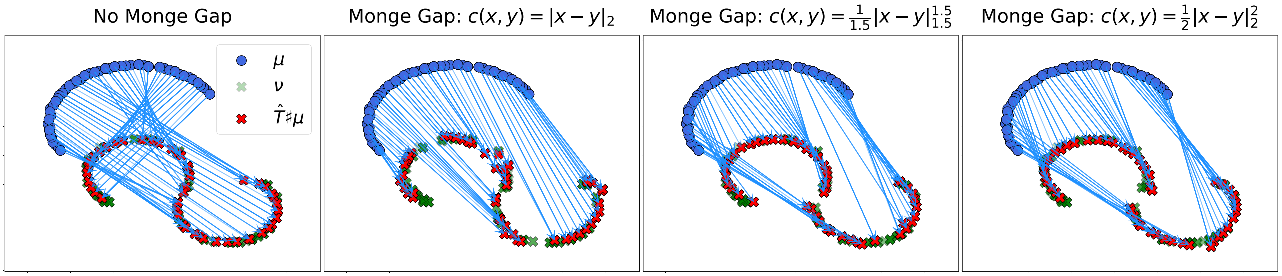


Figure 2. Fitting of transport maps between measures μ, ν in dimension $d = 2$, with the same fitting loss $\Delta = W_{2,\varepsilon}$ but Monge gap $\mathcal{M}_\mu^c$ instantiated with various costs c. We also fit an MLP without Monge gap, minimizing only the fitting loss. For $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$, we use the method for generic costs §4.1, directly parameterizing T_θ as an MLP and using $\lambda_{\text{MG}} = 5$. For strictly convex costs $c(\mathbf{x}, \mathbf{y}) = \frac{1}{1.5}\|\mathbf{x} - \mathbf{y}\|_{1.5}^{1.5}$ and $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, we use the method for costs with structure §4.2. Accordingly, we parameterize $T_\theta = \mathrm{I}_d - \nabla h^* \circ F_\theta$ with an MLP F_θ and penalize lack of conservativity with $\mathcal{C}_\mu$. Moreover, we use $\lambda_{\text{MG}} = 1$ and $\lambda_{\text{cons}} = 0.01$.

3.3. Properties of the Monge Gap.

When the Monge gap w.r.t. ρ of a map T is zero, then it will also be zero on *any* measure whose support is contained in that of ρ. This is a crucial property of our regularizer and a natural extension of (Brenier, 1987)'s result for the ℓ_2^2 cost, which states that a map is optimal between ρ onto $T\sharp\rho$, if and only if it is the gradient of a convex potential; assuming that is true, that map will therefore move optimally *any* measure whose support is contained in that of ρ. The proof of the following Prop. 3.3 can be found in Appendix A.2.

Proposition 3.3. *Let $\mu, \nu \in \mathcal{P}(\Omega)$ such that $\mathrm{Spt}(\mu) \subset \mathrm{Spt}(\rho)$, and a map T s.t. $T\sharp\mu = \nu$. Then $\mathcal{M}_\rho^c(T) = 0$ implies that T is a c-OT map between μ and ν.*

The Quadratic Case. We now focus on the Monge gap when $c(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_2^2$, abbreviated as $\mathcal{M}_\rho^2$. We study the properties of $\mathcal{M}_\rho^2$ and its empirical (entropic) counterpart $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$, for $\varepsilon \geq 0$, on $L_2(\rho)$ and $L_2(\hat{\rho}_n)$ respectively. Notably, they share the same appealing regularity properties.

Proposition 3.4. *Both $\mathcal{M}_\rho^2$ and $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$ for any $\varepsilon \geq 0$, are convex, sub-additive and positively homogeneous.*

Proof. We start by studying $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$ since it can be reformulated as a matrix input function. Indeed, it only depends on T via its values on the support of $\hat{\rho}_n$, namely $\mathbf{x}_1, ..., \mathbf{x}_n$. Therefore, we write $\mathbf{t}_i := T(\mathbf{x}_i)$ and study:

$$r(\mathbf{T}) := \tfrac{1}{2n}\|\mathbf{X} - \mathbf{T}\|_F^2 - W_{\ell_2^2,\varepsilon}(\hat{\rho}_n, \rho_{\mathbf{T}}),$$

where $\mathbf{X}, \mathbf{T} \in \mathbb{R}^{n \times d}$ contain observations $\mathbf{x}_i$ and $\mathbf{t}_i$ respectively, stored as rows, and $\rho_{\mathbf{T}}$ is the discrete measure supported on the $\mathbf{t}_i$. Expanding the squares yields:

$$r(\mathbf{T}) = \max_{\mathbf{P} \in U_n} \langle \mathbf{T}, (\mathbf{P} - \tfrac{1}{n} I_n)^\top \mathbf{X} \rangle + \varepsilon H(\mathbf{P}) \tag{9}$$

As a maximum of affine function, r is then convex, subaddtitive and positively homogeneous in $\mathbf{T}$, so is $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$ in T. We extendit to $\mathcal{M}_\rho^2$ using Prop. 3.2. See Appendix A.4. □

Insights of Prop. 3.4 are twofold. First, the convexity supports using $\mathcal{M}_\rho^2$ as a regularizer to gain Monge optimality. Furthermore, if $\rho \ll \mathcal{L}_d$, by Brenier (1987)'s Theorem, $\mathcal{M}_\rho^2$ is zero only on the set $\{F \mid \exists\, f : \Omega \to \mathbb{R} \text{ convex s.t. } F = \nabla f,\ \rho - \text{p.p.}\}$. Therefore, $\mathcal{M}_\rho^2$ behaves like a norm "up to the gradient of the convex functions".

Proposition 3.5. *For any $\gamma \in \mathcal{P}(\Omega)$, we note $C(\gamma) := 2\left(\lambda_{\max}\left(\mathbb{E}_\gamma[XX^\top]\right)\right)^{1/2}$. Then, $\mathcal{M}_\rho^2$ and $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$ are respectively $C(\rho)$ and $C(\hat{\rho}_n)$ Lispchitz continuous.*

Proof. We study the lipschitzness of the affine function over which maximization is performed in Equation (9). It allows to deduce the lipschitzness of r w.r.t. $\|\cdot\|_F$, hence of $\mathcal{M}_{\hat{\rho}_n,\varepsilon}^2$ w.r.t. $\|\cdot\|_{L_2(\hat{\rho}_n)}$. We then extend it to $\mathcal{M}_\rho^2$ w.r.t. $\|\cdot\|_{L_2(\rho)}$ using consistency Prop. 3.2. See Appendix A.5. □

Since $C(\rho) \geq 2\left(\lambda_{\max}(\mathrm{Cov}_\rho[X])\right)^{1/2}$, Prop. 3.5 provides insights about the choice of the reference measure ρ. It exhibits the expected trade-off w.r.t. Prop. 3.3: by choosing a bigger ρ, we trade the regularity of $\mathcal{M}_\rho^2$.

4. Learning with the Monge Gap

We show how the Monge gap can be used to learn approximately c-optimal parameterized maps, for any c.

4.1. Using directly the Monge gap as a regularizer.

We seek to learn a c-OT map between $\mu, \nu \in \mathcal{P}(\Omega)$. The Monge Prob. (1) has two parts: *(i)* fitting the marginal constraint $T\sharp\mu \approx \nu$, while *(ii)* minimizing the averaged c-cost of this displacement, i.e. achieving the c-optimality.

The Weaknesses of Naive Dualization. Because of the difficulty of handling the constraints in *(i)*, previous works have simply proposed to *dualize* that constraint, through a regularizer (Lu et al., 2020; Xie et al., 2019; Bousquet et al., 2017; Balaji et al., 2020). This consists, for simplicity, in introducing a fitting loss defined through any divergence Δ and a Lagrange multiplier λ; then solving:

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^d} \underbrace{\lambda\,\Delta(T\sharp\mu,\nu)}_{\text{fitting}} + \underbrace{\int c(\mathbf{x},T(\mathbf{x}))\,\mathrm{d}\mu(\mathbf{x})}_{\text{c-optimality}}\,. \tag{10}$$

Prob. (10) coincides with the OT Prob. (1) only if $\lambda\to+\infty$. Furthermore, suppose that a minimizer T_λ of Prob. (10) exists, then we can show that T_λ is optimal between μ and $T_\lambda\sharp\mu$. On the other hand, Gazdieva et al. (2022, Theorem 1.) states that $T_\lambda\sharp\mu\neq\nu$ and this bias increases drastically as λ decreases, so as we focus on the minimization of the displacement cost. Therefore, we have an intrinsic trade-off that prevents us from jointly: *(i)* fitting the marginal constraint while *(ii)* achieving the c-optimality.

The Monge gap as a debiased displacement cost. On the other hand, the Monge gap is a regularizer that can handle the c-optimality constraint elegantly. Provided that c-OT maps exist and $\mathrm{Spt}(\rho)\supset\mathrm{Spt}(\mu)$, the solutions of:

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^d}\mathcal{L}(T) := \underbrace{\Delta(T\sharp\mu,\nu)}_{\text{fitting}} + \underbrace{\mathcal{M}^c_\rho(T)}_{\text{c-optimality}} \tag{11}$$

are exactly those c-OT maps. Indeed, $\mathcal{L}(T)\geq 0$ with equality i.f.f. $\Delta(T\sharp\mu,\nu)=0$ and $\mathcal{M}^c_\rho(T)=0$, i.e. $T\sharp\mu=\nu$ and T is optimal between μ and $T_\sharp\mu=\nu$ using Proposition 3.3. For this reason, $\mathcal{M}^c_\rho$ can be seen as a "debiased" displacement cost that is 0 only when c-optimality is observed.

Monge gap-based learning. Introducing a family of parameterized maps $\{T_\theta\}_{\theta\in\mathbb{R}^p}$ and a regularization weight λ_{MG}, we can recover approximate c-OT maps by solving:

$$\min_{\theta\in\mathbb{R}^p}\mathcal{L}(\theta) := \Delta(T_\theta\sharp\mu,\nu) + \lambda_{\mathrm{MG}}\,\mathcal{M}^c_\rho(T_\theta)\,. \tag{12}$$

The introduction of λ_{MG} is purely related to practical considerations: balancing each objective function term and hence stabilizing training. In theory and according to the above discussion, any $\lambda_{\mathrm{MG}}>0$ allows recovering a c-OT map. Usually, one can simply set $\Delta=W_c$ and $\lambda_{\mathrm{MG}}=1$, so that Δ and $\lambda_{\mathrm{MG}}\mathcal{M}^c_\rho$ are naturally homogeneous. Therefore, λ_{MG} is easy to tune by construction.

4.2. Handling Costs with Structure.

Adapting the Map's Parametrization. The method described in § 4.1 can be refined when the cost introduces structure in the optimal map. As recalled in § 2, when $c(\mathbf{x},\mathbf{y})=h(\mathbf{x}-\mathbf{y})$ with h strictly convex, one has:

$$T^\star:\mathbf{x}\mapsto\mathbf{x}-\nabla h^*\circ\nabla\varphi^\star(\mathbf{x}) \tag{13}$$

where $\varphi^\star$ is a dual potential. Accordingly, we can adapt the map's parameterization, introducing a parametrized vector field F_θ to model directly the dual potential gradient $\nabla\varphi^\star$:

$$T_\theta:\mathbf{x}\mapsto\mathbf{x}-\nabla h^*\circ F_\theta(\mathbf{x}). \tag{14}$$

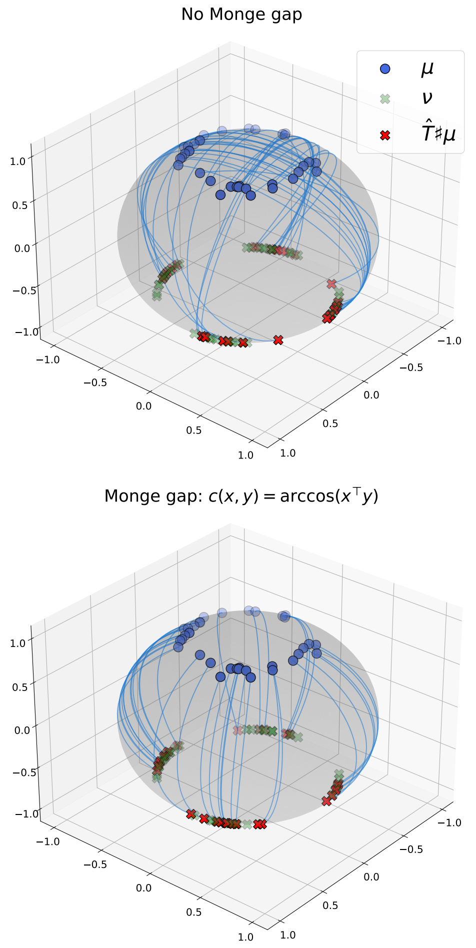


Figure 3. Fitting of transport maps between synthetic measures on the 2-sphere. In both cases, we parameterize the map as $T_\theta = F_\theta/\|F_\theta\|_2$ where F_θ is an MLP, and we use $\Delta=W_{\ell_2^2,\varepsilon}$ as fitting loss. On the upper plot, we do not use any regularize, while on the lower plot, we regularize with the Monge gap instantiated for the geodesic cost $c(\mathbf{x},\mathbf{y})=\arccos(\mathbf{x}^\top\mathbf{y})$ and use $\lambda_{\mathrm{MG}}=1$.

This case includes notably all $h=\frac{1}{p}\|\cdot\|_p^p$ with $p\geq 1$ and q s.t. $\frac{1}{p}+\frac{1}{q}=1$, for which $h^*=\frac{1}{q}\|\cdot\|_q^q$.

