Subspace Robust Wasserstein Distances

François-Pierre Paty [1] **Marco Cuturi** [2 1]

Abstract

Making sense of Wasserstein distances between discrete measures in high-dimensional settings remains a challenge. Recent work has advocated a two-step approach to improve robustness and facilitate the computation of optimal transport, using for instance projections on random real lines, or a preliminary quantization of the measures to reduce the size of their support. We propose in this work a "max-min" robust variant of the Wasserstein distance by considering the maximal possible distance that can be realized between two measures, assuming they can be projected orthogonally on a lower k-dimensional subspace. Alternatively, we show that the corresponding "min-max" OT problem has a tight convex relaxation which can be cast as that of finding an optimal transport plan with a low transportation cost, where the cost is alternatively defined as the sum of the k largest eigenvalues of the second order moment matrix of the displacements (or matchings) corresponding to that plan (the usual OT definition only considers the trace of that matrix). We show that both quantities inherit several favorable properties from the OT geometry. We propose two algorithms to compute the latter formulation using entropic regularization, and illustrate the interest of this approach empirically.

1. Introduction

The optimal transport (OT) toolbox (Villani, 2009) is gaining popularity in machine learning, with several applications to data science outlined in the recent review paper (Peyré & Cuturi, 2019). When using OT on high-dimensional data, practitioners are often confronted to the intrinsic instability of OT with respect to input measures. A well known result states for instance that the sample complexity of Wasserstein distances can grow exponentially in dimension (Dudley, 1969; Fournier & Guillin, 2015), which means that an irrealistic amount of samples from two continuous measures is needed to approximate faithfully the true distance between them. This result can be mitigated when data lives on lower dimensional manifolds as shown in (Weed & Bach, 2017), but sample complexity bounds remain pessimistic even in that case. From a computational point of view, that problem can be interpreted as that of a lack of robustness and instability of OT metrics with respect to their inputs. This fact was already a common concern of the community when these tools were first adopted, as can be seen in the use of ℓ_1 costs (Ling & Okada, 2007) or in the common practice of thresholding cost matrices (Pele & Werman, 2009).

Regularization The idea to trade off a little optimality in exchange for more regularity is by now considered a crucial ingredient to make OT work in data sciences. A line of work initiated in (Cuturi, 2013) advocates adding an entropic penalty to the original OT problem, which results in faster and differentiable quantities, as well as improved sample complexity bounds (Genevay et al., 2019). Following this, other regularizations (Dessein et al., 2018), notably quadratic (Blondel et al., 2018), have also been investigated. Sticking to an entropic regularization, one can also interpret the recent proposal by Altschuler et al. (2018b) to approximate Gaussian kernel matrices appearing in the regularized OT problem with Nyström-type factorizations (or exact features using a Taylor expansion (Cotter et al., 2011) as in (Altschuler et al., 2018a)), as robust approaches that are willing to tradeoff yet a little more cost optimality in exchange for faster Sinkhorn iterations. In a different line of work, quantizing first the measures to be compared before carrying out OT on the resulting distriutions of centroids is a fruitful alternative (Canas & Rosasco, 2012) which has been recently revisited in (Forrow et al., 2019). Another approach exploits the fact that the OT problem between two distributions on the real line boils down to the direct comparison of their generalized quantile functions (Santambrogio, 2015, §2). Computing quantile functions only requires sorting values, with a mere log-linear complexity. The *sliced* approximation of OT (Rabin et al., 2011) consists in projecting two probability distributions on a given line, compute the optimal transport cost between these projected values, and then repeat this procedure to average these distances over

[1]CREST-ENSAE, Palaiseau, France [2]Google Brain, Paris, France. Correspondence to: François-Pierre Paty <francois.pierre.paty@ensae.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

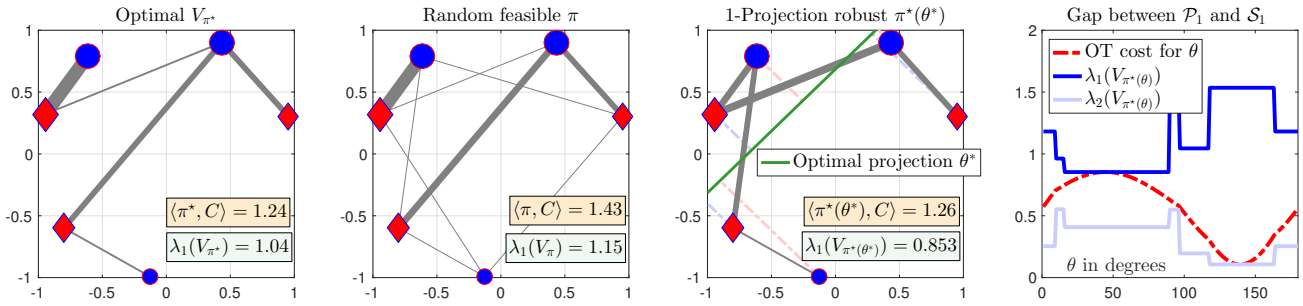

Figure 1. We consider two discrete measures (red and blue dots) on the plane. The left-most plot shows the optimal transport between these points, in which the width of the segment is proportional to the mass transported between two locations. The total cost is displayed in the lower right part of the plot as $\langle \pi^\star, C \rangle$, where C is the pairwise squared-Euclidean distance matrix. The largest eigenvalue of the corresponding second order moment matrix $V_{\pi^\star}$ of displacements, see (1), is given below. As can be expected and seen in the second plot, choosing a random transportation plan yields a higher cost. The third plot displays the most robust projection direction (green line), that upon which the OT cost of these point clouds is largest once projected. The maximal eigenvalue of the second order moment matrix (still in dimension 2) is smaller than that obtained with the initial OT plan. Finally, we plot as a function of the angle θ between $(0, 180)$ the OT cost (which, in agreement with the third plot, is largest for the angle corresponding to the green line of the third plot) as well as the corresponding maximal eigenvalue of the second order moment of the optimal plan corresponding to *each* of these angles θ. The maximum of the red curve, as well as the minimum reached by the dark blue one, correspond respectively to the values of the projection $\mathcal{P}_k$ and subspace $\mathcal{S}_k$ robust Wasserstein distances described in §3. They happen to coincide in this example, but one may find examples in which they do not, as can be seen in Figure 11 (supplementary material). The smallest eigenvalue is given for illustrative purposes only.

