Heat Flow through Pretrained Transformer

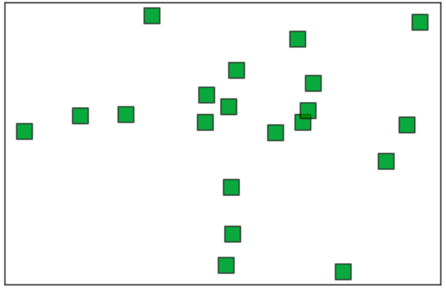
Joint Mathematics Meeting 2025

Medha Agarwal
Joint work with ***Garrett Mulcahy, Soumik Pal,*** *and* ***Zaid Harchaoui***

January 09, 2025

Transformer Self-Attention Mechanism

Transformer Self-Attention Mechanism



$\{x_1, \ldots, x_n\} \subset \mathbb{R}^d \sim \rho^0$

Self Attention Layer

$$x_i \leftarrow \sum_{j=1}^{n} K_{ij}^1 W_V x_j$$

Transformer Self-Attention Mechanism

$$x_i \leftarrow \sum_{j=1}^{n} K^1_{ij} W_V x_j$$

Value matrix $\in \mathbb{R}^{q \times d}$

$$K^1 = \texttt{Softmax}(C)$$

$$C_{i,j} = (W_Q x_i)^\top (W_K x_j)$$

Query matrix $\in \mathbb{R}^{p \times d}$

Key matrix $\in \mathbb{R}^{p \times d}$

Previous Work on Understanding Self-Attention

- [VBC20] formulated the self-attention mechanism as a non-linear transformation on probability measures.
- [GLPR24] derived the Lipschitz coefficient of self-attention mechanism.
- [CAP24] extended the analysis of Lipschitz coefficient to masked self-attention within a mean-field framework.
- We derive the mean-field limit of **Sinkformers**, proposed by [SABP22].

Recall

$$x_i \leftarrow \sum_{j=1}^{n} K_{ij}^1 W_V x_j$$

Value matrix $\in \mathbb{R}^{q \times d}$

$$K^1 = \texttt{Softmax}(C)$$

$$C_{i,j} = (W_Q x_i)^\top (W_K x_j)$$

Query matrix $\in \mathbb{R}^{p \times d}$

Key matrix $\in \mathbb{R}^{p \times d}$

Sinkformer [SABP22] Self-Attention Mechanism

$$x_i \leftarrow \sum_{j=1}^{n} K^{\infty}_{i,j} W_V x_j$$

Value matrix $\in \mathbb{R}^{q\times d}$

$$K^{\infty} = \mathtt{Sinkhorn}(C)$$

$$C_{i,j} = (W_Q \boldsymbol{x}_i)^{\top}(W_K \boldsymbol{x}_j)$$

Query matrix $\in \mathbb{R}^{p\times d}$

Key matrix $\in \mathbb{R}^{p\times d}$

Sinkformer Self-Attention Mechanism

$$x_i \leftarrow \sum_{j=1}^{n} K^{\infty}_{i,j} W_V x_j$$

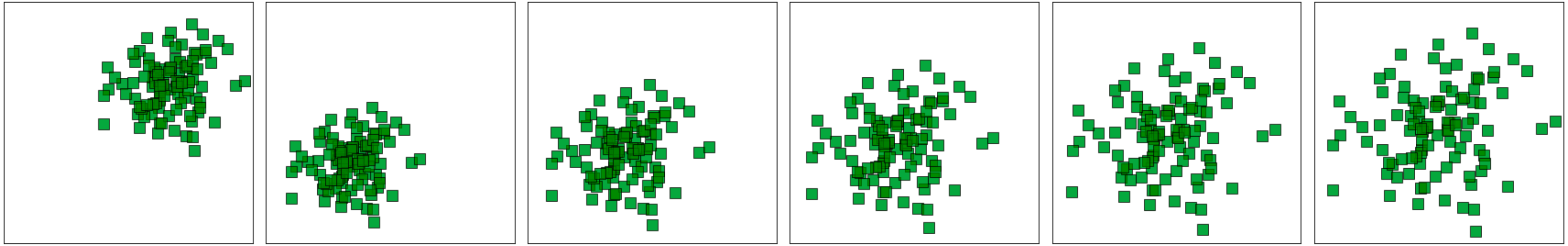
K^{∞} is obtained via Sinkhorn algorithm [Cut13]

- Initialize $K^0 = \exp(C)$.
- Update $K^{\ell+1} = \begin{cases} N_R(K^{\ell}) & \text{if } \ell \text{ is even,} \\ N_C(K^{\ell}) & \text{if } \ell \text{ is odd .} \end{cases}$
- N_R is row normalization and N_C is column normalization.

