Monge, Bregman and Occam: Interpretable Optimal Transport in High-Dimensions with Feature-Sparse Maps

Marco Cuturi¹, Michal Klein¹, Pierre Ablin¹

Abstract
Optimal transport (OT) theory focuses, among all maps \(T: \mathbb{R}^d \rightarrow \mathbb{R}^d\) that can morph a probability measure onto another, on those that are the “thriftiest”, i.e. such that the averaged cost \(c(x, T(x))\) between \(x\) and its image \(T(x)\) be as small as possible. Many computational approaches have been proposed to estimate such Monge maps when \(c\) is the \(\ell_2^2\) distance, e.g., using entropic maps (Pooladian and Niles-Weed, 2021), or neural networks (Makkuva et al., 2020; Korotin et al., 2020). We propose a new model for transport maps, built on a family of translation invariant costs \(c(x, y) := h(x - y)\), where \(h := \frac{1}{2}\| \cdot \|_2^2 + \tau\) and \(\tau\) is a regularizer. We propose a generalization of the entropic map suitable for \(h\), and highlight a surprising link tying it with the Bregman centroids of the divergence \(D_h\) generated by \(h\), and the proximal operator of \(\tau\). We show that choosing a sparsity-inducing norm for \(\tau\) results in maps that apply Occam’s razor to transport, in the sense that the displacement vectors \(\Delta(x) := T(x) - x\) they induce are sparse, with a sparsity pattern that varies depending on \(x\). We showcase the ability of our method to estimate meaningful OT maps for high-dimensional single-cell transcription data, in the \(34000\)-d space of gene counts for cells, without using dimensionality reduction, thus retaining the ability to interpret all displacements at the gene level.

1. Introduction
Learning a reliable mechanism to transfer observations from a source to a target probability measure is a fundamental task in machine learning. For such problems, optimal transport (OT) (Santambrogio, 2015) has emerged as a powerful toolbox, with both practical performance improvements and guiding theoretical principles. For instance, the computational approaches advocated in OT have been used to transfer knowledge across datasets in domain adaptation tasks (Courty et al., 2016; 2017), train generative models (Montavon et al., 2016; Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018), or realign datasets in natural sciences (Janati et al., 2019; Schiebinger et al., 2019).

High-dimensional Transport. OT has found success in many low-dimensional geometric domains (grids and meshes, graphs, etc...) for which the definition of a cost function is often straightforward and intuitive. This work focuses on the more challenging problems that arise when using OT on distributions in \(\mathbb{R}^d\), with \(d \gg 1\). In such problems, the ground cost \(c(x, y)\) between observations \(x, y\) is often chosen by default to be the \(\ell_2\) metric, or its square \(\ell_2^2\). However, that choice is rarely meaningful when used in large-\(d\) settings. This is due to the curse-of-dimensionality associated with OT estimation (Dudley et al., 1966; Weed and Bach, 2019) and the fact that the Euclidean distance loses its discriminative power as dimension grows. To mitigate this, practitioners rely on dimensionality reduction, either in two steps, before running OT solvers, using, e.g., PCA, a VAE, or a sliced-Wasserstein approach (Rabin et al., 2012; Bonneel et al., 2015); or jointly, by estimating both a projection and transport, e.g., on hyperplanes (Niles-Weed and Rigollet, 2022; Paty and Cuturi, 2019; Lin et al., 2020; Huang et al., 2021; Lin et al., 2021), lines (Deshpande et al., 2019; Kolouri et al., 2019), trees (Le et al., 2019) or more advanced featurizers (Salimans et al., 2018). However, an obvious drawback of these approaches is that transport maps estimated in reduced dimensions are hard to interpret in the original space (Muzellec and Cuturi, 2019).

Contributions. To target high \(d\) regimes, we introduce a radically different approach. We use the sparsity toolkit (Hastie et al., 2015; Bach et al., 2012) to build OT maps that are, adaptively to input \(x\), drastically simpler:

- We introduce a generalized entropic map (Pooladian and Niles-Weed, 2021) for translation invariant costs \(c(x, y) := h(x - y)\), where \(h\) is strongly convex. That entropic map \(T_{h, \epsilon}\) is defined almost everywhere (a.e.), and we show that it induces displacements \(\Delta(x) := T(x) - x\)
that can be cast as Bregman centroids relative to the Bregman divergence generated by $h$.

- When $h$ is an elastic-type regularizer, namely the sum of squared-Euclidean norm $\ell_2^2$ with a regularizer $\tau$, we show that such centroids are obtained using the proximal operator of $\tau$. Choosing for $\tau$ a sparsifying norm results therefore in sparse displacements $\Delta(x)$, with a sparsity pattern that depends locally on $x$, and that is controlled by the regularization strength set for $\tau$. To our knowledge, our formulation is the first in the computational OT literature that can produce feature-wise sparse OT maps.

