Iterated Schrödinger Bridge Approximation to Wasserstein Gradient Flows

Medha Agarwal, Zaid Harchaoui, Garrett Mulcahy, Soumik Pal

UNIVERSITY of WASHINGTON

IFML

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

Background

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 is

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

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

Sequence $\left(\rho^{\varepsilon}_k, k \in [T/\varepsilon]\right)$, $\rho^{\varepsilon}_k = \left(T_{\rho^{\varepsilon}_{k-1},\varepsilon}\right)_{\#} \rho^{\varepsilon}_{k-1}$ → Piece-wise continuous curve $\left(\rho^{\varepsilon}_t, t \in [0,T]\right)$, $\rho^{\varepsilon}_t = \rho^{\varepsilon}_{[t/\varepsilon]}, t \in [0,T]$

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

Setting

Consider the entropy regularized optimal transport (EOT) problem between measures μ and ν

$$\min_{\pi \in \Pi(\mu,\nu)} \left\{ \int \|x-y\|^2 d\pi + \varepsilon \text{KL}(\pi \,||\, \mu \otimes \nu) \right\} \longrightarrow \text{argmin } \pi_{\mu,\nu,\varepsilon} \text{ is the Schrödinger bridge}$$

Consider same marginal setting, i.e. $\mu = \nu = \rho$ with Schrödinger bridge $\pi_{\rho,\varepsilon}$, then

$$\mathscr{B}_{\rho,\varepsilon}(x) = \mathbb{E}_{\pi_{\rho,\varepsilon}}[Y \,|\, X = x] = \int y\, k^{\infty}(x,y)\, d\rho(y).$$

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

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

Main contribution: Under suitable conditions, $2x - \mathscr{B}_{\rho,\varepsilon} \approx x - \frac{\varepsilon}{2} \nabla \log \rho(x)$ $\Longrightarrow$ Claim: $(\rho^{\varepsilon}_k, k \in [T/\varepsilon])$ is a discrete approximate to the heat flow.

Wasserstein gradient flow is characterized by the continuity equation $\partial_t \rho_t + \nabla \cdot (v_t\, \rho_t) = 0$ where $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is the velocity field. For heat flow, $v_t = -0.5\, \nabla \log \rho_t$.

What about any gradient flow? Can we such approximations for them like self-attention for heat flow?

Key Contribution: (1) Under suitable conditions, self-attention iterations converge to the heat flow as $\varepsilon \to 0$. (2) We write the iterated scheme for a general functional $\mathscr{F}$. (3) We prove the uniform convergence of the scheme to the gradient flow of KL divergence with respect to a log-concave density.

Iterative Scheme

Consider absolutely continuous curve $(\rho_t, t \in [0,T])$ satisfying $\partial_t \rho_t + \nabla \cdot (v_t\, \rho_t) = 0$ where $v_t = v(\rho_t)$ is the velocity field. An example is gradient flow minimizing a functional $\mathscr{F} : \mathscr{P}(\mathbb{R}^d) \to \mathbb{R}$.

One-step Approximation

One discrete time approximation is analogous to explicit Euler scheme

$$S^1_{\varepsilon}(\rho) = \left(id + \varepsilon v\right)_{\#} \rho \text{ with } v = v(\rho).$$ Explicit Euler (EE) Step

Let $v = \nabla \psi$ for a smooth function ψ and there exists $\theta_{\varepsilon} \in \mathbb{R} \backslash \{0\}$ s.t. $\Lambda(\varepsilon) = \int_{\mathbb{R}^d} \exp(2\theta_{\varepsilon}\psi) < +\infty$.

Define the **surrogate measure** $\sigma_{\varepsilon} := \Lambda(\varepsilon)^{-1} \exp(2\,\theta_{\varepsilon}\,\psi)$ and the one-step update

$$SB^1_{\varepsilon}(\rho) := ((1 - \theta^{-1}_{\varepsilon})\, id + \theta^{-1}_{\varepsilon} \mathscr{B}_{\sigma_{\varepsilon},\varepsilon})_{\#} \rho$$ Schrödinger Bridge (SB) Step

Iterative Schemes

Explicit Euler Scheme: $\left(\rho, S^1_{\varepsilon}(\rho), \ldots, S^{N_{\varepsilon}}_{\varepsilon}(\rho)\right),\ N_{\varepsilon} = [T/\varepsilon].$

Schrödinger Bridge Scheme: $\left(\rho, SB^1_{\varepsilon}(\rho), \ldots, SB^{N_{\varepsilon}}_{\varepsilon}(\rho)\right),\ N_{\varepsilon} = [T/\varepsilon].$

Remark: For heat equation (gradient flow of entropy functional), $\sigma_{\varepsilon} = \rho$ and $\theta_{\varepsilon} = -1$. For Fokker Planck equation (gradient flow of KL divergence with respect to ν), $\sigma_{\varepsilon} = (\rho/\nu)^{-\theta_{\varepsilon}}$ where the sign of θ_{ε} depends on the integrability of σ_{ε}.

