Inference with Gromov-Wasserstein Distances

Kengo Kato (Cornell)

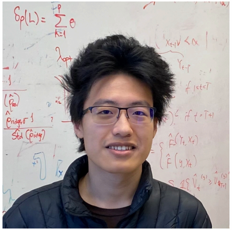
Cornell Bowers Statistics + Data Science

September 18, 2025

Collaborators

Ziv Goldfeld

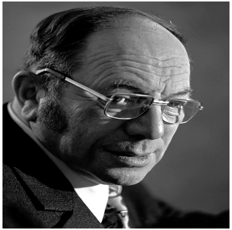
Gabriel Rioux



Boyu Wang

Optimal Transport

Monge



Kantorovich

Villani

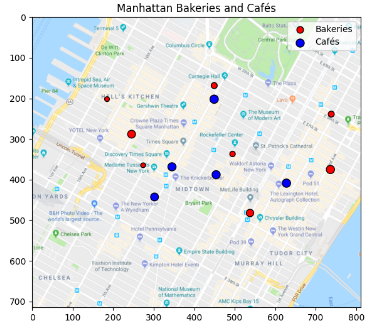
Monge Problem

- Given marginals $\color{red}{\mu}, \color{blue}{\nu} \in \mathcal{P}(\mathbb{R}^d)$ & cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$

Monge Problem (1781)

Find a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that solves

$$\inf_{T:\, \mu \circ T^{-1} = \nu} \mathbb{E}_{X \sim \mu}[c(X, T(X))]$$



Monge Problem

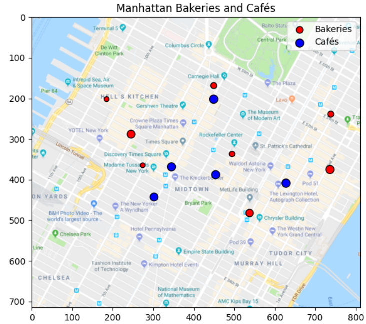
- Given marginals $\color{red}{\mu}, \color{blue}{\nu} \in \mathcal{P}(\mathbb{R}^d)$ & cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$

Monge Problem (1781)

Find a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that solves

$$\inf_{T:\, \mu \circ T^{-1} = \nu} \mathbb{E}_{X \sim \mu}[c(X, T(X))]$$

$\{T : \mu \circ T^{-1} = \nu\}$ may be empty...



Monge & Kantorovich Problems

Monge Problem (1781)

Find a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that solves

$$\inf_{T:\,\mu\circ T^{-1}=\nu} \mathbb{E}_{X\sim\mu}[c(X,T(X))]$$

$\{T : \mu \circ T^{-1} = \nu\}$ may be empty...

Kantorovich Problem (1930)

Find a coupling $\pi \in \mathcal{P}(\mathbb{R}^{2d})$ that solves

$$\underbrace{\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}[c(X,Y)]}_{\text{OT cost}}$$

Well-posed for lsc cost

- ✓ $\Pi(\mu,\nu) \neq \varnothing$
- ✓ $\inf = \min$

Wasserstein Distances

- OT cost induces a metric on the space of probability measures

Wasserstein Distance

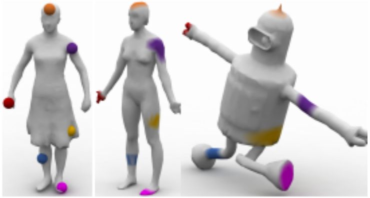
$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\pi}[\|X - Y\|^p]\right)^{1/p}, \; p \in [1, \infty)$$

