MK Optimal Transport and entropic relaxations

Soumik Pal
University of Washington, Seattle

Eigenfunctions seminar @ IISc Bangalore, August 30, 2019

Monge-Kantorovich Optimal Transport problem

Gaspard Monge 1781

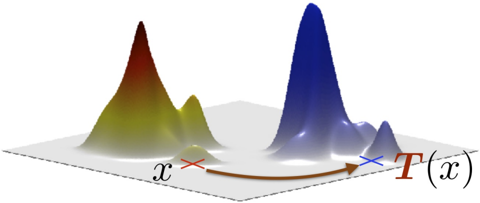

Figure: by M. Cuturi

- P, Q - probabilities on $\mathcal{X}, \mathcal{Y}$, respectively, say both $\mathbb{R}^d$.
- $c(x, y)$ - cost of transport. E.g., $c(x, y) = \|x - y\|$ or $c(x, y) = \frac{1}{2}\|x - y\|^2$.
- Monge problem: minimize among $T : \mathbb{R}^d \to \mathbb{R}^d$, $T_\# P = Q$,

$$\int c\,(x, T(x))\, dP.$$

Leonid Kantorovich 1939

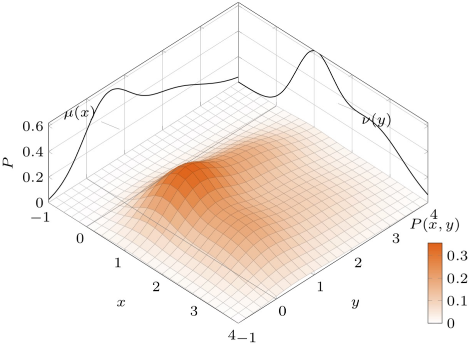

Figure: by M. Cuturi

- $\Pi(P, Q)$ - *couplings* of (P, Q) (joint dist. with given marginals).
- (Monge-) Kantorovich relaxation: minimize among $\nu \in \Pi(P, Q)$

$$\inf_{\nu \in \Pi(P,Q)} \left[\int c\,(x, y)\, d\nu \right].$$

Duality

- cost $\rightarrow$ price
- Among all functions $\phi(y), \psi(x)$ s.t. $\phi(y) - \psi(x) \leq c(x, y)$, maximize profit

$$\sup_{\phi,\psi} \left[\int \phi(y) Q(dy) - \int \psi(x) P(dx) \right].$$

- (Kantorovich duality) inf cost = sup profit.
- For the optimal "Kantorovich potentials"

$$\phi_c(x) - \psi_c(y) = c(x, y),$$

"optimal coupling" ν_c- almost surely.

Quadratic cost: Brenier's theorem

- How do OT looks like? Very special!
- $c(x,y) = \frac{1}{2}\|x-y\|^2$. Assume P has density ρ_0.
- (Y. Brenier) $\exists$ a convex F s.t. $(X, \nabla F(X))$, $X \sim \rho_0$ solves

$$(\mathrm{MK-OT}) \quad \mathbb{W}_2^2(P,Q) := \inf_{\Pi(P,Q)} \left[\int c(x,y)\, d\nu \right].$$

- K.- potentials? $F^*(y)$- Legendre convex dual of F.

$$\phi_c(x) = \frac{1}{2}\|x\|^2 - F(x), \quad -\psi_c(y) = \frac{1}{2}\|y\|^2 - F^*(y).$$

- $\phi_c(x) - \psi_c(y) = \frac{1}{2}\|x-y\|^2$, for $y = \nabla F(x)$, i.e., a.s. ν_c.

A generalized notion of convexity (Gangbo-McCann)

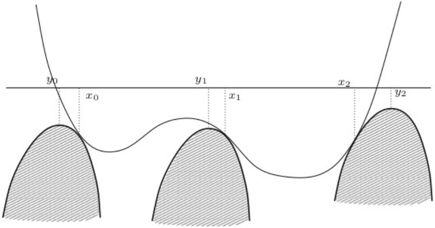

Figure: by C. Villani

- Convex functions lie above their tangents.
- c-convex function $\psi(x)$ lie above the cost curve $c(\cdot, y)$, $y \in \partial_c \psi(x)$.
- optimal Kantorovich potentials are c-concave.

$$\psi_c(x) = \sup_y \left[\phi_c(y) - c(x,y)\right], \quad \phi_c(y) - \psi_c(x) = c(x,y), \; y \in \partial_c \psi(x).$$

Convex cost: Gangbo - McCann '96

- $c(x, y) = g(x - y)$, g strictly convex +
- P has density ρ_0.
- $\exists$ c-concave function $\psi_c(x)$ for which

$$T(x) = x - (\nabla g)^{-1} \circ \nabla \psi_c(x)$$

  is s.t. $(X, T(X))$, $X \sim \rho_0$, ! solves the MK OT problem.
- $T(x) \in \partial_c \psi_c(x)$.
- Monge solution is also MK solution.
- Does not cover $g(z) = \|z\|$ or $g(z) = 1\{z \neq 0\}$.