Convergence

Theorem 1 (Tight Approximation of Same Marginal Schrödinger Bridge)

Let $\rho = e^{-g} \in \mathscr{P}(\mathbb{R}^d)$ with enough *regularity* such that there is strong solution to the Langevin SDE $dX_t = -\frac{1}{2}\nabla g(X_t) + dB_t$ with $X_0 \sim \rho$. Let $\ell_{\rho,\varepsilon} = \text{Law}(X_1, X_\varepsilon)$, then $H(\ell_{\rho,\varepsilon} \,|\, \pi_{\rho,\varepsilon}) + H(\pi_{\rho,\varepsilon} \,|\, \ell_{\rho,\varepsilon}) \le \varepsilon^2\, C(\varepsilon)$.

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

One-step Convergence

Theorem 2 (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 assumptions on surrogate measure σ_{ε}, $C(\varepsilon) = o(1)$.

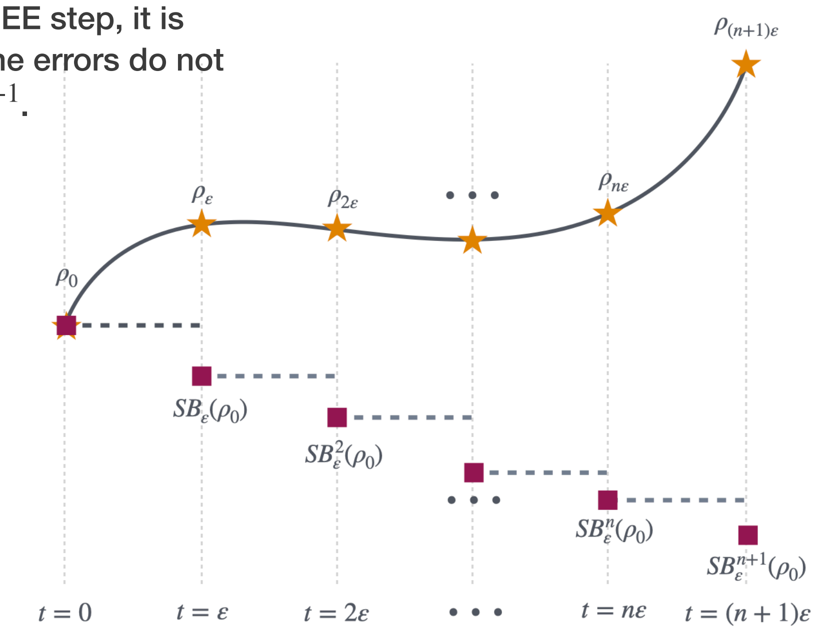
$$\lim_{\varepsilon \to 0} \varepsilon^{-1} \mathbb{W}_2(S^1_{\varepsilon}(\rho), SB^1_{\varepsilon}(\rho)) = 0$$

Remark: (1) The one step convergence relies on the close approximation of the same-marginal Schrödinger bridge by the Langevin diffusion.

Uniform Convergence

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

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 culminate by more than ε^{-1}.

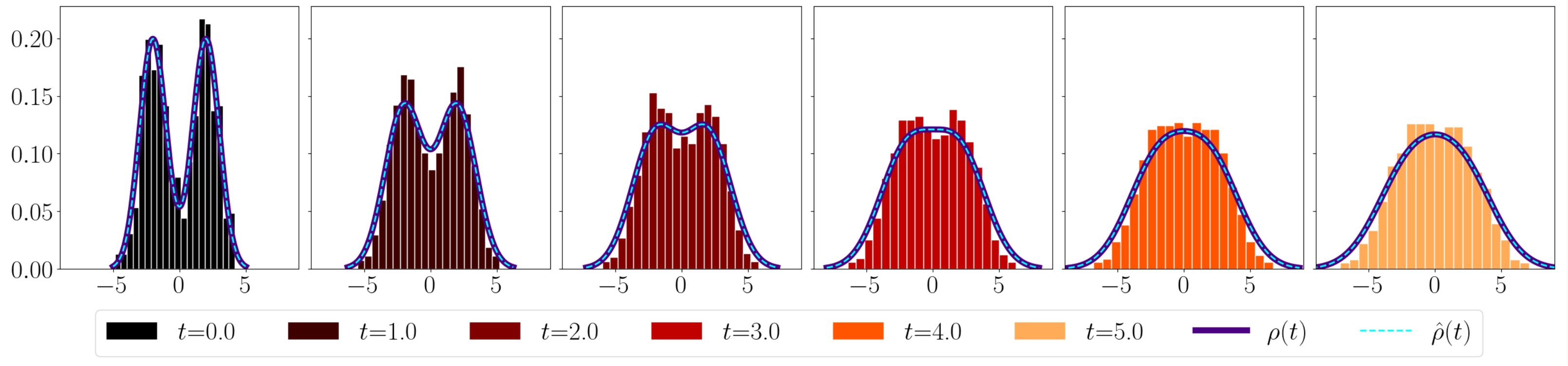


Preprint

Code

Experiments

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



Ornstein-Uhlenbeck process with $\varepsilon = 0.01$ and $\rho_0 = \mathcal{N}(0,4)$