- We apply our method to single-cell transcription data using two different sparsity-inducing proximal operators. We show that this approach succeeds in recovering simple and meaningful maps in extremely high-dimension.

**Not the Usual OT Sparsity.** Let us emphasize that the sparsity studied in this work is unrelated, and, in fact, orthogonal, to the numerous references to sparsity that are found in the computational OT literature. Such references arise because the most common formulation of the OT problem, the so-called Kantorovich formulation, involves computing an optimal coupling matrix between $n$ source and $m$ target points. When solved with linear program approach to solve that problem results in optimal coupling matrices that are sparse: any point in the source measure is only associated to one or a few points in the target measure. Such sparsity acts at the level of samples, and is a direct consequence of linear programming duality (Peyré and Cuturi, 2019, Proposition 3.4): the classical OT linear program (defined later in 2) results in $n \times m$ solutions that can only have up to $n + m - 1$ non-zero values, resulting in matrices that are largely filled with zeros. Such sparsity can also be enforced with additional regularization terms (Courty et al., 2016; Dessein et al., 2018; Blondel et al., 2018), or more advanced constraints (Liu et al., 2022). By contrast, sparsity in this work is unrelated to this so-called Kantorovich problem, and only occurs relative to the Monge problem (as in (1)), in the sense that it refers to features of the displacement vector $\Delta(x) \in \mathbb{R}^d$, when moving a given $x$. Our solution therefore results in displacements such that $\|\Delta(x)\|_0 \ll d$.

**Links to OT Theory with Degenerate Costs.** Starting with the seminal work by Sudakov (1979), who proved the existence of Monge maps for the original Monge problem, studying non-strongly convex costs with gradient discontinuities (Santambrogio, 2015, §3) has been behind many key theoretical developments (Ambrosio and Pratelli, 2003; Ambrosio et al., 2004; Evans and Gangbo, 1999; Trudinger and Wang, 2001; Carlier et al., 2010; Bianchini and Bardelloni, 2014). While these works have few practical implications, because they focus on the existence of Monge maps, constructed by stitching together OT maps defined pointwise, they did, however, guide our work in the sense that they shed light on the difficulties that arise from “flat” norms such as $\ell_1$. This has guided our focus in this work on elastic-type norms, which allow controlling the amount of sparsity through regularization strength, by analogy with the Lasso tradeoff where an $\ell_2^2$ loss is paired with an $\ell_1$ regularizer.

Figure 1. Plots of entropic map estimators $T_{h, \varepsilon}$, as defined in Prop. 4.2, to map a 2D measure supported on $(x^i)$ onto that supported on $(y^j)$, for various costs $h$. The displacements $\Delta(x) = T_{h, \varepsilon}(x) - x$ of unseen points are displayed as arrows. From left to right: standard $\ell_2^2$ norm, Elastic $\ell_1$, STVS, and $k$-support costs ($k = 1$). For each proposed cost, the regularization $\gamma$ is small on the top row and high on the bottom. Displacements are not sparse for the $\ell_2^2$ cost but become increasingly so as $\gamma$ grows, with a support that varies with input $x$. Note that Elastic $\ell_1$ and STVS tend to censor displacements as $\gamma$ grows, to the extent that they become null. In contrast, the $k$-support cost encourages sparsity but enforces displacements with at least $k$ non-zero values. See also Figure 2 for aggregate results.
Monge, Bregman and Occam

2. Background

2.1. Monge Kantorovich Duality.

Consider a translation-invariant cost function $c(x, y) := h(x - y)$, where $h : \mathbb{R}^d \to \mathbb{R}$. The Monge problem (1781) consists of finding, among all maps $T : \mathbb{R}^d \to \mathbb{R}^d$ that push-forward a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ onto $\nu \in \mathcal{P}(\mathbb{R}^d)$, the map which minimizes the average length (as measured by $h$) of its displacements:

$$
T^* := \arg \inf_{T_{\mu=\nu}} \int_{\mathbb{R}^d} h(x - T(x)) \, d\mu. \tag{1}
$$

Recovering Optimal Maps Using Duality. Problem (1) is notoriously difficult to solve directly; for instance, the set of admissible maps $T$ is not convex. Perhaps the most powerful result of OT theory lies in obtaining an optimal push-forward solution $T^*$ by casting Problem (1) differently, through a relaxation as a linear optimization problem. When relaxing the requirement that $x$ is mapped onto a single point $T(x)$, one can optimize instead over the space of couplings of $\mu, \nu$, namely on the set $\Pi(\mu, \nu)$ of probability distributions in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals $\mu, \nu$, to recover the Kantorovich problem:

$$
P^* := \arg \inf_{P \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c \, dP. \tag{2}
$$

If $T^*$ is optimal for Equation 1, then $\text{Id, } T^*$ is trivially an optimal coupling. Recovering a map $T^*$ from relaxed solutions can be achieved by focusing on the dual to (2):

$$
f^* , g^* \in \arg \sup_{f, g : \mathbb{R}^d \to \mathbb{R}} \int_{\mathbb{R}^d} f \, d\mu + \int_{\mathbb{R}^d} g \, d\nu , \tag{3}
$$

where $\forall x, y$ we write $(f \oplus g)(x, y) := f(x) + g(y)$.

