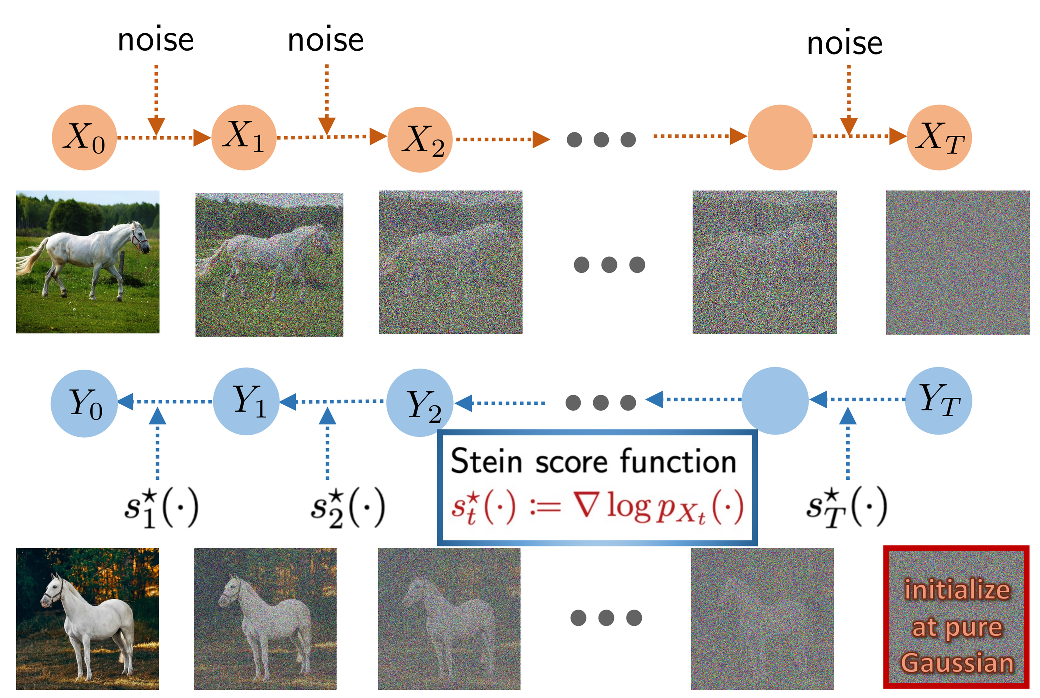
Improved Convergence Rate for Diffusion Probabilistic Models

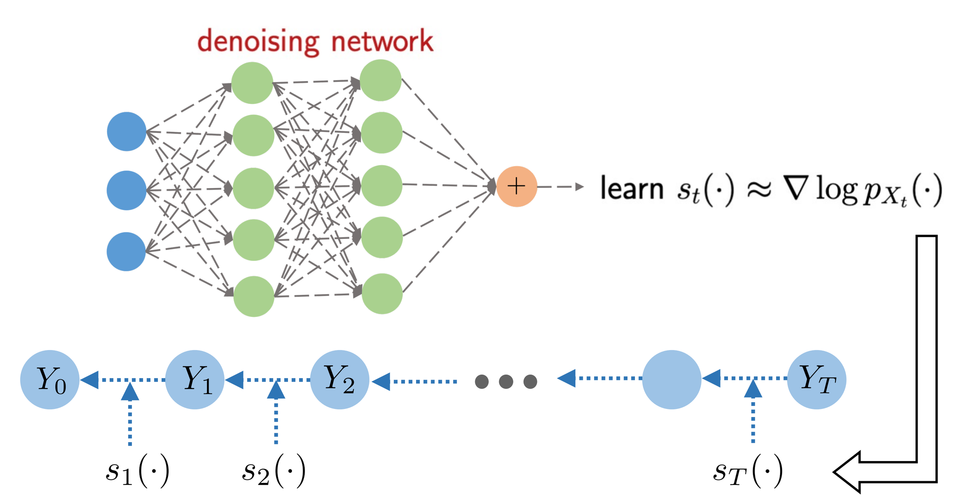
Gen Li — CUHK

Yuchen Jiao — CUHK

Diffusion models



- **forward process:** (progressively) diffuse data into noise
- **reverse process:** convert pure noise into data-like distributions



