Fused Gromov Wasserstein distance

Titouan Vayer Univ. Bretagne-Sud, IRISA
Laetitia Chapel Univ. Bretagne-Sud, IRISA
Remi Flamary Univ. Côte d'Azur, OCA Lagrange
Romain Tavenard Univ. Rennes, LETG
Nicolas Courty Univ. Bretagne-Sud, IRISA

June 24, 2018

Optimal Transport for structured data

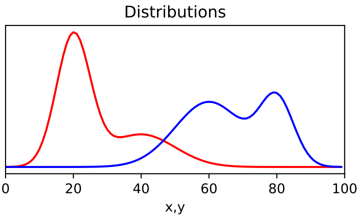
Classical Optimal Transport deals with distribution but can not leverage the specific relation amoung the component of the distribution.

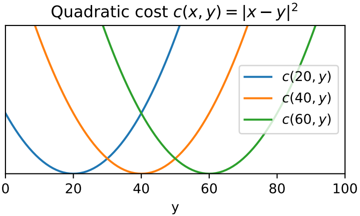
- How to include this structural information in the optimal transportation formulation ?
- How to use the new formulation in order to compare structured data (graphs, times series...)

Overview

1. Generality about Optimal Transport
   - Wasserstein distance
   - Gromov-Wasserstein distance
2. Structured data as distributions
3. A novel OT distance for structured data : Fused Gromov-Wasserstein distance

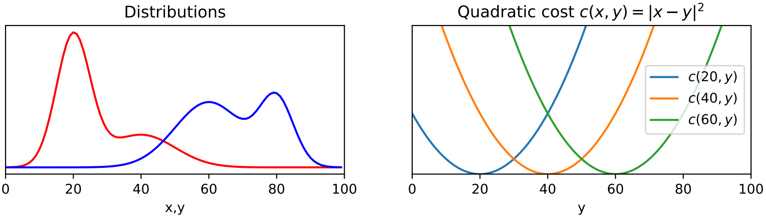
- Mathematical tools aiming at comparing distributions





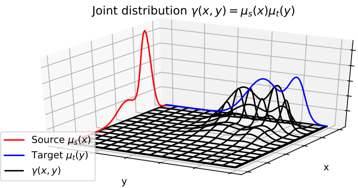
Optimal transport (Monge formulation)

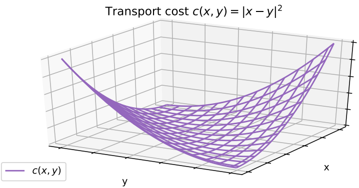
- Mathematical tools aiming at comparing distributions



- Probability measures μ_s and μ_t on Ω_s, Ω_t with a cost function $d : \Omega_s \times \Omega_t \rightarrow \mathbb{R}^+$.
- The Monge formulation aim at finding a mapping $T : \Omega_s \rightarrow \Omega_t$

$$\inf_{T\#\mu_s=\mu_t} \int_{\Omega_s} d(x, T(x))\mu_s(x)dx \tag{1}$$



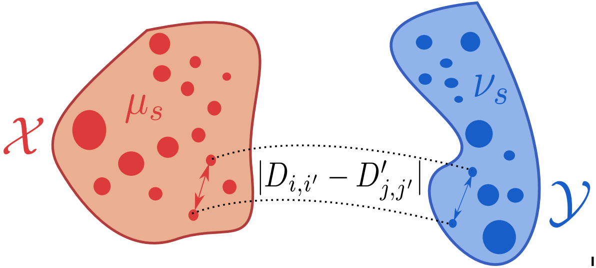


$\mu_s = \sum_{i=1}^n a_i \delta_{x_i}$ and $\mu_t = \sum_{j=1}^m b_j \delta_{y_j}$ on a commun ground space equipped with a distance

- The Kantorovich formulation seeks for a probabilistic coupling $\pi \in \Pi(\mu_s \times \mu_t)$ between μ_s and μ_t.
- π is a joint probability measure with prescribed marginals μ_s and μ_t.
- Computes the Wasserstein distance :

$$\mathcal{W}_p(\mu_s, \mu_t) = \left(\min_{\pi \in \Pi(\mu_s, \mu_t)} \sum_{i,j} d(x_i, y_j)^q \pi_{i,j} \right)^{\frac{1}{p}} \tag{2}$$

Optimal transport distance over measures with no common ground space. Compare the intrinsic distances in each space.