Finite particles

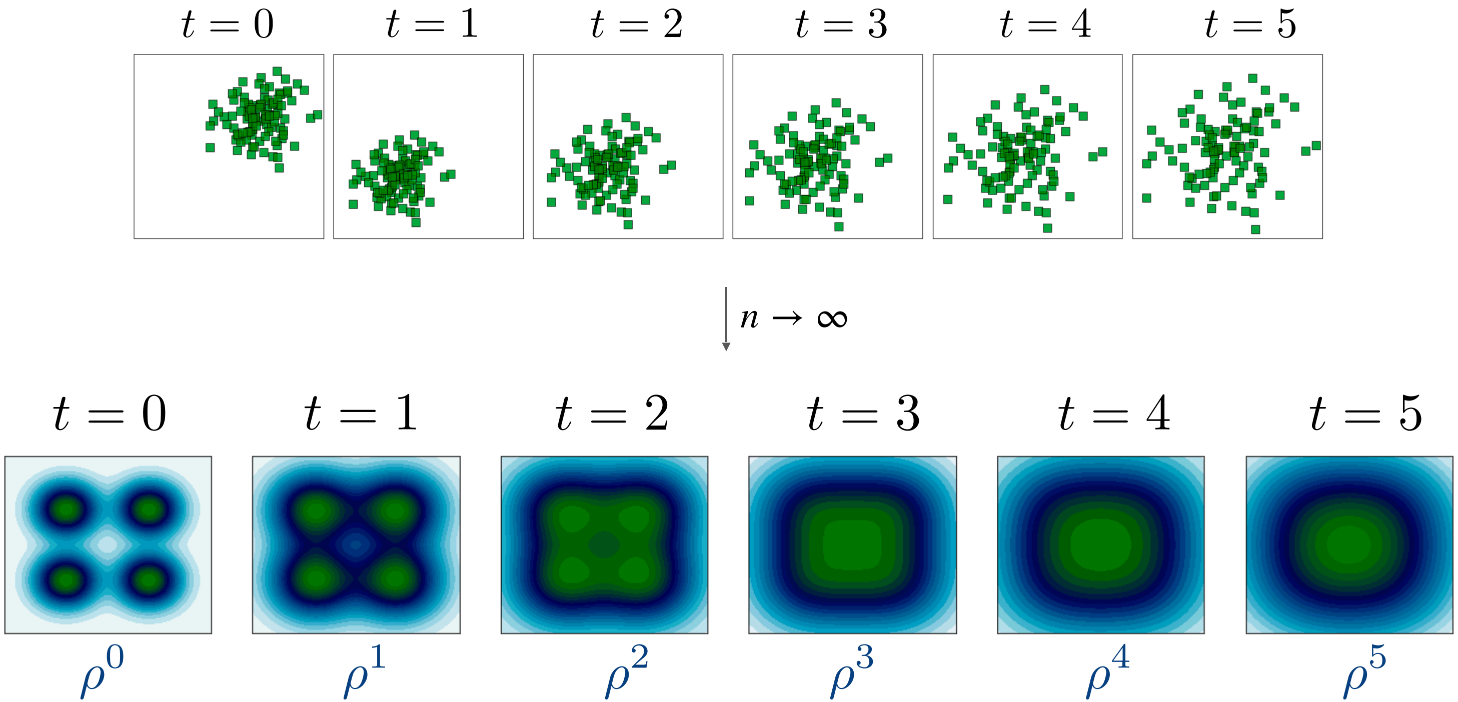
$$x_i = \sum_{j=1}^{n} K_{i,j}^{\infty} W_V x_j$$

$t = 0$ $t = 1$ $t = 2$ $t = 3$ $t = 4$ $t = 5$

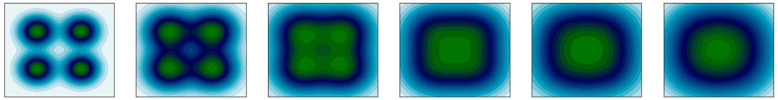


The position of each particle is influenced by the overall distribution.

