Subspace Robust Wasserstein Distances

FRANÇOIS-PIERRE PATY
CREST, ENSAE
francois.pierre.paty@ensae.fr

MARCO CUTURI
Google Brain and CREST, ENSAE
cuturi@google.com

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 unrealistic 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 $\ell_1$ 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 distributions 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 function 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 several random lines. This approach can be used to define kernels (Kolouri et al., 2016), compute...
The maximal eigenvalue of the second order moment matrix (still in dimension 2) is smaller than that obtained with the initial OT plan. We show that computing a maximally 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 to 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.

**Contributions** 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 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 to 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. 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.

---

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 $\langle \pi, C \rangle$, where $C$ is the pairwise squared-Euclidean distance matrix. The largest eigenvalue of the corresponding second order moment matrix $V_{\pi}$ 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 $\theta$ 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 $\theta$. 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 $P_\theta$ and subspace $S_\theta$ 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 15. The smallest eigenvalue is given for illustrative purposes only.
2 Background on Optimal Transport

For $d \in \mathbb{N}$, we write $[d] = \{1, \ldots, d\}$. Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures in $\mathbb{R}^d$, and let
\[
\mathcal{P}_2(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \middle| \int \|x\|^2 d\mu(x) < \infty \right\}.
\]

**Monge and Kantorovich Formulations of OT** For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we write $\Pi(\mu, \nu)$ for the set of couplings $\Pi(\mu, \nu) = \{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \mid \forall A, B \subset \mathbb{R}^d \text{ Borel,} \quad \pi(A \times \mathbb{R}^d) = \mu(A), \pi(\mathbb{R}^d \times B) = \nu(B) \}$, and their 2-Wasserstein distance is defined as
\[
W_2(\mu, \nu) := \left( \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^2 d\pi(x, y) \right)^{1/2}.
\]

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 $\mu \in \mathcal{P}(\mathcal{X})$, we denote by $T_\# \mu \in \mathcal{P}(\mathcal{Y})$ the pushforward of $\mu$ 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
\[
\mathcal{W}(\mu, \nu) = \left( \inf_{T: T_\# \mu = \nu} \int \|x - T(x)\|^2 d\mu(x) \right)^{1/2}.
\]

**W as Trace-minimization** For any coupling $\pi$, we define the $d \times d$ second order displacement matrix
\[
V_\pi := \int (x - y)(x - y)^T d\pi(x, y).
\]
(1)

Notice that when a coupling $\pi$ corresponds to a Monge map, namely $\pi = (\text{Id}, T)_\# \mu$, then one can interpret even more naturally $V_\pi$ as the second order moment of all displacement $(x - T(x))(x - T(x))^T$ weighted by $\mu$. With this convention, we remark that the total cost of a coupling $\pi$ is equal to the trace of $V_\pi$, using the simple identity 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 $V_\pi$. This simple observation will play an important role in the next section, and more specifically the study of $\lambda_1(V_\pi)$, the $l$-th largest eigenvalue of $V_\pi$.

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 $\mathbb{R}^d$:
\[
\mathcal{G}_k = \left\{ E \subset \mathbb{R}^d \mid \dim(E) = k \right\}.
\]

For $E \in \mathcal{G}_k$, we note $P_E$ the orthogonal projector onto $E$. Given two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, 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 $\mu$ and $\nu$ is
\[
P_k(\mu, \nu) = \sup_{E \in \mathcal{G}_k} \mathcal{W}(P_E \# \mu, P_E \# \nu).
\]

As we show in the appendix, 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 $\mu$ and $\nu$ is
\[
S_k(\mu, \nu) = \left( \inf_{\pi \in \Pi(\mu, \nu)} \sup_{E \in \mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 d\pi(x, y) \right)^{1/2}.
\]

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

**Lemma 1.** Optimal solutions for $S_k$ exist, i.e.,
\[
S_k(\mu, \nu) = \left( \min_{\pi \in \Pi(\mu, \nu)} \max_{E \in \mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 d\pi(x, y) \right)^{1/2}.
\]

We show next that the SRW variant $S_k$ can be elegantly reformulated as a function of the displacement second-order moment matrix $V_\pi$ (1):

**Lemma 2.** For $k \in [d]$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, one has
\[
S_k^2(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \max_{U \in \mathbb{R}^{d \times d}} \sum_{l=1}^k \lambda_l(V_\pi(UU^T - I_k)).
\]
Proof. A classical variational result by (Fan, 1949) states that
\[ \sum_{l=1}^{k} \lambda_l(V_\pi) = \max_{U \in \mathbb{R}^{x \times d}} \frac{\text{trace} \left( UV_\pi U^T \right)}{U^T I_k U} . \]
Then plugging the definition of \( V_\pi \) and taking the minimum over \( \pi \in \Pi(\mu, \nu) \) yields the result. \( \Box \)

This characterization as a sum of eigenvalues will be crucial to study theoretical properties of \( S_k \). 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 \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \),
\[ S_k^2(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \max_{0 \leq \Omega \preceq I} \int_{\text{trace}(\Omega) = k} d_\Omega^2 \, d\pi \]
(2)
\[ = \max_{0 \leq \Omega \preceq I} \min_{\pi \in \Pi(\mu, \nu)} \int_{\text{trace}(\Omega) = k} d_\Omega^2 \, d\pi \]
(3)
\[ = \max_{0 \leq \Omega \preceq I} \mathcal{W}^2 \left( \Omega^{1/2} \pi, \Omega^{1/2} \nu \right) \]
(4)
where \( d_\Omega \) stands for the Mahalanobis distance
\[ d_\Omega(x, y) = (x - y)^T \Omega (x - y) . \]

Proof. \( S_k^2(\mu, \nu) = (2) \): We fix \( \pi \in \Pi(\mu, \nu) \) and focus on the inner maximization in (2):
\[ \max_{0 \leq \Omega \preceq I} \int_{\text{trace}(\Omega) = k} d_\Omega^2 \, d\pi = \max_{0 \leq \Omega \preceq I} \langle \Omega | V_\pi \rangle . \]
A result by (Overton & Womersley, 1993) shows that this is equal to
\[ \max_{U \in \mathbb{R}^{x \times d}} \frac{\text{trace} \left( UV_\pi U^T \right)}{U^T I_k U} \]
which is nothing but the sum of the \( k \) largest eigenvalues of \( V_\pi \) by Fan’s result. By lemma 2, taking the minimum over \( \pi \in \Pi(\mu, \nu) \) yields the result. To prove \( (2) = (3) \) we use Sion’s minimax theorem to interchange the minimum and the maximum. For \( (3) = (4) \), we focus on the inner minimization and use Lemma 6 in Appendix to show that OT with ground cost \( d_\Omega^2 \) is equal to the Wasserstein distance between the pushforwards measures under \( \Omega^{1/2} \).