Penalizing Lack of Conservativity. By construction, F_θ intends to mimic a conservative vector field (i.e. a gradient field) $\nabla\varphi^\star$. To leverage this conservativity prior, we could impose the parameterization $F_\theta=\nabla f_\theta$. However, this hard gradient constraint can make training unstable and is one of the reasons why ICNNs are challenging to train (Saremi, 2019; Richter-Powell et al., 2021; Amos, 2022). Instead, we use a regularization to penalize a lack of conservativity (Chao et al., 2023). Introducing a reference measure ρ and considering a differentiable F, the regularizer penalizes the asymmetry of the jacobian $\mathrm{Jac}_{\mathbf{x}}F$ for $\mathbf{x}\sim\rho$:

$$\mathcal{C}_\rho(F)=\mathbb{E}_{X\sim\rho}\left[\|\mathrm{Jac}_X F-\mathrm{Jac}_X^T F\|_F^2\right]\,. \tag{15}$$

It can be estimated efficiently using the Hutchinson (1990)

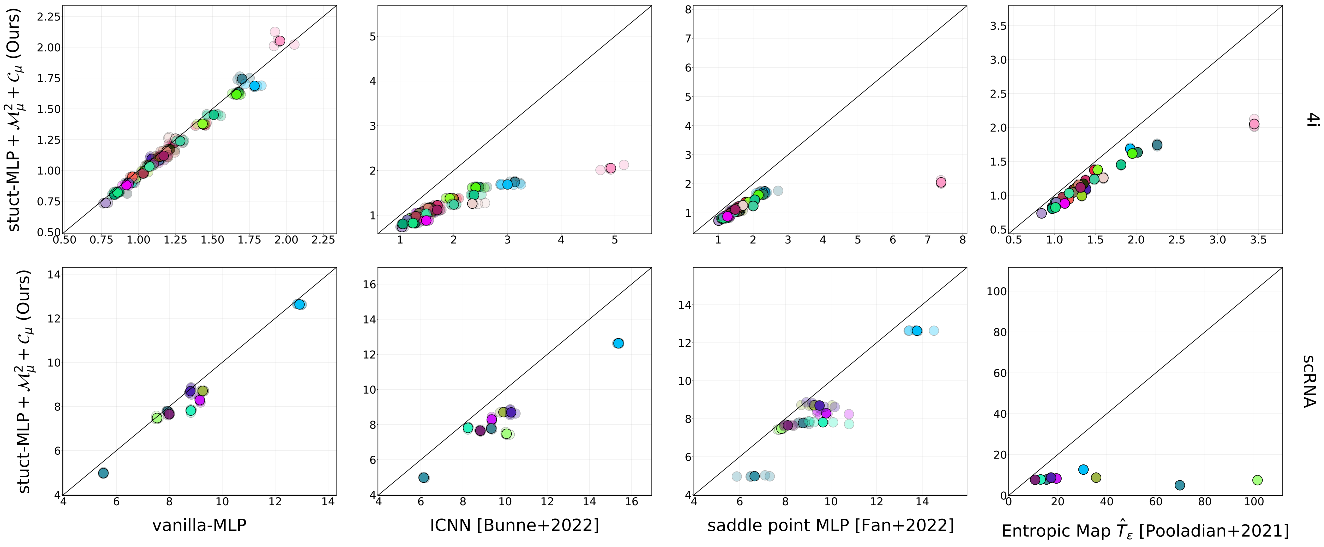


Figure 4. Fitting of a transport map $\hat{T}$ to predict the responses of cell populations to cancer treatments on 4i (upper plot) and scRNA (lower plot) datasets, providing respectively 34 and 9 treatment responses. For each profiling technology and each treatment, we compare the predictions provided by our method to those of the baselines listed in 6.1. We measure predictive performance using the Sinkhorn divergence between a batch of unseen (test) treated cells and a batch of unseen control cells mapped with $\hat{T}$, namely $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu_{\text{test}}, \nu_{\text{test}})$, see§ 6.4 and Appendix D.7 for details. Each scatter plot displays points $z_i = (x_i, y_i)$ where y_i is the divergence obtained by our method and x_i that of the other baseline on all treatments. A point below the diagonal $y = x$ refers to an experiment in which our methods outperform the baseline. We assign a color to each treatment and plot five runs, along with their mean (the brighter point). For a given color, a higher variability of points along the x axis means that our method is more stable than the baseline, and vice versa.

trace estimator, which turns pointwise Jacobians to pointwise Jacobian vector products (JVPs) and vector Jacobian products (VJPs). See Appendix B for details about this estimation procedure. Combining the parametrization trick and the conservativity regularizer, we then seek to solve:

$$\begin{aligned}\min_{\theta\in\mathbb{R}^p} &\mathcal{L}(\theta) := \Delta((\mathrm{I}_d - \nabla h^* \circ F_\theta)\sharp\mu, \nu) \\ &+ \lambda_{\mathrm{MG}}\, \mathcal{M}_\rho^c(\mathrm{I}_d - \nabla h^* \circ F_\theta) + \lambda_{\text{cons}}\, \mathcal{C}_\rho(F_\theta)\end{aligned} \tag{16}$$

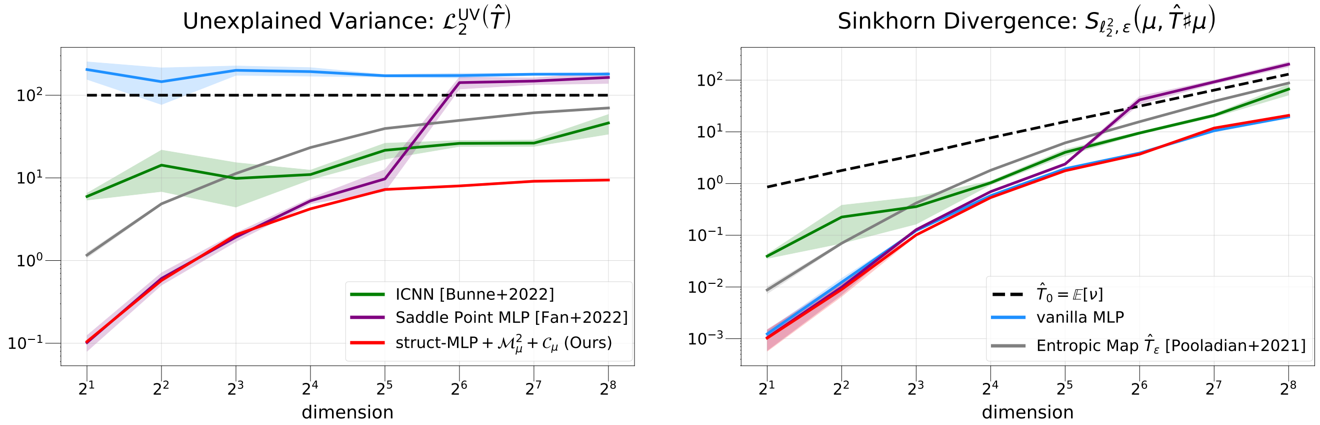
Note that the Monge gap and the conservative regularizer are not applied to the same vector field. While $\mathcal{M}_\rho^c$ is applied to $T_\theta := \mathrm{I}_d - \nabla h^* \circ F_\theta$, $\mathcal{C}_\rho$ is evaluated on F_θ, to mimic the gradient of a dual potential $\nabla\varphi^\star$. Since $\varphi^\star$ can always be taken c-concave (Santambrogio, 2015, Remark 1.13), F_θ can be thought as a *soft* Input c-concave Gradient Network.

Remark. We emphasize that the main ingredient of our method is the Monge gap. The conservative regularizer is, in theory, *not necessary*: using $\lambda_{\text{cons}} = 0$ and any $\lambda_{\mathrm{MG}} > 0$ in Prob. (16) is enough to recover a c-OT map. We propose to use it to stabilize training and reach better local minimas. We test its influence in detail in the Experiments section 6.

5. Related works

Neural OT map estimation. As recalled in the introduction, duality theory can guide the choice of neural OT architectures, using c-concavity. This motivates naturally ICNNs for the quadratic cost, but also more general c-concave neural potentials. These approaches are, however, fairly difficult to train and parameterize in practice. Fan et al. (2020) and propose an alternative approach, conceptually similar to a Wasserstein GAN (Arjovsky et al., 2017), where a Lagrange multiplier f is introduced in the Monge formulation defined in Eq. (1) to account for the marginal constraint $T\sharp\mu = \nu$. This results in a saddle point problem $\sup_f \inf_T \mathcal{L}(f, T)$, trading off two terms, a displacement cost, and a fitting loss error. The goal is then to make that displacement cost small while reaching a fitting loss as close as possible to zero. The proper trade-off between the two terms is, however, difficult to get right: the displacement cost cannot be minimized to zero (that term represents the "traveled" distance to go from source to target), and its scale will interfere with that the fitting loss (which should be, ideally, close to 0). By contrast, in our approach, both the fitting loss and the Monge gap (which can be interpreted as a "debiased" displacement cost) should be close to 0. In that sense, the Monge gap is truly a regularizer and not a displacement cost.

Beyond maps. The Kantorovitch formulation can also be reformulated as a saddle point problem, by relaxing $\pi \in \Pi(\mu, \nu)$ to $\pi \in \Pi(\mu)$ and introducing a Lagrange multiplier for the second marginal constraint. A recent line of work proposes to directly estimate non-deterministic parameterized couplings $\pi_\theta \in \Pi(\mu)$, modeling $\pi_\theta(\mathbf{y}|\mathbf{x})$


Unexplained Variance: $\mathcal{L}_2^{\mathrm{UV}}(\hat{T})$

Sinkhorn Divergence: $S_{\ell_2^2,\varepsilon}(\mu, \hat{T}\sharp\mu)$

Figure 5. Performances of our models and baselines on estimating the ground-truth Monge maps $T^\star$ between each pair of Gaussian mixtures μ, ν in dimension $d \in \{2, 4, 8, ..., 256\}$ of the Korotin et al. (2021)'s benchmark. We report both the unexplained variance $\mathcal{L}_2^{\mathrm{UV}}(\hat{T})$ (left) and the Sinkhorn divergence $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu, \nu)$ (right) of the fitted map $\hat{T}$. We average the results over 5 trainings.

via "one to many" stochastic maps (Yang and Uhler, 2019; Korotin et al., 2022; Asadulaev et al., 2022). More precisely, for a latent space $\mathcal{Z}$, take $\gamma \in \mathcal{P}(\mathcal{Z})$ and a stochastic map $T_{\pi_\theta} : \Omega \times \mathcal{Z} \to \Omega$, if $\mathbf{x} \sim \mu$, then for any $\mathbf{z} \sim \gamma$, $(\mathbf{x}, T_{\pi_\theta}(\mathbf{x}, \mathbf{z})) \sim \pi_\theta \in \Pi(\mu)$. Imposing deterministic couplings $\pi_\theta = (\mathrm{I}_d, T_\theta)\sharp\mu$, we recover the saddle point Monge problem of Fan et al. (2020), which is why we only consider Fan et al. (2020) as a baseline in our experiments.

6. Experiments

We evaluate the ability of our method to recover OT maps between both synthetic (§6.2,6.3) and real (§6.4) datasets.

6.1. Experimental Setting.

Using the Monge gap in practice. The `monge_gap`, along with a `MapEstimator` to estimate OT maps, are implemented in the OTT-JAX (Cuturi et al., 2022) package.[1]

Reference measure. Choosing the reference measure ρ is the first step in our construction. We provide preliminary experiments in Appendix C to investigate the influence of ρ. We settle in practice on the simplest choice of setting $\rho = \mu$ and leave other choices for future research.

Our models. We quote our models using this terminology:

- The prefix refers to the map's parameterization. `vanilla-MLP` indicates that we directly parameterize the map with an MLP and solve Prob. (12), with `algorithm` 1. `strcut-MLP` indicates that we use the parameterization trick and solve Prob. (16), see `algorithm` 2. The latter works for strictly convex costs.
- The suffix refers to the employed regularizers. We add $+\,\mathcal{M}_\mu^c$ to the name when we use the Monge gap, and $+\,\mathcal{C}_\mu$ when we use the conservative regularizer.

For strictly convex costs $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$, such as the quadratic cost, we use `struct-MLP` $+\,\mathcal{M}_\mu^c + \mathcal{C}_\mu$ by default. For generic costs, we use `vanilla-MLP` $+\,\mathcal{M}_\mu^c$. With or without regularizers, we fit our models with $\Delta = W_{\ell_2^2,\varepsilon}$. We adapt Bunne et al. (2022a) to define both Gaussian (for the quadratic cost) and Identity initialization schemes for our neural transport maps (see Appendix D.2). See Appendix D for other details about hyperparameters.

Metrics. To measure the predictive performances of an estimator $\hat{T}$ of $T^\star$, we rely on (i) the Sinkhorn Divergence between the target and the fitted target measures, namely $S_{\ell_2^2,\varepsilon}(\nu, \hat{T}\sharp\mu)$ and, when $T^\star$ is known, (ii) the $\mathcal{L}_2$ unexplained variance percentage (Makkuva et al., 2020), (Korotin et al., 2020), (Korotin et al., 2021) defined as:

$$\mathcal{L}_2^{\mathrm{UV}}(\hat{T}) := 100 \cdot \frac{\mathbb{E}_\mu[\|\hat{T}(X) - T^\star(X)\|_2^2]}{\mathrm{Var}_\nu(X)}\,. \tag{17}$$

$S_{\ell_2^2,\varepsilon}(\nu, \hat{T}\sharp\mu)$ quantifies the generative power of the method as a valid divergence between the reconstructed and the actual target. For all experiments, we use $\varepsilon = 0.1$. Instead, $\mathcal{L}_2^{\mathrm{UV}}(\hat{T})$ quantifies not only this generative power but the Monge optimality, measuring the deviation of $\hat{T}$ from $T^\star$. This deviation is normalized by the variance of ν, so that the constant baseline $\hat{T}_0 = \mathbb{E}_\nu[Y]$ provides $\mathcal{L}_2^{\mathrm{UV}}(\hat{T}_0) = 100\%$.

Baselines. We compare our methods to (i) a `vanilla-MLP` fitted without regularization, (ii) the ICNN neural dual formulation with Gaussian initializer (Bunne et al., 2022a), (iii) an MLP trained via the saddle point problem (Fan et al., 2020), (iv) the entropic map (Pooladian and Niles-Weed, 2021) and (v) the constant map $\hat{T}_0 = \mathbb{E}_\nu[Y]$.

