Learning Elastic Costs to Shape Monge Displacements

Michal Klein
Apple
michalk@apple.com

Aram-Alexandre Pooladian
NYU
aram-alexandre.pooladian@nyu.edu

Pierre Ablin
Apple
p_ablin@apple.com

Eugène Ndiaye
Apple
e_ndiaye@apple.com

Jonathan Niles-Weed
NYU
jnw@cims.nyu.edu

Marco Cuturi
Apple
cuturi@apple.com

Abstract

Given a source and a target probability measure, the Monge problem studies efficient ways to map the former onto the latter. This efficiency is quantified by defining a *cost* function between source and target data. Such a cost is often set by default in the machine learning literature to the squared-Euclidean distance, $\ell_2^2(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$. The benefits of using *elastic* costs, defined using a regularizer τ as $c(\mathbf{x}, \mathbf{y}) := \ell_2^2(\mathbf{x}, \mathbf{y}) + \tau(\mathbf{x} - \mathbf{y})$, was recently highlighted in [Cuturi et al., 2023]. Such costs shape the *displacements* of Monge maps T, namely the difference between a source point and its image $T(\mathbf{x}) - \mathbf{x}$, by giving them a structure that matches that of the proximal operator of τ. In this work, we make two important contributions to the study of elastic costs: *(i)* For any elastic cost, we propose a numerical method to compute Monge maps that are provably optimal. This provides a much-needed routine to create synthetic problems where the ground-truth OT map is known, by analogy to the Brenier theorem, which states that the gradient of any convex potential is always a valid Monge map for the ℓ_2^2 cost; *(ii)* We propose a loss to *learn* the parameter θ of a parameterized regularizer τ_θ, and apply it in the case where $\tau_A(\mathbf{z}) := \|A^\perp \mathbf{z}\|_2^2$. This regularizer promotes displacements that lie on a low-dimensional subspace of $\mathbb{R}^d$, spanned by the p rows of $A \in \mathbb{R}^{p \times d}$. We illustrate the soundness of our procedure on synthetic data, generated using our first contribution, in which we show near-perfect recovery of A's subspace using only samples. We demonstrate the applicability of this method by showing predictive improvements on single-cell data tasks.

1 Introduction

Finding efficient ways to map a distribution of points onto another is a low-level task that plays a crucial role across many machine learning (ML) problems. Optimal transport (OT) theory [Santambrogio, 2015] has emerged as a tool of choice to solve such challenging matching problems, notably in single-cell genomics [Schiebinger et al., 2019, Tong et al., 2020, Bunne et al., 2023, 2024, Klein et al., 2023]. We focus in this work on the numerical resolution of the Monge problem, which aims, using high-dimensional source and target data samples $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $(\mathbf{y}_1, \dots, \mathbf{y}_m)$, to recover a map $T : \mathbb{R}^d \to \mathbb{R}^d$ that is simultaneously *(i)* a *pushfoward* map, in the sense that T applied on source samples recovers the distribution of target samples; *(ii) efficient*, in the sense that $T(\mathbf{x}_i)$ is not too far, on average from $\mathbf{x}_i$. The notion of efficiency can be made precise by choosing a real-valued cost function c that compares a point $\mathbf{x}$ and its mapping via $c(\mathbf{x}, T(\mathbf{x})) \in \mathbb{R}$.

Challenges in the estimation of OT maps. When using standard cost functions such as $\ell_2^2(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, the estimation of OT maps is hindered, in principle, by the curse of dimensionality [Hütter

Preprint. Under review.

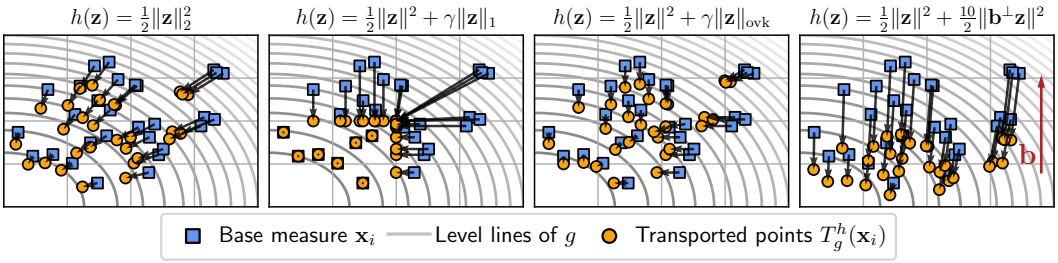

Figure 1: Illustration of ground-truth optimal transport maps with different costs h, for the same base function g. In this experiment, g is the negative of a random ICNN with 2-dimensional inputs, 3 layers and hidden dimensions of sizes $[8, 8, 8]$. All plots display the level lines of g. The optimal transport map T_g^h are recomputed four times using Prop. 1, with four different costs h, displayed above each plot. *(left)* When h is the usual ℓ_2^2 cost, we observe a typical OT map that follows from each $\mathbf{x}_i$, minus the gradient of g. With the ℓ_1 sparsity-inducing regularizer *(middle-left)*, we obtain sparse displacements: most arrows follow either of the two canonical axes, yet some points do not move at all. *(middle-right)* This is slightly different when using the k-overlap norm, which exhibits less shrinkage. With a cost that penalizes displacements that are orthogonal to a vector $\mathbf{b}$, we obtain displacements that push further to the bottom than in the *(left)* plot, as in the *(right)* plot, where displacements are almost parallel to $\mathbf{b}$. When $\mathbf{b}$ is not known beforehand, and both source and target samples are given, we present a procedure to learn adaptively such a parameter in § 5.

and Rigollet, 2021]. A simple workaround is to reduce the dimension of input data, using for instance a variational auto-encoder [Bunne et al., 2023], or learning hyperplane projections jointly with OT estimation [Paty and Cuturi, 2019, Niles-Weed and Rigollet, 2022, Lin et al., 2020, Huang et al., 2021, Lin et al., 2021]. We consider in this work another approach, which explores alternative choices for ground cost c. While OT theory is rife with rich cost structures [Ambrosio and Pratelli, 2003, Ma et al., 2005, Lee and Li, 2012, Figalli et al., 2010, Figalli and Rifford, 2010], that choice has comparatively received far less attention in machine learning, where for a vast majority of applications the cost function is often chosen as ℓ_2^2 and sometimes ℓ_2.

Cost structure impacts map structure. While the usage of Riemannian metrics within OT in ML has been considered [Cohen et al., 2021, Grange et al., 2023, Pooladian et al., 2023b], computational challenges restrict these approaches to low-dimensional manifolds. We argue in this work that costs that are translation invariant (TI), $c(\mathbf{x}, \mathbf{y}) := h(\mathbf{x} - \mathbf{y})$ with $h : \mathbb{R}^d \to \mathbb{R}$, can offer practitioners a reasonable middle ground, since many numerical schemes developed for the ℓ_2^2 cost can be extended to TI costs, both for static and dynamic formulations of OT (see e.g., Villani et al. [2009, Chap.7] or Liu [2022]). In particular, we propose to focus on *elastic costs* of the form $h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2 + \gamma\tau(\mathbf{z})$, with $\gamma > 0$ and $\tau : \mathbb{R}^d \to \mathbb{R}$ a regularizer, following the name of the elastic net regularization [Zou and Hastie, 2005] proposed in the context of regression. Cuturi et al. [2023] show that using elastic costs in OT map estimation results in Monge maps whose displacements satisfy $T(\mathbf{x}) - \mathbf{x} = -\operatorname{prox}_\tau \circ \nabla f(\mathbf{x})$, for some potential f, and are therefore shaped by the *proximal operator* of τ.

Contributions. While elastic costs offer the promise of obtaining OT maps with prescribed structure inherited from the proximal operator of a regularizer τ, our current understanding of how to use and exploit such costs is limited to the experimentation provided in [Cuturi et al., 2023]. This stands in stark contrast with the fine-grained characterization provided by Brenier that a map is optimal for the ℓ_2^2 cost if and only if it is the gradient of a convex potential. To this end:

- We show in § 3 that OT maps can be generated for any elastic cost h by running a proximal gradient descent scheme, through the proximal operator of τ, on a suitable objective. This results in, to our knowledge, the first visualization of Monge maps that extend beyond the usual grad-convex Brenier maps for ℓ_2^2 costs (see Figure 1), as well as synthetic generation in high-dimensions;
- We introduce *subspace* elastic costs in § 4, which promote displacements occurring in a low-dimensional subspace spanned by the line vectors of a matrix A, $A \in \mathbb{R}^{p\times d}$, $AA^T = I_p$, setting $\tau(\mathbf{z}) := \|A^\perp\mathbf{z}\|_2^2$. We prove sample-complexity estimates for the Monge-Bregman-Occam (MBO) estimator introduced in [Cuturi et al., 2023] with this cost (and more generally Mahalanobis costs), and establish a link with the spiked transport model [Niles-Weed and Rigollet, 2022] when $\gamma \to \infty$.

- Since the choice of the regularizer τ in the elastic cost gives rise to a diverse family of OT maps, whose structural properties are dictated by the choice of regularizer, we consider *parametrized families* τ_θ, and propose in § 5 a loss to select adaptively a suitable θ.
- We illustrate all above results, showing MBO estimator performance, *recovery* of A on the basis of i.i.d. samples, in both synthetic (using our first contribution) and single-cell data tasks, where we demonstrate an improved predictive ability compared to baseline estimators that do not learn A.