Suppose that we can access such optimal dual potentials $(f^*, g^*)$ (this will be discussed in the next section). By a standard duality argument (Santambrogio, 2015, §1.3), if a pair $(x^0, y^0)$ lies in the support $\supp(P^*)$ of an optimal coupling $P^*$ solution to (2), the constraint for dual variables is saturated, i.e.,

$$
f^*(x^0) + g^*(y^0) = h(x^0 - y^0) ,
$$

Additionally, by a so-called $c$-concavity argument one has:

$$
g^*(y^0) = \inf_x h(x - y^0) - f^*(x) .
$$

Assuming $f^*$ is differentiable at $x^0$, combining these two results yields perhaps the most pivotal result in OT theory:

$$
(x^0, y^0) \in \supp(P^*) \iff \nabla f^*(x^0) \in \partial h(x^0 - y^0) , \tag{4}
$$

where $\partial h$ denotes the subdifferential of $h$, see, e.g. (Carlier et al., 2010). Let $h^*$ be the convex conjugate of $h$,

$$
h^*(y) := \sup_{x \in \mathbb{R}^d} (x, y) - h(x) .
$$

Depending on $h$, two cases arise in the literature:

- If $h$ is differentiable everywhere, strictly convex, one has:

$$
\nabla f^*(x^0) = \nabla h(x^0 - y^0) .
$$

Thanks to the identity \( \nabla h^* = (\nabla h)^{-1} \), one can uniquely characterize the only point \( y^0 \) to which \( x^0 \) is associated in the optimal coupling \( P^* \) as \( x^0 - \nabla h^*(\nabla f^*(x^0)) \). More generally one recovers therefore for any \( x \in \text{supp}(\mu) \):

\[
T^*(x) = x - \nabla h^* \circ \nabla f^*(x). \tag{5}
\]

The Brenier theorem (1991) is a particular case of that result, which states that when \( h = \frac{1}{2} \parallel \cdot \parallel_2^2 \), we have \( T(x) = x - \nabla f^*(x^0) \), since in that case \( \nabla h = \nabla h^* = (\nabla h)^{-1} = \text{Id} \), see (Santambrogio, 2015, Theo. 1.22).

- If \( h \) is “only” convex, then one recovers the sub-differential inclusion \( y^0 \in x^0 + h^*(\nabla f^*(x^0)) \) (Ambrosio et al., 2004)(Santambrogio, 2015, §3).

In summary, given an optimal dual solution \( f^* \) to Problem (3), one can use differential (or sub-differential) calculus to define an optimal transport map, in the sense that it defines (uniquely or as a multi-valued map) where the mass of a point \( x \) should land.

### 2.2. Bregman Centroids

We suppose in this section that \( h \) is strongly convex, in which case its convex conjugate is differentiable everywhere and gradient smooth. The generalized Bregman divergence (or B-function) generated by \( h \) (Telgarsky and Dasgupta, 2012; Kiwel, 1997) is,

\[
D_h(x|y) = h(x) - h(y) - \sup_{w \in \partial h(y)} (w, x - y).
\]

Consider a family of \( k \) points \( z^1, \ldots, z^m \in \mathbb{R}^d \) with weights \( p^1, \ldots, p^m > 0 \) summing to 1. A point in the set

\[
\arg \min_{z \in \mathbb{R}^d} \sum_j p^j D_h(z, z^j),
\]

is called a Bregman centroid (Nielsen and Nock, 2009, Theo. 3.2). Assuming \( h \) is differentiable at each \( z^j \), one has that this point is uniquely defined as:

\[
C_h \left( (z^j)_{j=1}^m, (p^j)_{j=1}^m \right) := \nabla h^* \left( \sum_{j=1}^m p^j \nabla h(z^j) \right). \tag{6}
\]

### 2.3. Sparsity-Inducing Penalties

To form relevant functions \( h \), we will exploit the following sparsity-inducing functions: the \( \ell_1 \) and \( \parallel \cdot \parallel_{ovk} \) norms, and a handcrafted penalty that mimics the thresholding properties of \( \ell_1 \) but with less shrinkage.

- For a vector \( z \in \mathbb{R}^d \), \( \parallel z \parallel_p := (\sum_{i=1}^d |z_i|^p)^{1/p} \). We write \( \ell_2 \) and \( \ell_2^2 \) for \( \parallel \cdot \parallel_2 \) and \( \parallel \cdot \parallel_2^2 \) respectively.

- We write \( \ell_1 \) for \( \parallel \cdot \parallel_1 \). Its proximal operator \( \text{prox}_{\gamma \ell_1}(z) = \text{ST}_\gamma(z) = (1 - \gamma/|z|)_+ \odot z \) is called the soft-thresholding operator.

