

# Optimal transport, Gromov–Wasserstein and generative models

François-Xavier Vialard

LIGM, Université Gustave Eiffel



A focus on optimal transport

# Measures are dual objects I

Let  $X$  be a compact metric space and let  $C(X)$  carry the uniform norm. The Riesz–Markov–Kakutani theorem identifies finite signed Radon measures with continuous linear functionals:

$$\mathcal{M}(X) = C(X)^*, \quad \langle f, \mu \rangle := \int_X f(x) \, d\mu(x).$$

Positivity and unit mass define the probability measures

$$\mathcal{P}(X) = \{\mu \in \mathcal{M}(X) : \mu \geq 0, \langle \mathbf{1}, \mu \rangle = 1\}.$$

## Atomic measures

A point  $x \in X$  defines the Dirac measure

$$\langle f, \delta_x \rangle = f(x).$$

A sample  $x_1, \dots, x_n$  defines its empirical law

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \quad \langle f, \mu_n \rangle = \frac{1}{n} \sum_{i=1}^n f(x_i).$$

# Measures are dual objects II

## Measures with a density

Relative to a reference measure  $m$ ,

$$\mu = \rho m, \quad \rho \geq 0, \quad \int_X \rho \, dm = 1,$$

and

$$\langle f, \mu \rangle = \int_X f(x) \rho(x) \, dm(x).$$

Densities and Dirac masses belong to the same space  $\mathcal{P}(X)$ .

If  $Z$  has law  $\mu$ , then  $\langle f, \mu \rangle = \mathbb{E}[f(Z)]$ : measures encode every observable.

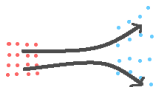


# Maps act on measures I

Let  $T : X \rightarrow Y$  be continuous. The pushforward  $T_{\#}\mu$  is characterized by

$$(T_{\#}\mu)(B) = \mu(T^{-1}(B)), \quad \langle f, T_{\#}\mu \rangle = \langle f \circ T, \mu \rangle.$$

Transport by a map



## Pullback, pushforward and atomic measures

Functions move backward; measures move forward.

$$C(X) \xleftarrow{T^*} C(Y)$$

$$\mathcal{M}(X) \xrightarrow{T_{\#}} \mathcal{M}(Y)$$

$$T_{\#} \left( \sum_i a_i \delta_{x_i} \right) = \sum_i a_i \delta_{T(x_i)}.$$

# Maps act on measures II

## Densities transform through the Jacobian

Let  $X, Y \subset \mathbb{R}^d$  and let  $T : X \rightarrow Y$  be a  $C^1$  diffeomorphism. If  $\mu = \rho(x) dx$ , then  $T_{\#}\mu = \rho_T(y) dy$ , where

$$\rho_T(y) = \rho(T^{-1}(y)) \left| \det DT^{-1}(y) \right|.$$

Equivalently, conservation of mass reads

$$\rho_T(T(x)) \left| \det DT(x) \right| = \rho(x).$$

## Weak-\* topology and compactness

For every  $f \in C(X)$ ,

$$\mu_n \rightharpoonup^* \mu \iff \langle f, \mu_n \rangle \rightarrow \langle f, \mu \rangle.$$

Continuous maps preserve this convergence:  $T_{\#}\mu_n \rightharpoonup^* T_{\#}\mu$ . On compact  $X$ ,  $\mathcal{P}(X)$  is weak-\* compact by Banach–Alaoglu. On a noncompact Polish space, tightness gives compactness through Prokhorov's theorem.

Compactness plus lower semicontinuity is the direct-method engine behind existence in optimal transport.

# Ways to compare probability measures

There is no canonical distance on  $\mathcal{P}(X)$ . Each construction decides which differences between distributions are visible.

## Observables and density ratios

- **Integral probability metrics (IPMs):** maximize the difference of expectations over a test class  $\mathcal{F}$ . Examples include total variation, MMD and  $W_1$ .
- **$f$ -divergences:** integrate a convex function of the density ratio  $d\mu/d\nu$ . Examples include KL, total variation, Hellinger and Jensen–Shannon.

## Geometry-aware constructions

- **Optimal transport:** minimizes the ground cost over couplings;  $W_p$  retains the geometry of  $X$ .
- **Sliced optimal transport:** averages inexpensive one-dimensional Wasserstein distances over projections.
- **Entropic OT and Sinkhorn divergences:** regularize transport to obtain faster computation and smoother sample losses.

The right discrepancy depends on the topology, statistical regime and computation required by the application.

# Integral probability metrics

For a symmetric class  $\mathcal{F}$  of bounded test functions, define

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f]| = \sup_{f \in \mathcal{F}} |\langle f, \mu - \nu \rangle|.$$

The class  $\mathcal{F}$  specifies the observables used to distinguish measures.

## Three important choices of test functions

- $\|f\|_{\infty} \leq 1$ : total variation, up to the factor  $1/2$ ;
- $\|f\|_{\mathcal{H}_k} \leq 1$ : maximum mean discrepancy (MMD);
- $\text{Lip}(f) \leq 1$ :  $W_1$  by Kantorovich–Rubinstein duality.

If  $\mathcal{F}$  separates probability measures,  $d_{\mathcal{F}}$  is a metric.

## Samples enter through empirical expectations

For  $\hat{\mu}_n = n^{-1} \sum_i \delta_{X_i}$ ,

$$d_{\mathcal{F}}(\hat{\mu}_n, \mu) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{\mu}[f] \right|.$$

The statistical rate is controlled by the complexity of  $\mathcal{F}$ : bounded RKHS balls often give an  $n^{-1/2}$  rate, while Lipschitz tests retain a strong dependence on dimension.

# $f$ -divergences compare density ratios

Let  $f : [0, +\infty) \rightarrow \mathbb{R} \cup \{+\infty\}$  be convex with  $f(1) = 0$ . If  $\mu \ll \nu$ , define

$$D_f(\mu \parallel \nu) := \int_X f\left(\frac{d\mu}{d\nu}\right) d\nu.$$

## Classical examples

- KL divergence:  $f(t) = t \log t$ ;
- total variation:  $f(t) = \frac{1}{2}|t - 1|$ ;
- squared Hellinger divergence:  $f(t) = (\sqrt{t} - 1)^2$ ;
- Jensen–Shannon divergence: a finite symmetrization of KL.

## Structural properties

Convexity gives  $D_f(\mu \parallel \nu) \geq 0$ , and measurable maps cannot increase an  $f$ -divergence. In general, however, it is asymmetric and does not satisfy the triangle inequality. For KL,

$$d_{\text{TV}}(\mu, \nu)^2 \leq \frac{1}{2} \text{KL}(\mu \parallel \nu).$$

Support mismatch is severe:  $\text{KL}(\delta_x \parallel \delta_y) = +\infty$  for  $x \neq y$ .

# What is a positive definite kernel?

## Definition (Kernel)

A symmetric function

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

is called a *positive kernel* (sometimes positive definite) if for any  $n \in \mathbb{N}$ , any  $x_1, \dots, x_n \in \mathcal{X}$ , and any  $c_1, \dots, c_n \in \mathbb{R}$ ,

$$\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0.$$

- The matrix  $K_{ij} = k(x_i, x_j)$  is positive semi-definite.
- For distinct  $x_i$ , if the inequality is strict for non-zero  $(c_i)_{i=1,\dots,n}$ , then  $K$  is said positive definite (also *strictly* positive definite).

$$\blacksquare \text{ Example: } k(x, y) = \langle x, y \rangle \implies \sum_{i,j=1}^n c_i c_j \langle x_i, x_j \rangle = \left\| \sum_{i,j=1}^n c_i x_i \right\|^2.$$

# Algebraic stability of positive definite kernels

## Algebraic stability properties

If  $k_1, k_2$  are PD kernels on  $\mathcal{X}$  and  $\alpha \geq 0$ , then:

- **Scaling:**  $\alpha k_1$  is PD.
- **Sum:**  $k_1 + k_2$  is PD.
- **Tensor products:**  $k_1(x_1, y_1)k_2(x_2, y_2)$ .
- **Product (Schur / Hadamard):**  $k_1 \cdot k_2$  is PD.
- **Pointwise limit:** If  $k_n \rightarrow k$  pointwise and each  $k_n$  is PD, then  $k$  is PD.
- **Exponential:**  $e^{k_1}$  is PD.
- **Function composition:** If  $f$  is a real entire function with non-negative coefficients in its power series, then  $f(k_1)$  is PD.

## Remark

Many classical kernels (Gaussian, Laplacian, Matérn, polynomial) arise from these operations.

