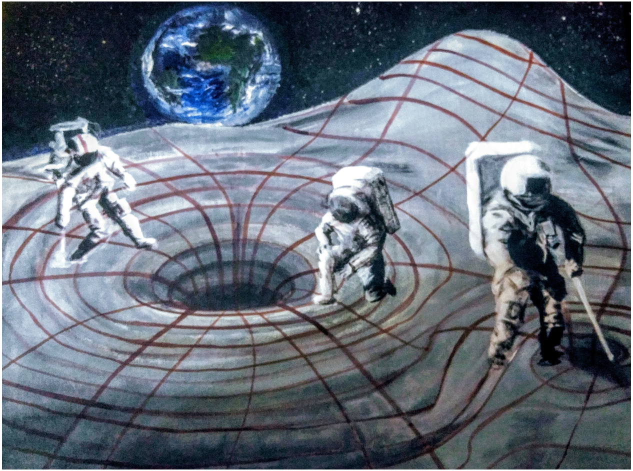
An introduction to optimal transport

Caroline Moosmüller

University of North Carolina at Chapel Hill
Department of Mathematics

Women in Optimal Transport, April 17 – 19, 2024

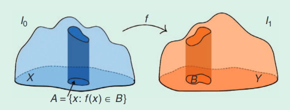
Earth Mover's Distance



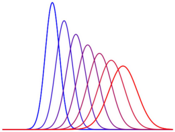
"Earth Mover's Distance" by Fana Hagos (Visual Arts undergraduate student, UCSD 2020)

Optimal transport

Analysis



Geometry

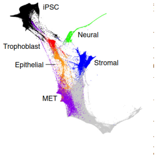


PDE

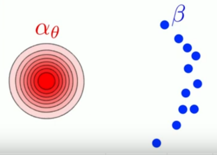
$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0$$

Data science/ML

Applications

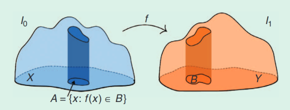


Statistics

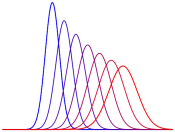


Optimal transport

Analysis



Geometry

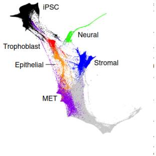


PDE

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{V}) = 0$$

Data science/ML

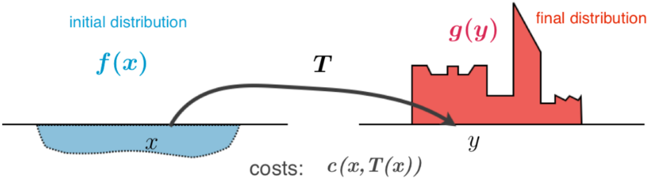
Applications



Statistics

Outline

Moving mass: The Monge problem



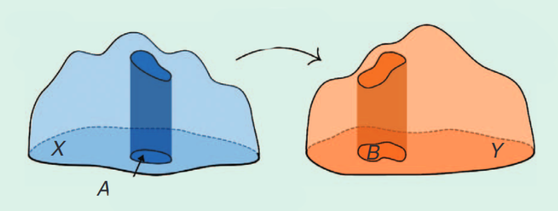
- Move “mass” f to g
- f, g are **probability densities** $\int_{\mathbb{R}^n} f(x)\,dx = \int_{\mathbb{R}^n} g(y)\,dy = 1$
- Find map $T : \mathbb{R}^n \to \mathbb{R}^n$ with **mass conservation**:

$$\int_A g(y)\,dy = \int_{T^{-1}(A)} f(x)\,dx, \quad A \subseteq \mathbb{R}^n,$$

  or equivalently $g(T(x))|\det(DT(x))| = f(x)$ for $x \in \mathbb{R}^n$
- There may be many such maps ... Find one with minimal work

Monge formulation: $\displaystyle \min_T \int_{\mathbb{R}^n} c(x, T(x))\, f(x)\,dx.$

Moving mass: The Monge problem



- More general: Consider **measures** μ and ν
- If μ is absolutely continuous (w.r.t. Lebesgue measure), then it has a density

$$\mu(A) = \int_A f(x)\,dx, \quad A \subseteq \mathbb{R}^n.$$

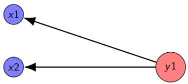
- $T : \mathbb{R}^n \to \mathbb{R}^n$ with **mass conservation** becomes