several random lines. This approach can be used to define kernels (Kolouri et al., 2016), compute barycenters (Bonneel et al., 2015) but also to train generative models (Kolouri et al., 2018; Deshpande et al., 2018). Beyond its practical applicability, this approach is based on a perhaps surprising point-of-view: OT on the real line may be sufficient to extract geometric information from two high-dimensional distributions. Our work builds upon this idea, and more candidly asks what can be extracted from a little more than a real line, namely a subspace of dimension $k \geq 2$. Rather than project two measures on several lines, we consider in this paper projecting them on a k-dimensional subspace that maximizes their transport cost. This results in optimizing the Wasserstein distance over the ground metric, which was already considered for supervised learning (Cuturi & Avis, 2014; Flamary et al., 2018).

Contributions This optimal projection translates into a "max-min" robust OT problem with desirable features. Although that formulation cannot be solved with convex solvers, we show that the corresponding "min-max" problem admits on the contrary a tight convex relaxation and also has an intuitive interpretation. To see that, one can first notice that the usual 2-Wasserstein distance can be described as the minimization of the *trace* of the second order moment matrix of the displacements associated with a transport plan. We show that computing a maximally discriminating optimal k dimensional subspace in this "min-max" formulation can be carried out by minimizing the sum of the k largest eigenvalues (instead of the entire trace) of that second order moment matrix. A simple example summarizing the link between these two "min-max" and "max-min" quantities is given in Figure 1. That figure considers a toy example where points in dimension $d = 2$ are projected on lines $k = 1$, our idea is designed to work for larger k and d, as shown in §6.

Paper structure We start this paper with background material on Wasserstein distances in §2 and present an alternative formulation for the 2-Wasserstein distance using the second order moment matrix of displacements described in a transport plan. We define in §3 our "max-min" and "min-max" formulations for, respectively, projection (PRW) and subspace (SRW) robust Wasserstein distances. We study the geodesic structure induced by the SRW distance on the space of probability measures in §4, as well as its dependence on the dimension parameter k. We provide computational tools to evaluate SRW using entropic regularization in §5. We conclude the paper with experiments in §6 to validate and illustrate our claims, on both simulated and real datasets.

2. Background on Optimal Transport

For $d \in \mathbb{N}$, we write $[\![d]\!] = \{1, ..., d\}$. Let $\mathscr{P}(\mathbb{R}^d)$ be the set of Borel probability measures in $\mathbb{R}^d$, and let

$$\mathscr{P}_2(\mathbb{R}^d) = \left\{ \mu \in \mathscr{P}(\mathbb{R}^d) \,\middle|\, \int \|x\|^2 \, d\mu(x) < \infty \right\}.$$

Monge and Kantorovich Formulations of OT For $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, we write $\Pi(\mu,\nu)$ for the set of couplings

$$\Pi(\mu,\nu) = \{\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \forall A, B \subset \mathbb{R}^d \text{ Borel}, \\ \pi(A \times \mathbb{R}^d) = \mu(A), \pi(\mathbb{R}^d \times B) = \nu(B)\},$$

and their 2-Wasserstein distance is defined as

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \, d\pi(x,y) \right)^{1/2}.$$

Because we only consider quadratic costs in the remainder of this paper, we drop the subscript 2 in our notation and will only use $\mathcal{W}$ to denote the 2-Wasserstein distance. For Borel $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, Borel $T : \mathcal{X} \to \mathcal{Y}$ and $\mu \in \mathscr{P}(\mathcal{X})$, we denote by $T_{\#}\mu \in \mathscr{P}(\mathcal{Y})$ the push-forward of μ by T, *i.e.* the measure such that for any Borel set $A \subset \mathcal{Y}$,

$$T_{\#}\mu(A) = \mu\left(T^{-1}(A)\right).$$

The Monge (1781) formulation of optimal transport is, when this minimization is feasible, equivalent to that of Kantorovich, namely

$$\mathcal{W}(\mu,\nu) = \left(\inf_{T : T_{\#}\mu = \nu} \int \|x - T(x)\|^2 \, d\mu(x) \right)^{1/2}.$$

$\mathcal{W}$ as Trace-minimization For any coupling π, we define the $d \times d$ second order displacement matrix

$$V_\pi := \int (x-y)(x-y)^T d\pi(x,y). \tag{1}$$

Notice that when a coupling π corresponds to a Monge map, namely $\pi = (\mathrm{Id}, T)_{\#}\mu$, then one can interpret even more naturally V_π as the second order moment of all displacement $(x - T(x))(x - T(x))^T$ weighted by μ. With this convention, we remark that the total cost of a coupling π is equal to the trace of V_π, using the simple identity $\mathrm{trace}(x-y)(x-y)^T = \|x-y\|^2$ and the linearity of the integral sum. Computing the $\mathcal{W}$ distance can therefore be interpreted as minimizing the trace of V_π. This simple observation will play an important role in the next section, and more specifically the study of $\lambda_l(V_\pi)$, the l-th *largest* eigenvalue of V_π.

