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

Anand Jerry George **Rodrigo Veiga** **Nicolas Macris**
École Polytechnique Fédérale de Lausanne (EPFL),
Lab for Statistical Mechanics of Inference in Large Systems (SMILS),
CH-1015 Lausanne,
Switzerland

Abstract

We theoretically investigate the phenomena of generalization and memorization in diffusion models. Empirical studies suggest that these phenomena are influenced by model complexity and the size of the training dataset. In our experiments, we further observe that the number of noise samples per data sample (m) used during Denoising Score Matching (DSM) plays a significant and non-trivial role. We capture these behaviors and shed insights into their mechanisms by deriving asymptotically precise expressions for test and train errors of DSM under a simple theoretical setting. The score function is parameterized by random features neural networks, with the target distribution being d-dimensional Gaussian. We operate in a regime where the dimension d, number of data samples n, and number of features p tend to infinity while keeping the ratios $\psi_n = \frac{n}{d}$ and $\psi_p = \frac{p}{d}$ fixed. By characterizing the test and train errors, we identify regimes of generalization and memorization as a function of ψ_n, ψ_p, and m. Our theoretical findings are consistent with the empirical observations.

1 INTRODUCTION

Generative models are at the heart of the ongoing revolution in artificial intelligence. Formally, they aim to generate new samples from an unknown probability distribution, given n i.i.d. samples drawn from it. Commercial generative models demonstrate remarkable capabilities across various modalities, including text, speech, images, and videos, with new advancements being reported regularly. Their impressive capabilities are mainly driven by two key architectures: *transformers* and *diffusion models*. While transformers excel primarily on text data, diffusion models show exceptional proficiency in generating natural-looking images from text prompts. Despite their success, the scientific community remains divided on whether these models are truly creative or merely imitate based on the examples that they have seen during training (Ukpaka, 2024; Kamb and Ganguli, 2024). An even more pressing concern is that of *memorization*, where the model's response partially or fully resembles training data. The memorization phenomenon raises serious implications, particularly regarding the privacy of data for training (Carlini et al., 2021, 2023). The limited theoretical understanding of these models prevents addressing such questions effectively.

In this study, we focus on the generalization and memorization behaviors of diffusion models. Our work is motivated by empirical observations of the factors affecting these properties in practical scenarios. Several prior works investigated the impact of model complexity and size of the training dataset on these behaviors (Somepalli et al., 2023, 2024; Carlini et al., 2023; Zhang et al., 2024; Yoon et al., 2023; Gu et al., 2023; Ross et al., 2024; Pham et al., 2024). In addition to this, our experiments also reveal that the number of noise samples per data sample (m) used in Denoising Score Matching (DSM) is another key factor impacting generalization and memorization. Details and results of our experiments with real and synthetic datasets can be found in Section 4 and in Appendix F. In a nutshell, the experiments suggest the following: 1) Memorization increases as the size of the training dataset decreases relative to model complexity (equivalently,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

it increases as model complexity increases relative to the number of data points), 2) Memorization increases as the number of noise samples per data sample increases in the overparametrized regime. Despite these observations, a theoretical study of memorization aspects of diffusion models has been done only with the *empirical optimal score* function (Biroli et al., 2024; Achilli et al., 2024, 2025; George et al., 2025) (see Section 2.2). Moreover, these analyses look at a regime in which the size of the training dataset scales exponentially with dimension, and do not shed much light on other regimes, e.g., a proportional one. This gap in theory and practice raises an important question: can the phenomenon of memorization be theoretically shown when using a parametric class of functions for the score function? This is precisely the context of our study. By analyzing the learning process of a specific diffusion model instance, we provide insights into generalization and memorization in diffusion models.

1.1 Main Contributions

This section provides a brief overview of our key contributions and findings. Our study focuses on the learning aspect of diffusion models, i.e., learning the score function of perturbed versions of a target distribution P_0 (see Eq. (2)). The task is to minimize a loss function known as *denoising score matching* (DSM) objective (see Eq. (5)) over a parametric class of functions, for which we consider random features neural networks (RFNNs). The target distribution is the d-dimensional Gaussian distribution, with n denoting the number of samples and p the number of features of the RFNN. We operate in a regime where $d, n, p \to \infty$, while the ratios $\psi_n = \frac{n}{d}$ and $\psi_p = \frac{p}{d}$ are kept fixed. The DSM objective involves an additional parameter m: the number of noise samples per data sample used in forming the loss function, Eq. (5). In this theoretical setting, we

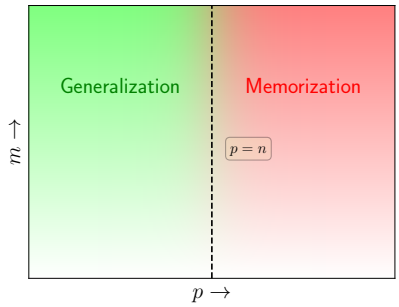


Figure 1: Phase diagram showing regimes of generalization and memorization. The gradient in color with m indicates the change in strength of the phenomenon.

make the following contributions, illustrated on the schematic phase diagram in Fig. 1:

- We analytically compute the test and train errors of the minimizer of the DSM loss.
- Using the obtained test and train errors, we study the generalization and memorization behavior in diffusion models.
- We show that a crossover transition between generalization and memorization occurs when $p = n$.
- We demonstrate that increasing the value of m enhances generalization when $p < n$, while it intensifies memorization when $p > n$.

The choice of a Gaussian target distribution P_0 provides an analytically tractable setting as long as we assume some simple structure for the covariance. It allows a precise theoretical analysis and captures phenomena observed in practice. We fully solve the case where P_0 is isotropic on a D-dimensional linear subspace, with $\psi_D = \frac{D}{d} \leq 1$ held fixed as $D, d \to \infty$.

1.2 Related Works

Diffusion models Diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2020) rely on the non-equilibrium dynamics of a diffusion process for generative modeling. Since their introduction, there were several improvements (Dhariwal and Nichol, 2021; Rombach et al., 2022; Ho and Salimans, 2021; Nichol et al., 2022) that led them to become the state-of-the-art in generative modeling of images. Further, the design aspects of diffusion models were studied in Karras et al. (2022).

Sampling accuracy, generalization, and memorization Several works have investigated the theoretical aspects of diffusion models. Sampling accuracy of the generative process in terms of distance from the target distribution was derived in Chen et al. (2022); Benton et al. (2023); Chen et al. (2023a); Bortoli (2022) for various settings. These works assume that the score function has been learned *a priori* with a certain level of accuracy. The score learning process was studied in Cui et al. (2023); Shah et al. (2023); Han et al. (2023); Zeno et al. (2025). The works Kadkhodaie et al. (2023); Chen et al. (2023b); Li et al. (2023); Wang et al. (2024); Cui et al. (2025) performed an end-to-end study of diffusion models giving a better understanding of their generalization properties. Specifically, Li et al. (2023) and Saha and Tran (2025) considers the score function parametrized by RFNNs and provides bounds on the KL divergence between the learned and target distributions. Further, the critical nature of feature emergence in diffusion models was studied in Li and Chen (2024); Sclocchi et al. (2025); Favero et al. (2025). The manifolds

learned by diffusion models were theoretically studied in Pidstrigach (2022). Statistical physics tools were employed to study memorization in high-dimensions using the empirical optimal score function in Biroli et al. (2024); Raya and Ambrogioni (2023); Ambrogioni (2024); Achilli et al. (2024, 2025). A geometric framework to understand memorization was proposed in Ross et al. (2024). More recently, Bonnaire et al. (2025) investigated how early stopping in training helps mitigate memorization. In particular, their theoretical analysis uses the same setting as ours (see Appendix G for further discussion).

1.3 Our Techniques

In the setting briefly described in Section 1.1 and to be detailed in Section 3, we analytically compute the test and train errors of the minimizer of the DSM loss. Our theoretical results are summarized in Theorems 3.2 and 3.6. Two main ingredients allowing the computation of exact asymptotic learning curves are the *Gaussian equivalence principle* (GEP) (Gerace et al., 2020; Goldt et al., 2022; Hu and Lu, 2023) and the theory of *linear pencils* (Far et al., 2006; Helton et al., 2018; Adlam and Pennington, 2020; Bodin, 2024). We briefly describe them here.

Gaussian Equivalence Principle Suppose $W \in \mathbb{R}^{p\times d}$, and $X \in \mathbb{R}^{d\times n}$ are random matrices with i.i.d. Gaussian entries. Let ϱ be a non-linear activation function and let $F = \varrho\left(\frac{W}{\sqrt{d}}X\right)$, where ϱ acts element-wise on a matrix. The GEP states that in the calculation of test and train errors, it is asymptotically equivalent to replace F by the matrix $\hat{F} = \mu_0 \mathbf{1}_p \mathbf{1}_n^T + \mu_1 \frac{W}{\sqrt{d}}X + v\Omega$, where $\Omega \in \mathbb{R}^{p\times n}$ is a random matrix with i.i.d. Gaussian entries, independent of W, X. The coefficients are given by $\mu_0 = \mathbb{E}_g[\varrho(g)]$, $\mu_1 = \mathbb{E}_g[\varrho(g)g]$, $v^2 = \mathbb{E}_g\left[\varrho(g)^2\right] - \mu_0^2 - \mu_1^2$, where $g \sim \mathcal{N}(0,1)$.

Linear Pencils The theory of linear pencils is a powerful technique to compute traces of rational functions involving random matrices with Gaussian entries. It entails constructing an appropriate block matrix called *linear pencil*, whose inverse gives the desired rational functions. We refer the reader to Chapter 3 of Bodin (2024) for further details on the method.

In our context, the test and train errors can be expressed as a sum of traces of rational functions of random matrices. Some of them have non-Gaussian entries due to ϱ. We circumvent this by using the GEP and, subsequently, the linear pencils method to compute the traces.

Notations The d-dimensional Gaussian distribution with mean μ and covariance Σ is denoted by $\mathcal{N}(\mu, \Sigma)$. The d-dimensional identity matrix is denoted by I_d, and $\mathbf{1}_d$ stands for the d-dimensional all-ones vector. $\|\cdot\|$ denotes the l_2 norm of a vector, while $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The operator ∇ represents the gradient of a scalar function.

2 PRELIMINARIES

2.1 Diffusion Models

Consider a set of n i.i.d. samples $\mathcal{S} = \{x_1, x_2, \cdots, x_n\}$ from an unknown distribution P_0 on $\mathbb{R}^d$. Diffusion models address the problem of leveraging the information in $\mathcal{S}$ to draw new samples from P_0 by time reversing a stochastic process transporting P_0 to a known distribution, such as the isotropic Gaussian. Letting this forward diffusion be an Ornstein-Uhlenbeck (OU) process (Gardiner, 2009), we have:

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \quad X_0 \sim P_0\,, \tag{1}$$

where B_t is a standard Brownian motion in $\mathbb{R}^d$. The distribution of X_t given X_0 can be computed in closed form from Eq. (1) and is simply $\mathcal{N}(a_t X_0, h_t I_d)$, where $a_t = e^{-t}$ and $h_t = 1 - e^{-2t}$ (hence, $a_t^2 + h_t = 1$). As $t \to \infty$, the distribution of X_t converges to the d-dimensional standard Gaussian, regardless of X_0, since $a_t \to 0$ and $h_t \to 1$. Let P_t denote the probability distribution of X_t:

$$P_t(x) = (2\pi h_t)^{-d/2} \int_{\mathbb{R}^d} \mathrm{d}P_0(x_0)\, e^{-\frac{\|x - a_t x_0\|^2}{2h_t}}\,. \tag{2}$$

Then, for a fixed $T > 0$ and $Y_T \sim P_T$, we define the *time reversal* of the forward process (1) as

$$-\mathrm{d}Y_t = (Y_t + 2\nabla \log P_t(Y_t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{B}_t\,, \tag{3}$$

where the process runs backward in time starting from Y_T, and $\tilde{B}_t$ is a different instance of standard Brownian motion. The term *time reversal* here means that the distributions of Y_t and X_t are identical for every t (Anderson, 1982; Haussmann and Pardoux, 1986). If the backward process is initiated with $Y_T \sim P_T$, the distribution of Y_0 will be P_0. However, since P_T is unknown due to the lack of knowledge of P_0, we instead start the reverse process with $Y_T \sim \mathcal{N}(0, I_d)$. This approximation incurs minimal error due to the OU process's exponential convergence to $\mathcal{N}(0, I_d)$.

The implementation of the backward process, Eq. (3), requires the so-called *score function* $\nabla \log P_t$, which is unknown as P_t is unknown. The learning task is to estimate the score function using the dataset $\mathcal{S}$. We consider minimizing the following score matching (Hyvärinen, 2005) objective for this: $\mathcal{L}_{\mathrm{SM}}(s) = \int_0^T \mathrm{d}t\ \mathbb{E}_{x_t\sim P_t}\left[\|s(t,x_t) - \nabla \log P_t(x_t)\|^2\right]$.

The $\mathcal{L}_{\text{SM}}$ loss function is not practical, as $\nabla \log P_t(x)$ is unknown. However, it is possible to construct an equivalent objective, the denoising score matching (DSM) loss (Vincent, 2011): $\mathcal{L}_{\text{DSM}}(s) = \int_0^T \mathrm{d}t\; w(t)\mathbb{E}\|s(t,x_t) - \nabla \log P_t(x_t|x_0)\|^2$, where w is a weighting function and the expectation is with respect to x_0 and x_t. In Appendix A.1, we show that $\mathcal{L}_{\text{DSM}}$ is equal to $\mathcal{L}_{\text{SM}}$ up to a time-dependent scaling factor and offset. Following Song et al. (2020), we choose $w(t) = (\mathbb{E}_{x_0,x_t}\|\nabla \log P_t(x_t|x_0)\|^2)^{-1}$. For OU process, we can compute $\nabla \log P_t(x_t|x_0)$ in closed form. We can write $x_t \sim P_t$ as $x_t = a_t x_0 + \sqrt{h_t} z$, where $x_0 \sim P_0$, $z \sim \mathcal{N}(0, I_d)$ are independent rvs and $a_t = e^{-t}$, $h_t = 1 - e^{-2t}$. Consequently, $\nabla \log P_t(x_t|x_0) = -\frac{(x_t - a_t x_0)}{h_t} = -\frac{z}{\sqrt{h_t}}$. The weight function is given by $w(t) = \frac{h_t}{d}$. Substituting these, we can write $\mathcal{L}_{\text{DSM}}(s) = \int_0^T \mathrm{d}t \frac{1}{d}\mathbb{E}\left\|\sqrt{h_t}s(t, a_t x_0 + \sqrt{h_t}z) + z\right\|^2$, where the expectation is with respect to x_0 and z. Since P_0 is unknown and only samples from it are available, we use an empirical estimate for the expectation with respect to x_0. Setting $y_i^t(z) = a_t x_i + \sqrt{h_t} z$, we define

$$\mathcal{L}^{\infty}_{\text{DSM}}(s) = \int_0^T \mathrm{d}t\; \frac{1}{dn}\sum_{i=1}^{n} \mathbb{E}_z \left\|\sqrt{h_t}s(t, y_i^t(z)) + z\right\|^2. \tag{4}$$

When s is a complicated function such as a neural network, which is the typical practical setting, computing the expectation over z might be intractable. In this case, one consideres m samples for z and uses an additional empirical estimate, obtaining the following loss function:

$$\mathcal{L}^{m}_{\text{DSM}}(s) = \int_0^T \mathrm{d}t\; \frac{1}{dnm}\sum_{i,j=1}^{n,m} \left\|\sqrt{h_t}s(t, y_{ij}^t) + z_{ij}^t\right\|^2 \tag{5}$$

where $y_{ij}^t(z) = a_t x_i + \sqrt{h_t} z_{ij}^t$.

