Computational Optimal Transport for Machine and Deep Learning

The Gromov-Wasserstein problem

Mathurin Massias, Titouan Vayer, Quentin Bertrand.

December 19, 2024

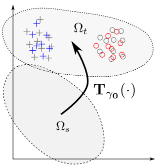
ENS DE LYON

Table of contents

The Gromov-Wasserstein distance

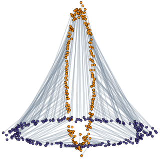
Applications

Three aspects of optimal transport



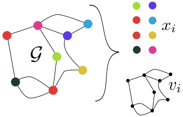
Transporting with optimal transport

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



Divergence between histograms

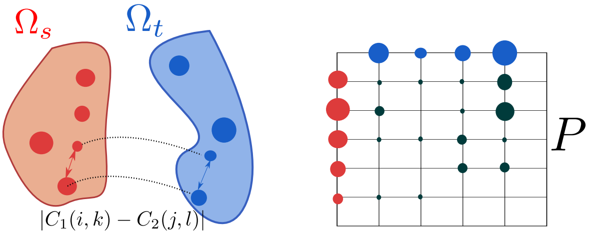
- Use the ground metric to encode complex relations between the bins of histograms for data fitting.
- OT losses are non-parametric divergences between non overlapping distributions.
- Used to train minimal Wasserstein estimators.



Divergence between graphs

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

Gromov-Wasserstein and extensions



Inspired from Gabriel Peyré

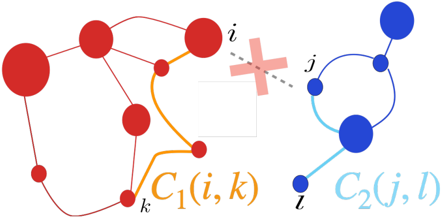
GW for discrete distributions Memoli 2011

$$\mathcal{GW}_p^p(\alpha, \beta) = \min_{P \in U(a,b)} \sum_{i,j,k,l} |C_1(i,k) - C_2(j,l)|^p P_{i,j} P_{k,l}$$

with $\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=}^{m} b_j \delta_{y_j}$

- $\mathbf{x}_i \in \Omega_s, \mathbf{y}_j \in \Omega_t$ with $\Omega_s \neq \Omega_t$.
- Distance between measures on different spaces w.r.t. isomorphism.
- OT plan that preserves the pairwise relationships between samples.
- Entropy regularized GW proposed in Peyré, Cuturi, and Solomon 2016.

Gromov-Wasserstein and extensions



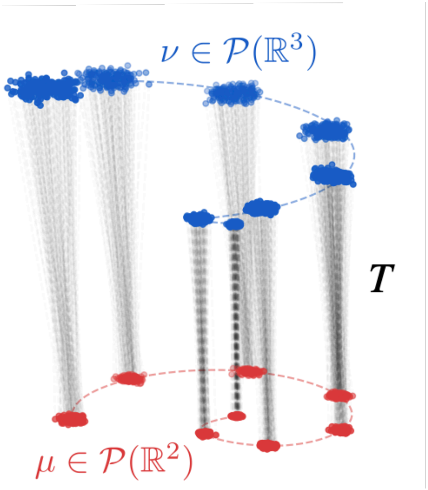
GW for discrete distributions

$$\mathcal{GW}_p^p(\alpha,\beta) = \min_{P \in U(a,b)} \sum_{i,j,k,l} |C_1(i,k) - C_2(j,l)|^p P_{i,j} P_{k,l}$$

with $\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=}^{m} b_j \delta_{y_j}$

- $\mathbf{x}_i \in \Omega_s, \mathbf{y}_j \in \Omega_t$ with $\Omega_s \neq \Omega_t$.
- Distance between measures on different spaces w.r.t. isomorphism.
- OT plan that preserves the pairwise relationships between samples.
- Entropy regularized GW proposed in Peyré, Cuturi, and Solomon 2016.

Examples



Solving the Gromov Wasserstein optimization problem

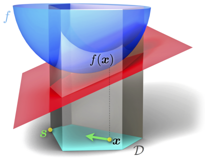
Optimization problem

$$\mathcal{GW}_p^p(\alpha, \beta) = \min_{P \in U(a,b)} \sum_{i,j,k,l} |C_1(i,k) - C_2(j,l)|^p P_{i,j} P_{k,l}$$

- 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.



The Frank-Wolfe algorithm

Solving a constrained problem

$$\min_{x \in C} \; f(x)$$

- C is convex, f is differentiable.
- Starts with $x_0 \in C$ and for $k \geq 0$ iterates

$$s_k \leftarrow \arg\min_{s \in C} \langle \nabla f(x_k), s \rangle \quad \text{(LMO step)}$$
$$x_{k+1} \leftarrow (1 - \gamma_k) x_k + \gamma_k s_k$$

Convergence guaranties

If f is convex and $x^\star$ is a minimizer then

$$f(x_k) - f(x^\star) \leq \frac{2}{k+2} M \,,$$

where $M = \sup_{\substack{\gamma \in [0,1] \\ x, s \in C}} f((1-\gamma)x + \gamma s) - f(x) - \gamma \langle \nabla f(x), s - x \rangle$.

