Pretrained Transformers are Heat Flows

Medha Agarwal*, Garrett Mulcahy†, Soumik Pal†, Zaid Harchaoui*

Institute for Foundations of Data Science
IFDS
Washington Wisconsin Santa Cruz Chicago

Self-attention mechanism

For initial d-dimensional particles/tokens $(x_1, \ldots, x_n) \sim \rho_0$, the Transformer update is

$$x_i \leftarrow \sum_{j=1}^{n} K^{\infty}_{ij} W_V x_j, \text{ where } K^{\infty} = \text{Sinkhorn}(C) \text{ and } C_{ij} = \left(W_Q x_i\right)^{\top}\left(W_K x_j\right)$$

Value Matrix

Query Matrix

Key Matrix

Under $(W_Q)^{\top} W_K = -W_V = I_d$, the infinite particles counterpart with bandwidth/temperature ε and $k_\varepsilon = \text{Sinkhorn}(c/\varepsilon)$ is

$$\rho \leftarrow \left(T_{\rho,\varepsilon}\right)_{\#} \rho \text{ where } T_{\rho,\varepsilon}(x) = 2x - \int k^{\infty}_{\varepsilon}(x,y)\; y \; d\rho(y)$$

Iteratively, with $\rho^{\varepsilon}_0 = \rho_0$,

Sequence $\left(\rho^{\varepsilon}_k, k \in [T/\epsilon]\right)$

$$\rho^{\varepsilon}_k = \left(T_{\rho^{\varepsilon}_{k-1},\varepsilon}\right)_{\#} \rho^{\varepsilon}_{k-1}$$

Continuous curve $\left(\rho^{\varepsilon}_t, t \in [0,T]\right)$

$$\rho^{\varepsilon}_t = \rho^{\varepsilon}_{\lfloor t/\varepsilon \rfloor}, t \in [0,T]$$

As $\varepsilon \to 0+$, what is the limit of the curve $(\rho^{\varepsilon}_t, t \in [0,T])$?

Answer: Heat equation!

In fact, if $(\rho_t, t \in [0,T])$ is heat equation starting from ρ_0,

$$\lim_{\varepsilon \to 0} \sup_{k \in [T/\varepsilon]} \mathbb{W}^2_2(\rho^{\varepsilon}_k, \rho_{k\varepsilon}) = 0$$

Background

Entropy-regularized optimal transport (EOT)

Consider the EOT problem between measures μ and ν

$$\min_{\pi \in \Pi(\mu,\nu)} \left\{ \frac{1}{2}\|x-y\|^2 d\pi + \varepsilon \mathrm{KL}(\pi \,||\, \mu \otimes \nu) \right\}$$

The argmin $\pi_{\mu,\nu,\varepsilon}$ of EOT is the Schrödinger bridge. Consider same marginal setting $\mu = \nu = \rho$ with Schrödinger bridge $\pi_{\rho,\varepsilon}$. Then if $(X,Y) \sim \pi_{\rho,\varepsilon}$ and $\pi_{\rho,\varepsilon}$ admits the disintegration $\pi_{\rho,\varepsilon} = \int \pi_{\rho,\varepsilon,x} d\rho(x)$, define barycentric projection

$$\int y\; k^{\infty}(x,y)\; d\rho(y) = \int y d\pi_{\rho,\varepsilon,x}(y) =: \mathscr{B}_{\rho,\varepsilon}(x)$$

Therefore, for infinite particles, the self-attention update for measures is

$$\rho^{\varepsilon}_k = \left(T_{\rho^{\varepsilon}_{k-1},\varepsilon}\right)_{\#} \rho^{\varepsilon}_{k-1} = \left(2I_d - \mathscr{B}_{\rho,\varepsilon}\right)_{\#} \rho^{\varepsilon}_{k-1}$$

Heat Equation

General Wasserstein gradient flows characterized by the continuity equation

$$\partial_t \rho_t + \nabla_x \cdot \left(v_t \rho_t\right) = 0$$

where $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is the velocity field.

Heat equation $(\rho_t, t \geq 0)$ is $\partial_t \rho_t = \frac{1}{2}\Delta \rho_t$, therefore, $v_t = -\frac{1}{2}\nabla \log \rho_t$.

Main contribution: Under suitable conditions,

$$2x - \mathscr{B}_{\rho,\varepsilon} \approx x - \frac{\varepsilon}{2}\nabla_x \log \rho(x)$$

And in general, $\frac{1}{\varepsilon}(\mathscr{B}_{\rho,\varepsilon} - I)$ approximates the score function $\frac{1}{2}\nabla_x \log \rho$.

Convergence

Claim: Iterations of self-attention mechanism $(\rho^{\varepsilon}_k, k \in \lfloor T/\varepsilon \rfloor)$ is a discretely approximate the heat flow starting from ρ_0.

Define $S^1_\varepsilon(\rho) = \left(id - 0.5\varepsilon \nabla_x \log \rho\right)_{\#} \rho$ to be the explicit Euler step (EE) and $SB^1_\varepsilon(\rho) := (2I_d - \mathscr{B}_{\sigma_\varepsilon,\varepsilon})_{\#}\rho$ be the Schrödinger bridge (SB) step.

