

On entropic cost — optimal transport cost

Soumik Pal
University of Washington, Seattle
[arxiv:1905.12206](https://arxiv.org/abs/1905.12206)

Eigenfunctions seminar @ IISc Bangalore, August 30, 2019

MK OT and entropic relaxation

- ρ_0, ρ_1 - probability densities on $\mathcal{X} = \mathbb{R}^d = \mathcal{Y}$.
- $c(x, y) = g(x - y)$, strictly convex, $g \geq 0$, $g(z) = 0$ iff $z = 0$.
- $\Pi(\rho_0, \rho_1)$ - set of couplings. Probabilities on $\mathcal{X} \times \mathcal{Y}$.
- Monge-Kantorovich (MK) OT problem:

$$\mathbb{W}_g(\rho_0, \rho_1) := \inf_{\nu \in \Pi} \nu(g(x - y)) = \inf_{\nu \in \Pi} \int g(x - y) d\nu.$$

- Entropic relaxation (Cuturi, Peyré). For $h > 0$,

$$K'_h := \inf_{\nu \in \Pi} [\nu(g(x - y)) + h \text{Ent}(\nu)], \quad \text{Ent}(\nu) = \int \nu(x) \log \nu(x) dx$$

- Fast algorithms for $h > 0$. Want $h \rightarrow 0$.

Entropic cost

- An equivalent form of entropic relaxation.
- Define “transition kernel”:

$$p_h(x, y) = \frac{1}{\Lambda_h} \exp\left(-\frac{1}{h}g(x - y)\right),$$

and joint distribution $\mu_h(x, y) = \rho_0(x)p_h(x, y)$.

- Relative entropy:

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

- Define **entropic cost**

$$K_h = \inf_{\text{couplings}(\rho_0, \rho_1)} H(\nu | \mu_h).$$

- $K_h = K'_h/h - \text{Ent}(\rho_0) + \log \Lambda_h$

Example: quadratic Wasserstein

- Consider $g(x - y) = \frac{1}{2} \|x - y\|^2$.
- $p_h(x, y)$ - transition of Brownian motion. $h = \text{temperature}$.

$$p_h(x, y) = (2\pi h)^{-d/2} \exp\left(-\frac{1}{2h} \|x - y\|^2\right).$$

- In general, there need not be a stochastic process for $p_h(x, y)$.

Theorem (Y. Brenier '87)

There exists unique convex ϕ such that $T(x) = \nabla\phi(x)$ solves both Monge and Kantorovich OT problems for (ρ_0, ρ_1) .

Schrödinger's problem

- Brownian motion X - temperature $h \approx 0$
- "Condition" $X_0 \sim \rho_0, X_1 \sim \rho_1$. Exponentially rare.
- On this rare event what do particles do?
- Schrödinger '31, Föllmer '88, Léonard '12.
- Particle initially at x moves close to $\nabla\phi(x)$ (Brenier map).
- In fact, $\lim_{h \rightarrow 0} hK_h = \frac{1}{2}\mathbb{W}_2^2(\rho_0, \rho_1)$.
- True in general. For any $g(x - y)$:

$$\lim_{h \rightarrow 0} hK_h = \mathbb{W}_g(\rho_0, \rho_1).$$

- Rate of convergence?

Pointwise convergence

Theorem (P. '19)

ρ_0, ρ_1 compactly supported and continuous (+ smoothness etc.).
Kantorovich potential uniformly convex.

$$\lim_{h \rightarrow 0^+} \left(K_h - \frac{1}{2h} \mathbb{W}_2^2(\rho_0, \rho_1) \right) = \frac{1}{2} (\text{Ent}(\rho_1) - \text{Ent}(\rho_0)).$$

- Complementary results known for gamma convergence. Pointwise convergence left open.
- Adams, Dirr, Peletier, Zimmer '11 (1-d), Duong, Laschos, Renger '13, Erbar, Maas, Renger '15 (multidimension, Fokker-Planck).

Divergence

- To state the result for a general g , need a new concept.
- For a convex function ϕ , Bregman divergence:

$$D[y | z] = \phi(y) - \phi(z) - (y - z) \cdot \nabla\phi(z) \geq 0.$$

- If $x^* = \nabla\phi(x)$,

$$D[y | x^*] = \frac{1}{2} \|y - x\|^2 - \phi_c(x) - \phi_c^*(y),$$

where ϕ_c, ϕ_c^* are c-concave functions:

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

- $y \approx x^*$, $D[y | x^*] \approx (y - x^*)^T A(x^*)(y - x^*)$, $A(z) = \nabla^2\phi^*(z)$.