- Schreck et al. (2015) propose the soft-thresholding operator with vanishing shrinkage (STVS),

\[
\tau_{\text{stvs}}(z) = \gamma^2 1_d^T \left( \sigma(z) + \frac{1}{2} - \frac{1}{2} e^{-2\sigma(z)} \right) \geq 0, \tag{7}
\]

with \( \sigma(z) := \text{asinh} \left( \frac{z}{\sqrt{d}} \right) \), and where all operations are element-wise. \( \tau_{\text{stvs}} \) is a non-convex regularizer, non-negative thanks to (our) addition of \( + \frac{1}{2} \), to recover a non-negative quantity that cancels if and only if \( z = 0 \). Schreck et al. show that the proximity operator \( \text{prox}_{\tau_{\text{stvs}}} \), written STVS for short, decreases the shrinkage (see Figure 3) observed with soft-thresholding:

\[
\text{STVS}_\gamma(z) = (1 - \gamma^2 / |z|^2)_{+} \odot z. \tag{8}
\]

The Hessian of \( \tau_{\text{stvs}} \) is a diagonal matrix with values \( \frac{1}{2} |z| / \sqrt{z^2 + \gamma^2 - \frac{1}{2}} \) and is therefore lower-bounded (with positive-definite order) by \( -\frac{1}{2} I_d \).

- Let \( G_k \) be the set of all subsets of size \( k \) within \( \{1, \ldots, d\} \). Argyriou et al. (2012) introduces the \( k \)-overlap norm:

\[
\parallel z \parallel_{ovk} = \min \left\{ \sum_{I \in G_k} \parallel v_I \parallel_2 \mid \text{supp}(v_I) \subset I, \sum_{I \in G_k} v_I = z \right\}.
\]

For any vector \( z \in \mathbb{R}^d \), we write \( z^+ \) for the vector composed with all entries of \( z \) sorted in a decreasing order. This formula can be evaluated as follows to exhibit a \( \ell_1/\ell_2 \) norm split between the \( d \) variables in a vector:

\[
\parallel z \parallel_{ovk}^2 = \sum_{i=1}^{d-k-1} (\parallel z \parallel_1^+)^2 + \left( \sum_{i=k-r}^d |z_i|^2 \right)^2 / (r + 1)
\]

where \( r \leq k - 1 \) is the unique integer such that

\[
|z_i|^2_{k-r-1} \leq \sum_{i=k-r}^d |z_i|^2 < \sum_{i=k-r}^d |z_i|^2_{k-r-1}.
\]
3. Generalized Entropic-Bregman Maps

**Generalized Entropic Potential.** When $h$ is the $\ell_2^2$ cost, and when $\mu$ and $\nu$ can be accessed through samples, i.e., $\hat{\mu}_m = \frac{1}{m} \sum_i \hat{\delta}_{x_i}, \hat{\nu}_m = \frac{1}{m} \sum_j \hat{\delta}_{y_j}$, a convenient estimator for $f^*$ and subsequently $T^*$ is the entropic map (Poodian and Niles-Weed, 2021; Rigollet and Stromme, 2022). We generalize these estimators for arbitrary costs $h$. Similar to the original approach, our construction starts by solving a dual entropy-regularized OT problem. Let $\epsilon > 0$ and write $K_{ij} = [\exp(-h(x_i^j - y^j)) / \epsilon]_{ij}$ the kernel matrix induced by cost $h$. Define (up to a constant):

$$ f^*, g^* = \arg \max_{f \in \mathbb{R}^m, g \in \mathbb{R}^m} \langle f, h \rangle + \langle g, \frac{1}{m} \rangle - \epsilon \langle f, K \bar{c} \rangle. \quad (9) $$

Problem (9) is the regularized OT problem in dual form (Peyré and Cuturi, 2019, Prop. 4.4), an unconstrained concave optimization problem that can be solved with the Sinkhorn algorithm (Cuturi, 2013). Once such optimal vectors are computed, estimators $f_{\epsilon}, g_{\epsilon}$ of the optimal dual functions $f^*, g^*$ of Equation 3 can be recovered by extending these discrete solutions to unseen points $x, y$:

$$ f_{\epsilon}(x) = \min_{\epsilon} (h(x - y^j) - g^*_j), \quad (10) $$

$$ g_{\epsilon}(y) = \min_{\epsilon} (h(x^i - y) - f^*_i), \quad (11) $$

where for a vector $u$ or arbitrary size $s$ we define the log-sum-exp operator as $\min_{\epsilon}(u) := -\epsilon \log(\frac{1}{s} \sum_i e^{-u(i) / \epsilon})$.

