A. Projection Robust Wasserstein Distances

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

$\mathcal{P}_k$ 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 $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$. Then $\Pi(f_\#\mu, f_\#\nu) = \{(f \otimes f)_\#\pi \,|\, \pi \in \Pi(\mu, \nu)\}$.*

This can be used to get the following result:

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

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

Proof. The Grassmannian $\mathcal{G}_k$ is compact, and we show that the application $E \mapsto \mathcal{W}\left(P_{E\#}\mu, P_{E\#}\nu\right)$ is upper semicontinuous, which gives existence. □

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

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

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

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

To prove the identity of indiscernibles, we use the following Lemma due to Rényi, generalizing Cramér-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 $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ 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 $\mathcal{G}_k$ 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

$$\sum_{l=1}^k \lambda_l(V_\pi) = \max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^T = I_k}} \operatorname{trace}\left(UV_\pi U^T\right).$$

Then using the linearity of the trace:

$$\begin{aligned}
\sum_{l=1}^k \lambda_l(V_\pi) &= \max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^T = I_k}} \int \operatorname{trace}\left[U(x-y)(x-y)^T U^T\right] \, d\pi(x, y) \\
&= \max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^T = I_k}} \int \|U(x-y)\|^2 \, d\pi(x, y) \\
&= \max_{E \in \mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x, y).
\end{aligned}$$

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

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

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \operatorname{trace}(\Omega) = k}} \int d_\Omega^2 \, d\pi = \max_{\substack{0 \preceq \Omega \preceq I \\ \operatorname{trace}(\Omega) = k}} \langle \Omega \,|\, V_\pi \rangle.$$

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

$$\max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^T = I_k}} \operatorname{trace}\left(UV_\pi U^T\right)$$

which is nothing but the sum of the k largest eigenvalues of V_π 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 $f(\Omega, \pi) = \int d_\Omega^2 \, d\pi$ and $\mathcal{R} = \left\{\Omega \in \mathbb{R}^{d \times d} \,|\, 0 \preceq \Omega \preceq I \,;\, \operatorname{trace}(\Omega) = k\right\}$. Note that $\mathcal{R}$ 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_Ω^2. Then $f(\Omega, \pi) = \sup_{j \in \mathbb{N}} \int \phi_j \, d\pi$. For $j \in \mathbb{N}$, ϕ_j is continuous and bounded, so $\pi \mapsto \int \phi_j \, d\pi$ is continuous for the

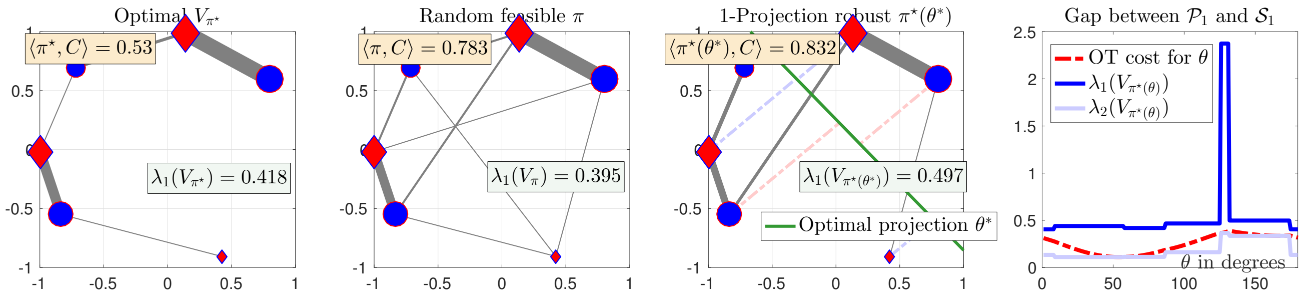

Figure 11. This figure should be compared to Figure 1. We also present an example for which the explicit computation of projection $\mathcal{P}_k$ and subspace $\mathcal{S}_k$ robust Wasserstein distances described in §3 can be carried out explicitly, by simple enumeration. Unlike Figure 11, and as can be seen in the rightmost plot, these two quantities do not coincide here. That plot reveals that the minimum across 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.

weak convergence. Then $f(\Omega, \cdot)$ is lower semicontinuous as the supremum of continuous functions. Then by Sion's minimax theorem,

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

which is exactly (2) = (3).