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

**Proposition 1.** For \( k \in [d] \), both \( P_k \) and \( S_k \) are distances over \( \mathcal{P}_2(\mathbb{R}^d) \).

Proof. Symmetry is clear for both objects, and for \( \mu \in \mathcal{P}_2(\mathbb{R}^d) \), \( S_k(\mu, \mu) = P_k(\mu, \mu) = 0 \). Let \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \) such that \( S_k(\mu, \nu) = 0 \). Then \( P_k(\mu, \nu) = 0 \) and for any \( E \in G_k \), \( \mathcal{W}(P_{E#\mu}, P_{E#\nu}) = 0 \), i.e. \( P_{E#\mu} = P_{E#\nu} \).

Lemma 7 (in Appendix) then shows that \( \mu = \nu \). For the triangle inequalities, let \( \mu_0, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d) \). Let \( \Omega_k \in \{ 0 \leq \Omega \leq I, \text{trace}(\Omega) = k \} \) be optimal between \( \mu_0 \) and \( \mu_2 \).

Using the triangle inequalities for the Wasserstein distance,
\[ S_k(\mu_0, \mu_2) = \mathcal{W} \left[ \Omega_0^{1/2} \mu_0, \Omega_2^{1/2} \mu_2 \right] \]
\[ \leq \mathcal{W} \left[ \Omega_0^{1/2} \mu_0, \Omega_0^{1/2} \mu_1 \right] + \mathcal{W} \left[ \Omega_0^{1/2} \mu_1, \Omega_2^{1/2} \mu_2 \right] \]
\[ \leq \sup_{0 \leq \Omega \preceq I} \mathcal{W} \left[ \Omega_0^{1/2} \mu_0, \Omega_0^{1/2} \mu_1 \right] \]
\[ + \sup_{0 \leq \Omega \preceq I} \mathcal{W} \left[ \Omega_0^{1/2} \mu_1, \Omega_2^{1/2} \mu_2 \right] \]
\[ = S_k(\mu_0, \mu_1) + S_k(\mu_1, \mu_2) . \]
The same argument, used this time with projections, yields the triangle inequalities for \( P_k \). \( \Box \)

**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 (see proof in Appendix):

**Lemma 3.** For \( x, y \in \mathbb{R}^d \) and \( k \in [d] \),
\[ S_k(\delta_x, \delta_y) = \| x - y \| . \]

**Metric Equivalence.** Subspace robust Wasserstein distances \( S_k \) are equivalent to the Wasserstein distance \( \mathcal{W} \):

**Proposition 2.** For \( k \in [d] \), \( S_k \) is equivalent to \( \mathcal{W} \). More precisely, for \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \),
\[ \sqrt{\frac{k}{d}} \, \mathcal{W}(\mu, \nu) \leq S_k(\mu, \nu) \leq \mathcal{W}(\mu, \nu) . \]
Moreover, the constants are tight since
\[ S_k(\delta_x, \delta_y) = \mathcal{W}(\delta_x, \delta_y) \]
\[ S_k(\delta_0, \sigma) = \sqrt{\frac{k}{d}} \mathcal{W}(\delta_0, \sigma) \]
where \( \delta_x, \delta_y, \delta_0 \) are Dirac masses at points \( x, y, 0 \in \mathbb{R}^d \) and \( \sigma \) is the uniform probability distribution over the centered unit sphere in \( \mathbb{R}^d \).

Proof. The upper bound comes from the fact that \( x \mapsto \Omega^{1/2}x \) is 1-Lipschitz for any \( 0 \leq \Omega \leq I \). For the lower bound, we consider the finite set \( \mathcal{B}_k \subset \mathcal{G}_k \) of subspaces spanned by \( k \) vectors of the canonical basis of \( \mathbb{R}^d \):
\[ \mathcal{B}_k = \left\{ \text{span}(e_{\sigma(1)}, \ldots, e_{\sigma(k)}) \mid \sigma \in \mathcal{S}_d \right\}. \]
Then taking the maximum over \( \mathcal{B}_k \) gives the result. \( \square \)

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

Lemma 4. Let \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \). For any \( k \in [d - 1] \),
\[ \lambda_{k+1}(V_{\pi_{k+1}}) \leq S_{k+1}^2(\mu, \nu) - S_k^2(\mu, \nu) \leq \lambda_{k+1}(V_{\pi_k}) \]
where for \( L \in [d] \), \( \pi_L = \arg \min_{\pi \in \Pi(\mu, \nu)} \sum_{i=1}^L \lambda_i(V_\pi) \).

Proposition 3. Let \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \). The sequence \( k \mapsto S_k^2(\mu, \nu) \) is increasing and concave. In particular, for \( k \in [d - 1] \),
\[ S_{k+1}^2(\mu, \nu) - S_k^2(\mu, \nu) \geq \frac{\mathcal{W}^2(\mu, \nu) - S_k^2(\mu, \nu)}{d - k}. \]
Moreover, for any \( k \in [d - 1] \),
\[ S_k(\mu, \nu) \leq S_{k+1}(\mu, \nu) \leq \sqrt{\frac{k+1}{k}} S_k(\mu, \nu). \]
The results are direct consequences of Lemma 4, calculations can be found in Appendix.

