Machine learning for graphs and with graphs

Optimal Transport for Graph Learning

Titouan Vayer & Pierre Borgnat
email: titouan.vayer@inria.fr, pierre.borgnat@ens-lyon.fr

October 22, 2024

ENS DE LYON

Table of contents

Acknowledgments

Slides adapted from those of Rémi Flamary.

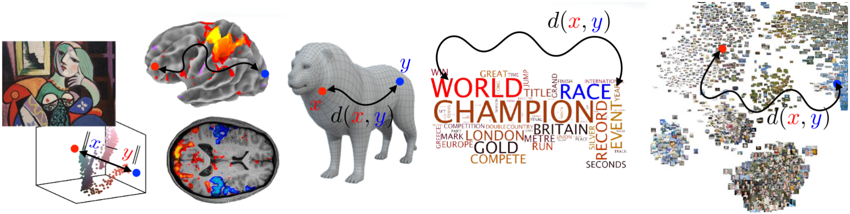
Distributions are everywhere

Distributions are everywhere in machine learning

- Images, vision, graphics, Time series, text, genes, proteins.
- Many datum and datasets can be seen as distributions.
- Important questions:
  - How to compare distributions?
  - How to use the geometry of distributions?
- Optimal transport provides many tools that can answer those questions.

Illustration from the slides of Gabriel Peyré.

Distributions are everywhere



Distributions are everywhere in machine learning

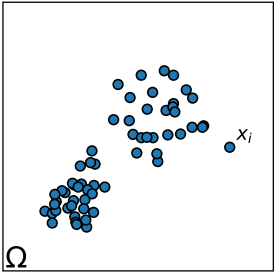
- Images, vision, graphics, Time series, text, genes, proteins.
- Many datum and datasets can be seen as distributions.
- Important questions:
  - How to compare distributions?
  - How to use the geometry of distributions?
- Optimal transport provides many tools that can answer those questions.

Illustration from the slides of Gabriel Peyré.

Discrete distributions: Empirical vs Histogram

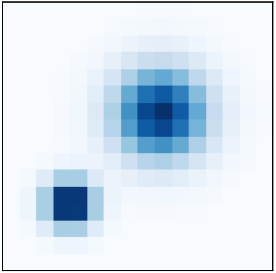
Discrete measure: $\mu = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1$

Lagrangian (point clouds)



- Constant weight: $a_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

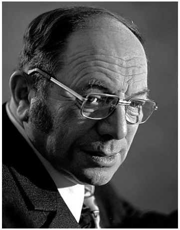
Eulerian (histograms)

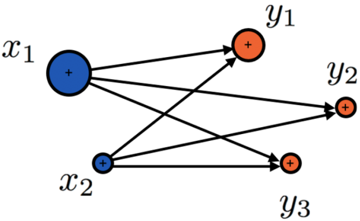


- Fixed positions $\mathbf{x}_i$ e.g. grid
- Convex polytope Σ_n (simplex): $\{(a_i)_i \geq 0; \sum_i a_i = 1\}$

Table of contents

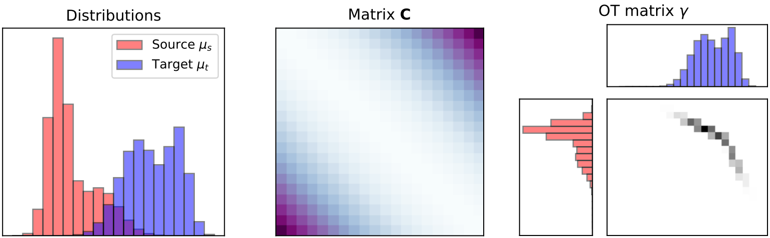
Optimal transport





- Problem introduced by Gaspard Monge in his memoire Monge 1781.
- How to move mass while minimizing a cost (mass + cost)
- Monge formulation seeks for a mapping between two mass distribution.
- Reformulated by Leonid Kantorovich (1912–1986), Economy nobelist in 1975
- Focus on where the mass goes, allow splitting Kantorovich 1942.
- Applications originally for resource allocation problems

Optimal transport between discrete distributions



Kantorovitch formulation : OT Linear Program

When $\mu_s = \sum_{i=1}^{n_s} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n_t} b_i \delta_{\mathbf{x}_i^t}$