- **score learning/matching:** learn estimates $s_t(\cdot)$ for $\nabla \log p_{X_t}(\cdot)$
- **data generation:** sampling w/ the aid of score estimates $\{s_t(\cdot)\}$

Learning rate (randomized midpoint)

$$X_t \overset{\mathsf{d}}{=} \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} W_t, \quad X_\tau \overset{\mathsf{d}}{=} \sqrt{1-\tau} X_0 + \sqrt{\tau} W_t, \qquad W_t \sim \mathcal{N}(0, I_d)$$

- Learning rate:

$$\widehat{\alpha}_{T+1} = \frac{1}{T^{c_0}}, \qquad \widehat{\alpha}_{t-1} = \widehat{\alpha}_t + \frac{c_1 \widehat{\alpha}_t (1-\widehat{\alpha}_t)\log T}{T}$$

$$\overline{\alpha}_t \sim \mathsf{Unif}(\widehat{\alpha}_t, \widehat{\alpha}_{t-1}), \quad \text{for } t = -\frac{N}{2}+1, \dots, T+1$$

- Continuous index:

$$\widehat{\tau}_{k,n} := 1 - \widehat{\alpha}_{T-\frac{kN}{2}-n}, \qquad \tau_{k,n} := 1 - \overline{\alpha}_{T-\frac{kN}{2}-n+1} \qquad \text{for} \quad k = 0, \cdots, K, \quad n = 0, \dots, N.$$

Lemma: The above learning rate satisfies:

$$1 - \tau_{0,0} \le \widehat{\alpha}_T \le \frac{2}{T^{c_0}}, \qquad \tau_{K,0} \le 1 - \widehat{\alpha}_1 \le \frac{1}{T^{c_0}}, \qquad \frac{\widehat{\tau}_{k,n-1} - \widehat{\tau}_{k,n}}{\widehat{\tau}_{k,n-1}(1-\widehat{\tau}_{k,n-1})} = \frac{c_1 \log T}{T}$$

Sampler

- Initialization: $Y_0 \sim \mathcal{N}(0, I_d)$.
- For each k ranging from 0 to $K-1$, iteratively update for $n = 1, \cdots, N$:

$$Y_{k,0} = Y_k,$$

$$\frac{Y_{k,n}}{\sqrt{1-\tau_{k,n}}} = \frac{Y_{k,0}}{\sqrt{1-\tau_{k,0}}} + \frac{s_{T-\frac{kN}{2}+1}(Y_{k,0})}{2(1-\tau_{k,0})^{3/2}}(\tau_{k,0} - \widehat{\tau}_{k,0}) + \sum_{i=1}^{n-1} \frac{s_{T-\frac{kN}{2}-i+1}(Y_{k,i})}{2(1-\tau_{k,i})^{3/2}}(\widehat{\tau}_{k,i-1} - \widehat{\tau}_{k,i}) + \frac{s_{T-\frac{kN}{2}-n+2}(Y_{k,n-1})}{2(1-\tau_{k,n-1})^{3/2}}(\widehat{\tau}_{k,n-1} - \tau_{k,n}),$$

- Noise injection:

$$Y_{k+1} = \sqrt{\frac{1-\tau_{k+1,0}}{1-\tau_{k,N}}} Y_{k,N} + \sqrt{\frac{\tau_{k+1,0} - \tau_{k,N}}{1-\tau_{k,N}}} Z_k, \quad Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$$

Instance-dependent convergence rate

- **Minimal data assumptions:**

$$\mathbb{E}_{X_0 \sim p_{\mathsf{data}}}[\|X_0\|_2^2] \le T^{c_R}$$

for arbitrarily large constant $c_R > 0$

- **ℓ_2 score estimation error:** $s_t^\star(X) := \nabla \log p_{X_t}(X)$,

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} \mathbb{E}_{Y_k \sim q_k}\big[\|s_{T-\frac{kN}{2}-n+1}(Y_{k,n}) - s^\star_{T-\frac{kN}{2}-n+1}(Y_{k,n})\|_2^2\big] \le \varepsilon_{\mathsf{score}}^2$$