**Generalized Entropic Maps.** Using the blueprint given in Equation 4, we use the gradient of these dual potential estimates to formulate maps. Such maps are properly defined on a subset of $\mathbb{R}^d$ defined as follows:

$$ \Omega_{\hat{\nu}_m}(h) := \{ x | \forall j \leq m, \nabla h(x - y^j) \text{ exists.} \} \subset \mathbb{R}^d. \quad (12) $$

However, because a convex function is a.e. differentiable, $\Omega_{\hat{\nu}_m}(h)$ has measure 1 in $\mathbb{R}^d$. With this, $\nabla f_{\epsilon}$ is properly defined for $x$ in $\Omega_{\hat{\nu}_m}(h)$, as:

$$ \nabla f_{\epsilon}(x) = \frac{1}{m} \sum_{j=1}^m p^j(x) \nabla h(x - y^j), \quad (13) $$

using the $x$-varying Gibbs distribution in the $m$-simplex:

$$ p^j(x) := \frac{\exp \left( - \left( h(x - y^j) - g^*_j \right) / \epsilon \right)}{\sum_{k=1}^m \exp \left( - \left( h(x - y^k) - g^*_k \right) / \epsilon \right)}. \quad (14) $$

One can check that if $h = \frac{1}{2} \ell_2^2$, Equation 13 simplifies to the usual estimator (Poodian and Niles-Weed, 2021):

$$ T_{2,\epsilon}(x) := x - \nabla f_{\epsilon}(x) = \sum_{j=1}^m p^j(x) y^j. \quad (15) $$

We can now introduce the main object of interest of this paper, starting back from Equation 5, to provide a suitable generalization for entropic maps of elastic-type:

**Definition 3.1.** The entropic map estimator for $h$ evaluated at $x \in \Omega_{\hat{\nu}_m}(h)$ is $x - \nabla h^* \circ \nabla f_{\epsilon}(x)$. This simplifies to:

$$ T_h(x) := x - C_h((x - y^j), (p^j(x)_j)) \quad (16) $$

**Bregman Centroids vs. $W_c$ Gradient flow.** To displace points, a simple approach consists of following $W_c$ gradient flows, as proposed, for instance, in (Cuturi and Doucet, 2014) using a primal formulation Equation 2. In practice, this can also be implemented by relying on variations in dual potentials $\nabla f_{\epsilon}$, as advocated in Feydy et al. (2019, §4). This approach arises from the approximation of $W_c(\hat{\mu}_m, \hat{\nu}_m$) using the dual objective Equation 3,

$$ S_{h,\epsilon} \left( \frac{1}{m} \sum \delta_{x_i}, \frac{1}{m} \sum \delta_{y_j} \right) = \frac{1}{m} \sum_i f_{\epsilon}(x_i) + \frac{1}{m} \sum_j g_{\epsilon}(y_j), $$

differentiated using the Danskin theorem. As a result, any point $x$ in $\mu$ is then pushed away from $\nabla f_{\epsilon}$ to decrease that distance. This translates to a gradient descent scheme:

$$ x \leftarrow x - \lambda \nabla f_{\epsilon}(x) $$

Our analysis suggests that the descent must happen relative to $D_h$, to use, instead, a Bregman update (here $\lambda = 1 - \lambda$):

$$ x \leftarrow -\nabla h^* (\lambda \nabla h(x) + \lambda \nabla h(x - \nabla h^* \circ \nabla f_{\epsilon}(x))) \quad (17) $$

Naturally, these two approaches are exactly equivalent as $h = \frac{1}{2} \ell_2^2$ but result in very different trajectories for other functions $h$ as shown in Figure 4.

4. Structured Monge Displacements

We introduce in this section cost functions $h$ that we call of elastic-type, namely functions with a $\ell_2^2$ term in addition to another function $\tau$. When $\tau$ is sparsity-inducing (minimized on sparse vectors, with kinks) and has a proximal operator in closed form, we show that the displacements induced by this function $h$ are feature-sparse.

4.1. Elastic-type Costs

By reference to (Zou and Hastie, 2005), we call $h$ of elastic-type if it is strongly convex and can be written as

$$ h(z) := \frac{1}{2} \| z \|^2 + \tau (z). \quad (18) $$
where \( \tau : \mathbb{R}^d \to \mathbb{R} \) is a function whose proximal operator is well-defined. Since OT algorithms are invariant to a positive rescaling of the cost \( c \), our elastic-type costs subsume, without loss of generality, all strongly-convex translation invariant costs with convex \( \tau \). They do also include useful cases arising when \( \tau \) is not, as for instance with \( \tau_{\text{stvs}} \).

**Proposition 4.1.** For \( h \) as in (18) and \( x \in \Omega_{\ell_{\infty}}(\tau) \) one has:

\[
T_{h,\epsilon}(x) := x - \text{prox}_x \left( x - \sum_{j=1}^m p_j(x) (y_j - \nabla \tau(x - y_j)) \right)
\]

(19)

**Proof.** The result follows from \( \nabla h^* = \text{prox}_r \). Indeed:

\[
h^*(w) = \sup_z w^T z - \frac{1}{2} \|z\|^2 - \tau(z)
\]

\[
= -\inf_z -w^T z + \frac{1}{2} \|z\|^2 + \tau(z)
\]

\[
= \frac{1}{2} \|w\|^2 - \inf_z \frac{1}{2} \|z - w\|^2 + \tau(z).
\]

Differentiating on both sides and using Danskin’s lemma, we gets the desired result by developing \( \nabla h \) and using the fact that the weights \( p_j(x) \) sum to 1.