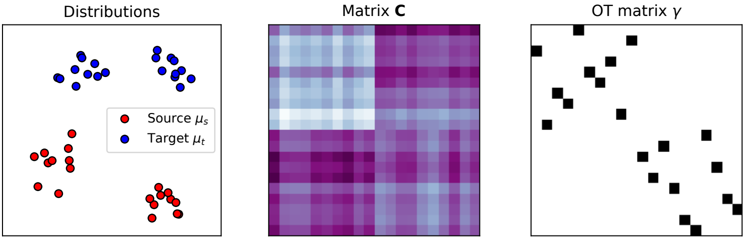
$$W_p^p(\mu_s, \mu_t) = \min_{\boldsymbol{T} \in \Pi(\mu_s, \mu_t)} \left\{ \langle \boldsymbol{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t) = \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^p$ and

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

- $(n = n_s = n_t)$ Solving OT with network simplex is $O(n^3 \log(n))$.
- $W_p(\mu_s, \mu_t)$ is called the Wasserstein distance (EMD for $p = 1$).

Optimal transport between discrete distributions



Kantorovitch formulation : OT Linear Program

When $\mu_s = \sum_{i=1}^{n_s} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n_t} b_i \delta_{\mathbf{x}_i^t}$

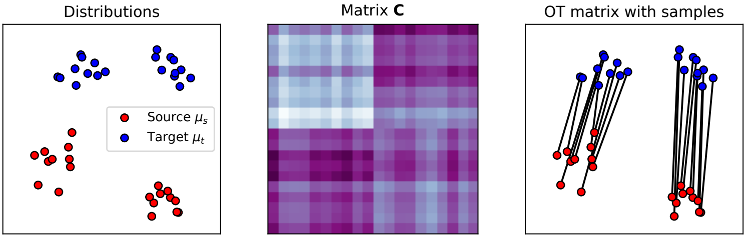
$$W_p^p(\mu_s, \mu_t) = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t) = \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^p$ and

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

- $(n = n_s = n_t)$ Solving OT with network simplex is $O(n^3 \log(n))$.
- $W_p(\mu_s, \mu_t)$ is called the Wasserstein distance (EMD for $p = 1$).

Optimal transport between discrete distributions



Kantorovitch formulation : OT Linear Program

When $\mu_s = \sum_{i=1}^{n_s} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n_t} b_i \delta_{\mathbf{x}_i^t}$

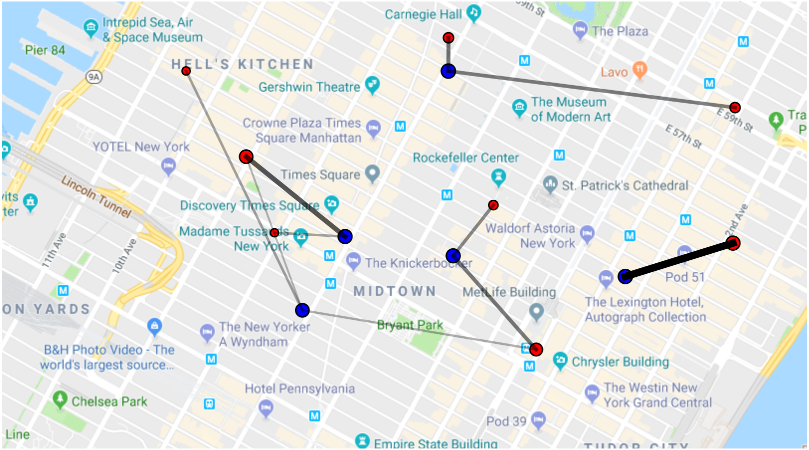
$$W_p^p(\mu_s, \mu_t) = \min_{\boldsymbol{T} \in \Pi(\mu_s, \mu_t)} \left\{ \langle \boldsymbol{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t) = \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^p$ and

