Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors

Geert-Jan Huizing, Laura Cantini, Gabriel Peyré

Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors

Geert-Jan Huizing [1 2] **Laura Cantini** [2] **Gabriel Peyré** [1]

Abstract

Defining meaningful distances between samples in a dataset is a fundamental problem in machine learning. Optimal Transport (OT) lifts a distance between *features* (the "ground metric") to a geometrically meaningful distance between *samples*. However, there is usually no straightforward choice of ground metric. Supervised ground metric learning approaches exist but require labeled data. In absence of labels, only ad-hoc ground metrics remain. Unsupervised ground metric learning is thus a fundamental problem to enable data-driven applications of OT. In this paper, we propose for the first time a canonical answer by simultaneously computing an OT distance between *samples* and between *features* of a dataset. These distance matrices emerge naturally as positive singular vectors of the function mapping ground metrics to OT distances. We provide criteria to ensure the existence and uniqueness of these singular vectors. We then introduce scalable computational methods to approximate them in high-dimensional settings, using stochastic approximation and entropic regularization. Finally, we showcase *Wasserstein Singular Vectors* on a single-cell RNA-sequencing dataset.

1. Introduction

Machine learning tasks like information retrieval and classification require a notion of distance between samples in a dataset $X \in \mathbb{R}^{n\times m}$. In particular, we will study the case of single-cell RNA-sequencing data (scRNA-seq).

[1]Département de mathématiques et applications de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure, Université PSL, 75005, Paris, France [2]Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, INSERM, Ecole Normale Supérieure, Université PSL, 75005, Paris, France. Correspondence to: Geert-Jan Huizing <huizing@ens.fr>, Gabriel Peyré <gabriel.peyre@ens.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

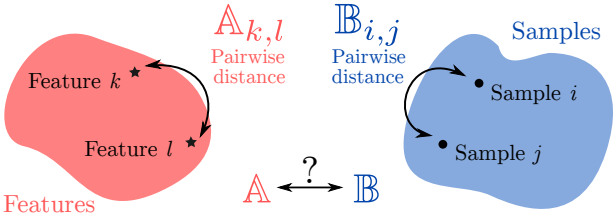

Figure 1. How to jointly define a distance $\mathbb{A}$ between features and a distance $\mathbb{B}$ between samples? Wasserstein Singular Vectors define natural Wasserstein distances $(\mathbb{A}, \mathbb{B})$ in an unsupervised manner.

Discrete histograms In scRNA-seq data, the sample $x_i \in \mathbb{R}_+^m$ represents the expression value of all genes in the i-th cell. This motivates two key assumptions in this paper: (i) samples are positive, which is natural when quantifying presence and quantity of physical objects (ii) samples can be normalized to discrete histograms, which is natural when the distribution over features (e.g. genes) is more important than the total mass. Indeed, gene expression in scRNA-seq data is usually normalized in some way as part of preprocessing (Luecken & Theis, 2019).

Optimal Transport distances Optimal Transport (OT) (Monge, 1781; Kantorovich, 1942) offers a geometrically meaningful distance between discrete probability distributions, and has recently emerged as a useful tool for machine learning applications (Frogner et al., 2015; Rolet et al., 2016). Contrarily to the Euclidean distance, OT does not compare distributions bin by bin. Instead, OT lifts a ground pairwise distance matrix $\mathbb{A} \in \mathbb{R}_+^{m\times m}$ between the m features to the "Wasserstein" OT distance between normalized samples $a_i := x_i/\|x_i\|_1$. It optimizes a transport plan $P \in \mathbb{R}^{m\times m}$ encoding the displacement of mass between the two m-dimensional histograms a_i, a_j.

$$\mathrm{W}_{\mathbb{A}}(a_i, a_j) := \min_{P\in\mathbb{R}_+^{m\times m}} \sum_{k,\ell} P_{k,\ell}\mathbb{A}_{k,\ell} \text{ s.t. } \begin{cases} P\mathbb{1}_m = a_i, \\ P^\top \mathbb{1}_m = a_j. \end{cases}$$

From supervised to unsupervised ground metric learning The crucial aspect of successful application of OT in ML is the design of a metric $\mathbb{A}$ which encodes the geometric relationships between features. In a supervised setting, one might take advantage of so-called *ground metric learning*

methods (Cuturi & Avis, 2014). In an unsupervised setting, one usually resorts to some ad-hoc choice of ground cost. For instance the Word Mover Distance (Kusner et al., 2015) uses Euclidean distances on Word2Vec embeddings (Mikolov et al., 2013). Similarly, the Gene Mover Distance (Bellazzi et al., 2021) uses Euclidean distances on Gene2Vec embeddings (Du et al., 2019).

In this work, we take a radically different route, by requiring that $\mathbb{A}_{k,\ell}$ is itself a Wasserstein distance between histograms b_k and $b_\ell \in \mathbb{R}^n_+$. The most intuitive case is to consider a_i (resp. b_k) to be the normalized row i (resp. column k) of a dataset $X \in \mathbb{R}_+^{n\times m}$. We will thus refer to a_i as a *sample* and b_k as a *feature*.

An intuitive way to motivate our method is to consider a bootstrapping approach. Given some initial (for instance random) metric $\mathbb{B}$, one can compute $\mathbb{A}_{k,\ell} = \mathrm{W}_{\mathbb{B}}(b_k, b_\ell)$. But there is no reason to stop here, and the metric $\mathbb{B}$ can be "improved" by updating $\mathbb{B}_{i,j} = \mathrm{W}_{\mathbb{A}}(a_i, a_j)$. By continuing this process of successively updating $\mathbb{A}$ and $\mathbb{B}$, one could hope to reach a limit where the pair of distance matrices $(\mathbb{A}, \mathbb{B})$ satisfies the following fixed point equation

Features — Distance between samples

$$\mathbb{A}_{k,\ell} = \tfrac{1}{\lambda}\mathrm{W}_{\mathbb{B}}(b_k, b_\ell), \quad \mathbb{B}_{i,j} = \tfrac{1}{\mu}\mathrm{W}_{\mathbb{A}}(a_i, a_j), \quad (1)$$

Distance between features — Samples

where $(\lambda, \mu) \in \mathbb{R}^2_+$ are scaling factors. This corresponds to casting ground metric learning as a non-linear singular vectors problem. In this work, we study theoretical properties (in particular existence and uniqueness) as well as the practical relevance of Wasserstein Singular Vectors for machine learning.

1.1. Previous works

Optimal Transport While the initial proposal of Monge (Monge, 1781) formulates the OT problem as an optimal matching problem, its modern and tractable formulation by Kantorovich (Kantorovich, 1942) is a linear program detailed in the Introduction. Besides its use to define matchings and couplings between distributions, the main feature of OT is that the transportation value induces a geometric distance on the space of probability distributions. This "Wasserstein" distance is thus parameterized by the underlying ground cost between pairs of points. We refer to the monographs (Villani, 2003; Santambrogio, 2015) for a detailed account of the theory of OT, and (Peyré et al., 2019) for its computational aspects. OT distances have been used for applications as diverse as image retrieval (Rubner et al., 2000), brain imaging (Gramfort et al., 2015; Janati et al., 2020), natural language processing (Kusner et al., 2015; Yurochkin et al., 2019), and generative models (Arjovsky et al., 2017; Tolstikhin et al., 2017). In recent years, many applications of OT to single-cell biology have been proposed (Hashimoto et al., 2016; Schiebinger et al., 2019; Bellazzi et al., 2021; Huizing et al., 2021; Tong et al., 2021).

Entropic regularization Entropic regularization of OT allows to scale to high-dimensional machine learning problems. It approximates OT distances using Sinkhorn's algorithm, which has quadratic complexity and streams well on GPU architectures. Entropic regularization was put forward in the seminal paper by Cuturi (Cuturi, 2013), who also emphasizes the smoothing effect, which is crucial when using Sinkhorn as a loss function to train deep learning models. Another benefit of this regularization is that it suffers less from the curse of dimensionality, as proved in (Genevay et al., 2019; Mena & Weed, 2019). This approach is also pivotal to scale our unsupervised metric learning method to tackle high-dimensional problems, for instance in genomics.

Metric Learning Metric learning is most often framed as the supervised problem of minimizing (resp. maximizing) the distance between points in a same (resp. different) class. Existing approaches are reviewed in (Kulis et al., 2012; Bellet et al., 2013). It is necessary to restrict the class of distances to make the problem tractable. The most common option is arguably to consider the class of Mahalanobis distances, which generalize the Euclidean distance and are equivalent to computing a vectorial embedding of the data points. See for instance (Xing et al., 2002; Weinberger et al., 2006; Davis & Dhillon, 2008). One can apply these methods for histogram data, or use instead of Euclidean distances more adapted discrepancies on the simplex, such as Chi-squared (Noh, 2012; Yang et al., 2015) and geodesic distances (Le & Cuturi, 2015). These methods however fail to capture the geometric nature of the problem, where histograms correspond to discrete distributions viewed as sums of localized Dirac masses.