Infinite particles



Infinite particles



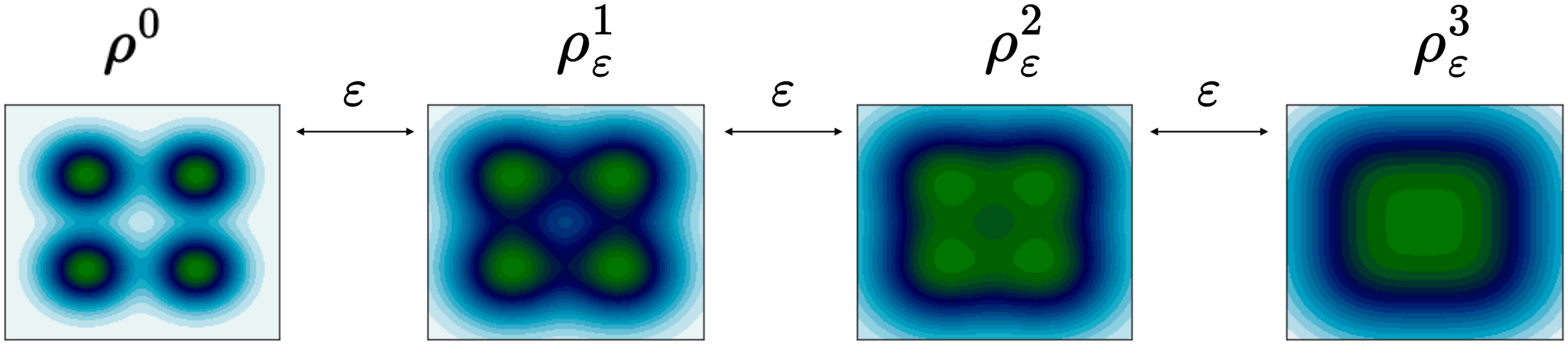
$$\rho^{k+1} = \left(T_{\rho^k}\right)_{\#} \rho^k \quad \text{where} \quad T_{\rho^k} = \int k^{\infty}(x, y)\, W_V y \, d\rho^k(y)$$

K^{∞} **is obtained via Sinkhorn algorithm [Cut 13]**

- Initialize $k^0 = \exp(c)$ where $c(x, y) = (W_Q x)^{\top}(W_K y)$.
- Update $k^{\ell+1}(x, y) = \begin{cases} \dfrac{k^{\ell}(x, y)}{\int k^{\ell}(x, y)\, d\rho^k(y)} & \text{if } \ell \text{ is even,} \\[2ex] \dfrac{k^{\ell}(x, y)}{\int k^{\ell}(x, y)\, d\rho^k(x)} & \text{if } \ell \text{ is odd .} \end{cases}$

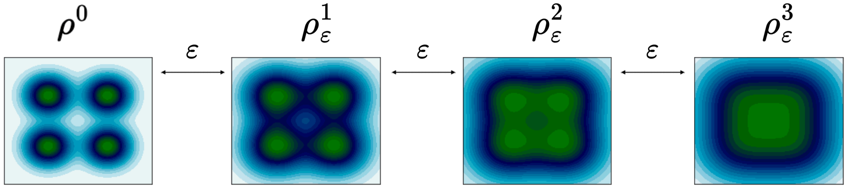
What is the continuous-time counterpart of this discrete-time process?

Temperature Parameter



What happens if $\varepsilon \to 0+$?

Temperature Parameter



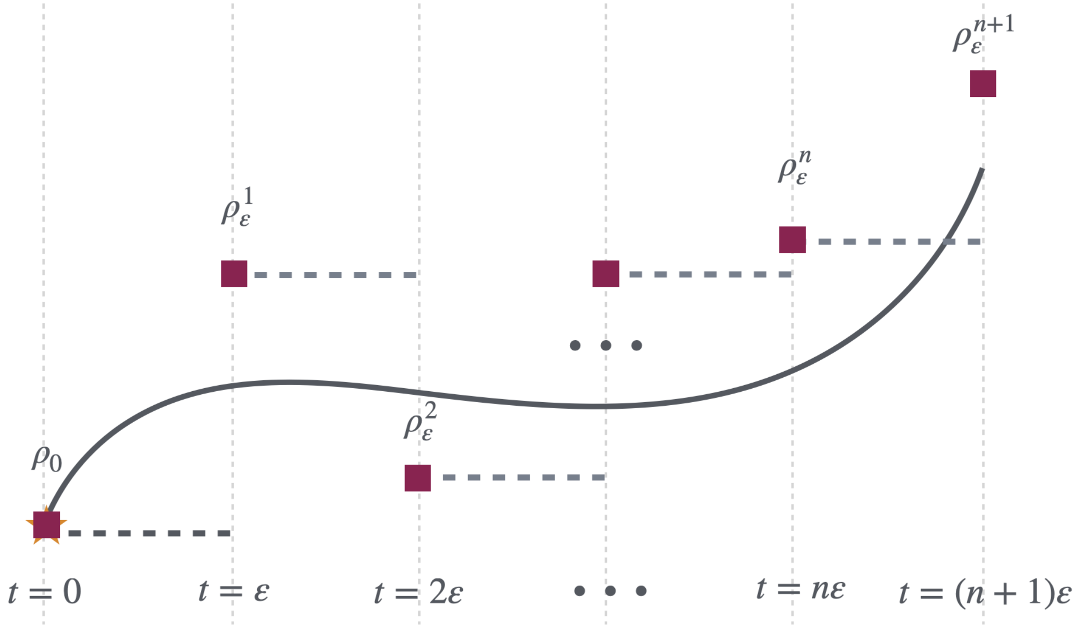
- Concretely, $\rho_\varepsilon^{k+1} = \left(T_{\rho_\varepsilon^k,\varepsilon}\right)_\# \rho_\varepsilon^k$,