\( \square \)

### 4.2. Sparsity-Inducing Functions \( \tau \)

We discuss in this section the three choices we introduced in §2 for proximal operators and their practical implications in the context of our generalized entropic maps.

1-Norm \( \ell_1 \). As a first example, we consider \( \tau(z) = \gamma \|z\|_1 \) in Equation 18. The associated proximal operator is the soft-thresholding operator \( \text{prox}_{\tau}(\cdot) = \text{ST}(\cdot, \gamma) \) mentioned in the introduction. We also have \( \nabla h(z) = z + \gamma \text{sign}(z) \) for \( z \) with no 0 coordinate. Plugging this in Equation 5, we find that the Monge map \( T_{\gamma_{\ell_1}}(x) \) is equal to

\[
x - \text{ST}_\gamma \left( x - \sum_{j=1}^m p_j(x) (y_j + \gamma \text{sign}(x - y_j)) \right),
\]

where the \( p_j(x) \) are evaluated at \( x \) using Equation 14.

Applying the transport consists in an element-wise operation on \( x \): for each of its features \( t \leq d \), one subtracts \( \text{ST}_\gamma \left( \sum_{j=1}^m p_j(x) \nabla h(x_t - y_t) \right) \). The only interaction between coordinates comes from the weights \( p_j(x) \).

The soft-thresholding operator sparsifies the displacement. Indeed, when for a given \( x \) and a feature \( t \leq d \) one has

\[
\left| x_t - \sum_{j=1}^m p_j(x)y_j + \gamma \text{sign}(x_t - y_j) \right| \leq \gamma,
\]

then there is no change on that feature: \( T_{\gamma_{\ell_1}}(x)_t = x_t \). That mechanism works to produce, locally, sparse displacements on certain coordinates. Another interesting phenomenon happens when \( x \) is too far from the \( y_j \)’s on some coordinates, in which case the transport defaults back to a \( \ell_2 \) average of the target points \( y_j \) that is seen when using the \( \ell_2 \) entropic estimator, with the important caveat that all weights \( p_j(x) \) rely in their computation on the \( \gamma_{\ell_1} \) regularization:

**Proposition 4.2.** If \( x \) is such that \( x_t \geq \max_j y_t \) or \( x_t \leq \min_j y_t \) then \( T_{\gamma_{\ell_1}}(x)_t = \sum_j p_j(x)y_j \).

**Proof.** For instance, assume \( x_t \geq \max_j y_t \). Then, for all \( j \), we have \( \text{sign}(x_t - y_j) = 1 \), and as a consequence \( \sum_{j=1}^m p_j(x) \nabla h(x - y_j)_t = x_t - \sum_j p_j(x)y_j + \gamma \). This quantity is greater than \( \gamma \), so applying the soft-thresholding gives \( \text{ST}_\gamma (\sum_{j=1}^m p_j(x) \nabla h(x - y_j)_t) = x_t - \sum_j p_j(x)y_j \), which gives the advertised result. Similar reasoning gives the same result when \( x_t \leq \min_j y_t \).

Interestingly, this property depends on \( \gamma \) only through the \( p_j(x) \)’s, and the condition that \( x_t \geq \max_j y_t \) or \( x_t \leq \min_j y_t \) does not depend on \( \gamma \) at all.

**Vanishing Shrinkage STVS:** The \( \ell_1 \) term added to form the elastic net has a well-documented drawback, notably for regression: on top of having a sparsifying effect on the displacement, it also shrinks values. This is clear from the soft-thresholding formula, where a coordinate greater than \( \gamma \) is reduced by \( \gamma \). This effect can lead to some “shortening” of displacement lengths in the entropic maps. We use the Soft-Thresholding with Vanishing Shrinkage (STVS) proposed by Schreck et al. (2015) to overcome this problem. The cost function is given by Equation 7, and its prox in Equation 8. When \( \|z\| \) is large, we have \( \text{prox}_{\tau_{\text{stvs}}}(z) = z + o(1) \), which means that the shrinkage indeed vanishes. Interestingly, even though the cost \( \tau_{\text{stvs}} \) is non-convex, it still has a proximal operator, and \( \frac{1}{2} \|\cdot\|^2 + \tau_{\text{stvs}} \) is \( \frac{1}{2} \)-strongly convex.

