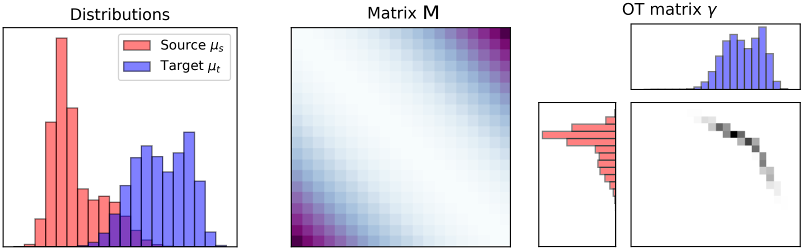
Sliced Gromov-Wasserstein

Titouan Vayer, Remi Flamary, Romain Tavenard,Laetitia Chapel and Nicolas Courty

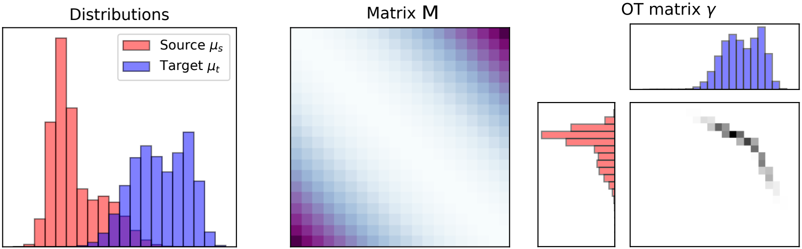
Optimal Transport in a nutshell

Optimal transport with discrete distributions



Let $\mu_s = \sum_{i=1}^{n_s} a_i \delta_{x_i}$ and $\mu_t = \sum_{j=1}^{n_t} b_j \delta_{y_j}$ be two discrete measures.

Optimal transport with discrete distributions



Let $\mu_s = \sum_{i=1}^{n_s} a_i \delta_{x_i}$ and $\mu_t = \sum_{j=1}^{n_t} b_j \delta_{y_j}$ be two discrete measures.

OT Linear Program

$$\boldsymbol{\pi}_0 = \underset{\boldsymbol{\pi} \in \Pi}{\operatorname{argmin}} \left\{ \langle \boldsymbol{\pi}, M \rangle_F = \sum_{i,j} \pi_{i,j} M_{i,j} \right\}$$

where M is a cost matrix with $M_{i,j} = c(x_i, y_j)$ and the marginals constraints are

$$\Pi = \left\{ \boldsymbol{\pi} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \boldsymbol{\pi} \mathbf{1}_{n_t} = a, \boldsymbol{\pi}^T \mathbf{1}_{n_s} = b \right\}$$

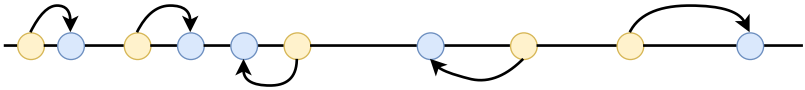
Solved with Network Flow solver of complexity $O(n^3 \log(n))$.

Special case: 1D distribution

We consider the case where $c(x, y)$ is a strictly convex and increasing function of $|x - y|$ and μ_s, μ_t are 1D distributions.

- if $x_1 < x_2$ and $y_1 < y_2$, it is easy to check that $c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1)$
- As such, any optimal transport plan respects the ordering of the elements, and the solution is given by the monotone rearrangement of μ_s onto μ_t

This gives very simple algorithm to compute the transport in $O(n \log n)$, by sorting both x_i and y_i and summing the absolute values of differences.



Sliced Radon Wasserstein

For $\mu_s = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\mu_t = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}$ with $x_i, y_j \in \mathbb{R}^p$

The principle is simple: slice the distribution along lines, project the measures onto it and compute $1D$ Wasserstein along those projections.

Sliced Radon Wasserstein

For $\mu_s = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\mu_t = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}$ with $x_i, y_j \in \mathbb{R}^p$

The principle is simple: slice the distribution along lines, project the measures onto it and compute $1D$ Wasserstein along those projections.