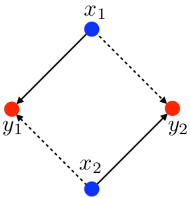
$$\nu = T_\sharp \mu, \quad T_\sharp\mu(A) = \mu(T^{-1}(A)), \quad A \subseteq \mathbb{R}^n.$$

- The **Monge problem** becomes

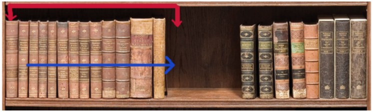
$$\min_{T : T_\sharp \mu = \nu} \int_{\mathbb{R}^n} c(x, T(x))\, d\mu(x).$$

- **Question 1:** What cost function c?
  $\rightarrow$ depends on the problem. Usually $c(x, y) = \|x - y\|^p, p \geq 1$; or geodesic distance $d(x, y)$ if measures supported on manifold.
- **Question 2:** Existence and uniqueness of solution?
  $\rightarrow$ In general: No and no.





- **Example**: The choice of cost influences uniqueness



$c(x, T(x)) = |x - T(x)|$ vs. $|x - T(x)|^2$ (strictly convex)

Moving mass: Brenier's theorem

Theorem (Brenier 1987)

Assume

- μ, ν be two measures on $\mathbb{R}^n$ with μ **absolutely continuous** (has density)
- Consider the cost $c(x, y) = \|x - y\|^2$

Then

- there exists a **unique map** T with $T_\sharp \mu = \nu$ that solves Monge
- T is uniquely defined as the **gradient of a convex function** φ, i.e. $T = \nabla \varphi$, where φ is the unique (up to constants) function with $(\nabla \varphi)_\sharp \mu = \nu$.

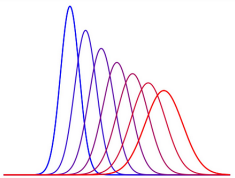
- Generalizations to other cost functions; Riemannian manifolds
- Note that with $T = \nabla \varphi$ the mass conservation property becomes the **Monge-Ampére equation**:

$$g(\nabla \varphi(x)) |\det(D^2 \varphi(x))| = f(x)$$

Convexity of φ leads to $D^2 \varphi(x) \geq 0$ is necessary for a solution.

Dynamic formulation

- Instead of looking for a (static) map T, we can try to **continuously move** from density f to g.



- Consider a path ρ_t with $\rho_0 = f$ and $\rho_1 = g$ and its velocity field v_t. **Conservation of mass** (continuity equation):

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

- Then find the pair (ρ_t, v_t) that minimizes the kinetic energy:

$$\textbf{dynamic formulation} = \min_{(\rho_t, v_t)} \int_0^1 \int_{\mathbb{R}^n} \|v_t(x)\|^2 \, d\rho_t(x) \, dt$$

- **Benamou-Brenier** (2000): If Monge solution exists, then *dynamic = Monge*, i.e. $\rho_t = ((1-t)\,\mathrm{id} + t\,T)_\sharp\, \rho_0$.

Outline

- Consider the space of (absolutely continuous) measures with finite 2-th moment $\mathcal{P}_2(\mathbb{R}^n) = \{\mu : \int_{\mathbb{R}^n} \|x\|^2 \, d\mu(x) < \infty\}$.
- The Monge/dynamic formulation define a **distance** on $\mathcal{P}_2(\mathbb{R}^n)$:

$$\begin{aligned} W_2^2(\mu,\nu) &= \min \left\{ \int_{\mathbb{R}^n} \|x - T(x)\|^2 \, d\mu(x) : T_\sharp \mu = \nu \right\} \\ &= \min \left\{ \int_0^1 \int_{\mathbb{R}^n} \|v_t(x)\|^2 \, d\rho_t(x) \, dt : (\rho_t, v_t) \text{ satisfy cont. equ} \right\} \\ &= \min \left\{ \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \, d\pi(x,y) : \pi \text{ has marginals } \mu, \nu \right\} \end{aligned}$$

- This is the **2-Wasserstein distance** or the **2-Monge-Kantorovich distance**. Also exists for other $p \geq 1$.
- The last formulation, is the Kantorovich formulation (more later).
- $\mathcal{P}_2(\mathbb{R}^n)$ has much more geometric structure. One can do (infinite dimensional) Riemannian-like geometry $\rightarrow$ F. Otto.