2 Background: Optimal transport with Elastic Costs

Monge Problem. Let $\mathcal{P}_2(\mathbb{R}^d)$ be the set of probability measures with finite second-order moment. We consider in this work cost functions c of the form $c(\mathbf{x}, \mathbf{y}) \coloneqq h(\mathbf{x} - \mathbf{y})$, where $h : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and, to simplify a few computations, symmetric, i.e., $h(\mathbf{z}) = h(-\mathbf{z})$. Given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Monge problem [1781] seeks a map $T : \mathbb{R}^d \to \mathbb{R}^d$ minimizing an average transport cost, as quantified by h, of the form:

$$T^\star \coloneqq \operatorname*{arg\,min}_{T_\sharp \mu = \nu} \int_{\mathbb{R}^d} h(\mathbf{x} - T(\mathbf{x}))\, \mu(\mathrm{d}\mathbf{x}) \tag{1}$$

Because the set of admissible maps T is not convex, solving (1) requires taking a detour that involves relaxing (1) into the so-called Kantorovich dual and semi-dual formulations, involving respectively two functions (or only one in the case of the semi-dual)[Santambrogio, 2015, §1.6]:

$$(f^\star, g^\star) \coloneqq \operatorname*{arg\,max}_{\substack{f,g:\mathbb{R}^d \to \mathbb{R} \\ f(\mathbf{x}) + g(\mathbf{y}) \leq h(\mathbf{x}-\mathbf{y})}} \int_{\mathbb{R}^d} f \mathrm{d}\mu + \int_{\mathbb{R}^d} g\, \mathrm{d}\nu \;=\; \operatorname*{arg\,max}_{\substack{f:\mathbb{R}^d \to \mathbb{R}, \\ f \text{ is } h\text{-concave}}} \int_{\mathbb{R}^d} f \mathrm{d}\mu + \int_{\mathbb{R}^d} \bar{f}^h \mathrm{d}\nu \tag{2}$$

A function f is said to be h-concave if there exists a function g such that f is the h-transform of g, i.e., $f = \bar{g}^h$, where for any function $g : \mathbb{R}^d \to \mathbb{R}$, we define its h-transform as

$$\bar{g}^h(\mathbf{x}) \coloneqq \inf_{\mathbf{y}} h(\mathbf{x} - \mathbf{y}) - g(\mathbf{y}). \tag{3}$$

We recall a fundamental theorem in optimal transport [Santambrogio, 2015, §1.3]. Assuming the optimal, h-concave, potential for (2), $f^\star$, is differentiable at $\mathbf{x}_0$ (this turns out to be a mild assumption since $f^\star$ is a.e. differentiable when h is), we have [Gangbo and McCann, 1996]:

$$T^\star(\mathbf{x}) = \mathbf{x} - (\nabla h)^{-1}(\nabla f^\star(\mathbf{x})) = \mathbf{x} - \nabla h^* \circ \nabla f^\star(\mathbf{x})\,, \tag{4}$$

where the convex conjugate of h reads: $h^*(\mathbf{w}) \coloneqq \sup_{\mathbf{z}} \langle \mathbf{z}, \mathbf{w} \rangle - h(\mathbf{z})$. The classic Brenier theorem [1991], which is by now a staple of OT estimation in machine learning [Korotin et al., 2019, Makkuva et al., 2020, Korotin et al., 2021, Bunne et al., 2023] through input-convex neural networks [Amos et al., 2017], is a particular example, stating for $h = \frac{1}{2}\|\cdot\|_2^2$, that $T(\mathbf{x}) = \mathbf{x} - \nabla f^\star(\mathbf{x}_0)$, since in this case, $\nabla h = (\nabla h)^{-1} = \mathrm{Id}$, see [Santambrogio, 2015, Theorem 1.22].

Maps and Elastic Costs. Cuturi et al. [2023] consider TI costs w.r.t. a regularizer τ: for $\gamma > 0$ they study *elastic costs* of the form

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + \gamma\tau(\mathbf{z}), \tag{5}$$

and show that the resulting Monge map is shaped by the proximal operator of τ:

$$T^\star(\mathbf{x}) = \mathbf{x} - \mathrm{prox}_{\gamma\tau} \circ \nabla f^\star(\mathbf{x})\,, \text{ where } \mathrm{prox}_{\gamma\tau}(\mathbf{w}) \coloneqq \operatorname*{arg\,min}_{\mathbf{z}} \frac{1}{2}\|\mathbf{w} - \mathbf{z}\|^2 + \gamma\tau(\mathbf{z})\,. \tag{6}$$

The MBO Estimator. While the result above is theoretical, as it assumes knowledge of an optimal $f^\star$, the Monge-Bregman-Occam (MBO) estimator proposes to plug into (6) an approximation of $f^\star$, recovered from samples from μ and ν. We write $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_m]$ for such samples in $\mathbb{R}^d$, possibly weighted by two probability vectors $\mathbf{a}$ and $\mathbf{b}$ of size n and m respectively. $f^\star$ can be estimated using entropy-regularized transport [Cuturi, 2013], with so-called entropic potentials [Pooladian and Niles-Weed, 2021]. This involves choosing a regularization strength $\varepsilon > 0$, and solving the following dual problem using the Sinkhorn algorithm [Peyré and Cuturi, 2019, § 4.2]:

$$(\mathbf{f}^\star, \mathbf{g}^\star) = D^\star(\mathbf{X}, \mathbf{a}, \mathbf{Y}, \mathbf{b}; h, \varepsilon) \coloneqq \operatorname*{arg\,max}_{\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle - \varepsilon \langle e^{\frac{\mathbf{f}}{\varepsilon}}, \mathbf{K} e^{\frac{\mathbf{g}}{\varepsilon}} \rangle\,. \tag{7}$$

where $\mathbf{K}_{ij} = \exp(-h(\mathbf{x}_i - \mathbf{y}_j)/\varepsilon)$. The entropy-regularized optimal transport matrix associated with that cost h and on those samples can be derived directly from these dual potentials [Peyré and Cuturi, 2019, Prop. 4.3] as $P^\star(\mathbf{X}, \mathbf{a}, \mathbf{Y}, \mathbf{b}; h, \varepsilon) \in \mathbb{R}^{n \times m}$ with entries at (i, j) equal to:

$$[P^\star(\mathbf{X}, \mathbf{a}, \mathbf{Y}, \mathbf{b}; h, \varepsilon)]_{i,j} = \exp\left(\frac{\mathbf{f}_i^\star + \mathbf{g}_j^\star - h(\mathbf{x}_i - \mathbf{y}_j)}{\varepsilon}\right). \tag{8}$$

We now introduce the soft-minimum operator, and its gradient, defined for any vector $\mathbf{u} \in \mathbb{R}^q$ as

$$\min_\varepsilon(\mathbf{u}) \coloneqq -\varepsilon \log \sum_{l=1}^q e^{-\mathbf{u}_l/\varepsilon}, \text{ and } \nabla \min_\varepsilon(\mathbf{u}) = \left[\frac{e^{-\mathbf{u}_k/\varepsilon}}{\sum_{l=1}^q e^{-\mathbf{u}_l/\varepsilon}}\right]_k.$$

Using vectors $(\mathbf{f}^\star, \mathbf{g}^\star)$, we can define estimators $\hat{f}_\varepsilon$ and $\hat{g}_\varepsilon$ for the optimal dual function $(f^\star, g^\star)$:

$$\hat{f}_\varepsilon : \mathbf{x} \mapsto \min_\varepsilon([h(\mathbf{x} - \mathbf{y}_j) - \mathbf{g}_j^\star]_j), \quad \hat{g}_\varepsilon : \mathbf{y} \mapsto \min_\varepsilon([h(\mathbf{x}_i - \mathbf{y}) + \mathbf{f}_i^\star]_i). \tag{9}$$

Plugging these approximations into (6) forms the basis for the MBO estimator outlined in Algo. 1.

Definition 1 (MBO Estimator)**.** *Given data, an elastic cost function $h = \ell_2^2 + \gamma\tau$ and solutions to Eq.(7), the MBO map estimator [Pooladian and Niles-Weed, 2021, Cuturi et al., 2023] is given by:*

$$T_\varepsilon(\mathbf{x}) = \mathbf{x} - \text{prox}_{\gamma\tau}\Big(\mathbf{x} + \sum_{j=1}^m \mathbf{p}_j(\mathbf{x})\left(\gamma \nabla\tau(\mathbf{x} - \mathbf{y}_j) - \mathbf{y}_j\right)\Big), . \tag{10}$$

where $\mathbf{p}(\mathbf{x}) \coloneqq \nabla \min_\varepsilon([h(\mathbf{x} - \mathbf{y}_j) - \mathbf{g}_j^\star]_j)$ is a probability vector.