Existence of Monge solution

- Sufficient conditions (Bernard-Buffoni, Villani, De Philippis)
- $\mathcal{X}, \mathcal{Y}$ bounded, open. P, Q have densities.
- $c(x, y) \in C^2$.
- $y \mapsto D_x c(x, y)$ is injective for each x (Twist condition).
- $x \mapsto D_y c(x, y)$ is injective for each y.
- See book by Villani Chapter 10.
- Smoothness of optimal T. Ma-Trudinger-Wang '05, Loeper '09 (see Villani, Chap 12).

Transport in one dimension

- Suppose $\mathcal{X} = \mathbb{R} = \mathcal{Y}$.
- for all convex $c(x, y) = g(x - y)$ the OT map is well-known.
- Monotone transport AKA inverse c.d.f. transform.

$$T(x) = G_1^{-1} \circ G_0(x),$$

  G_0, G_1 - c.d.f. of P, Q, resp, continuous.
- Optimal, unique if g is strict. (Homework)

Entropic Relaxation or Entropic Regularization

OT and statistics

- Goal: Fit data to model. Classical: MLE.
- Recent: minimize $\mathbb{W}_2^2(\mathrm{data}, \mathrm{model})$.
- Better estimates, more stable, high dimension, Adversarial Network training.
- Problem is computation. Discrete MK-OT.

OT and statistics

- Goal: Fit data to model. Classical: MLE.
- Recent: minimize $\mathbb{W}_2^2(\text{data}, \text{model})$.
- Better estimates, more stable, high dimension, Adversarial Network training.
- Problem is computation. Discrete MK-OT.
- Given two empirical distributions

$$\sum_{i=1}^{n} p_i \delta_{x_i}, \quad \sum_{j=1}^{n} q_j \delta_{y_j}, \quad \sum_i p_i = 1 = \sum_j q_j,$$

- minimize $\langle c, M \rangle := \sum_i \sum_j c(x_i, y_j) M_{ij}$, among all $n \times n$ matrices $M \geq 0$ with row sum p and col sum q.

Entropic relaxation, Cuturi '13

- Linear programing M. Simplex, interior point methods give complexity $O(n^3 \log n)$. Pretty bad.

Entropic relaxation, Cuturi '13

- Linear programing M. Simplex, interior point methods give complexity $O(n^3 \log n)$. Pretty bad.
- Define
$$\mathrm{Ent}(M) = \sum_{i,j} M_{ij} \log M_{ij}, \quad 0 \log 0 = 0.$$

Entropic relaxation, Cuturi '13

- Linear programing M. Simplex, interior point methods give complexity $O(n^3 \log n)$. Pretty bad.
- Define
$$\mathrm{Ent}(M) = \sum_{i,j} M_{ij} \log M_{ij}, \quad 0 \log 0 = 0.$$
- For $h > 0$, minimize $[\langle c, M \rangle + h\mathrm{Ent}(M)]$.
- Penalizes degenerate solutions (sparse M). Optimal $h \downarrow 0$.
- Computational complexity $\approx O\left(n^2 \log n\right)$. How?

Entropic relaxation: solution

- For $h > 0$, minimize $[\langle c, M \rangle + h\mathrm{Ent}(M)]$.
- Solution (Lagrange multipliers + calculus): $\exists u, v \in \mathbb{R}^n$

$$M_c = \mathrm{Diag}(u) \exp\left(-\frac{1}{h}c\right) \mathrm{Diag}(v), \text{ i.e.,}$$

$$M_c(i,j) = u_i \exp\left(-\frac{1}{h}c(x_i, y_j)\right) v_j, \quad 1 \le i, j \le n.$$

- Remember this form. Will get back in continuum.