- Applications: parametric estimation (W-GAN) and testing problems

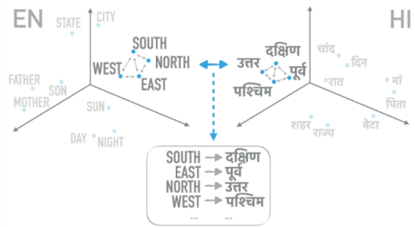
  Panaretos-Zemel '21; Chewi et al. '24

Beyond OT

- Various applications require matching heterogeneous & structured datasets



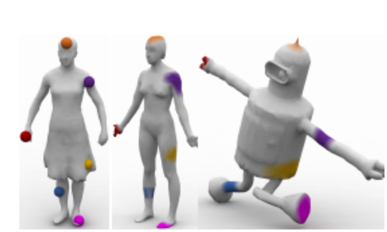
Solomon-Peyré-Kim-Sra '16



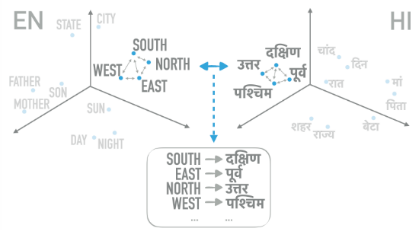
Alvarez-Melis-Jaakkola '18

Beyond OT

- Various applications require matching heterogeneous & structured datasets



Solomon-Peyré-Kim-Sra '16



Alvarez-Melis-Jaakkola '18

- Goals:
  1. Compare similarity btw two data sets;
  2. Obtain matching or alignment

Gromov-Wasserstein Distance

LIMIT LAWS FOR GROMOV-WASSERSTEIN ALIGNMENT WITH APPLICATIONS TO TESTING GRAPH ISOMORPHISMS

GABRIEL RIOUX, ZIV GOLDFELD, AND KENGO KATO

Abstract. The Gromov-Wasserstein (GW) distance enables comparing metric measure spaces based solely on their internal structure, making it invariant to isomorphic transformations. This property is particularly useful for comparing datasets that naturally admit isomorphic representations, such as unlabelled graphs or objects embedded in space. However, apart from the recently derived empirical convergence rates for the quadratic GW problem, a statistical theory for valid estimation and inference remains largely obscure. Pushing the frontier of statistical GW further, this work derives the first limit laws for the empirical GW distance across several settings of interest: (i) discrete, (ii) semi-discrete, and (iii) general distributions under moment constraints under the entropically regularized GW distance. The derivations rely on a novel stability analysis of the GW functional in the marginal distributions. The limit laws then follow by an adaptation of the functional delta method. As asymptotic normality fails to hold in most cases, we establish the consistency of an efficient estimation procedure for the limiting law in the discrete case, bypassing the need for computationally intensive resampling methods. We apply these findings to testing whether collections of unlabelled graphs are generated from distributions that are isomorphic to each other.

arXiv:2410.18006

CONVERGENCE OF EMPIRICAL GROMOV-WASSERSTEIN DISTANCE

KENGO KATO AND BOYU WANG

Abstract. We study rates of convergence for estimation of the Gromov-Wasserstein (GW) distance. For two marginals supported on compact subsets of $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively, with $\min\{d_x, d_y\} > 4$, prior work established the rate $n^{-\frac{2}{\min\{d_x,d_y\}}}$ in L^1 for the plug-in empirical estimator based on n i.i.d. samples. We extend this fundamental result to marginals with unbounded supports, assuming only finite polynomial moments. Our proof techniques for the upper bounds can be adapted to obtain sample complexity results for penalized Wasserstein alignment that encompasses the GW distance and Wasserstein Procrustes. Furthermore, we establish matching minimax lower bounds (up to logarithmic factors) for estimating the GW distance. Finally, we establish deviation inequalities for the error of empirical GW in cases where two marginals have compact supports, exponential tails, or finite polynomial moments. The deviation inequalities yield that the same rate $n^{-\frac{2}{\min\{d_x,d_y\}}}$ holds for empirical GW also with high probability.

arXiv:2508.03985

Gromov-Wasserstein Distance

- Given metric measure spaces $(\mathcal{X}, d_\mathcal{X}, \mu)$ & $(\mathcal{Y}, d_\mathcal{Y}, \nu)$

Gromov-Wasserstein Distance (Memoli '11; Sturm '12)

