Heat Flow through Doubly-Stochastic Self-Attention

Mathematics of Deep Learning Workshop
Institute for Foundations of Machine Learning

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

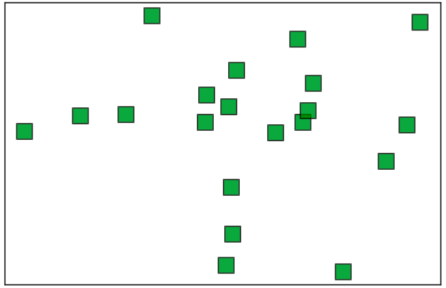
February 21, 2025

Outline

- Self-attention mechanism
- Doubly-stochastic self-attention mechanism
- Infinite-particles and time-discretized picture
- Relation between self-attention update and barycentric projections
- Three main results
- Simulations

Transformer Self-Attention Mechanism

Transformer Self-Attention Mechanism



$$\{x_1^0, \ldots, x_n^0\} \sim \rho^0$$

Self Attention Layer

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

$$x_i^k, i \in [n], k \in \mathbb{N}$$

Position of i-th particle after k-th self-attention update

Transformer Self-Attention Mechanism

$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^1 \, W_V x_j^k$$

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

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

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

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

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

Previous Work on Understanding Self-Attention

- [VBC20] formulated the self-attention mechanism as a non-linear transformation on probability measures.
- [GLPR23] derive the limiting geometry of particles undergoing self-attention updates for different configurations of query, key, and value matrices.
- [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].

Transformer Self-Attention Mechanism

$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^1 \, W_V x_j^k$$

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

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

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

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

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

Sinkformer [SABP22] Self-Attention Mechanism

$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^{\infty} W_V x_j^k$$

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

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

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

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

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

Sinkformer Self-Attention Mechanism

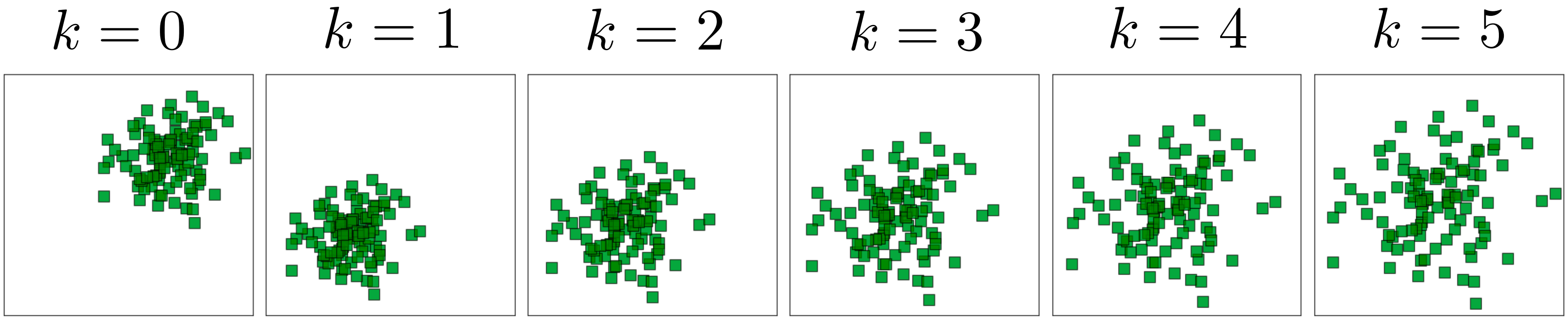
$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^{\infty} W_V x_j^k$$

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.

