Optimal Transport

A Statistical Perspective — From First Principles to the Frontier

Yuyao Wang
Department of Mathematics and Statistics

yuyaow@bu.edu

Outline

Motivation

A central question in statistics: how do we *compare* two probability distributions?

Classical tools you already know:

- Kullback–Leibler divergence: $\mathrm{KL}(\mu\|\nu) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu$
- Total variation: $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$
- Hellinger, χ^2, Jensen–Shannon, ...

A shared weakness:

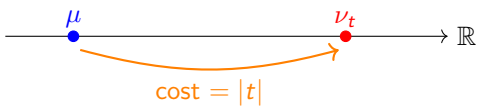
- They require μ and ν to share support (or $\mu \ll \nu$).
- They are *blind to geometry*: moving a Dirac mass by 1 m or by 1000 m yields $\mathrm{KL} = +\infty$ either way.
- In high dimensions, with finite samples, they are often unstable or undefined.

The promise of OT: a *geometry-aware* distance that uses the metric structure of $\mathbb{R}^d$ to compare distributions.

A toy example

Let $\mu = \delta_0$ and $\nu_t = \delta_t$ for $t \in \mathbb{R}$.

Divergence	Value as a function of t	Geometry aware?
$\mathrm{KL}(\mu \| \nu_t)$	$+\infty$ for all $t \neq 0$	✗
$\mathrm{TV}(\mu, \nu_t)$	1 for all $t \neq 0$	✗
$W_p(\mu, \nu_t)$	$\lvert t \rvert$ (Wasserstein-p)	✓



$\Rightarrow$ the Wasserstein distance gives the answer we *want* it to give.

A very short history

- **1781 – Gaspard Monge.** *Mémoire sur la théorie des déblais et des remblais*: what is the cheapest way to move a pile of soil into a target shape? (A military-engineering question.)
- **1940s – Leonid Kantorovich.** A linear-programming relaxation of Monge's problem; Nobel Prize in Economics, 1975.
- **1990s–2000s.** The modern mathematical theory: Brenier, McCann, Villani, Otto, Ambrosio, . . . (Villani awarded the Fields Medal in 2010.)
- **2013 – Cuturi.** Entropic regularization and the Sinkhorn algorithm make OT computationally tractable at ML scale.
- **Today.** WGANs, domain adaptation, single-cell RNA-seq alignment, distributionally robust optimization, fairness, causal inference, . . .

Monge's original formulation

Let μ, ν be probability measures on $\mathbb{R}^d$, and let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a cost function.

Monge problem (1781)

Find a *transport map* $T : \mathbb{R}^d \to \mathbb{R}^d$ minimising

$$\inf_{T:\, T_\sharp \mu = \nu} \int_{\mathbb{R}^d} c\big(x, T(x)\big)\, \mathrm{d}\mu(x),$$

where $T_\sharp \mu(B) := \mu(T^{-1}(B))$ is the *pushforward measure*.

Interpretation:

- Each source point x must be sent *as a whole* to $T(x)$ — no splitting of mass allowed.
- The constraint $T_\sharp \mu = \nu$ forces T to redistribute μ into exactly ν.

Two difficulties:

1. If $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, *no* map T can split an atom $\Rightarrow$ the problem is infeasible.
2. Even when feasible, existence and uniqueness of T are very hard (Brenier, 1987).

Brenier's theorem — existence of the map

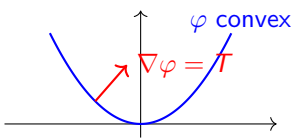
Brenier (1987, 1991)

Let $c(x, y) = \frac{1}{2}\|x - y\|^2$. Assume μ is absolutely continuous with respect to Lebesgue measure and that μ, ν have finite second moments. Then the Monge problem admits a unique solution of the form

$$T(x) = \nabla\varphi(x), \qquad \varphi : \mathbb{R}^d \to \mathbb{R} \text{ convex.}$$

Statistical implications:

- For quadratic cost, the optimal transport map is the *gradient of a convex potential*.
- This generalises monotone rearrangement on $\mathbb{R}$ to $\mathbb{R}^d$.
- It underlies *vector quantiles* and *center-outward quantiles* (Chernozhukov, Galichon, Hallin, ...).



Kantorovich's key idea

Core insight. Allow mass to *split*. Instead of looking for a map, look for a *joint distribution*.