Divergence

- Generalize to cost g . Monge solution given by (Gangbo - McCann)

$$x^* = x - (\nabla g)^{-1} \circ \nabla \psi,$$

for some c -concave function ψ . Dual c -concave function ψ^* .

- Divergence

$$D[y | x^*] = g(x - y) - \psi(x) - \psi^*(y) \geq 0.$$

- $y \approx x^*$, extract matrix $A(x^*)$ from the Taylor series.
- Divergence/ $A(\cdot)$ measures sensitivity of Monge map. Related to cross-difference of Kim & McCann '10, McCann '12, Yang & Wong '19.

Pointwise convergence

Theorem (P. '19)

ρ_0, ρ_1 compactly supported, continuous (+ smoothness etc.). $A(\cdot)$ "uniformly elliptic".

$$\lim_{h \rightarrow 0^+} \left(K_h - \frac{1}{h} \mathbb{W}_g(\rho_0, \rho_1) \right) = \frac{1}{2} \int \rho_1(y) \log \det(A(y)) dy - \frac{1}{2} \log \det \nabla^2 g(0).$$

- For $g(x - y) = \|x - y\|^2 / 2$, $\log \det \nabla^2 g(0) = 0$, for ϕ (Brenier)

$$\frac{1}{2} \int \rho_1(y) \log \det(A(y)) dy = \frac{1}{2} \int \rho_1(y) \log \det(\nabla^2 \phi^*(y)) dy,$$

which is $\frac{1}{2} (\text{Ent}(\rho_1) - \text{Ent}(\rho_0))$ by simple calculation a la McCann.

The Dirichlet transport

Dirichlet transport, P.-Wong '16

- Δ_n - unit simplex $\{(p_1, \dots, p_n) : p_i > 0, \sum_i p_i = 1\}$.
- Δ_n is an abelian group. $e = (1/n, \dots, 1/n)$
- If $p, q \in \Delta_n$, then

$$(p \odot q)_i = \frac{p_i q_i}{\sum_{j=1}^n p_j q_j}, \quad (p^{-1})_i = \frac{1/p_i}{\sum_{j=1}^n 1/p_j}.$$

- K-L divergence or relative entropy as “distance”:

$$H(q | p) = \sum_{i=1}^n q_i \log(q_i/p_i).$$

- Take $\mathcal{X} = \mathcal{Y} = \Delta_n$.

$$c(p, q) = H(e | p^{-1} \odot q) = \log \left(\frac{1}{n} \sum_{i=1}^n \frac{q_i}{p_i} \right) - \frac{1}{n} \sum_{i=1}^n \log \frac{q_i}{p_i} \geq 0.$$

Some economic motivation

- Market weights for n stocks: $\mu = (\mu_1, \dots, \mu_n)$.
- μ_i = Proportion of the total capital that belongs to i th stock.
- Investment portfolio: $\pi = (\pi_1, \dots, \pi_n) \in \Delta_n$.
- Portfolio weights:

π_i = Proportion of the total value that belongs to i th stock.

- Markovian investments $\pi = \pi(\mu) : \Delta_n \rightarrow \Delta_n$.
- How to build robust portfolios that compare with an index, say, S&P 500? ONLY solutions given by the Dirichlet transport.

Exponentially concave functions

- $\varphi : \Delta_n \rightarrow \mathbb{R} \cup \{-\infty\}$ is exponentially concave if e^φ is concave.
- $x \mapsto \frac{1}{2} \log x$ is e-concave, but not $x \mapsto 2 \log x$.
- Examples: $p, r \in \Delta_n$, $0 < \lambda < 1$.

$$\varphi(p) = \frac{1}{n} \sum_i \log p_i.$$

$$\varphi(p) = \log \left(\sum_i r_i p_i \right), \quad \varphi(p) = \frac{1}{\lambda} \log \left(\sum_i p_i^\lambda \right).$$

- (Fernholz '02, P. and Wong '15). Analog of Brenier's Theorem: If $(p, q = F(p))$ is the Monge solution, then

$$p^{-1} = \tilde{\nabla} \varphi(q), \quad \text{Kantorovich potential.}$$

- Smooth, MTW Khan & Zhang '19.