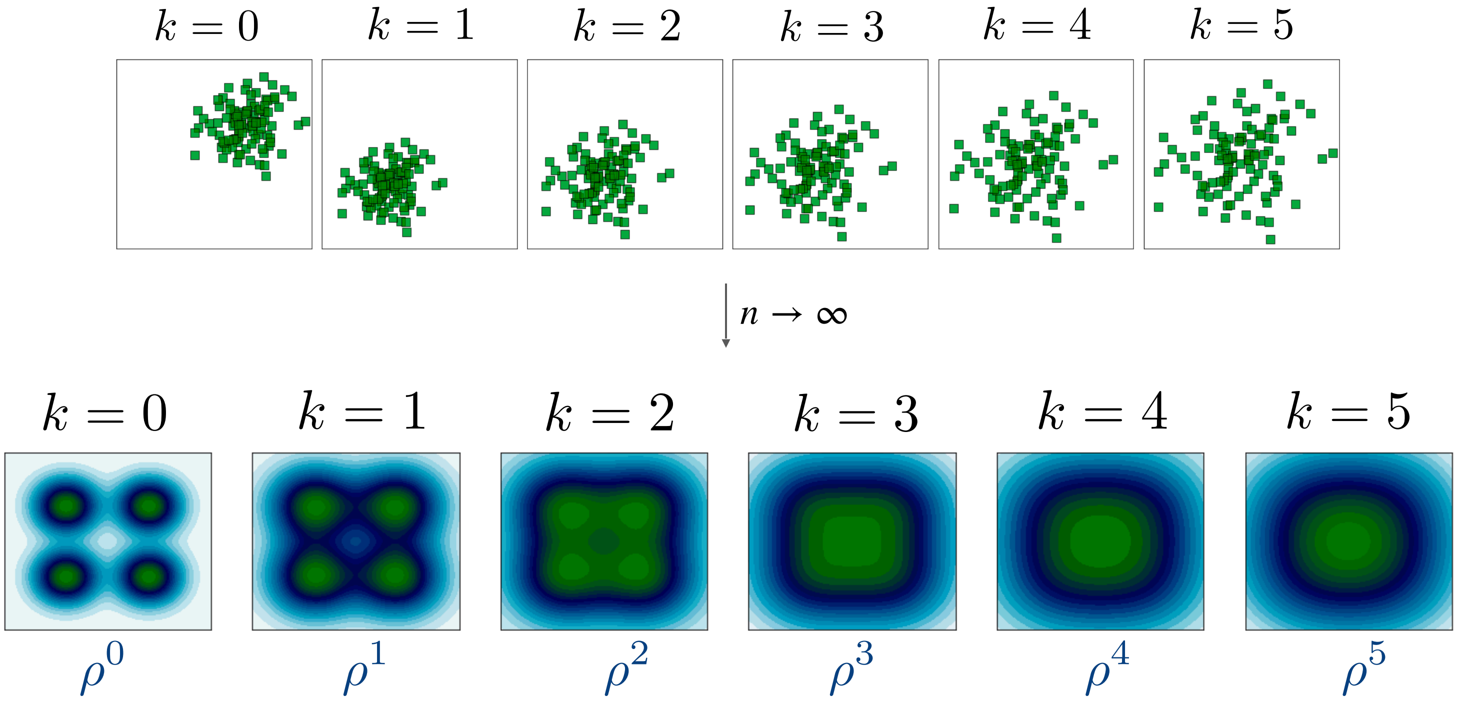
Evolution of interacting particles under doubly-stochastic self-attention

$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^{\infty} W_V x_j^k$$

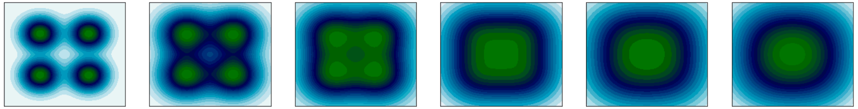


The position of each particle is influenced by the overall distribution.

Infinite particles



Infinite particles



$$x_i^{k+1} = x_i^k + \sum_{j=1}^{n} K_{ij}^{\infty} W_V x_j^k$$

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

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 $k^{\ell+1}(x, y) = \begin{cases} \dfrac{k^{\ell}(x, y)}{\int k^{\ell}(x, y)\, d\rho^k(y)} & \text{if } \ell \text{ is even,} \\ \dfrac{k^{\ell}(x, y)}{\int k^{\ell}(x, y)\, d\rho^k(x)} & \text{if } \ell \text{ is odd .} \end{cases}$

$$x^{k+1} = x^k + \int k^{\infty}(x^k, y)\, W_V y \, d\rho^k(y)$$

Normalized and time-discretized picture

Normalized and time-discretized picture

- The doubly-stochastic self-attention update is

$$x^{k+1} = x^k + \int k^\infty(x^k, y)\, W_V y \, d\rho^k(y)$$

- Viewing layers as time variable, the above update can be viewed as time discretization of the dynamic

$$\frac{d}{dt}x(t) = \int k^\infty(x(t), y)\, W_V y \, d\rho_t(y)$$

density of $x(t)$ at time t

- The ODE can be solved using forward Euler to obtain time-discretized self-attention updates

$$x((k+1)\varepsilon) = x(k\varepsilon) + \varepsilon \int k^\infty(x(k\varepsilon), y)\, W_V y \, d\rho_{k\varepsilon}(y)$$

