Taking a Big Step: Large Learning Rates in Denoising Score Matching Prevent Memorization

Yu-Han Wu[1], Pierre Marion[2], Gérard Biau[1], Claire Boyer[3]

01/07/2025

[1]LPSM, Sorbonne Université

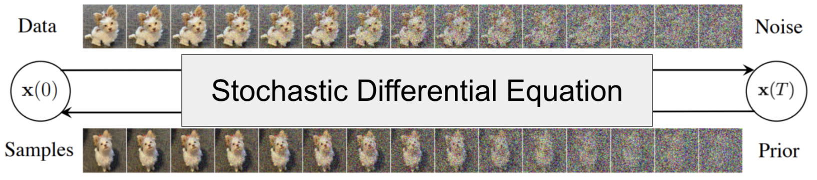
[2]Institute of Mathematics, EPFL

[3]LMO, Université Paris-Saclay

Introduction

Diffusion Model

Goal: Learn the distribution p_{true} of images.



- Forward SDE:

$$d\overrightarrow{X}_t = -\overrightarrow{X}_t dt + \sqrt{2} d\overrightarrow{B}_t, \quad \overrightarrow{X}_0 \sim p_{\text{true}},$$

- Backward SDE:

$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2\underbrace{\nabla \log p_{T-t}(\overleftarrow{X}_t)}_{\substack{\text{score function}\\ \text{unknown}}})dt + \sqrt{2} d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim p_T,$$

where $\overrightarrow{X}_t \sim p_t$ and $\overleftarrow{X}_{T-t} \overset{\mathcal{D}}{=} \overrightarrow{X}_t$.

Denoising Score Matching

- From now on we fix $t \in [0, T]$ and we denote σ (resp. μ) to be $\sigma(t)$ (resp. $\mu(t)$) for simplicity.
- Denoising score matching: given $x_1, \dots, x_n$ drawn from p_{true},

$$\mathcal{R}_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Y \sim \mathcal{N}(\mu x_i, \sigma^2)} \Big[\big(s(Y) + \frac{1}{\sigma^2}(Y - \mu x_i) \big)^2 \Big].$$

- Minimizer of $\mathcal{R}_n$:

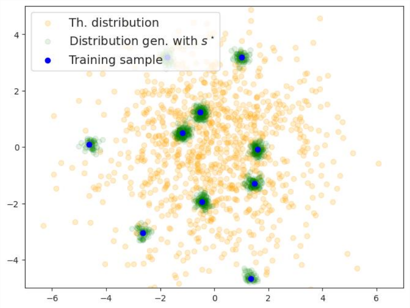
$$s^\star(y; \mu, \sigma) = \frac{\sum_{i=1}^{n} (\mu x_i - y) \exp(-(y - \mu x_i)^2 / 2\sigma^2)}{\sigma^2 \sum_{i=1}^{n} \exp(-(y - \mu x_i)^2 / 2\sigma^2)}, \quad y \in \mathbb{R}.$$

$s^\star$ is called the empirical optimal score function[1].

[1] Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. arXiv:2401.04856, 2024

Memorization

- Diffusion with $s^\star$:



- Reason: $s^\star$ is the score function of a Gaussian mixture distribution with a component centered at each data point.

Our Result

- Goal: explain the reason that in practice even without *explicit* regularization, we do not learn $s^\star$ and hence there is only a moderate amount of memorization.

Our Result

- Goal: explain the reason that in practice even without *explicit* regularization, we do not learn $s^\star$ and hence there is only a moderate amount of memorization.

Theorem (Informal)

Consider the denoising score matching objective $\mathcal{R}_n$ over the class of two-layer neural networks for one-dimensional data.

Our Result

- Goal: explain the reason that in practice even without *explicit* regularization, we do not learn $s^\star$ and hence there is only a moderate amount of memorization.

Theorem (Informal)

Consider the denoising score matching objective $\mathcal{R}_n$ over the class of two-layer neural networks for one-dimensional data.

Then, for a sufficiently small level of noise σ and a learning rate $\eta \gtrsim \sigma^2$, the stochastic gradient descent on $\mathcal{R}_n$ cannot stably converge to $s^\star$.

Our Result

- Goal: explain the reason that in practice even without *explicit* regularization, we do not learn $s^\star$ and hence there is only a moderate amount of memorization.

Theorem (Informal)

Consider the denoising score matching objective $\mathcal{R}_n$ over the class of two-layer neural networks for one-dimensional data.

Then, for a sufficiently small level of noise σ and a learning rate $\eta \gtrsim \sigma^2$, the stochastic gradient descent on $\mathcal{R}_n$ cannot stably converge to $s^\star$.

- Take home message: without vanishing learning rate, the model can not fully memorize the training data.

Problem Setup and Results

Model

- Two-layer ReLU network:

$$\mathcal{S} = \left\{ s_\theta : \mathbb{R} \to \mathbb{R} : s_\theta(y) = \frac{1}{m} \sum_{\ell=1}^{m} w_\ell^{(2)} \mathrm{ReLU}(w_\ell^{(1)} y + b_\ell) \right\}.$$

- Constraints on the weight for technical issue:

$$w_\ell^{(1)} \in \{\pm 1\}, w_\ell^{(2)} \in [-A, A],$$

  where $w_\ell^{(1)}$ is randomly initialized and fixed throughout training and we denote $\theta = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$.
- Training: SGD with learning rate η and mini-batch estimation $\hat{\mathcal{R}}_j$ of $\mathcal{R}_n$:

$$\theta_{j+1} = \theta_j - m\eta \nabla \hat{\mathcal{R}}_j(\theta_j).$$

Linear Stability

Definition

A local minimum $\theta^\star$ is said to be linearly stable if there is some $\varepsilon > 0$ such that, for any θ_0 in the ε-ball $\mathcal{B}_\varepsilon(\theta^\star)$, the following condition holds:

$$\limsup_{j\to\infty} \mathbb{E}\|\theta_j - \theta^\star\|_2 \leqslant \varepsilon.$$

In short, $\theta^\star$ can be converged by SGD if it is linearly stable.