# Examples of kernels

- **Linear kernel:**

$$k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$$

- **Polynomial kernel:**

$$k(x, y) = (\langle x, y \rangle + c)^p$$

- **Gaussian (RBF) kernel:**

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

- **Laplacian kernel:**

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$



# From a kernel to a space of functions

## Construction of the RKHS given the kernel $k$

- Consider the linear span

$$\mathcal{H}_0 := \text{span}\{k(\cdot, x) : x \in \mathcal{X}\}$$

- Define an inner product by

$$\left\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \beta_j k(\cdot, x_j) \right\rangle := \sum_{i,j} \alpha_i \beta_j k(x_i, x_j)$$

- Take the completion:  $\mathcal{H}_k := \overline{\mathcal{H}_0}$

## Key property (reproducing)

For all  $f \in \mathcal{H}_k$  and  $x \in \mathcal{X}$ ,

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}.$$

# Feature map representation

## Feature map

A kernel  $k$  is positive definite if and only if there exists a Hilbert space  $\mathcal{H}$  and a map

$$\Phi : \mathcal{X} \rightarrow \mathcal{H}$$

such that

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}.$$

- $\mathcal{H}$  can be infinite-dimensional
- Kernels allow implicit computations in  $\mathcal{H}$

# MMD compares "blurred" distributions I

A symmetric kernel  $k : X \times X \rightarrow \mathbb{R}$  is *positive semidefinite* (or nonnegative definite) if, for every  $N$ ,  $x_1, \dots, x_N \in X$  and  $a_1, \dots, a_N \in \mathbb{R}$ ,

$$\sum_{i,j=1}^N a_i a_j k(x_i, x_j) \geq 0.$$

For a bounded PSD kernel, let  $\mathcal{H}_k$  be its reproducing kernel Hilbert space (RKHS) and define the kernel mean

$$m_\mu := \int_X k(x, \cdot) d\mu(x) \in \mathcal{H}_k.$$

MMD is the corresponding integral probability metric:

$$\text{MMD}_k(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} |\langle f, \mu - \nu \rangle| = \|m_\mu - m_\nu\|_{\mathcal{H}_k}.$$

If  $X, X' \sim \mu$  and  $Y, Y' \sim \nu$  are mutually independent, (see quadratic computational complexity)

$$\text{MMD}_k^2(\mu, \nu) = \mathbb{E}k(X, X') + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y).$$

# MMD compares "blurred" distributions II

## When is it a metric?

If  $k$  is universal,  $m_\mu = m_\nu$  implies  $\mu = \nu$ . A continuous universal kernel on compact  $X$  also makes MMD metrize weak-\* convergence.

## Statistical complexity is parametric

For empirical measures, MMD uses only pairwise kernel evaluations. For bounded  $k$ ,  $\text{MMD}_k(\hat{\mu}_n, \mu) = O_{\mathbb{P}}(n^{-1/2})$ ; the kernel and its bandwidth determine what the test can see.

# The notion of closeness changes the problem

| Quantity | Test class            | Singular laws            | Topology                |
|----------|-----------------------|--------------------------|-------------------------|
| TV       | bounded functions     | finite; often saturated  | strong                  |
| KL       | density ratio         | often $+\infty$          | relative entropy        |
| MMD      | RKHS unit ball        | finite for bounded $k$   | weak-* if $k$ universal |
| $W_1$    | 1-Lipschitz functions | finite with first moment | weak + moments          |

## A simple example: two Dirac masses

On a metric space  $(X, d)$ , for  $x \neq y$ ,

$$d_{\text{TV}}(\delta_x, \delta_y) = 1, \quad \text{KL}(\delta_x \| \delta_y) = +\infty,$$

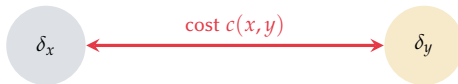
whereas

$$\text{MMD}_k^2(\delta_x, \delta_y) = k(x, x) + k(y, y) - 2k(x, y), \quad W_1(\delta_x, \delta_y) = d(x, y).$$

Optimal transport is designed to retain the geometry of where mass lives.

# Optimal transport lifts a ground cost to probability measures

Let  $c : X \times Y \rightarrow [0, +\infty]$  quantify the effort required to move one unit of mass from  $x$  to  $y$ .



## Transport between Dirac masses

There is only one possible move, hence

$$\text{OT}_c(\delta_x, \delta_y) = c(x, y).$$

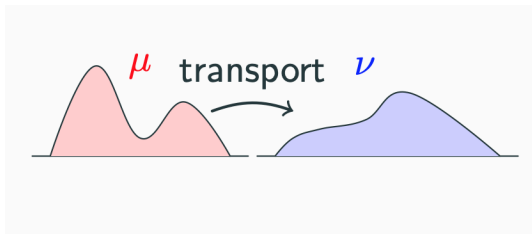
In particular, if  $X = Y$  is a metric space and  $c(x, y) = d(x, y)^p$ , then

$$W_p(\delta_x, \delta_y) = d(x, y).$$

## The lifting problem

How should the ground cost be extended from points to arbitrary  $\mu \in \mathcal{P}(X)$  and  $\nu \in \mathcal{P}(Y)$  while preserving the geometry of  $c$ ?

# Monge: move mass with a map



A map  $T : X \rightarrow Y$  assigns one destination  $T(x)$  to each particle and must satisfy  $T_{\#}\mu = \nu$ .

## Monge problem

$$M_c(\mu, \nu) := \inf_{T_{\#}\mu = \nu} \int_X c(x, T(x)) \, d\mu(x).$$

The optimization variable is a deterministic map, so mass cannot split.

## A map cannot split mass

No map sends  $\delta_0$  to  $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ . The Monge problem may therefore be infeasible.

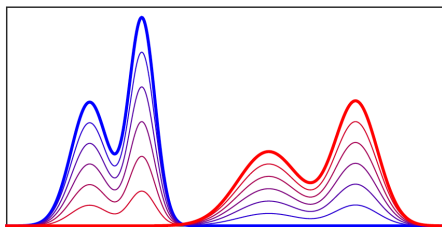
# Wasserstein interpolation moves mass

Between densities  $\rho_0$  and  $\rho_1$ , compare

$$\rho_t^{L^2} = (1-t)\rho_0 + t\rho_1, \quad \rho_t^{W_2} = ((1-t)\text{Id} + tT)_\# \rho_0,$$

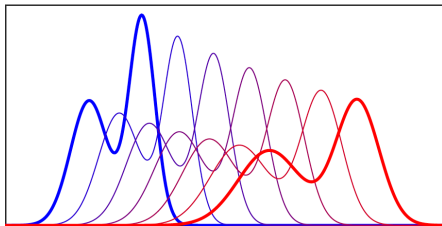
where  $T$  is the optimal map from  $\rho_0$  to  $\rho_1$ .

$L^2$  interpolation



- $L^2$  fades mass between fixed locations.
- $W_2$  displaces mass through the ground space.

Optimal transport



Illustrations adapted from Gabriel Peyré's OT course material.



# Kantorovich: optimize over couplings

A coupling is a joint probability law with prescribed marginals:

$$\Pi(\mu, \nu) := \{ \gamma \in \mathcal{P}(X \times Y) : (p_X)_\# \gamma = \mu, (p_Y)_\# \gamma = \nu \}.$$

The value  $\gamma(A \times B)$  describes how much mass travels from  $A$  to  $B$ .

## Kantorovich problem

$$\mathsf{K}_c(\mu, \nu) := \min_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) \, d\gamma(x, y).$$

This is a linear optimization problem over the convex set  $\Pi(\mu, \nu)$ .

## Monge is embedded in Kantorovich

A map  $T$  induces  $\gamma_T = (\text{Id}, T)_\# \mu$ . Therefore  $\mathsf{K}_c \leq \mathsf{M}_c$ , with equality when an optimal coupling is induced by a map.

## Existence and Wasserstein distance

Weak-\* compactness of  $\Pi(\mu, \nu)$  and lower semicontinuity of  $c$  give a minimizer. For  $c = d^p$ ,

$$W_p(\mu, \nu) := \mathsf{K}_{d^p}(\mu, \nu)^{1/p}.$$

# Equal-weight empirical OT is an assignment problem

Let

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \quad \nu_n = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}, \quad C_{ij} = c(x_i, y_j).$$

A coupling is represented by a bistochastic matrix

$$\mathcal{B}_n := \left\{ A \in \mathbb{R}_+^{n \times n} : A\mathbf{1} = \mathbf{1}, A^\top \mathbf{1} = \mathbf{1} \right\}, \quad \gamma_A = \frac{1}{n} \sum_{i,j=1}^n A_{ij} \delta_{(x_i, y_j)}.$$

## Finite-dimensional Kantorovich problem

$$\mathsf{K}_c(\mu_n, \nu_n) = \frac{1}{n} \min_{A \in \mathcal{B}_n} \sum_{i,j=1}^n C_{ij} A_{ij}.$$

## Birkhoff–von Neumann theorem

The objective is linear in  $A$ , and the extreme points of  $\mathcal{B}_n$  are the permutation matrices. Hence some optimum is a permutation  $\sigma \in S_n$ :

$$\mathsf{K}_c(\mu_n, \nu_n) = \frac{1}{n} \min_{\sigma \in S_n} \sum_{i=1}^n c(x_i, y_{\sigma(i)}).$$

Cost ties may produce other optimizers, but a permutation matching is optimal.