Normalized and time-discretized picture

- The doubly-stochastic self-attention update is

$$x^{k+1} = x^k + \int k^\infty(x^k, y)\, W_V y\, d\rho^k(y)$$

- Viewing layers as time variable, the above update can be viewed as time discretization of the dynamic

$$\frac{d}{dt}x(t) = \int k^\infty(x(t), y)\, W_V y\, d\rho_t(y)$$

density of $x(t)$ at time t

- The ODE can be solved using forward Euler to obtain time-discretized self-attention updates

$$x((k+1)\varepsilon) = x(k\varepsilon) + \varepsilon \int k^\infty(x(k\varepsilon), y)\, W_V y\, d\rho_{k\varepsilon}(y)$$

Normalized and time-discretized picture

- The doubly-stochastic self-attention update is

$$x^{k+1} = x^k + \int k^\infty(x^k, y)\, W_V y\, d\rho^k(y)$$

- Viewing layers as time variable, the above update can be viewed as time discretization of the dynamic

$$\frac{d}{dt}x(t) = \int k^\infty(x(t), y)\, W_V y\, d\rho_t(y)$$

density of $x(t)$ at time t

- The ODE can be solved using forward Euler to obtain time-discretized self-attention updates

$$x((k+1)\varepsilon) = x(k\varepsilon) + \varepsilon \int k^\infty(x(k\varepsilon), y)\, W_V y\, d\rho_{k\varepsilon}(y)$$

Normalized continuous-time dynamics

- The ODE can be solved using forward Euler to obtain time-discretized self-attention updates

$$x((k+1)\varepsilon) = x(k\varepsilon) + \varepsilon \int k^{\infty}(x(k\varepsilon), y)\, W_V y \, d\rho_{k\varepsilon}(y)$$

- Here $x(t)$ will diverge to $+\infty$. Following [SABP22], we rescale tokens as $z(t) = e^{-tW_V} x(t)$ and introduce the bandwidth parameter $\varepsilon > 0$ (that scales $W_Q^\top W_K$ and W_V by ε) to get the update

$$z((k+1)\varepsilon) = z(k\varepsilon) + \left[\int k_{\varepsilon}^{\infty}(z(k\varepsilon), y)\, W_V y \, d\rho_{k\varepsilon}(y) - W_V z(k\varepsilon)\right]$$

k_{ε}^{∞} → Sinkhorn(c/ε)

- For finite particles

$$z_i^{k+1} = z_i^k + \left[\sum_{j=1}^{n} K_{\varepsilon\, ij}^{\infty}\, W_V z_j^k - W_V z_i^k\right]$$

Normalized continuous-time dynamics

- The ODE can be solved using forward Euler to obtain time-discretized self-attention updates

$$x((k+1)\varepsilon) = x(k\varepsilon) + \varepsilon \int k^{\infty}(x(k\varepsilon), y)\, W_V y \, d\rho_{k\varepsilon}(y)$$

- Here $x(t)$ will diverge to $+\infty$. Following [SABP22], we rescale tokens as $z(t) = e^{-tV} x(t)$ and introduce the bandwidth parameter $\varepsilon > 0$ (that scales $W_Q^\top W_K$ and W_V by ε) to get the update

$$z((k+1)\varepsilon) = z(k\varepsilon) + \left[\int k_{\varepsilon}^{\infty}(z(k\varepsilon), y)\, W_V y \, d\rho_{k\varepsilon}(y) - W_V z(k\varepsilon) \right]$$

Sinkhorn(c/ε)