OT Ground Metric Learning This geometry is leveraged in (Cuturi & Avis, 2014) by introducing the problem of supervised OT ground metric learning and developing a nearest-neighbor based algorithm to solve it. This approach is further refined in (Wang & Guibas, 2012), which drops the triangular inequality constraint (as we do in our approach). It is possible to restrict the class of ground metrics, for instance using Mahalanobis (Xu et al., 2018; Kerdoncuff et al., 2021) or geodesic distances (Heitz et al., 2020) to develop more efficient learning schemes. (Zen et al., 2014) simultaneously perform ground metric learning and matrix factorization, and this finds applications to NLP (Huang et al., 2016). Metric learning can also be performed through adversarial optimization, where the metric is maximized over to perform generative model training (Genevay et al., 2018b), discriminant analysis (Flamary et al., 2018) and to

define robust transportation distances (Paty & Cuturi, 2019; Niles-Weed & Rigollet, 2019). Note that when imposing only convex constraints, adversarial ground metric learning is a concave maximization problem which finds applications in the modeling of crowd congestion phenomena (Benmansour et al., 2010). Another related question is the inverse problem of estimating a ground cost from the observation of matchings or couplings (Galichon & Salanié, 2020; Stuart & Wolfram, 2020; Li et al., 2019; Paty & Cuturi, 2020). This supervised metric learning problem can be regularized using sparsity or low-rank constraints, as explained in (Dupuy et al., 2019; Carlier et al., 2020). Finally, "hierarchical OT" (Yurochkin et al., 2019; Abrishami et al., 2020) uses OT to define the ground cost of a matching problem, using an intermediate level of meta-features.

1.2. Contributions

Our main contribution is the introduction in Section 2 of *Wasserstein singular vectors* as the positive singular vectors of monotone homogeneous "distance maps". The associated theoretical contributions, Theorem 2.3 and Theorem 2.5, state conditions ensuring the existence and uniqueness of such singular vectors.

Our second set of contributions allows to scale the method to large datasets. We first introduce in Section 3 a stochastic algorithm similar in spirit to Projected Stochastic Gradient Descent. Theorem 3.1 guarantees a convergence rate of $\mathcal{O}(\log t/\sqrt{t})$ under certain conditions. We then explain in Section 4 how to scale and parallelize this method even further by leveraging entropic regularization through the Sinkhorn algorithm. Proposition 4.4 shows that in the large regularization limit, our method computes metrics associated to 1-D and 2-D embeddings along the leading principal component axes.

Section 5 demonstrates the potential of Wasserstein Singular Vectors compared to ad-hoc applications of Optimal Transport, by studying a single-cell RNA-sequencing dataset.

A Python package implementing all algorithms in this paper is available at github.com/gjhuizing/wsingular. Optimal Transport distances were computed using the open-source POT library (Flamary et al., 2021). Appendix A lists the experiments' computation times and resources.

Notations

We denote $\mathcal{D}_m \subset \mathbb{R}_+^{m\times m}$ the set of pairwise distance matrices. In other words, $\mathbb{A} \in \mathcal{D}_m$ if (i) $\mathbb{A}_{k,\ell} = 0 \iff k = \ell$, (ii) $\mathbb{A}_{k,\ell} \leq \mathbb{A}_{k,s} + \mathbb{A}_{s,\ell}$ (iii) $\mathbb{A}_{k,\ell} = \mathbb{A}_{\ell,k}$. Its closure $\bar{\mathcal{D}}_m := \{\mathbb{A} \in \mathbb{R}_+^{m\times m} \text{ s.t. } \mathbb{A} = \mathbb{A}^\top, \mathrm{diag}(\mathbb{A}) = 0\}$ is a set of pseudo-distances.

2. Unsupervised Wasserstein Metric Learning

This section introduces the singular vectors of the Wasserstein distance map. Fortunately, this non-linear singular vector problem enjoys many desirable properties.

2.1. Wasserstein Singular Vectors

Wasserstein distance map The following map lifts a ground metric $\mathbb{A} \in \mathcal{D}_m$ to a pairwise distance matrix $\Phi_A(\mathbb{A}) \in \mathcal{D}_n$. A norm R operates as a regularizer to enforce strict positivity of the computed distances.

$$\Phi_A(\mathbb{A})_{i,j} := \mathrm{W}_{\mathbb{A}}(a_i, a_j) + \tau\|\mathbb{A}\|_\infty R(a_i - a_j) \qquad (2)$$

The map $\mathbb{B} \in \mathcal{D}_n \mapsto \Phi_B(\mathbb{B}) \in \mathcal{D}_m$ is defined similarly.

Role of regularization τ Let us insist that in practice our method can be applied in the unregularized setting $\tau = 0$, but some of the theoretical claims require $\tau > 0$. Other types of regularization could be considered, for instance a matrix with zeros on the diagonal and ones elsewhere.

Wasserstein singular vectors With this notation, our ground metric learning solves for a pair $\mathbb{A} \in \mathcal{D}_m$ and $\mathbb{B} \in \mathcal{D}_n$ of *Wasserstein singular vectors* satisfying

$$\exists(\lambda, \mu) \in (\mathbb{R}_+^*)^2 \text{ s.t. } \Phi_B(\mathbb{B}) = \lambda\mathbb{A},\ \Phi_A(\mathbb{A}) = \mu\mathbb{B}, \quad (3)$$

which corresponds to (1) when $\tau = 0$. The case $m = n$ and $A = B$ corresponds to the computation of an eigenvector $\mathbb{A} = \mathbb{B}$ of Φ_A with eigenvalue $\lambda = \mu$.

Power iterations algorithm The de-facto standard algorithm to extract singular vectors are "power iterations"

$$\mathbb{A}_{t+1} := \frac{\Phi_B(\mathbb{B}_t)}{\|\Phi_B(\mathbb{B}_t)\|_\infty}, \quad \mathbb{B}_{t+1} := \frac{\Phi_A(\mathbb{A}_{t+1})}{\|\Phi_A(\mathbb{A}_{t+1})\|_\infty}. \quad (4)$$

The complexity of performing a single power iteration is $\mathcal{O}(n^2m^2(n\log(n) + m\log(m)))$, since the computation of a single Wasserstein distance in $\mathbb{R}_+^n$ is $\mathcal{O}(n^3\log n)$ (Bonneel et al., 2011). To cope with large scale datasets, we propose in Section 3 and 4 to use stochastic optimization and entropic regularization.

Remark 2.1. *A remarkable property of this algorithm is that, in cases where the singular vector is unique (which is observed in practice and proved below for large τ), even if the initialization $\mathbb{B}_{t=0}$ is chosen arbitrarily, it converges toward a distance matrix (so in particular it satisfies the triangular inequality at convergence).*

2.2. Theoretical Properties

Properties of the Wasserstein distance map Non-linear singular vectors problems are notoriously difficult to study.

Fortunately, as explained in the Proposition 2.2, the Wasserstein distance map is a so-called topical mapping (i.e. positive and monotone) (Lemmens & Nussbaum, 2012). These mappings can be thought as non-linear generalizations of Markov chains. Problem (3) is thus an instance of a non-linear Perron-Frobenius problem, for which, contrarily to generic problems, existence and uniqueness of positive solutions is in general rather the rule than the exception. This explains in large part the practical success of our approach.

The following proposition lists some useful properties of the Wasserstein distance map Φ_A.

