Foundations of Optimal Transport

Computational Optimal Transport in Imaging Science
SIAM Conference on Imaging Science

Matthew Thorpe and **Yunan Yang**

28th May 2024

The Department of
Mathematics

Society for Industrial and Applied Mathematics

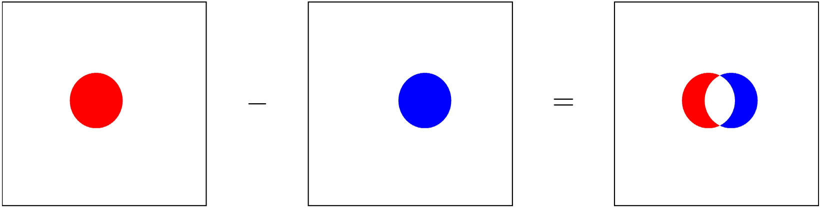
Office of Naval Research
Science & Technology

Contents

Contents

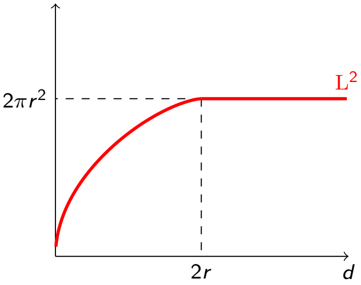
Euclidean Distances

A Euclidean distance considers a pixel-by-pixel difference.



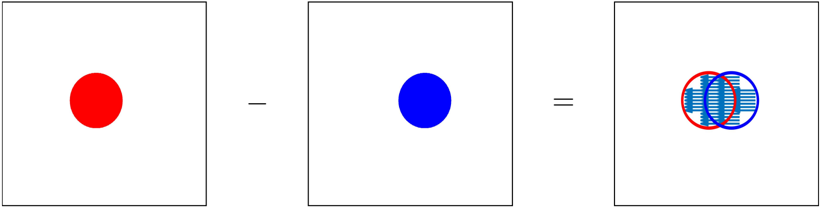
For example, the L^2 distance:

$$d_{\mathrm{L}^2}(f,g) = \sqrt{\int_\Omega |f(x) - g(x)|^2 \,\mathrm{d}x}.$$



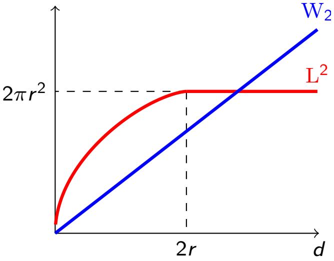
Lagrangian Distances

A Lagrangian distance considers the distance moved.



E.g., the Wasserstein distance measures
(1) the distance moved multiplied by
(2) the mass moved:

$$\mathrm{d}_{\mathrm{W}_2}(f, g) \sim \text{size of translation.}$$



Optimal Transport

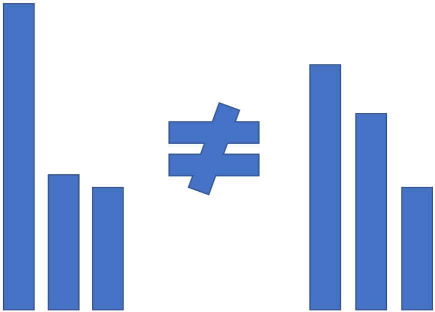


Figure: Synthetic data f (left) and observed data g (right)

Optimal Transport

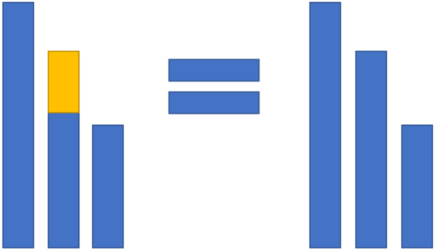


Figure: Synthetic data f (left) and observed data g (right)