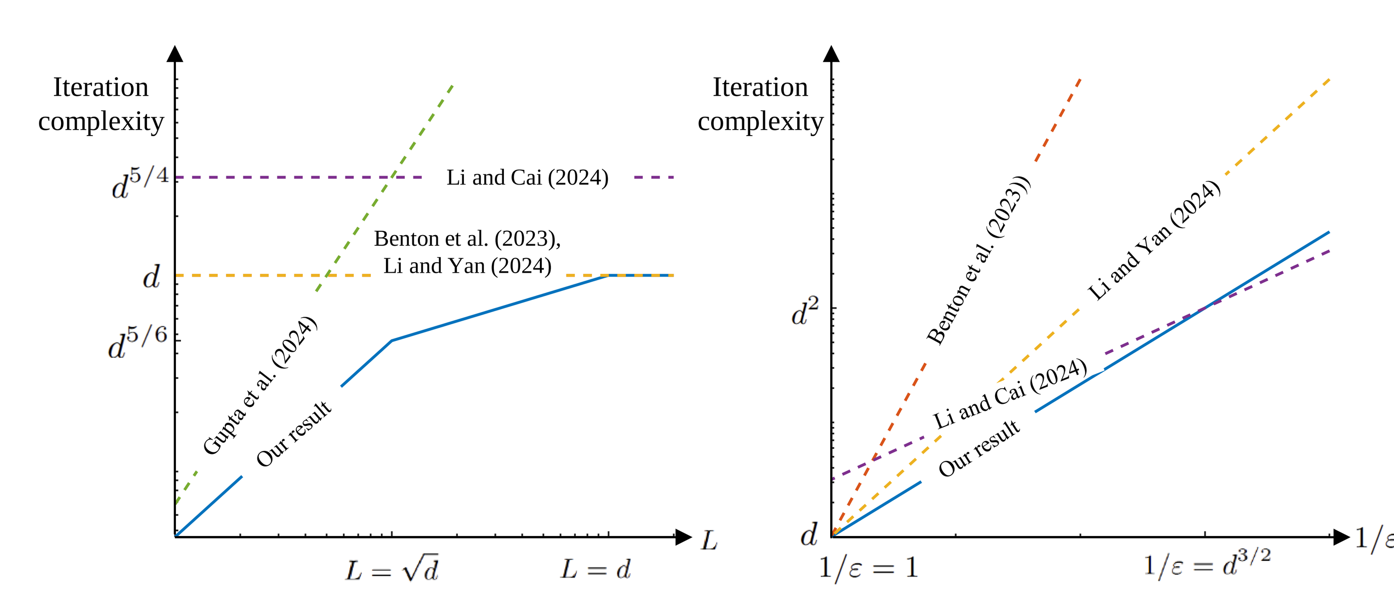
- **Lipschitz constant:** Define L as the smallest quantity such that

$$\|s_t^\star(x') - s_t^\star(x)\|_2 \le L\|x' - x\|_2, \quad \forall t$$

Theorem: The presented sampler based on randomized midpoint obeys

$$\mathsf{TV}(p_{X_1}, p_{Y_K}) \lesssim \frac{\min\{d^{3/2}, dL^{1/2}, d^{1/2}L^{3/2}\}\log^4 T}{T^{3/2}} + \varepsilon_{\mathsf{score}} \log^{1/2} T$$

- iteration complexity $\widetilde{O}(\min\{d, d^{2/3}L^{1/3}, d^{1/3}L\}\varepsilon^{-2/3})$ to yield TV dist $\le \varepsilon$
- accommodate ε score estimation error in averaged ℓ_2-norm
- comparison with prior works. left: $\varepsilon = O(1)$, right: $L = \infty$