Proposition 2.2. *(i) Φ_A is positively 1-homogeneous and monotone. (ii) Φ_A is continuous on $\mathcal{D}_m$ (iii) Φ_A is $(1+2\tau k_R)$-Lipschitz on $\mathcal{D}_m$ with regards to the $\|\cdot\|_\infty$ norm, where the constant $k_R > 0$ is such that $R(\cdot) \leq k_R\|\cdot\|_1$.*

Proof. (i) 1-homogeneity and monotony of Φ_A follows from the definition. (ii) Note that Φ_A is a vector-valued concave function (each coordinate being an infimum of linear forms) and hence is continuous on $\mathbb{R}^{m\times m}$. Actually, as we now show, it is Lipschitz for ℓ^∞. (iii) Let us prove that Φ_A is Lipschitz continuous for the ℓ^∞ norm on $\mathcal{D}_m$. Firstly, since $R(a_i - a_j) \leq k_R\|a_i - a_j\|_1$ and $|\|\mathbb{A}\|_\infty - \|\mathbb{A}'\|_\infty| \leq \|\mathbb{A} - \mathbb{A}'\|_\infty$, we have

$$|\Phi_A(\mathbb{A})_{i,j} - \Phi_A(\mathbb{A}')_{i,j}| \leq$$
$$|\mathrm{W}_{\mathbb{A}}(a_i, a_j) - \mathrm{W}_{\mathbb{A}'}(a_i, a_j)| + 2\tau k_R\|\mathbb{A} - \mathbb{A}'\|_\infty$$

Secondly, with $\Gamma(a, a')$ the set of valid couplings,

$$\begin{aligned}|\mathrm{W}_{\mathbb{A}}(a, a') - \mathrm{W}_{\mathbb{A}'}(a, a')| \\ = |\min_{P\in\Gamma(a,a')}\langle P, \mathbb{A}\rangle - \min_{P'\in\Gamma(a,a')}\langle P', \mathbb{A}'\rangle| \\ \leq \|\mathbb{A} - \mathbb{A}'\|_\infty.\end{aligned}$$

Indeed, $|\min(u) - \min(v)| \leq \max|u - v|$ and $\|P\|_1 = 1$. So Φ_A is $(1 + 2\tau k_R)$-Lipschitz. □

Existence of singular vectors The following proposition ensures the existence of positive (i.e. true distances) singular vectors. Its proof can be found in Appendix B. Note that the ℓ^∞ bound of Proposition 2.2 implies that all singular values are smaller than $1 + 2\tau k_R$.

Theorem 2.3. *When $\tau > 0$, there exist positive singular vectors $(\mathbb{A}, \mathbb{B}) \in \mathcal{D}_m \times \mathcal{D}_n$ solving the problem (3).*

Existence in the case $\tau = 0$ Extending Theorem 2.3 to the unregularized case is an open problem, and is out-of-reach using classical non-linear Perron-Frobenius theorems such as (Akian et al., 2016; 2018), which do not apply. The following remark exhibits solutions for a special case of the unregularized problem.

Remark 2.4 (Block-diagonal matrices)**.** *Let us consider the case $\tau = 0$ and a dataset $X = \mathrm{diag}(X_p)$, a block-diagonal matrix where $X_p \in \mathbb{R}_+^{n_p\times m_p}$. Let A and B be its normalizations along rows and columns respectively. Then all matrices $(\mathbb{A}, \mathbb{B}) \in \bar{\mathcal{D}}_m \times \bar{\mathcal{D}}_n$ of the form $\mathbb{A} = (c_{p,q}\mathbb{1}_{m_p\times m_q})_{p,q}$ and $\mathbb{B} = (c_{p,q}\mathbb{1}_{n_p\times n_q})_{p,q}$ for the same $c_{p\neq q} \in \mathbb{R}_+^*$ and $c_{p,p} = 0$ are dominant singular vectors with associated singular value 1. Indeed, the optimal transport plans for these ground costs also follow this block structure.*

Uniqueness of singular vectors If the dataset is too sparse and $\tau = 0$, one cannot hope to have uniqueness of the leading singular vectors. In fact, the previous remark exhibits an infinity of singular vectors when X is block-diagonal. It does not seem obvious to guarantee uniqueness by an a priori condition depending only on (A, B). The following proposition gives an a posteriori way to check the uniqueness of singular vectors inside $\mathcal{D}_m \times \mathcal{D}_n$. In the numerical simulations of Section 2.3, we checked *a posteriori* that the computed singular vectors are indeed unique.

Theorem 2.5. *Let $(\mathbb{A}, \mathbb{B}) \in \mathcal{D}_m \times \mathcal{D}_n$ a pair of Wasserstein Singular Vectors. We consider $P(a_i, a_j)$ and $P(b_k, b_\ell)$ optimal coupling solutions of the OT problems for the costs $\mathbb{A}$ and $\mathbb{B}$ respectively. These optimal couplings induce a graph on $\{(i, j), (k, \ell)\}$ by linking*

$$(k, \ell) \to (i, j) \text{ when } P(a_i, a_j)_{k,\ell} > 0$$
$$(i, j) \to (k, \ell) \text{ when } P(b_k, b_\ell)_{i,j} > 0.$$

If there exist optimal couplings such that this graph is strongly connected, then $(\mathbb{A}, \mathbb{B})$ are the unique Wasserstein singular vectors.

Proof. Let us consider $\Phi : (\mathbb{A}, \mathbb{B}) \mapsto (\Phi_B(\mathbb{B}), \Phi_A(\mathbb{A}))$ which maps $\mathcal{D}_m \times \mathcal{D}_m$ to itself. We use Theorem 7.5 of (Akian et al., 2016). It requires that the semi-differential of Φ at $(\mathbb{A}, \mathbb{B})$ has itself a unique positive eigenvector in $\mathcal{D}_m \times \mathcal{D}_n$. In fact, this eigenvector is also $(\mathbb{A}, \mathbb{B})$.

From the envelope theorem, upper-gradients of the concave function $\mathbb{A} \mapsto \mathrm{W}_{\mathbb{A}}(a_i, a_j)$ are the elements $P(a_i, a_j)$ (so if the optimal coupling is unique, then this map is differentiable). The semi-differential Ψ of Φ is then a block anti-diagonal matrix defining the graph detailed in the statement of the theorem. The (linear) Perron-Frobenius theorem for positive linear operators ensures the existence of a unique positive eigenvector of Ψ if this graph is connected. □

Convergence of power iterations In the case of linear positive maps, Perron-Frobenius theory ensures the convergence of (4) toward the unique positive singular vectors at a linear rate. Unfortunately, this result does not hold in general for the case of non-linear maps, and Φ_A is only non-expansive (and not necessarily contracting). The following

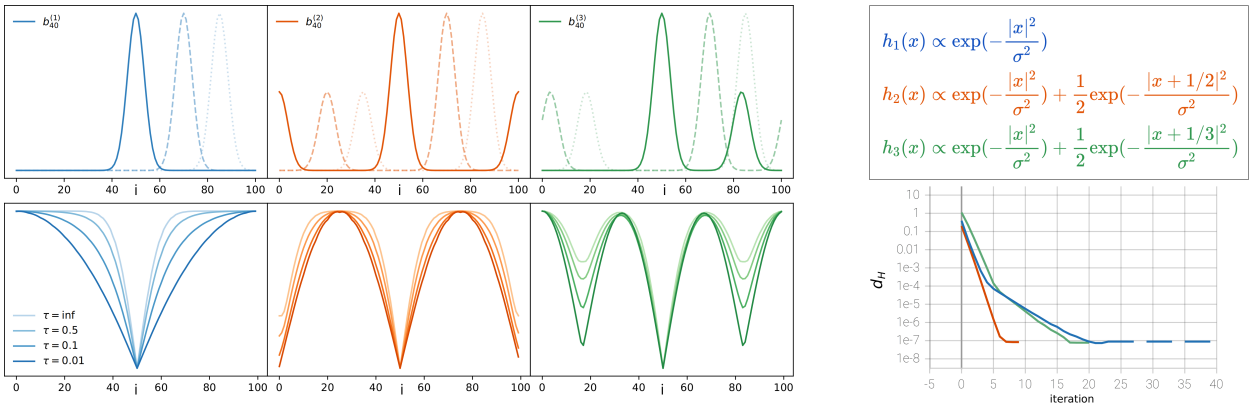

Figure 2. Illustration on the 1-D torus. (top, left) histograms whose translations form B_1, B_2, B_3 ; (bottom, left) distance maps c_1, c_2, c_3 associated to the singular vectors $\mathbb{B}_1$, $\mathbb{B}_2$, $\mathbb{B}_3$ for varying values of τ ; (top, right) functions h_1, h_2, h_3 generating the datasets ; (bottom, right) convergence rate of the power iterations for $\tau = 0.1$, according to the $d_{\mathcal{H}}$ metric.

proposition, proved in Appendix C, states that for large enough regularization, uniqueness and linear convergence are maintained.

Proposition 2.6. *For τ large enough, the singular vectors are unique and the power iterations (4) converge linearly for $\|\cdot\|_\infty$. When $\tau \to \infty$, the singular vectors converge to $\mathbb{A}_{k,\ell} \propto R(b_k - b_\ell)$ and $\mathbb{B}_{i,j} \propto R(a_i - a_j)$.*

The numerical simulations of Section 2.3 suggest that uniqueness and linear rates always hold in practical cases.