Algorithm 1 MBO-ESTIMATOR$(\mathbf{X}, \mathbf{Y}; \gamma, \tau, \varepsilon)$

1: Set $h = \frac{1}{2}\ell_2^2 + \gamma\tau$ ▷ if $\gamma = 0$, equivalent to [Pooladian+,'21]
2: $(\mathbf{f}^\star, \mathbf{g}^\star) = D^\star(\mathbf{X}, \mathbf{a}, \mathbf{Y}, \mathbf{b}; h, \varepsilon)$ ▷ Sinkhorn **(Eq. 7)**.
3: $\mathbf{p} =$ `lambda:` $\mathbf{x} \to \text{softmax}([\mathbf{g}_j^\star - h(\mathbf{x} - \mathbf{y}_j)]_j/\varepsilon)$
4: $\mathbf{M} =$`lambda:` $\mathbf{x} \to \sum_{j=1}^m \mathbf{p}(\mathbf{x})_j \left(\gamma \nabla\tau(\mathbf{x} - \mathbf{y}_j) - \mathbf{y}_j\right)$
5: $T_\varepsilon[\gamma, \tau, \varepsilon] =$ `lambda:` $\mathbf{x} \to \mathbf{x} - \text{prox}_{\gamma\tau}(\mathbf{x} + \mathbf{M}(\mathbf{x}))$.
6: **return:** $T_\varepsilon[\gamma, \tau, \varepsilon]$

3 On Ground-Truth Monge Maps for Elastic Costs

Our strategy to compute examples of ground-truth displacements for any elastic cost h rests on the following theorem, which is a direct consequence of [Santambrogio, 2015, Theorem 1.17].

Proposition 1. *Consider a potential $g : \mathbb{R}^d \to \mathbb{R}$ and its h-transform as defined in (3). Additionally, set $T_g^h \coloneqq \text{Id} - \nabla h^* \circ \nabla \bar{g}^h$. Then T_g^h is the OT Monge map for cost h between μ and $(T_g^h)_\sharp \mu$ for any measure μ in $\mathcal{P}(\mathbb{R}^d)$.*

The ability to compute an OT map for h therefore hinges on the ability to solve numerically the h-transform (3) of a potential function g. This can be done, provably, as long as g is concave and smooth, and prox_τ is available, as shown in the following result

Proposition 2. *Assume g is concave, L-smooth, and that $\lambda < 2/L$. Setting $\mathbf{y} = \mathbf{x}$ and iterating*

$$\mathbf{y} \leftarrow \mathbf{x} + \text{prox}_{\frac{\lambda\gamma}{\lambda+1}\tau}\left(\frac{\mathbf{y} - \mathbf{x} + \lambda \nabla g(\mathbf{y})}{1 + \lambda}\right) \tag{11}$$

converges to a point $\mathbf{y}^\star(\mathbf{x}) = \arg\min_{\mathbf{y}} h(\mathbf{x} - \mathbf{y}) - g(\mathbf{y})$. Furthermore, we have

$$\bar{g}^h(\mathbf{x}) = h(\mathbf{x} - \mathbf{y}^\star(\mathbf{x})) - g(\mathbf{y}^\star(\mathbf{x})), \nabla \bar{g}^h(\mathbf{x}) = \nabla h(\mathbf{x} - \mathbf{y}^\star(\mathbf{x})), \text{ and } T_g^h(\mathbf{x}) = \mathbf{y}^\star(\mathbf{x}). \tag{12}$$

Proof. Because h is the sum of a quadratic norm with $\gamma\tau$, the proximal operator of λh can be restated in terms of the proximal operator of τ [Parikh et al., 2014, §2.1.1]. The convergence of iterates (11) follows from [Beck and Teboulle, 2009, Thm. 1] or [Rockafellar, 1976, Thm. 1]. The final identities are given by [Bauschke and Combettes, 2011, Prop. 18.7]. □

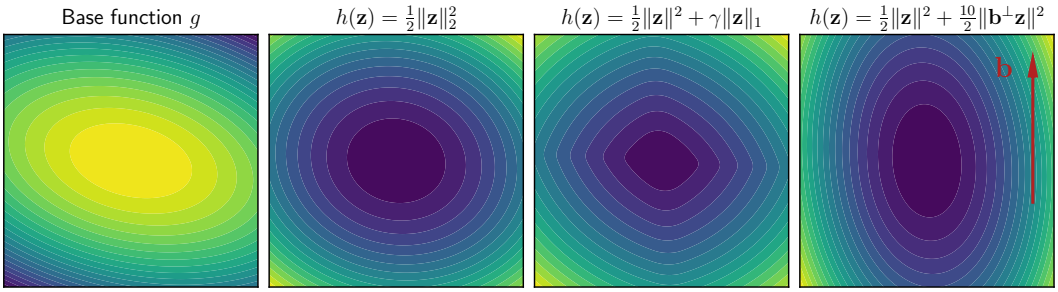

Figure 2: Illustration of the h-transform computation in 2d. *(left)*: base concave potential g, here a negative quadratic. *(other figures)* Level lines of the corresponding h-transform $\bar{g}^h$ for different choices of h. The h-transform is computed using the iterations described in Prop. 2.

As summarized in Algo.2, the proximal operator of τ is the only thing needed to implement iterations (11), and, as a result, the h-transform of a suitable concave potential. We can then plug the solution in (12) to evaluate the pushforward T_g^h. In practice, we use the JAXOPT [Blondel et al., 2021] library to run proximal gradient descent. We illustrate numerically in 2D the resulting transport maps for different choices of regularizer τ in Fig. 1. In this illustration, we use the same base function g, and see clearly the impact of c on h transforms.

Algorithm 2 GROUND-TRUTH OT MAP T_g^h

Inputs: point $\mathbf{x}$, elastic cost $h = \frac{1}{2}\ell_2^2 + \gamma\tau$, potential g.
$\mathbf{y} := \mathbf{x}$
while not converged **do**
  $\mathbf{y} \leftarrow \mathbf{x} + \mathrm{prox}_{\frac{\lambda\gamma}{\lambda+1}\tau}\left(\frac{\mathbf{y}-\mathbf{x}+\lambda\nabla g(\mathbf{y})}{1+\lambda}\right)$
end while
return: $\mathbf{y}$ ▷ $T_g^h(\mathbf{x})$ in **(Prop. 1)**

4 Subspace Elastic Costs

Recall that for a rank-p matrix $A \in \mathbb{R}^{p\times d}$, $p \leq d$, the projection matrix that maps it to its orthogonal is $A^\perp = I - A^T(AA^T)^{-1}A$. When A lies on the Stiefel manifold (i.e. $AA^T = I$), we have the simplification $A^\perp = I - A^TA$. This results in the Pythagorean identity $\|\mathbf{z}\|^2 = \|A^\perp\mathbf{z}\|^2 + \|A\mathbf{z}\|^2$. In order to promote displacements that happen *within* the span of A, we must set a regularizer that penalizes the presence of $\mathbf{z}$ within its *orthogonal complement*, namely

$$\tau_{A^\perp}(\mathbf{z}) := \tfrac{1}{2}\|A^\perp\mathbf{z}\|_2^2. \tag{13}$$

Since $\tau_{A^\perp}$ is quadratic, its proximal operator can be obtained by solving a linear system [Parikh et al., 2014, §6.1.1]; developing and using the matrix inversion lemma results in two equivalent quantities

$$\mathrm{prox}_{\gamma\tau_{A^\perp}}(\mathbf{z}) = \left(I_d + \gamma(A^\perp)^TA^\perp\right)^{-1}\mathbf{z} = \tfrac{1}{1+\gamma}\left(I_d + \gamma A^T(AA^T)^{-1}A\right)\mathbf{z}. \tag{14}$$

To summarize, given an orthogonal sub-basis A of p vectors (each of size d), promoting that a vector $\mathbf{z}$ lies in its orthogonal can be achieved by regularizing its norm in the space orthogonal to the span of A. That norm has a proximal operator that can be computed by parameterizing A *explicitly*, either as a full-rank $p \times d$ matrix, or more simply a $p \times d$ orthogonal matrix, to recover the suitable proximal operator for $\tau_{A^\perp}$ in (14). Because that operator is simpler when $A \in \mathcal{S}_{p,d}$ is in the Stiefel manifold,

$$\mathrm{prox}_{\gamma\tau_{A^\perp}}(\mathbf{z}) = \frac{1}{1+\gamma}\left(I_d + \gamma A^TA\right)\mathbf{z}. \tag{15}$$

We propose to restrict the study in this work to elastic costs of the form 14 where $A \in \mathcal{S}_{p,d}$. We also present in Appendix A alternative parameterizations left aside for future work.

4.1 Statistical Aspects of Subspace Monge Maps

The family of costs (13) is designed to promote transport maps whose displacements mostly lie in a low-dimensional subspace of $\mathbb{R}^d$. In this section, we consider the statistical complexity of estimating such maps from data, assuming A is known. The question of estimating transport maps was first

studied in a statistical context by Hütter and Rigollet [2021], and subsequent research has proposed alternative estimation procedures, with different statistical and computational properties [Deb et al., 2021, Manole et al., 2021, Muzellec et al., 2021, Pooladian and Niles-Weed, 2021]. We extend this line of work by considering the analogous problem for Monge maps with structured displacements.

We show that with a proper choice of ε, the MBO estimator outlined in Definition 1 is a consistent estimator of $T^\star$ as $n \to \infty$, and prove a rate of convergence in $L^2(\mu)$. We also give preliminary theoretical evidence that, as $\gamma \to \infty$, maps corresponding to the subspace structured cost $\frac{1}{2}\ell_2^2 + \gamma\tau_{A^\perp}$ can be estimated at a rate that depends only on the subspace dimension p, rather than on the ambient dimension d, thereby avoiding the *curse of dimensionality*.

