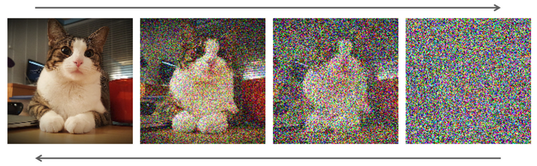
Improved Sample Complexity Bounds for Diffusion Model Training

Shivam Gupta Aditya Parulekar Eric Price Zhiyang Xun

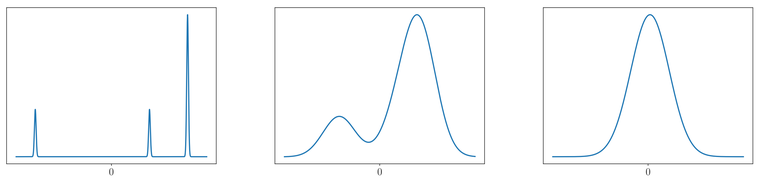
The University of Texas at Austin

Sampling Process of Diffusion Models

- Diffusion models learn a distribution q_0 by adding noise to training samples and learning to denoise.



- $x_0 \sim q_0$ evolves into $x_t \sim e^{-t}x_0 + \mathcal{N}(0, \sigma_t^2 I_d)$ at time t, where $\sigma_t^2 = 1 - e^{-2t}$. As t grows, distribution converges to $\mathcal{N}(0, I_d)$.



t=0 t = 0.5 t=1

- Need to learn the **score** function $s_t := \nabla \log q_t$.
- Given accurate enough s_t's, diffusion models can provably sample from q_0 with ε TV and γm_2 Wasserstein error, where m_2 is the second moment of q_0.

Question

How many training samples are required to learn score functions to enable accurate diffusion sampling?

- Traditionally, this is equivalent to: how many samples are required to learn each s_t's with ε^2 error in L^2 .

Background: Score Matching

- The *score matching* algorithm learns score function s_t using independent samples $x_1, \ldots, x_m$ drawn from q_0.
- Take Gaussian samples $z_1, \ldots, z_m \sim \mathcal{N}(0, \sigma_t^2 I_d)$. Then, the minimizer of the *score matching objective* is $\widehat{s}_t$:

$$\widehat{s}_t := \arg\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \left\| f(e^{-t}x_i + z_i) - \frac{-z_i}{\sigma_t^2} \right\|^2,$$

  where $\mathcal{F}$ is the class of functions represented by the neural network.
- As $m \to \infty$, s_t is provably the minimizer!
- We analyze its concentration: How large do we need m to be so that no *inadequate* score function becomes the minimizer?

Sample Complexity of Training

Let the candidate function class $\mathcal{F}$ be functions represented by a P-parameter, D-depth ReLU neural network.

Our Results

To train a diffusion model that achieves ε TV error and γm_2 Wasserstein error:

- $\mathrm{poly}(d, 1/\varepsilon, \log\frac{1}{\gamma}, D, P)$ training samples suffice, improving over previous $\mathrm{poly}(d, 1/\varepsilon, 1/\gamma, \exp(D), P)$.
- This matches the $\mathrm{poly}(d, 1/\varepsilon, \log\frac{1}{\gamma})$ number of iterations in the sampling process.
- It is impossible to get L^2 accurate scores using this number of samples. A new quantile measure is needed.

Work	Number of Samples	Notes
[OAS23]	$\widetilde{O}(\frac{1}{\varepsilon^{O(d)}})$	Density supported on $[-1,1]^d$, belongs to a Besov space
[CHZW23]	$\widetilde{O}(\frac{1}{(\varepsilon\gamma)^{O(d)}})$	Assuming density supported on d-dimensional subspace
[BMR20]	$\widetilde{O}\left(\frac{d^{5/2}R^3}{\gamma^3\varepsilon^2 m_2^3} P^D \sqrt{D}\right)$	Assuming NN can represent scores, distribution is bounded by R
Ours	$\widetilde{O}(\frac{d^2}{\varepsilon^3} PD \log^3\frac{1}{\gamma})$	Assuming NN can represent scores

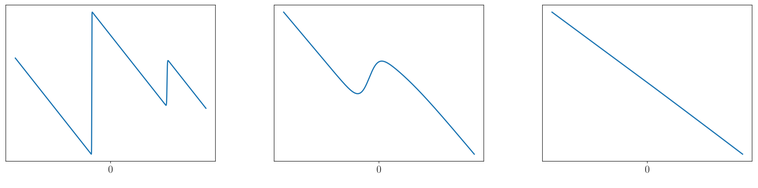
Proof Ideas

Exponential Improvement on D

- [BMR20] bounds the Rademacher complexity of the function class, which is exponential in depth.
- To circumvent this, we make use of a net argument.

Exponential Improvement on γ (Most Technical Part)

- The *score function* becomes simpler as noise level increases, so the score is hardest to learn for small t (when $\sigma_t = \gamma$):



t=0 t = 0.5 t=1

- We utilize the fact that, at time t, the ε accuracy requirement for s_t can be relaxed to $\varepsilon/\sigma_t \approx \varepsilon/\sqrt{\min(t, 1)}$, canceling out small t's hardness.

A $(1-\delta)$-Quantile Score Error Measure

- One key step in the proof is to relax the L^2 accuracy requirement for the scores.
- We prove that for score estimate $\widehat{s}_t$, we just need the $(1-\delta)$-quantile error of each $\widehat{s}_t$ to be smaller than ε/σ_t. That is,

$$\Pr_{x \sim q_t}\left[\|\widehat{s}_t(x) - s_t(x)\| > \varepsilon/\sigma_t\right] \le \delta,$$

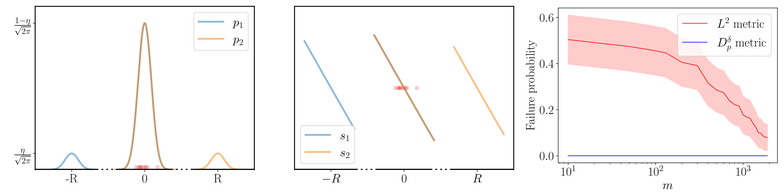
  for $\delta = \mathrm{poly}(\varepsilon)$.

For any function f with large $(1-\delta)$-quantile error:

- We guarantee that with m samples, f cannot be the minimizer of the score matching objective.
- The quantile error ensures the samples expose the high-error regions of f, making its value large in score matching objective.

Hardness of Learning L^2-Accurate Scores

There exist distributions needed $\mathrm{poly}(1/\gamma)$ samples to distinguish, but their scores have large L^2 distance.



- True distribution: $p_1 := (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(-R, 1)$, or $p_2 := (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(R, 1)$.
- The L^2 distance between the scores is about ηR^2.
- Given $o(1/\eta)$ samples from either p_1 or p_2 we will only see samples from the main Gaussian with high probability, and cannot distinguish them.

References

[BMR20] Adam Block, Youssef Mroueh, and Alexander Rakhlin, *Generative modeling with denoising auto-encoders and langevin sampling*, arXiv:2002.00107 [cs, math, stat] (2020), arXiv: 2002.00107.

[CHZW23] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang, *Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data*, Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[OAS23] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki, *Diffusion models are minimax optimal distribution estimators*, Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.