2.3. Numerical illustration on translated histograms

Generating translated histograms We generate three synthetic datasets X_1, X_2, $X_3 \in \mathbb{R}^{n\times m}$ by translating three different templates. We define the datasets by $[X_p]_{i,k} := h_p(i/n - k/m)$, and A_p (resp. B_p) is obtained by normalizing X_p along rows (resp. columns). By translational invariance of the problem, the singular vectors $\mathbb{A}_1$, $\mathbb{A}_2$ and $\mathbb{A}_3$ are of the form $(\mathbb{A}_p)_{k,l} = (c_p)_{k-l}$ where $c_p = (\mathbb{A}_p)_{0,\cdot}$ are periodic 1-D functions. The same argument applies to the singular vectors $\mathbb{B}_1$, $\mathbb{B}_2$ and $\mathbb{B}_3$. The templates h_1, h_2, h_3 are three different periodic functions on the 1-D torus (we use periodic boundary conditions). We use $n = 100$ samples and $m = 80$ features.

Wasserstein singular vectors We compute the Wasserstein singular vectors for different values of τ. Figure 2 displays the templates and the corresponding singular vectors $\mathbb{A}_1$, $\mathbb{A}_2$ and $\mathbb{A}_3$ obtained through power iterations. $\mathbb{B}_1$, $\mathbb{B}_2$ and $\mathbb{B}_3$ are symmetric and can be found in Appendix D. These results demonstrate that the learned metrics integrate geometrical properties (symmetries, multi-modalities, etc.) of the input datasets. For unimodal Gaussian-like distributions, the learned metric is close to $|\sin(i/n - j/n)|$, but exhibits non-monotonic behavior for multi-modal distributions.

Convergence rates Figure 2 also reports in logarithmic scale the convergence rate of power iterations according to the Hilbert metric $d_{\mathcal{H}}(\mathbb{B}, \mathbb{B}') := \|\log(\mathbb{B}/\mathbb{B}')\|_V$ where $\|Z\|_V := \max(Z) - \min(Z)$. This speed is always linear, suggesting that the maps Φ_{A_p} are contracting and that the singular vectors are unique (which is confirmed by running several initializations in $\mathcal{D}_n$, and through condition 2.5). The contractance rate (which is the slope of the error curves) is dependent on the geometry of the templates h_p. We also observed a steeper slope for larger values of τ.

3. Large Scale Stochastic Power Iterations

As n or m grows, the complexity of the power iterations (4) becomes prohibitive. In order to work around this issue we propose a stochastic power iteration scheme similar in spirit to stochastic gradient descent, which updates a single (or several if applied in a mini-batch setting) randomly chosen distance value at each step. This speeds up each iteration and leverages the correlations in the dataset.

Stochastic power iterations For some decreasing step size α_t and a scaling factors $\tilde{\lambda}_t, \tilde{\mu}_t > 0$, we define

$$\mathbb{A}_{t+1} := \Pi((1-\alpha_t)\mathbb{A}_t + \alpha_t \tilde{\mathbb{A}}_t),$$
$$\mathbb{B}_{t+1} := \Pi((1-\alpha_t)\mathbb{B}_t + \alpha_t \tilde{\mathbb{B}}_t),$$

$$\text{where} \quad (\tilde{\mathbb{B}}_t)_{i,j} := \begin{cases} \Phi_A(\mathbb{A}_t)_{i,j}/\tilde{\mu}_t \text{ if } (i,j) = (i_t, j_t), \\ (\mathbb{B}_t)_{i,j} \text{ otherwise.} \end{cases}$$

We define $\Pi(\mathbb{A}) := \mathbb{A}/\|\mathbb{A}\|_\infty$ and (i_t, j_t) is is drawn uniformly at random in $\{1, \dots, n\}^2$. It is the index updated at each step. $\tilde{\mathbb{A}}_t$ is computed by an analogous update rule, with $\tilde{\mu}_t$ replaced by $\tilde{\lambda}_t$.

Convergence of stochastic power iterations The following theorem, proved in Appendix E, guarantees that for a

large enough regularization parameter τ, these iterations converge to a pair of Wasserstein Singular Vectors. In practice, we observe that these iterations converge even for arbitrary small τ and for $\tilde{\lambda}_t = \tilde{\mu}_t = 1$.

Theorem 3.1. *For $\alpha_t = 1/\sqrt{t}$, for constant scaling factors $\tilde{\lambda}_t \leq \tau \min_{k\neq l} R(b_k - b_l)$ and $\tilde{\mu}_t \leq \tau \min_{i\neq j} R(a_i - a_j)$, and for τ large enough, the stochastic power iterations defined above converge to a pair $(\mathbb{A}, \mathbb{B}) \in \mathcal{D}_m \times \mathcal{D}_n$ of positive singular vectors with a convergence rate of $\mathcal{O}(\log(t)/\sqrt{t})$.*

Remark 3.2 (Adaptive selection of $\tilde{\lambda}_t$ and $\tilde{\mu}_t$)**.** *Tuning the values of the parameters $\tilde{\lambda}_t$ and $\tilde{\mu}_t$ is crucial to improving the convergence. Ideally, they should be as close as possible to the (unknown) singular values (λ, μ). Instead of fixing these scaling factors prior to running the algorithm, we propose using an estimation of the singular values. When using mini-batching, i.e. updating several indices $(i, j) \in \mathcal{I}$ at each iteration, one can use a least square estimate,*

$$\tilde{\mu}_{t+1} = (1 - \alpha_t)\tilde{\mu}_t + \alpha_t \frac{\sum_{i,j\in\mathcal{I}} \Phi_A(\mathbb{A}_t)_{i,j}(\mathbb{B}_t)_{i,j}}{\sum_{i,j\in\mathcal{I}} (\mathbb{B}_t)^2_{i,j}},$$

and similarly for $\tilde{\lambda}_{t+1}$. We found that in practice, $(\tilde{\lambda}_t, \tilde{\mu}_t)$ quickly converges to the singular values (λ, μ).

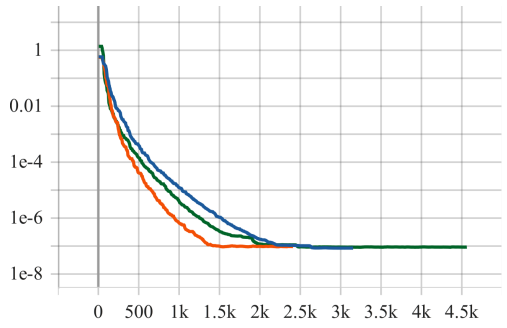

Figure 3. Convergence rates $d_{\mathcal{H}}(\mathbb{B}_t, \mathbb{B}_\infty)$ of stochastic updates.

Numerical illustration on translated histograms Figure 3 illustrates the convergence of stochastic power iterations in practice by comparing the approximated Wasserstein Singular Vectors for the synthetic experiments of Section 2.3 with the true singular vectors obtained using the non-stochastic power iterations (4). We used the approach outlined in Remark 3.2. As expected, the stochastic power iterations yield the same result as classical power iterations.

4. Parallelization With Entropic Regularisation

To further speed up the method, we propose to use the entropic regularization of Optimal Transport (Cuturi, 2013).

Sinkhorn's algorithm Entropic OT can be computed efficiently in $\mathcal{O}(n^2/\eta^2)$ using Sinkhorn's algorithm (detailed in Appendix F) at the expense of an approximation of order η (Altschuler et al., 2017). Beside speeding up the computation of OT, this enables embarrassingly parallel computations of the distance map on GPUs (Cuturi, 2013) and also reduces the curse of dimensionality which plagues OT (Genevay et al., 2018a).

Remark 4.1. *The theoretical guarantees listed above apply to entropic OT with Euclidean ground costs, but not necessarily to the setting described in this paper. However, in practice we observed that the benefits of Sinkhorn's algorithm were maintained in our situation.*

4.1. Sinkhorn divergence map

Sinkhorn divergence Entropic regularized OT is defined

$$\mathrm{W}^\varepsilon_{\mathbb{A}}(a_i, a_j) \coloneqq \min_{P\in\Gamma(a_i,a_j)} \langle P, \mathbb{A}\rangle + \varepsilon\|\mathbb{A}\|_\infty \langle P, \log P\rangle.$$

This quantity suffers from a bias, which is removed by using instead the Sinkhorn divergence (Genevay et al., 2018b)

$$\bar{\mathrm{W}}^\varepsilon_{\mathbb{A}}(a_i, a_j) \coloneqq \mathrm{W}^\varepsilon_{\mathbb{A}}(a_i, a_j) - \tfrac{1}{2}\left(\mathrm{W}^\varepsilon_{\mathbb{A}}(a_i, a_i) + \mathrm{W}^\varepsilon_{\mathbb{A}}(a_j, a_j)\right).$$

This debiasing is crucial to ensure that $\bar{\mathrm{W}}^\varepsilon_{\mathbb{A}}(a_i, a_i) = 0$, and it also reduces the approximation error to $|\bar{\mathrm{W}}^\varepsilon_{\mathbb{A}} - \bar{\mathrm{W}}_{\mathbb{A}}| \sim \varepsilon^2$ (Chizat et al., 2020).