One-step Convergence

Proposition 1 (Single Step Convergence)

If the surrogate measure σ_ε satisfies a set of regularity conditions, then there exists a constant $K > 0$ such that $\mathbb{W}_2\left(S^1_\varepsilon(\rho), SB^1_\varepsilon(\rho)\right) < K\varepsilon C(\varepsilon)$. Under

For Gaussian marginals, $C(\varepsilon) = \mathcal{O}(\varepsilon^2)$

$$\lim_{\varepsilon \to 0} \varepsilon^{-1}\mathbb{W}_2(S^1_\varepsilon(\rho), SB^1_\varepsilon(\rho)) = 0$$

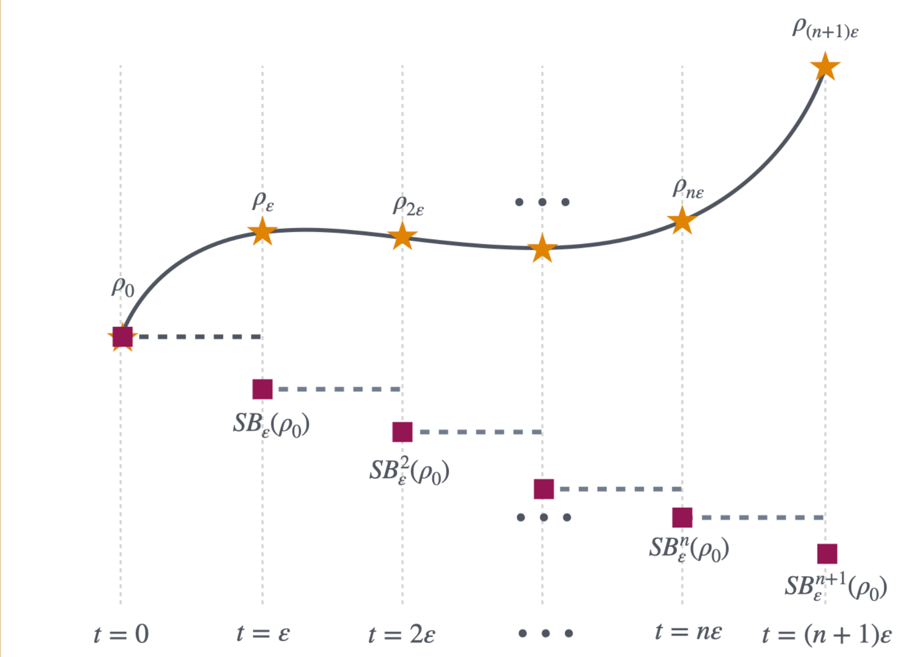
Uniform Convergence

Even though, we show that SB one step is $o(\varepsilon)$ approximation of the EE step, it is important to ensure that the errors do not accumulate.

Theorem (Uniform Convergence)

Under some regularity assumptions on the iterates $(S^k_\varepsilon, k \geq \lfloor T/\varepsilon \rfloor)$ and $(SB^k_\varepsilon, k \in \lfloor T/\varepsilon \rfloor)$, the sequence $(SB^k_\varepsilon, k \in \lfloor T/\varepsilon \rfloor)$ is a first order approximation of heat flow $(\rho_t, t \in [0,T])$.

$$\mathbb{W}_2(\rho_{k\varepsilon}, SB^k_\varepsilon(\rho)) \leq \mathbb{W}_2(\rho_{k\varepsilon}, SB^1_\varepsilon(\rho_{(k-1)\varepsilon})) + \mathbb{W}_2(SB^1_\varepsilon(\rho_{(k-1)\varepsilon}), SB^k_\varepsilon(\rho))$$

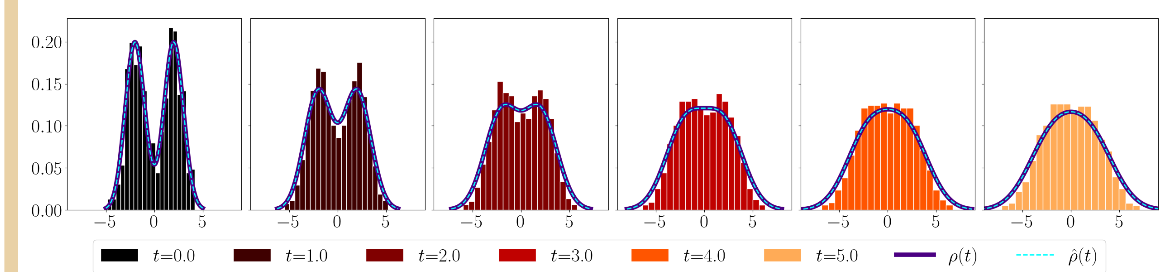


Preprint

Code

Experiments

Heat flow with $\varepsilon = 0.01$ and $\rho_0 = 0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$



Heat flow with $\varepsilon = 10^2$ and $\rho_0 \sim$ FashionMNIST