The Frank-Wolfe algorithm for GW

Finding a local solution to the GW problem

$$\min_{P \in U(a,b)} \sum_{i,j,k,l} |C_1(i,k) - C_2(j,l)|^p P_{i,j} P_{k,l} = \langle L(C_1, C_2) \otimes P, P \rangle = f(P)$$

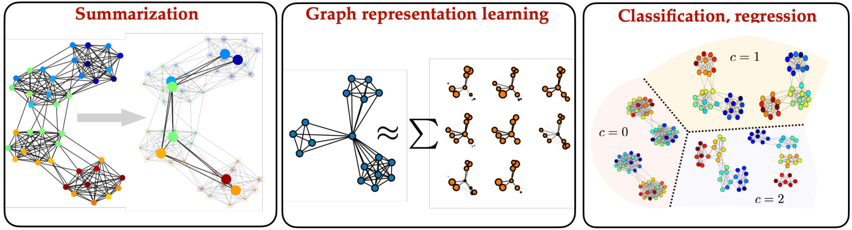
- $U(a, b)$ is convex, f is differentiable.
- $4D$-tensor $L(C_1, C_2) = (|C_1(i,k) - C_2(j,l)|^p)_{ijkl}$.
- If L is a tensor $L \otimes P = (\sum_{kl} L_{ijkl} P_{kl})_{ij}$.
- Starts with $P_0 \in U(a, b)$ and for $k \geq 0$ iterates

$$G_k = 2L(C_1, C_2) \otimes P_k \text{ (gradient of loss)}$$
$$S_k \leftarrow \arg\min_{S \in U(a,b)} \langle G_k, S \rangle \text{ (Linear OT problem)}$$
$$P_{k+1} \leftarrow (1 - \gamma_k) P_k + \gamma_k S_k$$

- Can be computed in $O(n^2 m + m^2 n)$ with $p = 2$.