Geodesics

- The dynamic path ρ_t actually defines the **geodesic** from ρ_0 to ρ_1:

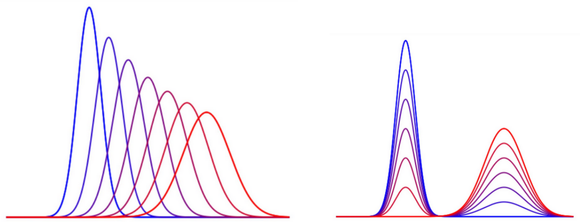
$$\rho_t = ((1-t)\,\mathrm{id} + t\,T)_\sharp\,\rho_0,$$

where T is the optimal Monge map.

- The geodesic is the "shortest path" in the sense of Riemannian geometry. It satisfies

$$W_2(\rho_s, \rho_t) = |s-t|\, W_2(\rho_0, \rho_1)$$

- Wasserstein vs. Euclidean path



Geodesics

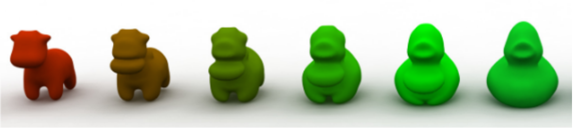
- The dynamic path ρ_t actually defines the geodesic from ρ_0 to ρ_1:

$$\rho_t = ((1-t)\,\mathrm{id} + t\,T)_\sharp\,\rho_0,$$

  where T is the optimal Monge map.
- The geodesic is the "shortest path" in the sense of Riemannian geometry. It satisfies

$$W_2(\rho_s, \rho_t) = |s-t|\, W_2(\rho_0, \rho_1)$$

- Geodesic between shapes

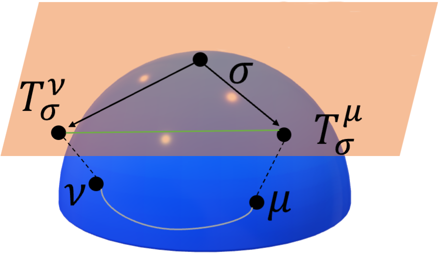


Tangent space

- Note that the geodesic path is **linear interpolation** in $L^2(\mathbb{R}^n, \rho_0)$ between id and T:

$$\rho_t = ((1-t)\,\mathrm{id} + t\,T)_\sharp\,\rho_0,$$

- $L^2(\mathbb{R}^n, \rho_0)$ is the **tangent space** at ρ_0. Monge maps $T = \nabla\varphi$ (or the velocity field v) are the "tangent vectors".



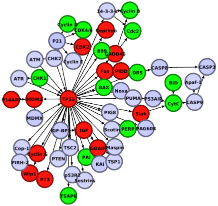
- We will use the tangent space later for **linearized OT**

Outline

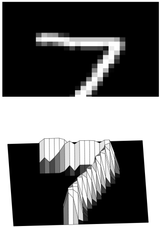
Data as point-clouds, histograms, densities

Bag-of-words

Gene expression data

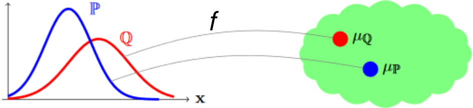


Images



- **Data**: Measures $\mu_k, k = 1, \ldots, N$ or points sampled from μ_k (point-cloud)
- **Compare and classify**: e.g. “Cancer” vs. “Healthy”
  - Supervised: Training data (μ_k, y_k), with classes $y_k \in \mathcal{C}$

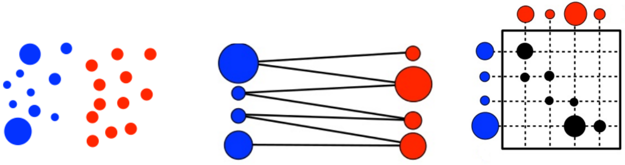
Learn a function



  - Unsupervised: Use $W_p(\mu_k, \mu_j) \rightarrow$ computational issues

Discrete measures: Kantorovich formulation

- Point-clouds/discrete measures: $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$:



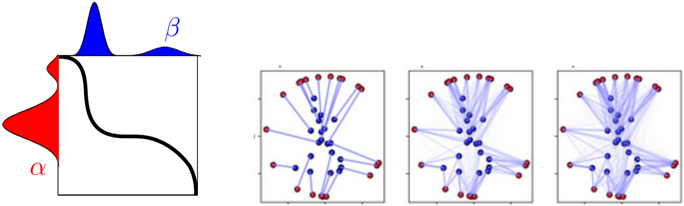
  with $a_i, b_j \geq 0$, $\sum a_i = \sum b_j = 1$ (probability vectors)
- Look for **coupling matrix** $P \in \mathbb{R}_+^{n \times m}$, where P_{ij} is the amount of mass moved from x_i to y_j. Mass can split!
- **Mass conservation**: $P1 = a, \quad P^T 1 = b$.
- **Kantorovich:** Find coupling matrix that minimizes work with given cost C_{ij}:

$$\min_P \sum_{ij} C_{ij} P_{ij} = \min_P \langle C, P \rangle$$

  Note this is a *linear* problem with linear constraints.
- **Cost:** Usually $C_{ij} = \|x_i - y_j\|^p$
- **Existence, Uniqueness:** Yes and no. $P = ab^T$ is feasible.