Sinkhorn divergence map Similarly to the map defined in (2), we define the Sinkhorn divergence map as

$$\Phi^\varepsilon_A(\mathbb{A} \in \mathcal{D}_m)_{i,j} \coloneqq \bar{\mathrm{W}}^\varepsilon_{\mathbb{A}}(a_i, a_j) + \tau\|\mathbb{A}\|_\infty R(a_i - a_j),$$

and similarly for Φ^ε_B. It reduces to (2) when $\varepsilon = 0$.

Sinkhorn singular vectors By analogy with (3) we define *Sinkhorn singular vectors* as $\mathbb{A}, \mathbb{B}$ such that

$$\exists(\lambda, \mu) \in {\mathbb{R}^*_+}^2 \text{ s.t. } \Phi^\varepsilon_B(\mathbb{B}) = \lambda\mathbb{A},\ \Phi^\varepsilon_A(\mathbb{A}) = \mu\mathbb{B}, \quad (5)$$

Remark 4.2 (Positivity property)**.** *The map Φ^ε is 1-homogeneous and monotone, but it is unclear that it always maps onto positive matrices. It is proved in (Feydy et al., 2019) that it is the case if $e^{-\mathbb{A}/(\varepsilon\|\mathbb{A}\|_\infty)}$ is a positive kernel (i.e. has positive eigenvalues). While it is unclear that such a condition is maintained during power iterations, we observed numerically that it is still the case in practice. We show below that this is true in the limit $\varepsilon \to +\infty$.*

4.2. Connection with PCA when $\varepsilon \to \infty$

Maximum Mean Discrepancy limit For simplicity, let us consider the case $\tau = 0$. We show in proposition 4.3 below that when $\varepsilon \to \infty$, our method operates over the set of squared Euclidean matrices

$$\mathbb{A} \in \mathcal{K}_m \subset \mathcal{D}_m \iff \exists (u_k \in \mathbb{R}^d)_{k=1}^m, \mathbb{A}_{k,\ell} = \|u_k - u_\ell\|_2^2.$$

Note that these matrices can be equivalently defined as conditionally negative matrices with zero diagonal, see (Schölkopf & Smola, 2002).

Proposition 4.3. *One has* $\Phi_A^\infty : \mathcal{K}_m \to \mathcal{K}_n$ *where*

$$\Phi_A^\infty(\mathbb{A}) := \lim_{\varepsilon \to \infty} \Phi_A^\varepsilon(\mathbb{A}) = (-\tfrac{1}{2}\langle \mathbb{A}(a_k - a_\ell), a_k - a_\ell \rangle)_{k,\ell}.$$

This property shows that in the large ε limits Φ_A^∞ is actually a linear map which computes Maximum Mean Discrepancies (Gretton et al., 2012) (a.k.a. Euclidean distances between probability distributions).

Connection with PCA For the sake of simplicity in the exposition, let us now assume $A = B^\top$. In this case, (3) is a classical linear singular vectors problem. While in general, ensuring existence of positive singular vectors is non-trivial, the following proposition, proved in Appendix G, shows that this is the case for Φ_A^∞.

Proposition 4.4 (Connection with PCA)**.** *Let us denote* $\tilde{A} := A - A\mathbb{1}_m\mathbb{1}_m^\top/m$ *the centered matrix. For any pair* (u, v) *of singular vectors of* $\tilde{A}$ *with singular value* λ*,*

$$\mathbb{A} = ((v_k - v_\ell)^2)_{k,\ell} \in \mathcal{K}_m, \ \mathbb{B} = ((u_i - u_j)^2)_{i,j} \in \mathcal{K}_n$$

are singular vectors of $(\Phi_A^\infty, \Phi_B^\infty)$*, with singular value* $2\lambda^2$.

This proposition shows that for $\varepsilon = +\infty$ a set of positive singular vectors is obtained as simply squared Euclidean distances over 1-D principal component embeddings of the data. Entropic regularization thus draws a link between our novel set of OT-based metric learning techniques and classical dimensionality reduction methods. This frames Sinkhorn singular vectors as a well-posed problem regardless of the value of ε.

5. Metric Learning for Single-Cell Genomics

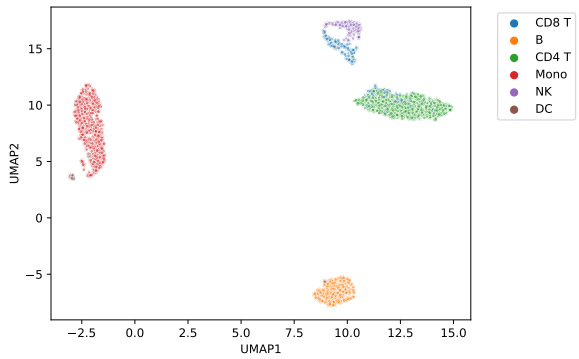

Figure 4. UMAP projection of the cells of a scRNA-seq dataset using the singular vector $\mathbb{B}$, with cells colored by cell type.

scRNA-seq data Single-cell RNA sequencing (*scRNA-seq*) is a high-throughput sequencing technology enabling the measurement of gene expression levels at single-cell resolution (Stegle et al., 2015). The analysis of scRNA-seq data has offered unprecedented insights in cellular heterogeneity and disease mechanisms (Tanay & Regev, 2017; Yuan et al., 2017). scRNA-seq data can be represented as a matrix of integer expression levels with cells on rows and genes on columns. One of the main uses of scRNA-seq is to identify cell populations through clustering or visualization. But these tasks rely on some notion of distance between cells. The most popular clustering and visualization tools, in particular Scanpy (Wolf et al., 2018) and Seurat (Stuart et al., 2019), rely on Euclidean distances on PCA embeddings of cells. Embeddings can also be provided by deep learning models like scVI (Gayoso et al., 2022). A good metric on the space of genes is also important because the phenotype of a cell is determined by the joint activity of all its expressed genes.

Optimal Transport distances between cells In order to take advantage of the biological relationships between genes, OT distances between cells have recently been proposed. The Gene Mover Distance (Bellazzi et al., 2021) is defined similarly to the Word Mover Distance (Kusner et al., 2015): the authors use as a ground cost the Euclidean distance between precomputed Gene2Vec (Du et al., 2019) embeddings. (Huizing et al., 2021) use a Sinkhorn divergence with a cosine distance between genes (i.e. vectors of cells) as a ground cost. In the present paper we compute OT distances using the Python package POT (Flamary et al., 2021).

Dataset A commonly analyzed scRNA-seq dataset is the "PBMC 3k" dataset produced by 10X Genomics, obtained through the function `pbmc3k` of Scanpy (Wolf et al., 2018). Details on preprocessing and cell type annotation are given in Appendix H. The processed dataset contains $m = 1030$ genes and $n = 2043$ cells, each belonging to one of 6 immune cell types: 'B cell', 'Natural Killer', 'CD4+ T cell', 'CD8+ T cell', 'Dendritic cell' and 'Monocyte'. The cell populations are heavily unbalanced. In addition, for each cell type we consider the set of canonical marker genes given by Azimuth (Hao et al., 2021), i.e. genes whose expression is characteristic of a certain cell type.