3. Subspace Robust Wasserstein Distances

With the conventions and notations provided in §2, we consider here different robust formulations of the Wasserstein distance. Consider first for $k \in [\![d]\!]$, the Grassmannian of k-dimensional subspaces of $\mathbb{R}^d$:

$$\mathcal{G}_k = \left\{ E \subset \mathbb{R}^d \,|\, \dim(E) = k \right\}.$$

For $E \in \mathcal{G}_k$, we note P_E the orthogonal projector onto E. Given two measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, a first attempt at computing a robust version of $\mathcal{W}(\mu,\nu)$ is to consider the worst possible OT cost over all possible low dimensional projections:

Definition 1. *For $k \in [\![d]\!]$, the k-dimensional projection robust 2-Wasserstein (PRW) distance between μ and ν is*

$$\mathcal{P}_k(\mu,\nu) = \sup_{E \in \mathcal{G}_k} \mathcal{W}\left(P_{E\#}\mu, P_{E\#}\nu\right).$$

As we show in the supplementary material, this quantity is well posed and itself worthy of interest, yet difficult to compute. In this paper, we focus our attention on the corresponding "min-max" problem, to define the k-dimensional subspace robust 2-Wasserstein (SRW) distance:

Definition 2. *For $k \in [\![d]\!]$, the k-dimensional subspace robust 2-Wasserstein distance between μ and ν is*

$$\mathcal{S}_k(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sup_{E \in \mathcal{G}_k} \left[\int \|P_E(x-y)\|^2 d\pi(x,y) \right]^{1/2}$$

Remark 1. *Both quantities $\mathcal{S}_k$ and $\mathcal{P}_k$ can be interpreted as robust variants of the $\mathcal{W}$ distance. By a simple application of weak duality we have that $\mathcal{P}_k(\mu,\nu) \leq \mathcal{S}_k(\mu,\nu)$.*

Lemma 1. *Optimal solutions for $\mathcal{S}_k$ exist,* i.e.

$$\mathcal{S}_k(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \max_{E \in \mathcal{G}_k} \left[\int \|P_E(x-y)\|^2 d\pi(x,y) \right]^{1/2}$$

We show next that the SRW variant $\mathcal{S}_k$ can be elegantly reformulated as a function of the eigendecomposition of the displacement second-order moment matrix V_π (1):

Lemma 2. *For $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, one has*

$$\begin{aligned} \mathcal{S}_k^2(\mu,\nu) &= \min_{\pi \in \Pi(\mu,\nu)} \max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^T = I_k}} \int \|Ux - Uy\|^2 \, d\pi(x,y) \\ &= \min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi). \end{aligned}$$

This characterization as a sum of eigenvalues will be crucial to study theoretical properties of $\mathcal{S}_k$. Subspace robust Wasserstein distances can in fact be interpreted as a convex relaxation of projection robust Wasserstein distances: they can be computed as the maximum of a concave function over a convex set, which will make computations tractable.

Theorem 1. *For $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$,*

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega) = k}} \int d_\Omega^2 \, d\pi \tag{2}$$

$$= \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega) = k}} \min_{\pi \in \Pi(\mu,\nu)} \int d_\Omega^2 \, d\pi \tag{3}$$

$$= \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega) = k}} \mathcal{W}^2\left(\Omega^{1/2}{}_{\#}\mu, \Omega^{1/2}{}_{\#}\nu\right) \tag{4}$$

where d_Ω stands for the Mahalanobis distance

$$d^2_\Omega(x,y) = (x-y)^T\Omega(x-y).$$

We can now prove that *both* PRW and SRW variants are, indeed, distances over $\mathscr{P}_2(\mathbb{R}^d)$.

Proposition 1. *For $k \in [\![d]\!]$, both $\mathcal{P}_k$ and $\mathcal{S}_k$ are distances over $\mathscr{P}_2(\mathbb{R}^d)$.*

Proof. Symmetry is clear for both objects, and for $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, $\mathcal{S}_k(\mu,\mu) = \mathcal{P}_k(\mu,\mu) = 0$. Let $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ such that $\mathcal{S}_k(\mu,\nu) = 0$. Then $\mathcal{P}_k(\mu,\nu) = 0$ and for any $E \in \mathcal{G}_k$, $\mathcal{W}(P_{E\#}\mu, P_{E\#}\nu) = 0$, *i.e.* $P_{E\#}\mu = P_{E\#}\nu$. Lemma 7 (in the supplementary material) then shows that $\mu = \nu$. For the triangle inequalities, let $\mu_0,\mu_1,\mu_2 \in \mathscr{P}_2(\mathbb{R}^d)$. Let $\Omega_\star \in \{0 \preceq \Omega \preceq I,\, \mathrm{trace}(\Omega) = k\}$ be optimal between μ_0 and μ_2. Using the triangle inequalities for the Wasserstein distance,

$$\begin{aligned}
&\mathcal{S}_k(\mu_0,\mu_2) = \mathcal{W}\left[{\Omega_\star^{1/2}}_\#\mu_0, {\Omega_\star^{1/2}}_\#\mu_2\right]\\
&\leq \mathcal{W}\left[{\Omega_\star^{1/2}}_\#\mu_0, {\Omega_\star^{1/2}}_\#\mu_1\right] + \mathcal{W}\left[{\Omega_\star^{1/2}}_\#\mu_1, {\Omega_\star^{1/2}}_\#\mu_2\right]\\
&\leq \sup_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \mathcal{W}\left[{\Omega^{1/2}}_\#\mu_0, {\Omega^{1/2}}_\#\mu_1\right]\\
&+ \sup_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \mathcal{W}\left[{\Omega^{1/2}}_\#\mu_1, {\Omega^{1/2}}_\#\mu_2\right]\\
&= \mathcal{S}_k(\mu_0,\mu_1) + \mathcal{S}_k(\mu_1,\mu_2).
\end{aligned}$$

The same argument, used this time with projections, yields the triangle inequalities for $\mathcal{P}_k$. □

4. Geometry of Subspace Robust Distances

We prove in this section that SRW distances share several fundamental geometric properties with the Wasserstein distance. The first one states that distances between Diracs match the ground metric:

Lemma 3. *For $x,y \in \mathbb{R}^d$ and $k \in [\![d]\!]$,*

$$\mathcal{S}_k(\delta_x,\delta_y) = \|x-y\|.$$

Metric Equivalence. Subspace robust Wasserstein distances $\mathcal{S}_k$ are equivalent to the Wasserstein distance $\mathcal{W}$:

Proposition 2. *For $k \in [\![d]\!]$, $\mathcal{S}_k$ is equivalent to $\mathcal{W}$. More precisely, for $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$,*

$$\sqrt{\frac{k}{d}}\mathcal{W}(\mu,\nu) \leq \mathcal{S}_k(\mu,\nu) \leq \mathcal{W}(\mu,\nu).$$

Moreover, the constants are tight since

$$\begin{aligned}
\mathcal{S}_k(\delta_x,\delta_y) &= \mathcal{W}(\delta_x,\delta_y)\\
\mathcal{S}_k(\delta_0,\sigma) &= \sqrt{\tfrac{k}{d}}\mathcal{W}(\delta_0,\sigma)
\end{aligned}$$

where $\delta_x,\delta_y,\delta_0$ are Dirac masses at points $x,y,0 \in \mathbb{R}^d$ and σ is the uniform probability distribution over the centered unit sphere in $\mathbb{R}^d$.

Dependence on the dimension. We fix $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ and we ask the following question : how does $\mathcal{S}_k(\mu,\nu)$ depend on the dimension $k \in [\![d]\!]$? The following lemma gives a result in terms of eigenvalues of V_{π_k}, where $\pi_k \in \Pi(\mu,\nu)$ is optimal for some dimension k, then we translate in Proposition 3 this result in terms of $\mathcal{S}_k$.

Lemma 4. *Let $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$. For any $k \in [\![d-1]\!]$,*

$$\begin{aligned}
\lambda_{k+1}(V_{\pi_{k+1}}) &\leq \mathcal{S}^2_{k+1}(\mu,\nu) - \mathcal{S}^2_k(\mu,\nu)\\
&\leq \lambda_{k+1}(V_{\pi_k})
\end{aligned}$$

where for $L \in [\![d]\!]$, $\pi_L \in \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{L} \lambda_l(V_\pi)$.

Proposition 3. *Let $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$. The sequence $k \mapsto \mathcal{S}^2_k(\mu,\nu)$ is increasing and concave. In particular, for $k \in [\![d-1]\!]$,*

$$\mathcal{S}^2_{k+1}(\mu,\nu) - \mathcal{S}^2_k(\mu,\nu) \geq \frac{\mathcal{W}^2(\mu,\nu) - \mathcal{S}^2_k(\mu,\nu)}{d-k}.$$

Moreover, for any $k \in [\![d-1]\!]$,

$$\mathcal{S}_k(\mu,\nu) \leq \mathcal{S}_{k+1}(\mu,\nu) \leq \sqrt{\frac{k+1}{k}}\mathcal{S}_k(\mu,\nu).$$

Geodesics We have shown in Proposition 2 that for any $k \in [\![d]\!]$, $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is a metric space with the same topology as that of the Wasserstein space $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{W}\right)$. We conclude this section by showing that $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is in fact a geodesic length space, and exhibits explicit constant speed geodesics. This can be used to interpolate between measures in $\mathcal{S}_k$ sense.

Proposition 4. *Let $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ and $k \in [\![d]\!]$. Take*

$$\pi^* \in \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi)$$

and let $f_t(x,y) = (1-t)x + ty$. Then the curve

$$t \mapsto \mu_t := {f_t}_\#\pi^*$$

is a constant speed geodesic in $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ connecting μ and ν. Consequently, $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is a geodesic space.

Proof. We first show that for any $s,t \in [0,1]$,

$$\mathcal{S}_k(\mu_s,\mu_t) = |t-s|\mathcal{S}_k(\mu,\nu)$$

by computing the cost of the transport plan $\pi(s,t) = (f_s,f_t)_\#\pi^* \in \Pi(\mu_s,\mu_t)$ and using the triangular inequality. Then the curve (μ_t) has constant speed

$$|\mu_t'| = \lim_{\epsilon\to 0} \frac{\mathcal{S}_k(\mu_{t+\epsilon},\mu_t)}{|\epsilon|} = \mathcal{S}_k(\mu,\nu),$$

and the length of the curve (μ_t) is

$$\sup \left\{ \sum_{i=0}^{n-1} \mathcal{S}_k(\mu_{t_i}, \mu_{t_{i+1}}) \,\middle|\, \begin{array}{c} n \geq 1 \\ 0 = t_0 < \ldots < t_n = 1 \end{array} \right\} = \mathcal{S}_k(\mu, \nu),$$

i.e. (μ_t) is a geodesic connecting μ and ν. $\square$

5. Computation

We provide in this section algorithms to compute the saddle point solution of $\mathcal{S}_k$. μ, ν are now discrete with respectively n and m points and weights a and b : $\mu := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu := \sum_{j=1}^m b_i \delta_{y_j}$. For $k \in [\![d]\!]$, three different objects are of interest: *(i)* the value of $\mathcal{S}_k(\mu, \nu)$, *(ii)* an optimal subspace E^* obtained through the relaxation for SRW, *(iii)* an optimal transport plan solving SRW. A subspace can be used for dimensionality reduction, whereas an optimal transport plan can be used to compute a geodesic, *i.e.* to interpolate between μ and ν.

5.1. Computational challenges to approximate $\mathcal{S}_k$

We observe that solving $\min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^k \lambda_l(V_\pi)$ is challenging. Considering a direct projection onto the transportation polytope

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathbb{R}^{n \times m} \,|\, \pi \mathbf{1}_m = a,\, \pi^T \mathbf{1}_n = b \right\}$$

would result in a costly quadratic network flow problem. The Frank-Wolfe algorithm, which does not require such projections, cannot be used directly because the application $\pi \mapsto \sum_{l=1}^k \lambda_l(V_\pi)$ is not smooth.