Sinkhorn algorithm AKA IPFP

- M_c can be solved by Iterative Proportional Fitting Procedure.
- Start with $M_0 = \exp\left(-\frac{1}{h}c\right)$. Inductively ...
- Rescale rows of M_k to get M_{k+1} with row sum p.
- Rescale columns of M_{k+1} to get M_{k+2} with col sum q.
- Limit $= M_c$.
- Called Sinkhorn iterations in Linear Algebra.

Entropic relaxation in continuum

- Recall $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$. Cost $c(x, y)$.
- P, Q have densities ρ_0, ρ_1.
- For density $\nu \in \Pi(\rho_0, \rho_1)$,

$$\mathrm{Ent}(\nu) = \int \nu(x, y) \log \nu(x, y) dxdy.$$

- Entropic relaxation: $h > 0$,

$$\mathrm{minimize} \left[\int c(x, y)\nu(x, y) dxdy + h\mathrm{Ent}(\nu),\ \nu \in \Pi(\rho_0, \rho_1) \right].$$

Entropic relaxation: continuum solution

- (Hobby - Pyke '65, Rüschendorff-Thomsen '93) Optimal solution

$$\nu_c(x,y) = \exp\left(a(x) + b(y) - \frac{1}{h}c(x,y)\right)$$

$$= u(x)\exp\left(-\frac{1}{h}c(x,y)\right)v(y).$$

- Just like the discrete case.
- Can be computed by IPFP. Unfortunately, very slow convergence.

Entropic duality

- Recall duality for MK-OT: $\inf_{\Pi(\rho_0,\rho_1)} \int c(x,y)\nu(x,y)dxdy$

$$= \sup_{\phi(y)-\psi(x)\le c(x,y)} \left[\int \phi(y)\rho_1(y)dy - \int \psi(x)\rho_0(x)dx \right].$$

- Duality for entropic relaxation: Solve

$$\sup \left[\int \phi(y)\rho_1(y)dy - \int \psi(x)\rho_0(x)dx - h \int e^{\phi(y)-\frac{1}{h}c(x,y)-\psi(x)} \right].$$

- Optimal solutions: $\psi(y) = b(y),\ \phi(x) = -a(x)$.
- a, b are Schrödinger potentials.

Schrödinger bridges, Large Deviations

Schrödinger's problem: Lazy gas experiment

- Imagine $N \approx \infty$ independent gas molecules in a cold chamber.
- Initial configuration of particles $L_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \approx P$.
- Each particle independent Brownian motion with $\sigma^2 \approx 0$.
- Condition of the terminal configuration $L_1 = \frac{1}{N}\sum_{j=1}^{N} \delta_{y_j} \approx Q$.
- (Schrödinger '32) What is the probability of the above event?
- What is the most likely path followed by an individual gas molecule?

Föllmer's reformulation '88

- Relative Entropy (RE) of μ w.r.t. ν

$$H(\mu \mid \nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu.$$

- R - Law of σ^2 BM on $C[0,1]$, initial distribution P.
- Among all probability μ on $C[0,1]$ s.t. $X_0 \sim P$, $X_1 \sim Q$,

$$\text{minimize}\ H\left(\mu \mid R\right).$$

- Solution is Schrödinger bridge between P and Q.
- Take $\sigma^2 \downarrow 0$.

Föllmer's disintegration

- Brownian transition

$$p_\sigma(x, y) = \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2\sigma^2} \|y - x\|^2\right).$$

- (Föllmer) Let R_{01} be the law of (X_0, X_1). Find $\nu \in \Pi(P, Q)$ to

$$\text{minimize } H(\nu \mid R_{01}).$$

- Generate (X_0, X_1) from the minimizer. Schrödinger bridge is σ^2 Brownian bridge given $X_0 = x_0, X_1 = x_1$.

Entropic relxation and Schrödinger bridge

- Minimize $H(\nu \mid R_{01})$ is the same problem as

$$\text{minimize}\left[\frac{1}{2}\int \|y-x\|^2\, d\nu + \sigma^2 \mathrm{Ent}(\nu)\right].$$

- Entropic relaxation $h = \sigma^2$ for the quadratic cost.
- Schrödinger bridge description: solve the entropic relaxation and join by Brownian bridge.
- What happens when $\sigma^2 \downarrow 0$?