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

Proposition 4. Let \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \) and \( k \in [d] \). Take
\[ \pi^* \in \arg \min_{\pi \in \Pi(\mu, \nu)} \sum_{i=1}^k \lambda_i(V_\pi) \]
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 \( (\mathcal{P}_2(\mathbb{R}^d), S_k) \) connecting \( \mu \) and \( \nu \). Consequently, \( (\mathcal{P}_2(\mathbb{R}^d), S_k) \) is a geodesic space.

Proof. We first show that for any \( s, t \in [0, 1] \),
\[ S_k(\mu_s, \mu_t) \leq |t - s| S_k(\mu, \nu) \]
by computing the cost of the transport plan \( \pi(s, t) = (f_s, f_t)_{\pi^*} \in \Pi(\mu_s, \mu_t) \). Then using the triangular inequality, we show that in fact
\[ S_k(\mu_s, \mu_t) = |t - s| S_k(\mu, \nu) \]
for all \( s, t \in [0, 1] \). Then the curve \( (\mu_t) \) has constant speed
\[ |\mu'_t| = \lim_{\epsilon \to 0} \frac{S_k(\mu_{t+\epsilon}, \mu_t)}{\epsilon} = S_k(\mu, \nu), \]
and the length of the curve \( (\mu_t) \) is
\[ \sup \left\{ \sum_{i=0}^{n-1} S_k(\mu_{t_i}, \mu_{t_{i+1}}) \mid 0 = t_0 < \ldots < t_n = 1 \right\} = S_k(\mu, \nu), \]
i.e. \( (\mu_t) \) is a geodesic connecting \( \mu \) and \( \nu \). \( \square \)

5 Computation

We provide in this section algorithms to compute the saddle point solution of \( S_k \). \( \mu, \nu \) 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 \( \nu := \sum_{j=1}^m b_j \delta_{y_j} \). For \( k \in [d] \), three
different objects are of interest: (i) the value of $S_k(\mu, \nu)$, (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 $\mu$ and $\nu$.

5.1 Computational challenges to approximate $S_k$

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

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

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 $\pi \mapsto \sum_{i=1}^{k} \lambda_i(V_{\pi})$ 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 $\mathcal{R} = \{ \Omega \in \mathbb{R}^{d \times d} \mid 0 \preceq \Omega \preceq I; \operatorname{trace}(\Omega) = k \}$ using Dykstra’s projection algorithm (Boyle & Dykstra, 1986). In this case, we will only get the value of $S_k(\mu, \nu)$ and an optimal subspace, but not necessarily to 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 $\pi^*$ 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).

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

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

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 $\Omega \in \mathcal{R}$,

$$\partial f(\Omega) = \operatorname{Conv} \left\{ V_{\pi^*} \mid \pi^* \in \arg \min_{\pi \in \Pi(\mu, \nu)} \langle \Omega \mid V_{\pi} \rangle \right\}.$$

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

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

over the convex set $\mathcal{R}$. Since for all $\Omega \in \mathcal{R}$, there is a unique $\pi^*$ minimizing $\pi \mapsto \langle \Omega \mid V_{\pi} \rangle + \gamma \sum_{i,j} \pi_{i,j} [\log(\pi_{i,j}) - 1]$, $f_{\gamma}$ 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, so it is easier to use. We only need to compute, for fixed $\pi \in \Pi(\mu, \nu)$, the maximum over $\mathcal{R}$ of $\langle \Omega \mid V_{\pi} \rangle$:

**Lemma 5.** For $\pi \in \Pi(\mu, \nu)$, compute the eigenvalue decomposition of $V_{\pi} = U \operatorname{diag}(\lambda_1, ..., \lambda_d) U^T$ with $\lambda_1 \geq ... \geq \lambda_d$. Then for $k \in \{0, 1, ..., d\}$, $\Omega = U \operatorname{diag}(\{1_k, 0_{d-k}\}) U^T$ solves

$$\max_{0 \preceq \Omega \preceq I} \sum_{k=0}^{d} \langle \Omega \mid V_{\pi} \rangle = \langle \Omega \mid V_{\pi} \rangle$$

This algorithm is outlined in algorithm 2.
We sometimes need to compute $S_k$. We propose to initialize Algorithms 1 and 2 with

\begin{align*}
\text{Algorithm 1 Frank-Wolfe algorithm for regularized SRW} \\
\text{Input: Measures } (x_i, a_i) \text{ and } (y_j, b_j) \text{, dimension } k \text{, regularization strength } \gamma > 0 \text{, precision } \varepsilon > 0 \\
\pi \leftarrow \text{reg}_\gamma \text{OT}((x,a), (y,b), \text{reg } = \gamma, \text{cost } = \| \cdot \|^2) \\
U \leftarrow \text{top } k \text{ eigenvectors of } V_{\pi} \\
\text{for } t = 0 \text{ to max_iter do} \\
\pi \leftarrow \text{reg}_\gamma \text{OT}((x,a), (y,b), \text{reg } = \gamma, \text{cost } = d_{\pi}^2) \\
U \leftarrow \text{top } k \text{ eigenvectors of } V_{\pi} \\
\text{if } \sum_{i=1}^k \lambda_i (V_{\pi}) - \langle \Omega \mid V_{\pi} \rangle \leq \varepsilon (\Omega \mid V_{\pi}) \text{ then} \\
\text{break} \\
\end{align*}

end if

\begin{align*}
\Omega \leftarrow U \text{diag } ([k, 0_{d-k}]) U^T \\
\tau = 2(2 + t) \\
\Omega \leftarrow (1 - \tau) \Omega + \tau \hat{\Omega} \\
\end{align*}

end for

Output: $\Omega, \pi, \langle \Omega \mid V_{\pi} \rangle$

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 $V_{\pi^*}$ and $\pi^*$ is an optimal transport plan between $\mu$ and $\nu$. In other words, $\Omega_0$ is the projection matrix onto the $k$ first principal components of the transport-weighted displacement vectors. Note that $\Omega_0$ would be optimal if $\pi^*$ 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 $S_k(\mu, \nu)$ 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 $\pi_{k-1}$ 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 $\varepsilon$. The computation error at iteration $t$ is:

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

where $\hat{S}_k(t)$ is the computed “max-min” value and $\Delta(t)$ is the duality gap at iteration $t$. We stop as soon as $\Delta(t) / \hat{S}_k(t) \leq \varepsilon$.