- $T_{\rho_\varepsilon^k,\varepsilon}(x) = \displaystyle\int k_\varepsilon^\infty(x,y)\, W_V y \, d\rho_\varepsilon^k(y)$

- The Sinkhorn kernel is $k_\varepsilon^\infty = \mathrm{Sinkhorn}(c/\varepsilon)$.

What happens if $\varepsilon \to 0+$?

Define $\rho_\varepsilon(t) = \rho_\varepsilon^k$ for $t \in [k\varepsilon, (k+1)\varepsilon)$.



What happens if $\varepsilon \to 0+$?

Is there a curve $(\rho(t), t \geq 0)$ such that $(\rho_\varepsilon(t), t \geq 0)$ converges uniformly it as $\varepsilon \to 0$?

Let's dive deeper...

Under assumption $W_K^\top W_Q = W_Q^\top W_K = -W_V = I$,

the infinite particles **Sinkformer self-attention update** is

$$\rho_\varepsilon^{k+1} = \left(2I - \int k_\varepsilon^\infty(x,\cdot)d\rho_\varepsilon^k(x)\right)_\# \rho_\varepsilon^k$$
$$= \underbrace{\left(2I - \mathcal{B}_{\rho_\varepsilon^k,\varepsilon}\right)}_{\text{Barycentric projection}}{}_\# \rho_\varepsilon^k$$

Claim

[SABP22] hypothesize that scheme $\left(\rho_\varepsilon^k, k \geq 0\right)$ converges uniformly to a heat flow. Consider,

Self-attention flow

$$\boxed{\rho_\varepsilon(t) = \rho_\varepsilon^k \text{ for } t \in [k\varepsilon, (k+1)\varepsilon)}$$

Heat flow

$$\boxed{\partial_t \rho(t,x) = \Delta_x \rho(t,x)}$$

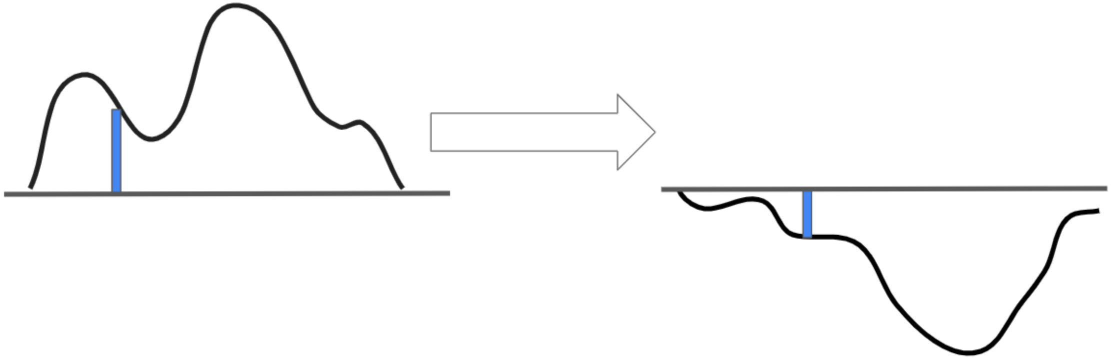
Concretely, let $\left(\rho(t), t \geq 0\right)$ be the heat flow. Then, for a fixed $T > 0$,

$$\boxed{\lim_{\varepsilon \to 0} \sup_{t \in [0,T]} \mathbb{W}_2\left(\rho_\varepsilon(t), \rho(t)\right) = 0}$$