Optimal Transport

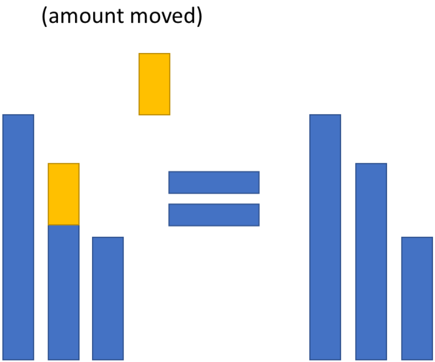


Figure: Synthetic data f (left) and observed data g (right)

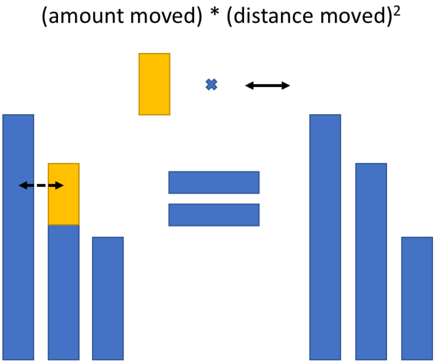
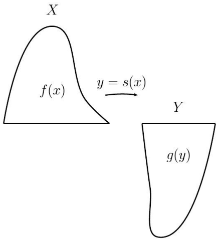


Figure: Synthetic data f (left) and observed data g (right)

A Brief History of Optimal Transport



Proposed by Monge in 1781

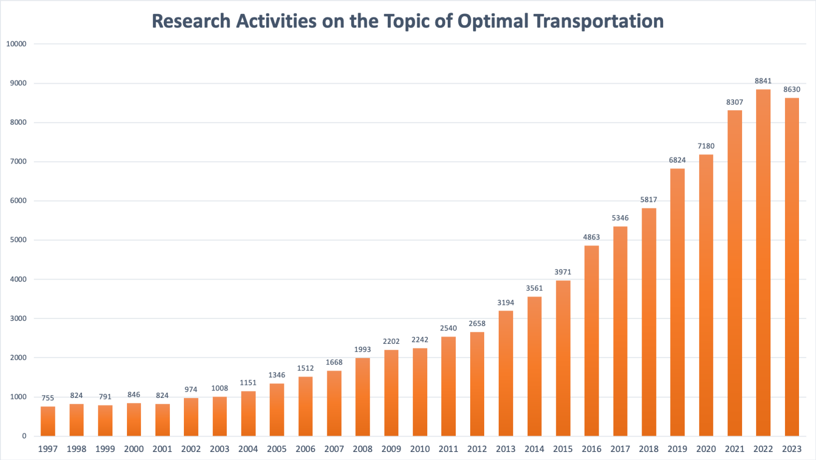
Mathematical Developments:

- Monge (1781)
- Kantorovich (1942)
- Brenier, Caffarelli, Gangbo, McCann, Benamou, Otto, Villani, Figalli, and many others (1990s - present)

Application Areas (see 2nd part):

- Data Assimilation
- Hyperbolic Model Reduction
- Image Processing
- Inverse Problems
- Machine Learning
- Sampling, and many more

The Development of Optimal Transport



The publication # containing keywords: ‘optimal transport’, ‘Wasserstein’, ‘Monge’, ‘Kantorovich’, or ‘earth mover’ according to Web of Sciences.

Contents

Rearranging Mass

1. Let μ and ν be two probability measures on spaces X and Y.
2. The cost is $c(x, y)$ to move one unit of mass from location $x \in X$ to location $y \in Y$.
3. The total cost of moving mass by $T : X \to Y$ is therefore

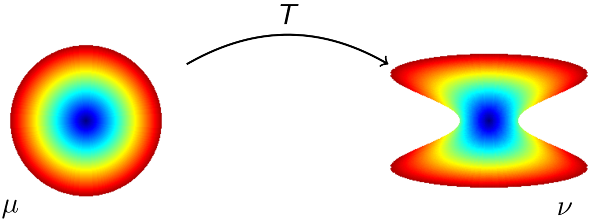
$$\int_X c(x, T(x))\,\mathrm{d}\mu(x).$$