For $p, q \in [1, \infty)$,

$$\mathsf{GW}_{p,q}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{\substack{(X,Y)\sim\pi \\ (X',Y')\sim\pi}} \left[\left| d_\mathcal{X}^q(X, X') - d_\mathcal{Y}^q(Y, Y') \right|^p \right] \right)^{1/p}$$

Gromov-Wasserstein Distance

- Given metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ & $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$

Gromov-Wasserstein Distance (Memoli '11; Sturm '12)

For $p, q \in [1, \infty)$,

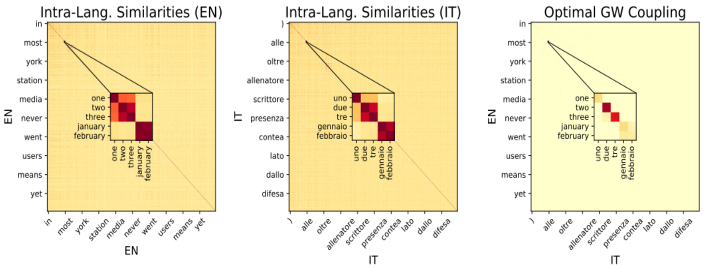
$$\mathsf{GW}_{p,q}(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{\substack{(X,Y)\sim\pi \\ (X',Y')\sim\pi}} \left[\left| d^q_{\mathcal{X}}(X,X') - d^q_{\mathcal{Y}}(Y,Y') \right|^p \right] \right)^{1/p}$$

- $\mathsf{GW}_{p,q}$ defines a metric on the space of all Polish metric measure spaces modulo measure-preserving isometries:

$$\mathsf{GW}_{p,q}(\mu,\nu) = 0 \Leftrightarrow \exists T : \mathrm{spt}(\mu) \to \mathrm{spt}(\nu) \text{ isometry s.t. } \nu = T_{\#}\mu$$

Discrete Marginals

- Suppose $\mu = \sum_{i=1}^{N_1} \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{N_2} \nu_j \delta_{y_j}$
- GW objective: $\sum_{i,k \in [N_1]} \sum_{j,l \in [N_2]} C_{ijkl} \pi_{ik} \pi_{jl}$ with $C_{ijkl} = |d_{\mathcal{X}}^q(x_i, x_j) - d_{\mathcal{Y}}^q(y_i, y_j)|^p$
- Cost tensor C_{ijkl} relies on relational rather than positional similarities

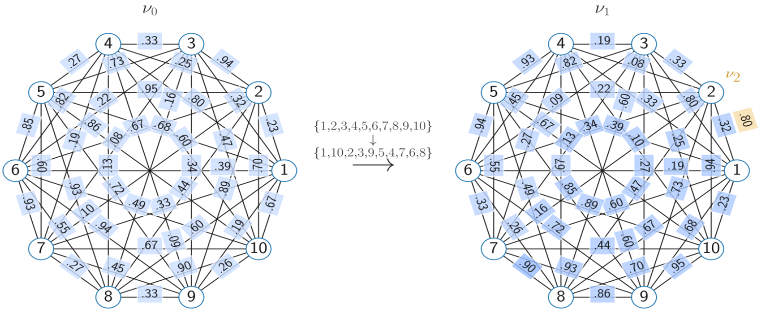


Alvarez-Melis-Jaakkola '18

Graph Isomorphism

- Consider a random weighted graph with N vertices with no self-loop
- Edges $\{i, j\}$ have independent weights $\sim \rho_{ij} \in \mathcal{P}([0, \infty))$ (zero-weight $\Leftrightarrow$ no edge)
- Identify a graph distribution ν with $(\rho_{ij})_{1 \le i < j \le N}$
- Call two graph distributions $\nu_0 \leftrightarrow (\rho_{0,ij})_{i<j}$ and $\nu_1 \leftrightarrow (\rho_{1,ij})_{i<j}$ isomorphic iff

$$\exists \sigma : [N] \to [N] \text{ permutation s.t. } \rho_{1,ij} = \rho_{0,\sigma(i)\sigma(j)}, \ \forall i < j.$$