# Convex duality: constraints become prices

Consider the convex problem

$$p^* = \inf_{x \in C} \{F(x) : Ax = b\}.$$

Associate a multiplier  $\lambda$  with the constraint and define

$$L(x, \lambda) = F(x) + \langle \lambda, b - Ax \rangle, \quad d(\lambda) = \inf_{x \in C} L(x, \lambda).$$

## Weak duality

If  $Ax = b$ , then  $L(x, \lambda) = F(x)$ . Hence every multiplier gives a lower bound:

$$d(\lambda) \leq p^*, \quad \underbrace{\sup_{\lambda} d(\lambda)}_{\text{dual problem}} \leq \underbrace{p^*}_{\text{primal problem}}.$$

## Strong duality

For a convex problem under mild regularity assumptions, the best lower bound is exact:  $\sup_{\lambda} d(\lambda) = p^*$ .

In optimal transport, the constraints prescribe two marginals. Their multipliers are the two Kantorovich potentials  $\varphi$  and  $\psi$ .

# Kantorovich duality: reduces the dimension of optimization I

For compact  $X, Y$  and continuous  $c$ , the primal transport problem equals a maximization over functions:

$$\begin{aligned} K_c(\mu, \nu) &= \min_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c \, d\gamma \\ &= \sup_{\substack{\varphi \in C(X), \psi \in C(Y) \\ \varphi(x) + \psi(y) \leq c(x, y)}} \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \right\}. \end{aligned}$$

## Why every dual pair is a certificate

For any feasible  $(\varphi, \psi)$  and any coupling  $\gamma$ ,

$$\begin{aligned} \int \varphi \, d\mu + \int \psi \, d\nu &= \int (\varphi(x) + \psi(y)) \, d\gamma \\ &\leq \int c(x, y) \, d\gamma. \end{aligned}$$

Strong duality says the best lower bound has no gap.

# Kantorovich duality: reduces the dimension of optimization II

## Structure of optimal potentials

- Adding a constant to  $\varphi$  and subtracting it from  $\psi$  leaves the problem unchanged.
- Given  $\varphi$ , the best choice is its  $c$ -transform

$$\varphi^c(y) := \inf_x \{c(x, y) - \varphi(x)\}.$$

- An optimal plan is concentrated on the contact set  $\varphi(x) + \psi(y) = c(x, y)$ .

$$\text{For } X = Y \text{ and } c = d: W_1(\mu, \nu) = \sup_{\text{Lip}(f) \leq 1} \langle f, \mu - \nu \rangle.$$

# Quadratic cost gives convex potentials I

For  $X = Y = \mathbb{R}^d$  and  $c(x, y) = \frac{1}{2} \|x - y\|^2$ ,

$$c(x, y) = \frac{1}{2} \|x\|^2 + \frac{1}{2} \|y\|^2 - \langle x, y \rangle.$$

Given feasible dual potentials  $(\varphi, \psi)$ , define

$$u(x) := \frac{1}{2} \|x\|^2 - \varphi(x), \quad v(y) := \frac{1}{2} \|y\|^2 - \psi(y).$$

## The dual constraint

$$\varphi(x) + \psi(y) \leq c(x, y)$$

is equivalent to

$$u(x) + v(y) \geq \langle x, y \rangle.$$

The best choice is  $v = u^*$ . Replacing  $u$  by  $u^{**}$  can only improve the dual value, so an optimizer may be taken convex.

# Quadratic cost gives convex potentials II

## Contact determines the transport

On the support of an optimal coupling,

$$u(x) + u^*(y) = \langle x, y \rangle \iff y \in \partial u(x).$$

Wherever the convex function  $u$  is differentiable, the destination is unique:

$$y = \nabla u(x).$$

For the quadratic cost, the optimal plan is supported by the graph of the gradient of a convex function

# Brenier's theorem

## Theorem (quadratic optimal transport)

Let  $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$  and assume that  $\mu$  is absolutely continuous with respect to Lebesgue measure. Then there exists a convex function  $u : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  such that

$$T = \nabla u, \quad T_{\#}\mu = \nu.$$

Moreover,

$$\gamma^* = (\text{Id}, \nabla u)_{\#}\mu$$

is the unique optimal coupling for  $c(x, y) = \frac{1}{2}\|x - y\|^2$ , and  $T$  is the unique optimal map  $\mu$ -almost everywhere.

- Conversely, every gradient of a convex function that pushes  $\mu$  to  $\nu$  is optimal.
- If  $\nu$  is also absolutely continuous, the reverse optimal map is  $\nabla u^*$ , with  $\nabla u^* \circ \nabla u = \text{Id}$   $\mu$ -almost everywhere.

Y. Brenier, *Commun. Pure Appl. Math.* 44 (1991), 375–417.



# Twisted costs also select a transport map I

Let  $c \in C^1(X \times Y)$ . The cost is *twisted* when, for every  $x$ ,

$$y \mapsto \nabla_x c(x, y) \quad \text{is injective.}$$

On the contact set of optimal dual potentials,  $\varphi(x) + \psi(y) = c(x, y)$ . At differentiability points of  $\varphi$ ,

$$\nabla \varphi(x) = \nabla_x c(x, y).$$

## Twist turns a plan into a map

Injectivity determines at most one destination

$$y = Y_c(x, \nabla \varphi(x)), \quad \nabla_x c(x, Y_c(x, p)) = p.$$

Under the standard regularity assumptions and  $\mu \ll dx$ , the optimal coupling is unique and concentrated on the graph of this map.

# Twisted costs also select a transport map II

## Examples

- $c(x, y) = h(x - y)$  with  $h$  strictly convex;
- $c(x, y) = \frac{1}{2}d_M(x, y)^2$  away from the cut locus.

## For regularity later

One also assumes nondegeneracy:

$$\det D_{xy}^2 c(x, y) \neq 0.$$

# Smoothness begins with Monge–Ampère I

Write  $\mu = \rho_0(x) \, dx$  and  $\nu = \rho_1(y) \, dy$ . If the Brenier map  $T = \nabla u$  is smooth and invertible, the change-of-variables formula gives

$$\rho_0(x) = \rho_1(\nabla u(x)) \det D^2 u(x).$$

## Monge–Ampère with the transport boundary condition

$$\det D^2 u(x) = \frac{\rho_0(x)}{\rho_1(\nabla u(x))}, \quad \nabla u(\Omega) = \Lambda,$$

where  $\Omega = \text{supp } \mu$  and  $\Lambda = \text{supp } \nu$ . Convexity selects the elliptic branch.

# Smoothness begins with Monge–Ampère II

## What Brenier gives

A convex potential and an almost-everywhere map. The equation first makes sense weakly, for instance in the Alexandrov sense.

## What smoothness requires

Positivity and regularity of the densities, together with suitable convexity and smoothness of the supports. Interior and boundary regularity are distinct issues.

The quadratic regularity theory is due in large part to Caffarelli.

# Can be generalized to other costs

Assume twist and  $\det D_{xy}^2 c \neq 0$ , and write

$$T(x) = Y_c(x, \nabla \varphi(x)), \quad \nabla \varphi(x) = \nabla_x c(x, T(x)).$$

Differentiating this identity gives

$$D^2 \varphi(x) = D_{xx}^2 c(x, T(x)) + D_{xy}^2 c(x, T(x)) DT(x).$$

Combining it with  $\rho_0(x) = \rho_1(T(x)) |\det DT(x)|$  yields

$$\left| \det \left( D^2 \varphi - D_{xx}^2 c(x, T) \right) \right| = \left| \det D_{xy}^2 c(x, T) \right| \frac{\rho_0(x)}{\rho_1(T(x))}.$$

## Same mechanism

Mass conservation still gives a Jacobian equation; the cost now enters both the map  $T$  and the nonlinear differential operator.

## New difficulty

Convexity of  $u$  is replaced by  $c$ -convexity of  $\varphi$ . Ellipticity and regularity depend on fourth-order geometry hidden in  $c$ .

# The MTW tensor is the curvature of the cost

Let  $c_{i,j} = \partial_{x_i} \partial_{y_j} c$  and let  $c^{i,j}$  denote the inverse mixed Hessian; indices before and after the comma denote  $x$ - and  $y$ -derivatives. In coordinates, the Ma–Trudinger–Wang tensor is

$$\mathfrak{S}_c(x, y)(\xi, \eta) = (c_{ij,r} c^{r,s} c_{s,kl} - c_{ij,kl}) \xi^i \xi^j \eta^k \eta^l.$$

The weak MTW condition requires

$$\mathfrak{S}_c(x, y)(\xi, \eta) \geq 0 \quad \text{whenever} \quad \xi^i c_{i,j} \eta^j = 0.$$

## Why it matters

MTW controls how  $c$ -support functions bend. Together with  $c$ -convexity of the domains and regular positive densities, it yields a priori estimates and continuity or smoothness of optimal maps.