4. **Where's the ν dependence?** We can't choose any T, after we move the mass $\mu(x)$ from x to $T(x)$, we obtain ν, i.e.,

$$\nu(B) = \mu(\underbrace{T^{-1}(B)}_{\{x \in X \colon T(x) \in B\}}) \quad \text{for all measurable sets } B \subset Y.$$

Transport Maps

If T satisfies $\nu(\cdot) = \mu(T^{-1}(\cdot))$ then we write $T_{\#}\mu = \nu$ and we say that T *pushes* μ *onto* ν.



- If $\mu = \sum_{i=1}^{N} w_i \, \delta_{x_i}$, then $\nu = T_{\#}\mu = \sum_{i=1}^{N} w_i \, \delta_{T(x_i)}$.
- If μ is absolutely continuous with respect to the Lebesgue measure, with probability density function $f(x)$, then ν is also a.c. with respect to Lebesgue with density $g(y)$.

$$\text{Change of Variable:} \quad g(T(x))|\det(\nabla T(x))| = f(x) \, .$$

Image Source: Thorpe, M., Park, S., Kolouri, S., Rohde, G. K., & Slepčev, D. (2017). A Transportation L^p Distance for Signal Analysis. Journal of mathematical imaging and vision, 59, 187-210.

Monge Form of Optimal Transport

1. There may be many such T satisfying $T_{\#}\mu = \nu$ (or there may also be none).
2. When there are many we choose the best possible T:

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

3. A special case is when $X = Y = \Omega$ and $c(x, y) = |x - y|^p$ which defines the p-**Wasserstein distance**:

$$\mathrm{d}_{\mathrm{W}_p}(\mu, \nu) = \left(\inf_{T \,:\, T_{\#}\mu=\nu} \int_\Omega |x - T(x)|^p \,\mathrm{d}\mu(x) \right)^{\frac{1}{p}}.$$

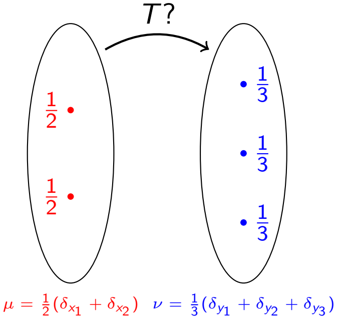
4. When $X = Y = \Omega$ and $c(x, y)$ is a metric then

$$\mathrm{d}_{\mathrm{EM}}(\mu, \nu) = \inf_{T \,:\, T_{\#}\mu=\nu} \int_\Omega c(x, T(x)) \,\mathrm{d}\mu(x)$$

is also a metric, called the **earth movers distance (EMD)**.

5. This formulation is called the $\boxed{\text{Monge form}}$.

Optimal Transport Between Diracs



Do transport maps always exist?

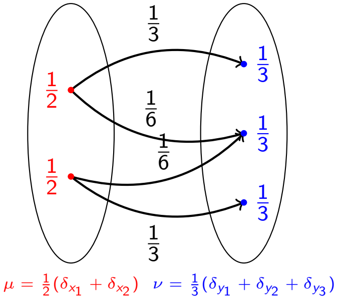
No!

Solution: Split mass.

Contents

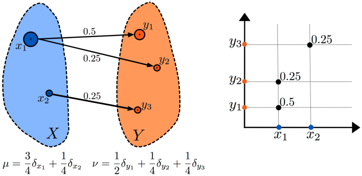
Kantorovich Form of Optimal Transport

1. Let $\pi(x, y)$ be the amount of mass that moves from x to y.
2. Total amount of mass leaving x should be $\pi(x, Y) = \mu(x)$.
3. Total amount of mass arriving at y should be $\pi(X, y) = \nu(y)$.
4. Let $\Pi(\mu, \nu)$ be the set of probability measures satisfying the above two conditions.
5. *Kantorovich form of OT*:



$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, \mathrm{d}\pi(x, y).$$

1. Kantorovich form of OT:
$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y)\, \mathrm{d}\pi(x,y).$$
2. If $\mu = \sum_{i=1}^{m} p_i \delta_{x_i}$, $\nu = \sum_{j=1}^{n} q_j \delta_{y_j}$, then we can view π as a matrix in the set $\Pi(\mu,\nu) \subset \mathbb{R}^{m \times n}$, with column sums equal to $(p_1, \ldots p_m)$ and row sums equal to $(q_1, \ldots, q_n)$.