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 $\mu = U([-1, 1]^d)$ to be uniform over an hypercube, and $\nu = T_{\#} \mu$ the pushforward of $\mu$ under the map $T(x) = x + 2 \text{sign}(x) \odot (\sum_{k=1}^d e_k)$, where sign is taken elementwise, $k^* \in \{d\}$ and $(e_1, \ldots, e_d)$ 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 $\mu$ and $\nu = T_{\#} \mu$ and

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

Note that for any $x$, the displacement vector $T(x) - x$ lies in the $k^*$-dimensional subspace span$\{e_1, \ldots, e_k\}$ in $\mathbb{R}^d$, which is optimal. This means that for $k \geq k^*$, $S_k^2(\mu, \nu)$ is constant equal to $4k^*$. We show the interest of plotting, based on two empirical distributions $\hat{\mu}$ from $\mu$ and $\hat{\nu}$ from $\nu$, the sequence $k \mapsto S_k^2(\hat{\mu}, \hat{\nu})$, 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.

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

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

$|W^2(\mu, \nu) - S_k^2(\hat{\mu}, \hat{\nu})|$ where $\hat{\mu}, \hat{\nu}$ are empirical measures from $\mu$ and $\nu$ 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 $\|\Omega^* - \hat{\Omega}\|$ depending on $n$, where $\Omega^*$ is the optimal projection matrix onto $\text{span}\{e_1, e_2\}$.

We consider next $k^* = 2$, and choose $k = 2$. We will need to compute $W^2(\mu, \nu)$. Although (Forrow et al.,

6.2 Transport from Disk to Annulus

Let $k^* \in [d]$. We now consider $\mu$ the uniform distribution over the $k^*$-dimensional disk embedded in $\mathbb{R}^d$,

$$\mu = \mathcal{U}\{x \in \mathbb{R}^d \mid \|x_{1,...,x_{k^*}}\| \leq 1, x_i \in [0,1] \text{ for } i = (k^* + 1),...,d\}$$

and $\nu$ the uniform distribution over a $k^*$-dimensional annulus (cylinder) embedded in $\mathbb{R}^d$,

$$\nu = \mathcal{U}\{x \in \mathbb{R}^d \mid 2 \leq \|x_{1,...,x_{k^*}}\| \leq 3, x_i \in [0,1] \text{ for } i = (k^* + 1),...,d\}.$$

We do the same experiments as for the fragmented hypercube. Based on two empirical distributions $\hat{\mu}$ from $\mu$ and $\hat{\nu}$ from $\nu$, we plot in Figure 6 the sequence $k \mapsto S_k^2(\hat{\mu}, \hat{\nu})$, for different values of $k^*$. An “elbow” shows at $k = k^*$, because the last $d-k^*$ dimensions only represent noise, which is recovered in our plot.

We also plot the optimal transport plan (in the sense of $W$, Figure 5 left) and the optimal transport plan (in the sense of $S_2$) between $\hat{\mu}$ and $\hat{\nu}$ (with $n = 250$ points each, Figure 5 right).
2019) seem to suggest that it is equal to 4, we find a different value (see proof in Appendix):

$$W^2(\mu, \nu) = \frac{14}{5} + \frac{8}{5\sqrt{5}} \log\left(\frac{3 + \sqrt{5}}{2}\right) \approx 3.48865.$$ 

We plot in Figure 7 the estimation error $|W^2(\mu, \nu) - S^2_k(\hat{\mu}, \hat{\nu})|$ depending on the number of points $n$ in the empirical measures $\hat{\mu}, \hat{\nu}$ from $\mu$ and $\nu$. In Figure 8, we plot the subspace estimation error $\|\Omega^* - \hat{\Omega}\|$ depending on $n$, where $\Omega^*$ is the optimal projection matrix onto span$\{e_1, e_2\}$.

We plot the optimal transport plan (in the sense of $W$, Figure 9 left) and the optimal transport plan (in the sense of $S^2$) between $\hat{\mu}$ and $\hat{\nu}$ (with $n = 250$ points each, Figure 9 right). Geodesics in the SRW space are robust to statistical noise.

For our experiment, we simulated 100 independent couples of covariance matrices $\Sigma_1, \Sigma_2$ 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 10, we plot the mean (over the 100 samples) of $l \mapsto S^2_l(\hat{\mu}, \hat{\nu}) / W^2(\hat{\mu}, \hat{\nu})$. 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.

### 6.3 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 $\mu$ and $\nu$ 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 $l \geq 2k$, $S^2_l(\mu, \nu) = W^2(\mu, \nu)$.

For our experiment, we simulated 100 independent couples of covariance matrices $\Sigma_1, \Sigma_2$ in dimension $d = 20$, each having independently a Wishart distribution with $k = 5$ degrees of freedom.

### 6.4 $S_k$ is Robust to Noise

As in experiment 6.3, 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 11, we plot the mean (over the 100 samples) of the relative errors
\[
\sigma \mapsto \left| S^2_2(\hat{\mu}_\sigma, \hat{\nu}_\sigma) - S^2_2(\hat{\mu}_0, \hat{\nu}_0) \right|
\]
and
\[
\sigma \mapsto \left| W^2(\hat{\mu}_\sigma, \hat{\nu}_\sigma) - W^2(\hat{\mu}_0, \hat{\nu}_0) \right|.
\]
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.

6.5 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 12 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 \( V_\pi \) in Algorithm 2.

6.6 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 13 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 14 (left).
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.

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.

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.

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., Jебара, 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.
A Projection Robust Wasserstein Distances

In this section, we prove some basic properties of projection robust Wasserstein distances $P_k$. First note that the definition of $P_k$ makes sense, since for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $k \leq \|d\|$ and $E \in \mathcal{G}_k$, $P_{E\#\mu}$ and $P_{E\#\nu}$ have a second moment (for orthogonal projections are 1-Lipschitz).