Graph Isomorphism

- Consider a random weighted graph with N vertices with no self-loop
- Edges $\{i, j\}$ have independent weights $\sim \rho_{ij} \in \mathcal{P}([0, \infty))$ (zero-weight $\Leftrightarrow$ no edge)
- Identify a graph distribution ν with $(\rho_{ij})_{1 \le i < j \le N}$
- Call two graph distributions $\nu_0 \leftrightarrow (\rho_{0,ij})_{i<j}$ and $\nu_1 \leftrightarrow (\rho_{1,ij})_{i<j}$ isomorphic iff

$$\exists \sigma : [N] \to [N] \text{ permutation s.t. } \rho_{1,ij} = \rho_{0,\sigma(i)\sigma(j)}, \ \forall i < j.$$

Theorem (Rioux-Goldfeld-K '24)

There is a known embedding $\iota : \mathcal{P}([0, \infty))^{N(N-1)/2} \to \mathcal{P}(\mathbb{R}^N)$ *such that* ν_0 *and* ν_1 *are isomorphic iff* $\mathrm{GW}_{2,2}(\iota(\nu_0), \iota(\nu_1)) = 0$

Challenges

- GW objective

$$\pi \mapsto \underbrace{\iint_{\mathcal{X}\times\mathcal{Y}} \iint_{\mathcal{X}\times\mathcal{Y}} |d_{\mathcal{X}}^q(x,x') - d_{\mathcal{Y}}^q(y,y')|^p \, d\pi(x,y) d\pi(x',y')}_{=:F(\pi)}$$

  is bilinear and nonconvex
- Lack of a comprehensive duality theory for general GW

Challenges

- GW objective

$$\pi \mapsto \underbrace{\iint_{\mathcal{X}\times\mathcal{Y}} \iint_{\mathcal{X}\times\mathcal{Y}} |d_{\mathcal{X}}^q(x,x') - d_{\mathcal{Y}}^q(y,y')|^p \, d\pi(x,y) d\pi(x',y')}_{=:F(\pi)}$$

  is bilinear and nonconvex

- Lack of a comprehensive duality theory for general GW
- One exception to the latter: $\mathrm{GW}_{2,2}$ with Euclidean metric measures spaces:

$$(\mathcal{X}, d_{\mathcal{X}}) = (\mathbb{R}^{d_x}, \|\cdot\|) \qquad (\mathcal{Y}, d_{\mathcal{Y}}) = (\mathbb{R}^{d_y}, \|\cdot\|)$$

$\mathrm{GW}_{2,2}$ case

- Assume w.l.o.g. μ, ν have mean zero
- $\mathrm{GW}_{2,2}$ objective can be decomposed as

$$F(\pi) = \big(\text{something independent of } \pi\big) \\ - \underbrace{\left\{4 \int \|x\|^2 \|y\|^2 \, d\pi(x,y) + 8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2 \right\}}_{=: G(\pi)}$$

$\mathsf{GW}_{2,2}$ case

- Assume w.l.o.g. μ, ν have mean zero
- $\mathsf{GW}_{2,2}$ objective can be decomposed as

$$F(\pi) = \big(\text{something independent of } \pi\big)$$
$$\underbrace{- \left\{ 4 \int \|x\|^2 \|y\|^2 \, d\pi(x,y) + 8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2 \right\}}_{=:G(\pi)}$$

Variational Representation (Zhang et al. '24)

For $c_A(x,y) = -4\|x\|^2\|y\|^2 - 32x^T A y$,

$$\inf_{\pi \in \Pi(\mu,\nu)} -G(\pi) = \inf_{A \in \mathbb{R}^{d_x \times d_y}} \left\{ 32\|A\|_F^2 + \mathsf{OT}_{c_A}(\mu,\nu) \right\}$$