## Geometry and counterexamples

- Euclidean quadratic cost:  $\mathfrak{S}_c = 0$ ;
- round spheres: positive cross-curvature away from the cut locus;
- failure of weak MTW can produce discontinuous maps even for smooth positive densities.

# Three burdens of optimal transport

- **Computation.** Dense discrete OT uses  $O(N^2)$  memory. Exact assignment has cubic scaling; dense auction methods cost  $O(N^3 \log(NC))$ .
- **Statistics.** For  $d > 2p$ ,
$$\mathbb{E} W_p(\hat{\mu}_n, \mu) \asymp n^{-1/d}, \quad n \gtrsim \varepsilon^{-d}.$$
- **Nonsmoothness.** Discrete OT is piecewise affine. Even  $W_1(\delta_x, \delta_0) = |x|$  is not differentiable at its minimum.

## Goal

A differentiable, statistically efficient proxy that retains the ground geometry and admits faster computation.

Bertsekas–Eckstein (1988); Fournier–Guillin (2015); Weed–Bach (2019).

# Entropic optimal transport smooths the coupling

For  $\alpha \in \mathcal{P}(X)$  and  $\beta \in \mathcal{P}(Y)$ , define

$$\text{OT}_\varepsilon(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \left\{ \int c(x, y) \, d\pi(x, y) + \varepsilon \text{KL}(\pi \| \alpha \otimes \beta) \right\}.$$

Here

$$\text{KL}(\pi \| \alpha \otimes \beta) = \int \log \left( \frac{d\pi}{d(\alpha \otimes \beta)} \right) d\pi.$$

## What $\varepsilon$ does

The KL term is strictly convex, hence the optimal coupling is unique. It favors couplings that are diffuse relative to  $\alpha \otimes \beta$  and smooths changes of the optimal plan.

## Two limiting regimes

$$\varepsilon \downarrow 0 : \quad \text{OT}_\varepsilon \longrightarrow \text{OT},$$

$$\varepsilon \uparrow \infty : \quad \pi_\varepsilon \longrightarrow \alpha \otimes \beta.$$

Schrödinger's bridge problem; Cuturi, NeurIPS 2013.



# The entropic dual uses a natural soft constraint

Entropic OT has the unconstrained smooth dual

$$\text{OT}_\varepsilon(\alpha, \beta) = \sup_{f, g} \left\{ \langle f, \alpha \rangle + \langle g, \beta \rangle - \varepsilon \int \left( e^{(f(x) + g(y) - c(x, y)) / \varepsilon} - 1 \right) d\alpha(x) d\beta(y) \right\}.$$

## Optimal coupling

First-order optimality gives

$$\frac{d\pi_\varepsilon}{d(\alpha \otimes \beta)}(x, y) = e^{(f_\varepsilon(x) + g_\varepsilon(y) - c(x, y)) / \varepsilon}.$$

Thus  $\pi_\varepsilon$  is a diagonal scaling of the Gibbs kernel  $K_\varepsilon(x, y) = e^{-c(x, y) / \varepsilon}$ .

## Differentiability

Under standard assumptions, the optimal dual potentials are first variations:

$$\nabla_\alpha \text{OT}_\varepsilon = f_\varepsilon, \quad \nabla_\beta \text{OT}_\varepsilon = g_\varepsilon,$$

modulo the usual additive constant.

The hard constraint  $f + g \leq c$  becomes a smooth exponential penalty.

# Sinkhorn's algorithm is alternating matrix scaling I

Let  $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ ,  $\beta = \sum_{j=1}^m b_j \delta_{y_j}$ , and

$$K_{ij} := e^{-c(x_i, y_j)/\varepsilon} > 0.$$

Absorbing the reference weights into two positive scaling vectors, the optimal coupling has the form

$$P_\varepsilon = \text{diag}(u) K \text{diag}(v).$$

## Enforce one marginal at a time

Starting from  $v^{(0)} > 0$ , iterate pointwise divisions

$$u^{(k+1)} = \frac{a}{K v^{(k)}}, \quad v^{(k+1)} = \frac{b}{K^\top u^{(k+1)}}.$$

The first update fixes the row sums; the second fixes the column sums.

# Sinkhorn's algorithm is alternating matrix scaling II

## Dual interpretation

Since the reference is  $\alpha \otimes \beta$ ,

$$f = \varepsilon \log(u/a), \quad g = \varepsilon \log(v/b).$$

Each update maximizes the smooth dual in one potential: Sinkhorn is block coordinate ascent.

## Cost of an iteration

Dense matrix-vector products cost  $O(nm)$  time and storing  $K$  costs  $O(nm)$  memory. Kernel structure can reduce these costs.

# When Sinkhorn struggles

- For  $K > 0$ , alternating scaling converges to the unique entropic coupling.
- Small  $\varepsilon$  makes  $K = e^{-C/\varepsilon}$  ill-conditioned.
- Consequences: slow convergence and numerical underflow.
- Stabilization: log-sum-exp updates,  $\varepsilon$ -scaling and warm starts.
- Stopping criteria: marginal residuals or a primal–dual gap.

Sinkhorn–Knopp (1967); Franklin–Lorenz, *Linear Algebra Appl.* 114–115 (1989), 717–735.

# Entropic OT is smooth, but it is biased

The raw regularized value is generally not a divergence:

$$\text{OT}_\varepsilon(\alpha, \alpha) \neq 0 \text{ in general, } \quad \nabla_\beta \text{OT}_\varepsilon(\alpha, \beta) \Big|_{\beta=\alpha} \neq 0.$$

Consequently, minimizing  $\beta \mapsto \text{OT}_\varepsilon(\alpha, \beta)$  need not recover  $\beta = \alpha$ .

## Origin of the bias

Entropy rewards diffuse couplings. The self-interaction of each measure contributes to the loss and can create shrinkage or blurring in optimization problems.

## The large- $\varepsilon$ limit

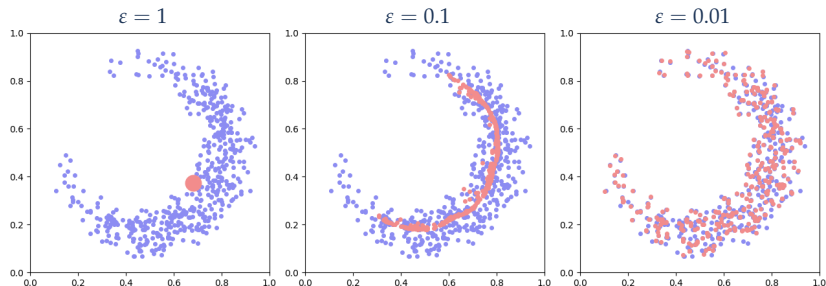
Since  $\pi_\varepsilon \rightarrow \alpha \otimes \beta$ ,

$$\text{OT}_\varepsilon(\alpha, \beta) \longrightarrow \iint c(x, y) \, \mathrm{d}\alpha(x) \mathrm{d}\beta(y),$$

an interaction energy rather than an OT distance.

The cure is to subtract the two entropic self-interactions.

# Entropic bias in simulations



The gradient flow of the red particles converges to a biased target.

The larger  $\epsilon$ , the larger the bias towards a sort of mean. The smaller  $\epsilon$ , the larger the computational cost.

The goal: Correct the bias while retaining the good properties.

Idea: study the large  $\epsilon$  limit, and mimick MMD.

# Sinkhorn divergence and CNT costs

For a symmetric continuous cost  $c : X \times X \rightarrow \mathbb{R}$ , define

$$S_\varepsilon(\alpha, \beta) := \text{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\text{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\text{OT}_\varepsilon(\beta, \beta).$$

## Conditionally negative type (CNT)

The cost  $c$  is CNT if, for every  $x_1, \dots, x_N \in X$ ,

$$\sum_i a_i = 0 \implies \sum_{i,j} a_i a_j c(x_i, x_j) \leq 0.$$

Equivalently (Schoenberg),  $K_\varepsilon = e^{-c/\varepsilon}$  is positive definite for every  $\varepsilon > 0$ .

## Theorem (Feydy et al.)

Let  $X$  be bounded;  $c$  need not vanish on the diagonal. Then

$$c \text{ is CNT} \iff S_\varepsilon(\alpha, \beta) \geq 0$$

for all  $\alpha, \beta \in \mathcal{P}(X)$  and every  $\varepsilon > 0$ . In particular,  $\|x - y\|^p$  is CNT for  $0 < p \leq 2$ .

CNT gives nonnegativity; strict separation and metrization of weak convergence require stronger kernel assumptions. Feydy et al., AISTATS 2019; see also Aubin, Ehrlacher and Todeschi, *Debiasing optimal transport: classical and entropic*.