$P_k$ is also well posed, since one can prove the existence of a maximizing subspace. To prove this, we will need the following lemma stating that the admissible set of couplings between the projected measures are exactly the projections of the admissible couplings between the original measures:

**Lemma 6.** Let $f : \mathbb{R}^d \to \mathbb{R}^d$ Borel and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Then $\Pi(f\#\mu, f\#\nu) = \{ f \otimes f \# \pi \mid \pi \in \Pi(\mu, \nu) \}$.

This can be used to get the following result:

**Proposition 5.** For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $k \leq \|d\|$, there exists a subspace $E^* \in \mathcal{G}_k$ such that

$$P_k(\mu, \nu) = \mathcal{W}(P_{E^*\#\mu}, P_{E^*\#\nu}).$$

**Proof.** The Grassmannian $\mathcal{G}_k$ is compact, and we show that the application $E \mapsto \mathcal{W}(P_{E\#\mu}, P_{E\#\nu})$ is upper semi-continuous, which gives existence. $\square$

Note that we could define projection robust Wasserstein distances for any $p \geq 1$ by:

$$\sup_{E \in \mathcal{G}_k} \mathcal{W}_p(P_{E\#\mu}, P_{E\#\nu}).$$

Then there is still existence of optimal subspaces, and it defines a distance over

$$\mathcal{P}_p(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \mid \int \|x\|^p \, d\mu(x) < \infty \right\}.$$ 

To prove the identity of indiscernibles, we use the following Lemma due to Rényi, generalizing Cramer-Wold theorem:

**Lemma 7.** Let $(E_j)_{j \in J}$ be a family of subspaces of $\mathbb{R}^d$ such that $\bigcup_{j \in J} E_j = \mathbb{R}^d$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ such that for all $j \in J$, $P_{E_j \# \mu} = P_{E_j \# \nu}$. Then $\mu = \nu$.

B Proofs

**Proof of Lemma 1.** For $\pi \in \Pi(\mu, \nu)$, the application $E \mapsto \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x, y)$ is continuous and $G_k$ is compact, so the supremum is a maximum. Moreover, the application $\pi \mapsto \max_{E \in \mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x, y)$ is lower semicontinuous as the maximum of lower semicontinuous functions. Since $\Pi(\mu, \nu)$ is compact (for any sequence in $\Pi(\mu, \nu)$ is tight), the infimum is a minimum.

**Proof of Lemma 2.** A classical variational result by (Fan, 1949) states that

$$\sum_{l=1}^k \lambda_l(V_{\pi}) = \max_{U \in \mathbb{R}^k} \text{trace} \left( UV_{\pi} U^T \right).$$

Then using the linearity of the trace:

$$\sum_{l=1}^k \lambda_l(V_{\pi}) = \max_{U \in \mathbb{R}^k} \int \text{trace} \left[ U(x - y)(x - y)^T U^T \right] \, d\pi(x, y)$$

$$= \max_{U \in \mathbb{R}^k} \int \|U(x - y)\|^2 \, d\pi(x, y)$$

$$= \max_{E \in \mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x, y).$$

Taking the minimum over $\pi \in \Pi(\mu, \nu)$ yields the result.

**Proof of Theorem 1.** $S^*_k(\mu, \nu) = (2) :$ We fix $\pi \in \Pi(\mu, \nu)$ and focus on the inner maximization in (2):

$$\max_{0 \leq \Omega \leq I, \text{trace}(\Omega) = k} \int d\Omega \, d\pi = \max_{0 \leq \Omega \leq I, \text{trace}(\Omega) = k} \langle \Omega | V_{\pi} \rangle.$$ 

A result by (Overton & Womersley, 1993) shows that this is equal to

$$\max_{U \in \mathbb{R}^k} \text{trace} \left( UV_{\pi} U^T \right)$$

which is nothing but the sum of the $k$ largest eigenvalues of $V_{\pi}$ by Fan’s result. By lemma 2, taking the minimum over $\pi \in \Pi(\mu, \nu)$ yields the result.

$(2) = (3) :$ We will use Sion’s minimax theorem to interchange the minimum and the maximum. Put $f(\Omega, \pi) = \int d\Omega \, d\pi$ and $\mathcal{R} = \left\{ \Omega \in \mathbb{R}^{d\times d} \mid 0 \leq \Omega \leq I; \text{trace}(\Omega) = k \right\}$. Note that $\mathcal{R}$
Subspace Robust Wasserstein Distances

Figure 15. This figure should be compared to Figure 1. We also present an example for which the explicit computation of projection $P_k$ and subspace $S_k$ robust Wasserstein distances described in §3 can be carried out explicitly, by simple enumeration. Unlike Figure 15, and as can be seen in the rightmost plot, these two quantities do not coincide here. That plot reveals that the minimum among all maximal eigenvalues of second order moment matrices computed on all optimal OT plans obtained by enumerating all lines (the subspace robust quantity) is strictly larger than the worst possible projection cost.

is convex and compact, and $\Pi(\mu, \nu)$ is convex (and actually compact, but we won’t need it here). Moreover, $f$ is bilinear and for any $\pi \in \Pi(\mu, \nu)$, $f(\cdot, \pi)$ is continuous. Let $\Omega \in \mathcal{R}$. Let us show that $f(\Omega, \cdot)$ is lower semicontinuous for the weak convergence. Let $(\phi_j)_{j \in \mathbb{N}}$ be an increasing sequence of bounded continuous functions, converging pointwise to $d^2\Omega$. Then $f(\Omega, \pi) = \sup_{j \in \mathbb{N}} \int \phi_j \, d\pi$. For $j \in \mathbb{N}$, $\phi_j$ is continuous and bounded, so $\pi \mapsto \int \phi_j \, d\pi$ is continuous for the weak convergence. Then $f(\Omega, \cdot)$ is lower semicontinuous as the supremum of continuous functions. Then by Sion’s minimax theorem,

$$\min_{\rho \in \Omega} \max_{\nu \in \mathcal{R}} f(\Omega, \pi) = \max_{\nu \in \mathcal{R}} \min_{\rho \in \Omega(\mu, \nu)} f(\Omega, \pi)$$

which is exactly (2) = (3).

