Understanding Generalizability of Diffusion Models Requires Rethinking the Hidden Gaussian Structure

Xiang Li, Yixiang Dai, Qing Qu

I. Introduction

Problem: Diffusion models generate images by progressively denoising a random noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma(T)^2)$ to its corresponding clean image with a probabilistic ODE:

$$d\boldsymbol{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t))dt, \tag{1}$$

where $\sigma(t)$ is a predefined schedule. In practice the score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t))$ is approximated by:

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t)) = (\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t)) - \boldsymbol{x})/\sigma(t)^2, \tag{2}$$

where $\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))$ is a deep network with parameters $\boldsymbol{\theta}$ trained with the denoising score matching objective:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}, \boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0}, \sigma(t)^2\boldsymbol{I})} \|\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x} + \boldsymbol{\varepsilon}; \sigma(t)) - \boldsymbol{x}\|_2^2. \tag{3}$$

Since we don't have access to the ground truth p_{data}, in practice the denoising score matching (3) is instead performed on a finite number of training samples. Suppose the training dataset contains a finite number of data points $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N$, a natural way to model the data distribution is to model it as a multi-delta distribution $p(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{x} - \boldsymbol{y}_i)$. In this case, the optimal denoiser takes the form:

$$\mathcal{D}_M = \frac{\sum_{i=1}^N \mathcal{N}(\boldsymbol{x}; \boldsymbol{y_i}, \sigma(t)^2\boldsymbol{I})\boldsymbol{y}_i}{\sum_{i=1}^N \mathcal{N}(\boldsymbol{x}; \boldsymbol{y_i}, \sigma(t)^2\boldsymbol{I})}. \tag{4}$$

, which is essentially a softmax-weighted combination of the finite data points. However, such optimal denoisers can only generate exact replicas of the training samples, therefore have no generalizability. In this work, we aim to understand what kind of function is learned by the $\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))$ in practice.

II. Emerging Linearity in Diffuison Models

It is well-known that Diffusion models transition from memorization to generalization as the training dataset size increases. Interestingly, we observe that this transition is accompanied by an emerging linearity of $\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))$, as shown in Figure 1. Here the generalization score (GL Score) is defined as $\frac{1}{k}\sum_{i=1}^k \frac{\|\boldsymbol{x}_i - \text{NN}_{Y(\boldsymbol{x}_i)}\|_2}{\|\boldsymbol{x}_i\|_2^2}$, where NN denotes the nearest neighbor of $\boldsymbol{x}_i$ in the dataset Y and the linearity is measured by computing the cosine similarity between $\mathcal{D}_{\boldsymbol{\theta}}(\alpha\boldsymbol{x}_1 + \beta\boldsymbol{x}_2; \sigma(t))$ and $\alpha\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}_1; \sigma(t)) + \beta\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}_2; \sigma(t))$.

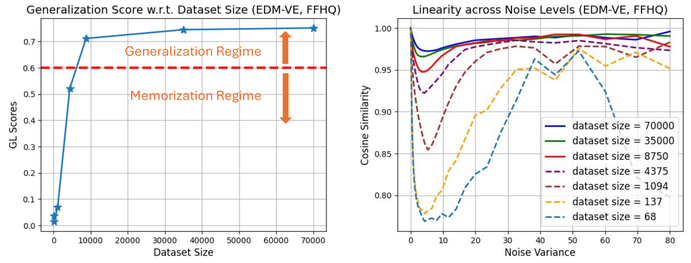

Figure 1: *Diffusion models exhibit increasing linearity as they transition from memorization to generalization.*

This emerging linearity motivates us to ask two questions: (i) to what extend can a diffusion model be approximated by a linear model and (ii) if diffusion models can be approximated linearly, what are the underlying characteristics of this linear approximation?