# Two elementary proofs

Let  $f_\alpha, f_\beta$  be optimal self-potentials and set  $\lambda_\alpha = e^{f_\alpha/\varepsilon}\alpha$ ,  $\lambda_\beta = e^{f_\beta/\varepsilon}\beta$  and  $K_\varepsilon = e^{-c/\varepsilon}$ .

## 1. Use the self-potentials as competitors in the cross problem

The entropic dual gives the suboptimality bound

$$S_\varepsilon(\alpha, \beta) \geq \frac{\varepsilon}{2} \iint K_\varepsilon(x, y) d(\lambda_\alpha - \lambda_\beta)(x) d(\lambda_\alpha - \lambda_\beta)(y) \geq 0.$$

The last inequality follows from Schoenberg's theorem when  $c$  is CNT. Conversely, as  $\varepsilon \rightarrow \infty$ ,

$$S_\varepsilon(\alpha, \beta) \longrightarrow -\frac{1}{2} \iint c(x, y) d(\alpha - \beta)(x) d(\alpha - \beta)(y).$$

Positivity for every pair of probabilities therefore implies CNT.

## 2. Use concavity – a stronger viewpoint

A change of variables in the dual shows that  $F_\varepsilon(\mu) := -\frac{1}{2}\text{OT}_\varepsilon(\mu, \mu)$  is convex. In fact,

$$S_\varepsilon(\alpha, \beta) \geq \frac{1}{2} \langle \alpha - \beta, \nabla F_\varepsilon(\alpha) - \nabla F_\varepsilon(\beta) \rangle \geq 0,$$

a symmetric Bregman lower bound.



# Sinkhorn divergences interpolate between OT and MMD

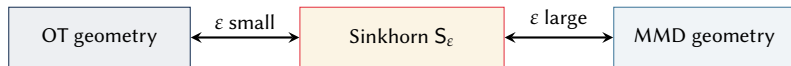
The regularization scale continuously changes the geometry:

$$\varepsilon \downarrow 0 : S_\varepsilon(\alpha, \beta) \longrightarrow \text{OT}_c(\alpha, \beta),$$

while

$$\varepsilon \uparrow \infty : S_\varepsilon(\alpha, \beta) \longrightarrow \frac{1}{2} \text{MMD}_{-c}^2(\alpha, \beta),$$

whenever  $-c$  induces the corresponding positive kernel discrepancy.



## The ground cost is never discarded

It enters through  $K_\varepsilon = e^{-c/\varepsilon}$ . For costs of negative type, the large-scale limit is an energy distance; for the squared Euclidean cost it degenerates to a discrepancy between means.

# Does the Sinkhorn divergence meet the wish list?

For fixed  $\varepsilon > 0$ , empirical Sinkhorn divergences have a parametric sampling exponent:

$$\mathbb{E} \left| S_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n) - S_\varepsilon(\alpha, \beta) \right| \leq \frac{C(\varepsilon, c, d)}{\sqrt{n}}.$$

## What is gained

- geometry inherited from  $c$ ;
- differentiable loss and gradients;
- $n^{-1/2}$  sampling exponent for fixed  $\varepsilon$ ;
- $O(nm)$  work per dense Sinkhorn iteration.

## What remains

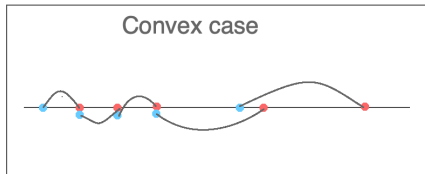
- $C(\varepsilon, c, d)$  deteriorates as  $\varepsilon \downarrow 0$ ;
- small  $\varepsilon$  worsens numerical conditioning;
- the dense kernel still costs  $O(nm)$  memory;
- debiasing requires cross and self transport solves.

$\varepsilon$  controls a geometry–statistics–computation tradeoff. Genevay, Chizat, Bach, Cuturi and Peyré, AISTATS 2019.

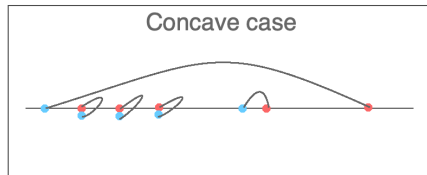
Generalization of OT and derived models: Sliced Wasserstein, Unbalanced OT, Gromov-Wasserstein

# In one dimension, sorting depends on the cost

For costs  $c(x, y) = f(y - x)$ , the shape of  $f$  determines the matching.



**Convex**  $f$ : monotone matching.



**Concave**  $f$ : long displacements may be cheaper.

## For the quadratic cost

With  $Q_\mu = F_\mu^{-1}$  and  $Q_\nu = F_\nu^{-1}$ ,

$$W_2^2(\mu, \nu) = \int_0^1 |Q_\mu(s) - Q_\nu(s)|^2 ds.$$

For equal-weight empirical measures, sorting gives  $W_2^2 = \frac{1}{n} \sum_i |x_{(i)} - y_{(i)}|^2$  in  $O(n \log n)$  time.

Sorting is optimal for convex displacement costs, not for every cost.

# Sliced Wasserstein: a competitive discrepancy

For  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  and  $\theta \in \mathbb{S}^{d-1}$ , let  $P_\theta(x) = \langle \theta, x \rangle$ . The sliced  $p$ -Wasserstein distance averages one-dimensional transport:

$$\text{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} W_p^p((P_\theta)_\# \mu, (P_\theta)_\# \nu) \, \text{d}\sigma(\theta).$$

For  $\mu_n = \frac{1}{n} \sum_i \delta_{x_i}$  and  $\nu_n = \frac{1}{n} \sum_i \delta_{y_i}$ , each slice only requires sorting:

$$W_p^p((P_\theta)_\# \mu_n, (P_\theta)_\# \nu_n) = \frac{1}{n} \sum_{i=1}^n \left| \langle \theta, x \rangle_{(i)} - \langle \theta, y \rangle_{(i)} \right|^p.$$

## Why it is competitive

- $K$  random slices cost  $O(Kn \log n)$ ;
- every subproblem is one-dimensional;
- $\text{SW}_p$  is a genuine metric.

## The price of slicing

In high dimension, many directions are needed, and the resulting loss does not recover a single transport coupling in the ambient space.

# Balanced transport cannot change mass

Classical optimal transport searches over  $\Pi(\alpha, \beta)$ , hence every admissible coupling satisfies

$$\gamma_0 = \alpha, \quad \gamma_1 = \beta, \quad |\alpha| = |\beta|.$$

## A possible bottleneck

- Different total masses make  $\Pi(\alpha, \beta)$  empty.
- Outliers must be transported somewhere, however far away.
- Appearance, disappearance and partial observations are excluded.

## Combine mass transport and mass growth

Replace exact marginals by a tradeoff:

$$\text{transport} \longleftrightarrow \text{creation / removal}.$$

The discrepancy between marginals determines the price of reaction.

The coupling is no longer forced to explain every observation.

# Static UOT softens marginal constraints

For positive measures, use the extended relative entropy

$$\text{KL}(\eta \mid \alpha) = \int_X (r \log r - r + 1) \, d\alpha, \quad r = \frac{d\eta}{d\alpha},$$

with value  $+\infty$  when  $\eta \not\ll \alpha$ . Then define

$$\text{UOT}_{c, \tau_0, \tau_1}(\alpha, \beta) = \inf_{\gamma \geq 0} \left\{ \int_{X \times Y} c \, d\gamma + \tau_0 \text{KL}(\gamma_0 \mid \alpha) + \tau_1 \text{KL}(\gamma_1 \mid \beta) \right\}.$$

## The transported submass

$\gamma_0$  and  $\gamma_1$  describe the mass that is actually matched. Their deviations from  $\alpha$  and  $\beta$  are penalized, not forbidden.

## Interpolates between optimal transport and Hellinger distance.

- $\tau_i \rightarrow \infty$ : recover hard marginal constraints;
- small  $\tau_i$ : reaction can be cheaper than long transport.

Chizat, Peyré, Schmitzer and Vialard, *An interpolating distance between optimal transport and Fisher–Rao*, FoCM 2018.

# Dynamic UOT adds a reaction direction

Allow the continuity equation to have a source term:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = a_t \rho_t, \quad \rho_{t=0} = \alpha, \quad \rho_{t=1} = \beta.$$

The Wasserstein–Fisher–Rao action is

$$\text{WFR}_\delta^2(\alpha, \beta) = \inf_{\rho, v, a} \int_0^1 \int_X (|v_t(x)|^2 + \delta^2 |a_t(x)|^2) \, d\rho_t(x) \, dt.$$

## Two tangent directions

- $v_t$ : advects mass in space;
- $a_t$ : locally grows or removes mass;
- $\delta$ : sets the transport–reaction length scale.