2.2 Empirical Optimal Score and Memorization

Consider the loss function given in (4). It has an unique minimizer: $s^e(t,x) = \nabla \log P_t^e(x)$, with

$$P_t^e(x) = \frac{1}{n(2\pi h_t)^{d/2}} \sum_{i=1}^{n} e^{-\frac{\|x - a_t x_i\|^2}{2h_t}}. \tag{6}$$

The score s^e is often referred to as the *empirical optimal score*. A backward process using s^e converges in distribution to the empirical distribution of the dataset $\mathcal{S}$ as $t \to 0$. That is, the backward process collapses to one of the data samples as $t \to 0$. This is related to the memorization phenomenon as studied in Biroli et al. (2024), although their study focuses on the regime where $n = O\left(e^{\Theta(d)}\right)$.

Inspired by this connection, we define memorization as occurring when the score function learned using DSM closely approximates the empirical optimal score function instead of the exact score.

2.3 Random Features Model

In practice, the score function s is typically chosen from a parametric class of functions, and the DSM objective (5) is minimized within this class, with appropriate regularization. In this work, we represent the score function using a *random features* neural network (RFNN) (Rahimi and Recht, 2007). A RFNN is a two-layer neural network in which the first layer weights are randomly chosen and fixed, while the second layer weights are learned during training. It is a function from $\mathbb{R}^d$ to $\mathbb{R}^d$ of the form $s_A(x|W) = \frac{A}{\sqrt{p}}\varrho\left(\frac{W}{\sqrt{d}}x\right)$, where $W \in \mathbb{R}^{p\times d}$ is a random matrix with its elements chosen i.i.d. from $\mathcal{N}(0,1)$, ϱ is an activation function acting element-wise and $A \in \mathbb{R}^{d\times p}$ are the second layer weights that need to be learned. The RFNN is a simple neural network amenable to theoretical analysis. It is able to capture attributes frequently observed in more complex models, such as the double descent curve related to overparametrized regimes (Mei and Montanari, 2022; Bodin and Macris, 2021).

3 MAIN RESULTS

We study the DSM loss $\mathcal{L}^m_{\text{DSM}}$ given in (5) when the score function is modeled using a RFNN and $P_0 \equiv \mathcal{N}(0, \mathcal{C})$ with $\frac{1}{d}\mathrm{tr}\,\mathcal{C}$ tending to a finite limit as $d \to \infty$. Our results characterize the asymptotic learning curves (test and train errors) of the minimizer of $\mathcal{L}^m_{\text{DSM}}$ (5) in this setting. We obtain closed form expressions for learning curves when $\mathcal{C}$ has certain structure. Specifically, we assume that $\mathcal{C} = MM^T$, where $M \in \mathbb{R}^{d\times D}$ has orthonormal columns. This corresponds to the situation where data lies in a D-dimensional subspace of $\mathbb{R}^d$, inspired by the well known manifold hypothesis. We consider two values of m: $m = 1$ and $m = \infty$, which correspond to the extremes of the number of noise samples per data sample used during score learning. Based on the derived learning curves, we discuss the generalization and memorization behaviors in diffusion models. For intermediate values of m, we obtain the learning curves numerically in Appendix E.

We assume that at each time instant t, a different instance of RFNN is used to learn the score function specific to that time. Although this is a simplification compared to the methods employed in practice, it has been adopted in prior theoretical studies, e.g., Cui et al. (2023). When using a different instance of RFNN

at each t, note that minimizing the DSM loss (5) is equivalent to minimizing the integrand therein at each time instant. Henceforth, we focus on minimizing the DSM loss for a fixed t. Introducing a regularization parameter $\lambda > 0$, the loss function becomes:

$$\mathcal{L}_t^m(A_t) = \frac{1}{dnm} \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} \frac{A_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y_{ij}^t \right) + z_{ij}^t \right\|^2 + \frac{h_t \lambda}{dp} \|A_t\|_F^2 , \tag{7}$$

with $z_{ij}^t \sim \mathcal{N}(0, I_d)$. We emphasize that W_t is a different and independent random matrix for each t, and A_t is learned separately at each t. Denote $\hat{A}_t$ as the output of some algorithm minimizing (7). The respective performance is evaluated by the test and train errors defined as follows:

$$\mathcal{E}_{\text{test}}^m(\hat{A}_t) = \frac{1}{d} \mathbb{E}_{x \sim P_t} \left\| \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} x \right) - \nabla \log P_t(x) \right\|^2 , \tag{8}$$

$$\mathcal{E}_{\text{train}}^m(\hat{A}_t) = \frac{1}{dnm} \sum_{i,j=1}^{n,m} \left\| \sqrt{h_t} \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y_{ij}^t \right) + z_{ij}^t \right\|^2 . \tag{9}$$

The quantities $\mathcal{E}_{\text{test}}^m$ and $\mathcal{E}_{\text{train}}^m$ are random, due to W_t, $\{x_i\}_{i=1}^n$, and $\{z_{ij}^t\}_{i,j=1}^{n,m}$. We expect them to concentrate on their expectations as $d \to \infty$.

The rationale behind studying the test and training errors of DSM to assess the performance of diffusion models stems from their direct connection to the model's generative accuracy. Specifically, as described in Song et al. (2021), the error in diffusion models can be upper bounded using the test error. Suppose $\hat{P}_t$ denotes the probability distribution of the backward process when we use the learned score function instead of the true score, and assume that $\hat{P}_T = P_T$. Then, the Kullback-Leibler (KL) divergence between P_0 and $\hat{P}_0$ can be upper bounded as $D_{KL}(P_0 || \hat{P}_0) \leq \frac{d}{2} \int_0^T \mathrm{d}t \, \mathcal{E}_{\text{test}}^m(\hat{A}_t)$.

3.1 Infinite Noise Samples per Data Sample

For $m = \infty$, the average over z_{ij}^t in the DSM loss becomes an expectation. We write the loss as

$$\mathcal{L}_t^\infty(A_t) = \frac{1}{dn} \sum_{i=1}^{n} \mathbb{E}_z \left\| \sqrt{h_t} \frac{A_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y_i^t(z) \right) + z \right\|^2 + \frac{h_t \lambda}{dp} \|A_t\|_F^2 , \tag{10}$$

and characterize the minimizer in the asymptotic regime $d, n, p \to \infty$, with $\psi_n = \frac{n}{d}$ and $\psi_p = \frac{p}{d}$ fixed. We make the following assumption on ϱ.

Assumption 3.1. The activation function ϱ has a Hermite polynomial expansion given by $\varrho(y) = \sum_{l=0}^{\infty} \mu_l \mathrm{He}_l(y)$, where He_l is the l^{th} Hermite polynomial. For ease of presentation, assume that $\mu_0 = 0$. The L_2 norm of ϱ with respect to the standard Gaussian measure is denoted by $\|\varrho\|$. The function c is defined as $c(\gamma) = \mathbb{E}_{u,v \sim P^\gamma}[\varrho(u)\varrho(v)]$, with P^γ denoting the bivariate standard Gaussian distribution with correlation coefficient γ (see Eq. (14) in the Appendix B).

Theorem B.2 in Appendix B characterizes the test and train errors when P_0 is a Gaussian supported on a D-dimensional subspace in $\mathbb{R}^d$. Here, for simplicity, we present the theorem for $\mathcal{C} = I_d$, which already captures the main phenomena.

Theorem 3.2. *Suppose $P_0 \equiv \mathcal{N}(0, I_d)$ and ϱ satisfies Assumption 3.1. Let $s^2 = \|\varrho\|^2 - c(a_t^2) - h_t \mu_1^2$, $v_0^2 = c(a_t^2) - a_t^2 \mu_1^2$, and $v^2 = \|\varrho\|^2 - \mu_1^2$. Let $\psi_n = \frac{n}{d}$, and $\psi_p = \frac{p}{d}$. Let $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ be the solution of the following set of algebraic equations as a function of q and z:*

$$\begin{aligned}
\zeta_2(\psi_n + v_0^2 \psi_p \zeta_1 - a_t \mu_1 \zeta_3) - \psi_n &= 0 , \\
a_t^2 \mu_1^2 \psi_p \zeta_1 \zeta_2 \zeta_4 + (1 + (h_t \mu_1^2 + q) \psi_p \zeta_1) \zeta_4 - 1 &= 0 , \\
\zeta_1(s^2 - z + a_t^2 \mu_1^2 \zeta_2 \zeta_4 + v_0^2 \zeta_2 + (h_t \mu_1^2 + q) \zeta_4) - 1 &= 0 , \\
a_t^2 \mu_1^2 \psi_p \zeta_1 \zeta_2 \zeta_4 + (1 + (h_t \mu_1^2 + q) \psi_p \zeta_1) \zeta_4 - 1 &= 0 .
\end{aligned}$$

Define the function $\mathcal{K}(q, z) = -\frac{\zeta_3(q,z)}{a_t \mu_1}$. Let $\varepsilon_{test}^\infty = 1 - 2\mu_1^2 \mathcal{K}(0, -\lambda) - \mu_1^4 \frac{\partial \mathcal{K}}{\partial q}(0, -\lambda) + \mu_1^2 v^2 \frac{\partial \mathcal{K}}{\partial z}(0, -\lambda)$, and $\varepsilon_{train}^\infty = 1 - \mu_1^2 h_t \mathcal{K}(0, -\lambda) - \mu_1^2 \lambda h_t \frac{\partial \mathcal{K}}{\partial z}(0, -\lambda)$. Then, for the minimizer of (10) $\hat{A}_t$, as $d, n, p \to \infty$:

$$\begin{aligned}
\lim_{d,n,p \to \infty} \mathbb{E}\left[\mathcal{E}_{test}^\infty(\hat{A}_t) \right] &= \varepsilon_{test}^\infty , \\
\lim_{d,n,p \to \infty} \mathbb{E}\left[\mathcal{E}_{train}^\infty(\hat{A}_t) \right] &= \varepsilon_{train}^\infty .
\end{aligned}$$

We defer the proof to Appendix B.

Remark 3.3. We expect $\mathcal{E}_{\text{test}}^\infty(\hat{A}_t)$ and $\mathcal{E}_{\text{train}}^\infty(\hat{A}_t)$ to concentrate on their expectations as $d \to \infty$, while a rigorous proof is beyond the scope of the current work. Henceforth, we use the term test (train) error for both $\mathcal{E}_{\text{test}}^\infty$ ($\mathcal{E}_{\text{train}}^\infty$) and $\mathbb{E}[\mathcal{E}_{\text{test}}^\infty]$ ($\mathbb{E}[\mathcal{E}_{\text{train}}^\infty]$), interchangeably.

Remark 3.4. In Appendix B, Lemma B.1, we first obtain expressions for non-averaged errors when $P_0 = \mathcal{N}(0, \mathcal{C})$. In order to derive closed-form formulas for their expected values, one has to assume some structure for the covariance $\mathcal{C}$. In Theorem B.2 we present the scenario where P_0 is a Gaussian distribution supported on a D-dimensional subspace in $\mathbb{R}^d$; and analytically obtained curves are illustrated for $\frac{D}{d} = 0.2$ in Appendix D.2

Theorem 3.2 allows the computation of test and train errors for different values of t, ψ_n, ψ_p. Fig. 2 shows the

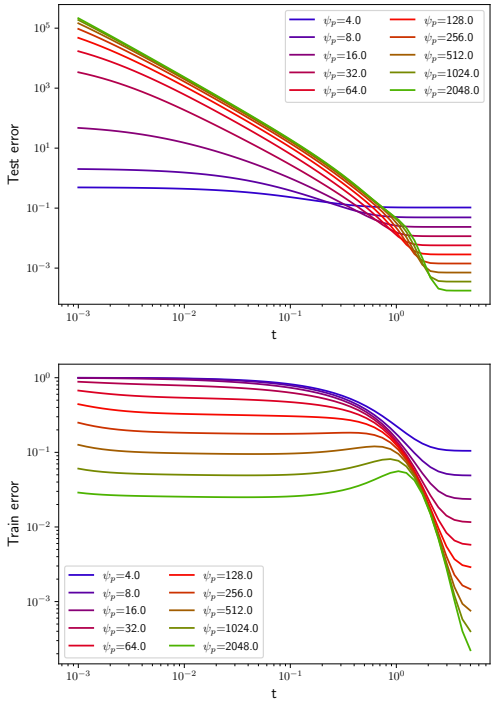


Figure 2: Learning curves for $m = \infty$, with $\psi_n = 20.0, \lambda = 0.001$. The activation function is ReLU.

results as a function of t for various ψ_p with fixed ψ_n. We can comprehend them by decomposing $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t)$ into bias and variance components through the following Lemma, which is proved in Appendix A.2.

Lemma 3.5. *Suppose $\mathcal{M}_t$ and $\mathcal{V}_t$ are given by*

$$\mathcal{M}_t = \frac{1}{dn}\sum_{i=1}^{n}\mathbb{E}_z\left\|\frac{\hat{A}_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}y_i^t(z)\right) - s^e(t, y_i^t(z))\right\|^2,$$

$$\mathcal{V}_t = \frac{1}{dn}\sum_{i=1}^{n}\mathbb{E}_z\left\|\sqrt{h_t}s^e(t, y_i^t(z)) + z\right\|^2,$$

where $s^e(t,x) = \nabla \log P_t^e(x)$ with P_t^e given by Eq. (6). Then, $\mathcal{E}^{\infty}_{train}(\hat{A}_t) = \mathcal{V}_t + h_t\mathcal{M}_t$.

