Statistical Optimal Transport via Factored Couplings

Aden Forrow
MIT

Jan-Christian Hütter
MIT

Mor Nitzan
Harvard University, Broad Institute

Philippe Rigollet
MIT

Geoffrey Schiebinger
MIT, Broad Institute

Jonathan Weed
MIT

Abstract

We propose a new method to estimate Wasserstein distances and optimal transport plans between two probability distributions from samples in high dimension. Unlike plug-in rules that simply replace the true distributions by their empirical counterparts, our method promotes couplings with low *transport rank*, a new structural assumption that is similar to the nonnegative rank of a matrix. Regularizing based on this assumption leads to drastic improvements on high-dimensional data for various tasks, including domain adaptation in single-cell RNA sequencing data. These findings are supported by a theoretical analysis that indicates that the transport rank is key in overcoming the curse of dimensionality inherent to data-driven optimal transport.

1 INTRODUCTION

Optimal transport (OT) was born from a simple question phrased by Gaspard Monge in the eighteenth century [Monge, 1781] and has since flourished into a rich mathematical theory two centuries later [Villani, 2003, 2009]. Recently, OT and more specifically Wasserstein distances, which include the so-called *earth mover's distance* [Rubner et al., 2000] as a special example, have proven valuable for varied tasks in machine learning [Bassetti et al., 2006, Cuturi, 2013, Cuturi and Doucet, 2014b, Frogner et al., 2015, Gao and Kleywegt, 2016, Genevay et al., 2016, 2017, Rigollet and Weed, 2018a,b, Rolet et al., 2016, Solomon et al., 2014b, Srivastava et al., 2015], computer graphics [Bonneel et al., 2011, 2016, de Goes et al., 2012, Solomon et al., 2014a, 2015], geometric processing [de Goes et al., 2011, Solomon et al., 2013], image processing [Gramfort et al., 2015, Rabin and Papadakis, 2015], and document retrieval [Kusner et al., 2015, Ma et al., 2014]. These recent developments have been supported by breakneck advances in computational optimal transport in the last few years that allow the approximation of these distances in near linear time [Altschuler et al., 2017, Cuturi, 2013].

In these examples, Wasserstein distances and transport plans are estimated from data. Yet the understanding of *statistical* aspects of OT is still in its infancy. In particular, current methodological advances focus on computational benefits but often overlook statistical regularization to address stability in the presence of sampling noise. Known theoretical results show that vanilla optimal transport applied to sampled data suffers from the curse of dimensionality [Dobrić and Yukich, 1995, Dudley, 1969, Weed and Bach, 2017] and there is an acute need for principled regularization techniques in order to scale optimal transport to high-dimensional problems, such as those arising in genomics.

At the heart of OT is the computation of Wasserstein distances, which consists of an optimization problem over the infinite dimensional set of *couplings* between probability distributions. (See (1) for a formal definition.) Estimation in this context is therefore nonparametric in nature and this is precisely the source of the curse of dimensionality. To overcome this limitation, and following a major trend in high-dimensional statistics [Candès and Plan, 2010, Liu et al., 2010, Markovsky and Usevich, 2012], we propose to impose low "rank" structure on the couplings. Interestingly, this technique can be implemented efficiently via *Wasserstein barycenters* [Agueh and Carlier, 2011, Cuturi and Doucet, 2014a] with finite support.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

We illustrate the performance of this new procedure for a truly high-dimensional problem arising in single-cell RNA sequencing data, where ad-hoc methods for domain adaptation have recently been proposed to couple datasets collected in different labs and with different protocols [Haghverdi et al., 2017], and even across species [Butler et al., 2018]. Despite a relatively successful application of OT-based methods in this context [Schiebinger et al., 2017], the very high-dimensional and noisy nature of this data calls for robust statistical methods. We show in this paper that our proposed method does lead to improved results for this application.

This paper is organized as follows. We begin by reviewing optimal transport in §2, and we provide an overview of our results in §3. Next, we introduce our estimator in §4. This is a new estimator for the Wasserstein distance between two probability measures that is statistically more stable than the naive plug-in estimator that has traditionally been used. This stability guarantee is not only backed by the theoretical results of §5, but also observed in numerical experiments in practice in §6.

Notation. We denote by $\|\cdot\|$ the Euclidean norm over $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, let δ_x denote the Dirac measure centered at x. For any two real numbers a and b, we denote their minimum by $a \wedge b$. For any two sequences u_n, v_n, we write $u_n \lesssim v_n$ when there exists a constant $C > 0$ such that $u_n \leq C v_n$ for all n. If $u_n \lesssim v_n$ and $v_n \lesssim u_n$, we write $u_n \asymp v_n$. We denote by $\mathbf{1}_n$ the all-ones vector of $\mathbb{R}^n$, and by e_i the ith standard vector in $\mathbb{R}^n$. Moreover, we denote by $\odot$ and $\oslash$ element-wise multiplication and division of vectors, respectively.

For any map $f : \mathbb{R}^d \to \mathbb{R}^d$ and measure μ on $\mathbb{R}^d$, let $f_\# \mu$ denote the pushforward measure of μ through f defined for any Borel set A by $f_\# \mu(A) = \mu\big(f^{-1}(A)\big)$, where $f^{-1}(A) = \{x \in \mathbb{R}^d : f(x) \in A\}$. Given a measure μ, we denote its support by $\mathrm{supp}(\mu)$.

2 BACKGROUND ON OPTIMAL TRANSPORT

In this section, we gather the necessary background on optimal transport. We refer the reader to recent books [Santambrogio, 2015, Villani, 2003, 2009] for more details.

Wasserstein distance Given two probability measures P_0 and P_1 on $\mathbb{R}^d$, let $\Gamma(P_0, P_1)$ denote the set of *couplings* between P_0 and P_1, that is, the set of joint distributions with marginals P_0 and P_1 respectively so that $\gamma \in \Gamma(P_0, P_1)$ iff $\gamma(U \times \mathbb{R}^d) = P_0(U)$ and $\gamma(\mathbb{R}^d \times V) = P_1(V)$ for all measurable $U, V \in \mathbb{R}^d$.