## Convex momentum variables

With  $m = \rho v$  and  $s = \rho a$ ,

$$\partial_t \rho + \nabla \cdot m = s, \quad \int \frac{|m|^2 + \delta^2 s^2}{\rho}.$$

The constraint is linear and the action is convex and 1-homogeneous.



# WFR/HK joins transport and Fisher–Rao I

For KL marginal penalties and a truncated-cosine transport cost, the static entropy–transport problem equals the dynamic WFR problem. It defines the Hellinger–Kantorovich metric on finite positive measures.

Yet a very strange transport cost!: it is a cone type of cost.

With the preceding normalization, let  $\cos_+(r) = \cos(\min\{r, \pi/2\})$ . Then

$$\text{WFR}_\delta^2(a\delta_x, b\delta_y) = 4\delta^2 \left( a + b - 2\sqrt{ab} \cos_+ \left( \frac{d(x, y)}{2\delta} \right) \right).$$

# WFR/HK joins transport and Fisher–Rao II

## Nearby particles

Transport and mass rescaling cooperate. At  $x = y$ ,

$$\text{WFR}_\delta(a\delta_x, b\delta_x) = 2\delta|\sqrt{a} - \sqrt{b}|,$$

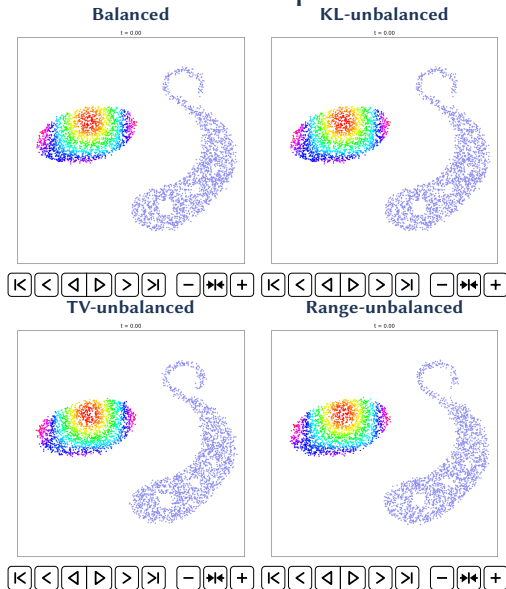
the Fisher–Rao/Hellinger geometry.

## Particles beyond the cut locus

When  $d(x, y) \geq \pi\delta$ , matching is no longer useful: the geodesic destroys mass at  $x$  and creates it at  $y$ .

Liero, Mielke and Savaré, *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance*, Invent. Math. 2018.

# Balanced and unbalanced transport flows

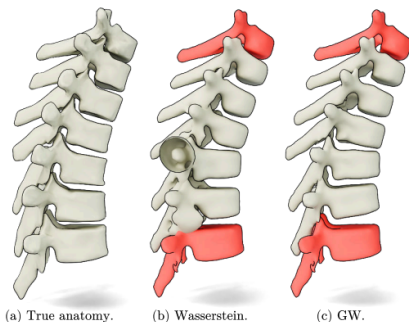


Balanced transport preserves all mass. Unbalanced flows may remove or create mass when displacement is too expensive.

# Wasserstein barycenters are Fréchet means

For weights  $\lambda_i \geq 0$  with  $\sum_i \lambda_i = 1$ , define

$$\text{Bar}_W(\mu_1, \dots, \mu_N) \in \operatorname{argmin}_{\rho \in \mathcal{P}_2(X)} \sum_{i=1}^N \lambda_i W_2^2(\rho, \mu_i).$$

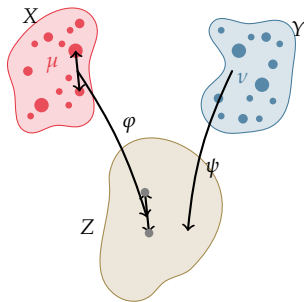


- Wasserstein barycenters average measures in a common ground space.
- GW barycenters instead average intrinsic metric structures.

# Sturm: embed the spaces, then transport

Let  $\mathbb{X} = (X, d_X, \mu)$  and  $\mathbb{Y} = (Y, d_Y, \nu)$  be metric-measure spaces. Sturm places isometric copies of both spaces in a common metric space  $(Z, d_Z)$  and applies classical optimal transport:

$$D_2^2(\mathbb{X}, \mathbb{Y}) = \inf_{\substack{(Z, d_Z), \varphi, \psi \\ \varphi, \psi \text{ isometric}}} W_2^2(\varphi_{\#}\mu, \psi_{\#}\nu).$$



## The cross-space geometry is part of the optimization

Choosing  $(Z, \varphi, \psi)$  is equivalent to choosing a metric on  $X \sqcup Y$  that restricts to  $d_X$  and  $d_Y$ . The induced cross-distance then supplies the cost needed by OT.

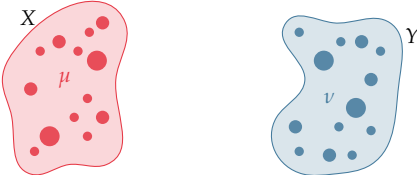
Sturm optimizes an ambient geometry before transporting mass.

Sturm, *Acta Mathematica*, 2006.

# Mémoli: minimize relational distortion

Mémoli keeps the spaces separate. A coupling  $\pi \in \Pi(\mu, \nu)$  proposes a soft correspondence, and two independent pairs drawn from  $\pi$  reveal how well it preserves distances:

$$\text{GW}_2^2(\mathbb{X}, \mathbb{Y}) = \inf_{\pi \in \Pi(\mu, \nu)} \iint |d_X(x, x') - d_Y(y, y')|^2 d\pi(x, y) d\pi(x', y').$$


$$\begin{aligned} &\Rightarrow (X \times X, \mu \otimes \mu) && \Rightarrow (Y \times Y, \nu \otimes \nu) \\ &\text{GW} = \inf_{\tilde{\pi}} \int |d_X - d_Y|^2 d\tilde{\pi} \\ &\text{under the } \textit{non-convex} \text{ constraint } \tilde{\pi} = \pi \otimes \pi \end{aligned}$$

Mémoli optimizes a correspondence and never constructs a common ambient space.

*Mémoli, Foundations of Computational Mathematics, 2011.*

# Computing Mémoli's GW distance

For discrete measures, write  $C_{ij}^X = c_X(x_i, x_j)$  and  $C_{kl}^Y = c_Y(y_k, y_\ell)$ . Introduce the bilinear form

$$G(\pi_1, \pi_2) := \sum_{i,j,k,\ell} |C_{ij}^X - C_{kl}^Y|^2 (\pi_1)_{ik} (\pi_2)_{j\ell}.$$

The quadratic Gromov–Wasserstein problem is

$$\text{GW}^2(\mathbb{X}, \mathbb{Y}) = \min_{\pi \in \Pi(\mu, \nu)} G(\pi, \pi).$$

## Linearize by freezing one copy of the coupling

Fixing  $\pi_2$  makes  $G(\pi_1, \pi_2)$  linear in  $\pi_1$ . The update of  $\pi_1$  is therefore a classical OT problem, often solved with entropic regularization and Sinkhorn iterations. Repeating this step gives an alternating or majorization scheme for the nonconvex GW objective.

## Computational complexity

A direct four-index contraction costs  $O(N^4)$  for two  $N$ -point spaces. For the squared loss, expanding  $|C_{ij}^X - C_{kl}^Y|^2$  reduces the interaction term to matrix multiplications, with a dense cost of  $O(N^3)$  per update.

**Scalability remains the central difficulty.**

Scetbon et al., *Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs*.

# CNT costs admit Hilbert embeddings

A symmetric cost  $c : X \times X \rightarrow \mathbb{R}_+$  with  $c(x, x) = 0$  is of *conditionally negative type* (CNT) when

$$\sum_i a_i = 0 \implies \sum_{i,j} a_i a_j c(x_i, x_j) \leq 0.$$

## Schoenberg's representation

The CNT condition is equivalent to the existence of a Hilbert space  $\mathcal{H}$  and a map  $\phi : X \rightarrow \mathcal{H}$  such that

$$c(x, x') = \|\phi(x) - \phi(x')\|_{\mathcal{H}}^2.$$

Thus a nonlinear cost becomes a squared Euclidean distance, possibly in infinite dimension.

## A broad and useful class

- $\|x - x'\|^p$  on  $\mathbb{R}^d$ , for  $0 < p \leq 2$ ;
- transformed costs such as  $1 - e^{-c/\ell}$  when  $c$  is CNT;
- $e^{-c/\varepsilon}$  is positive definite for every  $\varepsilon > 0$ .

CNT is the common structural thread between kernels, Sinkhorn divergences and scalable GW.