We discuss Fig. 2 by exploring the bias-variance decomposition above for different values of t. A general observation is that, since $\mathcal{V}_t$ is independent of $\hat{A}_t$, any change in $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t)$ as the relative number of features ψ_p varies must be attributed to changes in $\mathcal{M}_t$. A smaller $\mathcal{M}_t$ indicates that the learned score is closer to the empirical optimal score. Such closeness can harm the generalization performance of diffusion models. This can be understood as follows:

For small t, $h_t \approx 0$, so in the neighborhood of $a_t x_i$, P_t^e is dominated by the i^{th} term in (6). Therefore, in the vicinity of $a_t x_i$, we have $s^e(t,x) \approx -\frac{x - a_t x_i}{h_t}$. In the backward process, this translates to a component in the drift pointing towards $a_t x_i$. Consequently, the trajectories will tend to move toward training samples, causing the output of the backward process to resemble one of them. This behavior is referred to as *memorization*, occurring when the learned score closely approximates the empirical optimal one. With this in mind, we now qualitatively discuss the curves in Fig. 2 for different values of t.

- $t = \infty$: we have $a_t = 0, h_t = 1$ and $s^e(t,y) = -y$, leading to $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) = \mathcal{M}_t = \mathbb{E}_z\left\|\frac{\hat{A}_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}z\right) + z\right\|^2$. The train and test errors are then equal and depend only on how well the RFNN can approximate a linear function. As ψ_p increases, the approximation power of the RFNN increases, and thus the train and test errors decrease.
- $t \gg 1$: again $s^e(t,y) \approx -y$, leading to $\mathcal{V}_t \approx a_t^2$. Hence, $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) \approx h_t\mathcal{M}_t + a_t^2$. When ψ_p is large, $\mathcal{M}_t$ is small, and the train error is dominated by the a_t^2 term. Therefore, the train error increases rapidly as t decreases. However, the test error remains constant, since it is approximately equal to $\mathcal{M}_t$.
- $t \ll 1$: the empirical optimal score satisfies $s^e(t, a_t x_i + \sqrt{h_t}z) \approx -z/\sqrt{h_t}$ for any data point x_i. Consequently, $\mathcal{V}_t \approx 0$. This leads to $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) \approx h_t\mathcal{M}_t$. In this regime, train error depends on how well the RFNN approximates s^e, decreasing as ψ_p grows. By contrast, since the actual score significantly deviates from the empirical one, test error rises quickly with ψ_p. In the neighborhood of $a_t x_i$, learning s^e instead of the true score makes the drift in the backward process pull the trajectories towards $a_t x_i$, so if a trajectory enters the vicinity of $a_t x_i$ at time t, the model tends to recover the sample x_i.

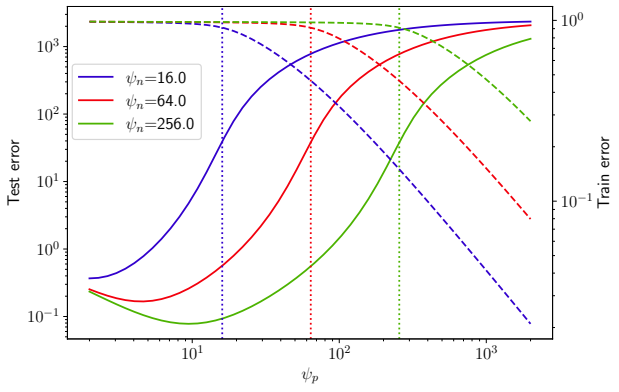


Figure 3: Learning curves for $m = \infty$, with $t = 0.01, \lambda = 10^{-3}$, $\varrho \equiv$ReLU. Solid (dashed) lines: test (train) error. Dotted vertical lines indicate $\psi_p = \psi_n$.

The analysis of Fig. 2 hints that the RFNN starts to show memorization behavior for $t < 1$ and large ψ_p.

To explore this further, we plot in Fig. 3 the learning curves as a function of ψ_p for different ψ_n, with t kept fixed and small. We observe: 1) For $\psi_p \ll \psi_n$, train error remains constant and test error is small, indicating generalization. 2) For $\psi_p \gg \psi_n$, train error is small and test error high, indicating the presence of memorization. 3) When $\psi_p \approx \psi_n$, test error rises rapidly, signaling the onset of memorization.

Therefore, $\psi_p = \psi_n$ acts as a crossover point at which the behavior of the score transitions from generalization phase to memorization phase. This is depicted in Fig. 1 as a phase diagram.

More curves illustrating the effects of λ and ϱ are found in Appendices D.3 and D.4. Our theory is consistent with previous findings on the roles of n and p on generalization and memorization in practical settings (Zhang et al., 2024; Yoon et al., 2023; Gu et al., 2023).

3.2 Single Noise Sample per Data Sample

For $m = 1$, the DSM loss (7) reduces to

$$\mathcal{L}^1(A_t) = \frac{1}{dn}\sum_{i=1}^{n}\left\|\sqrt{h_t}\frac{A_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}y_i^t\right) + z_i^t\right\|^2 + \frac{h_t\lambda}{dp}\|A_t\|_F^2 \,. \tag{11}$$

The train and test errors of its minimizer for $\mathcal{C} = I_d$ are characterized below by Theorem 3.6. The more general case, where P_0 is supported on a D-dimensional subspace in $\mathbb{R}^d$, is deferred to Appendix C .

Theorem 3.6. *Let $P_0 \equiv \mathcal{N}(0, I_d)$, and let ϱ satisfy Assumption 3.1. Define $v^2 = \|\varrho\|^2 - \mu_1^2$, $\psi_n = \frac{n}{d}$, and $\psi_p = \frac{p}{d}$. Let $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ be the solution to the following system of algebraic equations in q and z:*

$$\begin{aligned}
\zeta_1(-z + (q + \mu_1^2\zeta_4)\zeta_3 + v^2\zeta_4) - 1 &= 0 \,,\\
\zeta_2(1 + q\psi_p\zeta_1) + \mu_1^2\psi_p\zeta_1\zeta_2\zeta_4 + \psi_p a_t\mu_1\zeta_1 &= 0 \,,\\
\zeta_3(1 + q\psi_p\zeta_1) + \mu_1^2\psi_p\zeta_1\zeta_3\zeta_4 - 1 &= 0 \,,\\
\zeta_4(\psi_n + \psi_p v^2\zeta_1 - \mu_1\zeta_2/a_t) - \psi_n &= 0 \,.
\end{aligned} \tag{12}$$

Define the function $\mathcal{K}(q,z) = (1 - \zeta_4(q,z)(1 + \mu_1 h_t\zeta_4(q,z)\zeta_2(q,z)/a_t))$. Let $\varepsilon^1_{test} = 1 + 2\mu_1\zeta_2(0,-\lambda)\zeta_4(0,-\lambda)/a_t - \frac{\mu_1^2}{h_t}\frac{\partial\mathcal{K}}{\partial q}(0,-\lambda) + \frac{v^2}{h_t}\frac{\partial\mathcal{K}}{\partial z}(0,-\lambda)$, and $\varepsilon^1_{train} = 1 - \mathcal{K}(0,-\lambda) - \lambda\frac{\partial\mathcal{K}}{\partial z}(0,-\lambda)$. Then, for the minimizer of (11) $\hat{A}_t$, as $d,n,p \to \infty$:

$$\lim_{d,n,p\to\infty}\mathbb{E}\left[\mathcal{E}^1_{test}(\hat{A}_t)\right] = \varepsilon^1_{test} \,,$$
$$\lim_{d,n,p\to\infty}\mathbb{E}\left[\mathcal{E}^1_{train}(\hat{A}_t)\right] = \varepsilon^1_{train} \,.$$

The proof is given in Appendix C, and Remarks 3.3-3.4 apply here as well. Using Theorem 3.6, we compute test and train errors as functions of t, ψ_n and ψ_p, as shown in Fig. 5 in Appendix D.1. In Appendix D.2, we also illustrate the case where P_0 is supported on the D dimensional subspace with $\frac{D}{d} = 0.2$.

Fig. 5a shows learning curves as a function of t for different ψ_p with fixed ψ_n. Several notable trends emerge. Test error increases as t decreases but varies non-monotonically with ψ_p. Train error decreases monotonically with ψ_p for all t, indicating the model's progressive capacity to interpolate the training data.

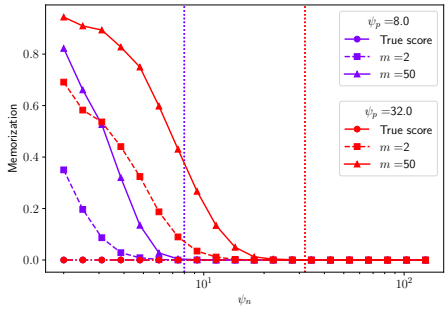
Note that for small t, the test error remains at least two orders of magnitude smaller than in the $m = \infty$ case. Further, it decreases as ψ_p increases beyond ψ_n. These observations suggest the model does not display memorization behavior when m equals 1. This contrasts the $m = \infty$ case, where memorization significantly impacts the test error. These findings indicate that larger values of m increase the propensity to memorize, justifying the illustration in Fig. 1.

The DSM loss (11) shares similarities with the loss used for RFNN in regression settings. Several works such as Mei and Montanari (2022); Bodin and Macris (2021); Hu et al. (2024), have analyzed the learning curves of RFNN in regression contexts. Notably, they predict the presence of a double descent phenomenon: test error peaks at $\psi_p = \psi_n$ and decreases for $\psi_p < \psi_n$ or $\psi_p > \psi_n$. The point $\psi_p = \psi_n$ is called the interpolation threshold. We also observe double descent in the DSM setting with $m = 1$, as shown in Fig. 5b.

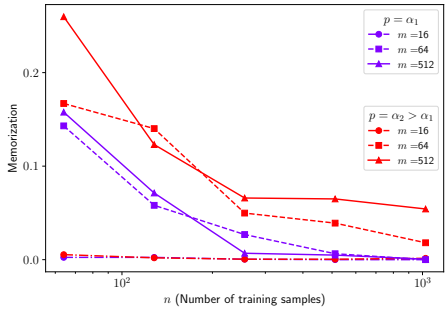
Intermediate values of m Simulations of learning curves for intermediate m are presented in Appendix E. They also confirm the predictions of Theorems 3.2 and 3.6. When $\psi_p > \psi_n$, the increase in memorization with a larger m can be explained as follows: For large m and small t, the learned score $\hat{s}(x)$ effectively approximates $\frac{x_1 - x}{h_t}$ in the vicinity of the training sample x_1. During the sampling phase, the drift term is given by $Y_t + 2\hat{s}(Y_t)$, implying that whenever the sampling trajectory is near a training sample, there is a drift towards it, leading to memorization. However, for $\psi_p < \psi_n$, the model's limited expressibility helps elude memorization even if m is large.

4 NUMERICAL EXPERIMENTS ON MEMORIZATION

We now conduct numerical experiments to illustrate the effects of n, p and m on memorization when a learned score function $s_{\hat{\theta}}$ is used in the backward diffusion (3). Specifically, we simulate N instances of $-\mathrm{d}Y_t = \big(Y_t + 2s_{\hat{\theta}}(t, Y_t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{B}_t$. We stop at $t_0 = 10^{-5}$ and check whether Y_{t_0} is closer to one of



(a) Gaussian data and RFNN score ($N = 5000$ and $\delta = 1/3$). The dotted vertical lines are for $\psi_n = \psi_p$.



(b) Fashion-MNIST dataset and U-Net score ($N = 2000$ and $\delta = 1/2$).

Figure 4: Results of experiments on memorization.

the training samples compared to the others (Yoon et al., 2023). Let $\mathrm{NN}_i(x)$ denote the i^{th} nearest neighbor of x in $\{x_1, x_2, \cdots, x_n\}$. We say the backward diffusion retrieves a data sample if $\|Y_{t_0} - \mathrm{NN}_1(Y_{t_0})\| < \delta\|Y_{t_0} - \mathrm{NN}_2(Y_{t_0})\|$, with $\delta > 0$. We measure *memorization* as the fraction of N backward diffusion instances that retrieves one of the data samples.

Gaussian data and RFNN score First, we measure memorization when the $P_0 \equiv \mathcal{N}(0, I_d)$ and $s_{\hat{\theta}}(t, y) = \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y\right)$ with $\hat{A}_t$ being the minimizer of (7). We demonstrate that, when $\psi_p > \psi_n$, once the backward diffusion process enters the vicinity of a data sample, it exhibits a tendency to remain within that neighborhood. More precisely, with $t_1 = 0.1$, we start the backward diffusion at $Y_{t_1} = a_{t_1} x_l + \sqrt{h_{t_1}} z$, where l is selected uniformly at random from the set $\{1, 2, \cdots, n\}$, and $z \sim \mathcal{N}(0, I_d)$. Fig. 4a shows the measure of memorization described in the previous paragraph as a function of ψ_n for different values of ψ_p and m. We note the following: 1) For $m = 50$ and $\psi_n < \psi_p$, the memorization is high. 2) The memorization decreases as m decreases. 3) Memorization is zero when $\psi_n > \psi_p$.

Real datasets and U-Net score Next, we measure memorization for Fashion-MNIST and MNIST datasets when $s_{\hat{\theta}}$ is a learned neural network with U-Net (Ronneberger et al., 2015) based architecture. We use a subset of MNIST/Fashion-MNIST dataset of size n for training. For a fixed n, m, and U-Net neural network, we train the U-Net for a large number of epochs (see Appendix F for more details). During the sampling phase, we use the learned score in the backward diffusion. In Fig. 4b, we plot the memorization for the Fashion-MNIST dataset as a function of n for different m, and for two U-Net neural networks with different number of parameters (p). Corresponding results for the MNIST dataset can be found in Fig. 12b in Appendix F. We draw the following conclusions from this experiment: 1) Memorization increases as $\frac{n}{p}$ decreases and 2) Memorization increases as m increases in the overparametrized regime.

Remarkably, our theoretical setting captures the effect of n, p and m on memorization in practical situations.

5 DISCUSSION AND FUTURE WORK

We studied the mechanisms underlying generalization and memorization in diffusion models by analyzing train and test errors of DSM with random features and Gaussian data. Memorization is typically observed in practice when highly overparameterized neural networks are trained long enough. Consistent with this, our findings indicate that model complexity (p) and the number of noise samples per data sample (m) used during DSM play a significant role in memorization. In practical implementations, m increases as the number of training epochs increases.