Sample Complexity Estimates for the MBO Estimator. The MBO estimator is a generalization of the entropic map estimator, originally defined by Pooladian and Niles-Weed [2021] for the quadratic cost $h = \frac{1}{2}\ell_2^2$. This estimator has been statistically analyzed in several regimes, see e.g., [Pooladian et al., 2023a, Rigollet and Stromme, 2022, del Barrio et al., 2022] and [Goldfeld et al., 2022]. We show that this procedure also succeeds for subspace structured costs of the form $h = \frac{1}{2}\ell_2^2 + \gamma\tau_{A^\perp}$. As a result of being recast as an estimation task for quadratic cost, the following sample-complexity result for the MBO estimator follows from [Pooladian and Niles-Weed, 2021, Theorem 3], and a computation relating the MBO estimator to a barycentric projection for the costs we consider (see Appendix B for the full statements, proofs, and applicability to general Mahalanobis norms).

Theorem 1. *Let $A \in \mathbb{R}^{p\times d}$ be fixed,and suppose ν has an upper- and lower- bounded density, and μ is upper-bounded, both supported over $\Omega \subseteq \mathbb{R}^d$ compact. Consider $T^\star$ of the form Equation (23) for some $\gamma \geq 0$ fixed, and suppose we have samples $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mu$ and $\mathbf{y}_1, \dots, \mathbf{y}_n \sim (T^\star)_\sharp\mu$. Let $\hat{T}_\varepsilon$ be the MBO estimator with $\varepsilon \asymp n^{-\frac{1}{d+4}}$. Then it holds that*

$$\mathbb{E}\|\hat{T}_\varepsilon - T^\star\|_{L^2(\mu)}^2 \lesssim n^{-\frac{2}{d+4}},$$

where the underlying constants depend on properties of μ, ν, γ and A.

4.2 Connection to the Spiked Transport Model

The additional structure we impose on the displacements allows us to closely relate our model to the "spiked transport model" as defined by Niles-Weed and Rigollet [2022]. The authors studied the estimation of the Wasserstein distance in the setting where the Brenier map between μ and ν takes the form,

$$T_{\text{spiked}}(\mathbf{x}) = \mathbf{x} - A^T(A\mathbf{x} - S(A\mathbf{x})), \tag{16}$$

where $A \in \mathcal{S}_{p,d}$ and $S : \mathbb{R}^p \to \mathbb{R}^p$ is the gradient of a convex function on $\mathbb{R}^p$. Divol et al. [2022] performed a statistical analysis of the map estimation problem under the spiked transport model. They constructed an estimator $\hat{T}_n$ such that the $L^2(\mu)$ risk decays with respect to the *intrinsic dimension* $p \ll d$; this is summarized in the following theorem.

Theorem 2 (Divol et al., 2022, Section 4.6)**.** *Suppose μ has compact support, with density bounded above and below. Suppose further that there exists a matrix $A \in \mathbb{R}^{p\times d}$ on the Stiefel manifold such that $\nu := (T_{spiked})_\sharp\mu$, with T_{spiked} defined as in Equation (16). Assume that μ is known explicitly. Given n i.i.d. samples from ν, there exists an estimator $\hat{T}_n$ satisfying*

$$\mathbb{E}\|\hat{T}_n - T_{spiked}\|_{L^2(\mu)}^2 \lesssim_{\log(n)} n^{-\Theta(\frac{1}{p})}. \tag{17}$$

We now argue that the spiked transport model can be recovered in the large γ limit of subspace structured costs. Indeed, if $\gamma \to \infty$, then displacements in the subspace orthogonal to A are heavily disfavored, so that the optimal coupling will concentrate on the subspace given by A, thereby recovering a map of the form (16), which by Theorem 2 can be estimated at a rate independent of the ambient dimension. Making this observation quantitative by characterizing the rate of estimation of $T^\star$ as a function of γ for γ large is an interesting question for future work.

5 A Bilevel Loss to Learn Elastic Costs

Following § 4, we propose a general loss to *learn* the parameter θ of a family of regularizers $\{\tau_\theta\}_\theta$ given source and target samples only. Our goal is to infer adaptively a θ that promotes

regular displacements, apply it within the estimation of Monge maps using MBO, and leverage this knowledge to improve prediction quality. Given input and target measures characterized by point clouds $\mathbf{X}, \mathbf{Y}$ and probability weights $\mathbf{a}, \mathbf{b}$, our loss follows a simple intuition: the ideal parameter θ should be such that the bulk of the OT cost bore by the optimal Monge map, for that cost, is dominated by displacements that have a *low* regularization value. Since the only moving piece in our pipeline will be θ, we consider all other parameters *constant* in the computation of the primal solution, to re-write (8) as:

$$P^\star(\theta) \coloneqq P^\star\left(\mathbf{X}, \mathbf{a}, \mathbf{Y}, \mathbf{b}; \tfrac{1}{2}\ell_2^2 + \gamma\tau_\theta, \varepsilon\right) \in \mathbb{R}^{n\times m}. \tag{18}$$

Each entry $[P^\star(\theta)]_{ij}$ quantifies the optimal association strength between a pair $(\mathbf{x}_i, \mathbf{y}_j)$ when the cost is parameterized by θ, where a given pair can be encoded as a displacement $\mathbf{z}_{ij} \coloneqq \mathbf{y}_j - \mathbf{x}_i$. For the regularizer θ to shape displacements, we expect $P^\star(\theta)$ to have a large entry on displacements $\mathbf{z}_{ij}$ that exhibit a low regularizer $\tau_\theta(\mathbf{z}_{ij})$ value. In other words, we expect that $\tau_\theta(\mathbf{z}_{ij})$ to be as small as possible when $P^\star_{ij}(\theta)$ is high. We can therefore consider the loss

Definition 2 (Elastic Costs Loss)**.** *Given two weighted point clouds* $\mathbf{a}, \mathbf{X}, \mathbf{b}, \mathbf{Y}$*, and* $P^\star(\theta)$ *defined implicitly, as an OT solution in Equation* (8)*, let*

$$\mathcal{L}(\theta) \coloneqq \langle P^\star(\theta), R(\theta)\rangle\,, \text{ with } [R(\theta)]_{ij} = \tau_\theta(\mathbf{z}_{ij}). \tag{19}$$

Because $P^\star(\theta)$ is itself obtained as the solution to an optimization problem, minimizing $\mathcal{L}$ is therefore a *bilevel* problem. To solve it, we must compute the gradient $\nabla\mathcal{L}(\theta)$, given by the vector-Jacobian operators $\partial P^\star(\cdot)^*[\cdot]$ and $\partial R(\cdot)^*[\cdot]$ of $P^\star$ and R respectively, borrowing notations from [Blondel and Roulet, 2024, §2.3] (see also § C for a walk-through of this identity)

$$\nabla\mathcal{L}(\theta) = \partial P^\star(\theta)^*[R(\theta)] + \partial R(\theta)^*[P^\star(\theta)] \tag{20}$$

The first operator $\partial P^\star(\cdot)^*[\cdot]$ requires differentiating the solution of an optimization problem $P^\star(\theta)$. This can be done [Blondel and Roulet, 2024, §10.3.3] using either unrolling of Sinkhorn iterations or using implicit differentiation. We rely on OTT-JAX [Cuturi et al., 2022] to provide that operator, using unrolling. The second operator $\partial R(\cdot)^*[\cdot]$ can be trivially evaluated, since it only involves differentiating the regularizer function $\tau_\theta(\cdot)$. These steps are summarized in Algo. 3.

Learning Subspace Costs. We focus in this section on the challenges arising when optimizing subspace costs, as detailed in Section 5. Learning matrix A in this context is equivalent to learning a subspace in which the displacement between the source and target measures happen mostly in the range of A. As discussed previously, the cost function $\mathcal{L}(A)$ should be optimized over the Stiefel manifold [Edelman et al., 1998]. We use Riemannian gradient descent [Boumal, 2023] for this task, which iterates, for a step-size $\eta > 0$

$$A \leftarrow \mathcal{P}(A - \eta\tilde{\nabla}\mathcal{L}(A))\,,$$

with the *Riemannian gradient* of $\mathcal{L}$ given by $\tilde{\nabla}\mathcal{L}(A) \coloneqq G - AG^TA$ where: $G \coloneqq \nabla\mathcal{L}(A)$ the standard Euclidean gradient of A computed with automatic differentiation provided in (20); $\mathcal{P}$ is the projection on the Stiefel manifold, with formula $\mathcal{P}(A) = (AA^\top)^{-1/2}A$. These updates ensure that one stays on the manifold [Absil and Malick, 2012].