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

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

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

Proof of Lemma 3. 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:

$$\begin{aligned}
\mathcal{S}_k(\delta_x, \delta_y) &= \max_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \mathcal{W}\left({\Omega^{1/2}}_\#\delta_x, {\Omega^{1/2}}_\#\delta_y\right) \\
&= \max_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \|\Omega^{1/2}(x-y)\|.
\end{aligned}$$

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

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

$$\begin{aligned}
&\mathcal{S}_k^2(\mu,\nu) \\
&= \max_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \min_{\pi\in\Pi(\mu,\nu)} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x,y) \\
&\le \max_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \, d\pi(x,y) \\
&= \mathcal{W}^2(\mu,\nu)
\end{aligned}$$

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

$$\mathcal{B}_k = \left\{\mathrm{span}(e_{\sigma(1)}, ..., e_{\sigma(k)}) \,|\, \sigma \in \mathfrak{S}_d\right\}.$$

Let us now bound $\mathcal{S}_k$ from below:

$$\begin{aligned}
&\mathcal{S}_k^2(\mu,\nu) \\
&= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{0\preceq\Omega\preceq I\\ \mathrm{trace}(\Omega)=k}} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x,y) \\
&\ge \min_{\pi\in\Pi(\mu,\nu)} \max_{E\in\mathcal{B}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x,y) \\
&= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{A\subset[\![d]\!]\\ |A|=k}} \int \sum_{i\in A} (x_i - y_i)^2 \, d\pi(x,y) \\
&= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{A\subset[\![d]\!]\\ |A|=k}} \sum_{i\in A} \int (x_i - y_i)^2 \, d\pi(x,y).
\end{aligned}$$

For $\pi \in \Pi(\mu,\nu)$,

$$\max_{\substack{A\subset[\![d]\!]\\ |A|=k}} \sum_{i\in A} \int (x_i - y_i)^2 \, d\pi(x,y)$$

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

$$\begin{aligned}\mathcal{S}_k^2(\mu,\nu) &\geq \frac{k}{d} \min_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \, d\pi(x,y) \\ &= \frac{k}{d} \mathcal{W}^2(\mu,\nu).\end{aligned}$$

Note that in the case of $\mu = \delta_0$ and $\nu = \sigma$, 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:

$$\begin{aligned}\mathcal{S}_{k+1}^2(\mu,\nu) &= \sum_{l=1}^{k} \lambda_l(V_{\pi_{k+1}}) + \lambda_{k+1}(V_{\pi_{k+1}}) \\ &\geq \sum_{l=1}^{k} \lambda_l(V_{\pi_k}) + \lambda_{k+1}(V_{\pi_{k+1}}) \\ &= \mathcal{S}_k^2(\mu,\nu) + \lambda_{k+1}(V_{\pi_{k+1}}).\end{aligned}$$

Let us now prove the upper bound:

$$\begin{aligned}\mathcal{S}_{k+1}^2(\mu,\nu) &= \min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^{k+1} \lambda_l(V_\pi) \\ &\leq \sum_{l=1}^{k+1} \lambda_l(V_{\pi_k}) \\ &= \mathcal{S}_k^2(\mu,\nu) + \lambda_{k+1}(V_{\pi_k}).\end{aligned}$$

Proof of Propositon 3. Increase is direct using lemma 4, since for any $\pi \in \Pi(\mu,\nu)$, V_π has only nonnegative eigenvalues.
Let $k \in [\![d-2]\!]$. Then using twice lemma 4,