As the complexity of the model increases, it can approximate more complex functions, leading to a better approximation of the empirical optimal score. Furthermore, as m increases, an optimizer of DSM loss (5) tends to be close to the empirical optimal score. These effects that lead to memorization are captured in our results and are illustrated as a phase diagram in Fig. 1. As expected, generalization occurs when the complexity of the model is limited, i.e., $p < n$.

Although our analysis relies on simplifications for analytical tractability: RFNNs for the score, unstructured data and an independent estimator for each time t, it is a first step toward understanding the generalization and memorization of DSM in diffusion models. In particular, we provide, to the best of our knowledge, the first theoretical study of the role of m in memorization. Furthermore, the behavior predicted by our theory under this simplified setting can indeed be seen on real data with practical models, as shown in Fig. 4b.

In our analysis, we adopt an Ornstein–Uhlenbeck

(OU) stochastic differential equation (SDE) as the forward process, corresponding to the so-called *variance-preserving* (VP) framework in the diffusion literature. An alternative and widely used choice is the *variance-exploding* (VE) framework, in which the forward process is given by a Brownian motion with increasing variance. Our analysis can be extended to this setting in a straightforward manner by appropriately choosing a_t and h_t. We leave a detailed investigation of this case for future work.

Several additional directions could further refine and generalize our analysis. For instance, one could incorporate explicit time dependence in the score function, consider richer and more expressive classes of score models, or relax the Gaussian assumption on the data distribution. Addressing these challenges would bring the theoretical framework closer to practical implementations of diffusion-based generative models and may shed further light on their empirical success.

Acknowledgements

The work of A. J. G. and R. V. has been supported by Swiss National Science Foundation grant number 200021-204119.

References

Achilli, B., Ambrogioni, L., Lucibello, C., Mézard, M., and Ventura, E. (2025). Memorization and Generalization in Generative Diffusion under the Manifold Hypothesis. arXiv:2502.09578 [cond-mat].

Achilli, B., Ventura, E., Silvestri, G., Pham, B., Raya, G., Krotov, D., Lucibello, C., and Ambrogioni, L. (2024). Losing dimensions: Geometric memorization in generative diffusion. arXiv:2410.08727.

Adlam, B. and Pennington, J. (2020). The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 74–84. PMLR. ISSN: 2640-3498.

Ambrogioni, L. (2024). In Search of Dispersed Memories: Generative Diffusion Models Are Associative Memory Networks. *Entropy*, 26(5):381.

Anderson, B. D. O. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326.

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2023). Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization. In *The Twelfth International Conference on Learning Representations*.

Biroli, G., Bonnaire, T., de Bortoli, V., and Mézard, M. (2024). Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957. Publisher: Nature Publishing Group.

Bodin, A. and Macris, N. (2021). Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. In *Advances in Neural Information Processing Systems*, volume 34, pages 21605–21617. Curran Associates, Inc.

Bodin, A. P. M. (2024). *Random matrix methods for high-dimensional machine learning models*. PhD thesis, EPFL.

Bonnaire, T., Urfin, R., Biroli, G., and Mézard, M. (2025). Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training. arXiv:2505.17638 [cs].

Bortoli, V. D. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, pages 5253–5270, USA. USENIX Association.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Chen, H., Lee, H., and Lu, J. (2023a). Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 4735–4763, Honolulu, Hawaii, USA. JMLR.org.

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4672–4712. PMLR. ISSN: 2640-3498.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. (2022). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*.

Cui, H., Krzakala, F., Vanden-Eijnden, E., and Zdeborova, L. (2023). Analysis of Learning a Flow-based Generative Model from Limited Sample Complexity. In *The Twelfth International Conference on Learning Representations*.

Cui, H., Pehlevan, C., and Lu, Y. M. (2025). A precise asymptotic analysis of learning diffusion models: theory and insights. arXiv:2501.03937 [cs].

Dhariwal, P. and Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

Far, R. R., Oraby, T., Bryc, W., and Speicher, R. (2006). Spectra of large block matrices. arXiv:cs/0610045.

Favero, A., Sclocchi, A., Cagnetta, F., Frossard, P., and Wyart, M. (2025). How compositional generalization and creativity improve as diffusion models are trained. arXiv:2502.12089 [stat].

Gardiner, C. W. (2009). *Stochastic methods: a handbook for the natural and social sciences*. Number 13 in Springer series in synergetics. Springer, Berlin Heidelberg, 4th ed edition.

George, A. J., Veiga, R., and Macris, N. (2025). Analysis of Diffusion Models for Manifold Data. arXiv:2502.04339 [math].

Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3452–3462. PMLR. ISSN: 2640-3498.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mezard, M., and Zdeborova, L. (2022). The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 426–471. PMLR. ISSN: 2640-3498.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. (2023). On Memorization in Diffusion Models. arXiv:2310.02664 [cs].

Han, Y., Razaviyayn, M., and Xu, R. (2023). Neural Network-Based Score Estimation in Diffusion Models: Optimization and Generalization. In *The Twelfth International Conference on Learning Representations*.

Haussmann, U. G. and Pardoux, E. (1986). Time Reversal of Diffusions. *The Annals of Probability*, 14(4):1188–1205. Publisher: Institute of Mathematical Statistics.

Helton, J. W., Mai, T., and Speicher, R. (2018). Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Ho, J. and Salimans, T. (2021). Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, H. and Lu, Y. M. (2023). Universality Laws for High-Dimensional Learning With Random Features. *IEEE Transactions on Information Theory*, 69(3):1932–1964. Conference Name: IEEE Transactions on Information Theory.

Hu, H., Lu, Y. M., and Misiakiewicz, T. (2024). Asymptotics of Random Feature Regression Beyond the Linear Scaling Regime. arXiv:2403.08160.

Hyvärinen, A. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (2023). Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*.

Kamb, M. and Ganguli, S. (2024). An analytic theory of creativity in convolutional diffusion models. arXiv:2412.20292 [cs].

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. *Advances in Neural Information Processing Systems*, 35:26565–26577.

Kibble, W. F. (1945). An extension of a theorem of Mehler's on Hermite polynomials. *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1):12–15.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].

Li, M. and Chen, S. (2024). Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 27474–27498. PMLR. ISSN: 2640-3498.

Li, P., Li, Z., Zhang, H., and Bian, J. (2023). On the Generalization Properties of Diffusion Models. *Advances in Neural Information Processing Systems*, 36:2097–2127.

Mei, S. and Montanari, A. (2022). The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen,

M. (2022). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR. ISSN: 2640-3498.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. (2024). Memorization to Generalization: The Emergence of Diffusion Models from Associative Memory.

Pidstrigach, J. (2022). Score-Based Generative Models Detect Manifolds. In *Advances in Neural Information Processing Systems*, volume 35, pages 35852–35865. Curran Associates, Inc.

Rahimi, A. and Recht, B. (2007). Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Raya, G. and Ambrogioni, L. (2023). Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 66377–66389. Curran Associates, Inc.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, New Orleans, LA, USA. IEEE.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Ross, B. L., Kamkari, H., Wu, T., Hosseinzadeh, R., Liu, Z., Stein, G., Cresswell, J. C., and Loaiza-Ganem, G. (2024). A Geometric Framework for Understanding Memorization in Generative Models. In *The Thirteenth International Conference on Learning Representations*.

Saha, E. and Tran, G. (2025). Generalization Bound for Diffusion Models using Random Features. arXiv:2310.04417 [stat].

Sclocchi, A., Favero, A., and Wyart, M. (2025). A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1):e2408799121.

Shah, K., Chen, S., and Klivans, A. (2023). Learning Mixtures of Gaussians Using the DDPM Objective. *Advances in Neural Information Processing Systems*, 36:19636–19649.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR. ISSN: 1938-7228.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058. ISSN: 2575-7075.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2024). Understanding and mitigating copying in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 47783–47803, Red Hook, NY, USA. Curran Associates Inc.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021). Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc.

Song, Y. and Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Ukpaka, P. M. (2024). The creative agency of large language models: a philosophical inquiry. *AI and Ethics*.

Vincent, P. (2011). A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674.

Wang, Y., He, Y., and Tao, M. (2024). Evaluating the design space of diffusion-based generative models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. (2023). Diffusion Probabilistic Models Generalize when They Fail to Memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*.

Zeno, C., Manor, H., Ongie, G., Weinberger, N., Michaeli, T., and Soudry, D. (2025). When Diffusion Models Memorize: Inductive Biases in Probability Flow of Minimum-Norm Shallow Neural Nets. arXiv:2506.19031 [stat].

Zhang, H., Zhou, J., Lu, Y., Guo, M., Wang, P., Shen, L., and Qu, Q. (2024). The Emergence of Reproducibility and Consistency in Diffusion Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60558–60590. PMLR. ISSN: 2640-3498.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A SCORE MATCHING

A.1 Proof that minimizers of $\mathcal{L}_{\text{DSM}}$ and $\mathcal{L}_{\text{SM}}$ are the same

From the L_2 minimization property of the conditional expectation, the minimizer of $\mathcal{L}_{\text{DSM}}$ is given by $\hat{s}(t,x) = \mathbb{E}[\,\nabla \log P_t(x_t|x_0) \mid x_t = x\,]$. We have

$$\begin{aligned}
\mathbb{E}[\,\nabla \log P_t(x_t|x_0) \mid x_t = x\,] &= \int \mathrm{d}x_0 \frac{\nabla P_t(x_t = x|x_0)}{P_t(x_t = x|x_0)} P_t(x_0|x_t = x) \\
&= \int \mathrm{d}x_0 \nabla P_t(x_t = x|x_0) \frac{P_0(x_0)}{P_t(x)} \\
&= \frac{1}{P_t(x)} \nabla \int \mathrm{d}x_0 P_t(x_t = x|x_0) P_0(x_0) \\
&= \nabla \log P_t(x)\ .
\end{aligned}$$

Hence, we have shown that the minimizer of $\mathcal{L}_{\text{DSM}}$ is same as the minimizer of $\mathcal{L}_{\text{SM}}$.

A.2 Proof of Lemma 3.5

Let P_t^e denote the joint probability distribution of (y_t, z), where $y_t = a_t x + \sqrt{h_t} z$, with $x \sim \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $z \sim \mathcal{N}(0, I_d)$. We have

$$\begin{aligned}
\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) &= \frac{1}{dn} \sum_{i=1}^n \mathbb{E}_z \left\| \sqrt{h_t} \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} (a_t x_i + \sqrt{h_t} z) \right) + z \right\|^2 \\
&= \frac{1}{d} \mathbb{E}_{P_t^e} \left[\left\| \sqrt{h_t} \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y_t \right) + z \right\|^2 \right] \\
&\overset{(a)}{=} \frac{1}{d} \mathbb{E}_{P_t^e} \left[\left\| \sqrt{h_t} \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y_t \right) - \sqrt{h_t} s^e(t, y_t) \right\|^2 + \left\| \sqrt{h_t} s^e(t, y_t) + z \right\|^2 \right] \\
&= h_t \mathcal{M}_t + \mathcal{V}_t\ .
\end{aligned}$$

The equality (a) follows from:

$$\begin{aligned}
&\mathbb{E}_{P_t^e} \left[\left\langle \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y \right) - s^e(t, y), \sqrt{h_t} s^e(t, y) + z \right\rangle \right] \\
&\quad = \mathbb{E}_{P_t^e} \left[\mathbb{E} \left[\left\langle \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y \right) - s^e(t, y), \sqrt{h_t} s^e(t, y) + z \right\rangle \,\middle|\, y \right] \right] \\
&\quad = \mathbb{E}_{P_t^e} \left[\left\langle \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y \right) - s^e(t, y), \sqrt{h_t} s^e(t, y) + \mathbb{E}[\,z \mid y\,] \right\rangle \right] \\
&\quad = \mathbb{E}_{P_t^e} \left[\left\langle \frac{\hat{A}_t}{\sqrt{p}} \varrho\left(\frac{W_t}{\sqrt{d}} y \right) - s^e(t, y), \sqrt{h_t} s^e(t, y) - \sqrt{h_t} \nabla \log P_t^e(y) \right\rangle \right] \\
&\quad = 0\ ,
\end{aligned}$$

since $\mathbb{E}[\,z \mid y\,] = -\sqrt{h_t} \nabla \log P_t^e(y)$.

B PROOF OF THEOREM 3.2

In this section, we present a more general version of Theorem 3.2, where P_0 is supported on a D-dimensional subspace in $\mathbb{R}^d$, and $\psi_D = \frac{D}{d}$ is kept fixed as $D, d \to \infty$. First we define generalization along a subspace as follows: Let $\mathcal{M}_\parallel$ be a subspace in $\mathbb{R}^d$ and $\mathcal{M}_\perp$ its orthogonal complement. Let Π_α denote a projection matrix that project onto $\mathcal{M}_\alpha$, for $\alpha \in \{\parallel, \perp\}$. Then, we define the test error along $\mathcal{M}_\alpha$ to be

$$\mathcal{E}^m_{\text{test},\alpha}(\hat{A}_t) = \frac{1}{d}\mathbb{E}_{x\sim P_t}\left\|\Pi_\alpha \frac{\hat{A}_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}x\right) - \Pi_\alpha \nabla \log P_t(x)\right\|^2 . \tag{13}$$

Note that the overall test error is given by $\mathcal{E}^m_{\text{test}}(\hat{A}_t) = \mathcal{E}^m_{\text{test},\parallel}(\hat{A}_t) + \mathcal{E}^m_{\text{test},\perp}(\hat{A}_t)$. Lemma B.1 expresses test and train errors as traces of rational functions of random matrices. For $\kappa > 0$, we define the Gaussian expectations: $\mu_0(\kappa) = \mathbb{E}_g[\varrho(\kappa g)]$, $\mu_1(\kappa) = \mathbb{E}_g[\varrho(\kappa g)g]$, $\|\varrho(\kappa\,\cdot)\|^2 = \mathbb{E}_g\left[\varrho(\kappa g)^2\right]$, $v^2(\kappa) = \|\rho(\kappa\,\cdot)\|^2 - \mu_0^2(\kappa) - \mu_1^2(\kappa)$, $c(\gamma,\kappa) = \mathbb{E}_{u,v\sim P^\gamma}[\varrho(\kappa u)\varrho(\kappa v)]$ where $g \sim \mathcal{N}(0,1)$ and P^γ denotes the bivariate standard Gaussian distribution with correlation coefficient γ. Explicitly,

$$P^\gamma(x,y) = \frac{1}{2\pi\sqrt{1-\gamma^2}} e^{-\frac{x^2+y^2-2\gamma xy}{2(1-\gamma^2)}} . \tag{14}$$