The 2-*Wasserstein distance*[1] between two probability measures P_0 and P_1 is defined as

$$W_2(P_0, P_1) := \inf_{\gamma \in \Gamma(P_0, P_1)} \sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\gamma(x, y)} \,. \quad (1)$$

Under regularity conditions, for example if both P_0 and P_1 are absolutely continuous with respect to the Lebesgue measure, it can be shown the infimum in (1) is attained at a unique coupling γ^*. Moreover γ^* is a deterministic coupling: it is supported on a set of the form $\{(x, T(x)) : x \in \mathrm{supp}(P_0)\}$. In this case, we call T a transport *map*. In general, however, γ^* is unique but for any $x_0 \in \mathrm{supp}(P_0)$, the support of $\gamma^*(x_0, \cdot)$ may not reduce to a single point, in which case, the map $x \mapsto \gamma^*(x, \cdot)$ is called a transport *plan*.

Wasserstein space The space of probability measures with finite 2nd moment equipped with the metric W_2 is called Wasserstein space and denoted by $\mathcal{W}_2$. It can be shown that $\mathcal{W}_2$ is a geodesic space: given two probability measures $P_0, P_1 \in \mathcal{W}_2$, the constant speed geodesic connecting P_0 and P_1 is the curve $\{P_t\}_{t \in [0,1]}$ defined as follows. Let γ^* be the optimal coupling defined as the solution of (1), and for $t \in [0, 1]$ let $\pi_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be defined as $\pi_t(x, y) = (1-t)x + ty$, then $P_t = (\pi_t)_\# \gamma^*$. We then call $P_{1/2}$ the geodesic midpoint of P_0 and P_1. It plays the role of an average in Wasserstein space, which, unlike the mixture $(P_0 + P_1)/2$, takes the geometry of $\mathbb{R}^d$ into account.

k-Wasserstein barycenters The now-popular notion of Wasserstein barycenters (WB) was introduced by Agueh and Carlier [2011] as a generalization of the geodesic midpoint $P_{1/2}$ to more than two measures. In its original form, a WB can be any probability measure on $\mathbb{R}^d$, but algorithmic considerations led Cuturi and Doucet [2014a] to restrict the support of a WB to a finite set of size k. Let $\mathcal{D}_k$ denote the set of probability distributions supported on k points:

$$\mathcal{D}_k = \left\{ \sum_{j=1}^k \alpha_j \delta_{x_j} \, : \, \alpha_j \geq 0, \sum_{j=1}^k \alpha_j = 1, x_j \in \mathbb{R}^d \right\}.$$

For a given integer k, the k-*Wasserstein Barycenter* $\bar{P}$ between N probability measures $P_0, \dots P_N$ on $\mathbb{R}^d$ is defined by

$$\bar{P} = \operatorname*{argmin}_{P \in \mathcal{D}_k} \sum_{j=1}^N W_2^2(P, P^{(j)}) \,. \quad (2)$$

In general (2) is not a convex problem but fast numerical heuristics have demonstrated good performance in

[1] In this paper we omit the prefix "2-" for brevity.

practice [Benamou et al., 2015, Claici et al., 2018, Cuturi and Doucet, 2014a, Cuturi and Peyré, 2016, Staib et al., 2017]. Interestingly, Theorem 4 below indicates that the extra constraint $P \in \mathcal{D}_k$ is also key to statistical stability.

3 RESULTS OVERVIEW

Ultimately, in all the data-driven applications cited above, Wasserstein distances must be estimated from data. While this is arguably the most fundamental primitive of all OT based machine learning, the statistical aspects of this question are often overlooked at the expense of computational ones. We argue that standard estimators of both $W_2(P_0, P_1)$ and its associated optimal transport plan suffer from statistical instability. The main contribution of this paper is to overcome this limitation by injecting statistical regularization.

Previous work Let $X \sim P_0$ and $Y \sim P_1$ and let $X_1, \ldots, X_n$ (resp. $Y_1, \ldots, Y_n$) be independent copies of X (resp. Y).[2] We call $\mathcal{X} = \{X_1, \ldots, X_n\}$ and $\mathcal{Y} = (Y_1, \ldots, Y_n)$ the *source* and *target* datasets respectively. Define the corresponding empirical measures:

$$\hat{P}_0 = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\,, \qquad \hat{P}_1 = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}\,.$$

Perhaps the most natural estimator for $W_2(P_0, P_1)$, and certainly the one most employed and studied, is the *plug-in* estimator $W_2(\hat{P}_0, \hat{P}_1)$. A natural question is to determine the accuracy of this estimator. This question was partially addressed by Sommerfeld and Munk [Sommerfeld and Munk, 2017], where the rate at which $\Delta_n := |W_2(\hat{P}_0, \hat{P}_1) - W_2(P_0, P_1)|$ vanishes is established. They show that $\Delta_n \asymp n^{-1/2}$ if $P_0 \neq P_1$ and $\Delta_n \asymp n^{-1/4}$ if $P_0 = P_1$. Unfortunately, these rates are only valid when P_0 and P_1 have finite support. Moreover, the plug-in estimator for distributions $\mathbb{R}^d$ has been known to suffer from the curse of dimensionality at least since the work of Dudley [Dudley, 1969]. More specifically, in this case, $\Delta_n \asymp n^{-1/d}$ when $d \geq 3$ [Dobrić and Yukich, 1995]. One of the main goals of this paper is to provide an alternative to the naive plug-in estimator by regularizing the optimal transport problem (1). Explicit regularization for optimal transport problems was previously introduced by Cuturi [Cuturi, 2013] who adds an entropic penalty to the objective in (1) primarily driven by algorithmic motivations. While entropic OT was recently shown [Rigollet and Weed, 2018b] to also provide statistical regularization, that result indicates that entropic OT does not alleviate the curse of dimensionality coming from sampling noise, but rather addresses the presence of additional measurement noise.