Understanding $\mathcal{B}_{\rho,\varepsilon}$ via Entropy-regularized Optimal Tranport

$$\rho_\varepsilon^{k+1} = \left(2I - \mathcal{B}_{\rho,\varepsilon}\right)_\# \rho_\varepsilon^k$$

Introduction to (Entropy Regularized) Optimal Transport



Monge Mass Transport Problem

Notation

Coupling of Measures

Given $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, we say $\gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is a coupling (transport plan) between μ and ν, denoted by $\gamma \in \Pi(\mu, \nu)$, if for all measurable $A, B \subset \mathbb{R}^d$

$$\gamma(A \times \mathbb{R}^d) = \mu(A) \text{ and } \gamma(\mathbb{R}^d \times B) = \nu(B)$$

Transport Map

A measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$ is a push forward from μ to ν, denoted by $T_\# \mu = \nu$, if for all measurable $A \subset \mathbb{R}^d$,

$$\nu(A) = \mu(T^{-1}(A))$$

$T_\# \mu = \nu$ if and only if $(Id, T)_\# \mu \in \Pi(\mu, \nu)$.

Optimal Transport

The **optimal transport problem** is then given by

$$\mathbb{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\gamma$$

Brenier's Theorem gives the structure of optimal coupling. Under moderate assumptions:

$$\gamma^* = (\mathrm{Id},\, T)_{\#}\mu$$

where T is the unique gradient of a convex function.

Entropic Regularization

Problem: Unable to efficiently calculate OT cost and OT maps.
Solution: Regularization by relative entropy (KL Divergence)

Entropy-regularized optimal transport problem

$$\boxed{\inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\gamma + \varepsilon H(\gamma | \mu \times \nu) \right)}$$

$$H(\alpha \,|\, \beta) = \int_{\mathbb{R}^d} \log(\alpha/\beta) d\alpha$$

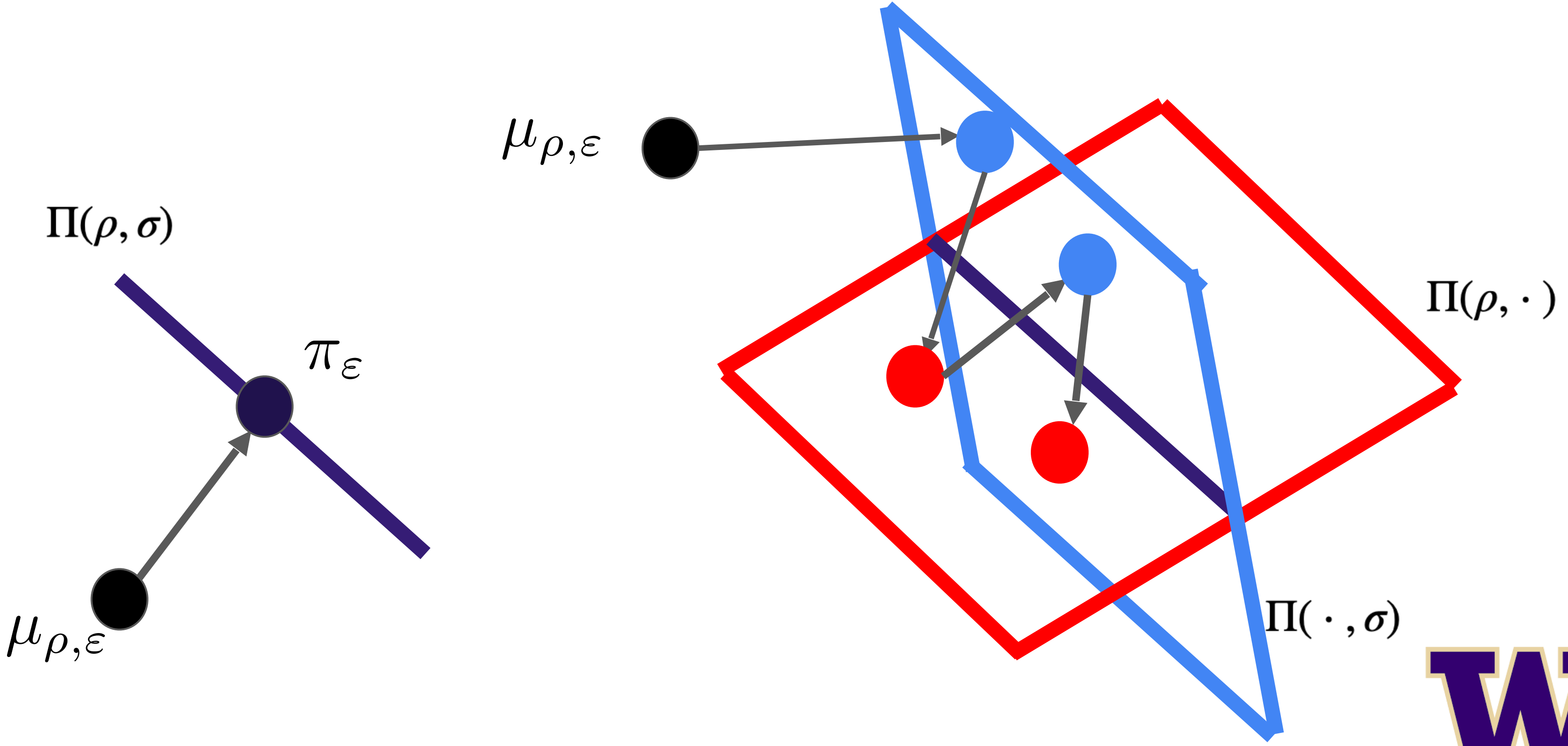
The argmin of the above problem, denoted by π_ε is the **Schrödinger bridge (SB)** from μ to ν.