Let $\Pi(\mu, \nu)$ denote the set of *couplings*:

$$\Pi(\mu,\nu) = \big\{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \ \pi(A \times \mathbb{R}^d) = \mu(A),\ \pi(\mathbb{R}^d \times B) = \nu(B)\big\}.$$

Kantorovich problem (1942)

$$\mathrm{OT}_c(\mu,\nu) \ = \ \inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\,\mathrm{d}\pi(x,y).$$

Why is this fundamental?

- $\Pi(\mu, \nu)$ is *never empty*: the product coupling $\mu \otimes \nu$ is always feasible.
- It is a convex problem: linear objective, convex feasible set.
- The relaxation is *tight*: if a Monge solution T^* exists, then $\pi^* = (\mathrm{id} \times T^*)_\sharp \mu$ solves the Kantorovich problem.

Discrete OT — a linear program

Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, with a, b probability vectors.
A coupling π is encoded by a matrix $P \in \mathbb{R}_+^{n \times m}$:

$$\begin{aligned} \min_{P \in \mathbb{R}_+^{n \times m}} \quad & \sum_{i,j} C_{ij} P_{ij} \\ \text{s.t.} \quad & P \mathbf{1}_m = a, \\ & P^\top \mathbf{1}_n = b, \end{aligned}$$

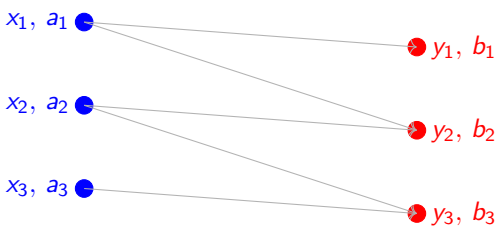
where $C_{ij} = c(x_i, y_j)$.

A standard linear program.

- nm variables, $n + m$ equality constraints.
- Simplex or network-flow methods: $O(n^3 \log n)$ (Hungarian-type algorithms).
- When $n = m$ and $a_i = b_j = 1/n$, this reduces to the *optimal assignment problem*.

A picture is worth a thousand words

$P_{ij} \geq 0$ is the mass sent from x_i to y_j



Row sums $= a$, column sums $= b$

- **Classical transportation problem**: factories $\rightarrow$ warehouses.
- **Statistical reinterpretation**: a is an empirical distribution, b is a target or reference.

The Wasserstein distance — definition

Specialize to cost $c(x, y) = \|x - y\|^p$, for some $p \geq 1$.

p-Wasserstein distance

$$W_p(\mu, \nu) \;=\; \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p}.$$

Defined on $\mathcal{P}_p(\mathbb{R}^d) = \{\mu : \int \|x\|^p \, \mathrm{d}\mu < \infty\}$.

Proposition (Metric properties)

W_p is a metric on $\mathcal{P}_p(\mathbb{R}^d)$:

1. $W_p(\mu, \nu) \geq 0$, *with* $W_p(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$;
2. *symmetry:* $W_p(\mu, \nu) = W_p(\nu, \mu)$;
3. *triangle inequality (via the gluing lemma).*

$\Rightarrow$ a genuine distance, not merely a divergence — so we can do *geometry* on the space of distributions.

Closed form in one dimension — a rare gift

Theorem (1D case)

Let μ, ν be probability measures on $\mathbb{R}$ with CDFs F_μ, F_ν. Then

$$W_p(\mu,\nu) \;=\; \left(\int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^p \,\mathrm{d}u\right)^{1/p}.$$

A beautifully clean interpretation: the optimal coupling is obtained by *matching quantiles with quantiles*.

In particular:

- $p = 1$: $W_1(\mu,\nu) = \displaystyle\int_{-\infty}^{\infty} |F_\mu(t) - F_\nu(t)| \,\mathrm{d}t$
  i.e. the *area between the two CDFs* — a quantity every statistician has met.
- **Empirical distributions:** $W_p(\hat{\mu}_n, \hat{\nu}_n) = \left(\frac{1}{n}\sum_{i=1}^n \left|x_{(i)} - y_{(i)}\right|^p\right)^{1/p}$, i.e. simply *sort and subtract*.

