Subspace Robust Wasserstein Distances

ENSAE ParisTech
Google AI

François-Pierre PATY (ENSAE ParisTech) *Marco CUTURI (Google Brain)*

Goal: Make sense of Wasserstein distances in high dimension by designing a robust variant of the Wasserstein.

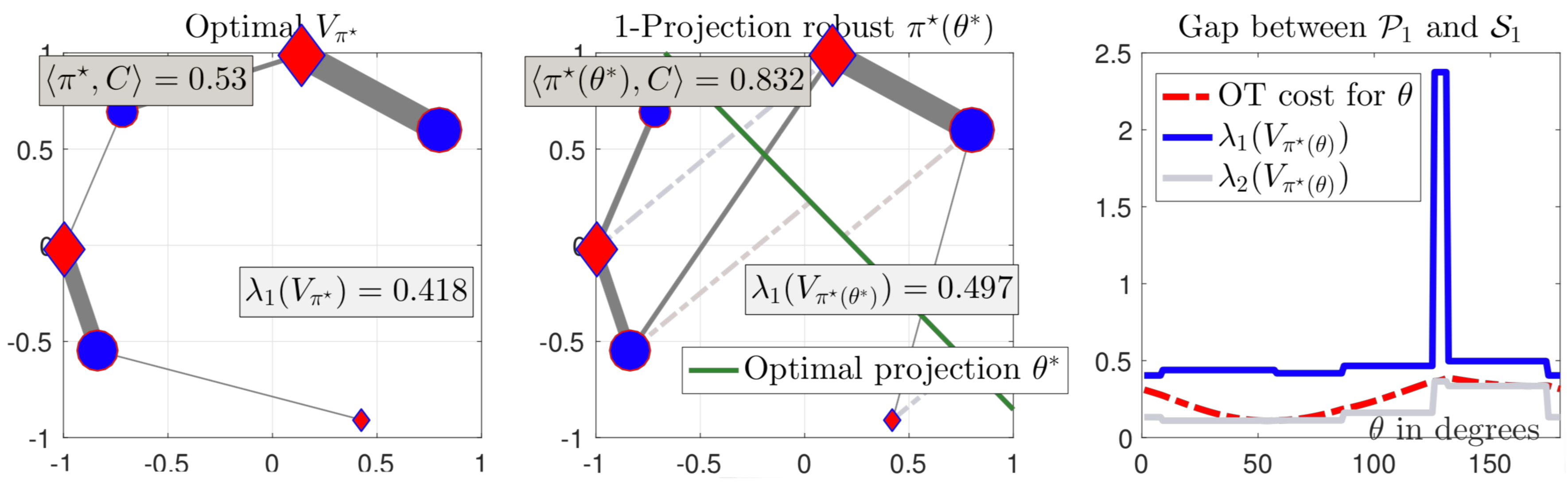
Approach: Project the measures onto a low-dimensional subspace and consider the maximum over all subspaces.

Results: Geodesic metric equivalent to $\mathcal{W}$. Efficient algorithms and use case on text data.

I. Wasserstein Distance in High Dimension

$$\mathcal{W}^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 d\pi(x,y)$$

Problems:
- $\|\cdot\|^2$ not informative in high dimension
- $\left|W_2^2(\hat\mu,\hat\nu) - W_2^2(\mu,\nu)\right| \sim \left(\frac{1}{n}\right)^{1/d}$



II. Projection and Subspace Robust Wasserstein Distances

Projection Robust Wasserstein Distance (PRW)

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

Not convex !

Subspace Robust Wasserstein Distance (SRW)

- Corresponding "min-max" problem:

$$\mathcal{S}_k^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \sup_{\dim(E)=k} \int \|P_E(x)-P_E(y)\|^2\, d\pi(x,y)$$

- SRW is a convex relaxation of PRW:

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

- SRW finds a coupling π minimizing the spectral cost:

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi)$$

Where V_π is the Second Order Moment Matrix of the Displacements:

$$V_\pi = \int (x-y)(x-y)^T d\pi(x,y)$$

III. The SRW Geometry

SRW is a distance between probability measures

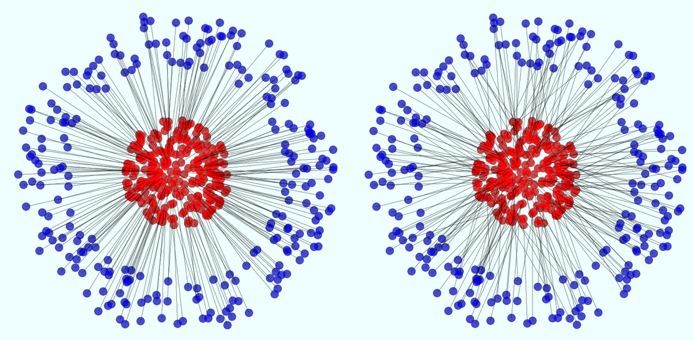
SRW is equivalent to the 2-Wasserstein distance

