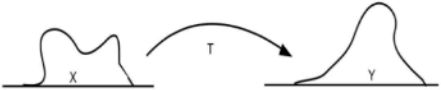
# An overview of entropy-regularized optimal transport and Schrödinger bridges

Soumik Pal
University of Washington, Seattle
CNI Seminar IISc Bangalore

Dec 10, 2024

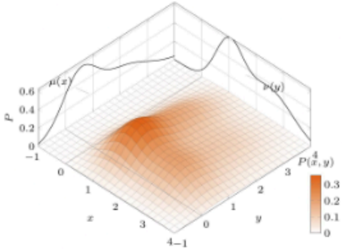
# The Monge problem 1781

- $P, Q$ - probabilities on $\mathcal{X} = \mathbb{R}^d = \mathcal{Y}$.
- Minimize among $T : \mathbb{R}^d \to \mathbb{R}^d$, $T(X) \sim Q$, if $X \sim P$, $\mathrm{E}\,\|T(X) - X\|^2$.



# Couplings

- $\mu, \nu$ probability measures on $\mathbb{R}^d$.
- Coupling of $(\mu, \nu)$ is a joint distribution with marginals $\mu$ and $\nu$.




- $\Pi(\mu, \nu)$ - set of couplings of $(\mu, \nu)$.
- $(X, T(X))$, if exists, is a coupling.

Image by M. Cuturi

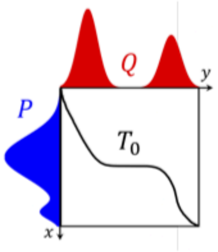
# The Monge-Kantorovich problem

- (Kantorovich '38) Minimize over $\Pi(\mu,\nu)$

$$\mathbb{W}_2^2(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \mathrm{E}_\gamma \left[\|Y - X\|^2\right].$$

- Linear optimization in $\gamma$ over convex $\Pi(\mu,\nu)$.
- Birth of linear programming. Dantzig '49.
- Lower semicontinuity + weak compactness $\rightarrow$ Existence of optimal coupling.
- How does the optimal coupling look like?

# Brenier's Theorem




- Suppose $\mu$ has density. Then unique solution to the MK problem.
- The optimal coupling is supported on a graph. **Monge map**.

$$\gamma = (\mathbf{id}, \nabla\phi)_{\#\mu} = \mathrm{Law}(X, \nabla\phi(X)), \quad X \sim \mu.$$

- $\phi : \mathbb{R}^d \to \mathbb{R}$ is a convex function.

Image by J.-C. Hütter

# Why the sudden interest of OT in statistics, ML, AI etc. ?

- OT is everywhere in stat/ML/generative AI
- More robust that Kullback-Leibler. $\mathbb{W}_2^2(\mu, \nu) < \infty$ even when disjoint support

$$\mathrm{KL}\left(Uni(2,3) \mid Uni(0,1)\right) = \infty, \ \mathbb{W}_2(Uni(2,3), Uni(0,1)) = 2.$$

- Manifold learning
- Regression with "uncoupled" data, e.g., single cell genomics
- Matching problems in continuum
- Computer vision and graphics
- Sampling, image generation
- Any problem with an underlying geometry $\mathbb{W}_2(\delta_x, \delta_y) = \|y - x\|$.

# Entropy

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

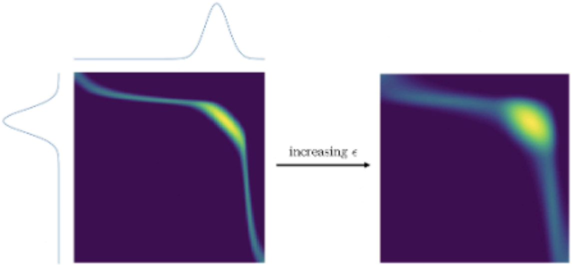
$$\mathrm{Ent}(\nu) := \begin{cases} \int f(x) \log f(x) dx, & \text{if } \nu \text{ has a density } f, \\ \infty, & \text{otherwise.} \end{cases}$$

- Example: Entropy of $N(0, \sigma^2)$ is $-\log \sigma +$ constant.
- Kullback-Leibler/ Relative entropy:

$$KL(P \mid R) = \int \log \frac{dP}{dR} dP,$$

if $P \ll R$ and $+\infty$, otherwise.

# Entropic regularization




- Föllmer '88, Galichon and Salanié '09, Cuturi '13 ... suggested penalizing MK OT with entropy.

$$EOT_\epsilon(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left[ \int \|y-x\|^2 \, d\gamma + \epsilon \mathrm{Ent}(\gamma) \right].$$

- Optimal coupling is called Schrödinger bridge at temperature $\epsilon$.