Algorithm 3 RECOVER-THETA: $(\mathbf{X}, \mathbf{Y}; \gamma, \theta_0)$

1: **for** $t = 0, \dots, T$ **do**
2: Sample mini-batches $\mathbf{X}_n, \mathbf{Y}_n$ from $\mathbf{X}, \mathbf{Y}$
3: Compute coupling: ▷ Sinkhorn **(Eq. 8)**

$$P(\theta_t) \leftarrow P^\star(\mathbf{X}_n, \tfrac{\mathbf{1}}{n}, \mathbf{Y}_n, \tfrac{\mathbf{1}}{n}; \tfrac{1}{2}\ell_2^2 + \gamma\tau_{\theta_t}, \varepsilon).$$

4: Compute loss: ▷ **(Eq. 19)**

$$\mathcal{L}(\theta_t) = \langle P(\theta_t), R(\theta_t)\rangle.$$

5: gradient $\mathbf{g} \leftarrow \nabla\mathcal{L}(\theta_t)$ using auto-diff.
6: $\theta_{t+1} \leftarrow$ GRAD-UPDATE$(\theta_t, \mathbf{g})$
7: **end for**
8: **return:** θ_T

6 Experiments

Thanks to our ability to compute ground-truth h-optimal maps presented in § 3, we generate benchmark tasks to measure the performance of Monge map estimators. We propose in § 6.1 to test the MBO estimator [Cuturi et al., 2023] when the ground-truth cost h that has generated those benchmarks is known. In §6.2, we consider the more difficult task of learning simultaneously, and as outlined in § 5, an OT map and the ground-truth parameter of a subspace-elastic cost defined by

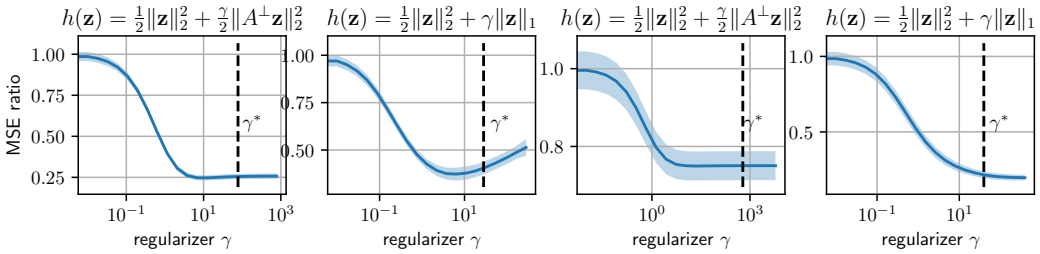

Figure 3: Performance of the MBO estimator on two ground-truth tasks involving the $\tau = \ell_1$ and $\tau_{A^\perp} = \|A^\perp\mathbf{z}\|_2^2$ structured costs, where $p = 2$ in dimension $d = 5$ (two figures to the *left*) and dimension $d = 10$ (two figures to the *right*). We display the MSE ratio between the MSE estimated with a regularizer strength $\gamma > 0$ and that in the absence of regularization (i.e., $\gamma = 0$). The level of regularization used for generating the ground-truth data is γ^*, whereas performance are shown varying w.r.t. γ. We display curves $\pm$ s.t.d. estimated over 10 random seeds.

a matrix A^* of size $p^* \times d$. The cost is parameterized by a matrix $\hat{A}$ of size $\hat{p} \times d$, where $\hat{p}$ is an estimate of the ground-truth subspace dimension p^* (usually not known), equal to or larger than p^*. We check with this synthetic task the soundness of the loss $\mathcal{L}(\theta)$, Definition 2, and of our Riemannian descent approach by evaluating to what extent the $\hat{p}$ vectors in $\hat{A}$ recovers the subspace spanned by A^*. Finally, we consider in § 6.3 a direct application of subspace elastic costs to real data, without any ground-truth knowledge, using perturbations of single-cell data. In this experiment, our pipeline learns both an OT map and a subspace. Our code implements a parameterized `RegTICost` class, added to OTT-JAX [Cuturi et al., 2022]. Such costs can be fed into the `Sinkhorn` solver, and their output cast as `DualPotentials` objects that can output the T_ε map given in Definition 1.

6.1 MBO on Synthetic Ground-Truth Displacement

In this section, we assume that the regularizer τ is *known*, using the same τ both for generation and estimation, but that the ground-truth regularization strength γ^* used to generate data is not known. We use that cost to evaluate the transport associated with $\bar{g}_\varepsilon^h$ on a sample of points, using Proposition 1, and then compare the performance of Sinkhorn based estimators, either with that cost or the standard $\frac{1}{2}\ell_2^2$ cost (which corresponds to $\gamma = 0$).

We consider the $\tau = \ell_1$ and $\tau_{A^\perp} = \|A^\perp\mathbf{z}\|_2^2$ regularizers, and their associated proximal operators. While we assume knowledge of τ in the MBO estimator, we do not use the ground-truth regularization γ^* which generated the data, and instead consider it a free parameter. The data is generated following § 3 by sampling a concave quadratic function $g(\mathbf{z}) := -\frac{1}{2}(\mathbf{z} - \mathbf{w})^T M(\mathbf{z} - \mathbf{w})$ where M is a Wishart matrix, sampled as $M = QQ^T$, where $Q \in \mathbb{R}^{d \times 2d}$ is multivariate Gaussian, and $\mathbf{w}$ is a random Gaussian vector. We then sample $n = 1024$ Gaussian points stored in $\mathbf{X}_T$ and transport each using the map defined in Proposition 1, computed in practice with Proposition 2. This recovers matched train data $\mathbf{X}_T$ and $\mathbf{Y}_T$. We do the same for a test fold $\mathbf{X}_t, \mathbf{Y}_t$ of the same size, to report our metric, the mean squared error (MSE), defined as $\|T_\varepsilon(\mathbf{X}_t) - \mathbf{Y}_t\|_2^2$, where T_ε is obtained from Definition 1 using $\mathbf{X}_T, \mathbf{Y}_T$. We plot this MSE as a function of γ, where $\gamma = 0$ corresponds exactly to the MBO using the naked ℓ_2^2 cost. We observe in Figure 3 that the MBO estimator with positive γ outperforms significantly that using ℓ_2^2 only, for any range of the parameter γ.

6.2 Recovery of Ground-Truth Subspace Parameters in Elastic Costs

We propose to test the ability of Algorithm 3 to recover the ground-truth A^* parameter of a regularizer $\tau_{A^\perp}$ as defined in (13). To do so, we proceed as follows: For dimension d, we build the ground-truth cost h by selecting A^*, sampling a $p^* \times d$ normal matrix that is then re-projected on the $p^* \times d$ Stiefel manifold. Next, we sample a random ICNN, and set the base function g to be its negative. We then sample a point cloud $\mathbf{X}$ of $n = 512$ standard Gaussian points, and apply, following Proposition 1, the corresponding ground-truth transport to obtain $\mathbf{Y}$ of the same size. We tune the regularization parameter γ for τ, to ensure that the p^* first singular values of displacements $\mathbf{Y} - \mathbf{X}$ captured either

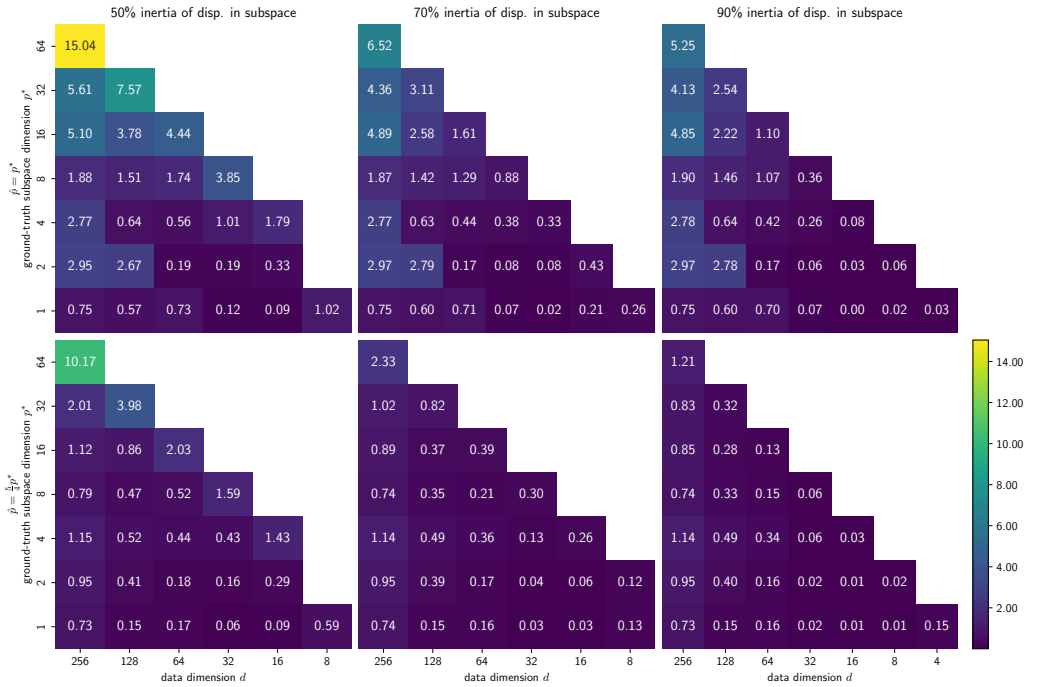

Figure 4: Error averaged over 5 seeded runs (lower is better) in $[0, 1]$ of the $\hat{p} \times d$ orthogonal matrix $\hat{A}$ recovered by our algorithm, compared to the ground-truth $p^* \times d$ cost matrix A^*. Error bars are not shown for compactness, but are negligible since all quantities are bounded below and close to 0. Dimensions d, p^* vary in each of these 6 plots, whereas $\hat{p}$ is fixed to either p^* (top row) or $1.25p^*$ (bottom row). Error is quantified as the normalized squared-residual error obtained when projecting the p^* basis vectors of A^* onto the span of $\hat{A}$. From left to right, the regularization strength γ^* increases to ensure that 50%, 70% and 90% of the total inertia of all displacements generated by the ground-truth Monge map are borne by the p^* highest singular values. As expected, recovery is easier when $\hat{p}$ is slightly larger than p^* (bottom) compared to being exactly equal (top). It is also easier as the share of inertia captured by p^* increases.