Proof I

$$\boxed{\inf_{\pi \in \Pi(\mu,\nu)} -G(\pi) = \inf_{A \in \mathbb{R}^{d_x \times d_y}} \left\{ 32\|A\|_F^2 + \mathsf{OT}_{c_A}(\mu,\nu) \right\}}$$

Recall:

$$G(\pi) = 4 \underbrace{\int \|x\|^2 \|y\|^2 \, d\pi(x,y)}_{\text{linear in } \pi} + 8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2$$

Proof (one line):

$$-b^2 \leq 4a^2 - 4ab \quad \text{equality holds iff } a = \frac{1}{2}b$$

Proof I

$$\boxed{\inf_{\pi \in \Pi(\mu,\nu)} -G(\pi) = \inf_{A \in \mathbb{R}^{d_x \times d_y}} \left\{ 32\|A\|_F^2 + \mathsf{OT}_{c_A}(\mu,\nu) \right\}}$$

Recall:

$$G(\pi) = 4 \underbrace{\int \|x\|^2 \|y\|^2 \, d\pi(x,y)}_{\text{linear in } \pi} + 8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2$$

Proof (one line):

$$-b^2 \leq 4a^2 - 4ab \quad \text{equality holds iff } a = \frac{1}{2}b$$

$$\Longrightarrow \quad -8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2 = \inf_A \left\{ 32\|A\|_F^2 - 32 \underbrace{\left\langle A, \int xy^T \, d\pi(x,y) \right\rangle_F}_{\text{linear in } \pi} \right\}$$

$$\text{and} \quad \inf_\pi \inf_A (\cdots) = \inf_A \inf_\pi (\cdots)$$

Proof II

$$\boxed{\inf_{\pi \in \Pi(\mu,\nu)} -G(\pi) = \inf_{A \in \mathbb{R}^{d_x \times d_y}} \left\{ 32\|A\|_F^2 + \mathsf{OT}_{c_A}(\mu,\nu) \right\}}$$

Recall:

$$G(\pi) = 4 \underbrace{\int \|x\|^2 \|y\|^2 \, d\pi(x,y)}_{\text{linear in } \pi} + 8 \left\| \int xy^T \, d\pi(x,y) \right\|_F^2$$

Observation (assume μ, ν are compactly supported):

$G(\pi)$ is convex and lsc wrt weak topology

$\Longrightarrow$ Bipolar theorem :

$$G(\pi) = \sup_h \left\{ \underbrace{\int h \, d\pi}_{\text{linear in } \pi} - G^*(h) \right\}$$

Estimation of GW Cost

- Variational representation paves a way for establishing formal guarantees for computation and estimation of GW cost
- Empirical distributions based on i.i.d. data:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \qquad \text{and} \qquad \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m}\delta_{Y_j}$$

Estimation of GW Cost

- Variational representation paves a way for establishing formal guarantees for computation and estimation of GW cost
- Empirical distributions based on i.i.d. data:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \qquad \text{and} \qquad \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m}\delta_{Y_j}$$

Sample complexity (Zhang et al. '24; K-Wang '25)

For $d_x \wedge d_y \geq 4$,

$$\mathbb{E}\left[\left|\mathsf{GW}_{2,2}^2(\widehat{\mu}_n,\widehat{\nu}_m) - \mathsf{GW}_{2,2}^2(\mu,\nu)\right|\right] \lesssim (n\wedge m)^{-\frac{2}{d_x\wedge d_y}}\left(\log(n\wedge m)\right)^{1(d_x\wedge d_y=4)}$$

provided that μ,ν have finite $(4q)$-th moments with

$$q > \frac{d_x\wedge d_y + 2d_xd_y}{d_x\wedge d_y - 2}$$

Minimax lower bounds

Set $D = \mathsf{GW}_{2,2}^2$

Minimax lower bounds (K-Wang '25)