- For finite particles

$$z_i^{k+1} = z_i^k + \left[\sum_{j=1}^{n} K_{\varepsilon\, ij}^{\infty} W_V z_j^k - W_V z_i^k \right]$$

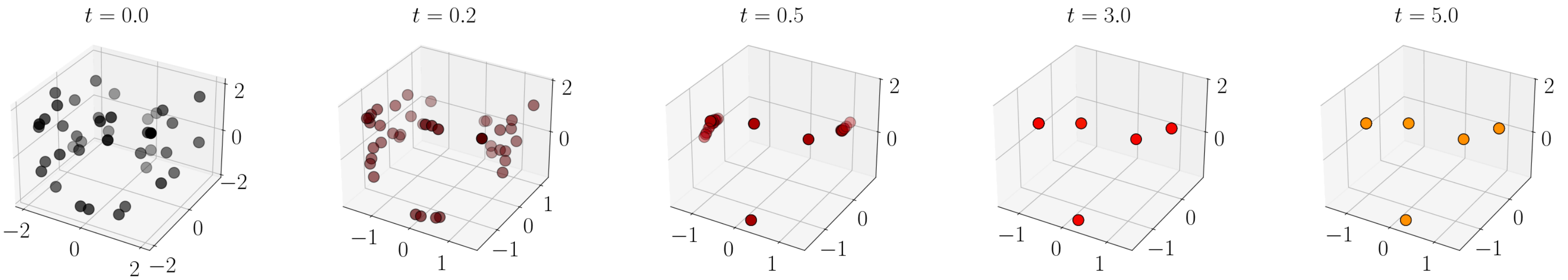
Evolution of interacting particles under normalized self-attention

$$z_i^{k+1} = z_i^k + \left[\sum_{j=1}^{n} K_{\varepsilon\, ij}^{\infty}\, W_V\, z_j^k - W_V\, z_i^k\right]$$

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity

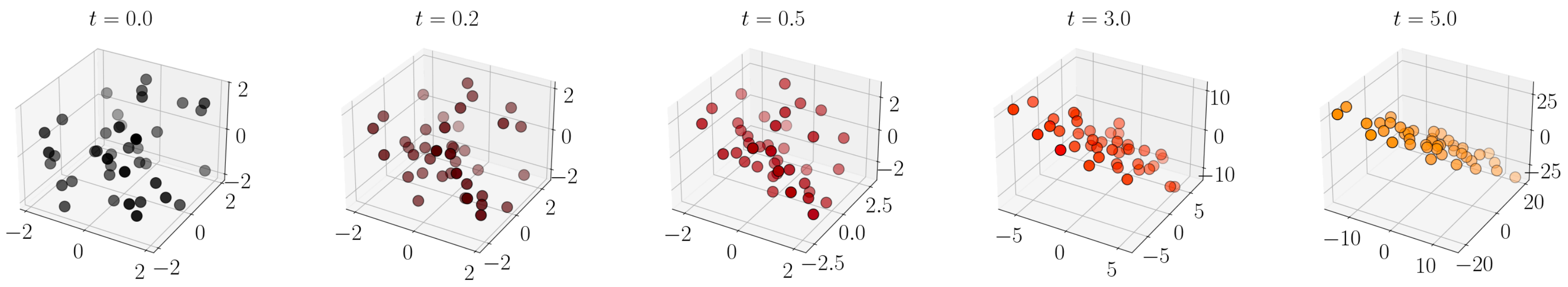
Evolution of interacting particles under self-attention