Closer to our setup are Courty et al. [2014] and Ferradans et al. [2014]; both consider sparsity-inducing structural penalties that are relevant for domain adaptation and computer graphics, respectively. While the general framework of Tikhonov-type regularization for optimal transport problems is likely to bear fruit in specific applications, we propose a new general-purpose *structural* regularization method, based on a new notion of complexity for joint probability measures.

Our contribution The core contribution of this paper is to construct an estimator of the Wasserstein distance between distributions that is more stable and accurate under sampling noise. We do so by defining a new regularizer for couplings, which we call the *transport rank*. As a byproduct, our estimator also yields an estimator of the optimal coupling in (1) that can in turn be used in domain adaptation where optimal transport has recently been employed [Courty et al., 2014, 2017].

To achieve this goal, we leverage insights from a popular technique known as nonnegative matrix factorization (NMF) [Lee and Seung, 2001, Paatero and Tapper, 1994] which has been successfully applied in various forms to many fields, including text analysis [Shahnaz et al., 2006], computer vision [Shashua and Hazan, 2005], and bioinformatics [Gao and Church, 2005]. Like its cousin factor analysis, it postulates the existence of low-dimensional latent variables that govern the high-dimensional data-generating process under study.

In the context of optimal transport, we consider couplings $\gamma \in \Gamma(P_0, P_1)$ such that whenever $(X, Y) \sim \gamma$, there exits a latent variable Z with *finite support* such that X and Y are conditionally independent given Z. To see the analogy with NMF, one may view a coupling γ as a doubly stochastic matrix whose rows and columns are indexed by $\mathbb{R}^d$. We consider couplings such that this matrix can be written as the product AB where A and $B^\top$ are matrices whose rows are indexed by $\mathbb{R}^d$ and columns are indexed by $\{1, \ldots k\}$. In that case, we call k the *transport rank* of γ. We now formally define these notions.

Definition 1. *Given $\gamma \in \Gamma(P_0, P_1)$, the* transport rank *of γ is the smallest integer k such that γ can be written*

$$\gamma = \sum_{j=1}^{k} \lambda_j (Q_j^0 \otimes Q_j^1)\,, \tag{3}$$

[2] Extensions to the case where the two sample sizes differ are straightforward but do not enlighten our discussion.

where the Q_j^0's and Q_j^1's are probability measures on $\mathbb{R}^d$, $\lambda_j \geq 0$ for $j = 1, \ldots, k$, and where $Q_j^0 \otimes Q_j^1$ indicates the (independent) product distribution. We denote the set of couplings between P_0 and P_1 with transport rank at most k by $\Gamma_k(P_0, P_1)$.

When P_0 and P_1 are finitely supported, the transport rank of $\gamma \in \Gamma(P_0, P_1)$ coincides with the nonnegative rank [Cohen and Rothblum, 1993, Yannakakis, 1991] of γ viewed as a matrix. By analogy with a nonnegative factorization of a matrix, we call a coupling written as a sum as in (3) a *factored coupling*. Using the transport rank as a regularizer therefore promotes simple couplings, i.e., those possessing a low-rank "factorization." To implement this regularization, we show that it can be constructed via k-Wasserstein barycenters, for which efficient implementation is readily available.

As an example of our technique, we show in §6 that this approach can be used to obtain better results on *domain adaptation* a.k.a *transductive learning*, a strategy in semi-supervised learning to transfer label information from a source dataset to a target dataset. Notably, while regularized optimal transport has proved to be an effective tool for *supervised* domain adaptation where label information is used to build an explicit Tikhonov regularization [Courty et al., 2014], our approach is entirely unsupervised, in the spirit of Gong et al. [2012] where unlabeled datasets are matched and then labels are transported from the source to the target. While both approaches, supervised and unsupervised, have their own merits, the unsupervised approach is more versatile and appropriate for the biological problem of single cell data integration.

4 REGULARIZATION VIA FACTORED COUPLINGS

To estimate the Wasserstein distance between P_0 and P_1, we find a low-rank factored coupling between the empirical distributions. As we show in §5, the bias induced by this regularizer provides significant statistical benefits. Our procedure is based on an intuitive principle: optimal couplings arising in practice can be well approximated by assuming the distributions have a small number of pieces moving nearly independently. For example, if distributions represent populations of cells, this assumption is that there are a small number of cell "types," each subject to different forces.

Before introducing our estimator, we note that a factored coupling induces coupled partitions of the source and target distributions. These clusterings are "soft" in the sense that they may include fractional points.

Definition 2. *Given $\lambda \in [0, 1]$, a* soft cluster *of a probability measure P is a sub-probability measure C of total mass λ such that $0 \leq C \leq P$ as measures. The* centroid *of C is defined by $\mu(C) = \frac{1}{\lambda}\int x \, \mathrm{d}C(x)$. We say that a collection $C_1, \ldots, C_k$ of soft clusters of P is a* partition *of P if $C_1 + \cdots + C_k = P$.*

The following fact is immediate.

Proposition 4.1. *If $\gamma = \sum_{j=1}^k \lambda_j (Q_j^0 \otimes Q_j^1)$ is a factored coupling in $\Gamma_k(P_0, P_1)$, then $\{\lambda_1 Q_1^0, \ldots, \lambda_k Q_k^0\}$ and $\{\lambda_1 Q_1^1, \ldots, \lambda_k Q_k^1\}$ are partitions of P_0 and P_1, respectively.*

We now give a simple characterization of the "cost" of a factored coupling.

Proposition 4.2. *Let $\gamma \in \Gamma_k(P_0, P_1)$ and let $C_1^0, \ldots, C_k^0$ and $C_1^1, \ldots, C_k^1$ be the induced partitions of P_0 and P_1, with $C_j^0(\mathbb{R}^d) = C_j^1(\mathbb{R}^d) = \lambda_j$ for $j = 1, \ldots k$. Then*

$$\int \|x - y\|^2 \, \mathrm{d}\gamma(x, y) = \sum_{j=1}^k \Big(\lambda_j \|\mu(C_j^0) - \mu(C_j^1)\|^2 + \sum_{l \in \{0,1\}} \int \|x - \mu(C_j^l)\|^2 \, \mathrm{d}C_j^l(x) \Big)$$

The sum over l in the above display contains intra-cluster variance terms similar to the k-means objective, while the first term is a transport term reflecting the cost of transporting the partition of P_0 to the partition of P_1. Since our goal is to estimate the transport distance, we focus on the first term. This motivates the following definition.