For $d_x \wedge d_y \geq 4$,

$$\inf_{\widehat{D}_{n,m}} \sup_{(\mu,\nu)\in\mathcal{P}(B(0,1))\times\mathcal{P}(B(0,1))} \mathbb{E}\left[\left|\widehat{D}_{n,m} - D(\mu,\nu)\right|\right] \gtrsim (n \wedge m)^{-\frac{2}{d_x \wedge d_y}} (\log(n \wedge m))^{-\frac{2}{d_x \wedge d_y}\mathbb{1}_{\{d_x \wedge d_y > 4\}}}$$

where the infimum is taken over all estimators of $D(\mu,\nu)$ constructed from i.i.d. samples from μ and ν of sizes n and m, respectively

$$\implies (n \wedge m)^{-\frac{2}{d_x \wedge d_y}} \text{ is sharp (up to a log factor)}$$

Semidiscrete case

Consider when μ is general but ν is finitely discrete:

$$\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$$

Semidsicrete case (K-Wang '25)

If μ has a finite $(2q)$-moment with $q \in (2, \infty)$,

$$\mathbb{E}\left[\left|\mathsf{GW}_{2,2}^2(\widehat{\mu}_n, \widehat{\nu}_m) - \mathsf{GW}_{2,2}^2(\mu, \nu)\right|\right] \lesssim (n \wedge m)^{-1/2} + n^{\frac{2-q}{q}}$$

- Parametric rate if μ has a finite 8-th moment
- $n^{\frac{2-q}{q}}$ is dominant if μ is heavy tailed

Limiting Law for Discrete GW

Statistical inference for GW?

Limiting law (Rioux-Goldfeld-K '24)

For finitely discrete μ and ν, there exist Gaussian processes G_μ and G_ν such that

$$\begin{aligned}&\sqrt{n}\big(\mathsf{GW}_{2,2}^2(\widehat{\mu}_n,\widehat{\nu}_n)-\mathsf{GW}_{2,2}^2(\mu,\nu)\big)\\ &\quad\xrightarrow{d}\inf_{A\in\mathcal{A}}\inf_{\pi\in\Pi_A}\sup_{(\varphi,\psi)\in\bar{\mathcal{D}}_A}G_\mu(f+g_\pi+\bar{\varphi})+G_\nu(f'+g'_\pi+\bar{\psi})\end{aligned}$$

Furthermore, when $\mathsf{GW}_{2,2}(\mu,\nu)=0$,

$$\sqrt{n}\mathsf{GW}_{2,2}^2(\widehat{\mu}_n,\widehat{\nu}_n)\xrightarrow{d}\sqrt{2}\|G_\mu\|_{\infty,\bar{\mathcal{H}}}$$

Proof Idea

- Identify μ & ν with finite dimensional vectors and establish the derivative of $(\mu,\nu) \mapsto \mathsf{GW}_{2,2}^2(\mu,\nu)$ using the variational representation:

$$\mathsf{GW}_{2,2}^2(\mu,\nu) = \underbrace{\mathsf{S}_1(\mu,\nu)}_{\substack{\text{depends only on}\\ \text{moments of marginals}}} + \inf_{A \in \mathbb{R}^{d_x \times d_y}} \left\{ 32\|A\|_F^2 + \mathsf{OT}_{c_A}(\mu,\nu) \right\},$$

where $c_A(x,y) = -4\|x\|^2\|y\|^2 - 32x^T A y$

- By duality,

$$\mathrm{OT}_{c_A}(\mu,\nu) = \sup_{\varphi_i + \psi_j \le c_A(x_i,y_j)} \varphi^T \mu + \psi^T \nu$$

- One needs to find the derivative of the map

$$(\mu, \nu) \mapsto \inf_{A \in \mathbb{R}^{dx \times dy}} \sup_{\varphi_i + \psi_j \leq c_A(x_i, y_j)} 32\|A\|_F^2 + \varphi^T \mu + \psi^T \nu$$

- One needs to find the derivative of the map

$$(\mu,\nu) \mapsto \inf_{A \in \mathbb{R}^{d_x \times d_y}} \sup_{\varphi_i + \psi_j \leq c_A(x_i, y_j)} 32\|A\|_F^2 + \varphi^T \mu + \psi^T \nu$$

- The constraint set for the inner maximization depends on A!