# Structure of the solution

- (Fortet '40, Rüschendorf & Thomsen '93) Schrödinger bridge admits a joint density. $\exists\, u^\epsilon, v^\epsilon : \mathbb{R}^d \to \mathbb{R}$,

$$\gamma^\epsilon(x,y) = \exp\left( -\frac{1}{2\epsilon}\|y-x\|^2 - \frac{1}{\epsilon}u^\epsilon(x) - \frac{1}{\epsilon}v^\epsilon(y) - f(x) - g(y) \right).$$

- $u^\epsilon, v^\epsilon$ - Schrödinger potentials. Unique up to constant.
- Typically not explicit. Determined by marginal constraints

$$\int \gamma^\epsilon(x,y)dy = e^{-f(x)}, \quad \int \gamma^\epsilon(x,y)dx = e^{-g(y)}.$$

- One approximate the Monge map by the barycentric projection

$$x \mapsto \mathrm{E}_{\gamma_\epsilon}(Y \mid X = x).$$

# Sinkhorn algorithm

- The proof by Fortet uses an iterative algorithm since called Sinkhorn/IPF.
- $\epsilon = 1$. $\mu, \nu$ uniform on $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, -1\}$. Initialize:

$$\begin{bmatrix} 1 & e^{-1/2} \\ e^{-1/2} & e^{-2} \end{bmatrix}.$$

- Make row sums $(1/2, 1/2)$.

$$\begin{bmatrix} \frac{1}{2}(1+e^{-1/2})^{-1} & \frac{1}{2}e^{-1/2}(1+e^{-1/2})^{-1} \\ \frac{1}{2}(e^{-1/2}+e^{-2})^{-1}e^{-1/2} & \frac{1}{2}(e^{-1/2}+e^{-2})^{-1}e^{-2}. \end{bmatrix} \approx \begin{bmatrix} 0.3 & 0.2 \\ 0.4 & 0.1 \end{bmatrix}.$$

- Make column sums $(1/2, 1/2)$.

$$\begin{bmatrix} 3/14 & 1/3 \\ 4/14 & 1/6. \end{bmatrix}$$

- And so on ....

# The Sinkhorn revolution

- Solving OT on finite data is an LP problem. Complexity $= \tilde{O}(n^3)$.
- Galichon & Salanié '09, Cuturi '13 proposed the Sinkhorn algorithm.
- Highly parallelizable on GPUs.
- (Altschuler et al. '17) Complexity$= \tilde{O}(n^2)$.

# Entropic Regularization

## Applications

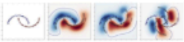
Distance between probability measures ($W_2$)


Bag-of-words models
(Rolet, Cuturi, Peyré, 2016)


Multi-label prediction
(Frogner et al., 2015)

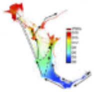
Wasserstein GAN
(Arjovsky, Chintala, Bottou, 2017)

Uncoupled function estimation ($T_0$)



Domain adaptation
(Courty, Flamary, Tuia, 2017)

Color transfer
(Rabin, Delon, Gousseau, 2010)



Trajectory inference in scRNA-Seq
(Schiebinger, Shu, Tabaka, et al., 2019)

Image by J.-C. Hütter

# Exponential convergence for $\epsilon > 0$

- The matrix algorithm is known to converge exponentially fast for fixed $\epsilon > 0$ under assumptions (Birkoff '57).
- Recent literature admits unbounded support with tail restrictions. See Conforti-Durmus-Greco '23, Ghosal-Nutz '22, Eckstein '23.
- All these results give convergence rates (in TV/ Wasserstein/ KL ) bounded by

$$C_\epsilon \kappa_\epsilon^n, \quad C_\epsilon > 0, \; \kappa_\epsilon \in (0,1), \; n = \text{iteration}.$$

- As $\epsilon \downarrow 0$, constants explode **badly**. Say $C_\epsilon = \exp(\mathrm{poly}(1/\epsilon))$.
- The "low teamperature" behavior is not understood. See Deb-Kim-P.-Schiebinger '23. Mirror gradient flows.

# Limiting results

$$EOT_\epsilon(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left[\int \|y-x\|^2 \, d\gamma + \epsilon \mathrm{Ent}(\gamma)\right].$$

- What happens as $\epsilon \to 0+$? (Mikami '04, Léonard '12)

$$\lim_{\epsilon\to 0+} EOT_\epsilon(\mu,\nu) = W_2^2(\mu,\nu)$$

  due to Large Deviations.
- Schrödinger bridge $\gamma_\epsilon \to$ Monge map.
- (P. '19, Conforti+Tamanini '19) Rate of convergence.

$$\lim_{\epsilon\to 0+} \frac{1}{\epsilon}\left(EOT_\epsilon(\mu,\nu) - W_2^2(\mu,\nu)\right) = \mathrm{Ent}(\mu) + \mathrm{Ent}(\nu).$$

# Schrödinger's lazy gas experiment

- $R$ = Law of reversible Brownian motion $X$ - diffusion $\epsilon$.
- "Condition" $X_0 \sim \mu$, $X_1 \sim \nu$. $P$ - Law on path space,
- Schrödinger '31, Föllmer '88. Dynamic Schrödinger bridge.
- The joint distribution $P\#(X_0, X_1)$ is the Schrödinger bridge.
- Given end points, particle follows Brownian bridge.