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

which shows that $k \mapsto \mathcal{S}_k^2(\mu,\nu)$ is concave.
Let $k \in [\![d-1]\!]$. Although the minoration of $\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu)$ is a direct consequence of concavity, we give a direct computation using lemma 4:

$$\begin{aligned}&\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu) \\ &\geq \lambda_{k+1}(V_{\pi_{k+1}}) \\ &\geq \frac{1}{d-k-1} \sum_{l=k+2}^{d} \lambda_l(V_{\pi_{k+1}}) \\ &= \frac{1}{d-k-1} \left[\sum_{l=1}^{d} \lambda_l(V_{\pi_{k+1}}) - \sum_{l=1}^{k+1} \lambda_l(V_{\pi_{k+1}})\right] \\ &\geq \frac{1}{d-k-1} \left[\mathcal{W}^2(\mu,\nu) - \mathcal{S}_{k+1}^2(\mu,\nu)\right],\end{aligned}$$

which implies that

$$\begin{aligned}(d-k)&\left[\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu)\right] \\ &\geq \mathcal{W}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu).\end{aligned}$$

Finally, the majoration of $\mathcal{S}_k(\mu,\nu)$ is a direct consequence of lemma 4:

$$\begin{aligned}\mathcal{S}_{k+1}^2(\mu,\nu) &\leq \mathcal{S}_k^2(\mu,\nu) + \lambda_{k+1}(V_{\pi_k}) \\ &\leq \mathcal{S}_k^2(\mu,\nu) + \lambda_k(V_{\pi_k}) \\ &\leq \mathcal{S}_k^2(\mu,\nu) + \frac{1}{k}\sum_{l=1}^{k} \lambda_l(V_{\pi_k}) \\ &= \frac{k+1}{k} \mathcal{S}_k^2(\mu,\nu).\end{aligned}$$

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

$$\begin{aligned}&\mathcal{S}_k^2(\mu_s,\mu_t) \leq \sum_{l=1}^{k} \lambda_l(V_{\pi(s,t)}) \\ &= \sum_{l=1}^{k} \lambda_l \left\{\int [f_s(x,y) - f_t(x,y)] \right. \\ &\qquad\qquad \left. [f_s(x,y) - f_t(x,y)]^T \, d\pi^*(x,y)\right\} \\ &= \sum_{l=1}^{k} \lambda_l \left((t-s)^2 V_{\pi^*}\right) \\ &= (t-s)^2 \mathcal{S}_k^2(\mu,\nu)\end{aligned}$$

where we have used

$$\begin{aligned}f_s(x,y) - f_t(x,y) &= (1-s)x + sy - (1-t)x - ty \\ &= (t-s)(x-y).\end{aligned}$$

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

$$\begin{aligned}&\mathcal{S}_k(\mu,\nu) \\ &\leq \mathcal{S}_k(\mu,\mu_s) + \mathcal{S}_k(\mu_s,\mu_t) + \mathcal{S}_k(\mu_t,\nu) \\ &\leq (s + (t-s) + (1-t))\, \mathcal{S}_k(\mu,\nu) = \mathcal{S}_k(\mu,\nu)\end{aligned}$$

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

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

which shows that 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 that the length of the curve (μ_t) is

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

i.e. that (μ_t) is a geodesic connecting μ and ν.

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

$$\max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}}\int d_\Omega^2\,d\pi=\max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}}\operatorname{trace}(\Omega V_\pi),$$

which is a SDP. Its dual writes

$$\min_{\substack{s\in\mathbb{R},Z\in\mathbb{R}^{d\times d}\\ Z\succeq 0\\ Z+sI\succeq V_\pi}}\operatorname{trace}(Z)+ks.$$