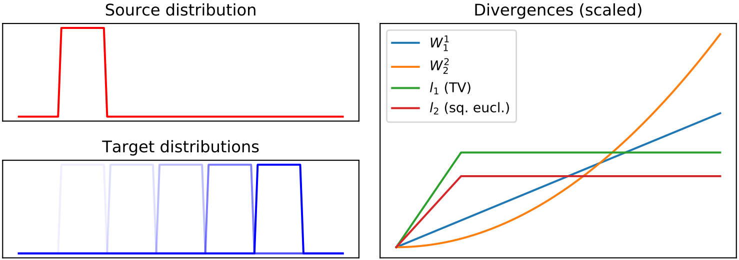
$$\Pi(\mu_s, \mu_t) = \left\{ \boldsymbol{T} \in (\mathbb{R}^+)^{n_s \times n_t} | \; \boldsymbol{T}\mathbf{1}_{n_t} = \mathbf{a}, \boldsymbol{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\}$$

- $(n = n_s = n_t)$ Solving OT with network simplex is $O(n^3 \log(n))$.
- $W_p(\mu_s, \mu_t)$ is called the Wasserstein distance (EMD for $p = 1$).

Boulangeries & Cafés



Wasserstein distance

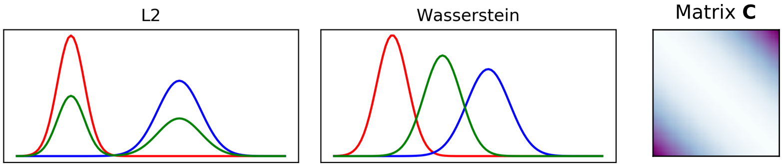


Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \int_{\Omega_s \times \Omega_t} \|\mathbf{x} - \mathbf{y}\|^p \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma}[\|\mathbf{x} - \mathbf{y}\|^p] \quad (1)$$

- Earth Mover's Distance (W_1^1) Rubner, Tomasi, and Guibas 2000.
- Useful between discrete distribution even without overlapping support.
- Smooth approximation can be computed with Sinkhorn Cuturi 2013.
- **Wasserstein barycenter**: $\overline{\mu} = \arg\min_\mu \sum_i w_i W_p^p(\mu, \mu_i)$

Wasserstein distance

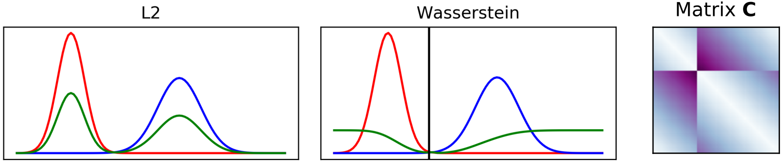


Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \int_{\Omega_s \times \Omega_t} \|\mathbf{x} - \mathbf{y}\|^p \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma}[\|\mathbf{x} - \mathbf{y}\|^p] \quad (1)$$

- Earth Mover's Distance (W_1^1) Rubner, Tomasi, and Guibas 2000.
- Useful between discrete distribution even without overlapping support.
- Smooth approximation can be computed with Sinkhorn Cuturi 2013.
- **Wasserstein barycenter**: $\bar{\mu} = \arg\min_\mu \sum_i w_i W_p^p(\mu, \mu_i)$

Wasserstein distance

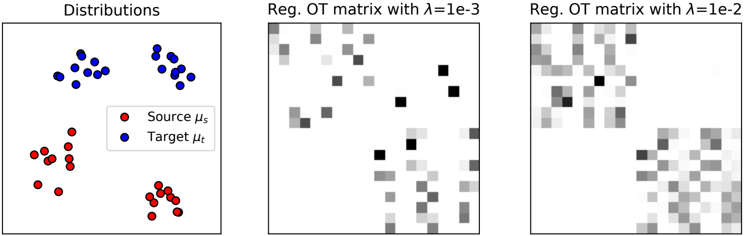


Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \int_{\Omega_s \times \Omega_t} \|\mathbf{x} - \mathbf{y}\|^p \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[\|\mathbf{x} - \mathbf{y}\|^p] \quad (1)$$

- Earth Mover's Distance (W_1^1) Rubner, Tomasi, and Guibas 2000.
- Useful between discrete distribution even without overlapping support.
- Smooth approximation can be computed with Sinkhorn Cuturi 2013.
- **Wasserstein barycenter**: $\overline{\mu} = \arg\min_{\mu} \sum_i w_i W_p^p(\mu, \mu_i)$

