Denoising Score Matching with Random Features: Insights on Diffusion Models from Precise Learning Curves

Anand Jerry George
EPFL

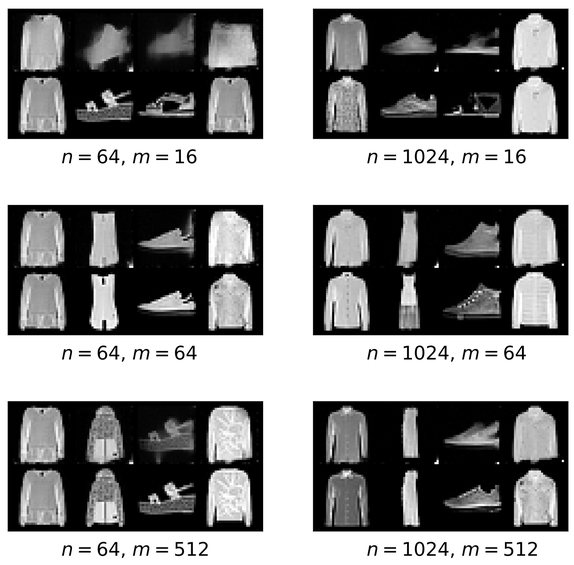
Rodrigo Veiga
University of Nottingham

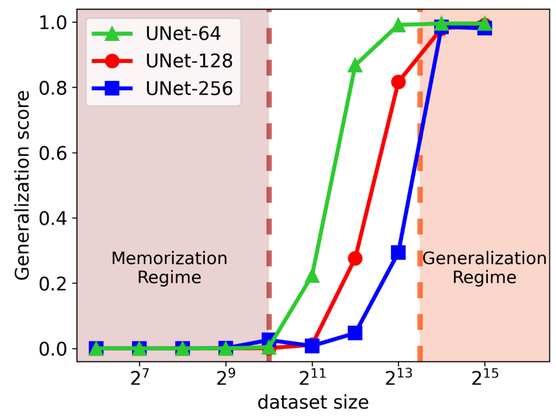
Nicolas Macris
EPFL

➢Motivation

- Diffusion models are powerful generative models but often **memorize** training data.
- Empirical findings suggest that **model complexity**, **dataset size** affect generalization and memorization.
- Furthermore, several studies show that there exists critical times (*speciation* and *collapse*) at which models starts to exhibit these behaviours.

Goal:
To understand the mechanisms behind these phenomena in an analytically tractable settings.





Zhang et. al. 2024

➢Problem Setup

- The forward process used is the standard Ornstein-Uhlenbeck process:. $dX_t = -X_t\,dt + \sqrt{2}\,dB_t\,, \quad X_0 \sim P_0(\mathbb{R}^d)$. The distribution of X_t given X_0 is simple: $X_t \sim \mathcal{N}(a_t X_0, h_t I_d)$, where $a_t = e^{-t}$, $h_t = 1 - e^{-2t}$
- We study the minimizer of the Denoising Score Matching objective:

$$\mathscr{L}^m_{\text{DSM}}(s) = \int_0^T dt \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} s(t, a_t x_i + \sqrt{h_t} z_{ij}^t) + z_{ij}^t \right\|^2 ; x_i \sim P_0,\ z_{ij} \sim \mathcal{N}(0, I_d)$$

- Score parameterized by *Random features* neural network $s_{A_t}(t,x) = A_t \varrho\left(W_t x\right)$.
- Data: $P_0 \sim \mathcal{N}(0, C)$ or comes from a *Hidden Manifold* Model; C can be isotropic, or can have low dimensional structure.
- Scaling regime: $d, D, n, p \to \infty$, $\psi_D = \dfrac{D}{d}$, $\psi_n = \dfrac{n}{d}$, $\psi_p = \dfrac{p}{d}$ fixed, where

  d : input dimension, D: dimension of latent space, n : # of training samples, p : # of random features.

➢Main Ideas

- Metrics used: Let $\hat{A}_t$ be the minimizer of the DSM loss. We define

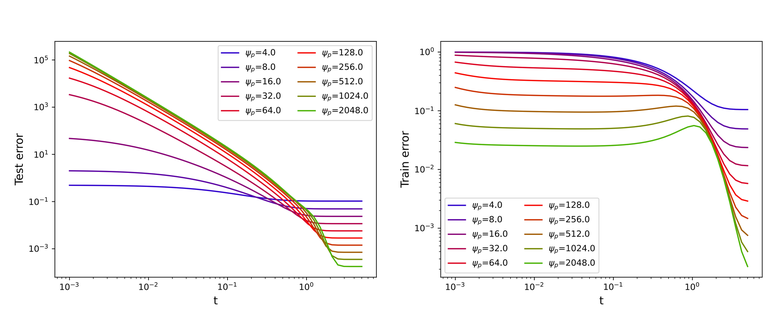
$$\mathscr{E}^m_{\text{test}}(\hat{A}_t) = \frac{1}{d}\mathbb{E}_{x\sim P_t} \left\| \hat{A}_t \varrho\left(W_t x\right) - \nabla \log P_t(x) \right\|^2$$

$$\mathscr{E}^m_{\text{train}}(\hat{A}_t) = \frac{1}{dnm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \sqrt{h_t} \hat{A}_t \varrho\left(W_t(a_t x_i + \sqrt{h_t} z_{ij}^t)\right) + z_{ij}^t \right\|^2$$