$\mathbf{S}^{p-1} = \{\theta \in \mathbb{R}^p : \|\theta\|_{2,p} = 1\}$ be the p-dimensional hypersphere and λ_{p-1} the uniform measure on $\mathbf{S}^{p-1}$. For θ we note P_θ the projection on θ, *i.e* $P_\theta(x) = \langle x, \theta\rangle$.

p-sliced Wasserstein distance pSW [Bonneel et al., 2015a]

$$pSW_2^2(\mu_s, \mu_t) = \int_{\mathbb{S}^{p-1}} W_2^2(P_\theta\#\mu_s, P_\theta\#\mu_t)d\lambda_{p-1}(\theta) \quad (1)$$

Many applications: barycenter computation [Bonneel et al., 2015b], classification [Kolouri et al., 2016] generative modeling [Kolouri et al., 2019, Deshpande et al., 2018].

Since $P_\theta\#\mu_s, P_\theta\#\mu_t$ are 1D distributions it can be computed in $O(Ln\log(n))$ with L the number of lines sampled for the Monte-Carlo estimation of (1).

Can handle millions of points!

Gromov-Wasserstein distance (GW)

Gromov-Wasserstein distance (GW)

Now what if μ_s, μ_t are not in the same metric space ?

$\mu_s = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\mu_t = \sum_{i=1}^{m} b_j \delta_{y_j}$ with $x_i \in X, y_j \in Y$ (e.g with $\mathbb{R}^p, \mathbb{R}^q$ with $p \leq q$).

Gromov-Wasserstein distance (GW)

Now what if μ_s, μ_t are not in the same metric space ?

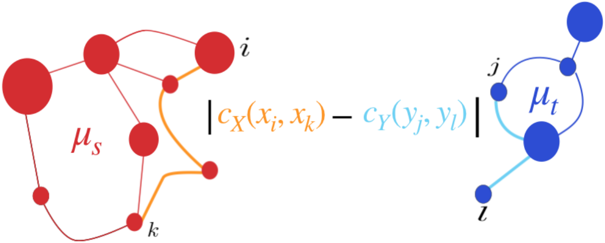
$\mu_s = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\mu_t = \sum_{i=1}^{m} b_j \delta_{y_j}$ with $x_i \in X, y_j \in Y$ (e.g with $\mathbb{R}^p, \mathbb{R}^q$ with $p \leq q$).

Let $c_X : X \times X \to \mathbb{R}_+$ (*resp.* $c_Y : Y \times Y \to \mathbb{R}_+$) measure the similarity between the samples. (GW) distance is defined as:

$$GW_2^2(c_X, c_Y, \mu_s, \mu_t) = \min_{\pi \in \Pi(a,b)} J(c_X, c_Y, \pi) \tag{2}$$

where

$$J(c_X, c_Y, \pi) = \sum_{i,j,k,l} \left| c_X(x_i, x_k) - c_Y(y_j, y_l) \right|^2 \pi_{i,j} \pi_{k,l}. \tag{3}$$



- Distance over measures with no common ground space *w.r.t* "isometric relations".
- Invariant to rotations and translation in either spaces.

Optimization

The optimization problem (2) is a non-convex Quadratic Program (QP) = notoriously hard.

- Conditional Gradient (aka Frank Wolfe) [Vayer et al., 2019]: $O(kn^3)$
- Entropic regularization [Peyré et al., 2016]: nearly $O(n^2)$ and implemented efficiently on GPU. The computation of the final cost is $O(n^3)$

Is there a way to define a sliced version of GW in order to speed up the computation of the underlying problem ?

Solving a Quadratic Assignement Problem in 1D

Koopmans-Beckmann form [Koopmans and Beckmann, 1957] a QAP takes as input matrices $A = (a_{ij})$, $B = (b_{ij})$.

Goal: find a permutation $\sigma \in S_n$ which minimizes the objective function

$$\sum_{i,j=1}^{n} a_{i,j} b_{\sigma(i),\sigma(j)} \tag{4}$$

$\implies$ Generally NP-hard

Some solutions when matrices A and B have simple known structures (for *e.g.* $a_{i,j} = \alpha_i \alpha_j$) [Çela et al., 2018, Çela et al., 2011, Çela et al., 2015]

In the paper we proved the following theorem:

Theorem

For real numbers $x_1 \leq ... \leq x_n$ *and* $y_1 \leq ... \leq y_n$,

$$\min_{\sigma \in S_n} \sum_{i,j} -(x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{5}$$

is achieved either by the identity permutation $\sigma(i) = i$ *or the anti-identity permutation* $\sigma(i) = n + 1 - i$.