Inspired from Gabriel Peyré

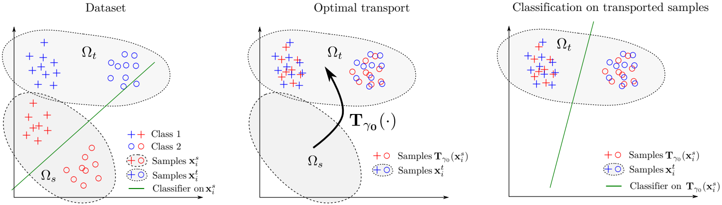
$\mu_s = \sum_i a_i \delta_{v_i}$ and $\mu_t = \sum_j b_j \delta_{w_j}$

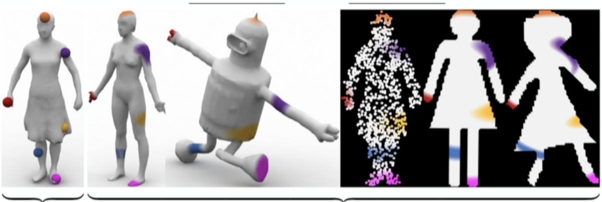
$$\mathcal{GW}_p(D, D', \mu_s, \mu_t) = \left(\min_{\pi \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p \pi_{i,j}\, \pi_{k,l} \right)^{\frac{1}{p}}$$

Optimal transport in Machine Learning

Numerous applications for this two distances. Some of them :

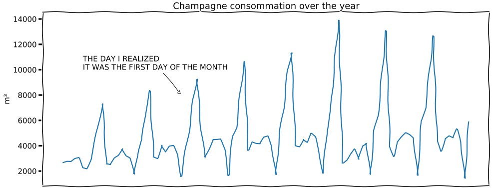
- Learning with Wasserstein Loss [FZM+15]
- Wasserstein GAN's [ACB17]
- Domain Adaptation [CFTR17]
- Image colorization [FPPA14], Dictionary Learning [RCP16] ...
- For GW : Shape comparaison [Mem11], shape barycenter [PCS16].

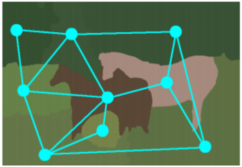




Why Structured data are interesting ?

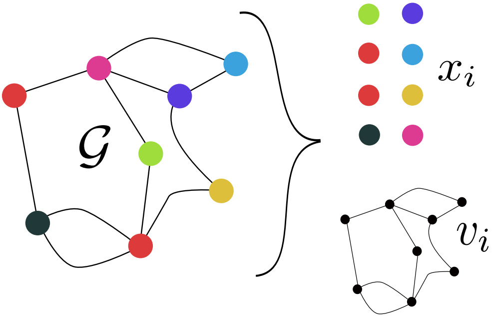
- Systems are usually complex compositions of entities and their interactions
- Crucial to include structural information in order to learn from small amounts of experience [BHB+18]
- A structure data is viewed as a combination of features informations linked within each other by some structural information.





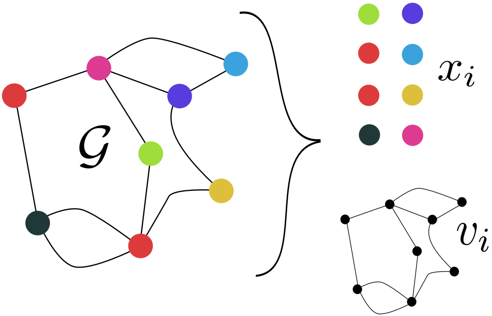
Structured data as distributions

Graphs are natural representations of discrete structured data of the type $\mu = \sum_{i=1}^{n} a_i \delta_{(x_i, v_i)}$:



Structured data as distributions

Graphs are natural representations of discrete structured data of the type $\mu = \sum_{i=1}^{n} a_i \delta_{(x_i, v_i)}$:



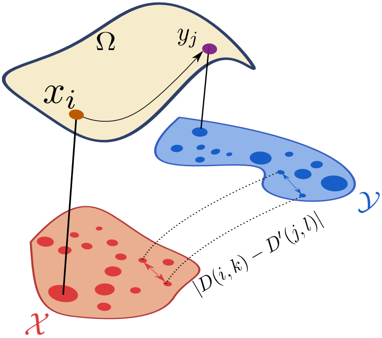
Problem when comparing two structured data on (x_i, v_i) and (y_j, w_j)

- Features value x_i and y_j can be compared through the common metric
- But no common between the structure points v_i and w_j