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity



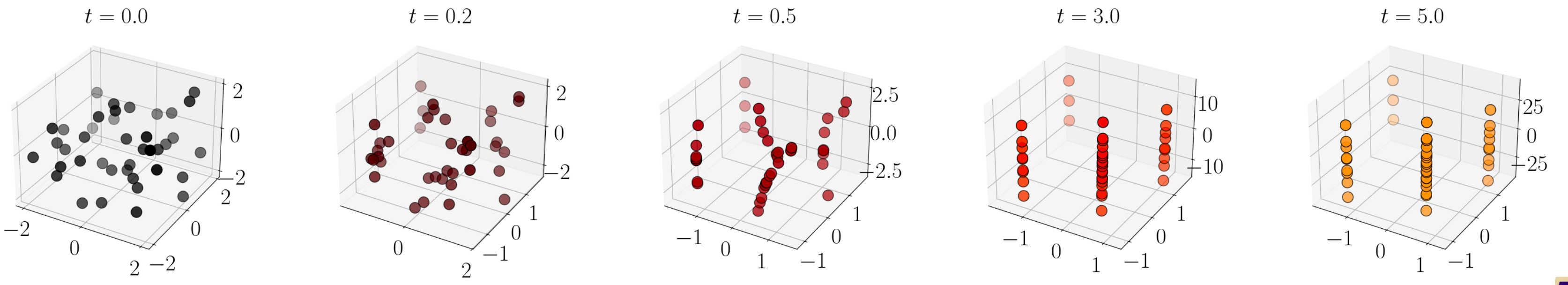
Evolution of interacting particles under self-attention

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity



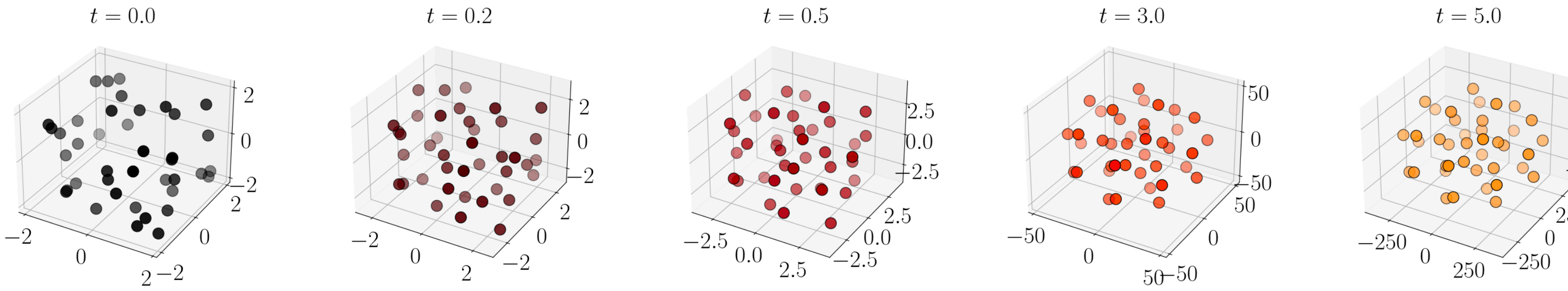
Evolution of interacting particles under self-attention

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity



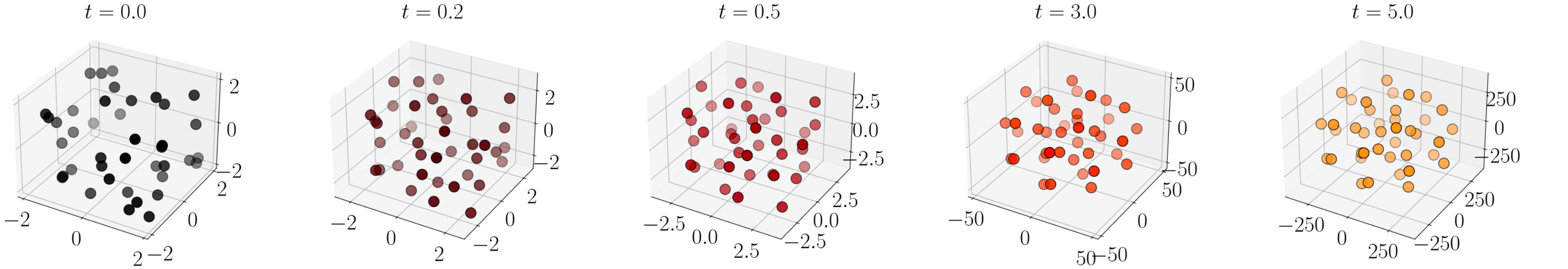
Evolution of interacting particles under self-attention