The Gaussian case — a second gift

Theorem (closed form for Gaussians)

Let $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$, with $m_i \in \mathbb{R}^d$. Then

$$W_2^2(\mu, \nu) \;=\; \|m_1 - m_2\|^2 + \operatorname{tr}\Big(\Sigma_1 + \Sigma_2 - 2\big(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\big)^{1/2}\Big).$$

Why this matters in statistics:

- **FID (Fréchet Inception Distance):** the standard evaluation metric for generative models is exactly W_2 between two *Gaussian approximations*.
- It is a continuous generalization of covariance matching / Procrustes alignment.
- When $\Sigma_1 = \Sigma_2$ it reduces to $\|m_1 - m_2\|$.
- When $d = 1$ it reduces to $(m_1 - m_2)^2 + (\sigma_1 - \sigma_2)^2$ — mean *and* scale.

The Kantorovich duality

Kantorovich duality

$$\mathrm{OT}_c(\mu,\nu) \;=\; \sup_{\substack{\varphi\in L^1(\mu),\,\psi\in L^1(\nu)\\ \varphi(x)+\psi(y)\le c(x,y)}} \int \varphi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu.$$

Economic interpretation (the shipper's problem):

- Imagine outsourcing to a shipper who charges $\varphi(x)$ at source x and pays $\psi(y)$ at destination y (or vice versa).
- The constraint $\varphi(x)+\psi(y) \le c(x,y)$ says that the shipper cannot charge more than you could pay yourself.
- The shipper maximizes profit $\Rightarrow$ the optimal profit equals the minimum total transport cost.

Why does this matter?

- Turns a *combinatorial/probabilistic* problem into one over *two functions* — they can be parameterised by neural networks (the basis of WGAN).
- The dual potentials φ, ψ have natural statistical interpretations (e.g. vector quantile maps).

Kantorovich–Rubinstein form ($p = 1$)

When $c(x, y) = \|x - y\|$, the dual collapses to a strikingly clean form.

KR duality

$$W_1(\mu, \nu) \;=\; \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{ \int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu \right\},$$

where $\|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|}$.

Statisticians should immediately recognise:

- This is an *integral probability metric* (IPM) over the *1-Lipschitz* test class.
- MMD is the same object but with a different test class (RKHS unit ball).
- It suggests a *Lipschitz neural network* as an estimator of W_1 — exactly what Arjovsky et al. (WGAN, 2017) did.

Exercise: use KR to show that $W_1(\delta_x, \delta_y) = \|x - y\|$ by taking $f(z) = \|z - y\|$.

The empirical Wasserstein distance

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mu$ with empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.
Basic question: at what rate does $W_p(\hat{\mu}_n, \mu) \to 0$?

Classical rates (Fournier–Guillin 2015; Weed–Bach 2019)

Suppose μ has compact support in $\mathbb{R}^d$. Then

$$\mathbb{E}\big[W_p(\hat{\mu}_n, \mu)\big] \asymp \begin{cases} n^{-1/2}, & d < 2p, \\ n^{-1/2}\log n, & d = 2p, \\ n^{-1/d}, & d > 2p. \end{cases}$$

A striking fact: for W_2 on $\mathbb{R}^d$, the *curse of dimensionality* kicks in already at $d = 5$:

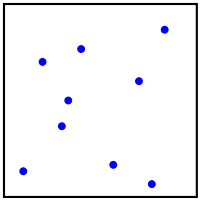
$$W_2(\hat{\mu}_n, \mu) \asymp n^{-1/d} \quad \Rightarrow \quad d = 10 \text{ requires } n \approx 10^{10} \text{ to reach } 0.1.$$

Why the curse? A quick intuition

One-dimensional case ($d = 1$): n samples partition the line into roughly n equal-probability intervals, each of width $\sim 1/n$; quantile matching gives $W_p \asymp n^{-1/2}$.

High dimensions ($d \geq 3$): n points in $[0,1]^d$ have *covering radius* $\sim n^{-1/d}$. Because W_p is geometric, it inherits this rate.

$d = 2$: 9 samples



largest "hole" radius $\sim n^{-1/d}$

d	n to reach $W_2 = 0.1$
1	$\sim 10^2$
2	$\sim 10^2$
5	$\sim 10^5$
10	$\sim 10^{10}$

Remedies: smoothing, slicing, entropic regularization, low-dimensional structure, projections.