Lemma B.1. *Let $\mathcal{M}$ be a subspace in $\mathbb{R}^D$. Let $\Pi_\parallel$ denote a projection matrix that project onto $\mathcal{M}$ and let $\Pi_\perp = I - \Pi_\parallel$. Suppose $P_0 \equiv \mathcal{N}(0,\mathcal{C})$, and let $X = [x_1, x_2, \cdots, x_n]$ be the data matrix with columns drawn i.i.d. from P_0. Consider the quantities $a_t, h_t, \lambda, \varrho$, and W_t as defined in (10), with ϱ satisfying Assumption 3.1. Define $\Sigma_t = a_t^2\mathcal{C} + h_t I_d$, $\sigma_t^2 = \frac{1}{d}\operatorname{tr}\{\Sigma_t\}$, and $\sigma_0^2 = \frac{1}{d}\operatorname{tr}\{\mathcal{C}\}$. Assume that σ_t converges to a finite limit as $d \to \infty$. Then, for the minimizer $\hat{A}_t$ of (10), for $\alpha \in \{\parallel, \perp\}$, in the limit $d, n, p \to \infty$, the test and train errors can be written as:*

$$\begin{aligned}
\mathcal{E}^\infty_{test,\alpha}(\hat{A}_t) &= E_{0,\alpha} - \frac{2\mu_1(\sigma_t)^2}{\sigma_t^2}E_{1,\alpha} + \frac{\mu_1(\sigma_t)^4}{\sigma_t^4}E_{2,\alpha} \\
&\quad + \frac{\mu_1(\sigma_t)^2 v(\sigma_t)^2}{\sigma_t^2}E_{3,\alpha} + \frac{\mu_1(\sigma_t)^2\mu_0(\sigma_t)^2}{\sigma_t^2}E_{4,\alpha} , \\
\mathcal{E}^\infty_{train}(\hat{A}_t) &= -\frac{h_t\mu_1^2(\sigma_t)}{\sigma_t^2}(E_{1,\parallel} + E_{1,\perp}) - h_t\lambda\frac{\mu_1^2(\sigma_t)}{\sigma_t^2}(E_{3,\parallel} + E_{3,\perp}) + 1 ,
\end{aligned}$$

where

$$\begin{aligned}
E_{0,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \Sigma^{-1}\right\} , \\
E_{1,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{W_t^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}\Pi_\alpha\right\} , \\
E_{2,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{W_t^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}\Sigma\frac{W_t^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}\Pi_\alpha\right\} , \\
E_{3,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{W_t^T}{\sqrt{d}}(U+\lambda I_p)^{-2}\frac{W_t}{\sqrt{d}}\Pi_\alpha\right\} , \\
E_{4,\alpha} &= \frac{1}{d}\boldsymbol{1}_p^T(U+\lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}\Pi_\alpha\frac{W_t^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\boldsymbol{1}_p ,
\end{aligned}$$

with

$$\begin{aligned}
U &= \frac{G}{\sqrt{n}}\frac{G^T}{\sqrt{n}} + \frac{h_t}{\sigma_t^2}\mu_1^2(\sigma_t)\frac{W_t}{\sqrt{d}}\frac{W_t^T}{\sqrt{d}} + s^2 I_p , \\
G &= \mu_0(\sigma_t)\boldsymbol{1}_p\boldsymbol{1}_n^T + \frac{a_t}{\sigma_t}\mu_1(\sigma_t)\frac{W_t}{\sqrt{d}}X + v_0\Omega ,
\end{aligned}$$

$$s^2 = \|\varrho(\sigma_t\,\cdot)\|^2 - c(a_t^2\sigma_0^2/\sigma_t^2, \sigma_t) - \frac{h_t}{\sigma_t^2}\mu_1^2(\sigma_t) ,$$

$$v_0^2 = c(a_t^2\sigma_0^2/\sigma_t^2, \sigma_t) - \mu_0(\sigma_t)^2 - \left(\frac{a_t\sigma_0}{\sigma_t}\mu_1(\sigma_t)\right)^2 .$$

Proof. We first find the minimizer of (10). When $m = \infty$, the denoising score matching loss is given by

$$\begin{aligned}\mathcal{L}_t^{\infty}(A_t) &= \frac{1}{dn}\sum_{i=1}^{n}\mathbb{E}_z\left[\left\|\sqrt{h_t}\frac{A_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{p}}(a_t x_i+\sqrt{h_t}z)\right)+z\right\|^2\right]+\frac{h_t\lambda}{dp}\|A_t\|_F^2\\ &= \frac{h_t}{d}\operatorname{tr}\left\{\frac{A_t}{\sqrt{p}}^T\frac{A_t}{\sqrt{p}}U\right\}+\frac{2\sqrt{h_t}}{d}\operatorname{tr}\left\{\frac{A_t}{\sqrt{p}}V\right\}+1+\frac{h_t\lambda}{d}\operatorname{tr}\left\{\frac{A_t^T}{\sqrt{p}}\frac{A_t}{\sqrt{p}}\right\},\end{aligned}$$

where

$$U=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_z\left[\varrho\left(\frac{W_t}{\sqrt{d}}(a_t x_i+\sqrt{h_t}z)\right)\varrho\left(\frac{W_t}{\sqrt{d}}(a_t x_i+\sqrt{h_t}z)\right)^T\right],$$

and

$$V=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_z\left[\varrho\left(\frac{W_t}{\sqrt{d}}(a_t x_i+\sqrt{h_t}z)\right)z^T\right].$$

Thus we get the optimal A_t as

$$\frac{\hat{A}_t}{\sqrt{p}}=-\frac{1}{\sqrt{h_t}}V^T(U+\lambda I_p)^{-1}. \tag{15}$$

We now proceed to the computation of test and train error for $P_0 \equiv \mathcal{N}(0,\mathcal{C})$. In this case, P_t is $\mathcal{N}(0,\Sigma_t)$ and $\nabla \log P_t(x) = -\Sigma_t^{-1}x$, where $\Sigma_t = a_t^2\mathcal{C} + h_t I_d$.

Test Error:

For $\alpha \in \{\|,\perp\}$, rewriting Eq. (13) for the particular case, we have:

$$\begin{aligned}\mathcal{E}_{\text{test},\alpha}^{\infty}(\hat{A}_t,\mathcal{M}) &= \frac{1}{d}\mathbb{E}_{x\sim P_t}\left[\left\|\Pi_\alpha\frac{\hat{A}_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}x\right)-\Pi_\alpha\nabla\log P_t(x)\right\|^2\right]\\ &= \frac{1}{d}\mathbb{E}_{x\sim P_t}\left[\left\|\Pi_\alpha\frac{\hat{A}_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{d}}x\right)+\Pi_\alpha\Sigma_t^{-1}x\right\|^2\right]\\ &= \frac{1}{d}\operatorname{tr}\{\Pi_\alpha\Sigma_t^{-1}\}-\frac{2}{d}\operatorname{tr}\left\{\frac{1}{\sqrt{h_t}}\Sigma_t^{-1}\Pi_\alpha V^T(U+\lambda I_p)^{-1}\underbrace{\mathbb{E}_x\left[\varrho\left(\frac{W_t}{\sqrt{d}}x\right)x^T\right]}_{:=\tilde{V}}\right\}\\ &\quad+\frac{1}{d}\operatorname{tr}\left\{\frac{1}{h_t}(U+\lambda I_p)^{-1}V\Pi_\alpha V^T(U+\lambda I_p)^{-1}\underbrace{\mathbb{E}_x\left[\varrho\left(\frac{W_t}{\sqrt{d}}x\right)\varrho\left(\frac{W_t}{\sqrt{d}}x\right)^T\right]}_{:=\tilde{U}}\right\}.\end{aligned}$$

Since we focus on a single time instant, we drop the subscript t in the above expressions. However, it is important to note that a and h depend on t and the relation $a^2+h=1$ is valid for all times. Additionally, define $\sigma_j^2 = \Sigma_{jj}$.

We need to compute $V, U, \tilde{V}, \tilde{U}$ in order to get an expression for $\mathcal{E}_{\text{test}}^{\infty}$. We will first consider $\tilde{V}$. Explicitly:

$$\tilde{V}=\mathbb{E}_x\left[\varrho\left(\frac{W}{\sqrt{d}}x\right)x^T\right].$$

Let w_i denote the i^{th} row of W and let $x(j)$ denote the j^{th} component of x. For large d, $\frac{\|w_i\|^2}{d}$ concentrates to 1. Note that, conditioned on W, $\frac{w_i^T x}{\sqrt{d}}$ is a zero-mean Gaussian random variable with variance $\mathbb{E}_x\left[\frac{(w_i^T x)^2}{d}\right] =$

$\frac{1}{d} w_i^T \Sigma w_i$. By the concentration property of quadratic forms, the variance concentrates to $\sigma^2 = \frac{1}{d}\operatorname{tr}\{\Sigma\}$ as $d \to \infty$. Thus we can write $(u_0, u_1) = \left(\frac{1}{\sigma}\frac{w_i^T x}{\sqrt{d}}, \frac{x(j)}{\sigma_j}\right) \sim P^{\gamma_{ij}}$, with $\gamma_{ij} = \frac{1}{\sigma\sigma_j}\frac{(W\Sigma)_{ij}}{\sqrt{d}}$.

$$\begin{aligned}
\tilde{V}_{ij} &= \mathbb{E}_x\left[\varrho\left(\frac{w_i^T x}{\sqrt{d}}\right) x(j)\right] \\
&= \mathbb{E}_{(u_0,u_1)\sim P^{\gamma_{ij}}}\left[\varrho\left(\sigma u_0\right)\sigma_j u_1\right] \\
&\overset{(a)}{=} \sum_{k=0}^{\infty} \frac{\gamma_{ij}^k}{k!}\mathbb{E}_{u_0}[\varrho\left(\sigma u_0\right)\mathrm{He}_k(u_0)]\sigma_j \mathbb{E}_{u_1}[u_1 \mathrm{He}_k(u_1)] \\
&= \frac{\mu_1(\sigma)}{\sigma}\frac{(W\Sigma)_{ij}}{\sqrt{d}} \ ,
\end{aligned}$$

where in (a) we used the Mehler Kernel formula Kibble (1945). Hence, $\tilde{V} = \frac{\mu_1(\sigma)}{\sigma}\frac{W}{\sqrt{d}}\Sigma$.

When considering $\tilde{U}$, a similar argument allows to write $(\tilde{u}_0, \tilde{u}_1) = (\frac{1}{\sigma}\frac{w_i^T x}{\sqrt{d}}, \frac{1}{\sigma}\frac{w_j^T x}{\sqrt{d}}) \sim P^{\tilde{\gamma}_{ij}}$ with $\tilde{\gamma}_{ij} = \frac{1}{\sigma^2}\frac{w_i^T \Sigma w_j}{\sqrt{d}}$, leading to

$$\begin{aligned}
\tilde{U}_{ij} &= \mathbb{E}_x\left[\varrho\left(\frac{w_i^T x}{\sqrt{d}}\right)\varrho\left(\frac{w_j^T x}{\sqrt{d}}\right)\right] \\
&= \mathbb{E}_{(\tilde{u}_0,\tilde{u}_1)\sim P^{\tilde{\gamma}_{ij}}}\left[\varrho(\sigma\tilde{u}_0)\varrho(\sigma\tilde{u}_1)\right] \\
&= \sum_{k=0}^{\infty}\frac{\tilde{\gamma}_{ij}^k}{k!}\,\mathbb{E}_{\tilde{u}_0}[\varrho(\sigma\tilde{u}_0)\mathrm{He}_k(u_0)]\mathbb{E}_{u_1}[\varrho(\sigma\tilde{u}_1)\mathrm{He}_k(\tilde{u}_1)] \\
&= \sum_{k=0}^{\infty}\frac{\tilde{\gamma}_{ij}^k}{k!}\,\mathbb{E}_{\tilde{u}_0}[\varrho(\sigma\tilde{u}_0)\mathrm{He}_k(\tilde{u}_0)]^2 \ .
\end{aligned}$$

In short:

$$\tilde{U}_{ij} = \begin{cases} \mu_0^2(\sigma) + \frac{\mu_1^2(\sigma)}{\sigma^2}\frac{w_i^T\Sigma w_j}{d} + O(1/d) & \text{if } i \neq j \ , \\ \|\varrho(\sigma\,\cdot)\|^2 & \text{if } i = j \ . \end{cases}$$

The $O(1/d)$ contribution cannot give rise to a $O(1)$ change in the asymptotic spectrum. Hence, we neglect it. We have

$$\tilde{U} = \mu_0^2(\sigma)\mathbf{1}_p\mathbf{1}_p^T + \frac{\mu_1^2(\sigma)}{\sigma^2}\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}} + v^2(\sigma)I_p \ ,$$

where $v^2(\kappa) = \|\rho(\kappa\,\cdot)\|^2 - \mu_0^2(\kappa) - \mu_1^2(\kappa)$. We now proceed to V. For the l^{th} data sample, we consider the matrix:

$$V^l = \mathbb{E}_z\left[\varrho\left(\frac{W}{\sqrt{d}}(ax_l + \sqrt{h}z)\right)z^T\right] \ ,$$

and write for each of its components:

$$\begin{aligned}
V_{ij}^l &= \mathbb{E}_z\left[\varrho\left(\frac{w_i^T(ax_l + \sqrt{h}z)}{\sqrt{d}}\right)z_j\right] \\
&= \mathbb{E}_{(u,v)\sim P^{\frac{w_{ij}}{\sqrt{d}}}}\left[\varrho\left(\frac{aw_i^T x_l}{\sqrt{d}} + \sqrt{h}u\right)v\right] \\
&= \sum_{k=0}^{\infty}\frac{(\frac{w_{ij}}{\sqrt{d}})^k}{k!}\mathbb{E}_u\left[\varrho\left(\frac{aw_i^T x_l}{\sqrt{d}} + \sqrt{h}u\right)\mathrm{He}_k(u)\right]\mathbb{E}_v[v\mathrm{He}_k(v)] \\
&= \frac{w_{ij}}{\sqrt{d}}\mathbb{E}_u\left[\varrho\left(\frac{aw_i^T x_l}{\sqrt{d}} + \sqrt{h}u\right)u\right] \\
&= \frac{w_{ij}}{\sqrt{d}}\varrho_1\left(\frac{aw_i^T x_l}{\sqrt{d}}\right) \ ,
\end{aligned}$$

where $\varrho_1(y) = \mathbb{E}_u\Big[\varrho(y+\sqrt{h}u)u\Big]$. Summing over the n data samples:

$$\begin{aligned}
V_{ij} &= \frac{1}{n}\sum_{l=1}^{n} V_{ij}^l \\
&= \frac{w_{ij}}{\sqrt{d}}\frac{1}{n}\sum_{l=1}^{n}\varrho_1\left(\frac{aw_i^T x_l}{\sqrt{d}}\right) \\
&= \frac{w_{ij}}{\sqrt{d}}\mathbb{E}_x\left[\varrho_1\left(\frac{aw_i^T x}{\sqrt{d}}\right)\right] + O(1/d) \\
&= \frac{w_{ij}}{\sqrt{d}}\mathbb{E}_g[\varrho_1(a\sigma_0 g)] + O(1/d) \\
&= \frac{w_{ij}}{\sqrt{d}}\mathbb{E}_{g,u}\Big[\varrho(a\sigma_0 g+\sqrt{h}u)u\Big] + O(1/d) \\
&= \frac{\sqrt{h}}{\sigma}\mu_1(\sigma)\frac{w_{ij}}{\sqrt{d}} + O(1/d)\ ,
\end{aligned}$$

where we recall that $\sigma_0^2 = a^2\frac{1}{d}\operatorname{tr}\{\mathcal{C}\} + h$. Neglecting $O(1/d)$ terms, we have $V = \frac{\sqrt{h}}{\sigma}\mu_1(\sigma)\frac{W}{\sqrt{d}}$.