Define the **barycentric projection** as the function

$$\boxed{\mathcal{B}_{\mu,\nu,\varepsilon}(x) := \mathbb{E}_{\pi_\varepsilon}[Y | X = x]}$$

Sinkhorn Algorithm

Compute Schrödinger Bridges in near linear time [Cut13] via **alternating projections**



Self-attention update via same marginal Schrödinger bridges

In this work, we assume $\mu = \nu$.

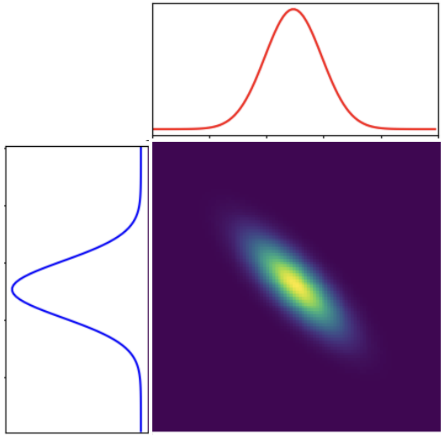
Let $\pi_{\rho,\varepsilon}$ be the Schrödinger bridge from ρ to itself and

$$\mathcal{B}_{\rho,\varepsilon}(x) = \mathbb{E}_{\pi_{\rho,\varepsilon}}[Y|X = x]$$

Recall,

$$\rho_\varepsilon^{k+1} = (2I - \mathcal{B}_{\rho,\varepsilon})_\# \rho_\varepsilon^k = \left(\mathrm{Id} - \varepsilon\left(\frac{\mathcal{B}_{\rho,\varepsilon} - Id}{\varepsilon}\right)\right)_\# \rho_\varepsilon^k$$

Want to calculate precisely the deviation of BP from identity.



Schrödinger Bridge between two $N(0,1)$ random variables with $\epsilon = 0.01$

Three main results

Result 1: Same Marginal Schrödinger Bridge is close to law of Langevin diffusion

Theorem [AHMP24, Theorem 1]

Let $\rho = e^{-g}$ be a probability density on $\mathbb{R}^d$ with enough regularity such that there is a strong solution to the Langevin SDE $dX_t = \frac{1}{2}\nabla g(X_t)dt + dB_t$ with initial distribution $X_0 \sim \rho$. Let $\ell_{\rho,\varepsilon} = \text{Law}(X_0, X_\varepsilon)$, then

$$H(\ell_{\rho,\varepsilon} \,|\, \pi_{\rho,\varepsilon}) + H(\pi_{\rho,\varepsilon} \,|\, \ell_{\rho,\varepsilon}) \leq C\varepsilon^2 \left(I(\rho) + \int_0^1 I(\rho_t^\varepsilon)dt \right)^{1/2}.$$

In particular, the right hand side is $o(\varepsilon^2)$. $I(\alpha) = \int_{\mathbb{R}^d} \|\nabla \log \alpha\|^2 d\alpha$.