[1] https://github.com/ott-jax/ott

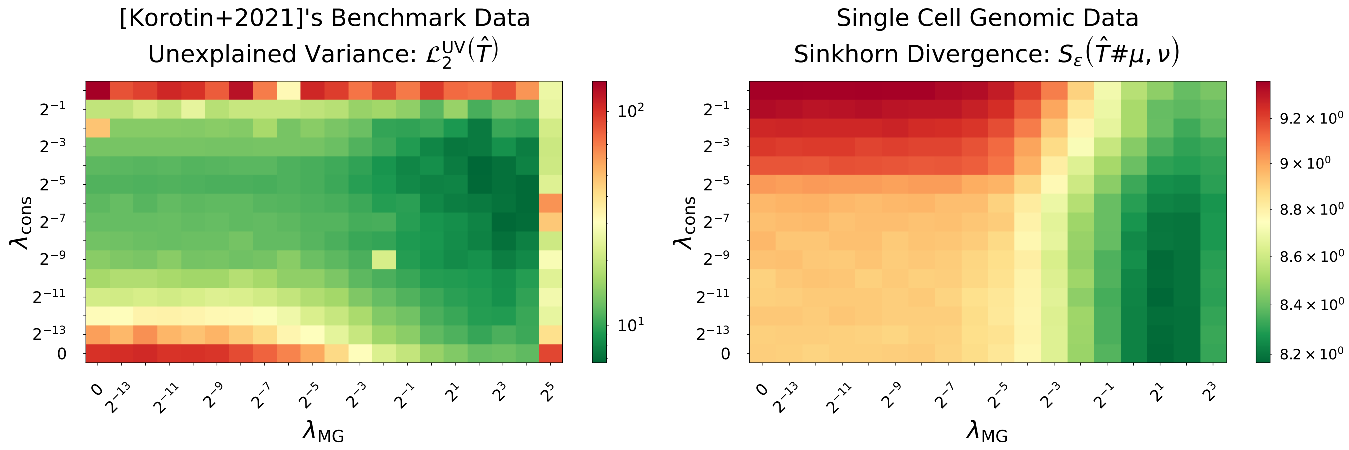


Figure 6. Heatmaps showing the influence of $\mathcal{M}_\mu^2$ and $\mathcal{C}_\mu$ on the performances of our model `struct-MLP`$+\mathcal{M}_\rho^2 + C_\mu$ on tow tasks: (left) learning the known Monge map between the Korotin et al. (2021) benchmark pair of dimension $d = 32$ and (right) predict the responses of cells populations to *abexinostat* drug with scRNAseq data. It corresponds to 1 of the 9 scRNAseq datasets considered in 6.4. For each task, we report the performances induced by each pair of regularization weights $(\lambda_{\mathrm{MG}}, \lambda_{\mathrm{cons}})$ on a regular grid. For the benchmark pair, since the Monge map is known, we measure performances using the unexplained variance. For the single-cell data, we use the Sinkhorn divergence between a batch of unseen treated cells and a batch of unseen control cells mapped with the fitted map.

6.2. Synthetic Data.

ℓ_p^q costs. We evaluate both § 4.1 and § 4.2 methods on $(\ell_p^q)_{p,q\geq 1}$ costs, see Figure 2. For $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$, We obtain a ”needle” alignment (without crossing lines) because c is a distance: this is known as the Monge Mather shortening principle Villani (2009, Chap. 8). Thus, the Monge gap allows to recover of OT maps when c is a distance.

Costs on the sphere. We consider measures supported on the 2-sphere along with $c(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x}^\top \mathbf{y})$, see Figure 3. Since c is the geodesic distance, we still observe the Monge Mather shortening principle (see above paragraph), which assesses the c-optimality of the fitted map.

6.3. High Dimensional Benchmark Pairs.

Experimental setting. To assess that our method allows us to recover Monge maps, we use the ICNN based Korotin et al. (2021)'s benchmark, providing pairs of Gaussians mixtures μ, ν in dimension $\{2, 4, 8, ..., 256\}$, with known Monge map for the squared Euclidean cost. We fix $\lambda_{\mathrm{cons}} = 0.01$, then $\lambda_{\mathrm{MG}} = 1$ for $d \leq 16$ and $\lambda_{\mathrm{MG}} = 10$ for $d \geq 32$. See § 6.5 for details on the influence of hyperparameters.

Results. are shown on Figure 5. A `vanilla-MLP` exhibits good generative power but, as expected, does not learn the Monge map $T^\star$. Our method performs uniformly better for $d \geq 16$. For $d \geq 64$, the Fan et al. (2020) estimator yields poor results, worse than the constant baseline. The ICNNs provide unstable and moderate performances, despite the Bunne et al. (2022a)'s Gaussian initializer, highlighting the difficulty of their training, even when the ground truth is explicitly known as the gradient of an ICNN.

6.4. Single-Cell Genomics.

Experimental setting. 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 $\hat{T}$. Following (Schiebinger et al., 2019), the use of OT theory to recover this map $\hat{T}$ has been used (Bunne et al., 2022b; 2021; 2022a; Lübeck et al., 2022; Eyring et al., 2022). We predict responses of cell populations to cancer treatments (perturbations) using the proteomic dataset used in (Bunne et al., 2021), consisting of two melanoma cell lines. Patient data is analyzed using (i) 4i (Gut et al., 2018), and scRNA sequencing (Tang et al., 2009). For each profiling technology, the response to respectively (i) 34 and (ii) 9 treatments are provided. As in (Bunne et al., 2021), (i) training is performed with the quadratic cost, in the data space for the 4i data and in a latent space learned by the scGen autoencoder (Lotfollahi et al.) for the scRNA data and (ii) both evaluations are carried in data space, selecting the top 50 marker genes for scRNA data using the `scanpy` (Wolf et al., 2018) package. We fix $\lambda_{\mathrm{cons}} = 0.01$, then set $\lambda_{\mathrm{MG}} = 1$ for 4i data and $\lambda_{\mathrm{MG}} = 10$ for scRNA data. See 6.5 for details on the influence of hyperparameters.

Results are shown on Figure 4. On both 4i and scRNA data, our method gives the best prediction among all models. These results also show that standard MLPs trained without regularization should not be discarded as a poor contenders since they perform consistently better than ICNNs. We believe this illustrates the rigidity of the ICNN architecture (Korotin et al., 2021; Amos, 2022)).

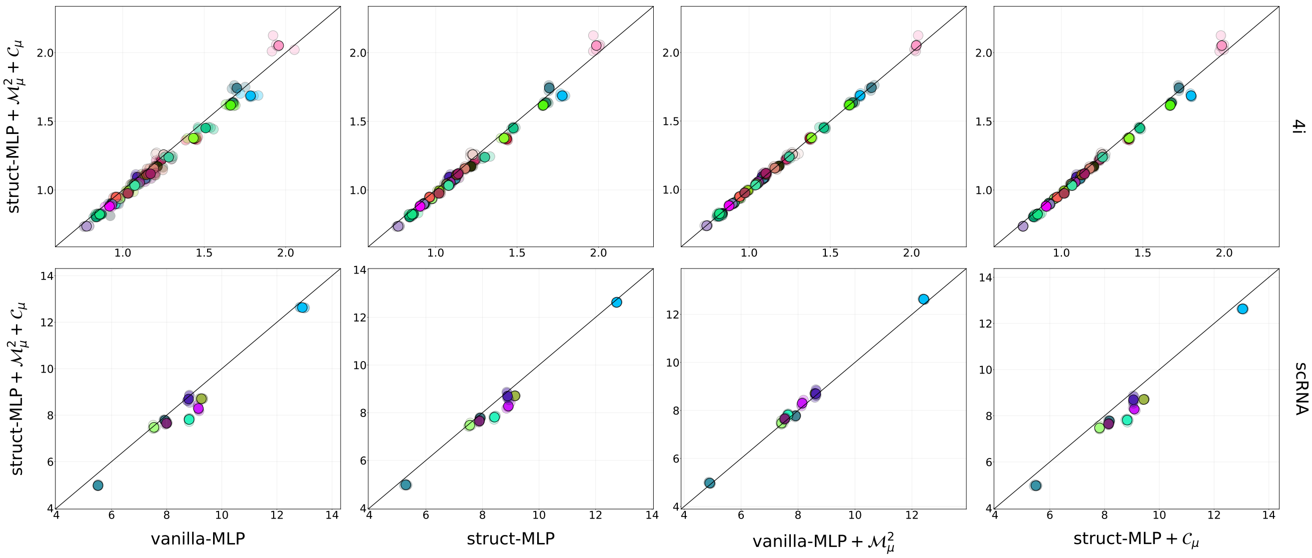


Figure 7. Ablation study on single-cell genomic data. On all 4i (upper plot) and scRNAseq (lower plot) datasets, we evaluate the effect of: (i) the parameterization trick, (ii) the Monge gap $\mathcal{M}_\mu^2$, and (iii) the conservative regularizer $\mathcal{C}_\mu$ (see Prob. (16)). The plot structure is the same as Figure 4, and we still evaluate each model using the Sinkhorn divergence between a batch of unseen treated cells and a batch of unseen control cells mapped with $\hat{T}$. We average the results over five runs. We remind the terminology of our models: the prefix `vanilla-MLP` indicates that we directly parameterize the map with an MLP (see Prob. (12)), while `struct-MLP` indicates that we use the parameterization trick (see Prob. (16)); we then add the employed regularizers to the name of the model.

6.5. Ablation Study.

We study the influence of each component of our models. In particular, we investigate the effect of each regularizer on real data to interpret the performance gains w.r.t. the baselines (especially ICNNs) shown in Figure 4.

Influence of $\mathcal{M}_\mu^2$ and $\mathcal{C}_\mu$. We assess the impact of $\mathcal{M}_\rho^2$ and $\mathcal{C}_\mu$ on the two following tasks: (i) learning the Monge map between the Korotin et al. (2021) benchmark pair of dimension $d = 32$ and (ii) predict the responses of cells populations to *abexinostat* drug with scRNAseq data. For each task, we report the performances induced by each pair of regularization weights $(\lambda_{\mathrm{MG}}, \lambda_{\mathrm{cons}})$ on a regular grid. Results are shown in Figure 6. First, on both tasks, the λ_{MG} regime providing the best performances is wide and corresponds to $\lambda_{\mathrm{MG}} \in [1, 10]$. This aligns with the discussion led in §4.1 and highlights that λ_{MG} is, by construction, easy to tune. However, the regularizers' influence on each task is very different. On the Korotin et al. (2021)'s benchmark, where the ground truth is explicitly known as the gradient of a convex potential, both $\mathcal{M}_\mu^2$ and $\mathcal{C}_\mu$ improve performance. However, on single-cell genomic data, performances only improve as we give greater importance to the Monge gap.

Ablation Study on Single-Cell Genomic Data. We conduct an ablation study of our models on single-cell genomic data. Results are shown in Figure 7. The performances with and without $\mathcal{C}_\mu$ are globally aligned. Therefore, it first shows that the performance gains visible in Figure 4 are obtained thanks to the Monge gap. These results thus differ from those obtained on the Korotin et al. (2021)'s benchmark, where $\mathcal{C}_\mu$ clearly helps since we explicitly target the gradient of a convex function. Therefore, while it helps with synthetic data, it is unclear whether leveraging the conservativity prior systematically helps with real data. This aspect explains, in particular, the poor performance of approaches parameterizing $T_\theta = \nabla f_\theta$ with f_θ ICNN on real data since they explicitly enforce the "gradient of a convex function" constraint. This is why we believe regularized approaches like ours make more sense in the finite sample regime and when Monge optimality is only a modeling assumption because we can adjust the weight given to each constraint. As a result, we hope the Monge gap and the learning paradigms we provide will facilitate the use of OT maps on real data.

Conclusion. In this paper, we provide a novel strategy to train optimal transport maps. Our approach is grounded on regularization rather than on constraints. We provide a regularizer, the Monge gap, with many favorable properties: lower-bounded by 0 and 0 when the property is observed, with a scale (as a difference between averaged distances) comparable to that of a fitting loss. That regularizer allows a more efficient trade-off to train maps that should be OT-like rather than exactly conforming to OT theory, so it facilitates the use of OT maps on real data. Furthermore, it adapts to any cost c but requires defining a reference measure ρ. An interesting direction lies in developing adaptive ways to define that measure, linking it to data measures of interest.

References

B. Amos. On amortizing convex conjugates for optimal transport. *arXiv preprint arXiv:2210.12153*, 2022.

B. Amos, L. Xu, and J. Z. Kolter. Input Convex Neural Networks. In *International Conference on Machine Learning (ICML)*, volume 34, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*. PMLR, 2017.

A. Asadulaev, A. Korotin, V. Egiazarian, and E. Burnaev. Neural optimal transport with general cost functionals, 2022. URL `https://arxiv.org/abs/2205.15403`.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation, 2020.

O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the vegan cookbook, 2017.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Y. Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305, 1987.

C. Bunne, S. G. Stark, G. Gut, J. S. del Castillo, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Ratsch. Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *bioRxiv*, 2021.

C. Bunne, A. Krause, and M. Cuturi. Supervised training of conditional monge maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

C. Bunne, L. Meng-Papaxanthos, A. Krause, and M. Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 25, 2022b.

C.-H. Chao, W.-F. Sun, B.-W. Cheng, and C.-Y. Lee. On investigating the conservative property of score-based generative models, 2023.

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

S. Cohen, B. Amos, and Y. Lipman. Riemannian convex potential maps. In *International Conference on Machine Learning*, pages 2028–2038. PMLR, 2021.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (9):1853–1865, 2016.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, pages 3733–3742, 2017.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

M. Cuturi, L. Meng-Papaxanthos, Y. Tian, C. Bunne, G. Davis, and O. Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022.

J. M. Danskin. *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, volume 5. Springer, 1967.

L. V. Eyring, D. Klein, G. Palla, S. Becker, P. Weiler, N. Kilbertus, and F. J. Theis. Modeling single-cell dynamics using unbalanced parameterized monge maps. *bioRxiv*, 2022. doi: 10.1101/2022.10.04.510766. URL `https://www.biorxiv.org/content/early/2022/10/05/2022.10.04.510766`.

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

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

M. Gazdieva, L. Rout, A. Korotin, A. Kravchenko, A. Filippov, and E. Burnaev. An optimal transport perspective on unpaired image super-resolution, 2022. URL `https://arxiv.org/abs/2202.01116`.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences, 2018. URL `https://arxiv.org/abs/1810.02733`.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample Complexity of Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

G. Gut, M. Herrmann, and L. Pelkmans. Multiplexed protein maps link subcellular organization to cellular state. *Science (New York, N.Y.)*, 361, 08 2018. doi: 10.1126/science.aar7042.

D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). 2016. doi: 10.48550/ARXIV.1606.08415. URL https://arxiv.org/abs/1606.08415.

N. J. Higham. Stable iterations for the matrix square root. 15(2):227–242, 1997. ISSN 1572-9265. doi: 10.1023/A:1019150005407. URL https://doi.org/10.1023/A:1019150005407.

M. Hutchinson. 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.

L. Kantorovich. On the transfer of masses (in Russian). In *Doklady Akademii Nauk*, volume 37, 1942.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

A. Korotin, V. Egiazarian, A. Asadulaev, A. Safin, and E. Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2020.

A. Korotin, L. Li, A. Genevay, J. Solomon, A. Filippov, and E. Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. 2021. doi: 10.48550/ARXIV.2106.01954. URL https://arxiv.org/abs/2106.01954.