(3) = (4) : Fix $\Omega \in \mathcal{R}$. One has:

$$\min_{\pi \in \Pi(\mu, \nu)} \int d^2\Omega \, d\pi = \min_{\pi \in \Pi(\mu, \nu)} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x, y)$$

$$= \min_{\pi \in \Pi(\mu, \nu)} \int \|x-y\|^2 \, d\left[\Omega^{1/2} \otimes \Omega^{1/2}\right]_\# \pi(x, y)$$

$$= \min_{\rho \in \Pi(\Omega^{1/2}_x, \omega^{1/2}_y)} \int \|x-y\|^2 \, d\rho(x, y)$$

$$= \mathcal{W}^2\left(\Omega^{1/2}_x, \omega^{1/2}_y\right)$$

where we have used Lemma 6. Taking the maximum over $\Omega \in \mathcal{R}$ gives the result.

Proof of Lemma 5. We use the fact that the pushforward by $f$ of a Dirac at $x$ is the Dirac at $f(x)$, and that the $\mathcal{W}$ distance between two Diracs is the Euclidean distance between the points:

$$S_k(\delta_x, \delta_y) = \max_{0 \leq \Omega \leq 1} \mathcal{W}\left(\Omega^{1/2}_x \delta_x, \Omega^{1/2}_y \delta_y\right)$$

$$= \max_{0 \leq \Omega \leq 1, \text{trace}(\Omega) = k} \|\Omega^{1/2}(x-y)\|$$

Since $\|\Omega^{1/2}(x-y)\| \leq \|x-y\|$ with equality for any orthogonal projection matrix $\Omega$ onto a subspace $E \in \mathcal{G}_k$ such that span$(y-x) \subset E$, the result follows.

Proof of Proposition 2. Let $k \in \|d\|$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Let us prove the upper bound on $S_k$. Using the change of variable formula and the fact that for any $\Omega \in \left\{\Omega \in \mathbb{R}^{d \times d} \mid 0 \leq \Omega \leq 1; \text{trace}(\Omega) = k\right\}$, $\Omega^{1/2}$ is 1-Lipschitz,

$$S_k^2(\mu, \nu) = \max_{0 \leq \Omega \leq 1} \min_{\text{trace}(\Omega) = k} \mathcal{W}^2\left(\Omega^{1/2}_x \delta_x, \Omega^{1/2}_y \delta_y\right)$$

$$\leq \max_{0 \leq \Omega \leq 1} \min_{\text{trace}(\Omega) = k} \|\Omega^{1/2}(x-y)\| \|x-y\|$$

$$= \mathcal{W}^2(\mu, \nu)$$

which gives the upper bound. For the lower bound, we define $B_k \subset \mathcal{G}_k$ the (finite) set of $k$-dimensional subspaces of $\mathbb{R}^d$ spanned by $k$ vectors of the canonical basis of $\mathbb{R}^d$:

$$B_k = \left\{\text{span}(e_{(1)}, \ldots, e_{(k)}) \mid \sigma \in \mathcal{G}_d\right\}.$$
Let us now bound $S_k$ from below:

$$S_k^2(\mu, v) = \min_{\pi \in \Pi(\mu, v)} \max_{0 \leq \Omega \leq I} \int_{\text{trace}(\Omega) = k} \|\Omega^{1/2}(x - y)\|^2 \, d\pi(x, y)$$

Note that in the case of $I$ being the sum of the $k$ largest elements of $I = \{\int (x_i - y_i)^2 \, d\pi(x, y) \mid i \in [d]\}$, so it is greater than $\frac{k}{d}$ times the sum of all the elements in $I$:

$$S_k^2(\mu, v) \geq \frac{k}{d} \min_{\pi \in \Pi(\mu, v)} \int \|x - y\|^2 \, d\pi(x, y) = \frac{k}{d} W^2(\mu, v).$$

Note that in the case of $\mu = \delta_0$ and $v = \sigma$, the two inequalities in the proof of the lower bound are equalities, hence the tight lower bound constant.

**Proof of Lemma 4.** Let us first prove the lower bound:

$$S_{k+1}^2(\mu, v) = \sum_{l=1}^{k} \lambda_l(V_{\pi_{k+1}}) + \lambda_{k+1}(V_{\pi_{k+1}}) \geq \sum_{l=1}^{k} \lambda_l(V_{\pi_k}) + \lambda_{k+1}(V_{\pi_{k+1}}) = S_k^2(\mu, v) + \lambda_{k+1}(V_{\pi_{k+1}}).$$

Let us now prove the upper bound:

$$S_{k+1}^2(\mu, v) = \min_{\pi \in \Pi(\mu, v)} \sum_{l=1}^{k+1} \lambda_l(V_{\pi_l}) \leq \sum_{l=1}^{k+1} \lambda_l(V_{\pi_l}) = S_k^2(\mu, v) + \lambda_{k+1}(V_{\pi_{k+1}}).$$

**Proof of Proposition 3.** Increase is direct using lemma 4, since for any $\pi \in \Pi(\mu, v)$, $V_{\pi}$ has only non-negative eigenvalues.

Let $k \in [d - 2]$. Then using twice lemma 4,

$$S_{k+2}^2(\mu, v) - S_{k+1}^2(\mu, v) \leq \lambda_{k+2}(V_{\pi_{k+1}}) \leq \lambda_{k+1}(V_{\pi_{k+1}}) \leq S_{k+1}^2(\mu, v) - S_k^2(\mu, v),$$

which shows that $k \mapsto S_k^2(\mu, v)$ is concave.