3. The OT problem can be written as
$$\inf_{\pi \in \Pi(\mu,\nu)} \sum_{i,j} C_{ij}\, \pi_{ij},$$
where $C_{ij} = c(x_i, y_j)$.
4. This is a linear programme!

Image Source: Kolouri, S., Park, S. R., Thorpe, M., Slepčev, D., & Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. IEEE signal processing magazine, 34(4), 43-59.

Contents

1. ("$M \geq K$".) The Monge formulation of optimal transport provides an upper bound for the Kantorovich formulation.
2. (**When "$M = K$".**) When the OT plan can be written in the form $\pi = (\mathrm{Id} \times T)_{\#}\mu$, then T is an optimal transport map and the Monge and Kantorovich forms **coincide**.
3. (**Existence of OT Plans.**) Assume X, Y are Polish and $c : X \times Y \to [0, \infty)$ is lower semi-continuous. Then for any $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, there exists a solution to Kantorovich's problem. ("inf" $\Longrightarrow$ "min")
4. (**Kantorovich Duality.**) Under conditions as above, define $\Phi_c = \{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) : \varphi(x) + \psi(y) \leq c(x, y) \, \text{a.e.}\}$, then the Kantorovich's problem admits a dual problem:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) \, \mathrm{d}\pi(x, y) = \sup_{(\varphi,\psi) \in \Phi_c} \int_X \varphi \, \mathrm{d}\mu + \int_Y \psi \, \mathrm{d}\nu.$$

Brenier's Theorem

Theorem

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $c(x, y) = \frac{1}{2}|x - y|^2$. Assume μ does not give mass to small sets. Then,

1. *There is a unique solution $\pi^* \in \Pi(\mu, \nu)$ to Kantorovich's optimal transport problem.*
2. *There exists an $\mathrm{L}^1(\mu)$, convex, lower semi-continuous function φ^* such that $\pi^* = (\mathrm{Id} \times \nabla\varphi^*)_{\#}\mu$.*
3. *$\nabla\varphi^*$ is the unique OT map to Monge's problem;*
4. *(**Sufficient Condition.**) If φ convex with $(\nabla\varphi)_{\#}\mu = \nu$, then $\nabla\varphi$ is the OT map that maps μ to ν.*

Monge-Ampère Equation & Differential Geometry

Assume μ and ν are absolutely continuous w.r.t. the Lebesgue measure, i.e., μ and ν have densities f and g, respectively.

$$\boxed{\text{Mass preserving}} \quad \nu = T_{\#}\mu \quad \Longrightarrow \quad g(T(x))|\det(\nabla T)| = f(x)$$

$$\boxed{\text{Mass preserving}} + \boxed{T = \nabla\varphi} \quad \Longrightarrow \quad \text{Monge-Ampère equation:}$$

$$\begin{cases} det(D^2\varphi(x)) = \dfrac{f(x)}{g(\nabla\varphi(x))} \\ \nabla\varphi : X \to Y \\ \varphi \text{ is convex} \end{cases}$$

The MA equation is already studied in the Weyl (1916) and Minkowski (1897) problems in differential geometry of surfaces. This connection is discovered in [Brenier, 1991].

— Connection to Geometric Measure Theory.