On the other hand, thanks to Theorem 1, solving the maximization problem is easier. Indeed, we can project onto the set of constraints $\mathcal{R} = \{\Omega \in \mathbb{R}^{d \times d} \,|\, 0 \preceq \Omega \preceq I\,;\, \mathrm{trace}(\Omega) = k\}$ using Dykstra's projection algorithm (Boyle & Dykstra, 1986). In this case, we will only get the value of $\mathcal{S}_k(\mu, \nu)$ and an optimal subspace, but not necessarily the actual optimal transport plan due to the lack of uniqueness for OT plans in general.

Smoothing It is well known that saddle points are hard to compute for a bilinear objective (Hammond, 1984). Computations are greatly facilitated by adding smoothness, which allows the use of saddle point Frank-Wolfe algorithms (Gidel et al., 2017). Out of the two problems, the maximization problem is seemingly easier. Indeed, we can leverage the framework of regularized OT (Cuturi, 2013) to output, using Sinkhorn's algorithm, a unique optimal transport plan $\pi^\star$ at each inner loop of the maximization. To save time, we remark that initial iterations can be solved with a low accuracy by limiting the number of iterations, and benefit from warm starts, using the scalings computed at the previous iteration, see (Peyré & Cuturi, 2019, §4).

Algorithm 1 Projected supergradient method for SRW

Input: Measures (x_i, a_i) and (y_j, b_j), dimension k, learning rate τ_0
$\pi \leftarrow \mathrm{OT}((x, a), (y, b), cost = \|\cdot\|^2)$
$U \leftarrow$ top k eigenvectors of V_π
Initialize $\Omega = UU^T \in \mathbb{R}^{d \times d}$
for $t = 0$ **to** max_iter **do**
  $\pi \leftarrow \mathrm{OT}((x, a), (y, b), cost = d^2_\Omega)$
  $\tau = \tau_0/(t+1)$
  $\Omega \leftarrow \mathrm{Proj}_{\mathcal{R}}\left[\Omega + \tau V_\pi\right]$
end for
Output: $\Omega, \langle \Omega \,|\, V_\pi \rangle$

5.2. Projected Supergradient Method for SRW

In order to compute SRW and an optimal subspace, we can solve equation (3) by maximizing the concave function

$$f : \Omega \mapsto \min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j} d^2_\Omega(x_i, y_j)\, \pi_{i,j} = \min_{\pi \in \Pi(\mu,\nu)} \langle \Omega \,|\, V_\pi \rangle$$

over the convex set $\mathcal{R}$. Since f is not differentiable, but only superdifferentiable, we can only use a projected supergradient method. This algorithm is outlined in Algorithm 1. Note that by Danskin's theorem, for any $\Omega \in \mathcal{R}$,

$$\partial f(\Omega) = \mathrm{Conv}\left\{ V_{\pi^*} \,\middle|\, \pi^* \in \operatorname*{arg\,min}_{\pi \in \Pi(\mu,\nu)} \langle \Omega \,|\, V_\pi \rangle \right\}.$$

5.3. Frank-Wolfe using Entropy Regularization

Entropy-regularized optimal transport can be used to compute a unique optimal plan given a subspace. Let $\gamma > 0$ be the regularization strength. In this case, we want to maximize the concave function

$$f_\gamma : \Omega \mapsto \min_{\pi \in \Pi(\mu,\nu)} \langle \Omega \,|\, V_\pi \rangle + \gamma \sum_{i,j} \pi_{i,j} \left[\log(\pi_{i,j}) - 1\right]$$

over the convex set $\mathcal{R}$. Since for all $\Omega \in \mathcal{R}$, there is a unique π^* minimizing $\pi \mapsto \langle \Omega \,|\, V_\pi \rangle + \gamma \sum_{i,j} \pi_{i,j} \left[\log(\pi_{i,j}) - 1\right]$, f_γ is differentiable. Instead of running a projected gradient ascent on $\Omega \in \mathcal{R}$, we propose to use the Frank-Wolfe algorithm when the regularization strength is positive. Indeed, there is no need to tune a learning rate in Frank-Wolfe, making it easier to use. We only need to compute, for fixed $\pi \in \Pi(\mu, \nu)$, the maximum over $\mathcal{R}$ of $\Omega \mapsto \langle \Omega \,|\, V_\pi \rangle$:

Lemma 5. *For $\pi \in \Pi(\mu, \nu)$, compute the eigendecomposition of $V_\pi = U \,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_d)\, U^T$ with $\lambda_1 \geq \ldots \geq \lambda_d$. Then for $k \in [\![d]\!]$, $\widehat{\Omega} = U \,\mathrm{diag}\,([\mathbf{1}_k, \mathbf{0}_{d-k}])\, U^T$ solves*

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega) = k}} \int d^2_\Omega \, d\pi.$$

This algorithm is outlined in algorithm 2.

Algorithm 2 Frank-Wolfe algorithm for regularized SRW

Input: Measures (x_i, a_i) and (y_j, b_j), dimension k, regularization strength $\gamma > 0$, precision $\epsilon > 0$
$\pi \leftarrow \text{reg\_OT}((x, a), (y, b), reg = \gamma, cost = \|\cdot\|^2)$
$U \leftarrow$ top k eigenvectors of V_π
Initialize $\Omega = UU^T \in \mathbb{R}^{d\times d}$
for $t = 0$ **to** max_iter **do**
  $\pi \leftarrow \text{reg\_OT}((x, a), (y, b), reg = \gamma, cost = d^2_\Omega)$
  $U \leftarrow$ top k eigenvectors of V_π
  if $\sum_{l=1}^k \lambda_l(V_\pi) - \langle \Omega \,|\, V_\pi\rangle \leq \epsilon \langle \Omega \,|\, V_\pi\rangle$ **then**
    break
  end if
  $\widehat{\Omega} \leftarrow U \,\text{diag}\,([\mathbf{1}_k, \mathbf{0}_{d-k}])\, U^T$
  $\tau = 2/(2 + t)$
  $\Omega \leftarrow (1 - \tau)\Omega + \tau\widehat{\Omega}$
end for
Output: $\Omega, \pi, \langle \Omega \,|\, V_\pi\rangle$

5.4. Initialization and Stopping Criterion

We propose to initialize Algorithms 1 and 2 with $\Omega_0 = UU^T$ where $U \in \mathbb{R}^{d\times k}$ is the matrix of the top k eigenvectors (*i.e.* the eigenvectors associated with the top k eigenvalues) of V_{π^*} and π^* is an optimal transport plan between μ and ν. In other words, Ω_0 is the projection matrix onto the k first principal components of the transport-weighted displacement vectors. Note that Ω_0 would be optimal is π^* were optimal for the min-max problem, and that this initialization only costs the equivalent of one iteration.