III. The Gaussian Inductive Bias

To address these questions, we propose to investigate the linear properties of diffusion models by finding their best linear approximations (with a bias term) $\mathcal{D}_L(\boldsymbol{x}; \sigma(t)) := \boldsymbol{W}_{\sigma(t)}\boldsymbol{x} + \boldsymbol{b}_{\sigma(t)}$ for a given diffusion denoiser $\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))$. Here $\boldsymbol{W}_{\sigma(t)}$ and $\boldsymbol{b}_{\sigma(t)}$ can be learned by solving the following optimization problem with gradient descent

$$\min_{\boldsymbol{W}_{\sigma(t)}, \boldsymbol{b}_{\sigma(t)}} \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}, \boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0}, \sigma(t)^2\boldsymbol{I})} \|\boldsymbol{W}_{\sigma(t)}(\boldsymbol{x}+\boldsymbol{\varepsilon}) + \boldsymbol{b}_{\sigma(t)} - \mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))\|_2^2. \tag{5}$$

After obtaining the linear denosiers, we can compare the differences between them and the actual diffusion denoisers $\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t))$ with the score approximation error defined as:

$$\text{Score-Difference(t)} := \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}, \boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0}, \sigma(t)^2\boldsymbol{I})} \sqrt{\frac{\|\mathcal{D}_{\boldsymbol{\theta}}(\boldsymbol{x}; \sigma(t)) - \mathcal{D}_L(\boldsymbol{x}; \sigma(t))\|_2^2}{d}}. \tag{6}$$

The results are shown in Figure 2, from which we observe the linear models generate samples that closely match those from the actual diffusion models, which highlights the important role of diffusion models' linear structure. Furthermore, the linear denoisers $\mathcal{D}_L(\boldsymbol{x}; \sigma(t))$ is nearly identical to $\mathcal{D}_G(\boldsymbol{x}; \sigma(t))$ with the following form:

$$\mathcal{D}_G(\boldsymbol{x}; \sigma(t)) := \boldsymbol{u} + \boldsymbol{U}\tilde{\Lambda}_{\sigma(t)}\boldsymbol{U}^{\boldsymbol{T}}(\boldsymbol{x} - \boldsymbol{u}), \tag{7}$$

where $\boldsymbol{u} = \frac{1}{N}\sum_i^N \boldsymbol{y}_i$, $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ are the mean and Covariance of the training dataset respectively, and $\tilde{\boldsymbol{\Lambda}}_{\sigma(t)} = \boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \sigma(t)^2\boldsymbol{I})^{-1}$.

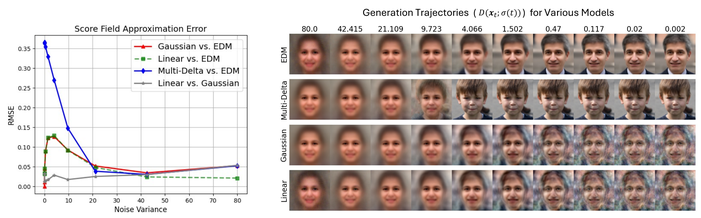

Figure 2: *Score approximation error and sampling Trajectory..*

Importantly, $\mathcal{D}_G(\boldsymbol{x}; \sigma(t))$ is the optimal solution to (3) under the assumption that $p_{\text{data}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{u}, \boldsymbol{\Sigma})$, i.e., a Multivariate Gaussian distribution. Our results demonstrate that diffusion models in practice have the inductive bias towards learning denoisers that are similar to the optimal denoisers under the Gaussian data assumption. We term this inductive bias as the Gaussian inductive bias.

IV. When does the inductive bias emerges?

Interestingly, the Gaussian inductive bias is most pronounced when the model capacity is relatively small and during the early training iterations. As illustrated in Figure 3, diffusion models generalize if we use a model with small capacity or applying early stopping. In such cases, the final generated images match those generated from the Gaussian denoisers $\mathcal{D}_G(\boldsymbol{x}; \sigma(t))$.

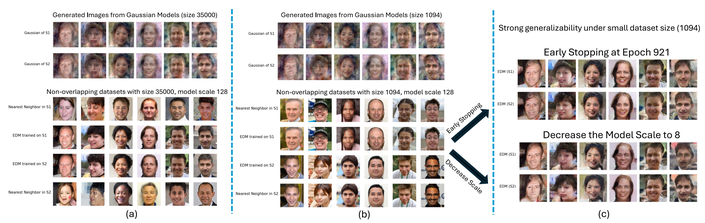

Figure 3: *Small model capacity and early stopping prompt generalization of diffusion models*