Evaluation We use the annotation on cells (resp. on marker genes) to evaluate the quality of distances between cells (resp. between marker genes). We report in Table 1 and Table 2 the Average Silhouette Width (ASW), computed using the function `silhouette_score` of Scikit-learn (Pedregosa et al., 2011). In addition, we visualize both of these distances using a 2-D UMAP projection (McInnes et al., 2018). We compare (i) Euclidean distances on PCA embeddings (ii) Euclidean distances on Kernel PCA embeddings using Scikit-learn's implementation with `kernel='rbf'` (iii) Euclidean distances on scVI (Gayoso et al., 2022) embeddings using default values (iv) Gene Mover Distance

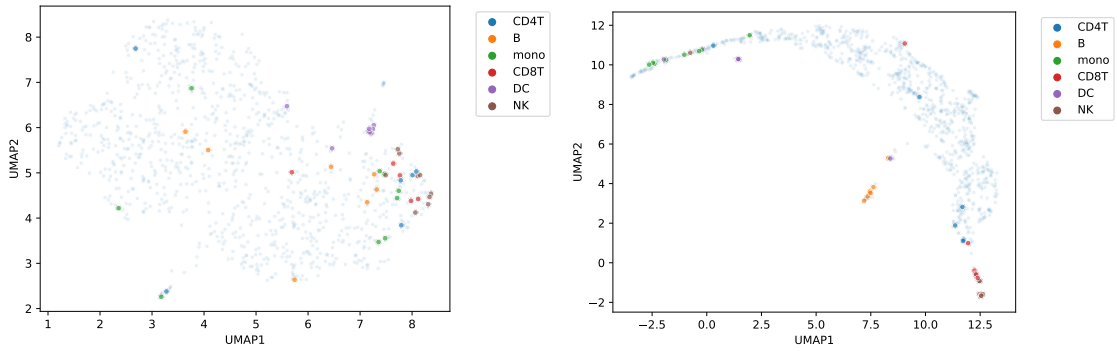

Figure 5. 2-D UMAP projection of marker genes, using the computed distances. Marker genes are colored by associated cell type, and other genes are faded out. Left, the Euclidean distance on Gene2Vec (Du et al., 2019) embeddings. Right, the singular vector $\mathbb{A}$.

(v) Sinkhorn divergence ($\varepsilon = .1$) with a cosine distance between genes as a ground cost (vi) Wasserstein Singular Vectors ($\tau = 0.001, \varepsilon = 0.1$), approached by 15 power iterations, which we found in this case to lead to better results than the stochastic power iterations. Note that for large regularisation ε, as shown in Section 4, the Wasserstein Singular Vectors are themselves (squared) Euclidean distances on PCA embeddings.

Results The results in Table 1 and Table 2 suggest that our method improves over all considered baselines. Figure 4 and Figure 5 shows the UMAP projection of the cells and the genes in the dataset. The Wassersein Singular Vectors clearly outperform the other metrics in terms of Average Silhouette Widths, both in the context of cells and of genes. Interestingly, in the case of marker genes we outperform the Euclidean distance on Gene2Vec embeddings, which are meant to contain "semantic" information about genes. These scores are also validated by the UMAP projection, where cells and marker genes cluster according to cell type. These results motivate further research in the biological implications of Wasserstein Singular Vectors. Let us highlight that we compute distances between cells or genes, but that we do not produce embeddings like PCA or scVI. In addition, scVI can handle complex tasks like the removal of unwanted sources of variation (Gayoso et al., 2022) which we do not consider in this article.

Table 1. Average Silhouette Width for cells

Method	ASW
PCA / ℓ^2	0.238
Kernel PCA / ℓ^2	0.241
scVI embedding / ℓ^2	0.168
Sinkhorn	0.003
Gene Mover Distance	0.066
WSV (ours)	**0.348**

Table 2. Average Silhouette Width for marker genes

ℓ^2	Gene2Vec / ℓ^2	WSV (ours)
-0.005	0.0186	**0.136**

6. Conclusion and Perspectives

Wasserstein Singular Vectors define a pair of "intrinsic" ground metrics associated to a given dataset. This elegantly solves the problem of unsupervised ground metric learning without resorting to ad hoc embeddings. Numerical results on single-cell RNA sequencing suggest that these metrics encode salient geometric structures of the data. This opens several avenues for future works, in particular an in-depth theoretical analysis when $\tau = 0$ and $\varepsilon > 0$. Our method can be extended to unbalanced optimal transport (Liero et al., 2015; Chizat et al., 2018), which has proved useful to increase the robustness of the metric for the analysis of biological sequencing datasets (Schiebinger et al., 2019). Lastly, our initial results regarding stochastic approximation of Wasserstein Singular Vectors would greatly benefit from further developments enabling better convergence rates.

Acknowledgements

We thank Stéphane Gaubert for very useful advice on nonlinear Perron-Frobenius theory.

This work was performed using HPC resources from GENCI-IDRIS [Grant 2021-AD011012285]. The project leading to this publication has received funding from the Agence Nationale de la Recherche (ANR) project scMOmix and Sanofi iTech Awards. The work of G. Peyré is supported by the European Research Council (ERC project NORIA) and by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Abrishami, T., Guillen, N., Rule, P., Schutzman, Z., Solomon, J., Weighill, T., and Wu, S. Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, 42(5):A3340–A3366, 2020.

Akian, M., Gaubert, S., and Nussbaum, R. Uniqueness of the fixed point of nonexpansive semidifferentiable maps. *Transactions of the American Mathematical Society*, 368 (2):1271–1320, 2016.

Akian, M., Gaubert, S., and Hochart, A. A game theory approach to the existence and uniqueness of nonlinear perron-frobenius eigenvectors. *arXiv preprint arXiv:1812.09871*, 2018.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1961–1971, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bellazzi, R., Codegoni, A., Gualandi, S., Nicora, G., and Vercesi, E. The gene mover's distance: Single-cell similarity via optimal transport. *arXiv preprint arXiv:2102.01218*, 2021.

Bellet, A., Habrard, A., and Sebban, M. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Benmansour, F., Carlier, G., Peyré, G., and Santambrogio, F. Derivatives with respect to metrics and applications: subgradient marching algorithm. *Numerische Mathematik*, 116(3):357–381, 2010.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pp. 1–12, 2011.

Carlier, G., Dupuy, A., Galichon, A., and Sun, Y. Sista: learning optimal transport costs under sparsity constraints. *arXiv preprint arXiv:2009.08564*, 2020.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. In *Proc. NeurIPS'20*, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Adv. in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, M. and Avis, D. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

Davis, J. V. and Dhillon, I. S. Structured metric learning for high dimensional problems. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 195–203, 2008.

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(1):7–15, 2019.

Dupuy, A., Galichon, A., and Sun, Y. Estimating matching affinity matrices under low-rank constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689, 2019.

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

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. POT: Python optimal transport, 2021.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Galichon, A. and Salanié, B. Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*, 2020.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., et al. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of Sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018a.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proc. AISTATS'18*, pp. 1608–1617, 2018b.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583. PMLR, 2019.

Gramfort, A., Peyré, G., and Cuturi, M. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pp. 261–272. Springer, 2015.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

Hashimoto, T., Gifford, D., and Jaakkola, T. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pp. 2417–2426. PMLR, 2016.

Heitz, M., Bonneel, N., Coeurjolly, D., Cuturi, M., and Peyré, G. Ground metric learning on graphs. *Journal of Mathematical Imaging and Vision*, pp. 1–19, 2020.

Huang, G., Quo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., and Sha, F. Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4869–4877, 2016.

Huizing, G.-J., Peyré, G., and Cantini, L. Optimal transport improves cell-cell similarity inference in single-cell omics data. *bioRxiv*, 2021.

Janati, H., Cuturi, M., and Gramfort, A. Spatio-temporal alignments: Optimal transport through space and time. In *International Conference on Artificial Intelligence and Statistics*, pp. 1695–1704. PMLR, 2020.

Kantorovich, L. On the transfer of masses (in Russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Kerdoncuff, T., Emonet, R., and Sebban, M. Metric Learning in Optimal Transport for Domain Adaptation. In *International Joint Conference on Artificial Intelligence*, Kyoto, Japan, January 2021.

Kulis, B. et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. Q. From word embeddings to document distances. In *Proc. of the 32nd Intern. Conf. on Machine Learning*, pp. 957–966, 2015.

Le, T. and Cuturi, M. Unsupervised riemannian metric learning for histograms using aitchison transformations. In *International Conference on Machine Learning*, pp. 2002–2011. PMLR, 2015.

Lemmens, B. and Nussbaum, R. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

Li, R., Ye, X., Zhou, H., and Zha, H. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, pp. 1–149, 2015.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mena, G. and Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *arXiv preprint arXiv:1905.11882*, 2019.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Niles-Weed, J. and Rigollet, P. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

Noh, S. χ^2 metric learning for nearest neighbor classification and its analysis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 991–995. IEEE, 2012.

Paty, F.-P. and Cuturi, M. Subspace robust wasserstein distances. In *International Conference on Machine Learning*, pp. 5072–5081. PMLR, 2019.

Paty, F.-P. and Cuturi, M. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pp. 7532–7542. PMLR, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 630–638, Cadiz, Spain, 09–11 May 2016. PMLR.