Going forward to U, we again consider the matrix relative to the l^{th} data sample:

$$U^l = \mathbb{E}_z\left[\varrho\left(\frac{W}{\sqrt{d}}(ax_l+\sqrt{h}z)\right)\varrho\left(\frac{W}{\sqrt{d}}(ax_l+\sqrt{h}z)\right)^T\right].$$

For the components with $i \neq j$, we have:

$$\begin{aligned}
U_{ij}^l &= \mathbb{E}_z\left[\varrho\left(\frac{w_i^T(ax_l+\sqrt{h}z)}{\sqrt{d}}\right)\varrho\left(\frac{w_j^T(ax_l+\sqrt{h}z)}{\sqrt{d}}\right)\right] \\
&= \mathbb{E}_{(u,v)\sim P^{\frac{w_i^T w_j}{d}}}\left[\varrho\left(a\frac{w_i^T x_l}{\sqrt{d}}+\sqrt{h}u\right)\varrho\left(a\frac{w_j^T x_l}{\sqrt{d}}+\sqrt{h}v\right)\right] \\
&= \sum_{k=0}^{\infty}\frac{(\frac{w_i^T w_j}{d})^k}{k!}\mathbb{E}_u\left[\varrho\left(a\frac{w_i^T x_l}{\sqrt{d}}+\sqrt{h}u\right)\mathrm{He}_k(u)\right]\mathbb{E}_v\left[\varrho\left(a\frac{w_j^T x_l}{\sqrt{d}}+\sqrt{h}v\right)\mathrm{He}_k(v)\right] \\
&= \varrho_0\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_0\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + \frac{w_i^T w_j}{d}\varrho_1\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_1\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + O(1/d)\ ,
\end{aligned}$$

where $\varrho_0(y) = \mathbb{E}_u\Big[\varrho(y+\sqrt{h}u)\Big]$ and $\varrho_1(y) = \mathbb{E}_u\Big[\varrho(y+\sqrt{h}u)u\Big]$. Summing over the n data samples:

$$\begin{aligned}
U_{ij} &= \frac{1}{n}\sum_{l=1}^{n} U_{ij}^l \\
&= \frac{1}{n}\sum_{l=1}^{n}\varrho_0\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_0\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + \frac{w_i^T w_j}{d}\frac{1}{n}\sum_{l=1}^{n}\varrho_1\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_1\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + O(1/d) \\
&= \frac{1}{n}\sum_{l=1}^{n}\varrho_0\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_0\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + \frac{w_i^T w_j}{d}\mathbb{E}_x\left[\varrho_1\left(a\frac{w_i^T x}{\sqrt{d}}\right)\varrho_1\left(a\frac{w_j^T x}{\sqrt{d}}\right)\right] + O(1/d) \\
&= \frac{1}{n}\sum_{l=1}^{n}\varrho_0\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_0\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + \frac{w_i^T w_j}{d}\mathbb{E}_g[\varrho_1(a\sigma_0 g)]^2 + O(1/d) \\
&= \frac{1}{n}\sum_{l=1}^{n}\varrho_0\left(a\frac{w_i^T x_l}{\sqrt{d}}\right)\varrho_0\left(a\frac{w_j^T x_l}{\sqrt{d}}\right) + \frac{h}{\sigma^2}\mu_1^2(\sigma)\frac{w_i^T w_j}{d} + O(1/d)\ .
\end{aligned}$$

For $i = j$, we have:

$$U_{ii}^l = \mathbb{E}_z \left[\left(\varrho \left(\frac{w_i^T (a x_l + \sqrt{h} z)}{\sqrt{d}} \right) \right)^2 \right] ,$$

and

$$\begin{aligned} U_{ii} &= \frac{1}{n} \sum_{l=1}^{n} \mathbb{E}_z \left[\left(\varrho \left(\frac{w_i^T (a x_l + \sqrt{h} z)}{\sqrt{d}} \right) \right)^2 \right] \\ &= \mathbb{E}_{z,x} \left[\left(\varrho \left(\frac{w_i^T (a x + \sqrt{h} z)}{\sqrt{d}} \right) \right)^2 \right] + O(1/\sqrt{d}) \\ &= \| \varrho(\sigma \, \cdot) \|^2 + O(1/\sqrt{d}) \, . \end{aligned}$$

The $O\Big(1/\sqrt{d}\Big)$ term in the above equation can be neglected, since there are only $O(d)$ terms on the diagonal. Let $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d \times n}$. We can write U as:

$$U = \frac{\varrho_0 \left(a \frac{W}{\sqrt{d}} X \right)}{\sqrt{n}} \frac{\varrho_0 \left(a \frac{W}{\sqrt{d}} X \right)^T}{\sqrt{n}} + \frac{h}{\sigma^2} \mu_1^2(\sigma) \frac{W}{\sqrt{d}} \frac{W^T}{\sqrt{d}} + s^2 I_p \, ,$$

where

$$\begin{aligned} s^2 &= \| \varrho(\sigma \, \cdot) \|^2 - \mathbb{E}_g \big[\varrho_0(a \sigma_0 g)^2 \big] - \frac{h}{\sigma^2} \mu_1^2(\sigma) \\ &= \| \varrho(\sigma \, \cdot) \|^2 - \mathbb{E}_g \Big[\mathbb{E}_u \Big[\varrho(a \sigma_0 g + \sqrt{h} u) \Big]^2 \Big] - \frac{h}{\sigma^2} \mu_1^2(\sigma) \\ &= \| \varrho(\sigma \, \cdot) \|^2 - c(a^2 \sigma_0^2 / \sigma^2, \sigma) - \frac{h}{\sigma^2} \mu_1^2(\sigma) \, , \end{aligned}$$

with $c(\gamma, \kappa) = \mathbb{E}_{u,v \sim P^\gamma} [\varrho(\kappa u) \varrho(\kappa v)]$.

We can use the Gaussian equivalence principle to replace the nonlinear term in U. A Gaussian equivalent for $\varrho_0(a \frac{W}{\sqrt{d}} X)$ is given by

$$\begin{aligned} G &= \mathbb{E}_g[\varrho_0(a \sigma_0 g)] \mathbf{1}_p \mathbf{1}_n^T + \mathbb{E}_g[\varrho_0(a \sigma_0 g) g] \frac{W}{\sqrt{d}} \frac{X}{\sigma_0} \\ &\quad + \Big(\mathbb{E}_g \big[\varrho_0(a \sigma_0 g)^2 \big] - \mathbb{E}_g[\varrho_0(a \sigma_0 g)]^2 - \mathbb{E}_g[\varrho_0(a \sigma_0 g) g]^2 \Big)^{1/2} \Omega \\ &= \mu_0(\sigma) \mathbf{1}_p \mathbf{1}_n^T + a \frac{\mu_1(\sigma)}{\sigma} \frac{W}{\sqrt{d}} X + \left(\underbrace{c(a^2 \sigma_0^2 / \sigma^2, \sigma) - \mu_0(\sigma)^2 - \Big(\frac{a \sigma_0}{\sigma} \mu_1(\sigma) \Big)^2}_{:= v_0^2} \right)^{1/2} \Omega \, , \end{aligned}$$

where $\Omega \in \mathbb{R}^{p \times n}$ is a random matrix with standard Gaussian entries. Hence we have

$$U = \frac{G}{\sqrt{n}} \frac{G^T}{\sqrt{n}} + \frac{h}{\sigma^2} \mu_1^2(\sigma) \frac{W}{\sqrt{d}} \frac{W^T}{\sqrt{d}} + s^2 I_p \, ,$$

with

$$G = \mu_0(\sigma) \mathbf{1}_p \mathbf{1}_n^T + a \frac{\mu_1(\sigma)}{\sigma} \frac{W}{\sqrt{d}} X + v_0 \Omega \, .$$

We now have expressions for all terms terms contributing to the test error:

$$\begin{aligned}\mathcal{E}^{\infty}_{\text{test},\alpha}(\hat{A}_t) &= \frac{1}{d}\operatorname{tr}\{\Pi_\alpha\Sigma^{-1}\} - \frac{2}{d}\operatorname{tr}\left\{\frac{1}{\sqrt{h}}\Sigma^{-1}\Pi_\alpha V^T(U+\lambda I_p)^{-1}\tilde{V}\right\} \\ &\quad + \frac{1}{d}\operatorname{tr}\left\{\frac{1}{h_t}(U+\lambda I_p)^{-1}V\Pi_\alpha V^T(U+\lambda I_p)^{-1}\tilde{U}\right\} \\ &= \frac{1}{d}\operatorname{tr}\{\Pi_\alpha\Sigma^{-1}\} - \frac{2\mu_1^2(\sigma)}{\sigma^2}\frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\right\} \\ &\quad + \frac{\mu_1^2(\sigma)}{\sigma^2}\frac{1}{d}\operatorname{tr}\Bigg\{(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1} \\ &\qquad \left(\mu_0(\sigma)^2\mathbf{1}_p\mathbf{1}_p^T + \frac{\mu_1(\sigma)^2}{\sigma^2}\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}} + v(\sigma)^2 I_p\right)\Bigg\}\,.\end{aligned}$$

We can then write

$$\mathcal{E}^{\infty}_{\text{test},\alpha}(\hat{A}_t) = E_{0,\alpha} - \frac{2\mu_1^2(\sigma)}{\sigma^2}E_{1,\alpha} + \frac{\mu_1^4(\sigma)}{\sigma^4}E_{2,\alpha} + \frac{\mu_1^2(\sigma)v(\sigma)^2}{\sigma^2}E_{3,\alpha} + \frac{\mu_1(\sigma)^2\mu_0(\sigma)^2}{\sigma^2}E_{4,\alpha}\quad,$$

with

$$\begin{aligned}E_{0,\alpha} &= \frac{1}{d}\operatorname{tr}\{\Pi_\alpha\Sigma^{-1}\}\,,\\ E_{1,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Pi_\alpha\right\}\,,\\ E_{2,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Pi_\alpha\right\}\,,\\ E_{3,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-2}\frac{W}{\sqrt{d}}\Pi_\alpha\right\}\,,\\ E_{4,\alpha} &= \frac{1}{d}\mathbf{1}_p^T(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Pi_\alpha\frac{W^T}{\sqrt{d}}(U+\lambda I_p)^{-1}\mathbf{1}_p\,,\end{aligned}$$

Next, we look at the train error.

Train error:

$$\begin{aligned}\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) &= \mathcal{L}(\hat{A}_t) - \frac{h_t\lambda}{pd}\left\|\hat{A}_t\right\|_F^2 \\ &= \frac{h}{d}\operatorname{tr}\left\{\frac{\hat{A}_t^T}{\sqrt{p}}\frac{\hat{A}_t}{\sqrt{p}}(U+\lambda I_p)\right\} + \frac{2\sqrt{h}}{d}\operatorname{tr}\left\{\frac{\hat{A}_t}{\sqrt{p}}V\right\} + 1 - \frac{h\lambda}{d}\operatorname{tr}\left\{\frac{\hat{A}_t^T}{\sqrt{p}}\frac{\hat{A}_t}{\sqrt{p}}\right\} \\ &= \frac{1}{d}\operatorname{tr}\{(U+\lambda I_p)^{-1}VV^T\} - \frac{2}{d}\operatorname{tr}\{V^T(U+\lambda I_p)^{-1}V\} + 1 - \frac{\lambda}{d}\operatorname{tr}\{V^T(U+\lambda I_p)^{-2}V\} \\ &= -\frac{h\mu_1^2(\sigma)}{d\sigma^2}\operatorname{tr}\left\{(U+\lambda I_p)^{-1}\frac{W}{\sqrt{d}}\frac{W^T}{\sqrt{d}}\right\} + 1 - \frac{h\mu_1^2(\sigma)\lambda}{d\sigma^2}\operatorname{tr}\left\{(U+\lambda I_p)^{-2}\frac{W}{\sqrt{d}}\frac{W^T}{\sqrt{d}}\right\} \\ &= -\frac{h\mu_1^2(\sigma)}{\sigma^2}(E_{1,\parallel}+E_{1,\perp}) - h\lambda\frac{\mu_1^2(\sigma)}{\sigma^2}(E_{3,\parallel}+E_{3,\perp}) + 1\,.\end{aligned}$$

This completes the proof of lemma B.1. □

We have the following theorem characterizing test and train errors in the $m=\infty$ case:

Theorem B.2. *Let $\mathcal{M}_\parallel$ be a D-dimensional subspace in $\mathbb{R}^D$, $\Pi_\parallel$ a projection matrix onto it, and $P_0 \equiv \mathcal{N}\left(0,\Pi_\parallel\right)$ with ϱ satisfying Assumption 3.1. Define $\sigma_0^2 = \frac{1}{d}\operatorname{tr}\{\Pi_\parallel\}$, $\sigma_t^2 = a_t^2\sigma_0^2 + h_t$, and $\mu_{1,t} = \mu_1(\sigma_t)/\sigma_t$. Set $s^2 =$*