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

Geodesics in SRW space

$\pi^* \in \Pi(\mu,\nu)$ minimizing $\pi \mapsto \sum_{l=1}^{k}\lambda_l(V_\pi)$

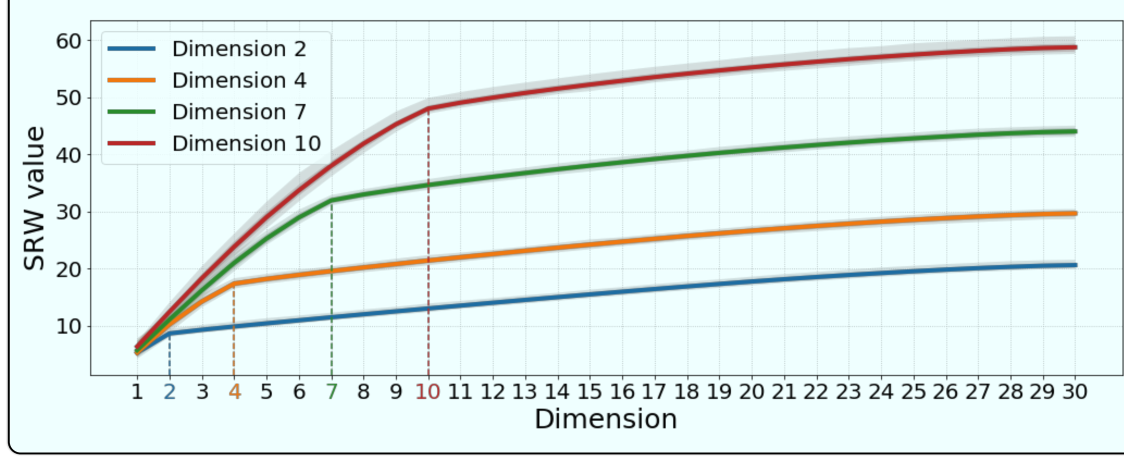
$$t \mapsto \mu_t := \{(1-t)x + ty\}_{\#}\pi^*$$



SRW (left) and W (right) geodesics in presence of noise (d=30)

Dependence on dimension

$k \mapsto \mathcal{S}_k^2(\mu,\nu)$ increasing and concave



- *Measures in dimension d=30, with transport only occurring in dimension 2, 4, 7 and 10 respectively.*
- *Use a 'elbow' rule of thumb to choose k in practice.*

IV. Computing SRW

Entropic Regularization

- Ensures uniqueness of optimal π^*
- Sinkhorn algorithm

Frank-Wolfe

Algorithm Frank-Wolfe algorithm for entropic SRW

Input: Measures (x_i,a_i) and (y_j,b_j), dimension k, regularization strength $\gamma>0$
Initialize Ω
for $t=0$ **to** max_iter **do**
  $\pi \leftarrow \text{reg\_OT}((x,a),(y,b), reg=\gamma, cost=d_\Omega^2)$
  $U \leftarrow$ top k eigenvectors of V_π
  $\tau = 2/(2+t)$
  $\Omega \leftarrow (1-\tau)\Omega + \tau\left[U \operatorname{diag}\left([\mathbf{1}_k, \mathbf{0}_{d-k}]\right) U^T\right]$
end for
Output: $\Omega, \pi, \langle\Omega \mid V_\pi\rangle$

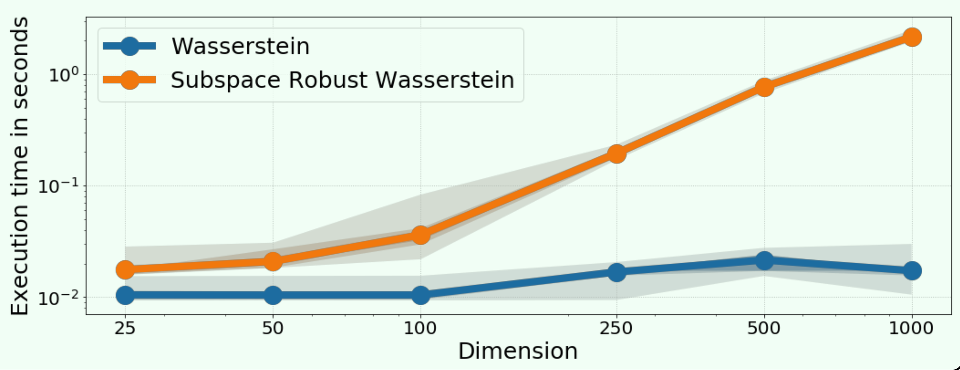
Projected Gradient Method

Algorithm Projected supergradient method for SRW

Input: Measures (x_i,a_i) and (y_j,b_j), dimension k
Initialize Ω
for $t=0$ **to** max_iter **do**
  $\pi \leftarrow \text{OT}((x,a),(y,b), cost=d_\Omega^2)$
  $\Omega \leftarrow \text{Proj}\left[\Omega + \frac{1}{t+1}V_\pi\right]$
end for
Output: $\Omega, \langle\Omega \mid V_\pi\rangle$