A novel OT distance for structured data : Fused Gromov-Wasserstein distance

→ Combining Wasserstein and Gromov-Wasserstein approach we define for measure on structured data $\mu_s = \sum_{i=1}^{n} a_i \delta_{x_i, v_i}$ and $\mu_t = \sum_{j=1}^{m} b_j \delta_{y_j, w_j}$

$$\mathcal{FGW}_{p,q,\alpha}(D, D', \mu_s, \mu_t) = \left(\min_{\pi \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) d(x_i, y_j)^q + \alpha |D_{i,k} - D'_{j,l}|^q \right)^p \pi_{i,j} \, \pi_{k,l} \right)^{\frac{1}{p}}$$



for $\alpha \in [0, 1]$ a trade off parameter between structure and features

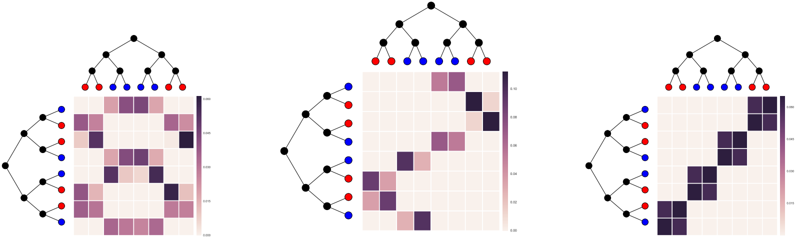
Algorithmic solution

- Optimization problem is a non-convex Quadratic program : can be solved with Conditional gradient [FPPA14] with OT solver.
- Convergence in local minima insured by Frank-Wolfe algorithm proprieties [LJ16].

Algorithmic solution

- Optimization problem is a non-convex Quadratic program : can be solved with Conditional gradient [FPPA14] with OT solver.
- Convergence in local minima insured by Frank-Wolfe algorithm proprieties [LJ16].
- Entropic regularization can be defined and allows Projected gradients with Sinkhorn [PCS16] using Bregman projections.

Optimal maps on toy trees as α increases



(c) $W = 0$ (d) $FGW \neq 0$ (e) $GW = 0$

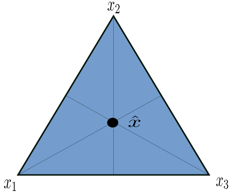
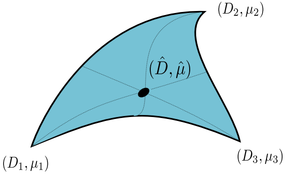
Application of FGW distance

We use our distance for graph classification and compare accuracies on classical graph datasets with state-of-the-art graph kernel approaches and CNN approach.

	Labeled Graphs			Social Graphs		Vector attributes Graph	
Dataset	MUTAG	PTC	NCI1	IMDB-B	IMDB-M	PROTEIN	ENZYMES
WL	80.72±3.0	56.97±2.0	80.22±0.5	-	-	72.9±0.5	53.7±1.4
GK	81.58±2.1	57.32±1.1	43.89±0.4	65.87±0.98	43.89±0.38	62.28±0.29	-
RW	83.68±1.66	57.26±1.30	-	-	-	74.22± 0.42	-
SP	85.79±2.51	58.53±2.55	73.00±0.51	-	-	75.07 ±0.54	-
WL-OA	84.5±1.7	**63.6±1.5**	**86.1±0.2**	-	-	76.4 ±0.4	59.9 ±1.1
PSCN $k = 10$	**88.95±4.37**	62.29±5.62	76.34±1.68	**71.00±2.29**	45.23±2.84	75.00±2.51	-
FGW CG	86.8±5.4	58.3±8.4	78.7±1.9	66.4±3.6	**48.5±3.0**	76.0±1.9	**66.3±6.5**
FGW SINK	81.6±5.9	56.9±6.6	75.3±2.3	70.6±2.8	45.5±2.8	**77.0±4.0**	55.5±5.1

FGW barycenter

We define barycenter of structured data using Fréchet mean :