A. Korotin, D. Selikhanovych, and E. Burnaev. Neural optimal transport. 2022. doi: 10.48550/ARXIV.2201.12220. URL https://arxiv.org/abs/2201.12220.

S. Lang. *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics. Springer New York. ISBN 978-0-387-98593-0. URL https://books.google.fr/books?id=AUL7sVhFZLkC.

J.-F. Le Gall. *Intégration, Probabilités et Processus Aléatoires*.

M. Lotfollahi, F. A. Wolf, and F. J. Theis. scGen predicts single-cell perturbation responses. 16 (8):715–721. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8. URL https://doi.org/10.1038/s41592-019-0494-8.

G. Lu, Z. Zhou, J. Shen, C. Chen, W. Zhang, and Y. Yu. Large-scale optimal transport via adversarial training with cycle-consistency, 2020.

F. Lübeck, C. Bunne, G. Gut, J. S. del Castillo, L. Pelkmans, and D. Alvarez-Melis. Neural unbalanced optimal transport via cycle-consistent semi-couplings. 2022. doi: 10.48550/ARXIV.2209.15621. URL https://arxiv.org/abs/2209.15621.

A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, volume 37, 2020.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning (ICML)*, 2015.

D. J. Rezende and S. Racanière. Implicit riemannian concave potential maps. *arXiv preprint arXiv:2110.01288*, 2021.

J. Richter-Powell, J. Lorraine, and B. Amos. Input convex gradient networks, 2021. URL https://arxiv.org/abs/2111.12187.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

F. Santambrogio. Optimal Transport for Applied Mathematicians. *Birkhäuser, NY*, 55(58-63):94, 2015.

S. Saremi. On approximating ∇f with neural networks, 2019. URL https://arxiv.org/abs/1910.12744.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4), 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-seq whole-transcriptome analysis of a single cell. 6(5), 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. URL https://doi.org/10.1038/nmeth.1315.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

F. Wolf, P. Angerer, and F. Theis. Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19, 02 2018. doi: 10.1186/s13059-017-1382-0.

Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha. On scalable and efficient computation of large scale optimal transport, 2019.

K. D. Yang and C. Uhler. Scalable unbalanced optimal transport using generative adversarial networks, 2019.

A. Proofs.

A.1. Proof of Proposition 3.2.

Let a measurable $T : \Omega \to \Omega$ and $\mathbf{x}_1, ..., \mathbf{x}_n \sim_{\text{i.i.d}} \rho$.

RHS. For $X \sim \rho$, since Ω is compact and c is continuous, $c(X, T(X))$ is bounded hence integrable, so $\frac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_i)) \to \int c(\mathbf{x}, T(\mathbf{x}))\,\mathrm{d}\rho(\mathbf{x})$ almost surely.

LHS. Since Ω is compact, T is bounded so for any bounded and continuous $f : \Omega \to \mathbb{R}$ and $X \sim \rho$, $f \circ T(X)$ is well defined and bounded so integrable. Afterwards, one can simply adapt the proof of the almost sure weak convergence of empirical measure based on the strong law of large numbers to show that, almost surely, $T\sharp\hat{\rho}_n \to T\sharp\rho$ weakly. See for instance (Le Gall, Theorem 10.4.1). Therefore, almost surely, both $\hat{\rho}_n \to \rho$ and $T\sharp\hat{\rho}_n \to T\sharp\rho$ weakly, so since c is continuous and Ω is compact, it almost surely holds that $W_c(\hat{\rho}_n, T\sharp\hat{\rho}_n) \to W_c(\rho, T\sharp\rho)$ (Santambrogio, 2015, Theorem 1.51).

A.2. Proof of 3.3.

Let T, μ, ν as described and suppose that $\mathcal{M}^c_\rho(T) = 0$. Then, $(\mathrm{Id}, T)\sharp\rho$ is an optimal coupling between ρ and $T\sharp\rho$. Since the cost c is continuous, $\mathrm{Spt}\left((\mathrm{Id}, T)\sharp\rho\right)$ is a c-cyclically monotone (c-CM) set by virtue of (Santambrogio, 2015, Theorem 1.38). Because $\mathrm{Spt}(\mu) \subset \mathrm{Spt}(\rho)$, one has $\mathrm{Spt}\left((\mathrm{Id}, T)\sharp\mu\right) \subset \mathrm{Spt}\left((\mathrm{Id}, T)\sharp\rho\right)$. Since the c-CM property is defined for sets, one has that $\mathrm{Spt}\left((\mathrm{Id}, T)\sharp\mu\right)$ is also c-CM. Moreover, since Ω is compact, c is uniformly continuous and bounded. Hence, cyclical monotonicity of its support implies that the coupling $(\mathrm{Id}, T)\sharp\mu$ is optimal between its marginals thanks to (Santambrogio, 2015, Theorem 1.49). Therefore, T is a c-OT map from μ to ν.

A.3. On the Positivity of $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}$.

Recall that

$$\begin{aligned}\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T) &:= \frac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_i)) - W_{c,\varepsilon}(\hat{\rho}_n, T\sharp\hat{\rho}_n)\\ &= \frac{1}{n}\sum_{i=1}^n c(\mathbf{x}_i, T(\mathbf{x}_i)) - \min_{\mathbf{P}\in U_n} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P})\end{aligned}$$

with $\mathbf{C} = [c(\mathbf{x}_i, T(\mathbf{x}_i)]_{1\le i,j\le n}$. For any coupling $\mathbf{P} \in U_n$, since $-\varepsilon H(\mathbf{P}) = -\varepsilon\sum_{i,j=1}^n \mathbf{P}_{ij}\log(\mathbf{P}_{ij}) < 0$, one has:

$$\langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon H(\mathbf{P}) < \langle \mathbf{P}, \mathbf{C}\rangle$$

As a result, applying minimization on both sides yields that $W_{c,\varepsilon}(\hat{\rho}_n, T\sharp\hat{\rho}_n) < W_{c,0}(\hat{\rho}_n, T\sharp\hat{\rho}_n)$, and therefore:

$$\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T) > \mathcal{M}^c_{\hat{\rho}_n,0}(T) = \mathcal{M}^c_{\hat{\rho}_n}(T) \geq 0.$$

A.4. Proof of 3.4.

We start by studying $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}$ since it can be reformulated as a matrix input function. Indeed, it only depends on T via its values on the support of $\hat{\rho}_n$, namely $\mathbf{x}_1, ..., \mathbf{x}_n$. Therefore, we write $\mathbf{t}_i := T(\mathbf{x}_i)$ and study:

$$r(\mathbf{T}) := \tfrac{1}{2n}\|\mathbf{X} - \mathbf{T}\|_F^2 - W_{\ell_2^2,\varepsilon}(\hat{\rho}_n, \rho_{\mathbf{T}}),$$

where $\mathbf{X}, \mathbf{T} \in \mathbb{R}^{n\times d}$ contain observations $\mathbf{x}_i$ and $\mathbf{t}_i$ respectively, stored as rows, and $\rho_{\mathbf{T}}$ is the discrete measure supported on the $\mathbf{t}_i$.

Since $\mathbf{C} = \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{t}_i\|_2^2\right]_{1\leq i,j\leq n}$. For any $\mathbf{P} \in U_n$:

$$\begin{aligned}
2\langle \mathbf{C}, \mathbf{P}\rangle &= \sum_{i,j=1}^{n} \mathbf{P}_{ij}\|\mathbf{x}_i - \mathbf{t}_i\|_2^2 \\
&= \sum_{i,j=1}^{n} \mathbf{P}_{ij}\|\mathbf{x}_i\|_2^2 + \sum_{i,j=1}^{n} \mathbf{P}_{ij}\|\mathbf{t}_j\|_2^2 - 2\sum_{i,j=1}^{n} \mathbf{P}_{ij}\langle \mathbf{x}_i, \mathbf{t}_j\rangle \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2 \sum_{j=1}^{n} \mathbf{P}_{ij} + \sum_{j=1}^{n} \|\mathbf{t}_j\|_2^2 \sum_{i=1}^{n} \mathbf{P}_{ij} - 2\sum_{i,j=1}^{n}\sum_{k=1}^{d} \mathbf{P}_{ij}\mathbf{X}_{ik}\mathbf{T}_{jk} \\
&= \tfrac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2 + \tfrac{1}{n}\sum_{j=1}^{n} \|\mathbf{t}_j\|_2^2 - 2\sum_{k=1}^{d}\sum_{j=1}^{n} \mathbf{T}_{jk}\sum_{i=1}^{n} \mathbf{P}_{ij}\mathbf{X}_{ik} \\
&= \tfrac{1}{n}\|\mathbf{X}\|_F^2 + \tfrac{1}{n}\|\mathbf{Y}\|_F^2 - 2\langle \mathbf{T}, \mathbf{P}^\top \mathbf{X}\rangle
\end{aligned}$$

Afterwards, we get:

$$\begin{aligned}
r(\mathbf{T}) &= \tfrac{1}{2n}\|\mathbf{X} - \mathbf{T}\|_F^2 - W_{\ell_2^2,\varepsilon}(\hat{\rho}_n, \rho_{\mathbf{T}}) \\
&= \tfrac{1}{2n}\|\mathbf{X} - \mathbf{T}\|_F^2 - \min_{\mathbf{P}\in U_n} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) \\
&= -\tfrac{1}{n}\langle \mathbf{T}, \mathbf{X}\rangle - \min_{\mathbf{P}\in U_n} -\langle \mathbf{T}, \mathbf{P}^\top \mathbf{X}\rangle - \varepsilon H(\mathbf{P}) \\
&= \max_{\mathbf{P}\in U_n} \langle \mathbf{T}, (\mathbf{P} - \tfrac{1}{n}I_n)^\top \mathbf{X}\rangle - \varepsilon H(\mathbf{P}) \\
&= \max_{\mathbf{P}\in U_n} r_{\mathbf{P}}(\mathbf{T})
\end{aligned}$$

where $r_{\mathbf{P}} : \mathbf{T} \mapsto \langle \mathbf{T}, (\mathbf{P} - \frac{1}{n}I_n)^\top \mathbf{X}\rangle - \varepsilon H(\mathbf{P})$.

Each $r_{\mathbf{P}}$ is affine, so r is convex, sub-additive and positively homogeneous as a maximum of affine functions. By construction, for all vector field T, $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}(T) = r(\mathbf{T})$; so $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}$ naturally enjoys these properties. Afterwards, since these properties are preserved under pointwise convergence, we extend the result to $\mathcal{M}^2_{\rho}$ using Proposition 3.2.

A.5. Proof of 3.5.

We first study the lipschitzness of $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}$ w.r.t. $\|\cdot\|_{L_2(\hat{\rho}_n)}$, which remains to study the lipschitzness of r w.r.t. $\|\cdot\|_F$. Then, to study the lipschitzness of r, we study the lipschitzness of each $r_{\mathbf{P}}$. As an affine map induced by the matrix $(\mathbf{P} - \frac{1}{n}I_n)^\top \mathbf{X}$, each $r_{\mathbf{P}}$ is $\|(\mathbf{P} - \frac{1}{n}I_n)^\top \mathbf{X}\|_{\text{op}}$-Lipschitz continuous. Moreover, one has:

$$\begin{aligned}
\|(\mathbf{P} - \tfrac{1}{n}I_n)^\top \mathbf{X}\|_{\text{op}} &\leq \|(\mathbf{P} - \tfrac{1}{n}I_n)^\top\|_{\text{op}}\|\mathbf{X}\|_{\text{op}} && (18) \\
&\leq (\|\mathbf{P}\|_{\text{op}} + \tfrac{1}{n}\|I_n\|_{\text{op}})\|\mathbf{X}\|_{\text{op}} && (19) \\
&= \tfrac{2}{n}\|\mathbf{X}\|_{\text{op}} && (20)
\end{aligned}$$

where 18 follows from the sub-multiplicativity of the operator norm, 19 from the triangular inequality and 20 from the fact that since $\mathbf{P} \in U_n$, $\|\mathbf{P}\|_{\text{op}} = \frac{1}{n}$. Indeed, one can write $\mathbf{P} = \frac{1}{n}\mathbf{Q}$ with $\mathbf{Q}$ a bi-stochastic matrix. Then, $\mathbf{Q}^\top\mathbf{Q}$ is also a bi-stochastic matrix, so $\lambda_{\max}(\mathbf{Q}^\top\mathbf{Q}) = 1$. Therefore, $\|\mathbf{Q}\|_{\text{op}} = 1$ and $\|\mathbf{P}\|_{\text{op}} = \frac{1}{n}$ by homogeneity.

Afterwards, each $r_{\mathbf{P}}$ is $\frac{2}{n}\|\mathbf{X}\|_{\text{op}}$-Lipschitz continuous, so is r. Indeed, for any $\mathbf{T}, \mathbf{T}' \in \mathbb{R}^{n\times d}$, one has:

$$\begin{aligned}
|r(\mathbf{T}) - r(\mathbf{T}')| &= |\max_{P\in U_n} r_{\mathbf{P}}(\mathbf{T}) - \max_{P\in U_n} r_{\mathbf{P}}(\mathbf{T}')| \\
&\leq \max_{P\in U_n} |r_{\mathbf{P}}(\mathbf{T}) - r_{\mathbf{P}}(\mathbf{T}')| && (21) \\
&\leq \tfrac{2}{n}\|\mathbf{X}\|_{\text{op}}\|\mathbf{T} - \mathbf{T}'\|_F
\end{aligned}$$

Now, let's reformulate the Lipschitz constant and deduce the Lipschitzness of $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}$ from the Lipschitzness of r. Reminding that $\mathbf{X}$ is the matrix containing the $\mathbf{x}_i$ as rows, one has:

$$\tfrac{1}{\sqrt{n}}\|\mathbf{X}\|_{\text{op}} = \tfrac{1}{\sqrt{n}}\left(\lambda_{\max}\left(\mathbf{X}^\top\mathbf{X}\right)\right)^{1/2} = \left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)\right)^{1/2} = \left(\lambda_{\max}\left(\mathbb{E}_{\hat{\rho}_n}[XX^\top]\right)\right)^{1/2} \tag{22}$$

Similarly, for two vector fields $T, T' : \Omega \to \Omega$, since $\mathbf{T}$ is the matrix containing the $\mathbf{t}_i = T(\mathbf{x}_i)$ as rows and similarly for $\mathbf{T}'$, one has:

$$\tfrac{1}{\sqrt{n}}\|\mathbf{T}-\mathbf{T}'\|_F = \left(\frac{1}{n}\sum_{i=1}^n \|T(\mathbf{x}_i) - T'(\mathbf{x}_i)\|_2^2\right)^{1/2} = \|T - T'\|_{L_2(\hat{\rho}_n)} \tag{23}$$

Therefore, combining Equations (22) and 23, Equation (21) can be reformulated as:

$$\begin{aligned} & |r(\mathbf{T}) - r(\mathbf{T}')| \leq \tfrac{2}{n}\|\mathbf{X}\|_{\text{op}}\|\mathbf{T}-\mathbf{T}'\|_F \\ \Leftrightarrow \quad & |\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}(T) - \mathcal{M}^2_{\hat{\rho}_n,\varepsilon}(T')| \leq \underbrace{2\left(\lambda_{\max}\left(\mathbb{E}_{\hat{\rho}_n}[XX^\top]\right)\right)^{1/2}}_{C(\hat{\rho}_n)}\|T-T'\|_{L_2(\hat{\rho}_n)} \end{aligned} \tag{24}$$

which proves the $C(\hat{\rho}_n)-$ Lipschitz continuity of of $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}$.