Discrete measures: Kantorovich formulation

- Kantorovich can also be formulated in continuous setting
- Kantorovich recovers Monge function in case it exists



- **Computation**: $\min_P \langle C, P \rangle$ is a linear program. Cost: $O(n^3 \log(n))$.
  $\rightarrow$ may be too slow for large data science problems.
- **Regularized version:** Provides approximate coupling & distance

$$\min_P \langle C, P \rangle - \varepsilon H(P)$$

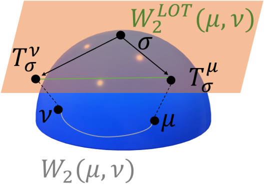
with $H(P) = -\sum P_{ij}(\log(P_{ij}) - 1)$ the entropy of P. This has a **unique solution** and can be solved in $O(n^2 \log(n))$ matrix scaling algorithms (Sinkhorn).

Think of transport coupling as a new set of features.

- **LOT embedding**: Pick a reference measure σ:

$$\begin{aligned} F_\sigma : \quad & \mathcal{P}(\mathbb{R}^n) \to L^2(\mathbb{R}^n, \sigma) \\ & \mu \mapsto T_\sigma^\mu \end{aligned}$$

- **Distance**: $W_2^{LOT}(\mu,\nu)^2 = \int_{\mathbb{R}^n} \| T_\sigma^\mu(x) - T_\sigma^\nu(x) \|^2 \, d\sigma(x)$

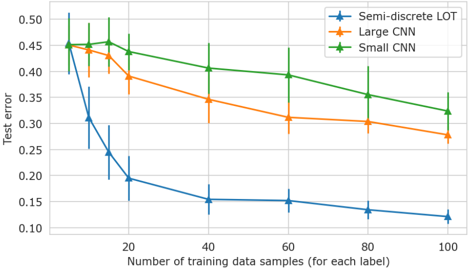


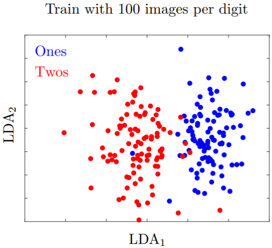
- **Learning**:

$$\begin{aligned} f_\mu : \quad & \mathcal{P}(\mathbb{R}^n) \to \mathcal{C} \\ & \mu \mapsto f(T_\sigma^\mu) \qquad \text{for } f : L^2(\mathbb{R}^n, \sigma) \to \mathcal{C} \end{aligned}$$

Learn a linear classifier in embedding space

MNIST Classification Between 7's and 9's





Learning in LOT embedding space

Theorem (Supervised learning in LOT (M., Cloninger 2023))

Let σ, τ_1, τ_2 absolutely continuous in $\mathcal{P}(\mathbb{R}^n)$, $\mathcal{H}$ convex set of ε-perturbations of elementary transformations. If

- $\mathcal{H}_\sharp \tau_1$, $\mathcal{H}_\sharp \tau_2$ compact, and
- minimal distance $W_2(h_{1\#}\tau_1, h_{2\#}\tau_2) > \delta$,

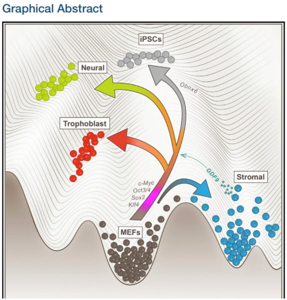
then $F_\sigma(\mathcal{H}_\sharp \tau_1)$ and $F_\sigma(\mathcal{H}_\sharp \tau_2)$ are linearly separable in $L^2(\mathbb{R}^d, \sigma)$.

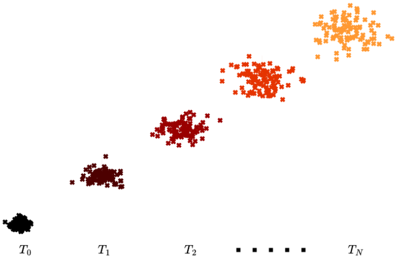
- Elementary transformations: Shifts, scalings, certain shearings
- δ can be given explicitly based on $\sigma, \tau_1, \tau_2, \varepsilon$.
- First version of this result by Rohde et. al. 2018 for $d = 1$ and $\varepsilon = 0$ ($\delta = 0$ in this case).
- Uses **Hahn-Banach theorem**. Key proof ingredient: Convexity of $\mathcal{H}$ is preserved via LOT.