Let us write the eigendecomposition of $V_\pi = U\operatorname{diag}(\lambda_1,...,\lambda_d)U^T$ with $\lambda_1\geq ...\geq\lambda_d$. Put $\widehat{\Omega}=U\operatorname{diag}\left([\mathbf{1}_k,\mathbf{0}_{d-k}]\right)U^T$, $\widehat{Z}=U\operatorname{diag}((\lambda_1-\lambda_k)_+,...,(\lambda_d-\lambda_k)_+)U^T$ and $\widehat{s}=\lambda_k$. Then $0\preceq\widehat{\Omega}\preceq I$, $\operatorname{trace}(\widehat{\Omega})=k$ and $(\widehat{s},\widehat{Z})$ is admissible for the dual problem, with corresponding primal and dual values

$$\operatorname{trace}(\widehat{\Omega}V_\pi)=\sum_{l=1}^k\lambda_l,$$
$$\operatorname{trace}(\widehat{Z})+k\widehat{s}=\sum_{l=1}^k(\lambda_l-\lambda_k)+k\lambda_k=\sum_{l=1}^k\lambda_l.$$

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

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

Proof of Lemma 6. Let $E=\operatorname{Im}(f)\subset\mathbb{R}^d$ and $\pi\in\Pi(\mu,\nu)$. 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.* $(f\otimes f)_\#\pi\in\Pi(f_\#\mu,f_\#\nu)$.
Conversely, let $\rho\in\Pi(f_\#\mu,f_\#\nu)$. Let us construct $\pi\in\Pi(\mu,\nu)$ such that $(f\otimes f)_\#\pi=\rho$. For any Borel sets $A,B\subset\mathbb{R}^d$, put

$$\pi(A\times B)=\frac{\rho(f(A)\times f(B))\mu(A)\nu(B)}{f_\#\mu(f(A))\,f_\#\nu(f(B))}$$

if $f_\#\mu(f(A))\neq 0$ and $f_\#\mu(f(B))\neq 0$, and $\pi(A,B)=0$ otherwise. Then $\pi\in\Pi(\mu,\nu)$. Indeed, for any Borel set $A\subset\mathbb{R}^d$, $\pi(A\times\mathbb{R}^d)=\rho(f(A)\times E)\frac{\mu(A)}{f_\#\mu(f(A))}=\mu(A)$ if $f_\#\mu(f(A))\neq 0$ and $\pi(A\times\mathbb{R}^d)=0$ if $f_\#\mu(f(A))=0$. But 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 $f(f^{-1}(B))=B$,

$$\begin{aligned}(f\otimes f)_\#\pi(A\times B)&=\pi(f^{-1}(A)\times f^{-1}(B))\\ &=\rho(A\times B)\frac{\mu(f^{-1}(A))}{f_\#\mu(A)}\frac{\nu(f^{-1}(B))}{f_\#\mu(B)}\\ &=\rho(A\times B)\end{aligned}$$

if $f_\#\mu(A)\neq 0$ and $f_\#\nu(B)\neq 0$. Otherwise, $(f\otimes f)_\#\pi(A\times B)=0$ and $\rho(A\times B)\leq\min\{\rho(A\times E),\rho(E\times B)\}=\min\{f_\#\mu(A),f_\#\nu(B)\}=0$, so $\rho(A\times B)=(f\otimes f)_\#\pi(A\times B)=0$. $\square$

Proof of Proposition 5. We endow the Grassmannian $\mathcal{G}_k$ 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 $\mathcal{G}_k$ is compact under this topology.

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

$$\begin{aligned}\mathcal{W}^2\left(P_\#\mu,P_\#\nu\right)&=\min_{\rho\in\Pi(P_\#\mu,P_\#\nu)}\int\|x-y\|^2\,d\rho(x,y)\\ &=\min_{\pi\in\Pi(\mu,\nu)}\int\|x-y\|^2\,d(P_E\otimes P_E)_\#\pi(x,y)\\ &=\min_{\pi\in\Pi(\mu,\nu)}\int\|P(x-y)\|^2\,d\pi(x,y).\end{aligned}$$

Since for $\pi\in\Pi(\mu,\nu)$, the application $P\mapsto\int\|P(x-y)\|^2\,d\pi(x,y)$ is continuous, the application $g:P\mapsto\mathcal{W}^2\left(P_\#\mu,P_\#\nu\right)$ is upper semicontinuous as the minimum of continuous functions. As the application $h:E\mapsto P_E$ is continuous, and $x\mapsto\sqrt{x}$ is nondecreasing, $f=\sqrt{g\circ h}$ is upper semicontinuous.