We now extend the result to $\mathcal{M}^2_\rho$. First, it almost surely holds that $\lim_{n\to+\infty}\|T-T'\|_{L_2(\hat{\rho}_n)} = \|T-T'\|_{L_2(\rho)}$. Then, since $\mathbf{A} \mapsto \lambda_{\max}(\mathbf{A})$ is a norm on $S_d^+(\mathbb{R})$, it is in particular continuous, so is $\mathbf{A} \mapsto (\lambda_{\max}(\mathbf{A}))^{1/2}$. Afterwards, since for each $n \geq 0$, $\mathbb{E}_{\hat{\rho}_n}[XX^\top] \in S_d^+(\mathbb{R})$, $\lim_{n\to+\infty}\left(\lambda_{\max}(\mathbb{E}_{\hat{\rho}_n}[XX^\top])\right)^{1/2} = \left(\lambda_{\max}(\mathbb{E}_{\rho}[XX^\top])\right)^{1/2}$ almost surely. Then, passing to the limit in Equation (24) with $\varepsilon = 0$ leads:

$$|\mathcal{M}^2_\rho(T) - \mathcal{M}^2_\rho(T')| \leq \underbrace{2\left(\lambda_{\max}(\mathbb{E}_\rho[XX^\top])\right)^{1/2}}_{C(\rho)}\|T-T'\|_{L_2(\rho)}$$

which proves the $C(\rho)-$ Lipschitz continuity of of $\mathcal{M}^2_\rho$.

B. Reminders and Details on the Conservative Regularizer.

In this section, we remind the definition of the conservative regularizer and detail its efficient estimation procedure based on the Hutchinson (1990) trace estimator.

Theoretical Motivations. First of all, let's remind the theoretical considerations behind the definition of this conservativity regularizer. By the Poincaré's lemma (Lang, Theorem 4.1, Chap. V), on a star-shaped domain $\Omega \subset \mathbb{R}^d$, any closed differential form is exact. Namely, any differentiable vector field whose Jacobian is symmetric on Ω is a gradient field, i.e. for any differentiable $F : \Omega \to \Omega$:

$$\forall \mathbf{x} \in \Omega,\ \text{Jac}_{\mathbf{x}}F = \text{Jac}_{\mathbf{x}}^\top F \ \Leftrightarrow\ \exists f : \mathbb{R}^d \to \mathbb{R},\ \text{s.t. } F = \nabla f\,.$$

Introducing a reference measure $\rho \in \mathcal{P}(\Omega)$ and considering a differentiable vector field $F : \Omega \to \Omega$, the regularizer hence penalizes the asymmetry of $\text{Jac}_{\mathbf{x}}F$ on the support of ρ:

$$\mathcal{C}_\rho(F) = \mathbb{E}_{X\sim\rho}\left[\|\text{Jac}_X F - \text{Jac}_X^T F\|_F^2\right]\,. \tag{25}$$

The regularizer $F \mapsto \mathcal{C}_\rho(F)$ is convex on differentiable vector fields. Indeed, for any $\mathbf{x} \in \text{Spt}(\rho)$, the functional $F \mapsto \|\text{Jac}_{\mathbf{x}}F - \text{Jac}_{\mathbf{x}}^\top F\|_2^2$ is convex as the composition of a linear operator and a convex function, so the convexity of $\mathcal{C}_\rho$ follows from linearity of the expectation.

Efficient Estimation. Similar to the Monge gap, we use an empirical estimator for $\mathcal{C}_{\hat{\rho}_n}(F)$. However, for large dimension d, computing the full Jacobian $\mathrm{Jac}_{\mathbf{x}}F$ might be too costly. We use instead the Hutchinson (1990) trace estimator, which turns pointwise Jacobians to pointwise Jacobian vector products (JVPs) and vector Jacobian products (VJPs). Indeed, for any $\mathbf{A} \in \mathbb{R}^{d\times d}$, using the fact that $\mathrm{I}_d = \mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[VV^\top]$, one has:

$$\begin{aligned}
\mathrm{Tr}(\mathbf{A}) &= \mathrm{Tr}(\mathbf{A}\,\mathrm{I}_d) \\
&= \mathrm{Tr}(\mathbf{A}\,\mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[VV^\top]) \\
&= \mathrm{Tr}(\mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[\mathbf{A}\,VV^\top]) \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[\mathrm{Tr}(\mathbf{A}\,VV^\top)] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[\mathrm{Tr}(V^\top\,\mathbf{A}\,V)] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}[V^\top\,\mathbf{A}\,V]
\end{aligned}$$

Therefore, applying the above formula inside the expectation defining the conservative regularizer, one has:

$$\begin{aligned}
\mathcal{C}_\rho(F) &= \mathbb{E}_{X\sim\rho}\left[\|\mathrm{Jac}_X F - \mathrm{Jac}_X^T F\|_F^2\right] \\
&= \mathbb{E}_{X\sim\rho}\left[\mathrm{Tr}\left((\mathrm{Jac}_X F - \mathrm{Jac}_X^T)^\top(\mathrm{Jac}_X F - \mathrm{Jac}_X^T)\right)\right] \\
&= \mathbb{E}_{X\sim\rho}\left[\mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}\left[V^\top(\mathrm{Jac}_X F - \mathrm{Jac}_X^T)^\top(\mathrm{Jac}_X F - \mathrm{Jac}_X^T)V\right]\right] \\
&= \mathbb{E}_{X\sim\rho}\left[\mathbb{E}_{V\sim\mathcal{N}(0,\mathrm{I}_d)}\left[\|\mathrm{Jac}_X FV - \mathrm{Jac}_X^T V\|_2^2\right]\right] \\
&= \mathbb{E}_{(X,V)\sim\rho\otimes\mathcal{N}(0,\mathrm{I}_d)}\left[\|\mathrm{Jac}_X FV - \mathrm{Jac}_X^T V\|_2^2\right]
\end{aligned}$$

Using samples $\mathbf{x}_1, ..., \mathbf{x}_n \sim_{\text{i.i.d}} \rho$ and $\mathbf{y}_1, ..., \mathbf{v}_m \sim_{\text{i.i.d}} \mathcal{N}(0, \mathrm{I}_d)$, its empirical counterpart translates to

$$\begin{aligned}
\mathcal{C}_{\hat{\rho}_n}(F) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathrm{Jac}_{\mathbf{x}_i}F\mathbf{v}_j - \mathrm{Jac}_{\mathbf{x}_i}F^\top\mathbf{v}_j\|_2^2 \\
&= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathtt{jvp}(F)(\mathbf{x}_i, \mathbf{v}_j) - \mathtt{vjp}(F)(\mathbf{x}_i, \mathbf{v}_j)\|_2^2\,,
\end{aligned} \tag{26}$$

where `jvp` and `vjp` denote respectively the Jacobian Vector Product and Vector Jacobian Product operators. Using the JAX framework (Bradbury et al., 2018), these operations can be carried out using the `jax.vjp` and `jax.jvp` primitives. Consequently, the Hutchinson (1990) trace estimator replaces the calculation of the full jacobians $\mathrm{Jac}_{\mathbf{x}_i}F$ and $\mathrm{Jac}_{\mathbf{x}_i}^\top F$ in each $\mathbf{x}_i$, by the calculation of m `jvp` and `vjp`. We need to choose $m \ll d$ to gain computational efficiency. Indeed, computing the full Jacobians $\mathrm{Jac}_{\mathbf{x}_i}F$ and $\mathrm{Jac}_{\mathbf{x}_i}F^\top$ requires the computation of respectively d JVPs and VJPs, instantiated along the vectors of the canonical basis of $\mathbb{R}^d$.

C. Additional Experiments.

C.1. Analysis of the Monge gap's Influence on Learning Dynamics.

In this section, we study the effect of the Monge gap on the learning dynamic induced by the optimization of Prob. (12). We remind it here, along with the goal we seek to achieve with each term of the loss:

$$\min_{\theta\in\mathbb{R}^p}\mathcal{L}(\theta) := \underbrace{\Delta(T_\theta\sharp\mu, \nu)}_{\text{fitting}} + \underbrace{\lambda_{\mathrm{MG}}\,\mathcal{M}_\rho^c(T_\theta)}_{\text{c-optimality}}\,.$$

Assume from now that c and Δ are differentiable and let $\varepsilon > 0$. The above optimization problem can be solved by sampling batches $\hat{\mu}_n, \hat{\nu}_n, \hat{\rho}_n$, and considering stochastic gradients, i.e. gradients of the estimated loss:

$$\hat{\mathcal{L}}_n(\theta) := \Delta(T_\theta\sharp\hat{\mu}_n, \hat{\nu}_n) + \lambda_{\mathrm{MG}}\,\mathcal{M}_{\hat{\rho}_n,\varepsilon}^c(T_\theta)$$

Monge gap's Influence on the Gradient Flow of $\mathcal{L}$. The effect of the fitting loss Δ is clear: using gradient steps with $\nabla_\theta \Delta(T_\theta \sharp \hat{\mu}_n, \hat{\nu}_n)$ will push the mapped source measure $T_\theta \sharp \mu$ to be as close as possible to the target measure ν. Then, to better understand the effect of the Monge gap, we take a closer look at the gradient of the entropic empirical Monge gap: $\nabla_\theta \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$. Since $\varepsilon > 0$, the optimal transport plan $\mathbf{P}^\varepsilon$ between $\hat{\rho}_n$ and $T_\theta \sharp \hat{\rho}_n$ is unique, so $W_{c,\varepsilon}(\hat{\rho}_n, T_\theta \sharp \hat{\rho}_n)$ is differentiable in its inputs thanks to the Danskin (1967) (see § 2). Especially, one can differentiate everywhere w.r.t. θ:

$$\nabla_\theta W_{c,\varepsilon}(\hat{\rho}_n, T_\theta \sharp \hat{\rho}_n) = \sum_{i,j=1}^{n} \mathbf{P}^\varepsilon_{ij} \nabla_\theta c(\mathbf{x}_i, T_\theta(\mathbf{x}_j))$$

Afterwards, $\theta \mapsto \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$ is differentiable and its gradient reads:

$$\nabla_\theta \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta) = \sum_{i,j=1}^{n} \left(\tfrac{1}{n}\delta_{ij} - \mathbf{P}^\varepsilon_{ij}\right) \nabla_\theta c(\mathbf{x}_i, T_\theta(\mathbf{x}_j))$$

One can notice that the magnitude of the gradient increases as $\mathbf{P}^\varepsilon$ deviates from the identity coupling $\frac{1}{n}\mathrm{I}_n$ which sends each $\mathbf{x}_i$ to $T_\theta(\mathbf{x}_i)$. More precisely, since $\mathbf{P}^\varepsilon \in U_n, \forall i,j,\ 0 \leq \mathbf{P}^\varepsilon_{ij} \leq 1/n$, so:

$$\begin{cases} (1/n)\delta_{ij} - \mathbf{P}^\varepsilon_{ij} \geq 0 & \text{if} \quad i = j \\ (1/n)\delta_{ij} - \mathbf{P}^\varepsilon_{ij} \leq 0 & \text{if} \quad i \neq j \end{cases}$$

Using gradient steps with $\nabla_\theta \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$ will therefore drive the $T_\theta(\mathbf{x}_i)$ to make $\mathbf{P}^\varepsilon$ as close as possible to the identity coupling by: decreasing the cost on the diagonal $c(\mathbf{x}_i, T_\theta(\mathbf{x}_i))$ while increasing the cost off the diagonal $c(\mathbf{x}_i, T_\theta(\mathbf{x}_j))$, $i \neq j$. We will therefore aim for the permutation giving the optimal assignment between $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\{T_\theta(\mathbf{x}_1), ..., T_\theta(\mathbf{x}_n)\}$ for cost c to be the identity. This is equivalent to reaching the cyclical monotonicity of the set $\{\mathbf{x}_1, ..., \mathbf{x}_n\} \times \{T_\theta(\mathbf{x}_1), ..., T_\theta(\mathbf{x}_n)\}$, which is in line with the discussion in § 3.2.

An experiment showing this dynamic on synthetic data in dimension $d = 2$ is provided in Figure 8. We use $\Delta = W_{\ell_2^2,\varepsilon}$ and the Monge gap instantiated with the cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ and $\rho = \mu$. We observe the effect of the Monge gap on the fitting: as the optimization proceeds, the assignment induced by T_θ tends to respect the Monge Mather shortening principle to get an assignment without crossing lines since c is a distance.

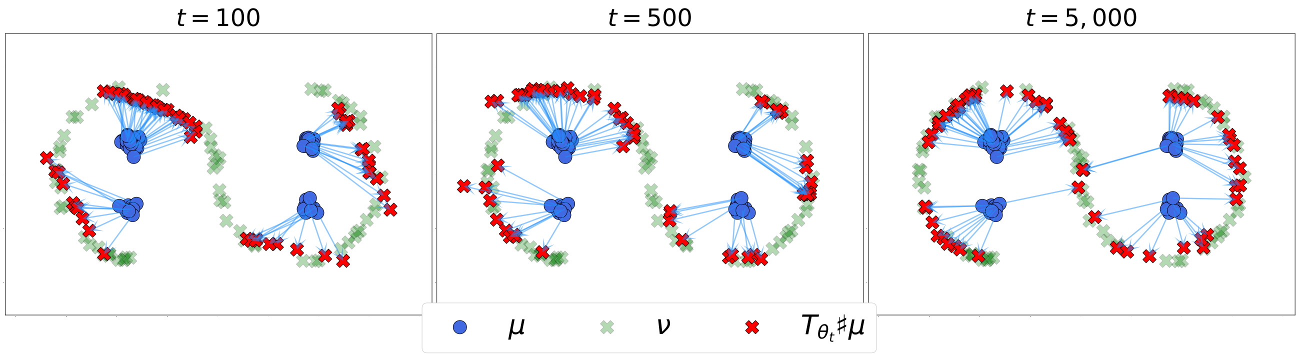


Figure 8. Transport map $(T_{\theta_t})_{t\geq 0}$ along the gradient flow of the loss $\mathcal{L}(\theta) = W_{\ell_2^2,\varepsilon}(T_\theta \sharp \mu, \nu) + \lambda_{\mathrm{MG}} \mathcal{M}^1_\mu(T_\theta)$ to fit an optimal map for cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ between two synthetic measures μ and ν. We use $\lambda_{\mathrm{MG}} = 1$. T_θ is directly parametrized as an MLP. We report three timestamps of the optimization at iterations 100, 500, and 10, 000.