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

Santambrogio, F. *Optimal Transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their applications*. Springer, 2015.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Stegle, O., Teichmann, S. A., and Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

Stuart, A. M. and Wolfram, M.-T. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 2020.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Tanay, A. and Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, 2017.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Tong, A., Huguet, G., Shung, D., Natik, A., Kuchroo, M., Lajoie, G., Wolf, G., and Krishnaswamy, S. Embedding signals on knowledge graphs with unbalanced diffusion earth mover's distance. *arXiv preprint arXiv:2107.12334*, 2021.

Villani, C. *Topics in C. Transportation*. Graduate studies in Math. AMS, 2003.

Wang, F. and Guibas, L. J. Supervised earth mover's distance learning and its computer vision applications. In *European Conference on Computer Vision*, pp. 442–455. Springer, 2012.

Weinberger, K. Q., Blitzer, J., and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pp. 1473–1480, 2006.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, pp. 12, 2002.

Xu, J., Luo, L., Deng, C., and Huang, H. Multi-level metric learning via smoothed wasserstein distance. In *IJCAI*, pp. 2919–2925, 2018.

Yang, W., Xu, L., Chen, X., Zheng, F., and Liu, Y. Chi-squared distance metric learning for histogram data. *Mathematical Problems in Engineering*, 2015, 2015.

Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. Challenges and emerging directions in single-cell analysis. *Genome biology*, 18(1):1–8, 2017.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, 32, 2019.

Zen, G., Ricci, E., and Sebe, N. Simultaneous ground metric learning and matrix factorization with earth mover's distance. In *2014 22nd International Conference on Pattern Recognition*, pp. 3690–3695. IEEE, 2014.

A. Computation Times and Numerical Resources

CPU computations were performed on a Dell Latitude 5420 with an 8 core 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz CPU. GPU computations were performed on Nvidia V100 SXM2 32 Go GPUs.

50 pairs of Wasserstein power iterations for the synthetic datasets in Section 2.3 run in about three minutes on CPU.

15 pairs of Sinkhorn ($\varepsilon = 0.1$) power iterations for the single-cell dataset in Section 5 run in about 1h50mn on GPU.

B. Proof of Theorem 2.3 (Existence of Singular Vectors)

We consider the set

$$\mathcal{K}_n^\rho := \{\mathbb{B} \in \mathcal{D}_n \text{ s.t. } \|\mathbb{B}\|_\infty = 1, \forall i \neq j, \mathbb{B}_{i,j} \geq \rho\}$$

Let $(\mathbb{A}, \mathbb{B}) \in \mathcal{K}_m^\rho \times \mathcal{K}_n^\rho$. A classical result states that

$$\frac{\rho}{2}\|a - a'\|_1 \leq \mathrm{W}_{\mathbb{A}}(a, a') \leq \tfrac{1}{2}\|a - a'\|_1$$

Thus with $(\mathbb{A}', \mathbb{B}') = \left(\dfrac{\Phi_B(\mathbb{B})}{\|\Phi_B(\mathbb{B})\|_\infty}, \dfrac{\Phi_A(\mathbb{A})}{\|\Phi_A(\mathbb{A})\|_\infty}\right)$,

$$\frac{\rho}{2}\|b_k - b_\ell\|_1 + \tau R(b_k - b_\ell) \leq \Phi_B(\mathbb{B})_{k,\ell} \leq \tfrac{1}{2}\|b_k - b_\ell\|_1 + \tau R(b_k - b_\ell)$$

and

$$\frac{\rho}{2}\|a_i - a_j\|_1 + \tau R(a_i - a_j) \leq \Phi_A(\mathbb{A})_{i,j} \leq \tfrac{1}{2}\|a_i - a_j\|_1 + \tau R(a_i - a_j).$$

Equivalence of norms implies $k_R^-\|\cdot\|_1 \leq R(\cdot) \leq k_R^+\|\cdot\|_1$. With $\gamma_A := \dfrac{\min_{i\neq j}\|a_i - a_j\|_1}{\max_{i\neq j}\|a_i - a_j\|_1}$ and $\gamma_B := \dfrac{\min_{k\neq \ell}\|b_k - b_\ell\|_1}{\max_{k\neq l}\|b_k - b_\ell\|_1}$,

$$\forall k \neq \ell, \mathbb{A}'_{k,\ell} \geq \frac{\rho + 2\tau k_R^-}{1 + 2\tau k_R^+}\gamma_B \quad \text{and} \quad \forall i \neq j, \mathbb{B}'_{i,j} \geq \frac{\rho + 2\tau k_R^-}{1 + 2\tau k_R^+}\gamma_A.$$

This shows that for

$$0 < \rho \leq \min\left(\frac{2\gamma_A \tau k_R^-}{2\tau k_R^+ + 1 - \gamma_A}, \frac{2\gamma_B \tau k_R^-}{2\tau k_R^+ + 1 - \gamma_B}\right)$$

one has $(\mathbb{A}', \mathbb{B}') \in \mathcal{K}_m^\rho \times \mathcal{K}_n^\rho$. So for such ρ, the function $(\mathbb{A}', \mathbb{B}') \mapsto \left(\dfrac{\Phi_B(\mathbb{B})}{\|\Phi_B(\mathbb{B})\|_\infty}, \dfrac{\Phi_A(\mathbb{A})}{\|\Phi_A(\mathbb{A})\|_\infty}\right)$ is a continuous map from the locally contractible compact $\mathcal{K}_m^\rho \times \mathcal{K}_n^\rho$ to itself so using the Brouwer theorem, it has a fixed point, which is a pair of singular vectors of (Φ_A, Φ_B).

C. Proof of Proposition 2.6 (Uniqueness of Singular Vectors for Large τ)

Denoting $s = 1/\tau$, we consider the map

$$\Psi_A(\mathbb{A}) := \frac{\|\mathbb{A}\|_\infty U + sW_A(\mathbb{A})}{\|\|\mathbb{A}\|_\infty U + sW_A(\mathbb{A})\|_\infty}$$

on the set of $\|\mathbb{A}\|_\infty = 1$, where $U := (R(a_i - a_j))_{i,j}$ is constant and $W_A(\mathbb{A}) := (\mathrm{W}_{\mathbb{A}}(a_i, a_j))_{i,j}$. Since W_A is 1-Lipschitz, and $\|\mathbb{A}\|_\infty = 1$, one needs to study the contractance of

$$W \mapsto \frac{U + sW}{\|U + sW\|_\infty}.$$

One can explicitly compute the derivative of this map, which is $O(s)$, so that Ψ_A is itself contractant for s small enough. This shows that $(\mathbb{A}, \mathbb{B}) \mapsto (\Psi_B(\mathbb{B}), \Psi_A(\mathbb{A}))$ is also contractant, which implies uniqueness of the singular vector and linear convergence for $\|\cdot\|_\infty$ of the power iterations.

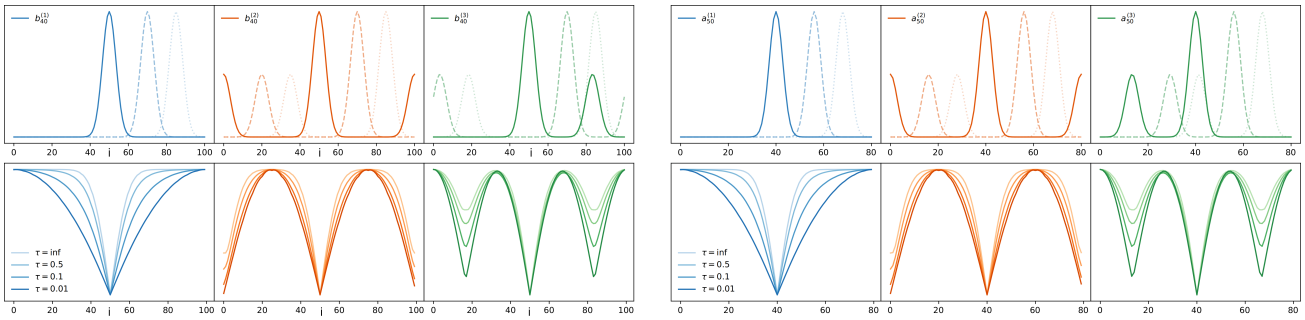

Figure 6. (left, top) Translated histograms template defining B_p (left, bottom) One element of the singular vector $\mathbb{A}_p$ (right, top) Translated histograms template defining A_p (right, bottom) One element of the singular vector $\mathbb{B}_p$.

D. Additional Figure for the Numerical Illustration of Section 2.3

Figure 6 shows that as predicted, the singular vectors $\mathbb{A}_p$ and $\mathbb{B}_p$ are identical up to rescaling.