Back to the Dirichlet transport

- What is the corresponding probabilistic picture for the cost function

$$c(p, q) = H(e \mid p^{-1} \odot q) \text{ on the unit simplex } \Delta_n?$$

- Symmetric Dirichlet distribution $\text{Dir}(\lambda)$:

$$\text{density} \propto \prod_{j=1}^n p_j^{\lambda/n-1}.$$

- Probability distribution on the unit simplex. If $U \sim \text{Dir}(\cdot)$,

$$\mathbb{E}(U) = e, \quad \text{Var}(U_i) = O\left(\frac{1}{\lambda}\right).$$

Dirichlet transition

- Haar measure on (Δ_n, \odot) is $\text{Dir}(0)$, $\nu(p) = \prod_{i=1}^n p_i^{-1}$.
- Consider transition probability: $p \in \Delta_n$, $U \sim \text{Dir}(\lambda)$, $Q = p \odot U$.

$$f_\lambda(p, q) = c\nu(q) \exp(-\lambda c(p, q)), \quad (\text{P.-Wong '18}).$$

- Temperature: $h = \frac{1}{\lambda}$. Let

$$p_h(p, q) = f_{1/h}(p, q).$$

- As $h \rightarrow 0+$, $p_h \rightarrow \delta_p$. As $h \rightarrow \infty$, $Q \rightarrow \text{Dir}(0)$, Haar measure.

Multiplicative Schrödinger problem

- Fix ρ_0, ρ_1 . Let $\mu_h(p, q) = \rho_0(p)\rho_h(p, q)$.
- Recall relative entropy: $H(\nu | \mu) = \int \log(d\nu/d\mu)d\mu$.
- Entropic cost

$$K_h = \inf_{\text{couplings}(\rho_0, \rho_1)} H(\nu | \mu_h)$$

- For ρ density on Δ_n , let

$$\text{Ent}_0(\rho) = H(\rho | \text{Dir}(0)).$$

Relative entropy w.r.t. Haar measure.

Pointwise convergence

Theorem (P. '19)

ρ_0, ρ_1 are compactly supported + exponentially concave potential is “uniformly convex”.

$$\lim_{h \rightarrow 0^+} \left(K_h - \left(\frac{1}{h} - \frac{n}{2} \right) \mathbf{C}(\rho_0, \rho_1) \right) = \frac{1}{2} (\text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0)).$$

- $\mathbf{C}(\rho_0, \rho_1)$ is the optimal cost of transport with cost c .
- Not a metric, but a divergence. Not symmetric in (ρ_0, ρ_1) .
- AFAIK, the only such example known.
- Related to Erbar '14 (jump processes), and Maas '11 (Markov chains).

Idea of the proof: approximate Schrödinger bridge

Idea of the proof: Brownian case

- Recall, want to condition Brownian motion to have marginals ρ_0, ρ_1 .
- $p_h(x, y)$ Brownian transition density at time h .

$$\mu_h(x, y) = \rho_0(x)p_h(x, y), \quad \text{joint distribution.}$$

- If I can “guess” this conditional distribution ν_h , then

$$K_h = \inf_{\text{couplings}(\rho_0, \rho_1)} H(\nu | \mu_h) = H(\tilde{\mu}_h | \mu_h).$$

- Can approximately do so for small h by a Taylor expansion in h .

Idea of the proof: Brownian case

- It is known (Rüschendorf) that $\tilde{\mu}_h$ must be of the form

$$\tilde{\mu}_h(x, y) = e^{a(x)+b(y)} \mu_h(x, y) \propto \exp\left(-\frac{1}{h}g(x-y) + a(x) + b(y)\right).$$

- ϕ - convex function from Brenier map.

$$a(x) = \frac{1}{h} \left(\frac{\|x\|^2}{2} - \phi(x) \right) + h\zeta_h(x), \quad b(y) = \frac{1}{h} \left(\frac{|y|^2}{2} - \phi^*(y) \right) + h\xi_h(y),$$

ζ_h, ξ_h are $O(1)$.