Fitting With and Without Monge gap. We now investigate the evolution of $\Delta(T_\theta \sharp \hat{\mu}_n)$ and $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$ throughout the iterations, in the cases where we solve Prob. (12) with and without using the Monge gap, i.e. with $\lambda_{\mathrm{MG}} > 0$ and $\lambda_{\mathrm{MG}} = 0$. To study this dynamic, we plot the evolution of $\Delta(T_\theta \sharp \hat{\mu}_n)$ and $\mathcal{M}_{\hat{\rho}_n,\varepsilon}(T_\theta)$ when fitting, with and without Monge gap, a map between the high-dimensional Korotin et al. (2021)'s benchmark pair of dimension $d = 128$. We remind that the Monge map $T^\star$, for the quadratic cost, between these two measures is known, so when using the Monge gap, we then instantiate it with the cost $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ to recover this Monge map. Moreover, we use $\lambda_{\text{MG}} = 1$, $\rho = \mu$ and $\Delta = W_{\ell_2^2,\varepsilon}$. To focus on the effect of the Monge gap, we use the model: vanilla-MLP+$\mathcal{M}_\rho^2$, i.e., we parameterize the map directly with an MLP and don't use the conservative regularizer. Since $T^\star$ is known in that setting, we also report unexplained variance $\mathcal{L}_2^{\text{UV}}(T_\theta)$ (see Eq. (17)), to quantify the deviation to $T^\star$ during training in both cases. Finally, we compute the metrics on $8,192$ unseen samples and average them across five runs.

The results are shown in Figure 9. When we fit with the Monge gap, both $\Delta(T_\theta \sharp \hat{\mu}_n)$ and $\mathcal{M}_{\hat{\rho}_n,\varepsilon}(T_\theta)$ are close to 0 at the end of training, which is in line with the discussion led in § 4.1. Moreover, $\mathcal{L}_2^{\text{UV}}(T_\theta)$ is also close to 0, so we effectively fit the Monge map $T^\star$. However, when we fit without the Monge gap, $\Delta(T_\theta \sharp \hat{\mu}_n)$ is close to 0 but both $\mathcal{M}_{\hat{\rho}_n,\varepsilon}(T_\theta)$ and $\mathcal{L}_2^{\text{UV}}(T_\theta)$ are high. The growth of $\mathcal{M}_{\hat{\rho}_n,\varepsilon}(T_\theta)$ and $\mathcal{L}_2^{\text{UV}}(T_\theta)$ shows that when we train without Monge gap, although the fitting loss is decreasing, $\mathcal{L}_2^{\text{UV}}(T_\theta)$ increases throughout iterations. We then fit an arbitrary push-forward between the source and the target measure, that has no reason to be c-optimal.

Note that in both cases, the Monge gap is small at initialization. This is because we initialize the neural map with the affine transport between the Gaussian approximations of the source and the target measures (see § D.2), which is an optimal map, thus having a small Monge gap. Indeed, at initialization, we have $T_{\theta_0} \approx T_{\mathbf{A},\mathbf{b}} : \mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \in S_d^+(\mathbb{R})$. Thus $T_{\theta_0} \approx \nabla \phi_{\mathbf{A},\mathbf{b}}$ where $\phi_{\mathbf{A},\mathbf{b}} : \mathbf{x} \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ is convex because $\mathbf{A} \in S_d^+(\mathbb{R})$ so $\mathcal{M}^2_{\hat{\rho}_n,\varepsilon}(T_{\theta_0}) \approx 0$. Furthermore, the Monge gap is not strictly equal to 0 at the end of the training because of the addition of the entropic regularization.

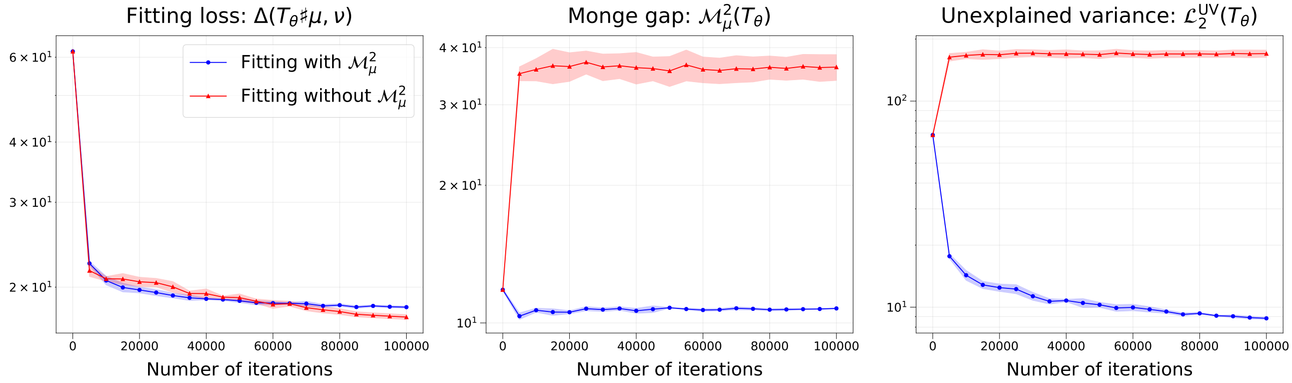


Figure 9. Learning dynamics when fitting a map, with and without Monge gap, between the high-dimensional Korotin et al. (2021)'s benchmark pair μ, ν of dimension $d = 128$. More precisely, we optimize the loss $\mathcal{L}(\theta) = W_{\ell_2^2,\varepsilon}(T_\theta \sharp \mu, \nu) + \lambda_{\text{MG}} \mathcal{M}_\mu^1(T_\theta)$ with $\lambda_{\text{MG}} = 0$ and $\lambda_{\text{MG}} = 1$ separately. In both cases, we report the values of $\Delta(T_\theta \sharp \hat{\mu}_n)$, $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$ and $\mathcal{L}_2^{\text{UV}}(T_\theta)$ throughout iterations. We compute the metrics on $8,192$ unseen samples and average them across five runs.

C.2. Preliminary Analysis of the Reference Measure's Influence.

In this section, we provide a preliminary analysis of the influence of the choice of reference measure ρ when we seek to fit a c-OT map between two measures μ, ν. Recall that by virtue of Prop. 3.3, we can choose any reference measure ρ such that $\text{Spt}(\rho) \supset \text{Spt}(\mu)$. We then study the evolution of performances when ρ is chosen more or less independently of μ while retaining the constraint $\text{Spt}(\rho) \supset \text{Spt}(\mu)$, on both synthetic data and single-cell genomic data.

Synthetic Data. We first test the influence of ρ in a simple case, for synthetic measures in dimension $d = 2$. We fit an OT map $\hat{T}$ for the cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$, using the Monge gap instantiated for this cost $\mathcal{M}_\rho^1$. We test several ρ verifying $\text{Spt}(\rho) \supset \text{Spt}(\mu)$ of different variance and shape, including $\rho = \mu$. For each fitting, we use $\Delta = W_{\ell_2^2,\varepsilon}$ and $\lambda_{\text{MG}} = 1$. We measure performances using $S_{\ell_2^2,\varepsilon}(\hat{T} \sharp \mu_{\text{test}}, \nu_{\text{test}})$ the Sinkhorn divergence between a batch of unseen source samples mapped by the fitted map and a batch of unseen target samples. Results are shown in Figure 10. We obtain similar performances for this simple low-dimensional task by choosing different reasonable ρ, more or less close to μ.

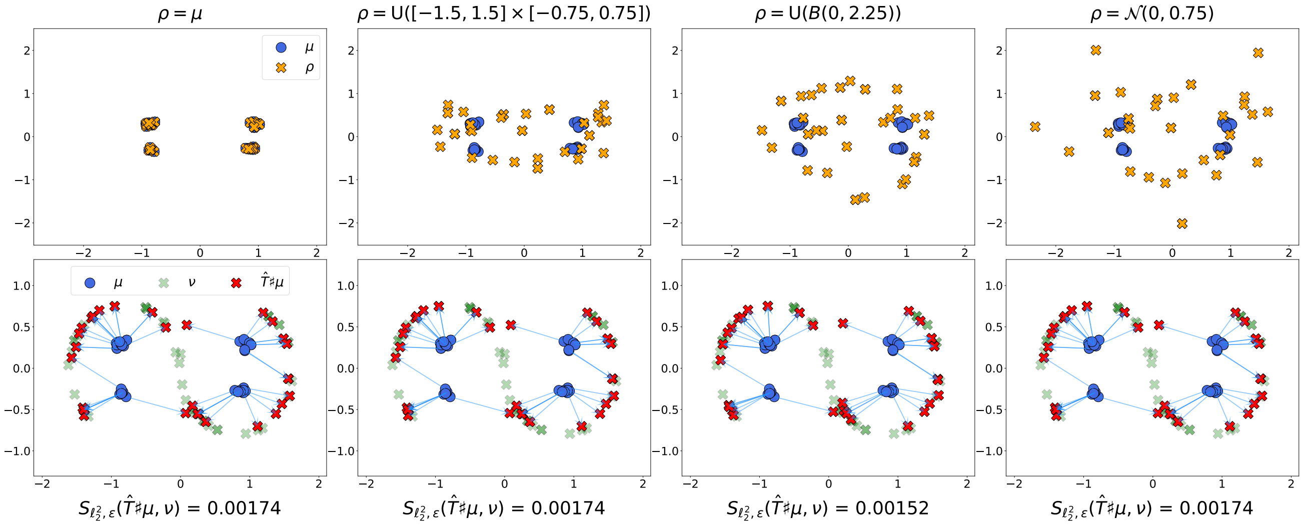


Figure 10. Influence of the reference measure ρ when fitting an OT map between two synthetic measures μ, ν for the cost $c(\mathbf{x}, \mathbf{y}) = \||\mathbf{x} - \mathbf{y}\|_2$, using the Monge gap induced by that cost $\mathcal{M}_\rho^1$. We estimate a map $\hat{T}$ for four different reference measures ρ satisfying $\mathrm{Spt}(\rho) \supset \mathrm{Spt}(\mu)$ and compare the results. The reference measure we test are: (i) $\rho = \mu$, (ii) $\rho = \mathrm{U}([-1.5, 1.5] \times [-0.75, 0.75])$, (iii) $\rho = \mathrm{U}(B(0, 2.25))$ and (iv) $\rho = \mathcal{N}(0, 0.75)$. For each fitting, we use $W_{\ell_2^2,\varepsilon}$ as the fitting loss and $\lambda_{\mathrm{MG}} = 1$, and we plot the Sinkhorn divergence $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu, \nu)$ computed on $8,192$ unseen samples.

Single-Cell Genomic Data. We then analyze the reference measure's influence on transport map fitting to predict the responses of cell populations to cancer treatments on the 4i and scRNA datasets considered in § 6.4, of respective dimension $d = 42$ and $d = 50$. We test $\rho = \mu$ and three other reference measures, defined in a more or less adaptive way w.r.t. μ. In this case of real, high-dimensional data, we use reference measures built from estimates of the source measure moments. For each dataset, we start by sampling a batch $\hat{\mu}_n$ of $n = 2,048$ samples from the source measure and compute $m_{\hat{\mu}_n}, \Sigma_{\hat{\mu}_n}$; the empirical mean and covariance. We then test ρ to be a Gaussian centered in the empirical mean $m_{\hat{\mu}_n}$ and whose covariance matrix is:

(i) the identity, i.e. $\rho_{\mathrm{stand}} = \mathcal{N}(m_{\hat{\mu}_n}, \mathrm{I}_d)$;

(ii) the diagonal of the empirical covariance, i.e. $\rho_{\mathrm{diag}} = \mathcal{N}(m_{\hat{\mu}_n}, \mathrm{diag}(\Sigma_{\hat{\mu}_n}))$;

(iii) the full empirical covariance, i.e. $\rho_{\mathrm{full}} = \mathcal{N}(m_{\hat{\mu}_n}, \Sigma_{\hat{\mu}_n})$.

Note that for all datasets, $\mathrm{diag}(\Sigma_{\hat{\mu}_n}) < \mathbf{1}_d$, so ρ_{stand} is the " largest " reference measure. We compare the predictions each reference measure provides for each profiling technology and treatment. To focus on the influence of the reference measure on the Monge gap, we use the model: `vanilla-MLP`+$\mathcal{M}_\rho^2$, i.e., we parameterize the map directly with an MLP and don't use the conservative regularizer. In all cases, we use $\Delta = W_{\ell_2^2,\varepsilon}$ and $\lambda_{\mathrm{MG}} = 1$. We measure predictive performance using the Sinkhorn divergence between a batch of unseen (test) treated cells and a batch of unseen control cells mapped with $\hat{T}$, namely $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu_{\mathrm{test}}, \nu_{\mathrm{test}})$ and average the results over five runs.

The results are shown in Figure 11. On 4i data, we observe the expected dynamic: the more the reference measure is chosen adaptively to the source measure μ, the better the performance. More precisely, we obtain the best performances with $\rho = \mu$, then ρ_{full}, ρ_{diag} and ρ_{stand}. On scRNAseq data, $\rho = \mu$ gives the best performance, then ρ_{full}, ρ_{diag} and ρ_{stand} give similar performances since $\Sigma_{\hat{\mu}_n}$ is closer to I_d in this case, so these three reference measures candidates are close.

This experiment shows that the reference measure might influence performances on real and high-dimensional data. Moreover, the choice $\rho = \mu$ seems to systematically give the best performances. On the other hand, as we observe that the more ρ is chosen adaptively to μ, the better the performances, an exciting line of work would be to dynamically adapt ρ, as a function of μ, during training, linking it to data measures of interest.