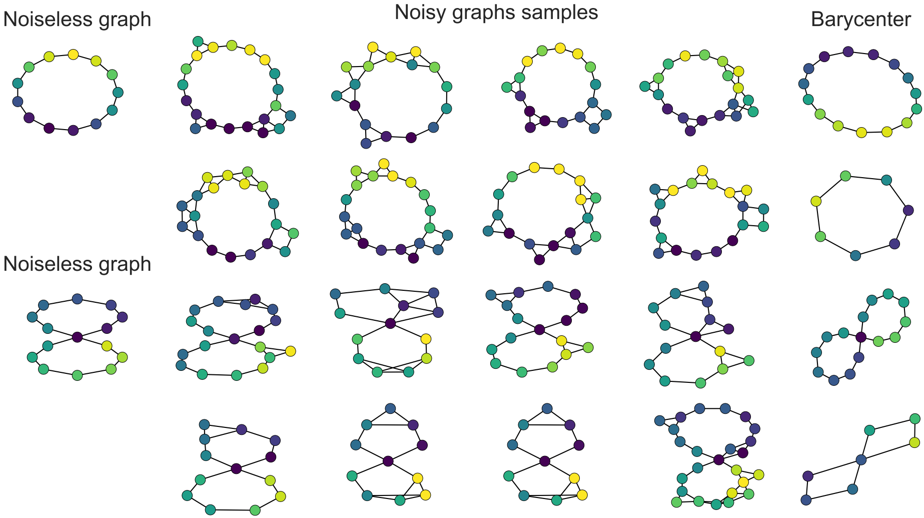
Euclidean barycenter	FGW barycenter
	
$\underset{x}{min} \sum_k \lambda_k \|x - x_k\|^2$	$\min_{D \in \mathbb{R}^{N \times N}, \mu} \sum_k \lambda_k \mathcal{FGW}_{1,q,\alpha}(D, D_k, \mu, \mu_k)$

Barycenters solved via block coordinate relaxation. Several variants of this problem :

- Computing the structure with fixed features
- Computing the features with fixed structure.
- Both features and structured unknown

FGW barycenter

We applied on toy noisy graphs :



Conclusion

- New versatile method for comparing structured data based on Optimal Transport
- Many desirable distance properties
- New notion of barycenter of structured data such as graphs or time series
- Promising applications for signal over graphs and deep learning for structured data

Martín Arjovsky, Soumith Chintala, and Léon Bottou, Wasserstein generative adversarial networks, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, 2017, pp. 214–223.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, Relational inductive biases, deep learning, and graph networks, ArXiv e-prints (2018).

Marco Cuturi and Mathieu Blondel, Soft-DTW: a differentiable loss function for time-series, Proceedings of the ICML (International Convention Centre, Sydney, Australia), vol. 70, PMLR, 06–11 Aug 2017, pp. 894–903.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, Optimal transport for domain adaptation, IEEETPAMI **39** (2017), no. 9, 1853–1865.

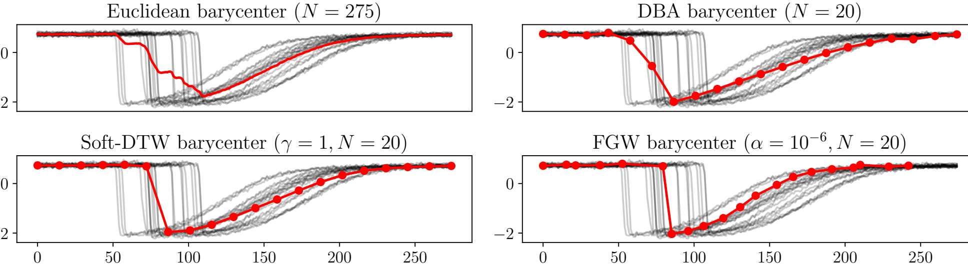
References II

- Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol, Regularized discrete optimal transport, SIAM Journal on Imaging Sciences **7** (2014), no. 3, 1853–1882.
- C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio, Learning with a Wasserstein Loss, ArXiv e-prints (2015).
- Simon Lacoste-Julien, Convergence rate of frank-wolfe for non-convex objectives, arXiv preprint arXiv:1607.00345 (2016).
- Facundo Memoli, Gromov wasserstein distances and the metric approach to object matching, Foundations of Computational Mathematics (2011), 1–71, 10.1007/s10208-011-9093-5.
- Gabriel Peyré, Marco Cuturi, and Justin Solomon, Gromov-wasserstein averaging of kernel and distance matrices, ICML, 2016, pp. 2664–2672.
- François Petitjean, Alain Ketterlin, and Pierre Gançarski, A global averaging method for dynamic time warping, with applications to clustering, Elsevier Pattern Recognition **44** (2011), no. 3, 678–693.

Antoine Rolet, Marco Cuturi, and Gabriel Peyré, Fast dictionary learning with a smoothed wasserstein loss, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Cadiz, Spain) (Arthur Gretton and Christian C. Robert, eds.), Proceedings of Machine Learning Research, vol. 51, PMLR, 09–11 May 2016, pp. 630–638.

FGW barycenter

We also applied our barycenter on real time serie dataset and compare with state-of-the-art methods [CB17] [PKG11]



FGW barycenter

Application in mesh interpolation :