Definition 3. *The cost of a factored transport $\gamma \in \Gamma_k(P_0, P_1)$ is*

$$\mathrm{cost}(\gamma) := \sum_{j=1}^k \lambda_j \|\mu(C_j^0) - \mu(C_j^1)\|^2$$

where $\{C_j^0\}_{j=1}^k$ and $\{C_j^1\}_{j=1}^k$ are the partitions of P_0 and P_1 induced by γ, with $C_j^0(\mathbb{R}^d) = C_j^1(\mathbb{R}^d) = \lambda_j$ for $j = 1, \ldots, k$.

Given empirical distributions $\hat{P}_0$ and $\hat{P}_1$, the (unregularized) optimal coupling between $\hat{P}_0$ and $\hat{P}_1$, defined as

$$\operatorname*{argmin}_{\gamma \in \Gamma(\hat{P}_0, \hat{P}_1)} \int \|x - y\|^2 \mathrm{d}\gamma(x, y) \,,$$

is highly sensitive to sampling noise. This motivates considering instead the regularized version

$$\operatorname*{argmin}_{\gamma \in \Gamma_k(\hat{P}_0, \hat{P}_1)} \int \|x - y\|^2 \mathrm{d}\gamma(x, y) \,, \tag{4}$$

where $k \geq 1$ is a regularization parameter. Whereas fast solvers are available for the unregularized problem [Altschuler et al., 2017], it is not clear how to find a solution to (4) by similar means. While alternating minimization approaches similar to heuristics for nonnegative matrix factorization are possible [Arora et al., 2012, Lee and Seung, 2001], we adopt a different approach which has the virtue of connecting (4) to k-Wasserstein barycenters.

Following Cuturi and Doucet [2014a], define the k-Wasserstein barycenter of $\hat{P}_0$ and $\hat{P}_1$ by

$$H = \operatorname*{argmin}_{P \in \mathcal{D}_k} \left\{ W_2^2(P, \hat{P}_0) + W_2^2(P, \hat{P}_1) \right\} . \quad (5)$$

As noted above, efficient procedures have been shown to work well in practice for this non-convex objective.

Strikingly, the k-Wasserstein barycenter of $\hat{P}_0$ and $\hat{P}_1$ implements a slight variant of (4). Given a feasible $P \in \mathcal{D}_k$ in (5), we first note that it induces a factored coupling in $\Gamma_k(\hat{P}_0, \hat{P}_1)$. Indeed, denote by γ_0 and γ_1 the optimal couplings between $\hat{P}_0$ and P and P and $\hat{P}_1$, respectively. Write $z_1, \ldots, z_j$ for the support of P. We can then decompose these couplings as follows:

$$\gamma_0 = \sum_{j=1}^{k} \gamma_0(\cdot \mid z_j) H(z_j), \quad \gamma_1 = \sum_{j=1}^{k} \gamma_1(\cdot \mid z_j) H(z_j)$$

Then for any Borel sets $A, B \subset \mathbb{R}^d$,

$$\gamma_P(A \times B) := \sum_{j=1}^{k} P(z_j) \gamma_0(A|z_j) \gamma_1(B|z_j) \in \Gamma_k(\hat{P}_0, \hat{P}_1)$$

and by the considerations above, this factored transport induces coupled partitions $C_1^0, \ldots, C_k^0$ and $C_1^1, \ldots, C_k^1$ of $\hat{P}_0$ and $\hat{P}_1$ respectively. We call the points $z_1, \ldots, z_j$ "hubs."

The following proposition gives optimality conditions for H in terms of this partition.

Proposition 4.3. *The partitions $C_1^0, \ldots, C_k^0$ and $C_1^1, \ldots, C_k^1$ induced by the solution H of (5) are the minimizers of*

$$\sum_{j=1}^{k} \left(\frac{\lambda_j}{2} \| \mu(C_j^0) - \mu(C_j^1) \|^2 + \sum_{l=0}^{1} \int \| x - \mu(C_j^l) \|^2 \, \mathrm{d}C_j^l(x) \right)$$

where $\lambda_j = C_j^0(\mathbb{R}^d) = C_j^1(\mathbb{R}^d)$. The minimum is over all partitions of $\hat{P}_0$ and $\hat{P}_1$ induced by feasible $P \in \mathcal{D}_k$.

Comparing this result with Proposition 4.2, we see that this objective agrees with the objective of (4) up to a multiplicative factor of $1/2$ in the transport term.

We therefore view (5) as a algorithmically tractable proxy for (4), expecting γ_H to be close to the optimal factored coupling. Hence, we propose the following estimator $\hat{W}$ of the squared Wasserstein distance:

$$\hat{W} := \operatorname{cost}(\gamma_H) , \qquad \text{where } H \text{ solves (5)} . \quad (6)$$

We can also use γ_H to construct an estimated transport map $\hat{T}$ on the points $X_1, \ldots, X_n \in \operatorname{supp}(\hat{P}_0)$ by setting

$$\hat{T}(X_i) = X_i + \frac{1}{\sum_{j=1}^{k} C_j^0(X_i)} \sum_{j=1}^{k} C_j^0(X_i)(\mu(C_j^1) - \mu(C_j^0)) .$$

Moreover, the quantity $\hat{T}_\sharp \hat{P}_0$ provides a stable estimate of the target distribution, which is particularly useful in domain adaptation.

Our core algorithmic technique involves computing a k-Wasserstein Barycenter as in (2). This problem is non-convex in the variables $\mathcal{M}$ and (γ_0, γ_1), but separately convex in each of the two. Therefore, it admits an alternating minimization procedure, Algorithm 1, similar to Lloyd's algorithm for k-means [Lloyd, 1982]. The update with respect to the hubs $\mathcal{H} = \{z_1, \ldots, z_k\}$, given plans γ_0 and γ_1, is a quadratic optimization problem with the explicit solution

$$z_j = \frac{\sum_{i=1}^{n} \gamma_0(z_j, X_i) X_i + \sum_{i=1}^{n} \gamma_1(z_j, Y_i) Y_i}{\sum_{i=1}^{n} \gamma_0(z_j, X_i) + \sum_{i=1}^{n} \gamma_1(z_j, Y_i)},$$

leading to Algorithm 2.