Entropic regularized optimal transport



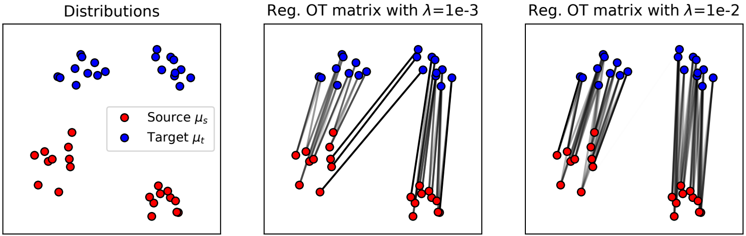
Entropic regularization Cuturi 2013

$$\mathbf{T}_0^\lambda = \underset{\mathbf{T}\in\Pi(\mu_s,\mu_t)}{\arg\min} \quad \langle \mathbf{T}, \mathbf{C}\rangle_F + \lambda \sum_{i,j} T_{i,j}(\log T_{i,j} - 1)$$

- Regularization with the negative entropy of $\boldsymbol{T}$.
- Looses sparsity but smooth and strictly convex optimization problem.
- Can be solved efficiently with Sinkhorn's matrix scaling algorithm with $\mathbf{u}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\mathbf{C}/\lambda)$ and $\mathbf{T} = \text{diag}(\mathbf{u}^\star)\mathbf{K}\text{diag}(\mathbf{v}^\star)$

$$\mathbf{v}^{(k)} = \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}^{(k-1)}, \quad \mathbf{u}^{(k)} = \mathbf{a} \oslash \mathbf{K}\mathbf{v}^{(k)}$$

Entropic regularized optimal transport



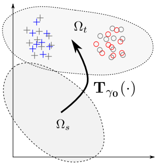
Entropic regularization Cuturi 2013

$$\mathbf{T}_0^\lambda = \underset{\mathbf{T}\in\Pi(\mu_s,\mu_t)}{\arg\min} \quad \langle \mathbf{T}, \mathbf{C}\rangle_F + \lambda \sum_{i,j} T_{i,j}(\log T_{i,j} - 1)$$

- Regularization with the negative entropy of $\boldsymbol{T}$.
- Looses sparsity but smooth and strictly convex optimization problem.
- Can be solved efficiently with Sinkhorn's matrix scaling algorithm with $\mathbf{u}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\mathbf{C}/\lambda)$ and $\mathbf{T} = \text{diag}(\mathbf{u}^\star)\mathbf{K}\text{diag}(\mathbf{v}^\star)$

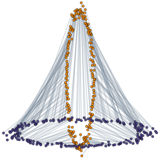
$$\mathbf{v}^{(k)} = \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}^{(k-1)}, \quad \mathbf{u}^{(k)} = \mathbf{a} \oslash \mathbf{K}\mathbf{v}^{(k)}$$

Three aspects of optimal transport



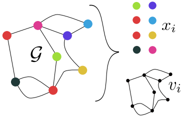
Transporting with optimal transport

- Learn to map between distributions.
- Estimate a smooth mapping from discrete distributions.
- Applications in domain adaptation.



Divergence between histograms

- Use the ground metric to encode complex relations between the bins of histograms for data fitting.
- OT losses are non-parametric divergences between non overlapping distributions.
- Used to train minimal Wasserstein estimators.

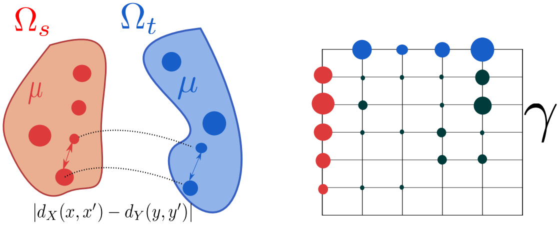


Divergence between graphs

- Modeling of structured data and graphs as distribution.
- OT losses (Wass. or (F)GW) measure similarity between distributions/objects.