$\|\varrho(\sigma_t\cdot)\|^2 - c(a_t^2\sigma_0^2/\sigma_t^2, \sigma_t) - h_t\mu_{1,t}^2$, $v_0^2 = c(a_t^2\sigma_0^2/\sigma_t^2, \sigma) - a_t^2\sigma_0^2\mu_{1,t}^2$ *and* $v^2 = \|\varrho(\sigma_t\cdot)\|^2 - \mu_1(\sigma_t)^2$. *Let* $\psi_D = \frac{D}{d}, \psi_n = \frac{n}{d}$, $\psi_p = \frac{p}{d}$ *and* $\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5$ *be the solution of the following system of algebraic equations in* q *and* z*:*

$$\begin{aligned}
\zeta_1(s^2 - z + (1-\psi_D)h_t(\mu_{1,t}^2 + q)\zeta_2 + \psi_D(h_t\mu_{1,t}^2 + q)\zeta_{33} + \psi_D a_t^2\mu_{1,t}^2\zeta_3\zeta_4 + v_0^2\zeta_4) - 1 = 0\,,\\
\zeta_2(1 + \psi_p h_t(\mu_{1,t}^2 + q)\zeta_1) - 1 = 0\,,\\
\zeta_3(1 + \psi_p(h_t\mu_{1,t}^2 + q)\zeta_1) + \psi_p a_t^2\mu_{1,t}^2\zeta_1\zeta_3\zeta_4 - 1 = 0\,,\\
\zeta_5(1 + \psi_p(h_t\mu_{1,t}^2 + q)\zeta_1) + (1 + a\mu_{1,t}\zeta_4\zeta_5)\psi_p a_t\mu_{1,t}\zeta_1 = 0\,,\\
\zeta_4\left(1 + \frac{\psi_p}{\psi_n}v_0^2\zeta_1 - \frac{\psi_D}{\psi_n}a_t\mu_{1,t}\zeta_5\right) - 1 = 0\,.
\end{aligned}$$

Let $e_{0,\|} = \psi_D$ *and* $e_{0,\perp} = \frac{1-\psi_D}{h_t}$. *Define functions* $\mathcal{K}_\|(q,z) = -\frac{\psi_D\zeta_5(q,z)}{a_t\mu_{1,t}}$ *and* $\mathcal{K}_\perp(q,z) = \frac{(1-\psi_D)(1-\zeta_2(q,z))}{h_t(\mu_{1,t}^2+q)}$. *For* $\alpha \in \{\|, \perp\}$, *let* $\varepsilon_{test,\alpha}^\infty = e_{0,\alpha} - 2\mu_{1,t}^2\mathcal{K}_\alpha(0,-\lambda) - \mu_{1,t}^4\frac{\partial\mathcal{K}_\alpha}{\partial q}(0,-\lambda) + \mu_{1,t}^2v^2\frac{\partial\mathcal{K}_\alpha}{\partial z}(0,-\lambda)$, *and* $\varepsilon_{train}^\infty = 1 - \mu_{1,t}^2h_t(\mathcal{K}_\|(0,-\lambda) + \mathcal{K}_\perp(0,-\lambda)) - \mu_{1,t}^2\lambda h_t(\frac{\partial\mathcal{K}_\|}{\partial z}(0,-\lambda) + \frac{\partial\mathcal{K}_\perp}{\partial z}(0,-\lambda))$. *Then, for the minimizer of (10)* $\hat{A}_t$, *as* $d,n,p \to \infty$*:*

$$\lim_{d,n,p\to\infty}\mathbb{E}\left[\mathcal{E}_{test,\|}^\infty(\hat{A}_t)\right] = \varepsilon_{test,\|}^\infty\,,\quad \lim_{d,n,p\to\infty}\mathbb{E}\left[\mathcal{E}_{test,\perp}^\infty(\hat{A}_t)\right] = \varepsilon_{test,\perp}^\infty\,,\quad \lim_{d,n,p\to\infty}\mathbb{E}\left[\mathcal{E}_{train}^\infty(\hat{A}_t)\right] = \varepsilon_{train}^\infty\,.$$

Remark B.3. Theorem 3.2 can be recovered by substituting $\psi_D = 1$.

Proof. First, for $P_0 \equiv \mathcal{N}(0, \mathcal{C})$, we use the following Lemma B.1 to express the test and train errors for $m = \infty$ as sums of traces of rational functions of random matrices. This will facilitate the use of linear pencils to compute the asymptotic test and train errors.

To prove Theorem 3.2, we begin by noting that, that without loss of generality, we may assume that $\mathcal{C} = \Pi_\|$ is a diagonal matrix with its first D diagonal entries equal to 1 and the remaining entries equal to 0. That is,

$$\mathcal{C} = \Pi_\| = \mathrm{diag}([\underbrace{1, 1, \cdots, 1}_{D}, 0, 0\cdots, 0])\,. \tag{16}$$

This simplification follows from the rotational invariance of the problem. We now invoke Lemma B.1 for this value of $\mathcal{C}$. As in the proof of the Lemma, since we focus on a single time instant, we drop the subscript t in the above expressions. For the $\mathcal{C}$ used here, note that $\sigma^2 = a^2\psi_D + h, \sigma_0^2 = \psi_D$. Also, let $\mu_0 = \mu_0(\sigma), \mu_1 = \mu_1(\sigma)/\sigma, \|\varrho(\sigma\,\cdot)\| = \|\varrho\|$.

Let the first D columns of W be denoted by $W_\|$ and $W_\perp$ the remaining columns, i.e., $W = [W_\|, W_\perp]$. For simplicity in presentation, we assume $\mu_0 = 0$ and write:

$$\begin{aligned}
\mathcal{E}_{\mathrm{test},\alpha}^\infty(\hat{A}_t) &= E_{0,\alpha} - 2\mu_1^2E_{1,\alpha} + \mu_1^4E_{2,\alpha} + \mu_1^2v^2E_{3,\alpha}\,,\\
\mathcal{E}_{\mathrm{train}}^\infty(\hat{A}_t) &= -h\mu_1^2(E_{1,\|} + E_{1,\perp}) - h\lambda\mu_1^2(E_{3,\|} + E_{3,\perp}) + 1\,,
\end{aligned}$$

with

$$\begin{aligned}
E_{0,\|} &= \psi_D, \quad E_{0,\perp} = \frac{1-\psi_D}{h}\,,\\
E_{1,\alpha} &= \frac{1}{d}\,\mathrm{tr}\left\{\frac{W_\alpha^T}{\sqrt{d}}(U + \lambda I_p)^{-1}\frac{W_\alpha}{\sqrt{d}}\right\}\,,\\
E_{2,\alpha} &= \frac{1}{d}\,\mathrm{tr}\left\{\frac{W_\alpha^T}{\sqrt{d}}(U + \lambda I_p)^{-1}\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}}(U + \lambda I_p)^{-1}\frac{W_\alpha}{\sqrt{d}}\right\}\,,\\
E_{3,\alpha} &= \frac{1}{d}\,\mathrm{tr}\left\{\frac{W_\alpha^T}{\sqrt{d}}(U + \lambda I_p)^{-2}\frac{W_\alpha}{\sqrt{d}}\right\}\,,
\end{aligned}$$

for $\alpha \in \{\|, \perp\}$. We now define the matrix:

$$\begin{aligned}
U(q) &:= \frac{G}{\sqrt{n}}\frac{G^T}{\sqrt{n}} + h\mu_1^2\frac{W}{\sqrt{d}}\frac{W^T}{\sqrt{d}} + q\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}} + s^2I_p\\
&= \frac{G}{\sqrt{n}}\frac{G^T}{\sqrt{n}} + (h\mu_1^2 + q)\frac{W_\|}{\sqrt{d}}\frac{W_\|^T}{\sqrt{d}} + (h\mu_1^2 + hq)\frac{W_\perp}{\sqrt{d}}\frac{W_\perp^T}{\sqrt{d}} + s^2I_p\,,
\end{aligned}$$

and consider its resolvent:

$$R(q,z) = (U(q) - zI_p)^{-1} .$$

Note that for our special choice of $\mathcal{C}$, the matrix G takes the form:

$$G = a\mu_1 \frac{W}{\sqrt{d}} X + v_0 \Omega = a\mu_1 \frac{W_\parallel}{\sqrt{d}} X_\parallel + v_0 \Omega ,$$

where $X_\parallel$ denotes the first D rows of X. Define

$$K_\alpha(q,z) = \frac{1}{d} \operatorname{tr}\Big\{ \frac{W_\alpha^T}{\sqrt{d}} R(q,z) \frac{W_\alpha}{\sqrt{d}} \Big\} .$$

Using the matrix derivative identities $\frac{\partial R}{\partial q} = -R(q,z)\frac{\mathrm{d}U}{\mathrm{d}q}R(q,z)$ and $\frac{\partial R}{\partial z} = R(q,z)^2$, we observe that

$$\begin{aligned} E_{1,\alpha} &= K_\alpha(0,-\lambda) , \\ E_{2,\alpha} &= -\frac{\partial K_\alpha}{\partial q}(0,-\lambda) , \\ E_{3,\alpha} &= \frac{\partial K_\alpha}{\partial z}(0,-\lambda) . \end{aligned}$$

Therefore, assuming we can take limit of the expectation inside the derivative, it suffices to consider the function $\mathcal{K}_\alpha(q,z) := \lim_{d\to\infty} \mathbb{E}[K_\alpha(q,z)]$ to have an expression for $\lim_{d\to\infty} \mathbb{E}\Big[\mathcal{E}^\infty_{\mathrm{test},\alpha}(\hat{A}_t)\Big]$. We then have

$$\begin{aligned} \lim_{d\to\infty} \mathbb{E}\Big[\mathcal{E}^\infty_{\mathrm{test},\alpha}(\hat{A}_t)\Big] &= E_{0,\alpha} - 2\mu_1^2 \lim_{d\to\infty} \mathbb{E}[E_{1,\alpha}] + \mu_1^4 \lim_{d\to\infty} \mathbb{E}[E_{2,\alpha}] + \mu_1^2 v^2 \lim_{d\to\infty} \mathbb{E}[E_{3,\alpha}] \\ &= E_{0,\alpha} - 2\mu_1^2 \mathcal{K}_\alpha(0,-\lambda) + \mu_1^4 \frac{\partial \mathcal{K}_\alpha}{\partial q}(0,-\lambda) + \mu_1^2 v^2 \frac{\partial \mathcal{K}_\alpha}{\partial z}(0,-\lambda) \\ &= E_{0,\alpha} - 2\mu_1^2 e_{1,\alpha} + \mu_1^4 e_{2,\alpha} + \mu_1^2 v^2 e_{3,\alpha} , \end{aligned} \tag{17}$$

where $e_{1,\alpha} = \mathcal{K}_\alpha(0,-\lambda)$, $e_{2,\alpha} = -\frac{\partial \mathcal{K}_\alpha}{\partial q}(0,-\lambda)$, $e_{3,\alpha} = \frac{\partial \mathcal{K}_\alpha}{\partial z}(0,-\lambda)$. The above computations can also be carried out without interchanging derivatives and expectations. This can be done by first differentiating $K_\alpha(q,z)$ (instead of $\mathcal{K}_\alpha(q,z)$), and then constructing a linear pencil matrix that is nearly twice the size of the one used here. Since this approach is more tedious, we opt to work with the smaller linear pencil below.

We now derive an expression for $\mathcal{K}_\alpha$ using the Linear Pencils method. To begin, we construct the following 5×5 block matrix:

$$L = \left[\begin{array}{c|cccc} (s^2 - z)I_p & (h\mu_1^2 + hq)\frac{W_2}{\sqrt{d}} & (h\mu_1^2 + q)\frac{W_1}{\sqrt{d}} & v_0\frac{\Omega}{\sqrt{n}} & a\mu_1\frac{W_1}{\sqrt{d}} \\ \hline -\frac{W_2^T}{\sqrt{d}} & I_{d-D} & 0 & 0 & 0 \\ -\frac{W_1^T}{\sqrt{d}} & 0 & I_D & 0 & 0 \\ -v_0\frac{\Omega^T}{\sqrt{n}} & 0 & -a\mu_1\frac{X_1^T}{\sqrt{n}} & I_n & 0 \\ 0 & 0 & 0 & -\frac{X_1}{\sqrt{d}} & I_D \end{array}\right] = \left[\begin{array}{cc} L_{11} & L_{12} \\ L_{21} & L_{22} \end{array}\right] .$$

First, we can invert L and verify that K_α can be obtained from the trace of one of the blocks in L^{-1}. In particular, we observe that $(L^{-1})^{2,2} = I - (h\mu_1^2 + hq)\frac{W_\perp^T}{\sqrt{d}} R(q,z) \frac{W_\perp}{\sqrt{d}}$ and $(L^{-1})^{3,5} = -a\mu_1 \frac{W_\parallel^T}{\sqrt{d}} R(q,z) \frac{W_\parallel}{\sqrt{d}}$, which yield the desired terms. We use the linear pencil formalism to derive the traces of the square blocks in L^{-1}. Let g be the matrix of traces of square blocks in L^{-1} divided by the block size. For example, if $L^{i,j}$ is a square matrix of dimension N, then $g_{ij} = \frac{1}{N}\operatorname{tr}\big\{(L^{-1})^{i,j}\big\}$. If $L^{i,j}$ is not a square matrix, we set $g_{ij} = 0$. In our setting, this gives:

$$g = \begin{bmatrix} g_{11} & 0 & 0 & 0 & 0 \\ 0 & g_{22} & 0 & 0 & 0 \\ 0 & 0 & g_{33} & 0 & g_{35} \\ 0 & 0 & 0 & g_{44} & 0 \\ 0 & 0 & g_{53} & 0 & g_{55} \end{bmatrix} .$$

Notice that all constant matrices appearing in L are multiples of identity. We can therefore encode their coefficients in a matrix B, given by

$$B = \begin{bmatrix} s^2 - z & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

If L^{il} and L^{jk} are square matrices, let σ_{ij}^{kl} denote the covariance between an element of L^{ij} and an element of L^{kl} multiplied by the block size of L^{jk}. Let L^{ij} be of dimension $N_i \times N_j$ and M denote the non-constant part of L. Then,

$$\sigma_{ij}^{kl} = N_j \mathbb{E}\left[M_{uv}^{ij} M_{vu}^{kl}\right] .$$

Let $S = \{(i,j) : N_i = N_j\}$ be the set of indices corresponding to square blocks in L. Then, a mapping η_L is defined such that

$$\eta_L(G)_{il} = \sum_{(jk)\in S} \sigma_{ij}^{kl} g_{jk} ,$$

for (il) in S. In our case, this yields:

$$\eta_L(g) = \begin{bmatrix} \sigma_{12}^{21} g_{22} + \sigma_{13}^{31} g_{33} + \sigma_{15}^{31} g_{53} + \sigma_{14}^{41} g_{44} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{12}^{21} g_{11} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{31}^{13} g_{11} & 0 & \sigma_{31}^{15} g_{11} \\ 0 & 0 & 0 & \sigma_{41}^{14} g_{11} + \sigma_{43}^{54} g_{35} & 0 \\ 0 & 0 & \sigma_{54}^{43} g_{44} & 0 & 0 \end{bmatrix} .$$