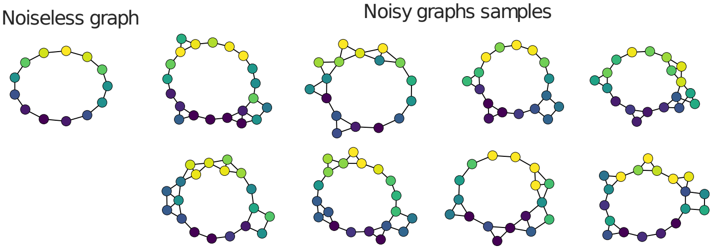
Applications of (F)GW

A tool for graphs

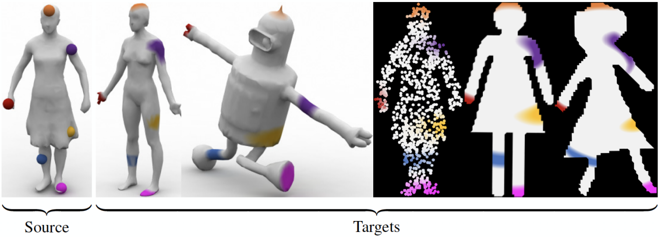


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

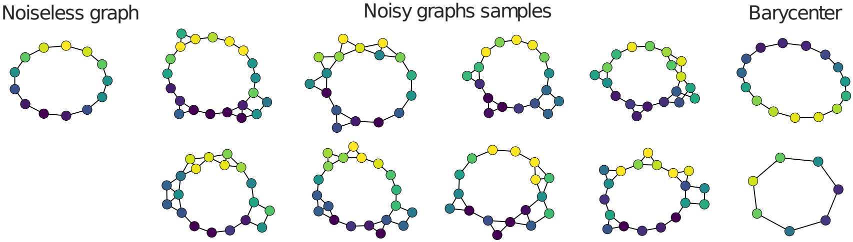


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

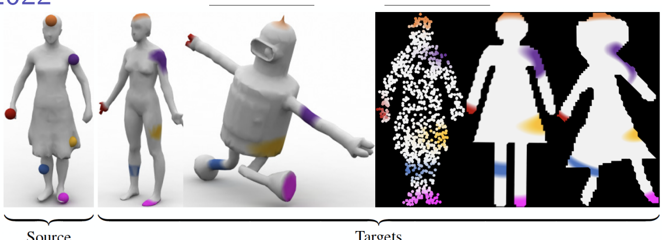


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

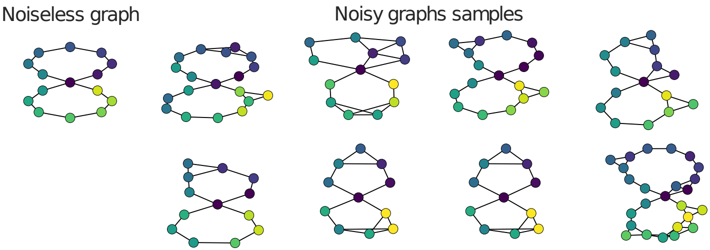


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

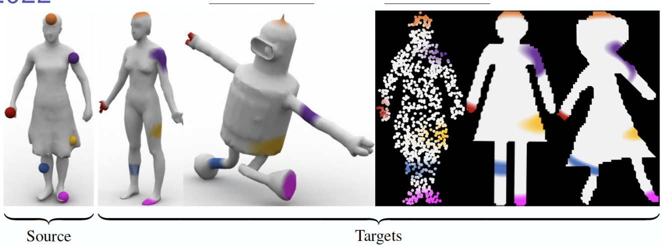


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

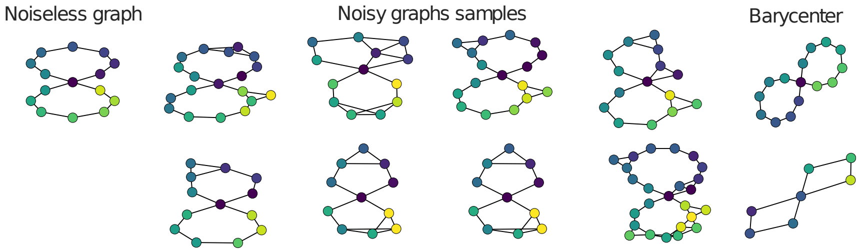


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

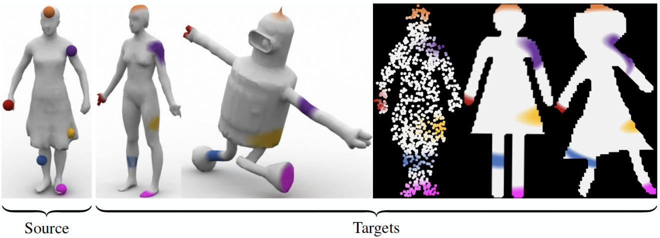


Applications of (F)GW

Barycenter/averaging of labeled graphs Vayer et al. 2018

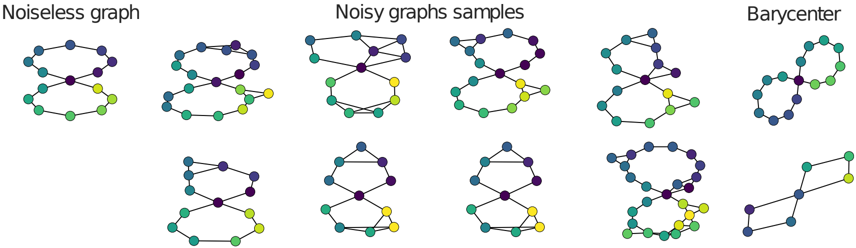


Shape matching between surfaces Solomon et al. 2016; Thual et al. 2022

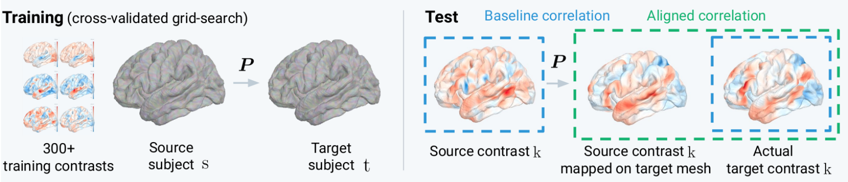


Applications of (F)GW

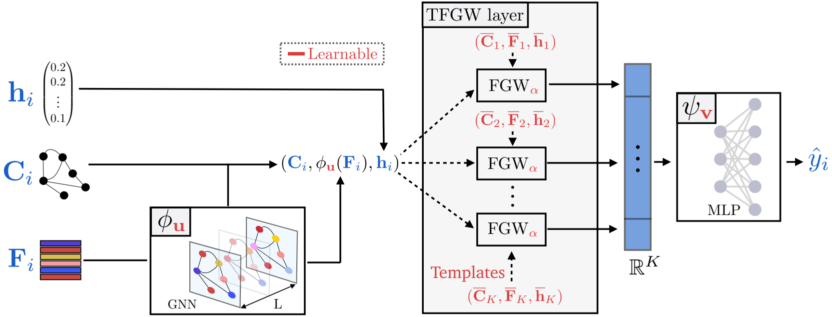
Barycenter/averaging of labeled graphs Vayer et al. 2018



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×#1, 3×#2 on paperwithcode).

References I

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

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

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

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.

References II

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