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity

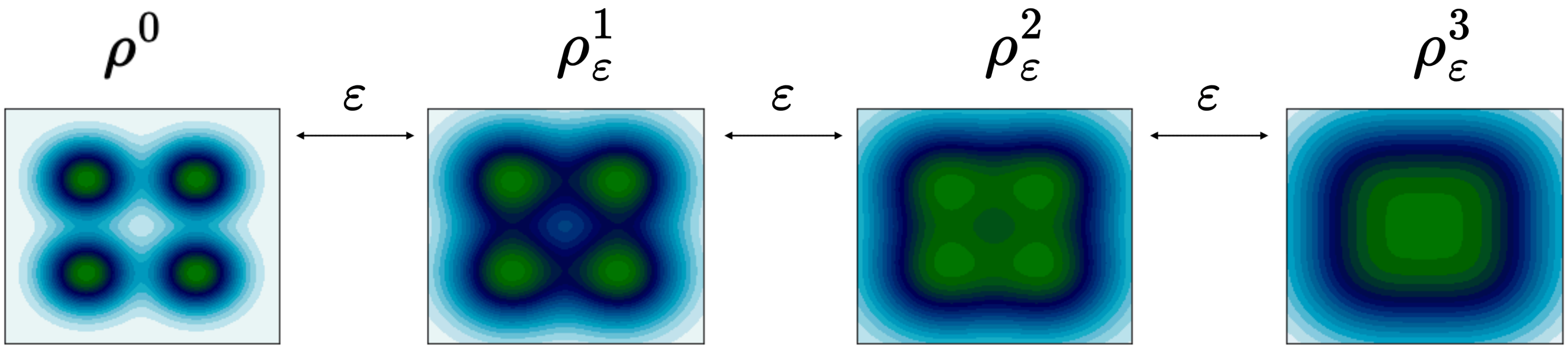


Evolution of interacting particles under self-attention

Key and Query	Value	Limit geometry
$W_Q^\top W_K \succ 0$	$W_V = I_d$	Vertices of a polytope
$W_Q^\top W_K \succ 0$	$\lambda(W_V) > 0$	Hyperplane
$W_Q^\top W_K \succ 0$	W_V is paranormal	Polytope X subspace
$W_Q^\top W_K \succ 0$	$W_V = -I_d$	Tokens diverge to infinity

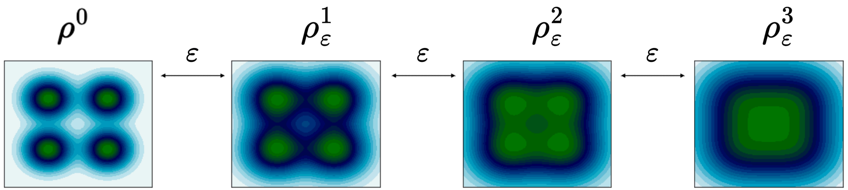


[AHMP24] prove that as $\varepsilon \to 0$, the discrete process converges to heat equation!



ρ^0
ε
ρ_ε^1
ε
ρ_ε^2
ε
ρ_ε^3

Motivating Question



- Concretely, $\rho_\varepsilon^{k+1} = \left(T_{\rho_\varepsilon^k,\varepsilon}\right)_\# \rho_\varepsilon^k$,
- $T_{\rho_\varepsilon^k,\varepsilon}(x) = x + \left(\int k_\varepsilon^\infty(x,y)\, W_V y \, d\rho_\varepsilon^k(y) - W_V x \right)$
- When $W_Q^\top W_V = I_d$ and $W_V = -I_d$, then $T_{\rho_\varepsilon^k,\varepsilon}(x) = 2I - \int k_\varepsilon^\infty(x,y)\, y \, d\rho_\varepsilon^k(y)$

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

Motivating Question

Under assumption $W_K^\top W_Q = W_Q^\top W_K = -W_V = I$, the **Sinkformer self-attention update** is