Heat Flow: Particle Approach

PDE (Evolution of Density)

$$\partial_t \rho(t,x) = \Delta_x \rho(t,x)$$

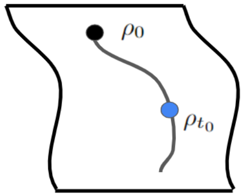
Particle Picture

Let $X_0 \sim \rho_0$ and consider the ODE

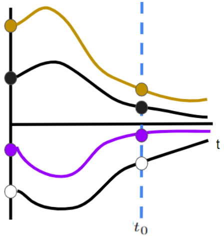
$$\dot{x}_t = v_t = -\frac{1}{2}\nabla \log \rho(t)$$

Then, $(x_t)_{\#}\rho_0 = \rho(t)$

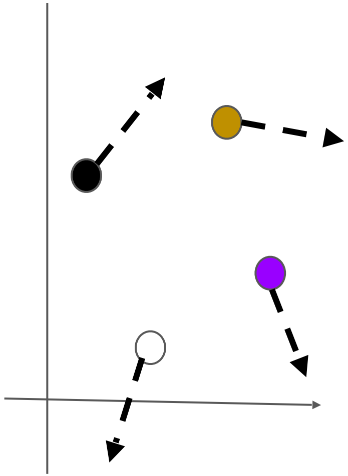
Flow of Measures



Particle Trajectories



$$\dot{x}_t(x) = -\frac{1}{2}\nabla \log \rho_t(x)$$



Result 2: One Step Approximation

Based on intuition from Result 1, $\pi_{\rho,\varepsilon} \approx \ell_{\rho,\varepsilon}$,

$$\mathcal{B}_{\rho,\varepsilon}(x) \approx \mathbb{E}_{\ell_{\rho,\varepsilon}}[Y|X=x] \approx x - \frac{\varepsilon}{2}\nabla g(x) = x + \frac{\varepsilon}{2}\nabla \log \rho(x)$$

Matches explicit Euler approximation from particle picture

Particle Picture

Let $X_0 \sim \rho_0$ and consider the ODE

$$\dot{x}_t = v_t = -\frac{1}{2}\nabla \log \rho(t)$$

Then, $(x_t)_{\#}\rho_0 = \rho(t)$

Takeaway: Can access **score function** via entropic OT objects, which can be **estimated** from samples!

Result 2: One Step Approximation

Explicit Euler Update

$$S_\varepsilon^1(\rho) = \left(\mathrm{Id} - \frac{\varepsilon}{2}\nabla \log \rho\right)_{\#} \rho$$

SB Update

$$SB_\varepsilon^1(\rho) = (2\,\mathrm{Id} - \mathcal{B}_{\rho,\varepsilon})_{\#}\,\rho$$

Theorem [AHMP24, Theorem 2]

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \mathbb{W}_2 \left(SB_\varepsilon^1(\rho), S_\varepsilon^1(\rho)\right) = 0$$

Result 3: Uniform Convergence

Define the SB and explicit Euler schemes for approximating $\left(\rho(t), t \in [0,T]\right)$. Let $N_\varepsilon = \lfloor N\varepsilon^{-1} \rfloor$, then for any $k \in [N_\varepsilon]$

Explicit Euler Update

$$SB_\varepsilon^{k+1}(\rho) = SB_\varepsilon^1(SB_\varepsilon^k(\rho))$$

SB Update

$$S_\varepsilon^{k+1}(\rho) = S_\varepsilon^1(S_\varepsilon^k(\rho))$$

Theorem 3 [AHMP 24]

The explicit Euler scheme converges to the heat equation uniformly from a starting measure $\rho_0 \in \mathscr{P}(\mathbb{R}^d)$ (satisfying some conditions), that is

$$\lim_{\varepsilon \downarrow 0} \sup_{k \in [N_\varepsilon]} \mathbb{W}_2\left(S_\varepsilon^k(\rho_0), \rho(k\varepsilon)\right) = 0$$

Result 3: Uniform Convergence

Theorem 3 [AHMP 24]

The explicit Euler scheme converges to the heat equation uniformly from a starting measure $\rho_0 \in \mathscr{P}(\mathbb{R}^d)$ (satisfying some conditions), that is

$$\lim_{\varepsilon \downarrow 0} \sup_{k \in [N_\varepsilon]} \mathbb{W}_2 \left(S_\varepsilon^k(\rho_0), \rho(k\varepsilon) \right) = 0$$