When entropic regularization is used, Sinkhorn algorithm is run at each iteration of Algorithms 1 and 2. We propose to initialize the potentials in Sinkhorn algorithm with the latest computed potentials, so that the number of iterations in Sinkhorn algorithm should be small after a few iterations of Algorithms 1 or 2.

We sometimes need to compute $\mathcal{S}_k(\mu, \nu)$ for all $k \in [\![d]\!]$, for example to choose the optimal k with an "elbow" rule. To speed the computations up, we propose to compute this sequence iteratively from $k = d$ to $k = 1$. At each iteration, *i.e.* for each dimension k, we initialize the algorithm with $\Omega_0 = UU^T$, where $U \in \mathbb{R}^{d\times k}$ is the matrix of the top k eigenvectors of $V_{\pi_{k+1}}$ and π_{k+1} is the optimal transport plan for dimension $k + 1$. We also initialize the Sinkhorn algorithm with the latest computed potentials.

Instead of running a fixed number of iterations in Algorithm 2, we propose to stop the algorithm when the computation error is smaller than a fixed threshold ϵ. The computation error at iteration t is:

$$\frac{|\mathcal{S}_k(\mu, \nu) - \widehat{\mathcal{S}_k}(t)|}{\mathcal{S}_k(\mu, \nu)} \leq \frac{\Delta(t)}{\widehat{\mathcal{S}_k}(t)}$$

where $\widehat{\mathcal{S}_k}(t)$ is the computed "max-min" value and $\Delta(t)$ is the duality gap at iteration t. We stop as soon as $\Delta(t)/\widehat{\mathcal{S}_k}(t) \leq \epsilon$.

6. Experiments

We first compare SRW with the experimental setup used to evaluate FactoredOT (Forrow et al., 2019). We then study the ability of SRW distances to capture the dimension of sampled measures by looking at their value for increasing dimensions k, as well as their robustness to noise.

6.1. Fragmented Hypercube

We first consider $\mu = \mathcal{U}([-1, 1])^d$ to be uniform over an hypercube, and $\nu = T_\#\mu$ the pushforward of μ under the map $T(x) = x + 2\,\text{sign}(x) \odot (\sum_{k=1}^{k^*} e_k)$, where sign is taken elementwise, $k^* \in [\![d]\!]$ and $(e_1, ..., e_d)$ is the canonical basis of $\mathbb{R}^d$. The map T splits the hypercube into four different hyperrectangles. T is a subgradient of a convex function, so by Brenier's theorem (1991) it is an optimal transport map between μ and $\nu = T_\#\mu$ and

$$\mathcal{W}^2(\mu, \nu) = \int \|x - T(x)\|^2 \, d\mu(x) = 4k^*.$$

Note that for any x, the displacement vector $T(x) - x$ lies in the k^*-dimensional subspace $\text{span}\{e_1, ..., e_{k^*}\} \in \mathcal{G}_{k^*}$, which is optimal. This means that for $k \geq k^*$, $\mathcal{S}^2_k(\mu, \nu)$ is constant equal to $4k^*$. We show the interest of plotting, based on two empirical distributions $\hat{\mu}$ from μ and $\hat{\nu}$ from ν, the sequence $k \mapsto \mathcal{S}^2_k(\hat{\mu}, \hat{\nu})$, for different values of k^*. That sequence is increasing concave by proposition 3, and increases more slowly after $k = k^*$, as can be seen on Figure 2. This is the case because the last $d - k^*$ dimensions only represent noise, but is recovered in our plot.

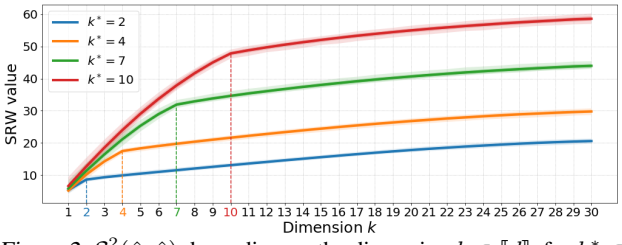

Figure 2. $\mathcal{S}^2_k(\hat{\mu}, \hat{\nu})$ depending on the dimension $k \in [\![d]\!]$, for $k^* \in \{2, 4, 7, 10\}$, where $\hat{\mu}, \hat{\nu}$ are empirical measures from μ and ν respectively with 100 points each. Each curve is the mean over 100 samples, and shaded area show the min and max values.

We consider next $k^* = 2$, and choose from the result of Figure 2, $k = 2$. We look at the estimation error

$|\mathcal{W}^2(\mu,\nu) - \mathcal{S}_k^2(\hat{\mu},\hat{\nu})|$ where $\hat{\mu},\hat{\nu}$ are empirical measures from μ and ν respectively with n points each. In Figure 3, we plot this estimation error depending on the number of points n. In Figure 4, we plot the subspace estimation error $\|\Omega^* - \widehat{\Omega}\|$ depending on n, where Ω^* is the optimal projection matrix onto span$\{e_1, e_2\}$.