Contents

Fluid Dynamics (Benamou–Brenier) Formulation

1. Consider μ as the initial state of a fluid and ν as the final state. We can reformulate W_2 as the minimum kinetic energy required to flow μ to ν [Benamou–Brenier, 2000]:

$$\begin{aligned} \rho(0,\cdot) &= \mu && \text{the initial state} \\ \rho(1,\cdot) &= \nu && \text{the final state} \\ \partial_t \rho(t,x) + \nabla\cdot(\rho(t,x)v(t,x)) &= 0 && \text{the conservation law} \\ & && \text{i.e., the continuity equation} \end{aligned}$$

2. The total kinetic energy of the flow is

$$\int_0^1 \int_\Omega |v(t,x)|^2 \,\mathrm{d}\rho(t,x).$$

   Note that the flow velocity $v(t,x)$ is not unique!

3. The squared quadratic Wasserstein distance is

$$\mathrm{d}^2_{\mathrm{W}_2}(\mu,\nu) = \inf_{\substack{\rho(0,\cdot)=\mu,\ \rho(1,\cdot)=\nu \\ \partial_t\rho+\nabla\cdot(\rho v)=0}} \int_0^1 \int_\Omega |v(t,x)|^2 \,\mathrm{d}\rho(t,x).$$

Contents

Interpolating Probability Distributions

Let $T(x)$ be the OT map between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$.

1. **On the particle level**, the geodesic interpolation between x and $T(x)$ in the Euclidean geometry is

$$T_t(x) = (1-t)x + tT(x)\,.$$

   The speed of this particle is

$$v(t,x) = \frac{\mathrm{d}}{\mathrm{d}t} T_t(x) = T(x) - x \implies \text{constant in time.}$$

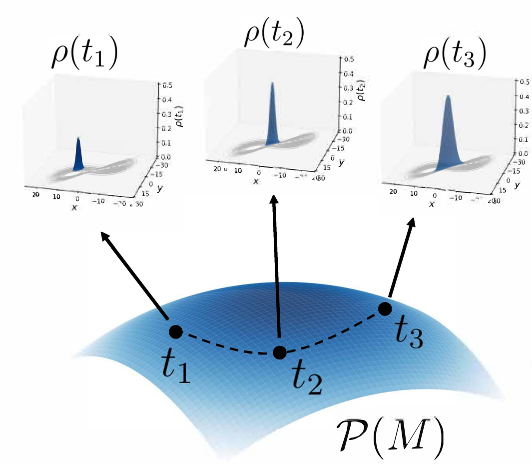
2. **On the measure level**, the geodesic interpolation between μ and ν in the Wasserstein geometry is given by $T_{t\#}$

$$\rho(t,\cdot) = T_{t\#}\mu.$$

   Note that $\rho(0,\cdot) = \mu$ and $\rho(1,\cdot) = \nu$.

Geodesics

1. A geodesic is the generalisation of a straight line, i.e. the shortest distance between two points.
2. In the Wasserstein space $\mathcal{P}(M)$, the "straight line" between μ and ν is induced by the optimal plan/map between them.



Wasserstein vs. Euclidean Geodesics

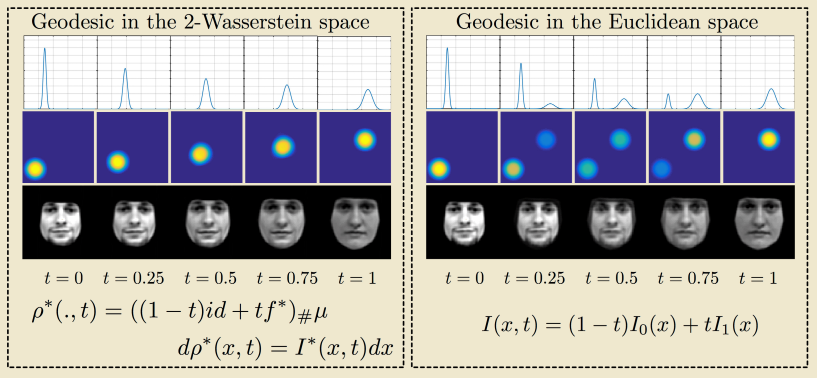


Image Source: Kolouri, S., Park, S. R., Thorpe, M., Slepčev, D., & Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. IEEE signal processing magazine, 34(4), 43-59.