**k-Overlap:** The \( k \)-overlap norm offers the distinctive feature that its proximal operator selects anywhere between \( d \) small \( \gamma \) and \( k \) large \( \gamma \) non-zero variables, see Figure 1. Applying this proximal operator is,
cost class, TICost, fed into a Sinkhorn solver, to output a DualPotentials object in OTT-JAX\(^1\) (Cuturi et al., 2022). We define a class of regularized translation invariant cost functions, one that specifies both the regularizer \(\tau\)—to run Sinkhorn and evaluate the gradient of a dual potential estimator—and its proximal operator, to evaluate a map as in (19). We call such costs RegTICost.

5.1. Synthetic experiments.

Constant sparsity-pattern. We measure the ability of our method to recover a sparse transport map using a setting inspired by (Pooladian and Niles-Weed, 2021). Here \(\mu = U_{[0,1]}\). For an integer \(s < d\), we set \(\nu = T^\star_{s\mu}\), where the map \(T^\star_{s\mu}\) acts on coordinates independently with the formula \(T^\star_{s\mu}(x) = [\exp(x_1), \ldots, \exp(x_s), x_{s+1}, \ldots, x_d]\): it only changes the first \(s\) coordinates of the vector, and corresponds to a sparse displacement when \(s \ll d\). Note that this sparse transport plan is much simpler than the maps our model can handle since, for this synthetic example, the sparsity pattern is fixed across samples. Note also that while it might be possible to detect that only the first \(s\) components have high variability using a 2-step pre-processing approach, or an adaptive, robust transport approach (Paty and Cuturi, 2019), our goal is to detect that support in a one-shot, thanks to our choice of \(h\). We generate \(n = 1,000\) i.i.d. samples \(x^i\) from \(\mu\), and \(y^j\) from \(\nu\) independently; the samples \(y^j\) are obtained by first generating fresh i.i.d. samples \(\tilde{x}^j\) from \(\mu\) and then pushing them: \(y^j := T^\star_{s\mu}(\tilde{x}^j)\). We use our three costs to compute \(T_{h,\varepsilon}\) from these samples, and measure our ability to recover \(T^\star_{s\mu}\) from \(T_{h,\varepsilon}\) using a normalized MSE defined as \(\frac{1}{md} \sum_{i=1}^{n} \|T^\star_{s\mu}(x^i) - T_{h,\varepsilon}(x^i)\|^2\). We also measure how well our method identifies the correct support: for each sample, we compute the support error as \(\sum_{i=s+1}^{d} \Delta_i^2 / \sum_{i=1}^{d} \Delta_i^2\) with \(\Delta\) the displacement \(T_{h,\varepsilon}(x^i) - x^i\). This quantity is between 0 and 1 and cancels if and only if the displacement

\(^1\)https://github.com/ott-jax/ott
happens only on the correct coordinates. We then average this quantity overall the $x^t$. Figure 5 displays the results as $d$ varies and $s$ is fixed. Here, $\gamma_{stvs}$ performs better than $\ell_1$.

**x-dependent sparsity-pattern.** To illustrate the ability of our method to recover transport maps whose sparsity pattern is adaptive, depending on the input $x$, we extend the previous setting as follows. To compute $F_s(x)$, we compute first the norms of two coordinate groups of $x$: $n_1 = \sum_{i=1}^{s} x_i^2$ and $n_2 = \sum_{i=s+1}^{d} x_i^2$. Second, we displace the coordinate group with the largest norm: if $n_1 > n_2$, $F_s(x) = [\exp(x_1), \ldots, \exp(x_s), x_{s+1}, \ldots, x_d]$, otherwise $F_s(x) = [x_1, \ldots, x_s, \exp(x_{s+1}), \ldots, \exp(x_{2s}), x_{2s+1}, \ldots, x_d]$. Obviously, the displacement pattern depends on $x$. Figure 6 shows the NMSE with different costs when the dimension $d$ increases while $s$ and $n$ are fixed. As expected, we observe a much better scaling for our costs than for the standard $\ell_2$ cost, indicating that sparsity-inducing costs mitigate the curse of dimensionality.

### 5.2. Single-Cell RNA-seq data.

We validate our approach on the single-cell RNA sequencing perturbation data from (Srivatsan et al., 2020). After removing cells with less than 200 expressed genes and genes expressed in less than 20 cells, the data consists of 579, 483 cells and 34, 636 genes. The raw counts are normalized (per cell) and $\log(x + 1)$ scaled. We select the 5 drugs (Belinostat, Dacnostat, Givinostat, Hesperadin, and Quisinostat) out of 188 drug perturbations that are highlighted in the original data (Srivatsan et al., 2020) as showing a strong effect. We consider 3 human cancer cell lines (A549, K562, MCF7) to each of which is applied each of the 5 drugs. We use our four methods to learn an OT map from control to perturbed cells in each of these $3 \times 5$ scenarios. For each cell line/drug pair, we split the data into 10 non-overlapping 80%/20% train/test splits, keeping the test fold to produce our metrics.