Computation Time

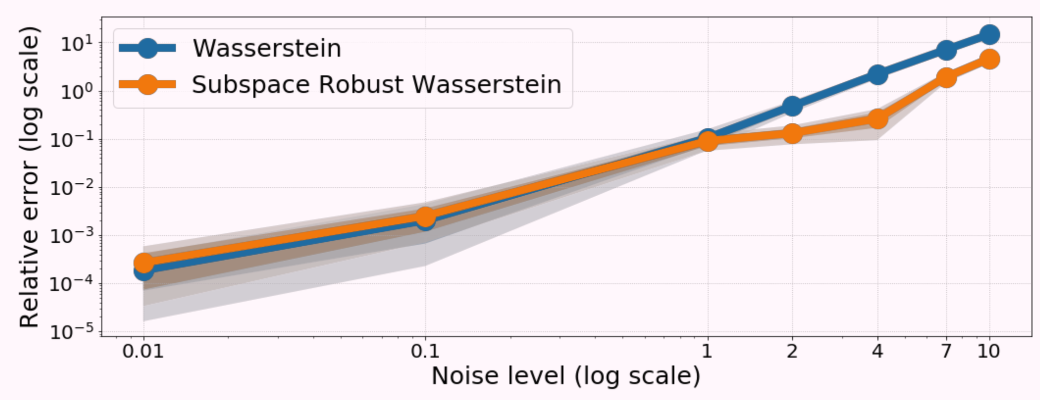
- Warm start in Sinkhorn
- Quadratic in dimension d



V. Applications

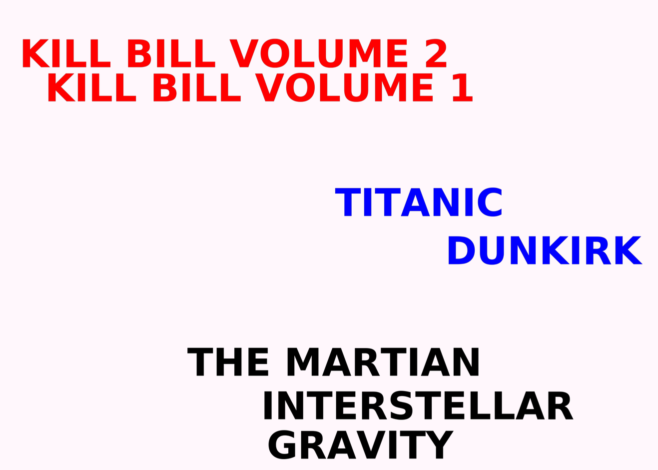
SRW is Robust to Noise

Low-dimensional Gaussians are added noise. We plot the relative error for SRW and W distances.

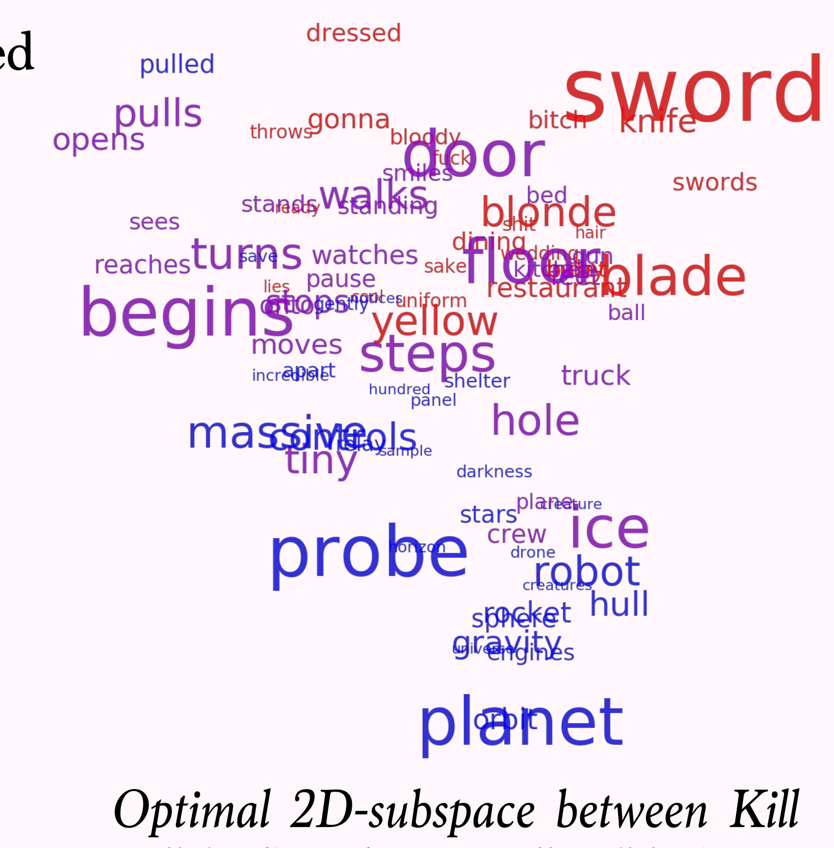


Movies

Scenarios of movies are transformed into measures in $\mathbb{R}^{300}$ using Word2vec



Metric MDS of SRW distances between movies



Optimal 2D-subspace between Kill Bill (red) and Interstellar (blue)