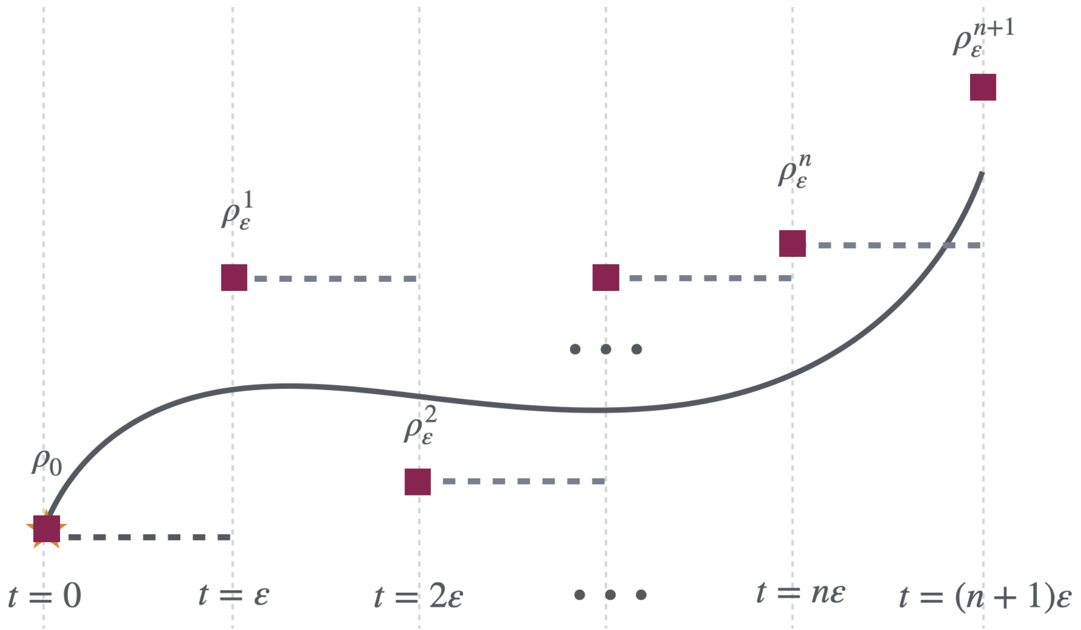
$$\rho_\varepsilon^{k+1} = \left(2I - \int k_\varepsilon^\infty(x,\cdot)d\rho_\varepsilon^k(x)\right)_\# \rho_\varepsilon^k$$

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

↓

Barycentric projection

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 $\varepsilon \to 0$?

Claim

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

Self-attention flow

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

Heat flow

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

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

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

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

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

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 $(Id, T)_\# \mu \in \Pi(\mu, \nu)$.

Entropy regularized optimal transport

Entropy-regularized optimal transport problem

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

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

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

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

$$(X, Y) \sim \pi_\varepsilon$$

Self-attention update via same marginal Schrödinger bridges

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

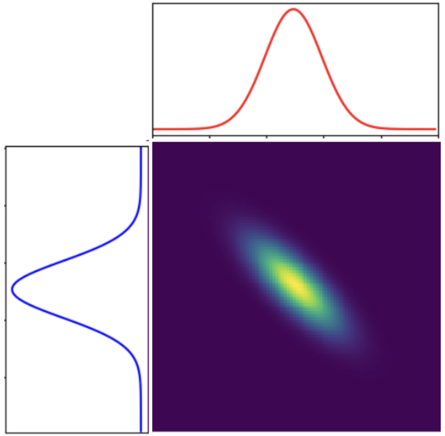
Let $\pi_{\rho,\varepsilon}$ be the Schrödinger bridge from ρ to itself and

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

Recall, the doubly-stochastic self-attention update

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

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)

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

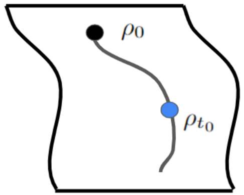
Particle Picture

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

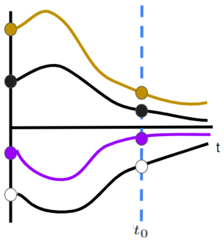
$$\dot{x}_t = v_t = -\frac{1}{2}\nabla \log \rho(t)$$

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

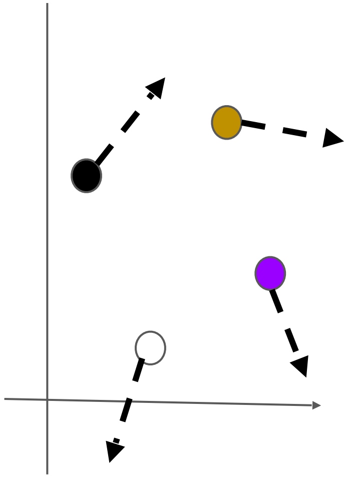
Flow of Measures



Particle Trajectories



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



Result 2: One Step Approximation - Intuition

From Result 1 ($\pi_{\rho,\varepsilon} \approx \ell_{\rho,\varepsilon}$), intuitively,