# CNT makes the GW expansion tractable

Embed  $c_X, c_Y$  as squared Hilbert distances and center the features:

$$\int \phi \, d\alpha = 0, \quad \int \psi \, d\beta = 0, \quad M_\pi := \int \phi(x) \otimes \psi(y) \, d\pi(x, y).$$

Set  $q_X = \int \|\phi\|^2 d\alpha$ ,  $q_Y = \int \|\psi\|^2 d\beta$  and  $S_\pi = \int \|\phi(x)\|^2 \|\psi(y)\|^2 d\pi$ . Then

$$\iint c_X(x, x') c_Y(y, y') \, d(\pi \otimes \pi) = 2q_X q_Y + 2S_\pi + 4\|M_\pi\|_{\text{HS}}^2.$$

## Only two coupling-dependent statistics remain

For the quadratic loss, all other terms depend only on the marginals:

$$\text{GW}^2(\mathbb{X}, \mathbb{Y}) = \inf_{\pi \in \Pi(\alpha, \beta)} \left\{ C_{\alpha, \beta} - 4S_\pi - 8\|M_\pi\|_{\text{HS}}^2 \right\}.$$

The Euclidean matrix expansion survives for every CNT cost.

# Entropic CNT–GW becomes repeated OT

Add  $\varepsilon \text{KL}(\pi \mid \alpha \otimes \beta)$  and dualize the negative quadratic term with a Hilbert–Schmidt operator  $\Gamma$ :

$$-8\|M_\pi\|_{\text{HS}}^2 = \min_{\Gamma} \left\{ 8\|\Gamma\|_{\text{HS}}^2 - 16\langle \Gamma, M_\pi \rangle_{\text{HS}} \right\}.$$

Consequently,

$$\text{GW}_\varepsilon^2 = C_{\alpha,\beta} + \min_{\Gamma} \left\{ 8\|\Gamma\|_{\text{HS}}^2 + \text{OT}_\varepsilon^{c_\Gamma}(\alpha, \beta) \right\},$$

where the inner problem is ordinary entropic OT with

$$c_\Gamma(x, y) = -4\|\phi(x)\|^2\|\psi(y)\|^2 - 16\langle \Gamma, \phi(x) \otimes \psi(y) \rangle_{\text{HS}}.$$

## Algorithmic consequence

Alternate an entropic OT solve in  $\pi$  with an update of  $\Gamma$ . In finite-dimensional feature spaces,  $\Gamma$  is a matrix.

## Beyond squared norms

CNT embeddings extend the same mechanism to nonlinear costs; kernel PCA supplies finite-rank approximations when needed.

Mathematical formulations behind generative models

# WGAN architecture: generator and critic

Let  $\mu \in \mathcal{P}(X)$  be the data distribution and  $\rho \in \mathcal{P}(Z)$  a simple latent distribution. The generator defines

$$Z \sim \rho \xrightarrow{G_\theta} \tilde{X} = G_\theta(Z) \sim \nu_\theta, \quad \nu_\theta = (G_\theta)_\# \rho.$$

## Generator $G_\theta : Z \rightarrow X$

$z \in \mathbb{R}^q \mapsto$  dense layer and reshape  $\mapsto$  upsampling blocks  $\mapsto \tilde{x} \in X$ .

For images, the upsampling blocks are convolutional or residual. Sampling requires one forward pass through this decoder-style network.

## Critic $f_\phi : X \rightarrow \mathbb{R}$

$x \in X \mapsto$  downsampling blocks  $\mapsto$  features  $\mapsto f_\phi(x) \in \mathbb{R}$ .

The same network scores real and generated samples. Its scalar output has no sigmoid: it estimates a Kantorovich potential, not a probability.

The critic supplies the loss; its gradient is propagated through the generator.

# Wasserstein-1 gives the adversarial objective

Kantorovich–Rubinstein duality turns transport into an optimization over scalar functions:

$$W_1(\mu, \nu_\theta) = \sup_{\text{Lip}(f) \leq 1} \{ \mathbb{E}_{X \sim \mu} [f(X)] - \mathbb{E}_{Z \sim \rho} [f(G_\theta(Z))] \}.$$

Replacing  $f$  by the critic network  $f_\varphi$  gives the sample-based game

$$\min_{\theta} \max_{\varphi: f_\varphi \text{ Lipschitz}} \left\{ \frac{1}{m} \sum_{i=1}^m f_\varphi(X_i) - \frac{1}{m} \sum_{i=1}^m f_\varphi(G_\theta(Z_i)) \right\}.$$

## Critic update

Increase the score gap between real and generated minibatches, subject to a Lipschitz bound on  $f_\varphi$ .

## Generator update

Hold the critic fixed and minimize  $-m^{-1} \sum_i f_\varphi(G_\theta(Z_i))$  by backpropagation through  $f_\varphi \circ G_\theta$ .

Arjovsky, Chintala and Bottou, *Wasserstein Generative Adversarial Networks*, ICML 2017.

# Original WGAN: Lipschitzness by weight clipping

The original algorithm constrains every critic parameter after each ascent step:

$$\varphi \leftarrow \text{clip}(\varphi, -c, c).$$

Hence the parameter set is compact. For a fixed feedforward architecture with 1-Lipschitz activations,

$$\text{Lip}(f_\varphi) \leq \prod_{\ell=1}^L \|W_\ell\|_{\text{op}} \leq K(c, \text{layer widths}).$$

## Alternating training

Perform several critic ascent steps, clipping after every update. Then take one generator descent step using fresh latent samples. The critic is trained more often so that it remains close to the inner maximizer.

## What clipping actually guarantees

Weight clipping gives a uniform  $K$ -Lipschitz family. It is not an exact projection onto  $\{f : \text{Lip}(f) \leq 1\}$ , and a small clipping range can severely restrict the critic. WGAN-GP later replaced clipping by a penalty on the input-gradient norm.

WGAN keeps one-pass generation, but training remains an alternating two-network problem.  
Arjovsky, Chintala and Bottou, ICML 2017; Gulrajani et al., NeurIPS 2017.

# Diffusion models split generation into regressions

Introduce a path of progressively smoother laws between the data and a simple noise distribution:

$$\rho_0 = \mu \longrightarrow \rho_1 \longrightarrow \cdots \longrightarrow \rho_T \simeq \gamma.$$

## Training: many small supervised problems

At a randomly selected noise level  $t$ , corrupt a data sample using a known random mechanism. The clean sample and the added noise are both known, so a time-conditioned neural network can learn the local denoising direction by least squares.

## Generation: reverse the small steps

Start from  $X_T \sim \gamma$  and successively apply the learned reverse updates until reaching the data scale. Each update only removes a small amount of noise.

The main optimization idea is to replace one difficult distribution-matching problem by regression problems at many noise levels.

Sohl-Dickstein et al., ICML 2015; Ho, Jain and Abbeel, NeurIPS 2020.

# Heat flow as advection by the score

Let  $\rho_0 = \mu$  and evolve it by the heat equation on  $\mathbb{R}^d$ :

$$\partial_t \rho_t = \Delta \rho_t = \nabla \cdot (\rho_t \nabla \log \rho_t), \quad \rho_t = \rho_0 * \mathcal{N}(0, 2tI).$$

Written as a continuity equation,

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \quad v_t = -\nabla \log \rho_t.$$

## Forward: data become noise

The score  $\nabla \log \rho_t$  points toward regions of high density, so the velocity  $-\nabla \log \rho_t$  spreads the mass. Increasing  $t$  progressively erases the fine structure of the data distribution.

## Backward: the score rebuilds structure

Reversing time changes the sign of the velocity. Starting from the smoothed terminal law requires the vector field  $+\nabla \log \rho_t$  at every noise level.

**The heat-flow picture identifies the quantity to learn: the score.**

F.-X. Vialard, *A word on generative diffusion models for the outsider mathematician*, 2025.



# Tweedie's formula turns the score into regression

Draw independent random variables

$$X \sim \mu, \quad Z \sim \mathcal{N}(0, I), \quad Y = X + \sigma Z,$$

and denote the density of  $Y$  by  $\rho_\sigma = \mu * \varphi_\sigma$ . Differentiating the Gaussian convolution and conditioning on  $Y = y$  gives

$$\nabla \log \rho_\sigma(y) = \frac{\mathbb{E}[X \mid Y = y] - y}{\sigma^2}.$$

## The random training pair is explicit

Sample a data point  $X$ , sample fresh Gaussian noise  $Z$ , and form the corrupted observation  $Y = X + \sigma Z$ . Both the input  $Y$  and target  $X$  are known during training.