# An extremely short review of statistical issues

- A lot of questions arise from estimation of OT and EOT from data.
- Consider $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$ and $EOT_\epsilon(\hat{\mu}_n, \hat{\nu}_n)$.

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad X_i \sim \mu. \quad \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}, \quad Y_j \sim \nu.$$

- (Fournier & Guillin '15) Convergence of $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$ to $W_2^2(\mu, \nu)$ is $O(n^{-2/d})$. Also see Horowitz and Karandikar '94.
- (Mena and Niles-Weed '19) If $\mu, \nu$ are sub-Gaussian, $EOT_\epsilon(\hat{\mu}_n, \hat{\nu}_n)$ converges at $O(n^{-1/2})$. Also see Strommae '22.
- CLTs are recently proved (Gonzalez-Sanz, Loubes and Niles-Weed '22) but LDs are not known.
- For other variants, see Harchaoui-Liu-P. '19. Explicit solutions. Similar properties.

Iterated Schrödinger bridge approximation to Wasserstein gradient flows.
Joint work with M. Agarwal, Z. Harchaoui and G. Mulcahy.

Arxiv [math.PR] 2406.10823

# Application of Theorem

Self-attention dynamics of Transformer neural architecture (Vaswani et al. '17, Sander et al '22, Geshkovski et al '24)

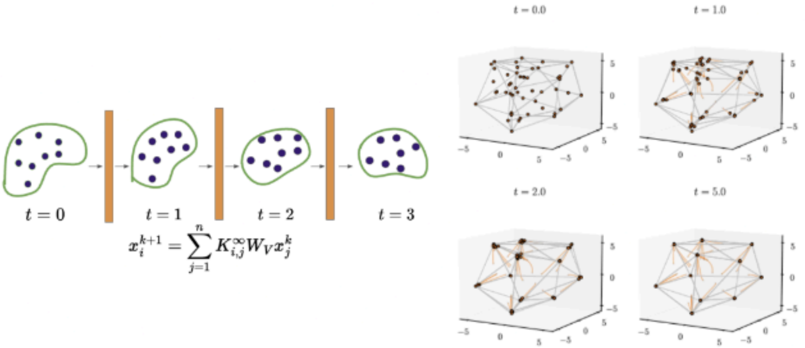



Figure: Self attention of Sinkformer SABP'22 (left) and Transformer GLPR '24 (right)

# A novel discrete scheme

- Start with $\rho_0$. Schrödinger Bridge $\gamma_\epsilon(\rho_0, \rho_0)$. Temperature $\epsilon \approx 0$.
- Compute barycentric projection

$$\mathcal{B}_0(x) = \mathrm{E}_{\gamma_\epsilon(\rho_0,\rho_0)}[Y \mid X = x] \approx x.$$

- Define

$$\rho_1(\epsilon) = (2\mathrm{id} - \mathcal{B}_0)\#\rho.$$

- I.e., if $X_0 \sim \rho_0$, then $X_1 := (2X_0 - \mathcal{B}_0(X_0)) \sim \rho_1$.

# A novel discrete scheme contd.

- Now iterate. For each $\rho_k(\epsilon)$, compute Schrödinger bridge $\gamma_\epsilon(\rho_k, \rho_k)$.
- Compute barycentric projection

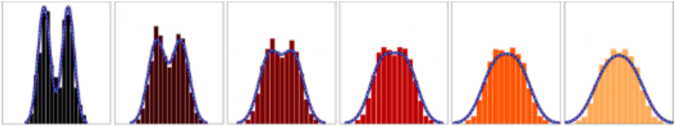
$$\mathcal{B}_k(x) = \mathrm{E}_{\gamma_\epsilon(\rho_k,\rho_k)}\left[Y \mid X = x\right].$$

- Define

$$\rho_{k+1}(\epsilon) = (2\mathrm{id} - \mathcal{B}_k)\,\#\rho_k.$$

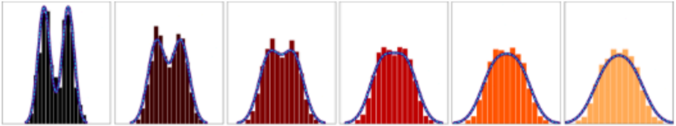
- I.e., if $X_k \sim \rho_k$, then $X_{k+1} := (2X_k - \mathcal{B}_k(X_k)) \sim \rho_{k+1}$.
- As $\epsilon \to 0+$, where does this sequence $(\rho_k)$ converge?