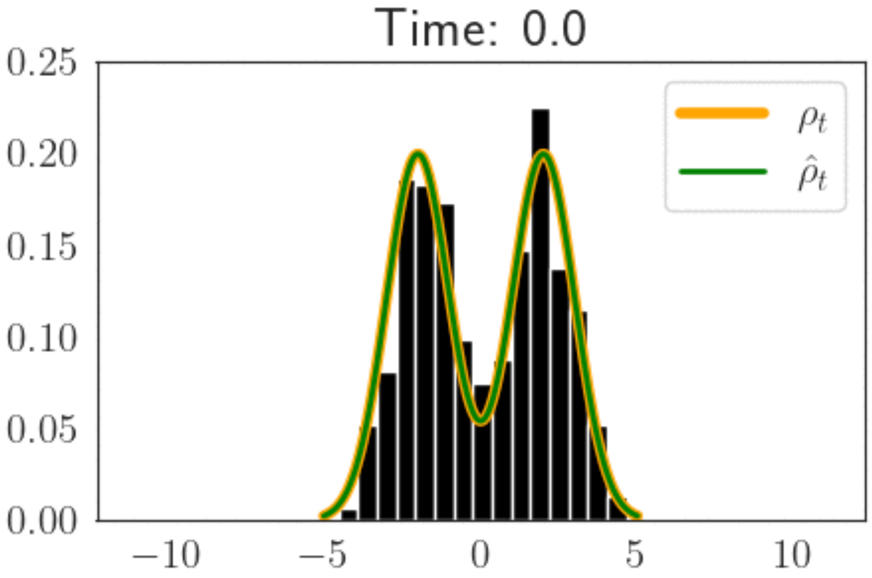
As a corollary,

$$\boxed{\lim_{\varepsilon \downarrow 0} \sup_{k \in [N_\varepsilon]} \mathbb{W}_2 \left(SB_\varepsilon^k(\rho_0), \rho(k\varepsilon) \right) = 0}$$

Simulations

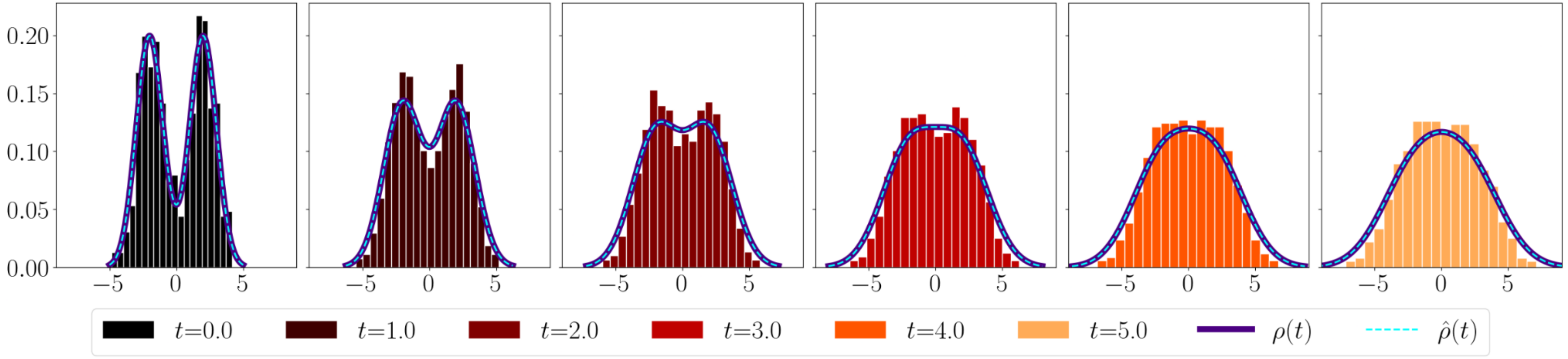
Mixture of Gaussians

$$\rho_0 = 0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1), \quad \varepsilon = 0.01$$



Mixture of Gaussians

$$\rho_0 = 0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1), \quad \varepsilon = 0.01$$



Thank you!
Questions?

The Team

Garrett Mulcahy
Mathematics, University of Washington

Zaid Harchaoui
Statistics, University of Washington

Soumik Pal
Mathematics, University of Washington

References

- [AHMP24] Agarwal, Medha, et al. "Iterated Schrödinger bridge approximation to Wasserstein Gradient Flows." *arXiv preprint arXiv:2406.10823* (2024).

- [CAP24] Val'erie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In ICML 2024, 2024.

- [Cut13] 3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26, 2013.

- [GLPR24] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. Advances in Neural Information Processing Systems, 36, 2024.

- [SABP22] Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyr´e. Sinkformers: Transformers with doubly stochastic attention. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 3515–3530. PMLR, 28–30 Mar 2022.

- [VBC20] James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. arXiv preprint arXiv:2007.02876, 2020.