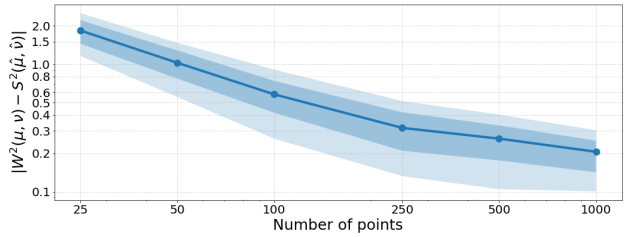

Figure 3. Mean estimation error over 500 random samples for n points, $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

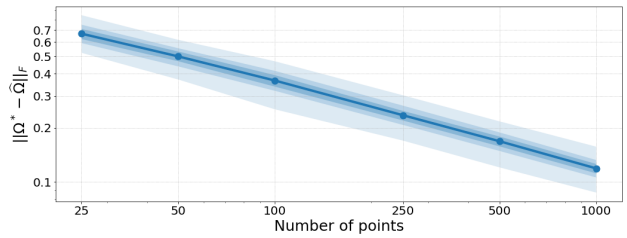

Figure 4. Mean estimation of the subspace estimation error over 500 samples, depending on $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

We also plot the optimal transport plan (in the sense of $\mathcal{W}$, Figure 5 left) and the optimal transport plan (in the sense of $\mathcal{S}_2$) between $\hat{\mu}$ and $\hat{\nu}$ (with $n = 250$ points each, Figure 5 right).

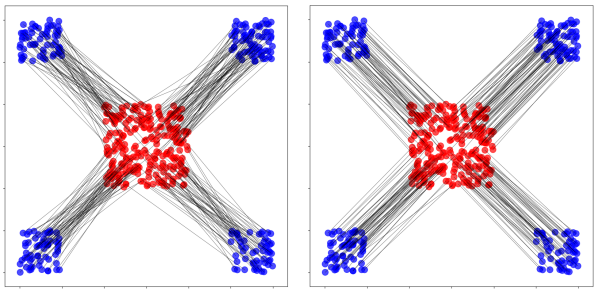

Figure 5. Fragmented hypercube, $n = 250$, $d = 30$. Optimal mapping in the Wasserstein space (left) and in the SRW space (right). Geodesics in the SRW space are robust to statistical noise.

6.2. Robustness, with 20-D Gaussians

We consider $\mu = \mathcal{N}(0, \Sigma_1)$ and $\nu = \mathcal{N}(0, \Sigma_2)$, with $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ semidefinite positive of rank k. It means that the supports of μ and ν are k-dimensional subspaces of $\mathbb{R}^d$. Although those two subspaces are k-dimensional, they may be different. Since the union of two k-dimensional subspaces is included in a $2k$-dimensional subspace, for any $l \geq 2k$, $\mathcal{S}_l^2(\mu,\nu) = \mathcal{W}^2(\mu,\nu)$.

For our experiment, we simulated 100 independent couples of covariance matrices Σ_1, Σ_2 in dimension $d = 20$, each having independently a Wishart distribution with $k = 5$ degrees of freedom. For each couple of matrices, we draw $n = 100$ points from $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ and considered $\hat{\mu}$ and $\hat{\nu}$ the empirical measures on those points. In Figure 6, we plot the mean (over the 100 samples) of $l \mapsto \mathcal{S}_l^2(\hat{\mu},\hat{\nu})/\mathcal{W}^2(\hat{\mu},\hat{\nu})$. We plot the same curve for noisy data, where each point was added a $\mathcal{N}(0, I)$ random vector. With moderate noise, the data is only approximately on the two $k = 5$-dimensional subspaces, but the SRW does not vary too much.

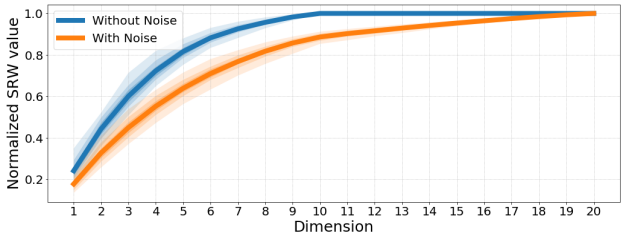

Figure 6. Mean normalized SRW distance, depending on the dimension. The shaded area show the 10%-90% and 25%-75% quantiles and the minimum and maximum values over the 100 samples.

6.3. $\mathcal{S}_k$ is Robust to Noise

As in experiment 6.2, we consider 100 independent samples of couples $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$, each following independently a Wishart distribution with $k = 5$ degrees of freedom. For each couple, we draw $n = 100$ points from $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ and consider the empirical measures $\hat{\mu}$ and $\hat{\nu}$ on those points. We then gradually add Gaussian noise $\sigma\mathcal{N}(0, I)$ to the points, giving measures $\hat{\mu}_\sigma, \hat{\nu}_\sigma$. In Figure 7, we plot the mean (over the 100 samples) of the relative errors

$$\sigma \mapsto \frac{|\mathcal{S}_5^2(\hat{\mu}_\sigma,\hat{\nu}_\sigma) - \mathcal{S}_5^2(\hat{\mu}_0,\hat{\nu}_0)|}{\mathcal{S}_5^2(\hat{\mu}_0,\hat{\nu}_0)}$$

and

$$\sigma \mapsto \frac{|\mathcal{W}^2(\hat{\mu}_\sigma,\hat{\nu}_\sigma) - \mathcal{W}^2(\hat{\mu}_0,\hat{\nu}_0)|}{\mathcal{W}^2(\hat{\mu}_0,\hat{\nu}_0)}.$$

Note that for small noise level, the imprecision in the computation of the SRW distance adds up to the error caused by the added noise. SRW distances seem more robust to noise than the Wasserstein distance when the noise has moderate to high variance.

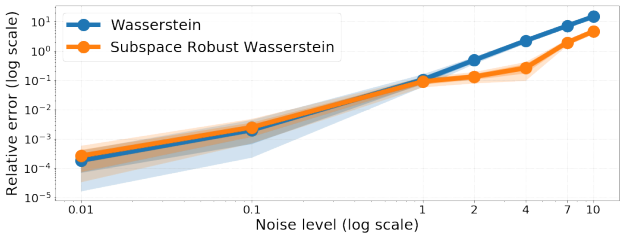

Figure 7. Mean SRW distance over 100 samples, depending on the noise level. Shaded areas show the min-max values and the 10%-90% quantiles.