Large deviation

- As $h = \sigma^2 \to 0+$, the optimal entropic coupling converges to the MK-optimal coupling.
- Recall Brenier: $P(dx) = \rho_0(x)dx$, $Q(dy) = \rho_1(y)dy$.
- $\exists$ F such that $y = \nabla F(x)$ gives Monge.
- σ^2 Brownian bridge converges to a constant velocity straight line joining x and y.
- Can be made precise by Large Deviation theory.
- Let ρ_t be law at time t of this limit. McCann interpolation between ρ_0 and ρ_1. Remember this name for later.

(f, g) transform of Markov processes

- How to describe the law of Schrödinger bridges? SDE? PDE?
- Markovian (f, g) transform of reversible Wiener measure $\mathbb{W}$:

$$d\mu = f(X_0)g(X_1)d\mathbb{W}, \quad E_{\mathbb{W}} f(X_0)g(X_1) = 1.$$

- Similar to Girsanov / Doob's h-transform, but on both sides.
- Markovian diffusion both forward and backward.

Generators for Schrödinger bridges

- Let μ_t be the law of the $\sigma^2 = 1$ Schrödinger bridge.
- Recall Schrödinger potentials: $a(x), b(y)$.
- Define, heat-flows

$$b_t(y) = \log \mathbb{W}\left(e^{b(X_1)} \mid X_t = y\right), \quad a_t(x) = \log \mathbb{W}\left(e^{a(X_0)} \mid X_t = x\right).$$

- Schrödinger bridge is BM with drift ∇b_t forward in time.
- Schrödinger bridge is BM with drift ∇a_t backward in time.
- Most properties are poorly understood.

Dynamics and geometry

McCann interpolation

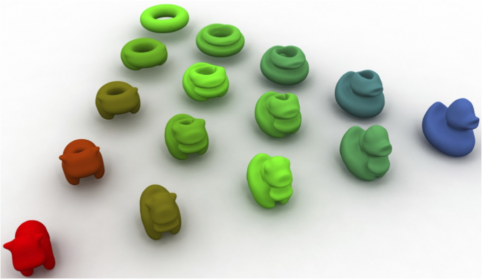

Figure: by M. Cuturi

- $\mathcal{P}_2\left(\mathbb{R}^d\right)$ - square integrable probabilities
- Recall: ρ_0 transported to ρ_1. $c(x,y) = \frac{1}{2}\|y-x\|^2$.
- Square-root optimal cost $\mathbb{W}_2(\rho_0, \rho_1)$ is a metric.
- $\rho_t =$ Law of $(1-t)X + tT(X)$, $X \sim \rho_0$, $0 \leq t \leq 1$.

Wasserstein geodesics

- Extend to Riemannian manifolds $(M, \mathbf{d})$.
- $c(x, y) = \frac{1}{2}\mathbf{d}^2(x, y)$. Metric $\mathbb{W}_2(\rho_0, \rho_1)$.
- (Otto + etc.) Riemannian geometry on $\mathcal{P}_2(M)$.
- $(\rho_t,\ 0 \le t \le 1)$ - geodesic (straight line) joining ρ_0 and ρ_1.
- (McCann + etc.) Many natural objects such as entropy are (semi-) convex functions over these lines.

Ricci curvature

- (McCann, Lott-Sturm-Villani) Synthetic view of Ricci curvature. Villani '09: Take a perfect gas in which particles do not interact, and ask to move from a certain prescribed density field at time $t = 0$, to another prescribed density field at time $t = 1$. Since the gas is lazy, it will find a way to do so that needs a minimal amount of work (least action principle). Measure the entropy of the gas at each time, and check that it always lies above the line joining the final and initial entropies. If such is the case, then we know that we live in a nonnegatively curved space.
- This is, of course, Schrödinger bridge in the limit. What about $\sigma^2 > 0$? (Conforti, Gigli)

Current activities

- In spite of their importance, Schrödinger bridges are (still) poorly understood. Many active areas:
- Generalizations to other cost functions (Léonard, Mikami)
- Discrete spaces, Markov chains, jump processes (Erbar, Maas)
- PDE, Smooth approximations to Wasserstein geodesics (Gigli-Tamimani etc.)
- Behavior of entropy, functional inequalities (Conforti, Conforti-Tamanini)
- Statistics (Cuturi, Peyré, Carlier, Rigollet, Weed)

References

- C. Léonard - A survey of the Schrödinger problem and some of its connections with optimal transport.
- Cédric Villani - Optimal Transport old and new.
- M. Cuturi and G. Peyré - Computational optimal transport.