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 $t\in\mathbb{R}^d$,

$$\begin{array}{rcl}\int\exp i\langle t|x\rangle\,dP_{E_j\#}\mu(x)&=&\int\exp i\langle t|x\rangle\,dP_{E_j\#}\nu(x)\\ \int\exp i\langle t|P_{E_j}x\rangle\,d\mu(x)&=&\int\exp i\langle t|P_{E_j}x\rangle\,d\nu(x)\\ \int\exp i\langle P_{E_j}t|x\rangle\,d\mu(x)&=&\int\exp i\langle P_{E_j}t|x\rangle\,d\nu(x)\end{array}$$

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

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:

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

We show that T is an optimal transport map between μ and ν. First, we show that $T_\#\mu=\nu$. Since T only operates on the first coordinate and μ and ν only differ on the first coordinate, we only have to prove that $T_{1\#}\mu_1$ and ν_1 have the same CDF, where T_1, μ_1 and ν_1 stand for the first coordinate projection of T, μ and ν. For any $r\in[2,3]$:

$$\begin{aligned}\mathbb{P}_{R\sim\mu_1}(T_1(R)\le r)&=\mathbb{P}_{R\sim\mu_1}(R\le T_1^{-1}(r))\\&=\mathbb{P}_{R\sim\mu_1}\left(R\le\sqrt{\frac{r^2-4}{5}}\right)\\&=\frac{\int_0^{\sqrt{\frac{r^2-4}{5}}}x\,dx}{\int_0^1 x\,dx}=\frac{r^2-4}{5}\end{aligned}$$

and

$$\mathbb{P}_{R\sim\nu_1}(R\le r)=\frac{\int_2^r x\,dx}{\int_2^3 x\,dx}=\frac{r^2-4}{5}$$

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

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

Then by Brenier's theorem, T is an optimal transport map between μ and ν, and

$$\begin{aligned}\mathcal{W}^2(\mu,\nu)&=\int\|x-T(x)\|^2\,d\mu(x)\\&=2\int_0^1\left(r-\sqrt{4+5r^2}\right)^2 r\,dr\\&=\frac{14}{5}+\frac{8}{5\sqrt{5}}\log\left(\frac{3+\sqrt{5}}{2}\right)\approx 3.48865.\end{aligned}$$

C. Experimental Details

C.1. Additional Experiment: Transport from Disk to Annulus

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

$$\begin{aligned}\mu=\mathcal{U}(\{x\in\mathbb{R}^d\,|\,\|(x_1,...,x_{k^*})\|\le 1,\\ x_i\in[0,1]\text{ for }i=(k^*+1),...,d\})\end{aligned}$$

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

$$\begin{aligned}\nu=\mathcal{U}(\{x\in\mathbb{R}^d\,|\,2\le\|(x_1,...,x_{k^*})\|\le 3,\\ x_i\in[0,1]\text{ for }i=(k^*+1),...,d\}).\end{aligned}$$

We do the same experiments as for the fragmented hypercube. Based on two empirical distributions $\hat\mu$ from μ and $\hat\nu$ from ν, we plot in Figure 12 the sequence $k\mapsto\mathcal{S}_k^2(\hat\mu,\hat\nu)$, 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.