Let $k \in [d - 1]$. Although the minoration of $S_{k+1}^2(\mu, v) - S_k^2(\mu, v)$ is a direct consequence of concavity, we give a direct computation using lemma 4:

$$S_{k+1}^2(\mu, v) - S_k^2(\mu, v) \geq \lambda_{k+1}(V_{\pi_{k+1}}) \geq \frac{1}{d - k - 1} \sum_{l=k+1}^{d} \lambda_l(V_{\pi_{k+1}}) = \frac{1}{d - k - 1} \left[ \sum_{l=1}^{d} \lambda_l(V_{\pi_{k+1}}) - \sum_{l=1}^{k+1} \lambda_l(V_{\pi_{k+1}}) \right] \geq \frac{1}{d - k - 1} [W^2(\mu, v) - S_{k+1}^2(\mu, v)],$$

which implies that

$$(d - k) \left[ S_{k+1}^2(\mu, v) - S_k^2(\mu, v) \right] \geq W^2(\mu, v) - S_k^2(\mu, v).$$

Finally, the majoration of $S_k(\mu, v)$ is a direct consequence of lemma 4:

$$S_{k+1}^2(\mu, v) \leq S_k^2(\mu, v) + \lambda_{k+1}(V_{\pi_{k+1}}) \leq S_k^2(\mu, v) + \lambda_k(V_{\pi_k}) \leq S_k^2(\mu, v) + \frac{k}{k+1} \sum_{l=1}^{k+1} \lambda_l(V_{\pi_l}) = \frac{k + 1}{k} S_k^2(\mu, v).$$

**Proof of Proposition 4.** For $s, t \in [0, 1]$, put $\pi(s, t) = (f_s, f_t) \# \pi^* \in \Pi(\mu_s, \mu_t)$, which is our candidate for an
optimal transport plan. Then

\[ S_k^2(\mu, \nu) = \sum_{l=1}^{k} \lambda_l (V_{\Pi(l)}) \]

\[ = \sum_{l=1}^{k} \lambda_l \left\{ \left( f_\Pi(x, y) - f_\Pi(x, y) \right)^T d\pi^* (x, y) \right\} \]

\[ = \sum_{l=1}^{k} \lambda_l \left( (t - s)^2 V_{\pi^*} \right) \]

\[ = (t - s)^2 S_k^2(\mu, \nu) \]

where we have used

\[ f_\Pi(x, y) - f_\Pi(x, y) = (1 - s)x + sy - (1 - t)x - ty \]

Then for \( 0 \leq s < t \leq 1 \), using the triangular inequality,

\[ S_k(\mu, \nu) \leq S_k(\mu, \mu) + S_k(\mu, \mu) + S_k(\mu, \mu) \leq (s + (t - s) + (1 - t)) S_k(\mu, \mu) = S_k(\mu, \mu) \]

which implies equality everywhere, and in particular optimality for \( \pi(s, t) \). Then for all \( s, t \in [0, 1] \),

\[ S_k(s, t) = |t - s| S_k(\mu, \nu), \]

which shows that the curve \( (\mu_t) \) has constant speed

\[ |\mu_t'| = \lim_{t \to 0} \frac{S_k(\mu_{t+c}, \mu_t)}{|c|} = S_k(\mu, \nu), \]

and that the length of the curve \( (\mu_t) \) is

\[ \sup \left\{ \sum_{i=0}^{n-1} S_k(\mu_{t_i}, \mu_{t_{i+1}}) \right\} \]

\[ = S_k(\mu, \nu), \]

i.e. that \( (\mu_t) \) is a geodesic connecting \( \mu \) and \( \nu \).

**Proof of Lemma 5.** Although this is a direct consequence of (Overton & Womersley, 1993), we give an explicit proof. Fix \( \pi \in P(\mu, \nu) \). Using the linearity of the trace,

\[ \max_{0 \leq \Omega \leq I} \int d\Omega^2 d\pi = \max_{0 \leq \Omega \leq I} \text{trace}(\Omega V_{\pi}), \]

which is a SDP. Its dual writes

\[ \min_{s \in \mathbb{R}, \Omega \in \mathbb{R}^{d \times d}} \text{trace}(\Omega) + ks. \]

Let us write the eigendecomposition of \( V_{\pi} = U \text{diag}(\lambda_1, \ldots, \lambda_d) U^T \) with \( \lambda_1 \geq \ldots \geq \lambda_d \). Put \( \Omega = U \text{diag}((1_k, 0_{d-k})) U^T, \tilde{Z} = U \text{diag}((\lambda_1 - \lambda_k, \ldots, \lambda_d - \lambda_k) U^T \) and \( \tilde{s} = \lambda_k \). Then \( 0 \leq \Omega \leq I, \text{trace}(\Omega) = k \) and \( (\tilde{s}, \tilde{Z}) \) is admissible for the dual problem, with corresponding primal and dual values

\[ \text{trace}(\Omega V_{\pi}) = \sum_{i=1}^{k} \lambda_i, \]

\[ \text{trace}(\tilde{Z}) + k\tilde{s} = \sum_{i=1}^{k} (\lambda_i - \lambda_k) + k\lambda_k = \sum_{i=1}^{k} \lambda_i. \]

We found primal and dual admissible variables that give the same value, so these variables are optimal. In particular, \( \Omega \) is solution to

\[ \max_{0 \leq \Omega \leq I} \text{trace}(\Omega) \int d\Omega^2 d\pi. \]

**Proof of Lemma 6.** Let \( E = \text{Im}(f) \subset \mathbb{R}^d \) and \( \pi \in P(\mu, \nu) \). Then \( (f \otimes f)_{\#} \pi \) is an admissible transport plan between \( f_{\#} \mu \) and \( f_{\#} \nu \). Indeed, for any Borel set \( A \subset E \), \( (f \otimes f)_{\#} \pi(A \times E) = \pi(f^{-1}(A) \times f^{-1}(E)) = \pi(f^{-1}(A) \times \mathbb{R}^d) = \mu(f^{-1}(A)) = f_{\#} \mu(A) \). So \( (f \otimes f)_{\#} \pi \) has first marginal \( f_{\#} \mu \), and likewise, has second marginal \( f_{\#} \nu \), i.e. \( (f \otimes f)_{\#} \pi \in P(f_{\#} \mu, f_{\#} \nu) \).

