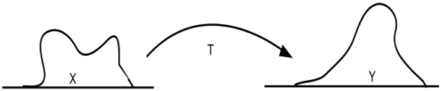
Divergence, Gibbs measures, and entropic regularizations of optimal transport

Soumik Pal
University of Washington, Seattle

Fields Institute, Feb 13, 2020

The Monge problem 1781



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

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

Kantorovich relaxation 1939

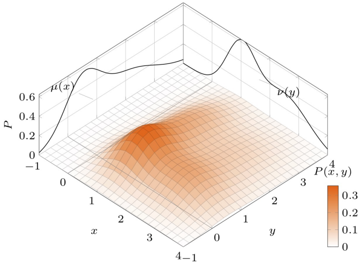


Figure: by M. Cuturi

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

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

- Linear optimization in ν over convex $\Pi(P,Q)$.

Example: quadratic Wasserstein

- Consider $c(x, y) = \frac{1}{2} \|x - y\|^2$.
- Assume P, Q has densities ρ_0, ρ_1.

$$\mathbb{W}_2^2(P, Q) = \mathbb{W}_2^2(\rho_0, \rho_1) = \inf_{\nu \in \Pi(\rho_0, \rho_1)} \left[\int \|x - y\|^2 \, d\nu \right].$$

Theorem (Y. Brenier '87)

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

When are MK solutions Monge?

- When transporting densities, other cost functions give Monge solutions.
- **Twist condition**: $y \mapsto \nabla_x c(x,y)$ is 1-1.
- Example: $c(x,y) = g(x-y)$, strictly convex.

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

Entropic regularization

- Monge solutions are highly degenerate; supported on a graph.
- Entropy as a measure of degeneracy:

$$\mathrm{Ent}(\nu) := \begin{cases} \int f(x)\log f(x)dx, & \text{if } \nu \text{ has a density } f, \\ \infty, & \text{otherwise.} \end{cases}$$

- Example: Entropy of $N(0,\sigma^2)$ is $-\log\sigma+$ constant.
- Monge solutions have infinite entropy.
- Föllmer '88, Rüschendorff-Thomsen '93, Cuturi '13, Gigli '19 ... suggested penalizing OT with entropy.
- Why? Fast algorithms. Statistical physics. Smooth approximations.

Entropic regularization

- MK OT problem with $c(x,y) = g(x-y)$, $g \geq 0$ str. cx.

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

- For $h > 0$,

$$K'_h := \inf_{\nu \in \Pi} \left[\nu(g(x-y)) + h\mathrm{Ent}(\nu) \right].$$

- Naturally,

$$K'_h(\rho_0, \rho_1) \approx \mathbb{W}_g(\rho_0, \rho_1), \quad \text{as } h \to 0+.$$

- What is the rate of convergence?

Entropic cost

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

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

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

- Relative entropy:

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

- Define **entropic cost**

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

- $K_h = K_h'/h - \mathrm{Ent}(\rho_0) + \log \Lambda_h$.

Example: quadratic Wasserstein

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

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

- Entropic cost, $K_h = \frac{K_h'}{h} - \mathrm{Ent}(\rho_0) + \frac{d}{2}\log(2\pi h)$.
- In general, there need not be a stochastic process for $p_h(x,y)$.

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).
- Recall: For any $g(x-y)$:

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

- Rate of convergence?

Pointwise convergence

Theorem (P. '19)

ρ_0, ρ_1 compactly supported (+ technical conditions). Kantorovich potential uniformly convex.

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

- 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:

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

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

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

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

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

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

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

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

- $y \approx x^*$, 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 (+ technical condition). $A(\cdot)$ "uniformly elliptic".

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

- 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 $\frac{1}{2}\left(\mathrm{Ent}(\rho_1) - \mathrm{Ent}(\rho_0)\right)$ by simple calculation par McCann.

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 $\widetilde{\mu}_h$, then

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

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

Idea of the proof: Brownian case

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

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

- ϕ - convex function from Brenier map.

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

ζ_h, ξ_h are $O(1)$.

Idea of the proof

- Thus, up to lower order terms,

$$\widetilde{\mu}_h(x,y) \propto \rho_0(x) \exp\left(-\frac{1}{h} g(x-y) + \frac{1}{h}\phi_c(x) + \frac{1}{h}\phi_c^*(y)\right)$$

$$= \rho_0(x) \exp\left(-\frac{1}{h} D[y \mid x^*]\right).$$

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

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

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

Idea of the proof

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

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

$$\exp\left(-\frac{1}{2h}(y-x^*)^T\nabla^2\phi^*(x^*)(y-x^*)\right).$$

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

$$H\left(\widetilde{\mu}_h \mid \mu_h\right) = \frac{1}{2}\int \det\nabla^2\phi^*(x^*)\rho_0(x)dx = \frac{1}{2}\left(\mathrm{Ent}(\rho_1) - \mathrm{Ent}(\rho_0)\right).$$

Divergence based methods

- Divergence based method is distinct from usual dynamic techniques.
- Usually: only quadratic cost, Benamou-Breiner, Otto calculus.
- See Conforti & Tamanini '19 for one more term for the quadratic cost.
- Higher order terms should be related to higher order derivatives of divergence.

The Dirichlet transport

Dirichlet transport, P.-Wong '16

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

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

- K-L divergence or relative entropy as "distance":

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

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

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

Exponentially concave functions

- $\varphi : \Delta_n \to \mathbb{R} \cup \{-\infty\}$ 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

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

- Smooth, MTW Khan & Zhang '19.

Back to the Dirichlet transport

- What is the corresponding probabilistic picture for the cost function

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

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

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

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

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

Dirichlet transition

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

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

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

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

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

Multiplicative Schrödinger problem

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

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

- For ρ density on Δ_n, let

$$\mathrm{Ent}_0(\rho) = H\left(\rho \mid \mathrm{Dir}(0)\right).$$

Relative entropy w.r.t. Haar measure.

A tabular comparison

Group	$(\mathbb{R}^n, +)$	$(\Delta_n, \odot)$
Id	0	$e = (1/n, \ldots, 1/n)$
Cost	$\|y - x\|^2$	$H(e \mid q \circ p^{-1})$
Potential	convex	exp-concave
Monge solution	$y = \nabla \phi(x)$	$q = \widetilde{\nabla} \varphi(p)$
Displacement	$y - x$	$\boldsymbol{\pi}(p) = q \circ p^{-1}.$
Stochastic transition	Add Gaussian	Multiply Dirichlet
Haar measure	Leb	$\mathrm{Dir}(0)$
Entropy	Standard	Ent_0

Pointwise convergence

Theorem (P. '19)

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

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

- $\mathbf{C}\left(\rho_0, \rho_1\right)$ 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).

Connections to gradient flow of entropy

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 $h > 0$.

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

- FOC:

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

Heat equation as a gradient flow of entropy

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

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

- 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,

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

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

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

- This "morally" implies gradient flow of entropy.

Gradient flow without a metric?

- Dirichlet transport has a similar structure.

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

- 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`