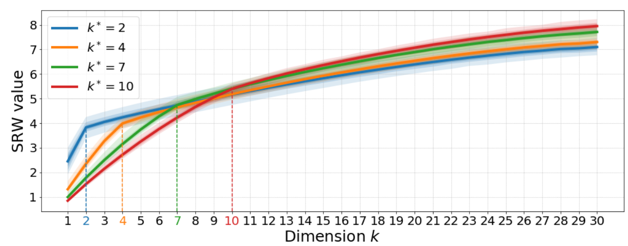

Figure 12. $\mathcal{S}_k^2(\hat\mu,\hat\nu)$ depending on the dimension $k\in[\![d]\!]$, for $k^*\in\{2,4,7,10\}$, where $\hat\mu,\hat\nu$ are empirical measures from μ and ν respectively with 100 points each. Each curve is the mean over 100 samples, and shaded area show the minimum and maximum values.

We consider next $k^*=2$, and choose $k=2$. We will need to compute $\mathcal{W}^2(\mu,\nu)$. Although (Forrow et al., 2019) seem to suggest that it is equal to 4, we find a different value :

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

We plot in Figure 13 the estimation error $|\mathcal{W}^2(\mu,\nu)-\mathcal{S}_k^2(\hat\mu,\hat\nu)|$ depending on the number of points n in the empirical measures $\hat\mu,\hat\nu$ from μ and ν. In Figure 14, we plot the subspace estimation error $\|\Omega^*-\widehat\Omega\|$ depending on n, where Ω^* is the optimal projection matrix onto $\mathrm{span}\{e_1,e_2\}$.

We plot the optimal transport plan (in the sense of $\mathcal{W}$, Figure 15 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 15 right).

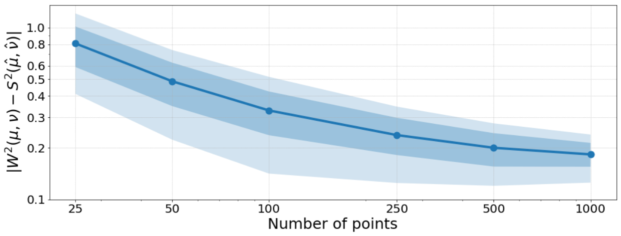

Figure 13. Mean estimation error over 500 random samples for $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

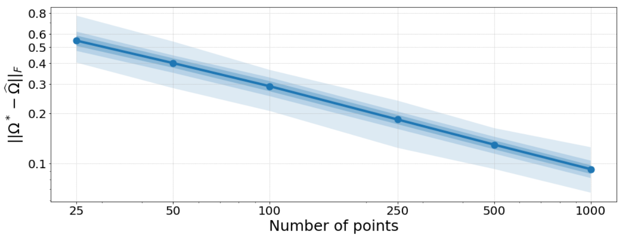

Figure 14. Mean estimation of the subspace estimation error over 500 samples, depending on $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

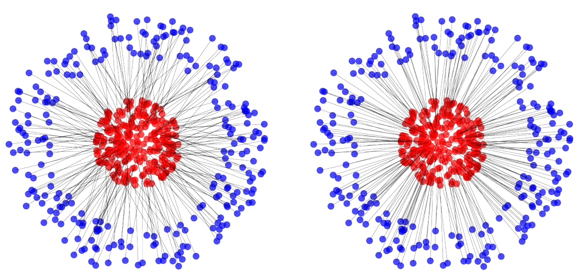

Figure 15. Disk to annulus, $n = 250$, $d = 30$. Optimal mapping in the Wasserstein space (left) and in the SRW space (right). Geodesics in the SRW space are robust to statistical noise.

C.2. Details about Experiment of Section 6.5

The complete vocabulary used consists of the 20000 most common words in English, except for the 2000 most common words, hence a total size of 18000 words. All the words in a movie script that whether do not belong to the vocabulary list, are digits or begin with a capital letter, are deleted. The remaining words form a discrete measure in $\mathbb{R}^{300}$, with the weights proportional to their frequency in the movie script.