- One needs to find the derivative of the map

$$(\mu, \nu) \mapsto \inf_{A \in \mathbb{R}^{d_x \times d_y}} \sup_{\varphi_i + \psi_j \leq c_A(x_i, y_j)} 32\|A\|_F^2 + \varphi^T \mu + \psi^T \nu$$

- The constraint set for the inner maximization depends on A!
- It turns out that the said map is not (Frechét) differentiable but directionally Hadamard differentiable with nonlinear derivative, which is enough to invoke the extended Delta method

- One needs to find the derivative of the map

$$(\mu, \nu) \mapsto \inf_{A \in \mathbb{R}^{d_x \times d_y}} \sup_{\varphi_i + \psi_j \leq c_A(x_i, y_j)} 32\|A\|_F^2 + \varphi^T \mu + \psi^T \nu$$

- The constraint set for the inner maximization depends on A!
- It turns out that the said map is not (Frechét) differentiable but directionally Hadamard differentiable with nonlinear derivative, which is enough to invoke the extended Delta method
- Used several techniques from optimization

Estimating Weak Limits

- Naive bootstrap is inconsistent

  Dümbgen '91; Fang-Santos '19
- Subsampling (m-out-of-n bootstrap with $m \ll n$) can be adapted to estimate the weak limit, but it is computationally intensive

Estimating Weak Limits

- Naive bootstrap is inconsistent

  Dümbgen '91; Fang-Santos '19
- Subsampling (m-out-of-n bootstrap with $m \ll n$) can be adapted to estimate the weak limit, but it is computationally intensive
- When $\mathrm{GW}_{2,2}(\mu, \nu) = 0$, the weak limit simplifies, and one can directly estimate it:

Algorithm 1 Sampling from L_n

Given samples $X_1, \dots, X_n$ from μ_0, construct $\hat{\mu}_{0,n}$,

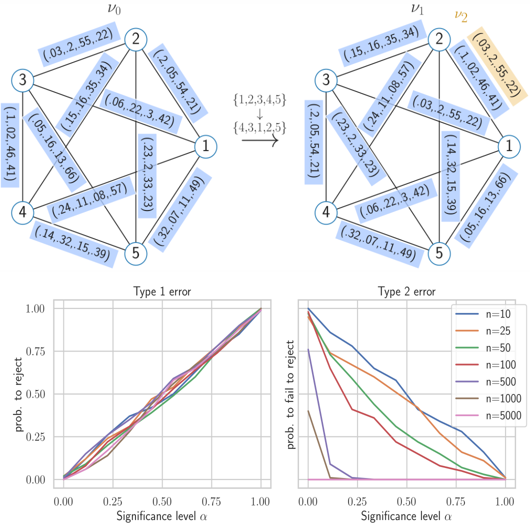
1: Let $(\bar{x}^{(i)})_{i=1}^{N_n} = \mathrm{spt}(\bar{\mu}_{0,n})$.

2: Let $b_n \in \mathbb{R}^{N_n(N_n-1)}$ and $\mathbf{A} \in \mathbb{R}^{N_n(N_n-1)\times N_n}$ be such that $\mathcal{H}_n = \{u \in \mathbb{R}^{N_n} : \mathbf{A}u \le b_n\}$.

3: Sample $Z_n \sim N(0, \mathbf{\Sigma}_{m_n})$ and solve $\sqrt{2}\sup_{u\in\mathcal{H}_n} Z_n^\intercal u \sim L_n$; repeat Step 3 to generate new samples.

- This method is consistent for estimating the null weak limit

Simulation Results



Summary

Gromov-Wasserstein Distance:

- Sample complexity upper bounds and minimax lower bounds for GW
- Limiting distributions for empirical GW
- Develop a method to estimate the null weak limit
- Other results:
  - Concentration inequalities
  - Weak limits for semidiscrete GW and entropic GW

Thank You!