In order to solve for the optimal (γ_0, γ_1) given a value for the hubs $\mathcal{H} = \{z_1, \ldots, z_k\}$ we add the following entropic regularization terms [Cuturi, 2013] to the objective function (5):

$$-\varepsilon \sum_{i,j} (\gamma_0)_{j,i} \log((\gamma_0)_{j,i}) - \varepsilon \sum_{i,j} (\gamma_1)_{j,i} \log((\gamma_1)_{j,i}),$$

where $\varepsilon > 0$ is a small regularization parameter. This turns the optimization over (γ_0, γ_1) into a projection problem with respect to the Kullback-Leibler divergence, which can be solved by a type of Sinkhorn iteration including updates of the hub weights λ_j at each step; see Benamou et al. [2015] and Algorithm 3. For small ε, this will yield a good approximation to the optimal value of the original problem, but the Sinkhorn iterations become increasingly unstable. We employ a numerical stabilization strategy due to Schmitzer [2016] and Chizat et al. [2016]. Also, an initialization for the hubs is needed, for which we suggest using a k-means clustering of either $\mathcal{X}$ or $\mathcal{Y}$.

Algorithm 1 FACTOREDOT

Input: Sampled points $\mathcal{X}, \mathcal{Y}$, parameter $\varepsilon > 0$
Output: Hubs $\mathcal{M}$, transport plans γ_0, γ_1

```
function FACTOREDOT(𝒳, 𝒴, ε)
    Initialize ℳ, e.g ℳ ← KMEANS(𝒳)
    while not converged do
        (γ0, γ1) ← UPDATEPLANS(𝒳, 𝒴, ℳ)
        ℳ ← UPDATEHUBS(𝒳, 𝒴, γ0, γ1)
    end while
    return (ℳ, γ0, γ1)
end function
```

Algorithm 2 UPDATEHUBS

function UPDATEHUBS($\mathcal{X}, \mathcal{Y}, \gamma_0, \gamma_1$)
  for $j = 1, \dots, k$ **do**
    $p^{(0)}_{i,j} = \gamma_0(z_j, X_i)$; $p^{(1)}_{i,j} = \gamma_1(z_j, Y_i)$
    $z_j \leftarrow \frac{\sum_{i=1}^n \{p^{(0)}_{i,j} X_i + p^{(0)}_{i,j} Y_i\}}{\sum_{i=1}^n \{p^{(0)}_{i,j} + p^{(1)}_{i,j}\}}$
  end for
end function

Algorithm 3 UPDATEPLANS

Require: Points $\mathcal{X}, \mathcal{Y}$, hubs $\mathcal{M}$, parameter $\varepsilon > 0$
  function UPDATEPLANS($\mathcal{X}, \mathcal{Y}, \mathcal{M}, \varepsilon$)
    $u_0 = u_1 = \mathbf{1}_k,\ v_0 = v_1 = \mathbf{1}_n$
    $(\xi_0)_{j,i} = \exp(\|z_j - X_i\|_2^2/\varepsilon)$
    $(\xi_1)_{j,i} = \exp(\|z_j - Y_i\|_2^2/\varepsilon)$
    while not converged **do**
      $v_0 = \frac{1}{n}\mathbf{1}_n \oslash (\xi_0^\top u_0)\quad v_1 = \frac{1}{n}\mathbf{1}_n \oslash (\xi_1^\top u_1)$
      $\lambda = (u_0 \odot (\xi_0 v_0))^{1/2} \odot (u_1 \odot (\xi_1 v_1))^{1/2}$
      $u_0 = \lambda \oslash (\xi_0 v_0)$; $u_1 = \lambda \oslash (\xi_1 v_1)$
    end while
  return $(\mathrm{diag}(u_0)\xi_0\,\mathrm{diag}(v_0), \mathrm{diag}(u_1)\xi_1\,\mathrm{diag}(v_1))$
  end function

5 THEORY

In this section, we give theoretical evidence that the use of factored transports makes our procedure more robust. In particular, we show that it can overcome the "curse of dimensionality" generally inherent to the use of Wasserstein distances on empirical data.

To make this claim precise, we show that the objective function in (5) is robust to sampling noise. This result establishes that despite the fact that the unregularized quantity $W_2^2(\hat{P}_0, \hat{P}_1)$ approaches $W_2^2(P_0, P_1)$ very slowly, the empirical objective in (5) approaches the population objective uniformly at the *parametric* rate, thus significantly improving the dependence on the dimension. Via the connection between (5) and factored couplings established in Proposition 4.3, this result implies that regularizing by transport rank yields significant statistical benefits. Specifically, $|\hat{W} - W_2^2(P_0, P_1)|$ will converge rapidly to the approximation errors from using factored couplings and switching (4) to (5).

Theorem 4. *Let P be a measure on $\mathbb{R}^d$ supported on the unit ball, and denote by $\hat{P}$ an empirical distribution comprising n i.i.d. samples from P. Then with probability at least $1 - \delta$,*

$$\sup_{\rho \in \mathcal{D}_k} |W_2^2(\rho, \hat{P}) - W_2^2(\rho, P)| \lesssim \sqrt{\frac{k^3 d \log k + \log(1/\delta)}{n}}\,.$$

A simple rescaling argument implies that this $n^{-1/2}$ rate holds for all compactly supported measures.

This result complements and generalizes known results from the literature on k-means quantization [Maurer and Pontil, 2010, Pollard, 1982, Rakhlin and Caponnetto, 2006]. Indeed, as noted above, the k-means objective is a special case of a squared W_2 distance to a discrete measure [Pollard, 1982]. Theorem 4 therefore recovers the $n^{-1/2}$ rate for the generalization error of the k-means objective; however, our result applies more broadly to *any* measure ρ with small support. Though the parametric $n^{-1/2}$ rate is optimal, we do not know whether the dependence on k or d in Theorem 4 can be improved. We discuss the connection between our work and existing results on k-means clustering in the supplement.