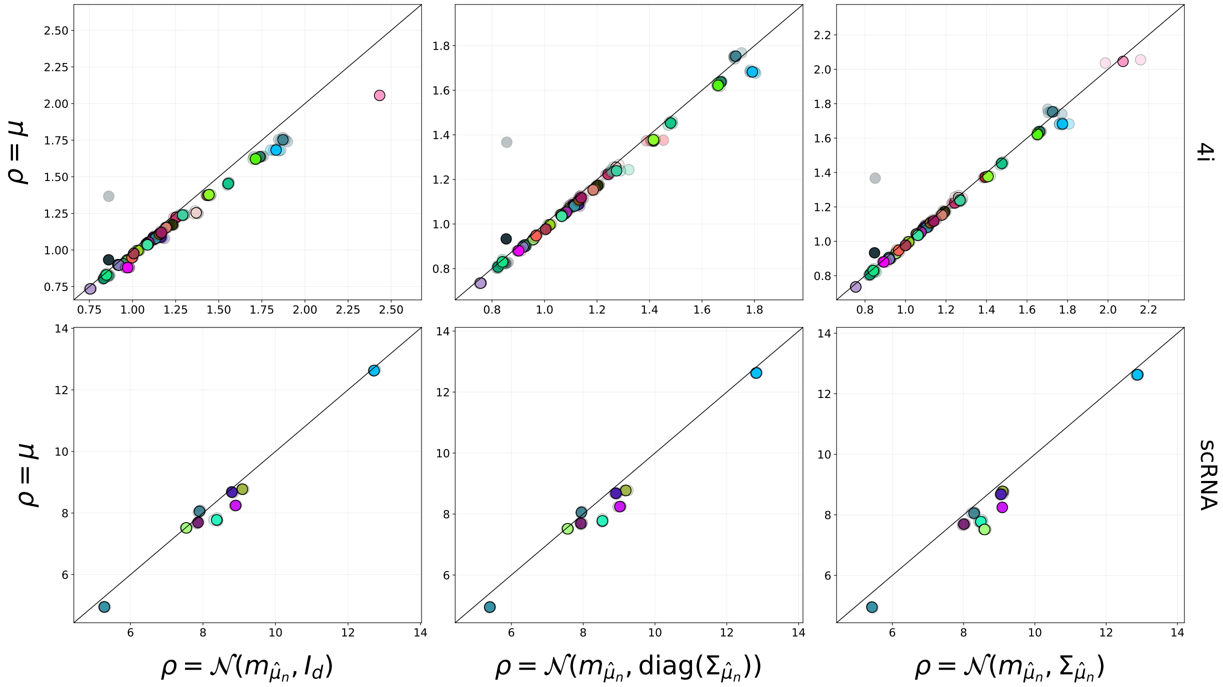


Figure 11. Fitting of a transport map $\hat{T}$ to predict the responses of cell populations to cancer treatments on 4i (upper plot) and scRNA (lower plot) datasets, with several reference measures. We test $\rho = \mu$ and three other reference measures chosen more or less adaptively to μ. For each profiling technology and each treatment, we compare the predictions provided by the model `vanilla-MLP`+$\mathcal{M}_\rho^2$ with each reference measure candidate ρ. We measure predictive performance using the Sinkhorn divergence between a batch of unseen (test) treated cells and a batch of unseen control cells mapped with $\hat{T}$, namely $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu_{\text{test}}, \nu_{\text{test}})$. The plot structure is similar to Figure Figure 4. We refer the reader to its caption for clear instructions on how to read it.

C.3. Additional Experiments on the 2-Sphere.

To show that the Monge gap can be used to fit an OT map for any cost on any domain, we use it for synthetic measures supported on $\Omega = \mathbb{R}^3_{++} \cap \mathbb{S}^2$. On this domain, we use $c(\mathbf{x}, \mathbf{y}) = -\log(\mathbf{x}^\top \mathbf{y})$. We can verify that this cost is a distance on Ω. While positivity, symmetry, and separation are clear, let's show the triangular inequality.

First, let us remark that for any $\mathbf{a}, \mathbf{b} \in \Omega$, since $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$, one has $\mathbf{a}^\top \mathbf{b} = 1 - \frac{1}{2}\|\mathbf{a} - \mathbf{b}\|_2^2$. Therefore:

$$\begin{aligned}
&\mathbf{x}^\top\mathbf{y} + \mathbf{x}^\top\mathbf{z} - (\mathbf{x}^\top\mathbf{y})(\mathbf{x}^\top\mathbf{z}) \\
=&\, 1 - \tfrac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + 1 - \tfrac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 - (1 - \tfrac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2)(1 - \tfrac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2) \\
=&\, 1 - \tfrac{1}{4}\|\mathbf{x} - \mathbf{y}\|_2^2\|\mathbf{x} - \mathbf{z}\|_2^2 \\
\geq&\, 0
\end{aligned}$$

The last inequality follows from $\|\mathbf{x} - \mathbf{y}\|_2^2 = 2(1 - \mathbf{x}^\top\mathbf{y}) \leq 2$ since $0 \leq \mathbf{x}^\top\mathbf{y} \leq 1$ because $\mathbf{x}, \mathbf{y} \in \Omega$ and similarly, $\|\mathbf{x} - \mathbf{z}\|_2^2 \leq 2$. Therefore, one has:

$$\begin{aligned}
& \mathbf{x}^\top\mathbf{y} + \mathbf{x}^\top\mathbf{z} \geq (\mathbf{x}^\top\mathbf{y})(\mathbf{x}^\top\mathbf{z}) \\
\Leftrightarrow\quad & -\log(\mathbf{x}^\top\mathbf{y} + \mathbf{x}^\top\mathbf{z}) \leq -\log(\mathbf{x}^\top\mathbf{y}) - \log(\mathbf{x}^\top\mathbf{z}) \\
\Leftrightarrow\quad & c(\mathbf{x}, \mathbf{y} + \mathbf{z}) \leq c(\mathbf{x}, \mathbf{y}) + c(\mathbf{x}, \mathbf{z})
\end{aligned}$$

The results are shown in Figure 12. The difference between fitting with and without the Monge gap is visually clear. Since c is a distance, the OT map should exhibit the Monge Mather shortening principle, so we should observe an assignment without crossing lines. This is the case for the map fitted with the Monge gap. However, when we fit without the Monge gap, the unique mode of the source measure is split into two parts, and each part is sent to the farthest of the two modes of the target measure. Therefore, we do observe crossing lines.

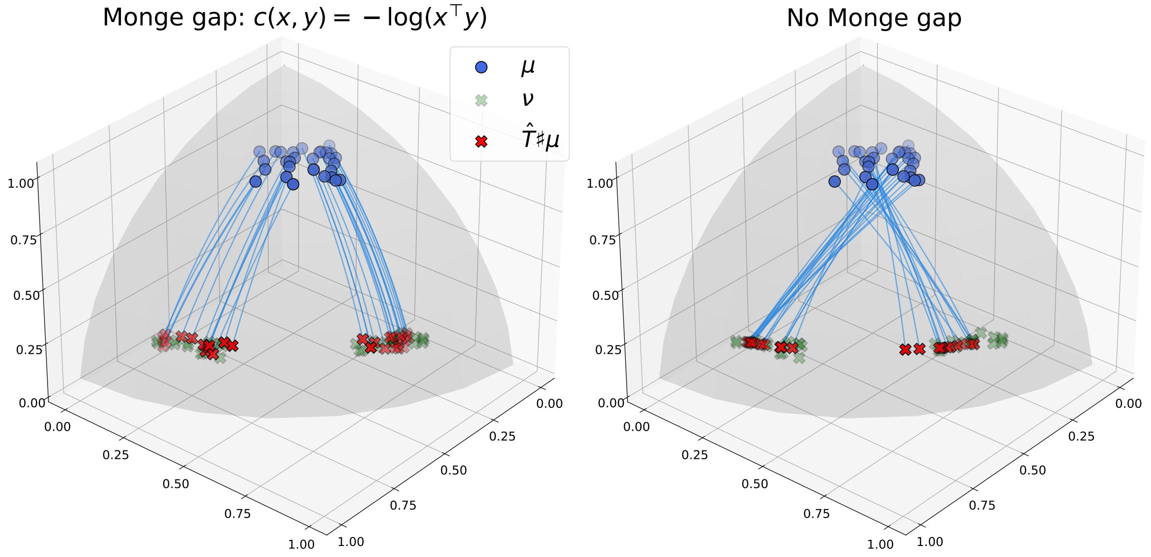


Figure 12. Fitting of 2 transport map between synthetic measures supported on $\mathbb{R}^3_{++} \cap \mathbb{S}^2$. For the left figure, we use the Monge gap instantiated cost $c(\mathbf{x}, \mathbf{y}) = -\log(\mathbf{x}^\top \mathbf{y})$ along with $\lambda_{\mathrm{MG}} = 1$ while we do not use regularizer for the left figure. In both cases, we parameterize the map as $T_\theta = F_\theta / \|F_\theta\|_2$ where F_θ is an MLP and use $W_{\ell_2^2, \varepsilon}$ as fitting loss.

D. Numerical Details.

D.1. Using the Monge gap in practice.

The monge_gap function is implemented in the OTT-JAX (Cuturi et al., 2022) package. We also provide a MapEstimator solver to fit c-OT maps, for an arbitrary cost function c. The user simply needs to instantiate the solver with a fitting loss Δ and a cost function c, then it approximates a c-OT map.

D.2. Initializer Schemes.

Let $F_\theta : \mathbb{R}^d \to \mathbb{R}^d$, $\theta \in \mathbb{R}^p$, an MLP. For any affine map $T_{\mathbf{A},\mathbf{b}} : \mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$, it is simple to choose θ_0 such that $F_{\theta_0} \approx T_{\mathbf{A},\mathbf{b}}$. One can initialize the feedforward weights randomly with relatively low variance and add a residual layer from the input layer to the output layer with parameters $(\mathbf{A}, \mathbf{b})$. This approach is described in Figure D.2.

- IdentityInit. For generic costs, we directly parameterize T_θ as an MLP, so we initialize with a residual layer parameterizing the identity. For structured costs $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$, since we parameterize $T_\theta = \mathrm{I}_d - \nabla h^* \circ F_\theta$ with F_θ an MLP, one typically has that for any $\mathbf{x}_0 \in \mathbb{R}^d$ close to 0, $\nabla h^*(\mathbf{x}_0) \approx 0$. Thus, in this case, we don't need to use a residual layer but initializing the feedforward weights randomly with a low variance provides $F_{\theta_0} \approx 0$ so $T_{\theta_0} = \mathrm{I}_d - \nabla h^\star \circ F_{\theta_0} \approx \mathrm{I}_d$.

- GaussianInit. This initializer uses the closed form of the OT map between Gaussian measures for the quadratic cost, which is affine. We denote $\hat{T}_{\mathcal{N}}$ the affine OT map between the Gaussian approximations of μ and ν. First, we estimate $\hat{T}_{\mathcal{N}}$ from samples, forming empirical means and covariances $(m_{\hat{\mu}_n}, \Sigma_{\hat{\mu}_n})$ and $(m_{\hat{\nu}_n}, \Sigma_{\hat{\nu}_n})$:

$$\hat{T}_{\mathcal{N}} : \mathbf{x} \mapsto \Sigma_{\hat{\mu}_n}^{-1/2} \left(\Sigma_{\hat{\mu}_n}^{1/2} \Sigma_{\hat{\nu}_n} \Sigma_{\hat{\mu}_n}^{1/2} \right)^{1/2} \Sigma_{\hat{\mu}_n}^{-1/2} (\mathbf{x} - m_{\hat{\mu}_n}) + m_{\hat{\nu}_n} \,. \tag{27}$$

  Square roots and inverse square roots of PSD matrices are computed with the OTT-JAX (Cuturi et al., 2022) implementation of the Higham (1997) algorithm. If we parameterize the map directly with an MLP, we then initialize $T_{\theta_0} \approx \hat{T}_{\mathcal{N}}$. If we use the parameterization trick, since we use the quadratic cost, one has $T_\theta = \mathrm{I}_d - F_\theta$ because $\nabla h^* = \mathrm{I}_d$. Then, we initialize $F_{\theta_0} \approx \mathrm{I}_d - \hat{T}_{\mathcal{N}}$, so that $T_{\theta_0} \approx \hat{T}_{\mathcal{N}}$.

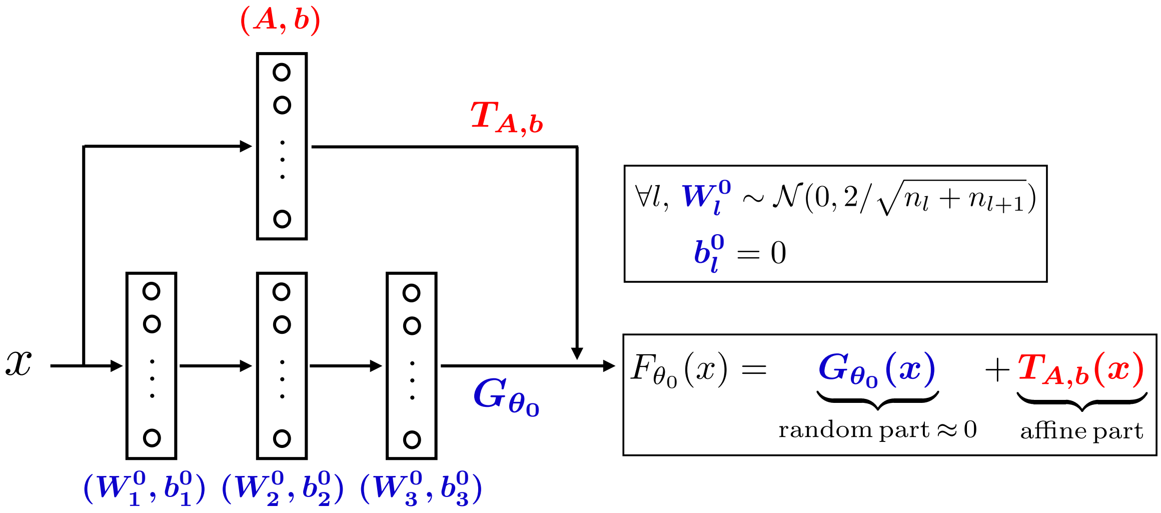

(A, b)
$T_{A,b}$
$\forall l,\ W_l^0 \sim \mathcal{N}(0, 2/\sqrt{n_l + n_{l+1}})$
$b_l^0 = 0$
x
G_{θ_0}
$F_{\theta_0}(x) = \underbrace{G_{\theta_0}(x)}_{\text{random part} \approx 0} + \underbrace{T_{A,b}(x)}_{\text{affine part}}$
(W_1^0, b_1^0)
(W_2^0, b_2^0)
(W_3^0, b_3^0)

Figure 13. Initialization scheme to match affine maps applied to a 3 hidden layers MLP. We initialize the feedforward weights using the Glorot and Bengio (2010) initialization technique and add a residual layer matching the targeted affine map.