E. Proof of Convergence for the Stochastic Power Iterations of Section 3.1

Proof. We follow steps similar to the proof of (Nemirovski et al., 2009) for projected stochastic gradient descent. The theorem supposes τ large enough, so we can consider a pair of unique Wasserstein Singular Vectors $(\mathbb{A}^\star, \mathbb{B}^\star)$. We study the quantity $\mathbb{E}[\ell_t]$ where $\ell_t := \|\mathbb{A}_t - \mathbb{A}^\star\|_2^2$. The proof for $\|\mathbb{B}_t - \mathbb{B}^\star\|_2^2$ is identical. We consider a constant scaling factor $\tilde{\lambda}$. Let us start by defining

$$(\mathbb{T}_t)_{k,\ell} = \begin{cases} (\mathbb{A}_t)_{k,\ell} - \Phi_B(\mathbb{B}_t)_{k,\ell}/\tilde{\lambda} \text{ if } (k,\ell) = (k_t, \ell_t) \\ 0 \text{ otherwise} \end{cases}$$

and $\mathbb{T}^\star := p(\mathbb{A}^\star - \Phi_B(\mathbb{B}^\star)/\tilde{\lambda})$ where $p = \frac{1}{mn}$ is the probability for an element to be updated.

By definition of the power iterations, $\mathbb{A}_{t+1} = \Pi(\mathbb{A}_t - \alpha_t \mathbb{T}_t)$.

By definition of the Wasserstein Singular Vectors, $\mathbb{A}^\star = \Pi(\mathbb{A}^\star - \alpha_t \mathbb{T}^\star)$.

Thus,

$$\|\mathbb{A}_{t+1} - \mathbb{A}^\star\|_2^2 = \|\Pi(\mathbb{A}_t - \alpha_t \mathbb{T}_t) - \Pi(\mathbb{A}^\star - \alpha_t \mathbb{T}^\star)\|_2^2.$$

The value of $\tilde{\lambda}$ proposed in the theorem ensures that $\|\mathbb{A}_t - \alpha_t \mathbb{T}_t\|_\infty \geq 1$ and $\|\mathbb{A}^\star - \alpha_t \mathbb{T}^\star\|_\infty \geq 1$.

The theorem of projection on a convex (the unit sphere for the norm $\|\cdot\|_\infty$) then ensures that

$$\|\mathbb{A}_{t+1} - \mathbb{A}^\star\|_2^2 \leq \|(\mathbb{A}_t - \mathbb{A}^\star) - \alpha_t(\mathbb{T}_t - \mathbb{T}^\star)\|_2^2.$$

Decomposing the squared norm, we get

$$\|(\mathbb{A}_t - \mathbb{A}^\star) - \alpha_t(\mathbb{T}_t - \mathbb{T}^\star)\|_2^2 =$$
$$\ell_t - 2\alpha_t \langle \mathbb{A}_t - \mathbb{A}^\star, \mathbb{T}_t - \mathbb{T}^\star \rangle + \alpha_t^2 \|\mathbb{T}_t - \mathbb{T}^\star\|_2^2.$$

The middle term can be simplified when taking its expectation:

$$\mathbb{E}_t[\langle \mathbb{A}_t - \mathbb{A}^\star, \mathbb{T}_t - \mathbb{T}^\star \rangle] =$$
$$p\ell_t - \frac{p}{\tilde{\lambda}} \langle \mathbb{A}_t - \mathbb{A}^\star, \mathrm{W}(\mathbb{B}_t) - \mathrm{W}(\mathbb{B}^\star) \rangle \geq (1 - L/\tilde{\lambda}) p\ell_t,$$

for some constant L, since W is 1-Lipschitz with regards to the infinite norm as proved earlier. The last term can be bounded as well:

$$\|\mathbb{T}_t - \mathbb{T}^\star\|_2^2 \leq \|\mathbb{T}_t\|_2^2 + \|\mathbb{T}^\star\|_2^2 \leq 2p \max(1, \tau \|R_B\|_\infty / \tilde{\lambda}).$$

Calling that term M, we have finally

$$\mathbb{E}_t[\ell_{t+1}] \leq \left(1 - 2\alpha_t p\left(1 - L/\tilde{\lambda}\right)\right) \times \ell_t + \alpha_t^2 M.$$

In the next steps we assume τ big enough for $p\left(1 - L/\tilde{\lambda}\right)$ to be positive and name the quantity Q.

Let us note that an overly pessimistic upper bound for L is m, which would require a very large value of τ. However, this upper-bound of L is obtained by juggling between different norms. In practice, the algorithm converges for arbitrarily small values of τ. This suggests a much smaller constant, that does not depend on the data's dimensionality.

Taking the expectation over all times t,

$$\mathbb{E}[\ell_{t+1}] \leq (1 - 2\alpha_t Q) \times \mathbb{E}[\ell_t] + \alpha_t^2 M.$$

Reformulating,

$$2\alpha_t Q \mathbb{E}[\ell_t] \leq \mathbb{E}[\ell_t] - \mathbb{E}[\ell_{t+1}] + \alpha_t^2 M.$$

Summing along $t = 1...T$,

$$\min_{t=1...T} \mathbb{E}[\ell_t] \leq \left(\sum_{t=1}^{T} \alpha_t\right)^{-1} \left(\frac{\ell_1}{2Q} + \frac{M}{2Q}\sum_{t=1}^{T} \alpha_t^2\right).$$

For $\alpha_t = 1/\sqrt{t}$, we thus have classically a convergence rate of $\mathcal{O}(\log(t)/\sqrt{t})$ □

F. Sinkhorn Algorithm

The Sinkhorn cost can be computed by the dual formula

$$W_{\mathbb{A}}^{\varepsilon}(a_i, a_j) = \varepsilon\left(\langle \log(u), a_i\rangle + \langle \log(v), a_j\rangle - \langle Kv, u\rangle\right),$$

where $K \coloneqq \exp(-\frac{\mathbb{A}}{\varepsilon \|\mathbb{A}\|_\infty})$ and (u, v) are obtained by iterating the following Sinkhorn fixed point

$$u \leftarrow \frac{a_i}{K^\top v} \quad \text{and} \quad v \leftarrow \frac{a_j}{Ku}.$$

This allows one to compute with precision ε the n^2 entries of $\Phi_A^{\varepsilon}(\mathbb{A})$ in $\mathcal{O}((mn)^2/\varepsilon^2)$ operations (Altschuler et al., 2017), using a parallelizable algorithm that is well suited for GPU computations.

G. Proof of Proposition 4.4 (Connection with PCA)

Proof. We define the operator mapping correlation kernels to Euclidean distance

$$\Delta(K) \coloneqq -(K + K^\top) + \mathrm{diag}(K)\mathbb{1}_n^T + \mathbb{1}_n \mathrm{diag}(K)^\top.$$

One has the convenient formula

$$\Phi_A^{\infty}(\mathbb{A}) = -\Delta(A^\top \mathbb{A}\, A).$$

Let the centering operator $J \coloneqq \mathrm{Id}_n - \mathbb{1}_{n\times n}/n$, which satisfies $J^2 = J$ and $\ker(J) = \mathrm{Span}(\mathbb{1}_n)$.

Let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ be a pair of singular vectors, so that there exists $\lambda \in \mathbb{R}$ such that

$$AJu = \lambda v \quad \text{and} \quad JA^\top v = \lambda u.$$

By linearity and using the fact that

$$\ker(\Phi_A^{\infty}) = \{C \text{ s.t. } C = -C^\top\} \cup \{a\mathbb{1}_n^\top + \mathbb{1}_n b^\top\}$$

one has

$$\begin{aligned}\Phi_A^\infty(\Delta(vv^\top)) &= -2\Phi_A^\infty(vv^\top)\\ &= 2\Delta(A^\top vv^\top A)\\ &= 2\Delta(JA^\top vv^\top AJ)\\ &= 2\lambda^2\Delta(uu^\top),\end{aligned}$$

where we used the fact that $\Delta(JKJ) = \Delta(K)$. The same reasoning for Φ_B^∞ yields the advertised result. □

H. Details on Data Processing

We recovered the 'pbmc3k' dataset using the function `pbmc3k` of Scanpy (Wolf et al., 2018). Cell types were annotated using the Azimuth (Hao et al., 2021) web tool, which projects it onto large-scale reference atlases. We removed the cluster 'other T cells' and cells for which the annotation was less than 90% confident. Cells were selected using a standard quality filtering pipeline. The data was CPM-normalized, log1p-transformed, and then the 1000 most varying genes were selected. To those genes we added the canonical markers given in the documentation of Azimuth (Hao et al., 2021).