Two-sample problems and a CLT

Two-sample Wasserstein (equality testing, goodness-of-fit):

$$W_p(\hat{\mu}_n, \hat{\nu}_m), \qquad X_i \sim \mu,\ Y_j \sim \nu \text{ iid.}$$

Limit theorem (Sommerfeld–Munk 2018; del Barrio et al.): in $d = 1$ (and more generally under low-dimensional/discrete structure), after suitable centering,

$$\sqrt{n}\big(W_p(\hat{\mu}_n, \mu) - \mathbb{E}[W_p(\hat{\mu}_n, \mu)]\big) \xrightarrow{d} \mathcal{L},$$

where the limit $\mathcal{L}$ is a Gaussian-type law or a functional of a Gaussian process.

Why statisticians care:

- *Wasserstein goodness-of-fit* tests for $H_0 : \mu = \mu_0$.
- *Two-sample* tests for $H_0 : \mu = \nu$ sensitive to both location and shape.
- Inference tools: bootstrap (Sommerfeld–Munk), debiased estimators (Bigot et al.), resampling schemes.

Why regularize?

Exact OT (discrete): an LP with $O(n^3 \log n)$ complexity — infeasible for $n = 10^5$.

The Cuturi (2013) idea: add an *entropic penalty* to the objective.

Entropic OT

$$\mathrm{OT}_c^{\varepsilon}(\mu, \nu) \;=\; \inf_{\pi \in \Pi(\mu,\nu)} \int c \,\mathrm{d}\pi + \varepsilon \,\mathrm{KL}\big(\pi \,\|\, \mu \otimes \nu\big).$$

- $\varepsilon \to 0$: recovers classical OT.
- $\varepsilon \to \infty$: collapses to the independent coupling $\mu \otimes \nu$.
- *Strictly convex* $\Rightarrow$ unique solution with a *product structure*.

Structure of the optimum: in the discrete case, $P^*_{ij} = u_i K_{ij} v_j$ with $K_{ij} = e^{-C_{ij}/\varepsilon}$.

The Sinkhorn algorithm

Algorithm 1 Sinkhorn–Knopp

Input: cost $C \in \mathbb{R}^{n \times m}$, marginals $a \in \Delta^n$, $b \in \Delta^m$, regularizer $\varepsilon > 0$
$K \leftarrow \exp(-C/\varepsilon)$; $v \leftarrow \mathbf{1}_m$
while not converged **do**
    $u \leftarrow a \oslash (Kv)$; $v \leftarrow b \oslash (K^\top u)$
end while
Return $P = \operatorname{diag}(u)\, K \operatorname{diag}(v)$

Why does it work?

- Equivalent to *matrix scaling* / *IPFP* — familiar from contingency-table analysis.
- Equivalent to *alternating maximization* of the dual potentials (φ, ψ).
- Linear convergence (contraction in the Hilbert projective metric).
- Each iteration is a matrix-vector product $\Rightarrow$ excellent on GPUs.
- Complexity: $\tilde{O}(n^2/\varepsilon^2)$ (Altschuler–Weed–Rigollet, 2017).

The *statistical* benefits of regularization

The bias–variance–dimension triangle:

- **Breaking the curse:** Genevay et al. (2019) showed that for the Sinkhorn divergence,

$$\mathbb{E}\big|S_\varepsilon(\hat{\mu}_n,\nu) - S_\varepsilon(\mu,\nu)\big| \;\lesssim\; \tfrac{1}{\sqrt{n}}\left(1+\varepsilon^{-\lceil d/2\rceil}\right).$$

  $\Rightarrow$ the *parametric* rate $n^{-1/2}$ at the price of an $\varepsilon^{-d/2}$ constant.

- **Sinkhorn divergence:** debiased version,

$$S_\varepsilon(\mu,\nu) = \mathrm{OT}_c^\varepsilon(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_c^\varepsilon(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_c^\varepsilon(\nu,\nu).$$

  Positive-definite, symmetric, $\to \mathrm{OT}$ as $\varepsilon \to 0$, and $\to$ MMD as $\varepsilon \to \infty$.

- **Differentiability:** Sinkhorn iterations are differentiable in the inputs $\Rightarrow$ backpropagation through OT.