Finally note that while this analysis is a strong indication of the stability of our procedure, it does not provide explicit rates of convergence for $\hat{W}$ defined in (6). This question requires a structural description of the optimal coupling between P_0 and P_1 and is beyond the scope of the present paper.

6 EXPERIMENTS

We illustrate our theoretical results with numerical experiments on both simulated and real high-dimensional data. Further details about the experimental setup are included in the appendix §F.

6.1 Synthetic data

Two synthetic examples show the improved performance of our estimator for the W_2 distance.

Fragmented hypercube We consider $P_0 = \mathsf{Unif}([-1,1]^d)$, the uniform distribution on a hypercube in dimension d and $P_1 = T_\#(P_0)$, the push-forward of P_0 under a map T, defined as the distribution of $Y = T(X)$ if $X \sim P_0$. We choose $T(X) = X + 2\,\mathrm{sign}(X) \odot (e_1 + e_2)$, where the sign is taken element-wise. As can be seen in Figure 1, this splits the cube into four pieces which drift away. This map is the subgradient of a convex function and hence an optimal transport map by Brenier's Theorem [Villani, 2003, Theorem 2.12]. This observation allows us to compute explicitly $W_2^2(P_0, P_1) = 8$. We compare the results of computing optimal transport on samples and the associated empirical optimal transport cost with the estimator (6), as well as with a simplified procedure that consists in first performing k-means on both $\hat{P}_0$ and $\hat{P}_1$ and subsequently calculating the W_2 distance between the centroids.

The bottom left subplot of Figure 1 shows that FactoredOT provides a substantially better estimate of

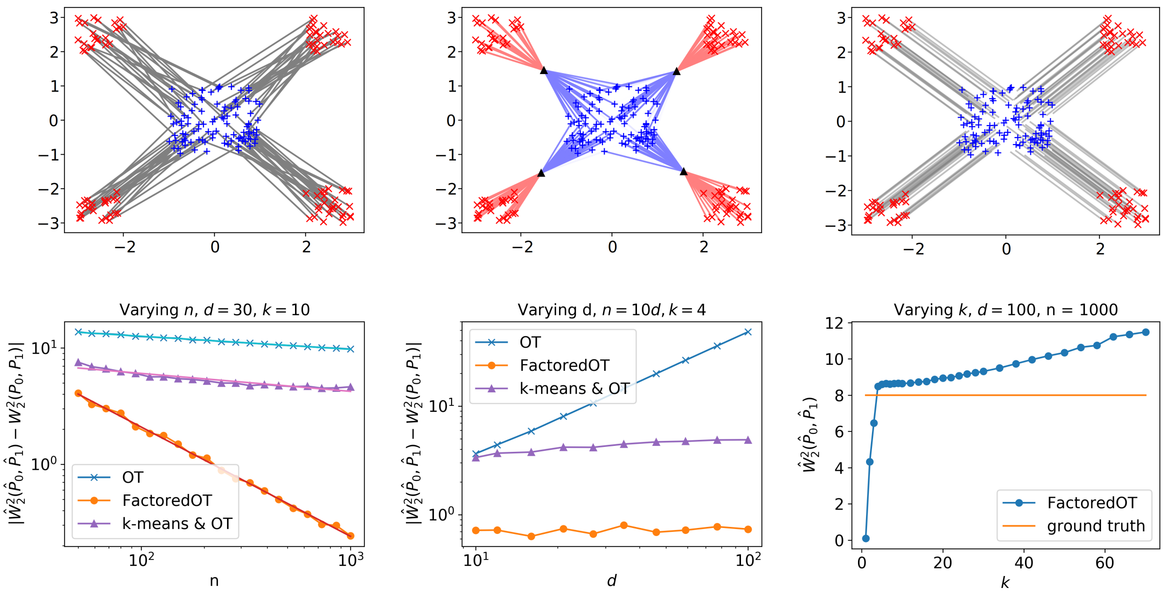

Figure 1: Fragmenting hypercube example. **Top row:** Projections to the first two dimensions (computed for $d = 30$) of (left) the OT coupling of samples from P_0 (in blue) to samples from P_1 (red), (middle) the FactoredOT coupling (factors in black), and (right) the FactoredOT coupling rounded to a map. **Bottom row:** Performance comparisons for (left) varying n and (middle) varying d with $n = 10d$, as well as (right) a diagnostic plot with varying k. All points are averages over 20 samples.

the W_2 distance compared to the empirical optimal transport cost, especially in its scaling with the sample size. Moreover, from the bottom center subplot of the same figure, we deduce that a linear scaling of samples with respect to the dimension is enough to guarantee bounded error for FactoreedOT, unlike for an empirical coupling. Finally, the bottom right plot indicates that the estimator is rather stable to the choice of k above a minimum threshold. We suggest choosing k to match this threshold.

Disk to annulus To show the robustness of our estimator in the case where the ground truth Wasserstein distance is not exactly the cost of a factored coupling, we calculate the optimal transport between the uniform measures on a disk and on an annulus. In order to turn this into a high-dimensional problem, we consider the 2D disk and annulus as embedded in d dimensions and extend both source and target distribution to be independent and uniformly distributed on the remaining $d-2$ dimensions. In other words, we set

$$
\begin{aligned}
P_0 &= \mathsf{Unif}(\{x \in \mathbb{R}^d : \|(x_1, x_2)\|_2 \leq 1, \\
&\qquad\qquad x_i \in [0,1] \text{ for } i = 3, \ldots, d\}) \\
P_1 &= \mathsf{Unif}(\{x \in \mathbb{R}^d : 2 \leq \|(x_1, x_2)\|_2 \leq 3, \\
&\qquad\qquad x_i \in [0,1] \text{ for } i = 3, \ldots, d\})
\end{aligned}
$$

Figure 2 shows that the performance is similar to that obtained for the fragmenting hypercube.