50%, 70% or 90% of the total inertia. We expect that the larger this percentage, the easier recovery should be. See § D for details.

We launch our solver fed with these datasets with a subspace dimension $\hat{p}$ preset in advance to either $\hat{p} = p^*$ (matching ground truth) or $\hat{p} = \frac{5}{4}p^*$ (overbudgeting). We measure recovery of A^* by $\hat{A}$ through the average (normalized by the basis size) of the residual error, when projecting the vectors in A^* in the span of the basis $\hat{A}$, namely $\|A^* - \hat{A}\hat{A}^T A^*\|_2^2/p^*$. For simplicity, we report performance after 1000 iterations of Riemannian gradient descent, with a stepsize η of $0.1/\sqrt{i+1}$ at iteration i. All results in Figure 4 agree with intuition in the way performance varies with $d, p^*, \hat{p}$. More importantly, with errors that are often below one percent, we can be confident that our algorithm is sound. We observe that most underperforming experiments could be improved using early stopping.

6.3 Learning Displacement Subspaces for Single-Cell Transport

We borrow the experimental setting in [Cuturi et al., 2023], using single-cell RNA sequencing data from [Srivatsan et al., 2020]. The original dataset shows the responses of cancer cell lines to 188 drug perturbations, downsampled to the 5 drugs (Belinostat, Dacinostat, Givinostat, Hesperadin, and Quisinostat) that have the largest effect. After various standard pre-processings (dropping low-variability genes, and using a $\log(1 + \cdot)$ scaling), we project the dataset to $d = 256$ directions using PCA. In Table 1, we report the total number of cells used for experiments after pre-processing. We then use 80% train/20% test folds to benchmark two MBO estimators: that computed using the ℓ_2^2 cost, and ours, using an elastic subspace cost, following the learning pipeline outlined in § 5. We plot

the Sinkhorn divergence (cf. Feydy et al. [2019]) for the ℓ_2^2 cost for reference (see the documentation in OTT-JAX [Cuturi et al., 2022]).

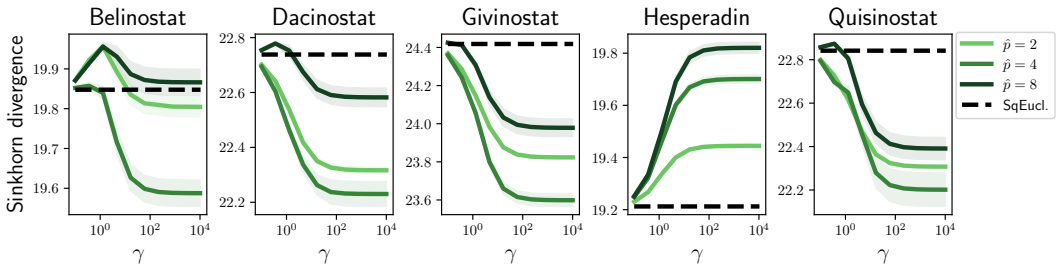

Figure 5: Predictive performance of the MBO estimator on single-cell datasets, $d = 256$, using either the naive baseline ℓ_2^2 cost (black dotted line) or elastic subspace cost (13), with varying γ and $\hat{p}$. Remarkably, promoting displacements to happen in a subspace of much lower dimension improves predictions, even when measured in the squared-Euclidean distance.

Table 1: Number of cells for each cell line and drug/control splits.

	Control	Dac.	Giv.	Bel.	Hes.	Quis.
A549	3274	558	703	669	436	475
K562	3346	388	589	656	624	339
MCF7	6346	1562	1805	1684	882	1520

Conclusion. In this work, we proposed an algorithmic mechanism to design ground-truth transports for elastic costs. As a first application, we were able to successfully benchmark the MBO estimator of [Cuturi et al., 2023] on two tasks (involving the ℓ_1 and an orthogonal projection norm), showcasing the versatility of the MBO framework. Next, we demonstrated our ability to leverage subspace-penalizing costs to *learn displacement subspaces* by solving an *inverse OT problem*. We showed successful numerical performance of the MBO estimator when the subspace is known but the regularization strength is not, but also that we were able to learn the ground-truth subspace. We foresee several open directions, the most encouraging being considering other learnable proximal operators beyond subspace approaches, and cementing connections and distinctions between subspace regularized transport (where *displacements* happen in a subspace) vs. the spiked transport model (where all points are projected on a subspace prior to being transported)

References

P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

Luigi Ambrosio and Aldo Pratelli. Existence and stability results in the l1 theory of optimal transportation. *Optimal Transportation and Applications: Lectures given at the CIME Summer School, held in Martina Franca, Italy, September 2-8, 2001*, 2003.

Brandon Amos, Lei Xu, and J Zico Kolter. Input Convex Neural Networks. volume 34, 2017.

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Mathieu Blondel and Vincent Roulet. The elements of differentiable programming. *arXiv preprint arXiv:2403.14606*, 2024.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4), 1991. doi: 10.1002/cpa.3160440402.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20, 2023.

Charlotte Bunne, Geoffrey Schiebinger, Andreas Krause, Aviv Regev, and Marco Cuturi. Optimal transport for single-cell and spatial omics. *Nature Reviews Methods Primers*, 4(1):58, 2024.

Samuel Cohen, Brandon Amos, and Yaron Lipman. Riemannian convex potential maps. In *International Conference on Machine Learning*, pages 2028–2038. PMLR, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Marco Cuturi, Michal Klein, and Pierre Ablin. Monge, bregman and occam: Interpretable optimal transport in high-dimensions with feature-sparse maps. In *Proceedings of the 40th ICML*, 2023.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *arXiv preprint arXiv:2107.01718*, 2021.

Eustasio del Barrio, Alberto Gonzalez-Sanz, Jean-Michel Loubes, and Jonathan Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*, 2022.

Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

Alessio Figalli and Ludovic Rifford. Mass transportation on sub-riemannian manifolds. *Geometric and functional analysis*, 20:124–159, 2010.

Alessio Figalli, Ludovic Rifford, and Cédric Villani. On the ma–trudinger–wang curvature on surfaces. *Calculus of Variations and Partial Differential Equations*, 39:307–332, 2010.

Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and the sinkhorn divergence. *arXiv preprint arXiv:2207.08683*, 2022.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Daniel Grange, Mohammad Al-Jarrah, Ricardo Baptista, Amirhossein Taghvaei, Tryphon T Georgiou, Sean Phillips, and Allen Tannenbaum. Computational optimal transport and filtering on riemannian manifolds. *IEEE Control Systems Letters*, 2023.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A riemannian block coordinate descent method for computing the projection robust wasserstein distance. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4446–4455. PMLR, 2021.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimée Bastidas-Ponce, Marta Tarquis-Medina, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, and Fabian J. Theis. Mapping cells through time and space with moscot. *bioRxiv*, 2023. doi: 10.1101/2023.05.11.540374. URL `https://www.biorxiv.org/content/early/2023/05/12/2023.05.11.540374`.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. 2019.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark. 2021.

Paul WY Lee and Jiayong Li. New examples satisfying ma–trudinger–wang conditions. *SIAM Journal on Mathematical Analysis*, 44(1):61–73, 2012.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems*, 33: 9383–9397, 2020.

Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael I Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Xi-Nan Ma, Neil S Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Archive for rational mechanics and analysis*, 177:151–183, 2005.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. volume 37, 2020.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, 1781.

Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1 (3):127–239, 2014.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. *arXiv preprint arXiv:2301.11302*, 2023a.

Aram-Alexandre Pooladian, Carles Domingo-Enrich, Ricky TQ Chen, and Brandon Amos. Neural optimal transport with Lagrangian costs. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023b.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 2020.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 2020.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A More on Subspace Elastic Costs

Recall that for a rank-p matrix $A \in \mathbb{R}^{p\times d}$, $p \le d$, the projection matrix that maps it to its orthogonal is $A^\perp = I - A^T(AA^T)^{-1}A$. When A lies in the Stiefel manifold (i.e. $AA^T = I$), we have the simplification $A^\perp = I - A^TA$. This results in the Pythagorean identity $\|\mathbf{z}\|^2 = \|A^\perp\mathbf{z}\|^2 + \|A\mathbf{z}\|^2$, as intended. In order to promote displacements that happen *within* the span of A, we must set a regularizer that penalizes the presence of $\mathbf{z}$ within its *complement*:

$$\tau_{A^\perp}(\mathbf{z}) := \tfrac{1}{2}\|A^\perp\mathbf{z}\|_2^2 = \tfrac{1}{2}\mathbf{z}^T(A^\perp)^TA^\perp\mathbf{z} = \tfrac{1}{2}\mathbf{z}^T(I_d - A^T(AA^T)^{-1}A)\mathbf{z}.$$

Since $\tau_{A^\perp}$ is a quadratic form, its proximal operator can be obtained by solving a linear system [Parikh et al., 2014, §6.1.1]; developing and using the matrix inversion lemma results in

$$\operatorname{prox}_{\gamma\tau_{A^\perp}}(\mathbf{z}) = \left(I_d + \gamma(A^\perp)^TA^\perp\right)^{-1}\mathbf{z} = \tfrac{1}{1+\gamma}(I + \gamma A^T(AA^T)^{-1}A)\,\mathbf{z}. \tag{21}$$

To summarize, given an orthogonal sub-basis A of p vectors (each of size d), promoting that a vector $\mathbf{z}$ lies in its orthogonal can be achieved by regularizing its norm in the orthogonal of A. That norm has a proximal operator that can be computed either by

1. Parameterizing A *implicitly*, through an *explicit* parameterization of an orthonormal basis B for $A^\perp$, as a matrix directly specified in the $(d-p)\times p$ Stiefel manifold. This can alleviate computations to obtain a closed form for its proximal operator:

$$\operatorname{prox}_{\gamma\tau_{A^\perp}}(\mathbf{z}) = \operatorname{prox}_{\gamma\tau_B}(\mathbf{z}) = \mathbf{z} - B^T\left(B\mathbf{z} - \frac{1}{1+\gamma}B\mathbf{z}\right) = \left(I_d - \frac{\gamma}{1+\gamma}B^TB\right)\mathbf{z},$$

   but requires storing B, a $(d-p)\times d$ orthogonal matrix, which is cumbersome when $p \ll d$.