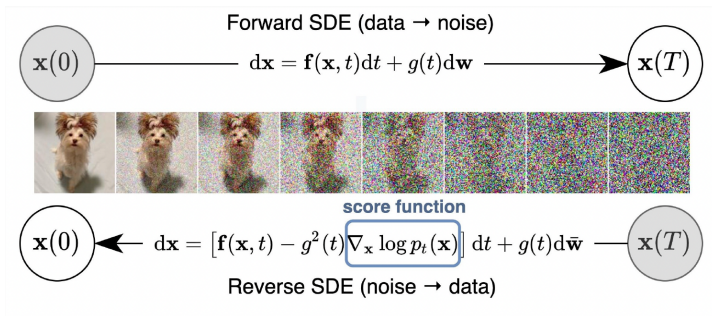
## Least squares estimates the score

$$\min_D \mathbb{E} \|D(Y) - X\|^2 \implies D^*(y) = \mathbb{E}[X \mid Y = y],$$

hence  $\nabla \log \rho_\sigma(y) = (D^*(y) - y)/\sigma^2$ . One network can handle all  $\sigma$  by taking the noise level as an additional input.

Denoising regression is score estimation in disguise.

# DDPM: an exact Gaussian noising chain



With  $\alpha_t = 1 - \beta_t$ , the forward chain is

$$X_t = \sqrt{\alpha_t} X_{t-1} + \sqrt{\beta_t} \varepsilon_t, \quad \varepsilon_t \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I).$$

Writing  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$  gives the exact marginal

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I).$$

Training samples  $(X_0, t, \varepsilon)$ , constructs  $X_t$ , and minimizes

$$\mathbb{E} \|\varepsilon - \varepsilon_\theta(X_t, t)\|^2, \quad s_\theta(X_t, t) = -\frac{\varepsilon_\theta(X_t, t)}{\sqrt{1 - \bar{\alpha}_t}}.$$

Illustration: Yang Song, Ho, Jain and Abbeel, *Denoising Diffusion Probabilistic Models*, NeurIPS 2020.

# Flow matching: regress a velocity field I

Let  $\rho_0 = \mathcal{N}(0, I)$  be the source and  $\rho_1 = \mu$  the data law. Choose a coupling  $\pi \in \Pi(\rho_0, \rho_1)$ , sample  $(X_0, X_1) \sim \pi$ , and lift the endpoints to a path, for instance

$$X_t = (1 - t)X_0 + tX_1, \quad \dot{X}_t = X_1 - X_0, \quad t \in [0, 1].$$

Denote the law of  $X_t$  by  $\rho_t$ .

## From random paths to an Eulerian flow

Average the path velocities that meet at  $x$ :

$$v_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x].$$

For every smooth test function  $f$ ,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[f(X_t)] &= \mathbb{E}[\langle \nabla f(X_t), \dot{X}_t \rangle] \\ &= \int \langle \nabla f(x), v_t(x) \rangle d\rho_t(x). \end{aligned}$$

Hence  $\partial_t \rho_t + \operatorname{div}(\rho_t v_t) = 0$ .

# Flow matching: regress a velocity field II

## Regression, then generation

Sample also  $t \sim \text{Unif}[0, 1]$  and solve

$$\min_{\theta} \mathbb{E}[\|v_{\theta}(X_t, t) - \dot{X}_t\|^2].$$

Least squares recovers the conditional mean  $v_t$ . To generate, integrate the learned ODE

$$\dot{Y}_t = v_{\theta}(Y_t, t), \quad Y_0 \sim \rho_0,$$

so that  $Y_1$  approximately follows  $\rho_1$ .

The coupling and path interpolation are design choices; the velocity is learned by regression on a finite time interval.

Lipman, Chen, Ben-Hamu, Nickel and Le, *Flow Matching for Generative Modeling*, ICLR 2023.

# Drifting models move the iteration to training

Let  $f_{\theta_i}$  be a one-pass generator and  $q_i = (f_{\theta_i})_{\#} p_{\varepsilon}$  its current output law. Training evolves  $q_i$  toward the data law  $p$ ; inference uses only  $\varepsilon \mapsto f_{\hat{\theta}}(\varepsilon)$ .

## Idea 1: drifting field

For a law  $r$ , define the kernel-weighted field

$$V_r(x) = \frac{\mathbb{E}_{Y \sim r}[k(x, Y)(Y - x)]}{\mathbb{E}_{Y \sim r}[k(x, Y)]}, \quad V_{p,q}(x) = V_p(x) - V_q(x).$$

Generated samples are attracted toward nearby data and repelled from nearby generated samples. Since  $V_{p,p} = 0$ , distribution matching is an equilibrium.

## Idea 2: step-by-step fitting

Set  $x_i = f_{\theta_i}(\varepsilon)$ , freeze the drifted target, and regress:

$$\tilde{x}_i = \text{sg}[x_i + \eta V_{p,q_i}(x_i)], \quad \mathcal{L}_i(\theta) = \mathbb{E}_{\varepsilon} \|f_{\theta}(\varepsilon) - \tilde{x}_i\|^2.$$

The optimizer updates the generator one local fitting problem at a time.

**The distribution evolves during training; the learned map gives one-pass generation.**

Original paper: Mingyang Deng, He Li, Tianhong Li, Yilun Du and Kaiming He, *Generative Modeling via Drifting*, arXiv:2602.04770, 2026.

A personal viewpoint

# Design the flow first, then regress it

Let  $\rho_s$  be an easy source law,  $\rho_t$  the target, and  $f_\theta$  a one-pass generator:

$$\rho_\theta = (f_\theta)_\# \rho_s, \quad \mathcal{L}_D(\theta) = F(\rho_\theta), \quad F(\rho) = D(\rho, \rho_t).$$

## 1. Design the flow

Work first in  $\mathcal{H} = L^2(\rho_s; \mathbb{R}^d)$  and lift the objective:

$$\tilde{F}(f) := F(f_\# \rho_s).$$

Choose a map-space evolution

$$\dot{f}_t = V[f_t] \circ f_t$$

whose geometry and convergence can be understood independently of a neural parametrization.

## 2. Regress the trajectory

At step  $k$ , turn the desired increment into a frozen target:

$$f_{\theta_k} \longmapsto \text{sg}[f_{\theta_k} + \tau V_k \circ f_{\theta_k}].$$

The network learns to reproduce this motion while retaining one-pass inference after training.

Separate the design of the dynamics from its approximation by a neural network.  
Dumont, Lacombe and Vialard, *Gauss-Newton drifting extends natural gradient descent*, 2026.

# One Euclidean step loses the map-space geometry

Let  $\overline{\nabla}F(\rho)$  denote the Wasserstein gradient field of  $F$ . Differentiating the direct objective gives

$$\nabla_{\theta} \mathcal{L}_D(\theta) = \int (D_{\theta} f_{\theta}(z))^{\top} \overline{\nabla}F(\rho_{\theta})(f_{\theta}(z)) \, d\rho_s(z).$$

Now freeze the target and define the stepwise regression

$$\mathcal{L}_{\text{drift},k}(\theta) := \frac{1}{2} \|f_{\theta} - \text{sg}[f_{\theta_k} + \tau V_k \circ f_{\theta_k}]\|_{L^2(\rho_s)}^2.$$

For a gradient drift, the two parameter gradients coincide

If  $V_k = -\overline{\nabla}F(\rho_{\theta_k})$ , then

$$\nabla_{\theta} \mathcal{L}_{\text{drift},k}(\theta_k) = \tau \nabla_{\theta} \mathcal{L}_D(\theta_k).$$

## Consequence

One Euclidean gradient step on the regression is merely Euclidean gradient descent on  $\mathcal{L}_D$ , up to the step size. Convexity of  $F$  in measure space therefore gives no automatic convexity, nor convergence guarantee, in the parameters  $\theta$ .

The stop-gradient target fixes the field; the optimizer decides how that field is projected onto the model.



# Gauss–Newton performs the regression in map space I

Linearize the generator at  $\theta_k$  and solve for the desired map increment:

$$\delta\theta_k \in \operatorname{argmin}_{\delta\theta} \|D_{\theta}f_{\theta_k} \delta\theta - \tau V_k \circ f_{\theta_k}\|_{L^2(\rho_s)}^2.$$

The normal equations involve the pullback metric

$$G_{\theta} := \int (D_{\theta}f_{\theta}(z))^{\top} D_{\theta}f_{\theta}(z) \, \mathrm{d}\rho_s(z).$$

# Gauss–Newton performs the regression in map space II

## Gradient case: natural descent

For  $V_k = -\overline{\nabla}F(\rho_{\theta_k})$ ,

$$\delta\theta_k = -\tau G_{\theta_k}^\dagger \nabla_\theta \mathcal{L}_D(\theta_k).$$

Thus Gauss–Newton drifting is natural gradient descent for the  $L^2(\rho_s)$  geometry of maps.

## Why this is different

- It finds the best tangent approximation of the designed flow.
- $G_\theta^\dagger$  handles redundant parameters.
- $(G_\theta + \lambda I)^{-1}$  gives a stable Levenberg–Marquardt variant.
- The same projection works for non-gradient drifts.

Natural gradient is the geometry induced by changing the generated map, not by changing its coordinates.

# Conclusion

- Modeling power must be balanced against computational efficiency and scalability.
- Mathematical structures help formulate new models.
- Continuous and discrete representations and algorithms inform one another.