6.2 Batch correction for single cell RNA data

The advent of single cell RNA sequencing is revolutionizing biology with a data deluge. Biologists can now quantify the cell types that make up different tissues and quantify the molecular changes that govern development (reviewed in Wagner et al. [2016] and Kolodziejczyk et al. [2015]). As data is collected by different labs, and for different organisms, there is an urgent need for methods to robustly integrate and align these different datasets [Butler et al., 2018, Crow et al., 2018, Haghverdi et al., 2018].

Cells are represented mathematically as points in a several-thousand dimensional vector space, with a dimension for each gene. The value of each coordinate represents the expression-level of the corresponding gene. Here we show that optimal transport achieves state of the art results for the task of aligning single cell datasets. We align a pair of haematopoietic datasets collected by different scRNA-seq protocols in different laboratories (as described in Haghverdi et al. [2018]). We quantify performance by measuring the fidelity of cell-type label transfer across data sets. This information is available as ground truth in both datasets, but is not involved in computing the alignment.

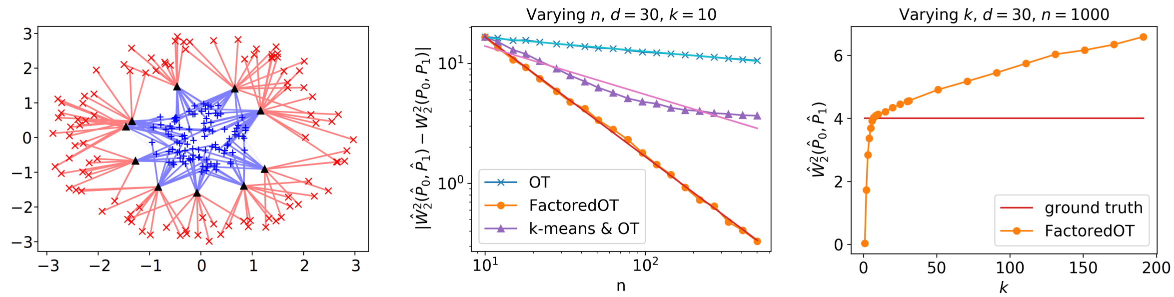

Figure 2: Disk to annulus example, $d = 30$. **Left:** Visualization of the cluster assignment in first two dimensions. **Middle:** Performance for varying n. **Right:** Diagnostic plot when varying k.

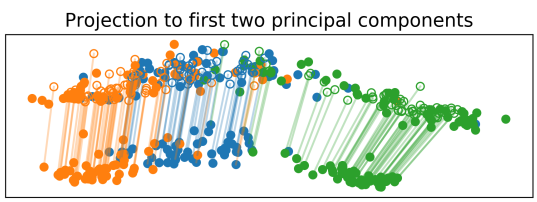

Figure 3: Domain adaptation for scRNA-seq. Both source and target data set are subsampled (50 cells/type) and colored by cell type. Empty circles indicate the inferred label with 20NN classification after FactoredOT.

Table 1: Mean mis-classification percentage (Error) and standard deviation (Std) for scRNA-Seq batch correction

Method	Error	Std
FOT	**14.10**	4.44
MNN	17.53	5.09
OT	17.47	3.17
OT-ER	18.58	6.57
OT-L1L2	15.47	5.35
kOT	15.37	4.76
SA	15.10	3.14
TCA	24.57	7.04
NN	21.98	4.90

We compare the performance of FactoredOT (FOT) to the following baselines: (a) independent majority vote on k nearest neighbors in the target set (NN), (b) optimal transport (OT), (c) entropically regularized optimal transport (OT-ER), (d) OT with group lasso penalty (OT-L1L2) [Courty et al., 2014], (e) a two-step method in which we first perform k-means and then use OT on the k-means centroids (kOT), (f) Subspace Alignment (SA) [Fernando et al., 2013], (g) Transfer Component Analysis (TCA) [Pan et al., 2011], and (h) mutual nearest neighbors (MNN) [Haghverdi et al., 2018]. After projecting the source data onto the target set space, we predict the label of each source single cell by using a majority vote over the 20 nearest neighbor cells in the target dataset (Figure 3). FactoredOT outperforms the baselines for this task, as shown in Table 1, where we report the percentage of mislabeled data.

7 DISCUSSION

We have made a first step towards statistical regularization of optimal transport with the objective of estimating both the Wasserstein distance and the optimal coupling between two probability distributions. Such regularization remains largely unexplored and many other forms of inductive bias may be envisioned, including latent distributions with infinite support but low complexity. The method proposed here generically applies to various tasks associated to optimal transport, leads to a good estimator of the W_2 distance even in high dimension, and is also competitive with state-of-the-art domain adaptation techniques. Our theoretical results demonstrate that the curse of dimensionality in statistical optimal transport can be overcome by imposing structural assumptions. This is an encouraging step towards the deployment of optimal transport as a tool in high-dimensional data analysis.

Acknowledgements

M.N. is supported by the James S. McDonnell Foundation, Schmidt Futures, Israel Council for Higher Education, and the John Harvard Distinguished Science Fellows Program; P.R. by NSF grants DMS-1712596 and TRIPODS-1740751 and IIS-1838071, ONR grant N00014-17-1-2147, the Chan Zuckerberg Initiative DAF 2018-182642, and the MIT Skoltech Seed Fund; G.S. by a Burroughs Welcome Fund Career Award at the Scientific Interface and the Klarman Cell Observatory; and J.W. by the Josephine de Karman fellowship.

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

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

S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 145–162, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1245-5.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76(12):1298–1302, 2006.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich. Displacement interpolation using Lagrangian mass transport. In *ACM Transactions on Graphics*, volume 30, pages 158:1–158:12, 2011.

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016.

A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 2018.

G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. doi: 10.1109/JPROC.2009.2035722. URL https://doi.org/10.1109/JPROC.2009.2035722.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein barycenters. *International Conference on Machine Learning (ICML), to appear*, 2018.