**Methods.** We ran experiments in two settings, using the whole 34,000-d gene space and subsampling to the top 5k highly variable genes using SCANPY (Wolf et al., 2018). We consider entropic map estimators with the following cost functions and pre-processing approaches: $\frac{1}{2} \ell_2$ cost; the $\frac{1}{2} \ell_2$ cost on 50-d PCA space (PCA directions are recomputed on each train fold); Elastic with $\gamma_{\ell_1}$; Elastic with $\gamma_{stvs}$ cost. We vary $\gamma$ for these two methods. In these approaches we follow OTT-JAX strategy to set $\varepsilon$ to scale as 10% of the mean of the entries of the cost matrix. We did not use the $\|\cdot\|_{owk}$ norm because of memory challenges when handling such a high-dimensional dataset. For the non-PCA-based approaches, we can also measure their performance in PCA space by projecting their high-dimensional predictions onto the 50-d space. The $\varepsilon$ regularization parameter for all these approaches is set for each cost and experiment to 10% of the mean value of the cost matrix between the train folds of control and treated cells, respectively.

**Evaluation.** We evaluate methods using these metrics:

- the $\ell_2^2$-Sinkhorn divergence (using $\varepsilon$ to be 10% of the mean of pairwise $\ell_2^2$ cost matrix of treated cells) between transferred points (from test fold of control) and test points (from perturbed state); lower is better.
- Ranked biased overlap (Webber et al., 2010) with $p = 0.9$, between the 50 perturbation marker genes as computed on all data with SCANPY, and the following per-gene statistic, computed using a map as follows: average (on fold) expression of (predicted) perturbed cells from original control cells (this tracks changes in log-expression before/after predicted treatment); higher is better.
- Coefficient of determination ($R^2$) between the average ground-truth / predicted gene expression on the 50 perturbation markers (Lotfollahi et al., 2019); higher is better.

These results are summarized in Figure 7 across various costs, perturbations, and hyperparameter choices.

| Table 1. Per cell line, sample sizes of control + drug perturbation. |
|-------------------------|------------|------------|------------|------------|------------|------------|------------|
| A549  | 3274 | 558 | 703 | 669 | 436 | 475 |
| K562  | 3346 | 388 | 589 | 656 | 624 | 339 |
| MCF7  | 6346 | 1562 | 1805 | 1684 | 882 | 1520 |

**Conclusion.**

We consider structured translation-invariant ground costs $h$ for transport problems. After examining an entropic potential with such costs, we plugged it in Brenier’s approach to construct a generalized entropic map. We highlighted a surprising connection between that map and the Bregman centroids associated with the divergence generated by $h$, resulting in a more natural approach to gradient flows defined by $W_\varepsilon$, illustrated in a simple example. By selecting costs $h$ of elastic type (a sum of $\ell_2^2$ and a sparsifying term), we show that our maps mechanically exhibit sparsity, in the sense that they have the ability to only impact adaptively $x$ on a subset of coordinates. We have proposed two simple generative models where this property helps estimation and applied this approach to high-dimensional single-cell datasets where we show, at a purely mechanical level, that we can recover meaningful maps. Many natural extensions of our work arise, starting with more informative sparsity-inducing norms (e.g., group lasso), and a more general approach leveraging the Bregman geometry for more ambitious $W_\varepsilon$ problems, such as barycenters.
Figure 7. Top row: performance, for all 15 experiments, of the elastic-$\gamma \ell_1$ estimator vs. the $\ell_2^2$ entropic map. We consider 6 values for $\gamma$. Each of the $15 \times 6$ crosses denotes the mean, over 10 random 80%/20% splits of that cell line/drug experiment, of a quantity of interest. To facilitate reading, rather than reporting the $\gamma$ value, we report the average percentage of non-zero displacements (below $10^{-8}$ in absolute value) across all displaced points in that fold (yellow means 40% dense displacements, dark-blue displacements only happen on $\approx 5\%$ of genes). While all these maps are estimated in full genes space ($\approx 34k$), we provide a simplified measure of their ability to reconstruct the measures by computing a $\ell_2^2$-Sinkhorn divergence in PCA space. This picture shows that one can sparsify significantly $\ell_2^2$ maps and still get a similar reconstruction error. Next, we picture separately the $R^2$ (see text body for details) computed on marker genes on low ($10nM$) and high ($10\mu M$) dosages of the drug. For low dosages, inducing sparsity in displacements seems to help, whereas this may no longer be the case when the effect of perturbations becomes large. Finally, the RBO metric shows that sparsity does help to select marker genes based only on map estimation. Bottom row: Close up on Hesperadin/MCF7 and Givinostat/K562 experiments. For each, we provide the test $\ell_2^2$-Sinkhorn divergence in full gene space for 4 different baselines: The usual $\ell_2^2$ approach, and two ICNN baselines, computed either on 50-d PCA space, or on full space. We also report the proportion of non-zero displacements w.r.t $\gamma$.

References


Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycen-


Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust wasserstein distance
Monge, Bregman and Occam


