Subspace Robust Wasserstein Distances

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

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 the Wasserstein distance. Efficient algorithms and use case on text data.

github.com/francoispierrepaty/SubspaceRobustWasserstein

I. Wasserstein Distance in High Dimension

Wasserstein Distance

The 2-Wasserstein distance between two probability measures μ and ν is defined by

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

where $\Pi(\mu,\nu)$ is the set of couplings between μ and ν.

The Wasserstein distance has become a classical tool in several machine learning applications:

- Domain adaptation
- Color transfer
- WGAN
- NLP

Curse of Dimensionality

In data sciences, we only have access to data, *i.e.* to random samples $x_1,\dots,x_n \sim \mu$; $y_1,\dots,y_m \sim \nu$

What is really computed is the Wasserstein distance between the empirical measures on the samples $\mathcal{W}(\hat{\mu},\hat{\nu})$

In high-dimension, this can be very different from the real Wasserstein:

$$|\mathcal{W}(\mu,\nu) - \mathcal{W}(\hat{\mu},\hat{\nu})| \sim (1/n)^{1/d}$$

Two-step approach

In applications, people use a two-step approach:

1. Reduce dimension using PCA
2. Compute the Wasserstein distance between the projected point clouds

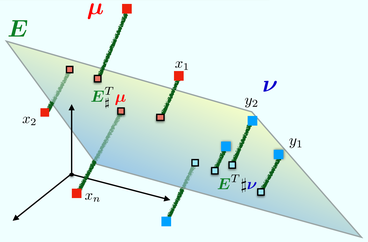
We propose a one-step approach, combining dimensionality reduction and optimal transport.

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:

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

- 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

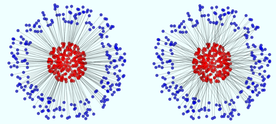
SRW is equivalent to the 2-Wasserstein distance

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

Geodesics in SRW space

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

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

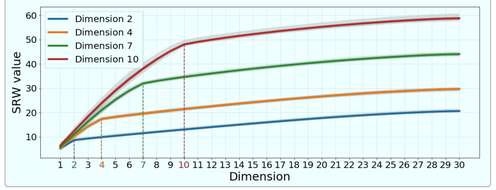


SRW (left) and W (right) geodesics in presence of noise (d=30)

Dependence on dimension

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

- *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**
  $\pi \leftarrow \mathrm{reg\_OT}((x,a),(y,b), reg=\gamma, cost=d^2_\Omega)$
  $U \leftarrow$ top k eigenvectors of V_π
  $\tau = 2/(2+t)$
  $\Omega \leftarrow (1-\tau)\Omega + \tau\left[U\,\mathrm{diag}\left([\mathbf{1}_k, \mathbf{0}_{d-k}]\right)U^T\right]$
end for
Output: $\Omega, \pi, \langle\Omega \,|\, V_\pi\rangle$

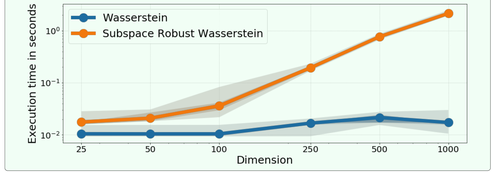
Projected Gradient Method

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**
  $\pi \leftarrow \mathrm{OT}((x,a),(y,b), cost=d^2_\Omega)$
  $\Omega \leftarrow \mathrm{Proj}\left[\Omega + \frac{1}{t+1}V_\pi\right]$
end for
Output: $\Omega, \langle\Omega \,|\, V_\pi\rangle$

Computation Time

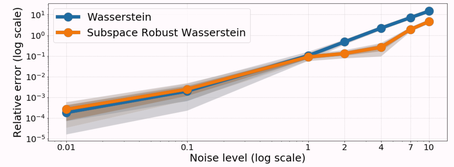
- Warm start in Sinkhorn
- Quadratic in dimension d



V. Experiments

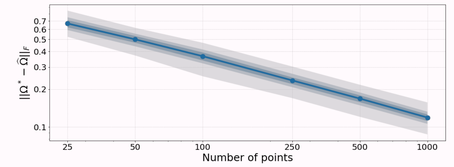
SRW is Robust to Noise

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



Convergence of Subspaces

When the OT plan lies on a subspace, it is retrieved by SRW.



Application in NLP

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)