Table of contents

Gromov-Wasserstein and extensions



Inspired from Gabriel Peyré

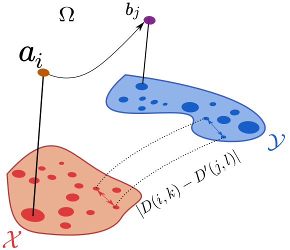
GW for discrete distributions Memoli 2011

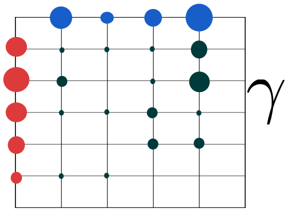
$$\mathcal{GW}_p^p(\mu_s,\mu_t) = \min_{T\in\Pi(\mu_s,\mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j}\, T_{k,l}$$

with $\mu_s = \sum_i a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_j b_j \delta_{x_j^t}$

- Distance between metric measured spaces : across different spaces.
- OT plan that preserves the pairwise relationships between samples.
- Entropy regularized GW proposed in Peyré, Cuturi, and Solomon 2016.

Gromov-Wasserstein and extensions





FGW for discrete distributions Vayer et al. 2018

$$\mathcal{FGW}_p^p(\mu_s, \mu_t) = \min_{T \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) C^q_{i,j} + \alpha |D_{i,k} - D'_{j,l}|^q\right)^p T_{i,j}\, T_{k,l}$$

with $\mu_s = \sum_i a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_j b_j \delta_{x_j^t}$

- Distance between metric measured spaces : across different spaces.
- OT plan that preserves the pairwise relationships between samples.
- Entropy regularized GW proposed in Peyré, Cuturi, and Solomon 2016.

Solving the Gromov Wasserstein optimization problem

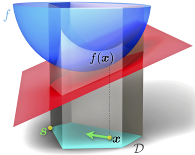
Optimization problem

$$\mathcal{GW}_p^p(\mu_s, \mu_t) = \min_{\boldsymbol{T} \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j} T_{k,l}$$

- Quadratic Program (Wasserstein is a linear program).
- Nonconvex, NP-hard, related to Quadratic Assignment Problem.
- Large problem and non convexity forbid standard QP solvers.

Optimization algorithms

- Local solution with conditional gradient algorithm (Frank-Wolfe) Frank and Wolfe 1956.
- Each FW iteration requires solving an OT problems.
- With entropic regularization, one can use mirror descent Peyré, Cuturi, and Solomon 2016.



Entropic Gromov-Wasserstein

Optimization Problem

$$\mathcal{GW}_{p,\epsilon}^{p}(\mu_s, \mu_t) = \min_{\boldsymbol{T}\in\Pi(\mu_s,\mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j}\ T_{k,l} + \epsilon \sum_{i,j} T_{i,j} \log T_{i,j} \quad (2)$$

- Smoothing the original GW with a convex and smooth entropic term.

Solving the entropic $\mathcal{GW}$ Peyré, Cuturi, and Solomon 2016

- Problem (2) can be solved using a KL mirror descent.
- This is equivalent to solving at each iteration t

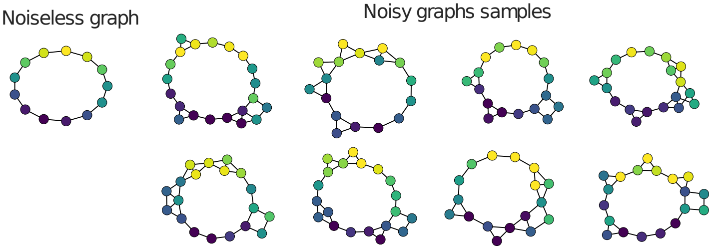
$$\boldsymbol{T}^{(t+1)} = \min_{\boldsymbol{T}\in\mathcal{P}} \quad \left\langle \boldsymbol{T}, \mathbf{G}^{(t)} \right\rangle_F + \epsilon \sum_{i,j} T_{i,j} \log T_{i,j}$$

  Where $G_{i,j}^{(t)} = 2\sum_{k,l} |D_{i,k} - D'_{j,l}|^p T_{k,l}^{(t)}$ is the gradient of the GW loss at previous point $\boldsymbol{T}^{(k)}$.
- Problem above solved using a Sinkhorn-Knopp algorithm of entropic OT.