Idea of the proof

- Thus, up to lower order terms,

$$\begin{aligned}\tilde{\mu}_h(x, y) &\propto \rho_0(x) \exp\left(-\frac{1}{h}g(x-y) + \frac{1}{h}\phi_c(x) + \frac{1}{h}\phi_c^*(y)\right) \\ &= \rho_0(x) \exp\left(-\frac{1}{h}D[y | x^*]\right).\end{aligned}$$

- If $y - x^*$ is large, it gets penalized exponentially. Hence

$$\tilde{\mu}_h(x, y) \propto \rho_0(x) \exp\left(-\frac{1}{2h}(y - x^*)^T \nabla^2 \phi^*(x^*)(y - x^*)\right)$$

- Gaussian transition kernel with mean x^* and covariance $h(\nabla^2 \phi^*(x^*))^{-1}$.

Idea of the proof

- For $h \approx 0$, the Schrödinger bridge is approximately Gaussian. Sample $X \sim \rho_0$, generate $Y \sim N(x^*, h(\nabla^2 \phi^*(x^*))^{-1})$.

$$\tilde{\mu}_h(x, y) \approx \rho_0(x) \frac{1}{\sqrt{\det(\nabla^2 \phi^*(x^*))}} (2\pi h)^{-d/2} \times \\ \exp\left(-\frac{1}{2h}(y - x^*)^T \nabla^2 \phi^*(x^*)(y - x^*)\right).$$

- Y is not exactly ρ_1 . Lower order corrections.
- Nevertheless,

$$H(\tilde{\mu}_h | \mu_h) = \frac{1}{2} \int \det \nabla^2 \phi^*(x^*) \rho_0(x) dx = \frac{1}{2} (\text{Ent}(\rho_1) - \text{Ent}(\rho_0)).$$

Gradient flow of entropy

- Ambrosio-Gigli-Savaré; recent survey by Santambrogio.
- Consider the Cauchy problem in \mathbb{R}^n :

$$x'(t) = -\nabla F(x(t)), \quad x(0) = x_0.$$

- Gradient flow with potential F .
- Euler discretization: fix small step parameter $h > 0$.

$$x_{k+1}^h = \operatorname{argmin}_x \left[\frac{\|x - x_k^h\|^2}{2h} + F(x) \right].$$

- FOC:

$$\frac{x_{k+1}^h - x_k^h}{h} = -\nabla F(x_k^h), \text{ converges to gradient flow as } h \rightarrow 0+.$$

Heat equation as a gradient flow of entropy

- Start with $\rho(0) = \rho_0$ density. Fix $h > 0$.

$$\rho^{(k+1)} = \operatorname{argmin}_{\rho} \left[\frac{1}{2h} \mathbb{W}_2^2(\rho, \rho^k) + \operatorname{Ent}(\rho) \right].$$

- Define interpolation

$$\rho^h(t) = \rho^{(k)}, \quad kh \leq t < (k+1)h.$$

- Jordan-Kinderlehrer-Otto (JKO) '98: $\rho^h(t)$ “converges” to heat equation.

$$\frac{\partial \rho}{\partial t} = \frac{\partial^2 \rho}{\partial x^2}, \quad \rho(0, x) = \rho_0.$$

- Gradient flow of entropy in Wasserstein metric space.

Entropic cost to gradient flow

- How does entropic cost imply gradient flow for the heat equation?
- Brownian motion starting from ρ_0 .
- $\rho(t)$ - density at time t . Obviously,

$$\rho_h = \operatorname{argmin} K_h(\rho_0, \rho), \quad \rho_{(k+1)h} = \operatorname{argmin}_\rho K_h(\rho_{kh}, \rho).$$

Relative entropy **is** minimized by the exact transition density.

- But

$$J_h(\rho_0, \rho) \approx \frac{1}{2h} \mathbb{W}_2^2(\rho_0, \rho) + \frac{1}{2} (\operatorname{Ent}(\rho) - \operatorname{Ent}(\rho_0)).$$

- This “morally” implies gradient flow of entropy.

Gradient flow without a metric?

- Dirichlet transport has a similar structure.

$$K_h(\rho, \rho_0) \approx \left(\frac{1}{h} - \frac{n}{2} \right) \mathbf{C}(\rho_0, \rho) + \frac{1}{2} (\text{Ent}_0(\rho) - \text{Ent}_0(\rho_0)).$$

- Hence, successively multiplying \odot by symmetric Dirichlet should be a gradient flow of entropy.
- BUT ... $\mathbf{C}(\rho_0, \rho)$ is not a metric. No such theory exists.
- Is there even a stochastic process?

Thank you very much for your attention

arxiv math.PR:1905.12206