D.3. Algorithms.

In this section, we provide the algorithms corresponding to each proposed method.

- The `algorithm` 1 corresponds to the `vanilla-MLP+`$\mathcal{M}_\rho^c$ method detailed in § 4.1, that works for generic costs. It solves Prob. (12).
- The `algorithm` 2 corresponds to the `struct-MLP+`$\mathcal{M}_\rho^c + \mathcal{C}_\rho$ method detailed in § 4.2, that works for strictly convex costs inducing structure in the OT map, and using the conservative regularizer. It solves Prob. (16).

Algorithm 1 The `vanilla-MLP+`$\mathcal{M}_\rho^c$ method for OT map estimation with generic costs.

```
Data. Source $\mu$, target $\nu$ and reference measure $\rho$; cost function $c$; regularization weight $\lambda_{\mathrm{MG}}$; entropic regularization strength $\varepsilon$; learning rate $\eta$; batch size $n$; number of iterations $K_{\mathrm{iters}}$; and an MLP $T_\theta$ to model the OT map.
Results. Estimated OT map $\hat{T}_\theta$.
Initialization. $T_{\theta_0} \leftarrow$ GaussianInit if $c = \frac{1}{2}\|\cdot - \cdot\|_2^2$ else IdentityInit (see § D.2).
For $k = 1, ..., K_{\mathrm{iters}}$ do:
    • Sample batches $\hat{\mu}_n$, $\hat{\nu}_n$, $\hat{\rho}_n$.
    • Compute $\Delta(T_\theta \sharp \hat{\mu}_n, \hat{\nu}_n)$.
    • Compute $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$ by running Sinkhorn$(\hat{\rho}_n, T_\theta \sharp \hat{\rho}_n, c, \varepsilon)$ (see Eq. (8)).
    • $\hat{\mathcal{L}}_n(\theta) \leftarrow \Delta(T_\theta \sharp \hat{\mu}_n, \hat{\nu}_n) + \lambda_{\mathrm{MG}}\, \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta)$
    • Update $\theta$ to minimize $\hat{\mathcal{L}}_n(\theta)$.
```

D.4. Fixed hyperparameters across experiments.

Entropic regularization. Whenever we run the Sinkhorn algorithm on a cost matrix $\mathbf{C}$, we set $\varepsilon = 0.01 \cdot \mathrm{mean}(\mathbf{C})$. The only case where we use a different ε value is for evaluation, when we compute the Sinkhorn divergence $S_{\ell_2^2,\varepsilon}$, for which we set $\varepsilon = 0.1$ across all experiments. We use the `OTT-JAX` (Cuturi et al., 2022) implementation of the Sinkhorn algorithm.

Algorithm 2 The `struct-MLP`+$\mathcal{M}_\rho^c + \mathcal{C}_\rho$ method for OT map estimation with strictly convex costs.

```
Data. Source $\mu$, target $\nu$ and reference measure $\rho$; cost function $c(\cdot,\cdot) = h(\cdot - \cdot)$ with $h$ strictly convex; regularization weights $\lambda_{\mathrm{MG}}$, $\lambda_{\mathrm{cons}}$; entropic regularization strength $\varepsilon$; number of Hutchinson vectors $m$; learning rate $\eta$; batch size $n$; number of iterations $K_{\mathrm{iters}}$; and an MLP $F_\theta$ to model the dual potential gradient, s.t. $T_\theta = \mathrm{I}_d - \nabla h^* \circ F_\theta$.
Results. Estimated OT map $\hat{T}_\theta = \mathrm{I}_d - \nabla h^* \circ \hat{F}_\theta$.
Initialization. Initialize $F_{\theta_0}$ s.t. $T_{\theta_0} \leftarrow$ GaussianInit if $c = \frac{1}{2}\|\cdot - \cdot\|_2^2$ else IdentityInit (see § D.2).
For $k = 1, ..., K_{\mathrm{iters}}$ do:
    • Sample batches $\hat{\mu}_n$, $\hat{\nu}_n$, $\hat{\rho}_n$ and Hutchison vectors $\mathbf{v}_1, ..., \mathbf{v}_m \sim \mathcal{N}(0, \mathrm{I}_d)$.
    • Compute $\Delta((\mathrm{I}_d - \nabla h^* \circ F_\theta)\sharp\hat{\mu}_n, \hat{\nu}_n)$.
    • Compute $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(\mathrm{I}_d - \nabla h^* \circ F_\theta)$ by running Sinkhorn$(\hat{\rho}_n, (\mathrm{I}_d - \nabla h^* \circ F_\theta)\sharp\hat{\rho}_n, c, \varepsilon)$ (see Eq. (8)).
    • Compute $\mathcal{C}_{\hat{\rho}_n}(F_\theta)$ by computing the vjp$(F_\theta)(\mathbf{x}_i, \mathbf{v}_j)$ and jvp$(F_\theta)(\mathbf{x}_i, \mathbf{v}_j)$, for each point $\mathbf{x}_i$ in the batch $\hat{\rho}_n$ and each Hutchison vectors $\mathbf{v}_j$ (see Eq. (26)).
    • $\hat{\mathcal{L}}_n(\theta) \leftarrow \Delta((\mathrm{I}_d - \nabla h^* \circ F_\theta)\sharp\hat{\mu}_n, \hat{\nu}_n) + \lambda_{\mathrm{MG}}\, \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(\mathrm{I}_d - \nabla h^* \circ F_\theta) + \lambda_{\mathrm{cons}}\, \mathcal{C}_{\hat{\rho}_n}(F_\theta)$.
    • Update $\theta$ to minimize $\hat{\mathcal{L}}_n(\theta)$.
```

Number of Hutchsinon vectors. Whenever we use the conservative regularizer, the number of hutchinson vectors m is fixed to the upper integer part of 20% of the dimension d.

ICNNs. All ICNNs are trained with the `NeuralDualSolver` of `OTT-JAX` which implements the Bunne et al. (2022a)'s Gaussian initializer and hence the induced specific architecture, using quadratic potentials injected in the first hidden vector. Furthermore, as suggested by Makkuva et al. (2020) and used in Bunne et al. (2021; 2022a):

- To represent discontinuous transport maps, it uses ReLu as activation function.
- It relaxes the positivity constraint on the feedforward weights W_k^z of the ICNN g_θ s.t. $T_\theta = \nabla g_\theta$ with the penalty:

$$R(\theta) = \sum_{W_k^z \in \theta} \|\max(-W_k^z, 0)\|_F^2 \tag{28}$$

MLPs. All MLPs use vanilla fully connected layers. To train MLPs within the Fan et al. (2020) saddle point problem, we follow their choice of using the PRelu activation function for both the Lagrange multiplier f and the map T. For all our MLPs, we use the GeLu activation (Hendrycks and Gimpel, 2016).

Calibration of NN sizes. As the employed ICNN architecture uses (i) linear residual layers from the input layer to each hidden layer and (ii) specific layers designed for Gaussian and identity initializers scheme, if we fix the number of layers and hidden units, they naturally have more parameters than the MLP with same number of layers and hidden units. In particular, the layers suited to the initializers scheme are quadratic in the input, so the difference in parameters explodes as the dimension increases. For instance, for data in dimension $d = 64$, an ICNN with hidden layer sizes [128, 64, 64] has 33,345 parameters, while an MLP with same hidden layer sizes and a residual layer from input layer to output layer for Gaussian initialization (see § D.2) has 24,896 parameters. Thus, the ICNN has about 33% more parameters than the MLP. To mitigate this difference, for each experiment where we use both an ICNN and an MLP, we first fix the ICNN size, then we use an MLP with the same number of layers but we adapt the number of hidden units on each of its layers to match the number of parameters up to 1%. In the previous example, this leads to an MLP with hidden layer sizes [146, 82, 82] which leads to 33,662 parameters.

D.5. Synthetic Data

ℓ_p^q **costs.** We train all MLPs with $W_{\ell_2^2,\varepsilon}$ as fitting loss. For $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$, we parametrize T_θ as an MLP and use $\lambda_{\mathrm{MG}} = 5$. For $c(\mathbf{x}, \mathbf{y}) = \frac{1}{1.5}\|\mathbf{x} - \mathbf{y}\|_{1.5}^{1.5}$ and $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, we parametrize $T_\theta = \mathrm{I}_d - \nabla h^* \circ F_\theta$ with an MLP F_θ and add conservativity regularizer $\mathcal{C}_\mu$. We use $\lambda_{\mathrm{MG}} = 1$ and $\lambda_{\mathrm{cons}} = 0.01$ in both cases. All MLPs are initialized with the Identity initializer and have hidden layer sizes [128, 64, 64]. They are trained with ADAM (Kingma and Ba, 2014) for $K_{\mathrm{iters}} = 10,000$ iterations with a learning rate $\eta = 0.01$ and a batch size $n = 1024$.

Costs on the sphere. We parameterize the maps with $T_\theta = \frac{F_\theta}{\|F_\theta\|_2}$ where F_θ is an MLP. We train the MLPs with $W_{\ell_2^2,\varepsilon}$ as fitting loss and set $\lambda_{\mathrm{MG}} = 1$ for both $c(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x}^\top \mathbf{y})$ and $c(\mathbf{x}, \mathbf{y}) = -\log(\mathbf{x}^\top \mathbf{y})$. All MLPs have hidden layer sizes [128, 64, 64] and are initialized with the Identity initializer. They are trained with ADAM for $K_{\mathrm{iters}} = 10,000$ iterations with a learning rate $\eta = 0.01$ and a batch size $n = 1024$.

D.6. Korotin Benchmark

Evaluation. We compute both the Sinkhorn divergence $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu, \nu)$ and the unexplained variance $\mathcal{L}_2^{\mathrm{UV}}(\hat{T})$ to evaluate the models on $8,192$ unseen samples from the source and the target measures.

ICNNs. We initialize the ICNNs using the Gaussian initializer scheme instantiated on $4,096$ samples. We optimize them using ADAM for $_{\mathrm{iters}} = 100,000$ and $K_{\mathrm{inner\_iters}} = 10$, with a learning rate $\eta = 10^{-4}$ a batch size $n = 1024$. For all experiments, we us ICNNs with hidden layer sizes [max(2d, 128), max(d, 64), max(d, 64)] where d is the dimension of the data. Note that we use the ICNN architecture provided by (Bunne et al., 2022a, Section 4), which is not the one used for ψ_1, ψ_2. This slightly mitigates the bias favoring ICNN-based methods induced by the benchmark pair design.

Our MLPs. We initialize the MLPs testing both Gaussian and Identity initializer scheme instantiated on $4,096$ samples. We also test the Identity initializer because it generalizes to generic costs. We train MLPs with $W_{\ell_2^2,\varepsilon}$ as fitting loss. When using regularization, we set $\lambda_{\mathrm{MG}} = 1$ and $\lambda_{\mathrm{cons}} = 0.01$ for $d \leq 64$, and $\lambda_{\mathrm{MG}} = 10$ and $\lambda_{\mathrm{cons}} = 0.1$ for $d \geq 128$. With or without regularizations, we train the MLPs for $K_{\mathrm{iters}} = 100,000$ iterations with a batch size $n = 1024$ and the Adam optimizer. 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 $d \geq 64$ 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 $4,096$ samples. We set the hidden layer sizes size according to the size of the ICNNs.

Saddle Point Problem Fan et al. (2020) MLPs. We train the saddle point problem (Fan et al., 2020) with two MLPs of hidden layer sizes adapted to the ICNN ones. We optimize them using ADAM for $K_{\mathrm{iters}} = 100,000$ and $K_{\mathrm{inner\_iters}} = 10$, with a batch size $n = 1024$ and a learning rate $\eta = 10^{-4}$, which is the learning rate mostly used in their experiments. For the dimensions $d \geq 64$, we did not succeed in tuning the learning rate to improve the performance.

Entropic map. We train the entropic map using $8,192$ from the source and the target measures and using $\varepsilon = 0.01 \cdot \mathrm{mean}(\mathbf{C})$ where $\mathbf{C}$ is the cost matrix.

D.7. Single Cell Genomics

Evaluation. For each dataset, 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. We perform such a strong train-test split because the datasets are unbalanced: they contain fewer treated cells than control cells. As we evaluate the performances with $S_{\ell_2^2,\varepsilon}$ which is a distributional metric, we need a number of test samples high enough to make this quantity meaningful. To counteract this unbalancedness, Bunne et al. (2021) makes a 80%-20% train-test split but concatenates the training and treated cells for evaluation. We do not follow this strategy to evaluate the models only on unseen treated cells.

MLPs. We train all MLPs with $W_{\ell_2^2,\varepsilon}$ as fitting loss. When using regularization, we set $\lambda_{\mathrm{MG}} = 1$ and $\lambda_{\mathrm{cons}} = 0.01$ for the 4i data, and $\lambda_{\mathrm{MG}} = 10$ and $\lambda_{\mathrm{cons}} = 0.1$ for the scRNA data. With or without regularizations, we train the MLPs for $K_{\mathrm{iters}} = 10,000$ iterations with a batch size $n = 512$ and the ADAM optimizer (Kingma and Ba, 2014) using a learning rate $\eta = 0.001$, along with a polynomial schedule of power $p = 1.5$ to decrease it to 10^{-5}. When using regularization, we initialize with the Gaussian initailizer scheme trained on half of the training set. We set the hidden layer sizes according to the ones of the ICNNs.

ICNNs. We use the Gaussian initializer scheme trained on half of the training set. We train the ICNNs using ADAM and learning rate $\eta = 10^{-4}$. Bunne et al. (2021) optimize the ICNNs on $K_{\mathrm{iters}} = 100,000$ and $K_{\mathrm{inner\_iters}} = 10$, with a batch size $n = 256$. On the other hand, since we use a batch size $n = 512$ for our models and it is a fundamental hyperparameter whose increase can drastically improve performances, especially in OT based models, we adpat the batch size while keeping the same number of epochs: we train the ICNNs on $K_{\mathrm{iters}} = 50,000$ and $K_{\mathrm{inner\_iters}} = 10$ with $B = 512$. We initialize

ICNNs with Gaussian initializer (Bunne et al., 2022a) using half of the training set. For all experiments, we use ICNNs with hidden layer sizes [128, 128, 64, 64].