A panorama of applications

Application	**Role of OT**
Wasserstein GAN	Train generator G to minimise $W_1(\mathbb{P}_G, \mathbb{P}_{\text{data}})$
Domain adaptation	Align source and target feature distributions
Barycenters	Fréchet mean of several distributions ($=$ "average histogram")
Distributionally robust opt. (DRO)	Worst-case over $\{\nu : W_p(\nu, \hat{\mu}_n) \leq \rho\}$
Single-cell RNA-seq alignment	Match cells across time/batches
Color transfer / image morphing	A geometric version of histogram matching
Matching estimators in causal inference	Generalization of propensity-score matching
Algorithmic fairness	Equalize predicted-score distributions across groups

Application I — Wasserstein GAN

A generator $G_\theta : \mathcal{Z} \to \mathbb{R}^d$ maps latent $z \sim \rho$ to samples $G_\theta(z)$. Let $\mathbb{P}_\theta$ be the induced distribution.

Classical GAN loss (JS divergence): unstable when $\mathbb{P}_\theta$ and $\mathbb{P}_{\text{data}}$ have disjoint support (common at initialization).

Wasserstein GAN (Arjovsky et al. 2017):

$$\min_\theta \; W_1(\mathbb{P}_\theta, \mathbb{P}_{\text{data}}) \;=\; \min_\theta \sup_{\|f\|_{\text{Lip}} \le 1} \mathbb{E}_{x \sim \mathbb{P}_{\text{data}}}[f(x)] - \mathbb{E}_{z \sim \rho}[f(G_\theta(z))].$$

Parameterise f by a neural "critic" f_w; constrain it to be Lipschitz via weight clipping or gradient penalty.

Why it helps:

- W_1 is finite and continuous in θ even when supports are disjoint (recall $W_1(\delta_0, \delta_t) = |t|$).
- Dual form $\Rightarrow$ trainable critic, informative gradients.
- Empirically: more stable training, fewer mode-collapse pathologies.

Application II — Wasserstein barycenters

Given $\mu_1, \ldots, \mu_K$ on $\mathbb{R}^d$ and weights $\lambda \in \Delta^K$, the *Wasserstein barycenter* is

$$\bar{\mu} \;=\; \arg\min_{\mu} \sum_{k=1}^{K} \lambda_k \, W_2^2(\mu, \mu_k).$$

Compare with the usual Fréchet mean in Euclidean space:

$$\bar{x} \;=\; \arg\min_{x} \sum_{k} \lambda_k \, \|x - x_k\|^2 \;=\; \sum_{k} \lambda_k x_k.$$

The Wasserstein barycenter is its analogue *on the space of distributions*.

Uses in statistics:

- Combining histograms or empirical distributions from different sources (a principled "average").
- Template estimation in shape analysis.
- Aggregating posteriors from distributed Bayesian computation (e.g. Srivastava et al.).
- Population-level summaries in functional data analysis.

Application III — Distributionally robust optimization

Given data $\hat{\mu}_n$, solve

$$\min_{\theta} \sup_{\nu:\, W_p(\nu,\hat{\mu}_n)\le\rho} \mathbb{E}_{(X,Y)\sim\nu}\big[\ell(\theta;X,Y)\big].$$

Interpretation. Protect yourself against *all* distributions within Wasserstein-radius ρ of the empirical distribution.

A beautiful theorem (Blanchet–Murthy, Gao–Kleywegt, Esfahani–Kuhn): for many losses, Wasserstein DRO is *exactly equivalent* to a *regularized* version of the ERM problem.

For example, with linear regression and $p = 2$:

$$\sup_{\nu:W_2(\nu,\hat{\mu}_n)\le\rho} \mathbb{E}_{\nu}[(Y - X^\top\theta)^2] \;=\; \left(\sqrt{\tfrac{1}{n}\textstyle\sum_i (y_i - x_i^\top\theta)^2} + \rho\,\|\theta\|_2\right)^2.$$

$\Rightarrow$ *Ridge-like* regularization emerges *from first principles* of distributional robustness.

Similarly, W_∞ DRO yields Lasso / ℓ_∞-type regularization.

Application IV — Causal inference and matching

Classical matching (Rubin, Rosenbaum). In observational studies, match treated and control units on covariates to approximate a randomized experiment.

OT viewpoint. Matching is exactly a transport problem:

$$\min_{\pi \in \Pi(\hat{\mu}_T, \hat{\mu}_C)} \int \|x - x'\|^2 \, \mathrm{d}\pi(x, x').$$

Advantages over classical matching:

- Handles many-to-many matches with *fractional weights*.
- Global optimum (not greedy).
- Provides an estimate of the *distributional imbalance* between groups (the transport cost itself).
- Naturally extends to continuous treatments and longitudinal data.