Contents

Gradient $\neq$ Derivative!

If (M, d) denotes the metric space,

$$(\nabla_M E(u), v)_M = \lim_{\epsilon \to 0} \frac{E(u + \epsilon v) - E(u)}{\epsilon}, \quad \forall v \in M. \qquad (1)$$

The gradient $\nabla_M E(u)$ depends on the inner product $(\cdot\,,\,\cdot)_M$.

- If $M = L^2(\mathbb{R}^d)$ and $d = \|\cdot\|_{L^2}$, $\nabla_M E(u) = \frac{\delta E}{\delta u}$.
- If $M = H^1(\mathbb{R}^d)$ and $d = \|\cdot\|_{\dot{H}^1}$, $\nabla_M E(u) = \Delta^{-1}\frac{\delta E}{\delta u}$.

If $M = \mathcal{P}_2(\mathbb{R}^d)$ (all probability distributions with finite 2nd-order moment) and $d = \mathrm{d}_{\mathrm{W}_2}$ (the W_2 metric),

$$\nabla_{\mathrm{d}_{\mathrm{W}_2}} E(u) = -\nabla \cdot (u \nabla \frac{\delta E}{\delta u}).$$

The Wasserstein gradient flow for energy $E(u)$ is

$$\boxed{\frac{\partial u}{\partial t} = \nabla \cdot (u \nabla \frac{\delta E}{\delta u}), \quad u(0) = u_0\,.}$$

The connection started with the seminal work:

> **Jordan**, R., **Kinderlehrer**, D., & **Otto**, F. (1998). The variational formulation of the Fokker–Planck equation. SIAM journal on mathematical analysis, 29(1), 1-17. (the "JKO Paper/Scheme")

$$\rho_{n+1}^{\tau} = \operatorname*{argmin}_{\rho} \left\{ E(\rho) + \frac{\mathrm{d}^2_{\mathrm{W}_2}(\rho, \rho_n^{\tau})}{2\tau} \right\}$$

Well-known kinetic PDEs reinterpreted as WGFs:

Energy Functional	**Gradient Flow**	**Well-known PDE**
$E(\rho) = \int \rho \log(\rho)$	$\rho_t = \Delta \rho$	Heat Equation
$E(\rho) = \int \rho \log(\rho) + \int \rho V$	$\rho_t = \Delta \rho + \nabla \cdot (\rho \nabla V)$	Fokker-Planck Eq.
$E(\rho) = \frac{1}{m-1} \int \rho^m$	$\rho_t = \Delta \rho^m$	Porous Medium Eq.
$E(\rho) = \frac{1}{2} \int \rho(x)\rho(y)W(x-y)$	$\rho_t = \nabla \cdot \rho(\nabla(\rho * W))$	McKean-Vlasov Eq.

— New variational formulation for analyzing & solving kinetic PDEs.

Recommended Books on Optimal Transport (Incomplete)

- Villani, C. (2003). Topics in optimal transportation (Vol. 58). American Mathematical Soc.
- Villani, C. (2009). Optimal transport: old and new (Vol. 338). Berlin: springer.
- Santambrogio, F. (2015). Optimal transport for applied mathematicians. Birkäuser, NY, 55(58-63), 94.
- Galichon, A. (2018). Optimal transport methods in economics. Princeton University Press.
- Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. Foundations and Trends in Machine Learning, 11(5-6), 355-607.
- Figalli, A., & Glaudo, F. (2021). An invitation to optimal transport, Wasserstein distances, and gradient flows.
- Ambrosio, L., Brué, E., & Semola, D. (2021). Lectures on optimal transport (Vol. 130). Cham: Springer.
- Maggi, F. (2023). Optimal mass transport on Euclidean spaces. Cambridge University Press.