Outline

Inferring cell trajectories

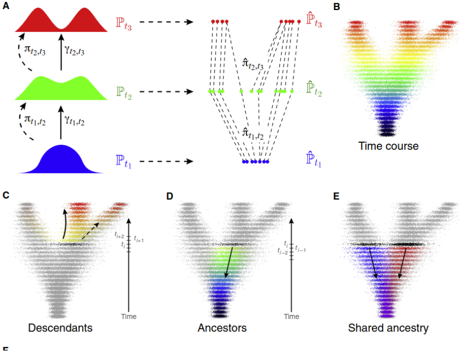
- Single cells are modeled as point-clouds in gene-expression space. Their “development” over time can be interpreted as a curve in Wasserstein space.



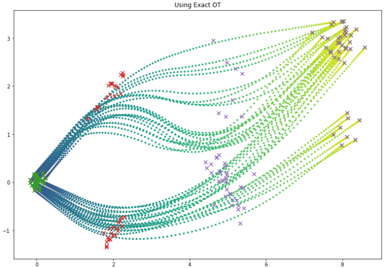


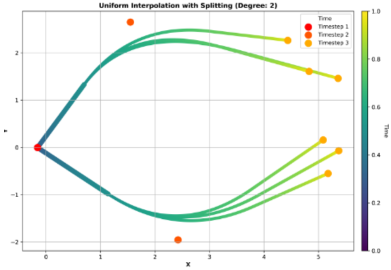
- Interpolate to e.g. understand development into certain cell types and identify responsible genes (reprogramming)

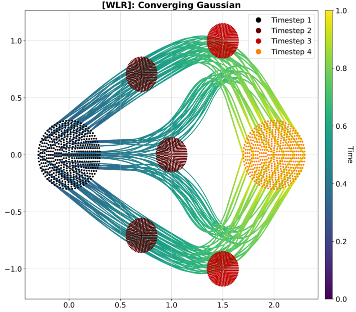
- Schiebinger et. al. original paper (2019): use **linear interpolation**

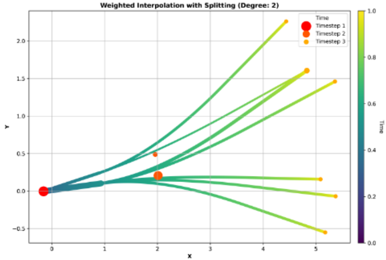


- To infer **smoother** trajectories, spline methods have been proposed.
- **New method**: spline-like, smooth, fast, intrinsic, and can deal with non-uniform mass and trajectory splitting (on arXiv soon!)


Using Exact OT


Uniform Interpolation with Splitting (Degree: 2)
Time
Timestep 1
Timestep 2
Timestep 3
x
Time


[WLR]: Converging Gaussian
Timestep 1
Timestep 2
Timestep 3
Timestep 4
Time


Weighted Interpolation with Splitting (Degree: 2)
Time
Timestep 1
Timestep 2
Timestep 3
x
y
Time

Thank you! - Questions?

OT papers

- G. Schiebinger et al. *Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming*, Cell 2019.
- M. Cuturi, G. Peyre *Computational optimal transport*, Foundations and Trends in Machine Learning, 2019.
- J. Solomon et. al. *Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains*, ACM Transactions on Graphics 2015.
- S. Kolouri et al. *Optimal Mass Transport: Signal processing and machine-learning applications*. IEEE signal process Mag 2017.
- M. Thorpe, *Introduction to Optimal Transport*, lecture notes 2018.

Our recent papers

- V. Khurana, H. Kannan, A. Cloninger, C. Moosmüller. *Learning sheared distributions using linearized optimal transport*, Sampling Theory, Signal Processing, and Data Analysis, 2023.
- A. Cloninger, K. Hamm, V. Khurana, C. Moosmüller, *Linearized Wasserstein dimensionality reduction with approximation guarantees*, arXiv 2023.
- C. Moosmüller, A. Cloninger. *Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations,* Information and Inference: A Journal of the IMA, 2023.
- S. Li, C. Moosmüller, *Measure transfer via stochastic slicing and matching*, arXiv 2023.