Conversely, let \( \rho \in P(f_{\#} \mu, f_{\#} \nu) \). Let us construct \( \pi \in P(\mu, \nu) \) such that \( (f \otimes f)_{\#} \pi = \rho \). For any Borel sets \( A, B \subset \mathbb{R}^d \), put

\[ \pi(A \times B) = \rho(f(A) \times f(B)) \mu(A) \nu(B) / f_{\#} \mu(f(A)) f_{\#} \nu(f(B)). \]

if \( f_{\#} \mu(f(A)) \neq 0 \) and \( f_{\#} \mu(f(A)) \neq 0 \), and \( \pi(A, B) = 0 \) otherwise. Then \( \pi \in P(\mu, \nu) \). Indeed, for any Borel set \( A \subset \mathbb{R}^d \), \( \pi(A \times \mathbb{R}^d) = \rho(f(A) \times \mathbb{R}^d) \mu(A) \). If \( f_{\#} \mu(f(A)) \neq 0 \) and \( \mu(A \times \mathbb{R}^d) = 0 \), then \( \mu(A) \leq \mu(f^{-1}(f(A))) = f_{\#} \mu(f(A)) = 0 \) so \( \mu(A) = 0 = \pi(A \times \mathbb{R}^d) \). The same calculations give the result for the second marginal.

There remains to prove that \( (f \otimes f)_{\#} \pi = \rho \). For any Borel sets \( A, B \subset E \), noting that \( f(f^{-1}(A)) = A \) and
Proof of Lemma 7. Let $j \in J$. Since $P_{E_j \# \mu} = P_{E_j \# \nu}$, their characteristic functions are equal, i.e., for all $t \in \mathbb{R}^d$,

$$
\begin{align*}
\int \exp i \langle t, x \rangle \, d\mu(x) &= \int \exp i \langle t, x \rangle \, d\nu(x) \\
\int \exp i \langle t, P_{E_j} x \rangle \, d\mu(x) &= \int \exp i \langle t, P_{E_j} x \rangle \, d\nu(x) \\
\int \exp i \langle P_{E_j} t, x \rangle \, d\mu(x) &= \int \exp i \langle P_{E_j} t, x \rangle \, d\nu(x)
\end{align*}
$$

i.e. the characteristic functions of $\mu$ and $\nu$ coincide on $E_j$, for all $j \in J$. Since the subspaces $(E_j)_{j \in J}$ cover the whole space $\mathbb{R}^d$, $\mu$ and $\nu$ have the same characteristic functions on $\mathbb{R}^d$, hence $\mu = \nu$.

Proof of Proposition 5. We endow the Grassmannian $G_k$ with the metric topology associated with metric $d : (E, F) \mapsto \|P_E - P_F\|$, where $P_E$ and $P_F$ are the linear projectors onto $E$ and $F$. Then it is well known that $G_k$ is compact under this topology.

We only have to show that, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the map $f : E \mapsto \mathcal{W}(P_{E \# \mu}, P_{E \# \nu})$ is upper semicontinuous. For any orthogonal projector $P$, using lemma 6,

$$
\mathcal{W}^2(P_{E \# \mu}, P_{E \# \nu}) = \min_{\rho \in \Pi(P_{E \# \mu}, P_{E \# \nu})} \int \|x - y\|^2 \, d\rho(x, y)
$$

Therefore, the application $P \mapsto \mathcal{W}^2(P_{E \# \mu}, P_{E \# \nu})$ is continuous, the application $g : P \mapsto \mathcal{W}^2(P_{E \# \mu}, P_{E \# \nu})$ is upper semicontinuous as the minimum of continuous functions. As the application $h : E \mapsto P_{E \# \mu}$ is continuous, and $x \mapsto \sqrt{x}$ is nondecreasing, $f = \sqrt{g \circ h}$ is upper semicontinuous.

Proof of the value of $\mathcal{W}^2(\mu, \nu)$ for the Disk to Annulus setup. Let us define a map $T$ using polar coordinates for the first two coordinates, and cartesian coordinates for the remaining $d - 2$, as follows:

$$
T(r, \theta, x_3, ..., x_d) = \left( \sqrt{4 + 5r^2}, \theta, x_3, ..., x_d \right).
$$

We show that $T$ is an optimal transport map between $\mu$ and $\nu$. First, we show that $T_{\# \mu} = \nu$. Since $T$ only operates on the first coordinate and $\mu$ and $\nu$ only differ on the first coordinate, we only have to prove that $T_{1 \# \mu_1}$ and $v_1$ have the same CDF, where $T_1, \mu_1$ and $v_1$ stand for the first coordinate projection of $T, \mu$ and $\nu$. For any $r \in [2, 3]$:  

$$
\mathbb{P}_{R \sim \mu_1}(T_1(R) \leq r) = \mathbb{P}_{R \sim \mu_1}(R \leq T_1^{-1}(r)) = \mathbb{P}_{R \sim \mu_1}(R \leq \sqrt{\frac{r^2 - 4}{5}}) = \int_0^{\sqrt{\frac{r^2 - 4}{5}}} x \, dx = \frac{r^2 - 4}{5}
$$

which shows that $T_{\# \mu} = \nu$. Moreover, $T$ is a subgradient of a convex function, since its gradient is semidefinite positive:

$$
\nabla T(r, \theta, x_3, ..., x_d) = \text{Diag} \left( \frac{5r}{\sqrt{4 + 5r^2}}, 1, ..., 1 \right) \succeq 0.
$$

Then by Brenier’s theorem, $T$ is an optimal transport map between $\mu$ and $\nu$, and

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

$$
= 2 \int_0^1 \left( r - \sqrt{4 + 5r^2} \right)^2 r \, dr
$$

$$
= 14 \left( \frac{5}{5} + \frac{8}{5\sqrt{5}} \log \left( \frac{3 + \sqrt{5}}{2} \right) \right) \approx 3.48865.
$$