So for any real numbers finding the solution to (5) is $O(n \log(n))$

Gromov-Wasserstein distance on the real line

Gromov-Monge (GM)

When $n = m$ and $a_i = b_j = \frac{1}{n}$ we look for the *hard assignment* version of the GW distance resulting on the Gromov-Monge problem [Mémoli and Needham, 2018]:

$$GM_2(c_X, c_Y, \mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n^2} \sum_{i,j} \left| c_X(x_i, x_j) - c_Y(y_{\sigma(i)}, y_{\sigma(j)}) \right|^2 \tag{6}$$

$\sigma \in S_n$ is a one-to-one mapping

Gromov-Monge (GM)

When $n = m$ and $a_i = b_j = \frac{1}{n}$ we look for the *hard assignment* version of the GW distance resulting on the Gromov-Monge problem [Mémoli and Needham, 2018]:

$$GM_2(c_X, c_Y, \mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n^2} \sum_{i,j} \left| c_X(x_i, x_j) - c_Y(y_{\sigma(i)}, y_{\sigma(j)}) \right|^2 \tag{6}$$

$\sigma \in S_n$ is a one-to-one mapping

Using recent advances in graph matching we can prove [Maron and Lipman, 2018]:

Theorem

Let $\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \in \mathcal{P}(\mathbb{R})$ *and* $\nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_j} \in \mathcal{P}(\mathbb{R})$ *with* $d(x, x') = |x - x'|$. *Then:*

$$GW_2(d^2, \mu, \nu) = GM_2(d^2, \mu, \nu)$$

For euclidean distances, uniform weights and same number of atoms, the minimum is in the corner of the Birkhoff polytope! (as for Wass)

Using the two previous theorems:

Theorem (Closed form for GW between 1D discrete measures)

For $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \in \mathcal{P}(\mathbb{R})$ *and* $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_j} \in \mathcal{P}(\mathbb{R})$ *the* GW *distance can be computed in* $O(n\log(n))$ *using simple sorts.*

Using the two previous theorems:

Theorem (Closed form for GW between 1D discrete measures)
For $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \in \mathcal{P}(\mathbb{R})$ and $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_j} \in \mathcal{P}(\mathbb{R})$ the GW distance can be computed in $O(n\log(n))$ using simple sorts.

Indeed:

$$
\begin{aligned}
GW_2(d^2,\mu,\nu) = GM_2(d^2,\mu,\nu) &= \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i,j}\left|(x_i-x_j)^2-(y_{\sigma(i)}-y_{\sigma(j)})^2\right|^2 \\
&= C + \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i,j} -(x_i-x_j)^2(y_{\sigma(i)}-y_{\sigma(j)})^2
\end{aligned} \tag{7}
$$

Then finding σ is $O(n\log(n))$ and computing the cost is provably $O(n)$ (by developing the sum using binomial expansion) so that the overall complexity is $O(n\log(n))$.

Using the two previous theorems:

Theorem (Closed form for GW between 1D discrete measures)
For $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \in \mathcal{P}(\mathbb{R})$ *and* $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_j} \in \mathcal{P}(\mathbb{R})$ *the* GW *distance can be computed in* $O(n\log(n))$ *using simple sorts.*

Indeed:

$$\begin{aligned} GW_2(d^2,\mu,\nu) = GM_2(d^2,\mu,\nu) &= \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i,j}\left|(x_i-x_j)^2-(y_{\sigma(i)}-y_{\sigma(j)})^2\right|^2 \\ &= C + \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i,j} -(x_i-x_j)^2(y_{\sigma(i)}-y_{\sigma(j)})^2 \end{aligned} \tag{7}$$

Then finding σ is $O(n\log(n))$ and computing the cost is provably $O(n)$ (by developing the sum using binomial expansion) so that the overall complexity is $O(n\log(n))$.

$\equiv$ On the real line GW is as difficult as W!!

Sliced Gromov Wasserstein

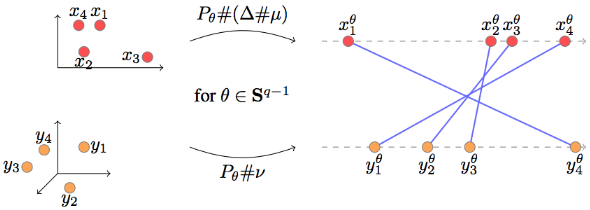
Sliced Gromov Wasserstein (SGW)

Let $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_j}$ with $x_i \in \mathbb{R}^p, y_j \in \mathbb{R}^q$.