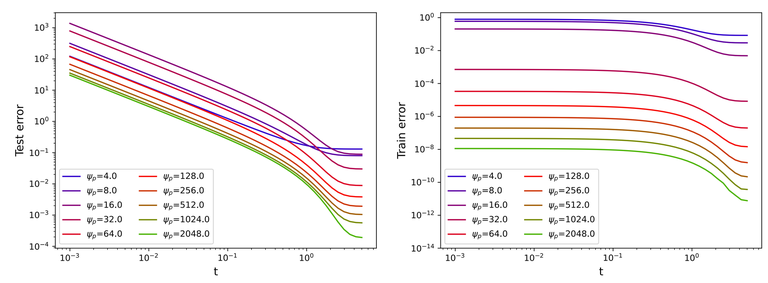
- We derive asymptotically precise values of $\mathscr{E}^m_{\text{test}}(\hat{A}_t)$ and $\mathscr{E}^m_{\text{train}}(\hat{A}_t)$ as a function of ψ_D, ψ_n, ψ_p, and t, for $m = 1, \infty$. We interpret the obtained curves to identify regimes of memorization and generalization.
- *Linear pencils* technique from Random Matrix Theory is used to derive the learning curves.

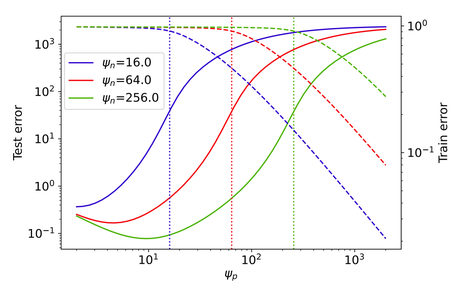
➢Results

- $m = \infty$

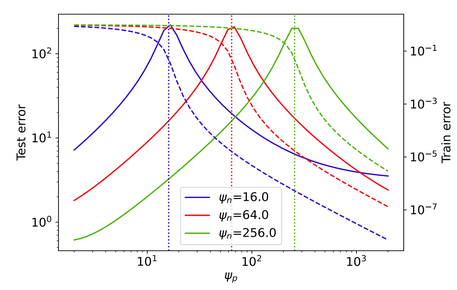


- $m = 1$





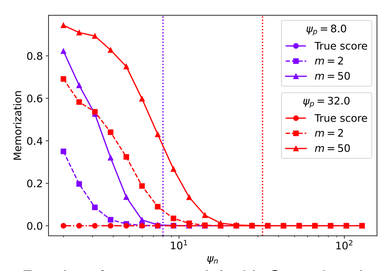
$m = \infty$



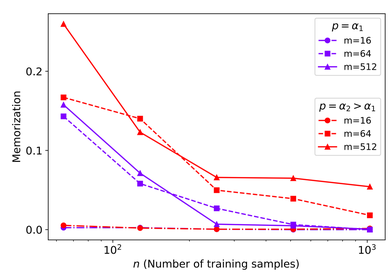
$m = 1$

Numerical results on memorization

- We simulate N backward diffusion instances Y_t, and check whether $Y_0$is closer to one of the training samples compared to the others.
- Let $\text{NN}_i(x)$ denote the ith nearest training sample of x. We say the backward diffusion retrieves a data sample if $\|Y_0 - \text{NN}_1(Y_0)\| < \delta \|Y_0 - \text{NN}_2(Y_0)\|$ with $\delta > 0$.
- We measure *memorization* as the fraction of N backward diffusion instances that retrieves one of the data samples.



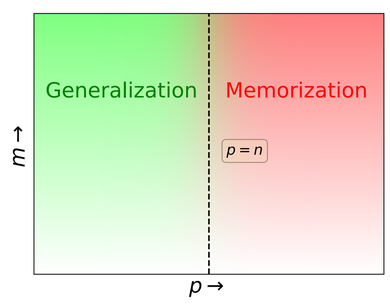
Random features model with Gaussian data



U-Net with FMNIST data

Phase diagram and key take aways

- Memorization increases as n/p decreases
- Memorization increases as m increases in the overparametrized regime.



SNSF Grant n° 200021_204119 / 1