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

$$\implies \frac{\mathcal{B}_{\rho,\varepsilon}(x) - x}{\varepsilon} \approx \frac{\nabla \log \rho(x)}{2}$$

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

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

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

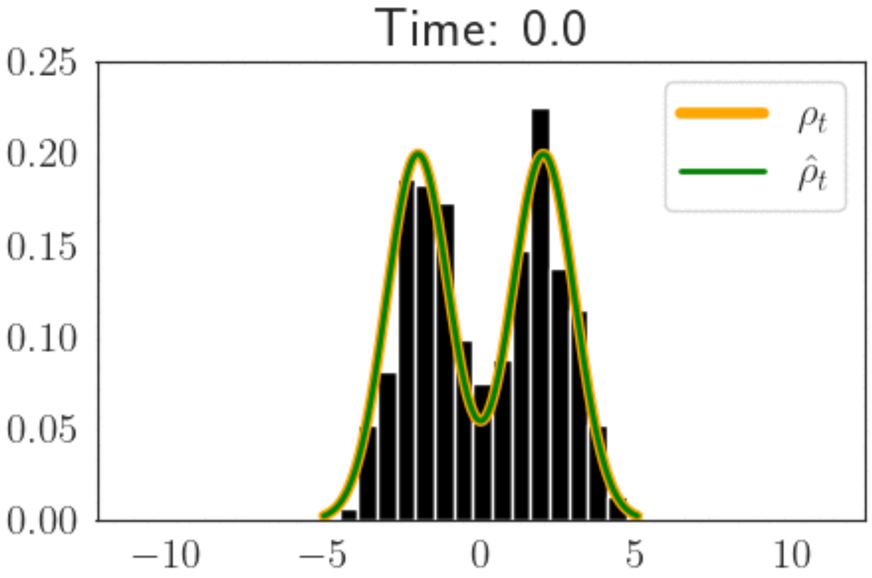
As a corollary,

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

Simulations

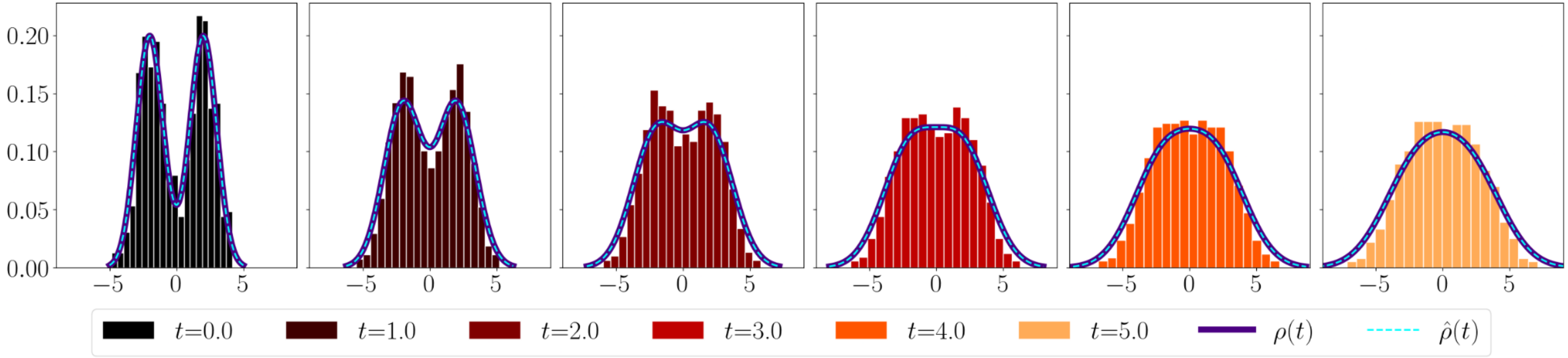
Mixture of Gaussians

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



Mixture of Gaussians

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



Thank you for your attention!

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.