For a linear map $\Delta \in \mathbb{R}^{q\times p}$ we define the Sliced Gromov-Wasserstein (SGW) as follows:

$$SGW_{\Delta}(\mu,\nu) = \int_{\mathbf{S}^{q-1}} GW_2^2(d^2, P_\theta\#\mu_\Delta, P_\theta\#\nu)d\lambda_{q-1}(\theta) : \quad (8)$$

where $\mu_\Delta = \Delta\#\mu \in \mathcal{P}(\mathbb{R}^q)$. Can be computed in $O(Ln\log(n))$ as SW.



Δ

Δ acts as a mapping for a point in $\mathbb{R}^p$ of the measure μ onto $\mathbb{R}^q$. One straightforward choice the "uplifting" operator which pads each point of the measure with zeros: $\Delta_{pad}(x) = (x_1, \ldots, x_p, \underbrace{0, \ldots, 0}_{q-p})$.

Δ acts as a mapping for a point in $\mathbb{R}^p$ of the measure μ onto $\mathbb{R}^q$. One straightforward choice the "uplifting" operator which pads each point of the measure with zeros: $\Delta_{pad}(x) = (x_1, \dots, x_p, \underbrace{0, \dots, 0}_{q-p})$.

Fixing $\Delta \implies$ loose some property of GW.

We define Rotation Invariant SGW ($RISGW$):

$$RISGW(\mu, \nu) = \min_{\Delta \in \mathbb{V}_p(\mathbb{R}^q)} SGW_\Delta(\mu, \nu) \tag{9}$$

We propose to minimize SGW_Δ with respect to Δ in the Stiefel manifold [Absil et al., 2009].

SGW holds various properties of the GW distance as summarized in the following theorem:

Theorem

Properties of SGW

- *For all Δ, SGW_Δ and $RISGW$ are translation invariant. $RISGW$ is also rotational invariant when $p = q$, more precisely if $Q \in \mathcal{O}(p)$ is an orthogonal matrix, $RISGW(Q\#\mu, \nu) = RISGW(\mu, \nu)$*
- *SGW and $RISGW$ are pseudo-distances on $\mathcal{P}(\mathbb{R}^p)$, i.e they are symetric, satisfy the triangle inequality and $SGW(\mu, \mu) = RISGW(\mu, \mu) = 0$.*
- *For $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ as defined previously, if $SGW(\mu, \nu) = 0$ then μ and ν are isomorphic for the distance induce by the ℓ_1 norm on $\mathbb{R}^p$. In particular this implies $GW_2(d_{\|.\|_{1,p}}, \mu, \nu) = 0$.*

Experiments

Spiral Dataset

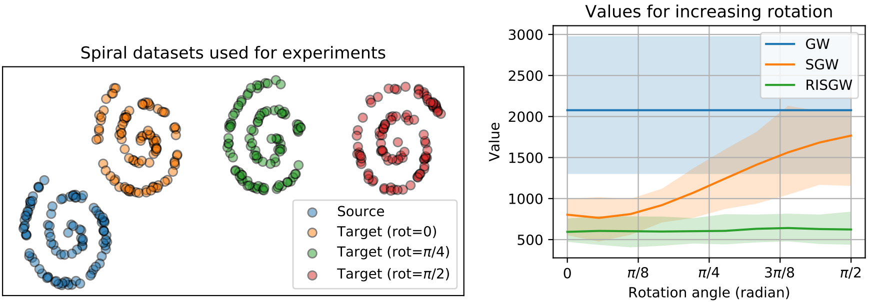


Figure 1: Illustration of SGW, $RISGW$ and GW on spiral datasets for varying rotations on discrete 2D spiral datasets. (left) Examples of spiral distributions for source and target with different rotations. (right) Average value of SGW, GW and $RISGW$ with $L = 20$ as a function of rotation angle of the target. Colored areas correspond to the 20% and 80% percentiles.