Key ingredients of our analysis

- Probability flow ODE

Lemma: Define $\Phi_{\tau_1 \to \tau_2}(x) := x_{\tau_2}|_{x_{\tau_1}=x}$ through the following ODE, where $s_\tau^\star(x) = \nabla \log p_{X_\tau}(x)$

$$\mathrm{d}\frac{x_\tau}{\sqrt{1-\tau}} = -\frac{s_\tau^\star(x_\tau)}{2(1-\tau)^{3/2}}\mathrm{d}\tau.$$

We have

$$\widehat{X}_{k+1} = \sqrt{\frac{1-\tau_{k+1,0}}{1-\tau_{k,N}}}\Phi_{\tau_{k,0}\to\tau_{k,N}}(\widehat{X}_k) + \sqrt{\frac{\tau_{k+1,0}-\tau_{k,N}}{1-\tau_{k,N}}} Z_k \overset{\mathrm{d}}{=} X_{\tau_{k+1,0}}$$

- Error decomposition

$$\mathsf{TV}(p_{X_1}, p_{Y_K}) \le \sqrt{\frac{1}{2}\mathsf{KL}(p_{\widehat{X}_0}\|p_{Y_0}) + \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_{x_k \sim p_{\widehat{X}_k}}\Big[\mathsf{KL}\big(p_{\widehat{X}_{k+1}|\widehat{X}_k}(\cdot|x_k) \,\|\, p_{Y_{k+1}|Y_k}(\cdot|x_k)\big)\Big]} + 2\sum_{k=0}^{K-1} P(\widetilde{X}_k \in \mathcal{E}_k^c)$$

($\mathcal{E}_k$: some "typical" set, $\widetilde{X}_k$: auxiliary sequence based on $\mathcal{E}_k$)

- Analysis on one-step error $\xi_{k,n}$

$$\mathbb{E}_{x_{\tau_{k,0}} \sim p_{\widehat{X}_k}}\big[\|\xi_{k,n}(x_{\tau_{k,0}})\|_2^2\big] \lesssim \frac{d\log^4 T}{T^3}\min\Big\{\frac{Nd\widehat{\tau}_{k,-1}\log T}{T(1-\widehat{\tau}_{k,-1})} + \int_{\tau_{k,n}}^{\tau_{k,0}} \frac{\mathbb{E}[\mathsf{Tr}(\Sigma_\tau^2(x_\tau))]}{(1-\tau)^2}\mathrm{d}\tau, \frac{NL^2\widehat{\tau}_{k,-1}\log T}{T(1-\widehat{\tau}_{k,-1})}\Big\} + \frac{N\log^2 T}{T^2}\sum_{i=0}^{n-1}\widehat{\tau}_{k,i}(1-\widehat{\tau}_{k,i})^{-1}\varepsilon_{k,i}^2,$$

with $\Sigma_\tau(x) = \mathsf{Cov}[Z | \sqrt{1-\tau}X_0 + \sqrt{\tau}Z = x]$, $\quad \frac{1}{T}\sum_{k=0}^{K-1}\sum_{i=0}^{N-1}\varepsilon_{k,i}^2 = \varepsilon_{\mathsf{score}}^2$

- Error propagation control

$$\sum_{k=0}^{K-1}\mathbb{E}_{x_{\tau_{k,0}} \sim p_{\widehat{X}_k}}\Big[\mathsf{KL}\Big(p_{\widehat{X}_{k+1}|\widehat{X}_k}\big(\cdot|x_{\tau_{k,0}}\big) \,\|\, p_{Y_{k+1}|Y_k}\big(\cdot|x_{\tau_{k,0}}\big)\Big)\Big] \lesssim \frac{Kd\log^5 T}{T^3}\min\big\{d, L^2\big\} + \varepsilon_{\mathsf{score}}^2 \log T$$

Future directions

- **instance-dependent bounds for other settings:**
  - accelerated samplers
  - data distributions with low-dimensional structures
- **estimating the Lipschitz constant in real-world case**
  - broadening the applicability of our results
  - Gaussian mixture models, singular Gaussian distributions,...

"Improved Convergence Rate for Diffusion Probabilistic Models," G. Li, Y. Jiao, arxiv:2410.13738, ICLR.