Our Result

- Let $\Delta = \min_{1 \leq i < j \leq n} |x_i - x_j|$.

Theorem

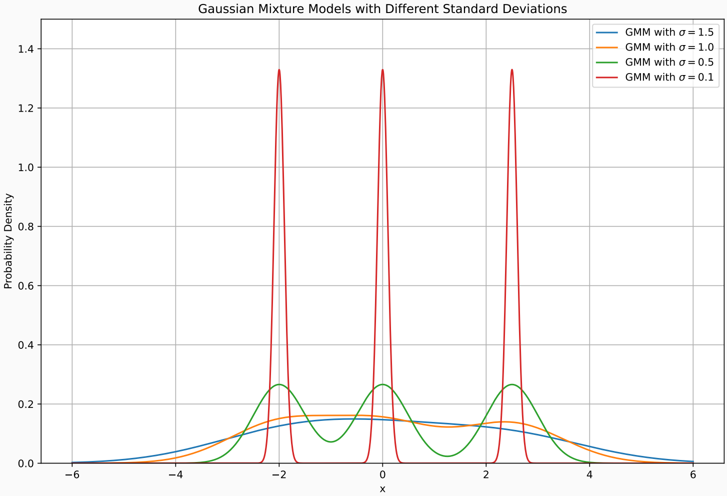
Let $\theta^\star \in \mathbb{R}^{2m}$ be a linearly stable local minimum of $\mathcal{R}_n$. Then there exist $\sigma_0, C > 0$, depending on μ and the training sample, such that if $\sigma \leqslant \sigma_0$ and $\eta > \frac{2^{12}\sigma^2}{\mu n^2 \Delta}$, one has

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y \sim \mathcal{N}(\mu x_i, \sigma^2)}[(s_{\theta^\star}(Y_i) - s^\star(Y_i))^2] > \frac{Cn^5\Delta^3}{A^4\sigma^4}.$$

- Interpretation: large learning rates prevent memorization.
- Remark: The more seperated the training data is (i.e., Δ is larger), the more $s_{\theta^\star}$ is away from $s^\star$.

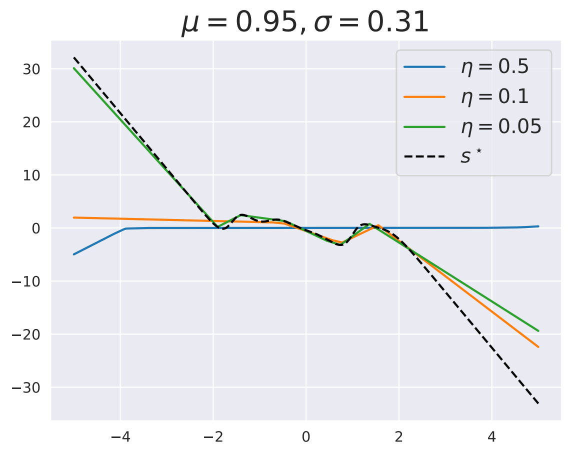
Proof Idea

Key observation: Probability density of mixtures of Gaussians becomes less and less regular as $\sigma \to 0$.



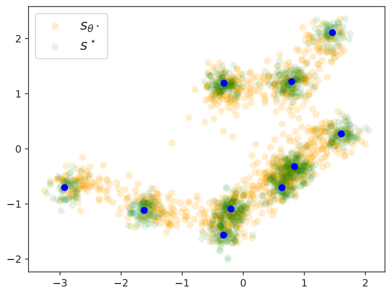
Proof Idea

Regularization effect of learning rate. Larger the learning rate, smoother the learnt score becomes.

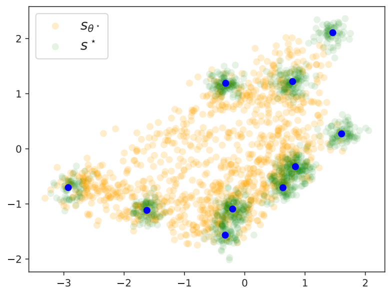


Experiments

Experiment: large learning rate leads to less memorization

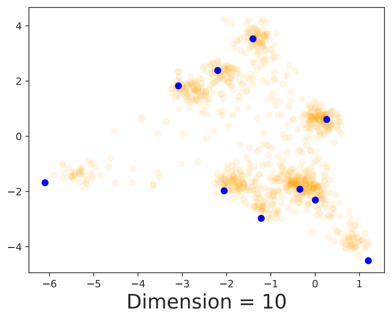


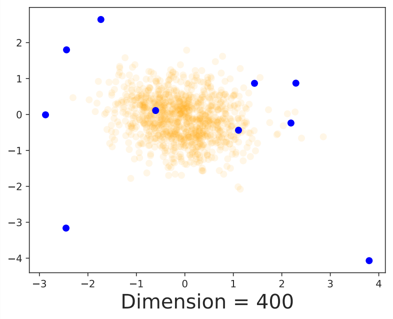
$\eta = 0.05$



$\eta = 2.0$

Experiment: increasing dimension reduces memorization





Conclusion

Conclusion

Summary:

- Non-vanishing learning rate prevents diffusion model from fully memorizing training data.
- Experiments suggest our result also applies to multidimensional data.

Future research interests:

- Generalization to high dimension data.
- Effects of sample size n and dimension d.
- Balancing η and quality of generated data.

Thank you