2. Parameterizing A *explicitly*, either as a full-rank $p \times d$ matrix, or more simply a $p \times d$ orthogonal matrix, to recover the suitable proximal operator for $\tau_{A^\perp}$, by either

(a) Falling back on the right-most expression in (14) in the linear solve, which can be handled using sparse conjugate gradient solvers, since the application of the right-most linear operator has complexity $(p+1) \times d$ and is positive definite, in addition to the linear solve of complexity $O(p^3)$. This simplifies when A is orthogonal, $A \in \mathcal{S}_{p,d}$ since in that case,

$$\operatorname{prox}_{\gamma \tau_{A^\perp}}(\mathbf{z}) = \frac{1}{1+\gamma}\left(I_d + \gamma A^T A\right)\mathbf{z}. \tag{22}$$

(b) Alternatively, compute a matrix in the $(d-p) \times p$ Stiefel manifold that spans the same linear space as, through the Gram-Schmidt process [Golub and Van Loan, 2013, p.254] of the $d \times d$ matrix $A^\perp$ or rank $d-p$, $B :=$ Gram-Schmidt$(A^\perp)$, to fall back on the expression above.

B Proofs from Section 4.1

To perform this analysis, we rely on the following characterization of optimal maps for subspace structured costs, which reveals a close connection with optimal maps for the standard ℓ_2^2 cost.

Proposition 3. *Let $T^\star$ be the optimal map between μ and ν for the cost $h = \frac{1}{2}\ell_2^2 + \gamma\tau_{A^\perp}$. Denote by W the linear map $\mathbf{x} \mapsto ((1+\gamma)I - \gamma A^T A)^{1/2}\mathbf{x}$. Then $W \circ T^\star \circ W^{-1}$ is the Brenier map (i.e., ℓ_2^2 optimal map) between $W_\sharp\mu$ and $W_\sharp\nu$. Equivalently, $T^\star$ is h-optimal if and only if it can be written*

$$T^\star = W^{-1} \circ \tilde{T} \circ W, \tag{23}$$

where $\tilde{T}$ is the gradient of a convex function.

Proof of Proposition 3. The cost $h = \frac{1}{2}\ell_2^2 + \gamma\tau_{A^\perp}$ can be written as

$$\tfrac{1}{2}\left[z^\top(I + \gamma(A^\perp)^\top A^\perp)z\right] = \tfrac{1}{2}\|Wz\|^2\,.$$

The optimal transport problem we consider is therefore equivalent to minimizing

$$\min_{\pi \in \Gamma(\mu,\nu)} \int \tfrac{1}{2}\|Wx - Wy\|^2 \mathrm{d}\pi(x,y) = \min_{\pi \in \Gamma(W_\sharp\mu, W_\sharp\nu)} \int \tfrac{1}{2}\|x' - y'\|^2 \mathrm{d}\pi(x',y')\,. \tag{24}$$

Brenier's theorem implies that the solution to the latter problem is given by the gradient of a convex function, and that this property uniquely characterizes the optimal map. Writing this function as $\tilde{T}$, we obtain that the optimal coupling between $W_\sharp\mu$ and $W_\sharp\nu$ is given by $y' = \tilde{T}(x')$, which implies that the optimal h-coupling between μ and ν is given by $T^\star = W^{-1} \circ \tilde{T} \circ W$, as desired.

□

The proof of Theorem 1 requires the following two lemmas.

Lemma 1. *For costs of the form $h(z) = \frac{1}{2}z^\top Bz$ where B is positive definite, the MBO estimator between two measures μ and ν can be written as the barycentric projection of the corresponding optimal entropic coupling.*

Proof. Note that $h^*(w) = \frac{1}{2}w^\top B^{-1}w$, and thus $\nabla h^*(w) = B^{-1}w$. Let $(f_\varepsilon, g_\varepsilon)$ denote the optimal entropic potentials for this cost, with corresponding coupling π_ε. Borrowing computations from e.g., Proposition 2 of Pooladian and Niles-Weed [2021], we can compute

$$\nabla f_\varepsilon(x) = \int B(x-y)\mathrm{d}\pi_\varepsilon^x(y) = Bx - B\int y \mathrm{d}\pi_\varepsilon^x(y)\,,$$

where $\pi_\varepsilon^x(y)$ is the conditional entropic coupling (given x). The proof concludes by taking the expression of the MBO estimator and expanding:

$$T_\varepsilon(x) = x - (\nabla h^*) \circ (\nabla f_\varepsilon(x)) = x - B^{-1}\left(Bx - B\int y\mathrm{d}\pi_\varepsilon^x(y)\right) = \int y \mathrm{d}\pi_\varepsilon^x(y)\,,$$

which is the definition of the barycentric projection of π_ε for a given x. □

Lemma 2 (Pre-conditioning of MBO). *Let T_ε be the MBO estimator between μ and ν for the cost $h = \frac{1}{2}\ell^2 + \gamma\tau_{A^\perp}$. Let W be denoted as in Proposition 3. Then the MBO estimator is written as*

$$T_\varepsilon = W^{-1} \circ \tilde{T}_\varepsilon \circ W\,, \tag{25}$$

where $\tilde{T}_\varepsilon$ is the barycentric projection between $W_\sharp\mu$ and $W_\sharp\nu$.

Proof. The proof here is similar to Proposition 3, which we outline again for completeness. As before, we are interested in solutions to the optimization problem

$$\min_{\pi\in\Gamma(\mu,\nu)} \int \tfrac{1}{2}\|Wx - Wy\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi\|\mu\otimes\nu)\,,$$

with optimal coupling $\pi^\star_\varepsilon$. Performing a change of variables $\pi' = (W\otimes W)_\sharp\pi$, we have

$$\min_{\pi'\in\Gamma(\mu',\nu')} \int \tfrac{1}{2}\|x - y\|^2 \mathrm{d}\pi'(x,y) + \varepsilon \mathrm{KL}(\pi'\|\mu'\otimes\nu')\,,$$

where $\mu' := W_\sharp\mu$ (and similarly for ν'), where now the optimizer reads $(\pi'_\varepsilon)^\star$. The two optimal plans are related as

$$\pi^\star = (W^{-1}\otimes W^{-1})_\sharp(\pi'_\varepsilon)^\star\,.$$

It was established in Lemma 1 that the MBO estimator $T^\star_\varepsilon$ is given by the barycentric projection

$$T^\star_\varepsilon(x) = \mathbb{E}_{\pi^\star_\varepsilon}[Y|X=x]\,.$$

Performing the change of variables $Y' = WY$ and $X' = WX$, we can re-write this as a function of π'_ε instead:

$$\begin{aligned} T^\star_\varepsilon(x) &= \mathbb{E}_{\pi^\star_\varepsilon}[Y|X=x] \\ &= \mathbb{E}_{(\pi'_\varepsilon)^\star}[W^{-1}Y'|W^{-1}X'=x] \\ &= W^{-1}\mathbb{E}_{(\pi'_\varepsilon)^\star}[Y'|X'=Wx] \\ &= W^{-1}\tilde{T}_\varepsilon(Wx)\,, \end{aligned}$$

where we identify $\tilde{T}_\varepsilon(\cdot) := \mathbb{E}_{(\pi'_\varepsilon)^\star}[Y'|X'=\cdot]$; this completes the proof. □

We are now ready to present the main proof.

Proof of Theorem 1. Let $T_{\varepsilon,n}$ denote the MBO estimator between samples from μ and ν, and let $\tilde{T}_{\varepsilon,n}$ denote the entropic map estimator from samples $\mu' := W_\sharp\mu$ and $\nu' := W_\sharp\nu$, where W has spectrum $0 < \lambda_{\min}(W) \le \lambda_{\max}(W) < +\infty$, where we have access to W since A is known.