# Where does it converge?



- Scale iterations by $\epsilon$.
- What is the limit of $\left(\rho^{\epsilon}_{\lfloor t/\epsilon \rfloor},\ t \geq 0\right)$ as $\epsilon \to 0$?

# Where does it converge?



- Scale iterations by $\epsilon$.
- What is the limit of $\left(\rho^{\epsilon}_{\lfloor t/\epsilon \rfloor},\ t \geq 0\right)$ as $\epsilon \to 0$?
- Theorem. (P. et al. '24) Under assumptions, heat flow starting with $\rho_0$.

$$\dot{\rho}_t = \frac{1}{2}\Delta \rho_t.$$

- Originally observed by Sander-Ablin-Blondel-Peyré '22 in their analysis of Transformers.

# Brief idea of proof

- For $\epsilon \approx 0$,

$$\mathrm{E}_{\gamma_\epsilon(\rho,\rho)}[Y \mid X = x] \approx x + \frac{1}{2}\epsilon \nabla \log \rho(x).$$

- Hence,

$$2x - \mathrm{E}_{\gamma_\epsilon(\rho,\rho)}[Y \mid X = x] \approx x - \frac{1}{2}\epsilon \nabla \log \rho(x).$$

- $X_{k+1} \approx X_k - \frac{\epsilon}{2}\nabla \log \rho_k(X_k)$. Euler iterations for the ODE:

$$\dot{x}_t = -\nabla \log \rho_t(x), \quad \rho_t = \rho_0 \# x_t.$$

- $(\rho_t,\ t \geq 0)$ satisfies the heat equation

$$\dot{\rho}_t = \frac{1}{2}\nabla \cdot (\rho_t \nabla \log \rho_t) = \frac{1}{2}\Delta \rho_t.$$

# Brief idea of proof

- How do we approximate the Schrödinger bridge at low temperatures?
- Let $(Z_t,\ t \geq 0)$ denote the stationary Langevin diffusion with law $\rho$.

$$dZ_t = \frac{1}{2}\nabla \log \rho(Z_t)dt + dB_t,\ Z_0 \sim \rho.$$

- Theorem. (P. et al '24) $\gamma_\epsilon(\rho, \rho) \approx$ the law $\ell_\epsilon(\rho)$ of $(Z_0, Z_\epsilon)$,

$$H\left(\gamma_\epsilon \mid \ell_\epsilon\right) + H\left(\ell_\epsilon \mid \gamma_\epsilon\right) = o(\epsilon^2).$$

- From the diffusion SDE

$$\mathrm{E}\left(Z_\epsilon \mid Z_0 = x\right) \approx x + \frac{\epsilon}{2}\nabla \log \rho(x).$$

# Concluding remarks

- Sander et al '22 proposed changing the weight matrix to be doubly stochastic.
- As an output of the Sinkhorn algorithm.
- The main claim: dynamics of the self-attention converges to the heat flow.
- Our theorems in P. et al '24 justify the claim in continuum.
- Convergence of the particle system remains open.
- The main challenge is to prove consistency of the estimation of score function.

# A curious example

- For each $\rho_k(\epsilon)$, compute Schrödinger bridge $\gamma_\epsilon(\rho_k, \rho_k)$.
- Compute barycentric projection

$$\mathcal{B}_k(x) = \mathrm{E}_{\gamma_\epsilon(\rho_k,\rho_k)}\left[Y \mid X = x\right].$$

- Define

$$\rho_{k+1}(\epsilon) = (\mathcal{B}_k)\#\rho_k.$$

# Reversing the heat flow

- If $X_k \sim \rho_k$, then $X_{k+1} := \mathcal{B}_k(X_k) \sim \rho_{k+1}$.
- As $\epsilon \to 0+$, where does this sequence $(\rho_k)$ converge?
- Backward heat equation, for small enough $\epsilon$!
- No proof. Gaussian computations in P. et al '24.

# Generalizations

- We *can* generalize to other AC curves. General idea:

$$\dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = \nabla \phi_t.$$

- Define a "surrogate density" $\sigma_t \propto \exp(\pm 2\phi_t)$. Assume integrable.

$$\mathrm{E}_{\gamma_\epsilon(\sigma_t, \sigma_t)}[Y \mid X = x] \approx x + \frac{\epsilon}{2} v_t(x).$$

- "Geodesic approximation" may be substituted by Sinkhorn algorithm.
- Does not require estimating the "score function".

Thank you for your attention