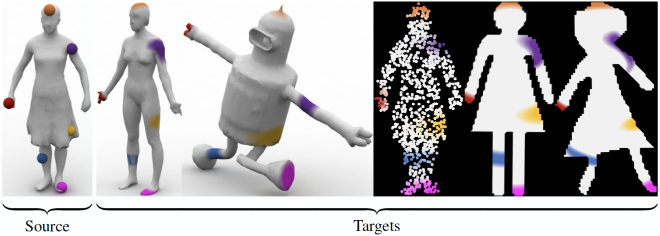
Table of contents

Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

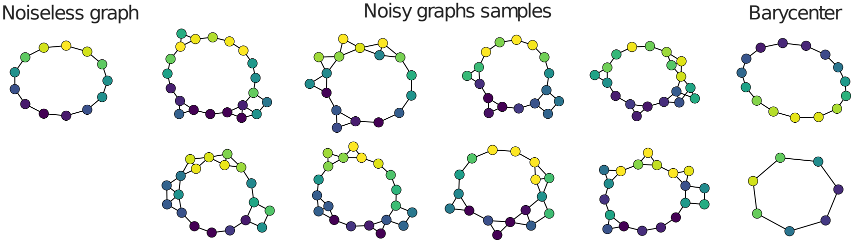


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

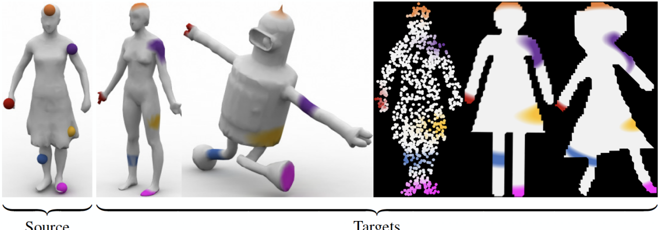


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

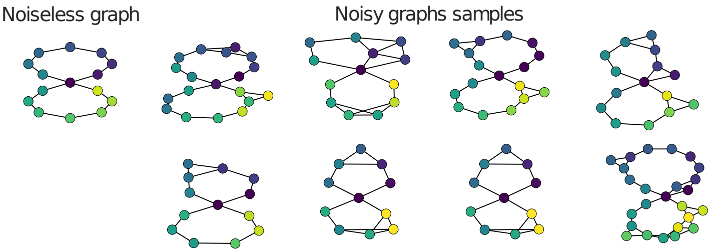


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

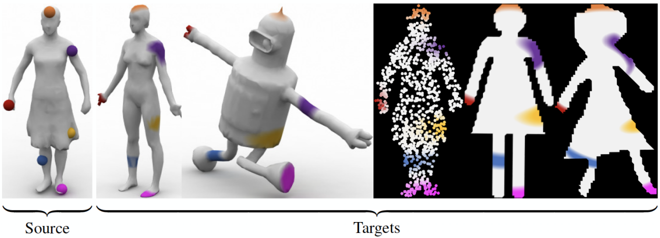


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

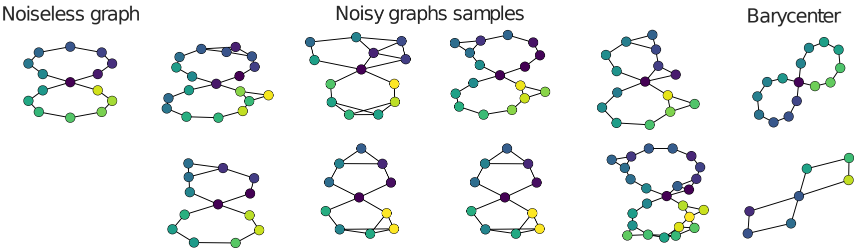


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

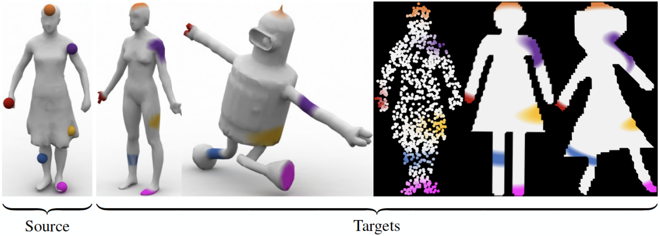


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

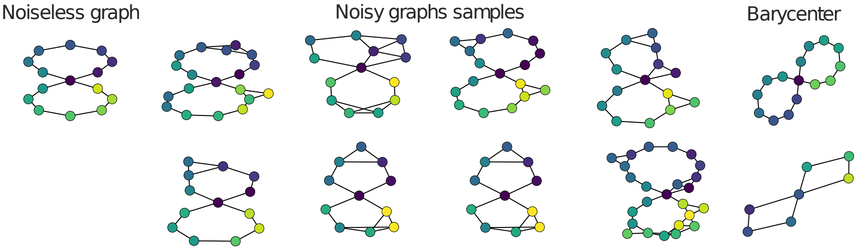


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

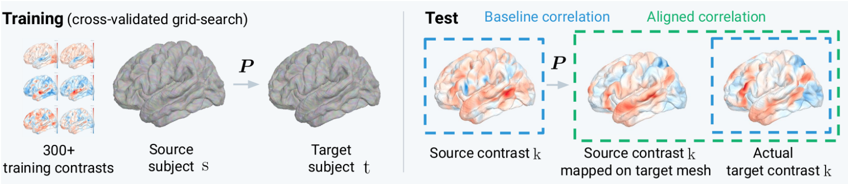


Applications of (F)GW

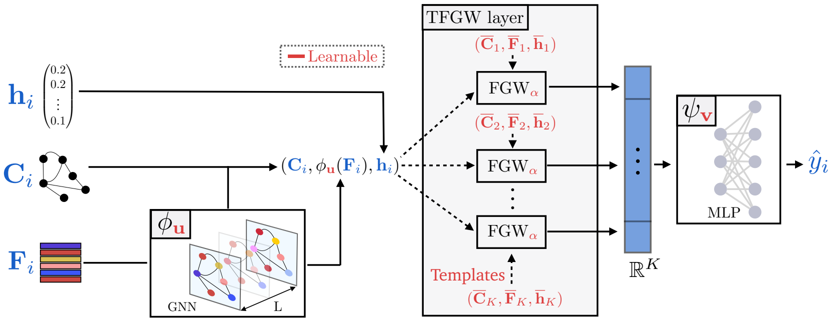
Barycenter/averaging of labeled graphs Vayer et al. 2018



Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022



FGW for a pooling layer in GNN



Template based FGW layer (TFGW) Vincent-Cuaz et al. 2022

- Principle: represent a graph through its distances to learned templates.
- Learnable parameters are illustrated in red above.
- New end-to-end GNN models for graph-level tasks.
- Sate-of-the-art (still!) on graph classification ($1\times\#1$, $3\times\#2$ on paperwithcode).

References I

Cuturi, Marco (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *NeurIPS*, pp. 2292–2300.