6.4. Computation time

We consider the Fragmented Hypercube experiment, with increasing dimension d and fixed $k^* = 2$. Using $k = 2$ and Algorithm 2 with $\gamma = 0.1$ and stopping threshold $\epsilon = 0.05$, we plot in Figure 8 the mean computation time of both SRW and Wasserstein distances on GPU, over 100 random samplings with $n = 100$. It shows that SRW computation is quadratic in dimension d, because of the eigendecomposition of matrix V_π in Algorithm 2.

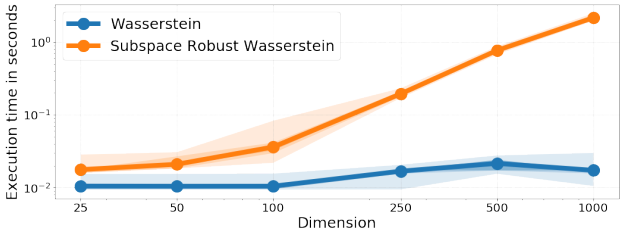

Figure 8. Mean computation times on GPU (log-log scale). The shaded areas show the minimum and maximum values over the 100 experiments.

6.5. Real Data Experiment

We consider the scripts of seven movies. Each script is transformed into a list of words, and using word2vec (Mikolov et al., 2018), into a measure over $\mathbb{R}^{300}$ where the weights correspond to the frequency of the words. We then compute the SRW distance between all pairs of films: see Figure 9 for the SRW values. Movies with a same genre or thematic tend to be closer to each other: this can be visualized by running a two-dimensional metric multidimensional scaling (mMDS) on the SRW distances, as shown in Figure 10 (left).

In Figure 10 (right), we display the projection of the two measures associated with films *Kill Bill Vol.1* and *Interstellar* onto their optimal subspace. We compute the first (weighted) principal component of each projected measure, and find among the whole dictionary their 5 nearest neighbors in terms of cosine similarity. For *Kill Bill Vol.1* , these are: 'swords', 'hull', 'sword', 'ice', 'blade'. For *Interstellar*, they are: 'spacecraft', 'planets', 'satellites', 'asteroids',

	D	*G*	*I*	*KB1*	*KB2*	*TM*	*T*
D	0	0.186	0.186	0.195	0.203	0.186	**0.171**
G	0.186	0	**0.173**	0.197	0.204	0.176	0.185
I	0.186	0.173	0	0.196	0.203	**0.171**	0.181
KB1	0.195	0.197	0.196	0	**0.165**	0.190	0.180
KB2	0.203	0.204	0.203	**0.165**	0	0.194	0.180
TM	0.186	0.176	**0.171**	0.190	0.194	0	0.183
T	**0.171**	0.185	0.181	0.180	0.180	0.183	0

Figure 9. $\mathcal{S}_k^2$ distances between different movie scripts. Bold values correspond to the minimum of each line. *D*=Dunkirk, *G*=Gravity, *I*=Interstellar, *KB1*=Kill Bill Vol.1, *KB2*=Kill Bill Vol.2, *TM*=The Martian, *T*=Titanic.

'planet'. The optimal subspace recovers the semantic dissimilarities between the two films.

Figure 10. Left: Metric MDS projection for the distances of Figure 9. *Right:* Optimal 2-dimensional projection between *Kill Bill Vol.1* (red) and *Interstellar* (blue). Words appearing in both scripts are displayed in violet. For clarity, only the 30 most frequent words of each script are displayed.

7. Conclusion

We have proposed in this paper a new family of OT distances with robust properties. These distances take a particular interest when used with a squared-Euclidean cost, in which case they have several properties, both theoretical and computational. These distances share important properties with the 2-Wasserstein distance, yet seem far more robust to random perturbation of the data and able to capture better signal. We have provided algorithmic tools to compute these SRW distance. They come at a relatively modest overhead, given that they require using regularized OT as the inner loop of a FW type algorithm. Future work includes the investigation of even faster techniques to carry out these computations, eventually automatic differentiation schemes as those currently benefitting the simple use of Sinkhorn divergences.

Acknowledegments. Both authors acknowledge the support of a "Chaire d'excellence de l'IDEX Paris Saclay". We thank P. Rigollet and J. Weed for their remarks and their valuable input.

References

Altschuler, J., Bach, F., Rudi, A., and Weed, J. Approximating the quadratic transportation metric in near-linear time. *arXiv preprint arXiv:1810.10046*, 2018a.

Altschuler, J., Bach, F., Rudi, A., and Weed, J. Massively scalable sinkhorn distances via the nyström method. *arXiv preprint arXiv:1812.05189*, 2018b.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 880–889, 2018.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Boyle, J. P. and Dykstra, R. L. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pp. 28–47. Springer, 1986.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2492–2500. 2012.

Cotter, A., Keshet, J., and Srebro, N. Explicit approximations of the gaussian kernel. *arXiv preprint arXiv:1109.4603*, 2011.

Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

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

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced wasserstein distance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Dessein, A., Papadakis, N., and Rouas, J.-L. Regularized optimal transport and the rot mover's distance. *Journal of Machine Learning Research*, 19(15):1–53, 2018.

Dudley, R. M. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Fan, K. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

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

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. Statistical optimal transport via factored couplings. 2019.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. 2019.

Gidel, G., Jebara, T., and Lacoste-Julien, S. Frank-Wolfe algorithms for saddle point problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Hammond, J. H. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Kolouri, S., Martin, C. E., and Rohde, G. K. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

Ling, H. and Okada, K. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (5):840–853, 2007.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

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

Overton, M. L. and Womersley, R. S. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357, 1993.

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pp. 460–467, 2009.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Santambrogio, F. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

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