See: Gunsilius (2023), Torous–Gunsilius–Rigollet (2024) and the literature on *optimal transport for causal inference*.

Sliced Wasserstein — beating the curse cheaply

Idea. Project onto random 1D directions and average:

$$\mathrm{SW}_p^p(\mu,\nu) \;=\; \int_{\mathbb{S}^{d-1}} W_p^p\big(u_\sharp^\top \mu,\; u_\sharp^\top \nu\big)\,\mathrm{d}\sigma(u),$$

where $u_\sharp^\top \mu$ is the pushforward of μ by $x \mapsto u^\top x$.

Properties:

- It is a proper metric on $\mathcal{P}_p(\mathbb{R}^d)$, equivalent to W_p on compact sets.
- Each 1D slice has a closed form (sort + subtract), so SW is essentially free to compute.
- **Parametric rate:** $\mathbb{E}[\mathrm{SW}_p(\hat{\mu}_n,\mu)] \lesssim n^{-1/2}$ in *every* dimension d.
- Variants: max-sliced, generalized-sliced, tree-sliced Wasserstein.

Unbalanced OT — when mass is not conserved

Standard OT requires $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d) = 1$. In many applications (cell birth/death, ecology, images with occlusion) the *total mass changes*.

Unbalanced OT (Chizat–Peyré–Schmitzer–Vialard):

$$\inf_{\pi \geq 0} \int c \, \mathrm{d}\pi + \tau_1 \, \mathrm{KL}(\pi_1 \| \mu) + \tau_2 \, \mathrm{KL}(\pi_2 \| \nu),$$

where π_1, π_2 are the marginals of π, and $\tau_1, \tau_2 > 0$.

- Trades off transport cost against mass creation / destruction.
- Leads to the *Hellinger–Kantorovich* distance.
- Has a Sinkhorn-style algorithm.
- Widely used in computational biology (e.g. Waddington-OT for developmental trajectories in single-cell data).

Gromov–Wasserstein — comparing distributions on different spaces

Want to compare μ on $\mathcal{X}$ and ν on $\mathcal{Y}$ when *the spaces are different*? Standard OT requires a common cost $c(x, y)$ — impossible.

Gromov–Wasserstein (Mémoli 2011):

$$\mathrm{GW}^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint \big(d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')\big)^2 \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y').$$

- Compares only *internal* distances within each space — invariant to rotations, isometries.
- Non-convex (quadratic in π), but solvable by iterative Sinkhorn-type methods.
- Applications: graph matching, shape comparison, cross-species genomics, NLP embedding alignment.

The Wasserstein geometry of distributions

$\big(\mathcal{P}_2(\mathbb{R}^d), W_2\big)$ is a *Riemannian-like* space with deep structure.

- **Geodesics (McCann's displacement interpolation):** between μ_0, μ_1, the geodesic is

$$\mu_t = \big((1-t)\mathrm{id} + tT\big)_\sharp \mu_0,$$

  where T is the optimal transport map $\mu_0 \to \mu_1$.
- **Tangent space:** gradient fields, Otto calculus.
- **Gradient flows:** the Fokker–Planck equation is the W_2-gradient flow of relative entropy (the JKO scheme, 1998). Many PDEs are gradient flows on Wasserstein space.
- **Statistics on Wasserstein space:** PCA, regression, and hypothesis testing when the data points are themselves distributions.

This perspective has reshaped large parts of analysis, PDE theory, and probability.

Summary

What OT gives us:

1. A geometry-aware *distance* between probability measures.
2. A *coupling* that is optimal in a precise sense — i.e. a principled matching.
3. A *dual formulation* that turns the problem into optimisation over functions.

Why OT matters to statistics:

- Unifies goodness-of-fit, two-sample testing, matching, and many divergence-based methods under one geometric roof.
- Provides principled *geometric* regularization (ridge/lasso emerge from DRO).
- Underlies modern tools: WGAN, FID, Sinkhorn divergences, differentiable sorting, soft matching.
- Suggests a generalization of quantiles and ranks to $\mathbb{R}^d$.

Live research frontiers: sample complexity in high dimensions; inference and CLTs for W_p; neural OT; unbalanced and Gromov–Wasserstein; causal and fair ML; OT for diffusion models and flow matching.