Our goal is to establish upper bounds on

$$\|T_{\varepsilon,n} - T^\star\|^2_{L^2(\mu)} = \|W^{-1}\circ(\tilde{T}_{\varepsilon,n}\circ W - \tilde{T}\circ W)\|^2_{L^2(\mu)}\,.$$

Paying for constants that scale like $\lambda_{\max}(W^{-1})$, we have the bound

$$\|T_{\varepsilon,n} - T^\star\|^2_{L^2(\mu)} \lesssim_W \|\tilde{T}_{\varepsilon,n} - \tilde{T}\|^2_{L^2(\mu')}\,,$$

where we can now directly use the rates of convergence from [Pooladian and Niles-Weed, 2021, Theorem 3], as μ' satisfies our regularity assumptions under the conditions we have imposed on W. this completes the proof. □

C Gradient of Elastic Cost Loss

The gradient of the loss $\mathcal{L}$ in Definition 2 can be recovered through a simple aggregation of weighted gradients

$$\nabla\mathcal{L}(\theta) = \sum_{ij} [R(\theta)]_{ij}\,\nabla_\theta\,[P^\star(\theta)]_{ij} + [P^\star(\theta)]_{ij}\,\nabla_\theta\,[R(\theta)]_{ij}\,. \tag{26}$$

To write this formula in a more compact way, it is sufficient to notice that, adopting the convention that θ be a parameter in $\mathbb{R}^q$, and introducing an arbitrary vector $\omega \in \mathbb{R}^q$,

$$\begin{aligned}\langle \nabla \mathcal{L}(\theta), \omega \rangle &= \sum_{ij} [R(\theta)]_{ij} \langle \nabla_\theta [P^\star(\theta)]_{ij}, \omega \rangle + [P^\star(\theta)]_{ij} \langle \nabla_\theta [R(\theta)]_{ij}, \omega \rangle \\ &= \langle R(\theta), \left[\langle \nabla_\theta [P^\star(\theta)]_{ij}, \omega \rangle \right]_{ij} \rangle + \langle P^\star(\theta), \left[\langle \nabla_\theta [R(\theta)]_{ij}, \omega \rangle \right]_{ij} \rangle .\end{aligned}$$

The products of all coordinate wise gradients with ω is equivalent to the application of the Jacobians of R and $P^\star$. We write $J_\theta P^\star$ and $J_\theta R$ for these Jacobians, both being maps taking θ as input, and outputting a linear map $J_\theta R : \mathbb{R}^q \to \mathbb{R}^{n \times m}$, i.e. $J_\theta R(\theta)$ is a $n \times m$ matrix. As a consequence one has

$$\langle \nabla \mathcal{L}(\theta), \omega \rangle = \langle R(\theta), J_\theta P^\star(\theta)\, \omega \rangle + \langle P^\star(\theta), J_\theta R(\theta)\, \omega \rangle$$

because these maps are linear, one also has

$$\begin{aligned}\langle \nabla \mathcal{L}(\theta), \omega \rangle &= \langle J_\theta^T P^\star(\theta) R(\theta), \omega \rangle + \langle J_\theta^T R(\theta) P^\star(\theta), \omega \rangle \\ &= \langle J_\theta^T P^\star(\theta) R(\theta) + J_\theta^T R(\theta) P^\star(\theta), \omega \rangle\end{aligned}$$

which gives the identification given in the main text.

D Additional Details on Experiments

In § 6.2, and unlike Figure 3, we do not choose a predefined value for $\gamma^\star$, but instead select it with the following procedure: we start with a small value for $\gamma_0 = 0.1$, and increase it gradually, until a certain desirable criterion on these displacements goes above a threshold. To measure this, we first compute the (paired) matrix of displacements on a given sample,

$$D = [T_g^h(\mathbf{x}_i) - \mathbf{x}_i]_i \in \mathbb{R}^{n \times d}$$

We then consider ratio of singular values on p^* subspace (to select γ for $\|A^\perp \cdot\|_2^2$), writing σ for the vector of singular values of D, ranked in decreasing order, to compute

$$\text{sv-ratio}(\gamma) = \sum_{i=1}^{p} \sigma_i / \sum_i \sigma_i \in [0, 1] . \tag{27}$$

D.1 Synthetic experiments

In Algorithm 5, we outline a way to generate ground-truth data used in synthetic experiments mentioned in § 6.1 and Figure 3. For a given regularization strength γ and a regularizer τ_θ, we first sample a concave function g, along with the ground-truth parameters for the regularizer θ^*. We then sample n source points $\mathbf{X}$ from the standard normal distribution and push them through the ground-truth OT map to get n target points $\mathbf{Y}$, as described in Algorithm 4.

Algorithm 4 GT-SAMPLES$(\mathbf{X}; h, g_0)$

Inputs: points $\mathbf{X}$, elastic cost h, potential $g_0 : \mathbb{R}^d \to \mathbb{R}$
for $i, \mathbf{x}$ in `enumerate`$(\mathbf{X})$ **do**
  $\mathbf{y}_i \leftarrow T_{g_0}^h(\mathbf{x})$ ▷ Algorithm 2
end for
return: $\mathbf{Y} := [\mathbf{y}_1, \dots, \mathbf{y}_n]$

Algorithm 5 SYNTHETIC-OT-TASK(γ), on regularizers τ_θ.

Sample g concave function (e.g. minus ICNN).
Sample θ^* parameter ▷ when τ admits parameters
Sample n source points $\mathbf{X}$ (e.g. Gaussian).
Create n targets $\mathbf{Y} =$ GT-SAMPLES$(\mathbf{X}; \frac{1}{2}\ell_2^2 + \gamma \tau_{\theta^*}, g)$
return: $\mathbf{X}, \mathbf{Y}$, optionally GT parameters θ^*.

D.2 Experimental Procedure

Experiments in Section 6.1, *MBO on Synthetic Ground Truth Displacement*

- For both $\tau := \ell_1$ and $\tau_A := \|A^\perp \cdot\|^2$, run **SYNTHETIC-OT-TASK** (Algo. 5) to form samples $\mathbf{X}, \mathbf{Y}$ (store A^* for τ_A).

- Using these samples, benchmark MBO estimator informed by the cost structure: varying γ, using automatically scaled ε, and ground-truth parameter A^* when studying τ_{A^*}, for 10 random splits of $\mathbf{X}, \mathbf{Y}$ into $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$ and $\mathbf{X}^{\text{test}}, \mathbf{Y}^{\text{test}}$.
  - Run **MBO-ESTIMATOR** (Alg. 1 using regularization) on $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$, store $\text{MSE}_1 = \sum_i \|T_\varepsilon[\gamma, \tau_{\theta^*}, \varepsilon](\mathbf{x}_i^{\text{test}}) - \mathbf{y}_i^{\text{test}}\|^2$
  - Run **MBO-ESTIMATOR** (Alg. 1 *without* regularization) on $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$, store $\text{MSE}_2 = \sum_i \|T_\varepsilon[0, 0, \varepsilon](\mathbf{x}_i^{\text{test}}) - \mathbf{y}_i^{\text{test}}\|^2$
  - Report ratio $\text{MSE}_1/\text{MSE}_2$, showing MBO seeded with the right regularizer always outperforms original entropic map.

Experiments in Section 6.2, *Recovery of Ground-Truth Subspace Parameters in Elastic Costs*

- For $\tau_A := \|A^\perp \cdot\|^2$, run **SYNTHETIC-OT-TASK** (Algo. 5) to generate paired samples $\mathbf{X}, \mathbf{Y}$ in $\mathbb{R}^d$. Tune γ to have displacements concentrated (50%, 70%, 90%) in a subspace, as described in Appendix D. Recover A^* of dimension $p^* \times d$.
- Run **RECOVER-THETA** on $\mathbf{X}, \mathbf{Y}$ using arbitrary γ to output $\hat{A}$. Display Average recovery error $\|A^* - \hat{A}\hat{A}^T A^*\|_2^2/p^*$

Experiments in Section 6.3, *Learning Displacement Subspaces for Single-Cell Transport*

- Here data come from real measurements, $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^{m\times d}$, $n \neq m$. Split into $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$ and $\mathbf{X}^{\text{test}}, \mathbf{Y}^{\text{test}}$.
- Run **RECOVER-THETA** on $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$ using arbitrary γ and random initialization with varying $\hat{p}$. Output $\hat{A}$.
- Run **MBO-ESTIMATOR** (Alg. 1 using regularization) on $\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{train}}$. Form predictions $\tilde{\mathbf{y}}_i = T_\varepsilon[\gamma, \tau_{\hat{A}}, \varepsilon](\mathbf{x}_i^{\text{test}})$.
- Compute ℓ_2^2 Sinkhorn divergence between point cloud $(\tilde{\mathbf{y}}_i)_{i=1}^n$ and $(\mathbf{y}^{\text{test}_j})_{j=1}^m$.
- To benchmark (dotted-line in Fig. 5), run two steps above, but setting $\gamma = 0$ (no subspace regularization).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract describes the challenges underlying the selection of a cost in optimal transport theory. We advertise in that abstract a new way to benchmark such OT problems for a fairly wide class of exotic costs, and claim statistical consistency results and a new algorithm to learn that cost' parameters. All these claims are substantiated by simple experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is discussed in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: While theory is only a part of our contribution (our paper proposes also a new algorithm to identify parameters, experimented on synthetic and real data), we have provided, to the best of our best ability, the proofs of these theoretical results in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: By interfacing our code with the OTT-JAX toolbox, and using public datasets, we believe these resuts and claims can be fully reproduced, with some effort. We will release the entire codebase for experiments in coming weeks, as python notebooks/tutorials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Data is open. We provide code to use our tools. We do not share all of the more complex config files that were required to run all computations on GPU nodes.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we have provided such details in the description of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have for a part of our synthetic results. We also have error stds stored for Figure 4 but decided not to plot them given space constraints and the very low value obtained by *positive* quantities (i.e. a mean close to 0 means the std is also close to 0).

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Although no claim is made in terms of compute performance, the fairly small scale of the experiments allows to execute these runs on a single GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed these ethics guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss these impacts in the conclusion.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data resources have permissive licenses. We refer to the citations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.