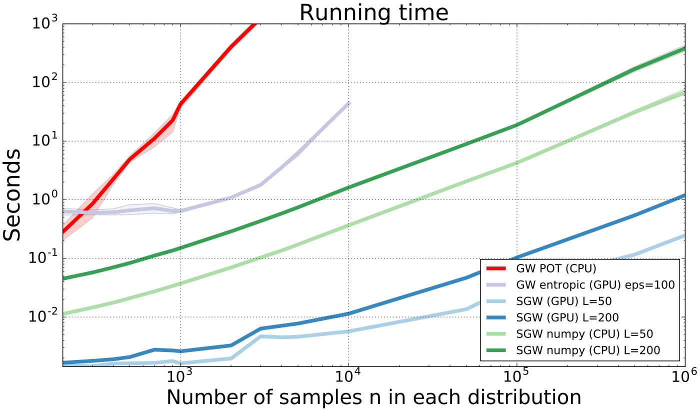


Figure 2: Runtimes comparison between SGW, GW, entropic-GW between two 2D random distributions with varying number of points from 0 to 10^6 in log-log scale. The time includes the calculation of the pair-to-pair distances.

Meshes comparison

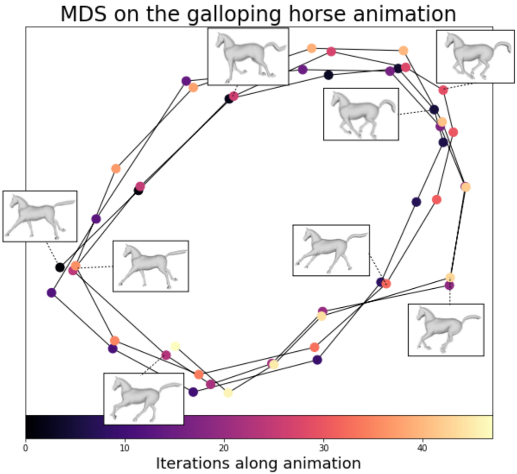


Figure 3: Each sample in this Figure corresponds to a mesh and is colored by the corresponding time iteration. One can see that the cyclical nature of the motion is recovered.

GAN

$$G^* = \operatorname{argmin} GW_2^2(c_X, c_{G(Z)}, \mu, \nu_G), \tag{10}$$

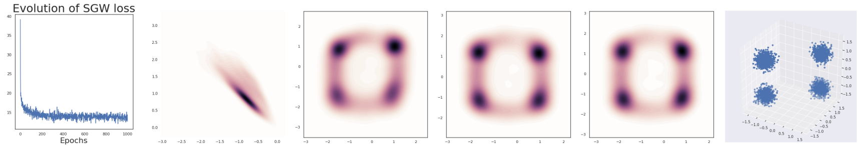


Figure 4: Using SGW in a GAN loss. First image shows the loss value along epochs. The next 4 images are produced by sampling the generated distribution ($3,000$ samples, plotted as a continuous density map). Last image shows the target 3D distribution.

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009).
Optimization algorithms on matrix manifolds.
Princeton University Press.

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

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

Çela, E., Deineko, V., and Woeginger, G. J. (2018).
New special cases of the Quadratic Assignment Problem with diagonally structured coefficient matrices.
European journal of operational research, 267(3):818–834.

Çela, E., Schmuck, N. S., Wimer, S., and Woeginger, G. J. (2011).
The Wiener maximum quadratic assignment problem.
Discrete Optimization, 8:411–416.

Çela, E., Deineko, V. G., and Woeginger, G. J. (2015).
Well-solvable cases of the QAP with block-structured matrices.
Discrete applied mathematics, 186:56–65.

Deshpande, I., Zhang, Z., and Schwing, A. G. (2018).
Generative modeling using the sliced wasserstein distance.
In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. (2019).
Sliced wasserstein auto-encoders.
In *International Conference on Learning Representations*.

Kolouri, S., Zou, Y., and Rohde, G. K. (2016).
Sliced wasserstein kernels for probability distributions.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koopmans, T. and Beckmann, M. J. (1957).
Assignment problems and the location of economic activities.
Econometrica: journal of the Econometric Society, (53–76).

Maron, H. and Lipman, Y. (2018).
(probably) concave graph matching.
In *Advances in Neural Information Processing Systems*, pages 408–418.

Mémoli, F. and Needham, T. (2018).
Gromov-Monge quasi-metrics and distance distributions.
arXiv:1810.09646.

Peyré, G., Cuturi, M., and Solomon, J. (2016).
Gromov-Wasserstein Averaging of Kernel and Distance Matrices.
In *ICML*, Proc. 33rd ICML, New-York, United States.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2019).
Optimal Transport for structured data with application on graphs.
In *International Conference on Machine Learning*, volume 97.