Frank, Marguerite and Philip Wolfe (1956). "An algorithm for quadratic programming". In: *Naval research logistics quarterly* 3.1-2, pp. 95–110.

Kantorovich, L. (1942). "On the translocation of masses". In: *C.R. (Doklady) Acad. Sci. URSS (N.S.)* 37, pp. 199–201.

Memoli, F. (2011). "Gromov Wasserstein Distances and the Metric Approach to Object Matching". In: *Foundations of Computational Mathematics*, pp. 1–71. ISSN: 1615-3375.

Monge, Gaspard (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

Peyré, Gabriel, Marco Cuturi, and Justin Solomon (2016). "Gromov-Wasserstein averaging of kernel and distance matrices". In: *ICML*, pp. 2664–2672.

Rubner, Yossi, Carlo Tomasi, and Leonidas J Guibas (2000). "The earth mover's distance as a metric for image retrieval". In: *International journal of computer vision* 40.2, pp. 99–121.

References II

Solomon, Justin et al. (2016). "Entropic metric alignment for correspondence problems". In: *ACM Transactions on Graphics (TOG)* 35.4, p. 72.

Thual, Alexis et al. (2022). "Aligning individual brains with Fused Unbalanced Gromov-Wasserstein". In: *Neural Information Processing Systems (NeurIPS)*.

Vayer, Titouan et al. (2018). "Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties". In:

Vincent-Cuaz, Cédric et al. (2022). "Template based Graph Neural Network with Optimal Transport Distances". In: *Neural Information Processing Systems (NeurIPS)*.