J. E. Cohen and U. G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra Appl.*, 190: 149–168, 1993. ISSN 0024-3795. doi: 10.1016/0024-3795(93)90224-C. URL https://doi.org/10.1016/0024-3795(93)90224-C.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

M. Crow, A. Paul, S. Ballouz, Z. J. Huang, and J. Gillis. Characterizing the replicability of cell types defined by single cell rna-sequencing data using metaneighbor. *Nature communications*, 9(1): 884, 2018.

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

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014a.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proc. ICML*, pages 685–693, 2014b.

M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

F. de Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun. An optimal transport approach to robust reconstruction and simplification of 2d shapes. In *Computer Graphics Forum*, volume 30, pages 1593–1602, 2011.

F. de Goes, K. Breeden, V. Ostromoukhov, and M. Desbrun. Blue noise through optimal transport. *ACM Transactions on Graphics*, 31(6):171, 2012.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

V. Dobrić and J. E. Yukich. Asymptotics for transportation cost in high dimensions. *J. Theoret. Probab.*, 8(1):97–118, 1995. ISSN 0894-9840. doi:

10.1007/BF02213456. URL https://doi.org/10.1007/BF02213456.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist*, 40:40–50, 1969. ISSN 0003-4851. doi: 10.1214/aoms/1177697802. URL https://doi.org/10.1214/aoms/1177697802.

R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929 (1979), 1978. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(197812)6:6<899:CLTFEM>2.0.CO;2-0&origin=MSN.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Computer Vision (ICCV), 2013 IEEE International Conference On*, pages 2960–2967. IEEE, 2013.

X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. pages 1–96. Lecture Notes in Math., Vol. 480, 1975.

S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2044–2052, 2015.

R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:1604.02199*, 2016.

Y. Gao and G. Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21(21):3970–3975, 2005.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Sinkhorn-autodiff: Tractable Wasserstein learning of generative models. *arXiv:1706.00292*, 2017.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016. ISBN 978-1-107-04316-9. doi: 10.1017/CBO9781107337862. URL https://doi.org/10.1017/CBO9781107337862.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, pages 2066–2073. IEEE, 2012.

S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. ISBN 3-540-67394-6. doi: 10.1007/BFb0103945. URL https://doi.org/10.1007/BFb0103945.

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

L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Correcting batch effects in single-cell rna sequencing data by matching mutual nearest neighbours. *bioRxiv*, page 165118, 2017.

L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 2018.

D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

M. Kusner, Y. Sun, N. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proc. ICML*, pages 957–966, 2015.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 663–670. Omnipress, 2010. URL http://www.icml2010.org/papers/521.pdf.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

J. Ma, Q. Z. Sheng, L. Yao, Y. Xu, and A. Shemshadi. Keyword search over web documents based on earth mover's distance. In *Web Information Systems Engineering*, pages 256–265. 2014.

I. Markovsky and K. Usevich. *Low rank approximation*. Springer, 2012.

A. Maurer and M. Pontil. *K-dimensional coding schemes in Hilbert spaces*. *IEEE Trans. Inform. Theory*, 56(11):5839–5846, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2069250. URL https://doi.org/10.1109/TIT.2010.2069250.

C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mém. de l'Ac. R. des Sc.*, pages 666–704, 1781.

A. Müller. Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2):429–443, 1997. ISSN 0001-8678. doi: 10.2307/1428011. URL https://doi.org/10.2307/1428011.

S. Nestorowa, F. K. Hamey, B. P. Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31, 2016.

M. K. Ng. A note on constrained k-means algorithms. *Pattern Recognition*, 33(3):515–519, 2000.

A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, second edition, 2000. ISBN 0-471-98635-6. doi: 10.1002/9780470317013. URL https://doi.org/10.1002/9780470317013. With a foreword by D. G. Kendall.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.

S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171, 2014.

D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982.

J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *Scale Space and Variational Methods in Computer Vision*, pages 256–269. 2015.

A. Rakhlin and A. Caponnetto. Stability of k-means clustering. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1121–1128. MIT Press, 2006. ISBN 0-262-19568-2. URL http://papers.nips.cc/paper/3116-stability-of-k-means-clustering.

P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *arXiv preprint arXiv:1809.05572*, 2018a.

P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *arXiv preprint arXiv:1806.10648*, 2018b.

A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *AISTATS*, 2016.

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

F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. ISBN 978-3-319-20827-5; 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2. URL https://doi.org/10.1007/978-3-319-20828-2. Calculus of variations, PDEs, and modeling.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. Lander. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *bioRxiv*, 2017.

B. Schmitzer. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *arXiv:1610.06519 [cs, math]*, Oct. 2016.

F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Inf. Process. Manage.*, 42(2):373–386, 2006. doi: 10.1016/j.ipm.2004.11.005. URL https://doi.org/10.1016/j.ipm.2004.11.005.

A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In L. D. Raedt and S. Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 792–799. ACM, 2005. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102451. URL http://doi.acm.org/10.1145/1102351.1102451.

D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1962.tb02419.x. URL https://doi.org/10.1002/j.1538-7305.1962.tb02419.x.

J. Solomon, L. Guibas, and A. Butscher. Dirichlet energy for analysis and synthesis of soft maps. In *Computer Graphics Forum*, volume 32, pages 197–206. Wiley Online Library, 2013.

J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics*, 33(4):67, 2014a.

J. Solomon, R. M. Rustamov, L. J. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *ICML*, pages 306–314, 2014b.

J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66, 2015.

M. Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2017.

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

M. Staib, S. Claici, J. M. Solomon, and S. Jegelka. Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 2644–2655, 2017.

V. N. Sudakov. Gaussian random processes, and measures of solid angles in Hilbert space. *Dokl. Akad. Nauk SSSR*, 197:43–45, 1971. ISSN 0002-3264.

V. N. Vapnik and A. J. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971. ISSN 0040-361x.

R. Vershynin. High-dimensional probability. *An Introduction with Applications*, 2016.

C. Villani. *Topics in Optimal Transportation.* Number 58. American Mathematical Soc., 2003.

C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. Old and new.

A. Wagner, A. Regev, and N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145, 2016.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

M. Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441 – 466, 1991.