The block dimensions are $N_1 = p, N_2 = d - D, N_3 = D, N_4 = n, N_5 = D$, and the non-zero covariances are

$$\sigma_{12}^{21} = -(1-\psi_D)(h\mu_1^2 + hq) , \qquad \sigma_{21}^{12} = -\psi_p(h\mu_1^2 + hq) , \tag{18}$$

$$\sigma_{13}^{31} = -\psi_D(h\mu_1^2 + q) , \qquad \sigma_{31}^{13} = -\psi_p(h\mu_1^2 + q) , \tag{19}$$

$$\sigma_{14}^{41} = -v_0^2 , \qquad \sigma_{41}^{14} = -\frac{\psi_p}{\psi_n} v_0^2 , \tag{20}$$

$$\sigma_{15}^{31} = -\psi_D a\mu_1 , \qquad \sigma_{31}^{15} = -\psi_p a\mu_1 , \tag{21}$$

$$\sigma_{43}^{54} = \frac{\psi_D}{\psi_n} a\mu_1 , \qquad \sigma_{54}^{43} = a\mu_1 . \tag{22}$$

Finally, the matrix g satisfies the fixed point equation

$$(B - \eta_L(g))g = I ,$$

which explicitly reads:

$$\begin{bmatrix} s^2 - z + (1-\psi_D)(h\mu_1^2 + hq)g_{22} + \psi_D(h\mu_1^2 + q)g_{33} + \psi_D a\mu_1 g_{53} + v_0^2 g_{44} & 0 & 0 & 0 & 0 \\ 0 & 1 + \psi_p(h\mu_1^2 + hq)g_{11} & 0 & 0 & 0 \\ 0 & 0 & 1 + \psi_p(h\mu_1^2 + q)g_{11} & 0 & \psi_p a\mu_1 g_{11} \\ 0 & 0 & 0 & 1 + \frac{\psi_p}{\psi_n} v_0^2 g_{11} - \frac{\psi_D}{\psi_n} a\mu_1 g_{35} & 0 \\ 0 & 0 & -a\mu_1 g_{44} & 0 & 1 \end{bmatrix} \times \begin{bmatrix} g_{11} & 0 & 0 & 0 & 0 \\ 0 & g_{22} & 0 & 0 & 0 \\ 0 & 0 & g_{33} & 0 & g_{35} \\ 0 & 0 & 0 & g_{44} & 0 \\ 0 & 0 & g_{53} & 0 & g_{55} \end{bmatrix} = I .$$

This leads to the following set of equations:

$$\begin{aligned}
g_{11}(s^2 - z + (1-\psi_D)(h\mu_1^2 + hq)g_{22} + \psi_D(h\mu_1^2+q)g_{33} + \psi_D a\mu_1 g_{53} + v_0^2 g_{44}) &= 1 \,, \\
g_{22}(1 + \psi_p(h\mu_1^2 + hq)g_{11}) &= 1 \,, \\
g_{33}(1 + \psi_p(h\mu_1^2 + q)g_{11}) + g_{53}\psi_p a\mu_1 g_{11} &= 1 \,, \\
g_{35}(1 + \psi_p(h\mu_1^2 + q)g_{11}) + g_{55}\psi_p a\mu_1 g_{11} &= 0 \,, \\
g_{44}(1 + \frac{\psi_p}{\psi_n}v_0^2 g_{11} - \frac{\psi_D}{\psi_n}a\mu_1 g_{35}) &= 1 \,, \\
-g_{33}a\mu_1 g_{44} + g_{53} &= 0 \,, \\
-g_{35}a\mu_1 g_{44} + g_{55} &= 1 \,.
\end{aligned}$$

This system of seven equations can be reduced to the following five equations:

$$\begin{aligned}
\zeta_1(s^2 - z + (1-\psi_D)h(\mu_1^2+q)\zeta_5 + \psi_D(h\mu_1^2+q)\zeta_4 + \psi_D a^2\mu_1^2\zeta_4\zeta_2 + v_0^2\zeta_2) - 1 &= 0 \,, \\
\zeta_4(1 + \psi_p(h\mu_1^2+q)\zeta_1) + \psi_p a^2\mu_1^2\zeta_1\zeta_4\zeta_2 - 1 &= 0 \,, \\
\zeta_3(1 + \psi_p(h\mu_1^2+q)\zeta_1) + (1 + a\mu_1\zeta_2\zeta_3)\psi_p a\mu_1\zeta_1 &= 0 \,, \\
\zeta_2(1 + \frac{\psi_p}{\psi_n}v_0^2\zeta_1 - \frac{\psi_D}{\psi_n}a\mu_1\zeta_3) - 1 &= 0 \,, \\
\zeta_5(1 + \psi_p h(\mu_1^2+q)\zeta_1) - 1 &= 0 \,,
\end{aligned}$$

where $\zeta_1 = g_{11}$, $\zeta_2 = g_{44}$, $\zeta_3 = g_{35}$, $\zeta_4 = g_{33}$, $\zeta_5 = g_{22}$. We solve this system numerically to obtain $\zeta_1, \zeta_2, \zeta_3, \zeta_4$, and ζ_5. We compute $\mathcal{K}_\alpha = \lim_{d\to\infty}\mathbb{E}[K_\alpha]$ by using the expressions $\mathcal{K}_\parallel(q,z) = -\frac{\psi_D\zeta_3(q,z)}{a\mu_1}$ and $\mathcal{K}_\perp(q,z) = \frac{(1-\psi_D)(1-\zeta_5(q,z))}{h\mu_1^2+hq}$. Finally, we use (17) to obtain $\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^\infty_{\text{test},\alpha}(\hat{A}_t)\right]$.

Next, we compute the train error. From the above:

$$\begin{aligned}
\mathcal{E}^\infty_{\text{train}}(\hat{A}_t) &= -h\mu_1^2(E_{1,\parallel} + E_{1,\perp}) - h\lambda\mu_1^2(E_{3,\parallel} + E_{3,\perp}) + 1 \,, \\
&= -h\mu_1^2(K_\parallel(0,-\lambda) + K_\perp(0,-\lambda)) - h\lambda\mu_1^2(\frac{\partial K_\parallel}{\partial z}(0,-\lambda) + \frac{\partial K_\perp}{\partial z}(0,-\lambda)) + 1 \,.
\end{aligned}$$

Thus,

$$\begin{aligned}
\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^\infty_{\text{train}}(\hat{A}_t)\right] &= -h\mu_1^2(\mathcal{K}_\parallel(0,-\lambda) + \mathcal{K}_\perp(0,-\lambda)) - h\lambda\mu_1^2(\frac{\partial \mathcal{K}_\parallel}{\partial z}(0,-\lambda) + \frac{\partial \mathcal{K}_\perp}{\partial z}(0,-\lambda)) + 1 \\
&= 1 - h\mu_1^2(e_{1,\parallel} + e_{1,\perp}) - h\lambda\mu_1^2(e_{3,\parallel} + e_{3,\perp}) \,, \qquad (23)
\end{aligned}$$

where $e_{1,\alpha} = \mathcal{K}_\alpha(0,-\lambda)$, $e_{3,\alpha} = \frac{\partial\mathcal{K}_\alpha}{\partial z}(0,-\lambda)$. This concludes the proof of Theorem 3.2. □

C PROOF OF THEOREM 3.6

We use Lemma C.1 to express the test and train errors for $m = 1$ case as sums of traces of rational functions of random matrices when $P_0 \equiv \mathcal{N}(0,\mathcal{C})$. We recall the definition of the following entities: $\mu_0(\kappa) = \mathbb{E}_g[\varrho(\kappa g)]$, $\mu_1(\kappa) = \mathbb{E}_g[\varrho(\kappa g)g]$, $\|\varrho(\kappa\,\cdot)\|^2 = \mathbb{E}_g\left[\varrho(\kappa g)^2\right]$, $v^2(\kappa) = \|\rho(\kappa\,\cdot)\|^2 - \mu_0^2(\kappa) - \mu_1^2(\kappa)$, $c(\gamma,\kappa) = \mathbb{E}_{u,v\sim P^\gamma}[\varrho(\kappa u)\varrho(\kappa v)]$, where $g \sim \mathcal{N}(0,1)$ and P^γ denote the bivariate standard Gaussian distribution with correlation coefficient γ. Explicitly,

$$P^\gamma(x,y) = \frac{1}{2\pi\sqrt{1-\gamma^2}}e^{-\frac{x^2+y^2-2\gamma xy}{2(1-\gamma^2)}} \,.$$

Lemma C.1. *Let $\mathcal{M}$ be a subspace in $\mathbb{R}^D$. Let $\Pi_\parallel$ denote a projection matrix onto $\mathcal{M}$, and let $\Pi_\perp = I - \Pi_\parallel$. Suppose $P_0 \equiv \mathcal{N}(0,\mathcal{C})$. Let $a_t, h_t, \lambda, \varrho, W_t$ be as in (11), and define $\Sigma_t = a_t^2\mathcal{C} + h_t I_d$, $\sigma_t^2 = \frac{1}{d}\operatorname{tr}\{\Sigma_t\}$, and*

$\sigma_0^2 = \frac{1}{d}\operatorname{tr}\{\mathcal{C}\}$. *Assume that* σ_t *converges to a finite limit as* $d \to \infty$. *Let* $Z = [z_1, z_2, \cdots, z_n]$ *with* $z_i \sim \mathcal{N}(0, I_d)$ *i.i.d., and let* $Y = [y_1, y_2, \cdots, y_n]$ *for* $y_i = a_t x_i + \sqrt{h_t} z_i$, *with* $x_i \sim P_0$ *i.i.d. Assume* ϱ *satisfies Assumption 3.1. Then, for the minimizer of (11)* $\hat{A}_t$, *for* $\alpha \in \{\|, \perp\}$, *as* $d, n, p \to \infty$, *the test and train errors can be represented as:*

$$\mathcal{E}^1_{test,\alpha}(\hat{A}_t) = E_{0,\alpha} - \frac{2\mu_1(\sigma_t)}{\sigma_t\sqrt{h_t}} E_{1,\alpha} + \frac{\mu_1(\sigma_t)^2}{\sigma_t^2 h_t} E_{2,\alpha} + \frac{v(\sigma_t)^2}{h_t} E_{3,\alpha} + \frac{\mu_0(\sigma_t)^2}{h_t} E_{4,\alpha} ,$$
$$\mathcal{E}^1_{train}(\hat{A}_t) = -E_5 - \lambda(E_{3,\|} + E_{3,\perp}) + 1 ,$$

where

$$E_{0,\alpha} = \frac{1}{d}\operatorname{tr}\{\Pi_\alpha \Sigma^{-1}\} ,$$
$$E_{1,\alpha} = \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}(U + \lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}, \Pi_\alpha\right\} ,$$
$$E_{2,\alpha} = \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}(U + \lambda I_p)^{-1}\frac{W_t}{\sqrt{d}}\frac{W_t^T}{\sqrt{d}}(U + \lambda I_p)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\right\} ,$$
$$E_{3,\alpha} = \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}(U + \lambda I_p)^{-2}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\right\} ,$$
$$E_{4,\alpha} = \frac{1}{d}\boldsymbol{1}_p^T(U + \lambda I_p)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}(U + \lambda I_p)^{-1}\boldsymbol{1}_p ,$$
$$E_5 = \frac{1}{d}\operatorname{tr}\left\{\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}(U + \lambda I_p)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\right\} ,$$

with

$$U = \frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}} ,$$

and

$$F = \mu_0(\nu_t)\boldsymbol{1}_p\boldsymbol{1}_n^T + \frac{\mu_1(\sigma_t)}{\sigma_t}\frac{W_t}{\sqrt{d}}Y + v(\sigma_t)\Omega .$$

Proof. When $m = 1$, the loss function reduces to:

$$\mathcal{L}^1_t(A_t) = \frac{1}{dn}\sum_{i=1}^n \left\| \sqrt{h_t}\frac{A_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{p}}(a_t x_i + \sqrt{h_t} z_i)\right) + z_i \right\|^2 + \frac{h_t\lambda}{dp}\|A_t\|_F^2 . \tag{24}$$

Let $X = [x_1, x_2, \cdots, x_n]$, $Z = [z_1, z_2, \cdots, z_n]$, and let Y be the matrix whose i^{th} column is $y_i = a_t x_i + \sqrt{h_t} z_i$. Also let $F = \varrho(\frac{W}{\sqrt{d}}Y)$. We then write:

$$\begin{aligned}\mathcal{L}^1_t(A_t) &= \frac{1}{dn}\sum_{i=1}^n \left\| \sqrt{h_t}\frac{A_t}{\sqrt{p}}\varrho\left(\frac{W_t}{\sqrt{p}}(a_t x_i + \sqrt{h_t} z_i)\right) + z_i \right\|^2 + \frac{h_t\lambda}{dp}\|A_t\|_F^2 \\ &= \frac{h_t}{d}\operatorname{tr}\left\{\frac{A_t}{\sqrt{p}}^T\frac{A_t}{\sqrt{p}}U\right\} + \frac{2\sqrt{h_t}}{d}\operatorname{tr}\left\{\frac{A_t}{\sqrt{p}}V\right\} + 1 + \frac{h_t\lambda}{d}\operatorname{tr}\left\{\frac{A_t^T}{\sqrt{p}}\frac{A_t}{\sqrt{p}}\right\} ,\end{aligned}$$

where

$$U = \frac{1}{n}\varrho\left(\frac{W_t}{\sqrt{d}}Y\right)\varrho\left(\frac{W_t}{\sqrt{d}}Y\right)^T = \frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}} ,$$

and

$$V = \frac{1}{n}\varrho\left(\frac{W_t}{\sqrt{d}}Y\right)Z^T = \frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}} .$$

Thus, the optimal A_t is written as

$$\frac{\hat{A}_t}{\sqrt{p}} = -\frac{1}{\sqrt{h_t}} V^T (U + \lambda I_p)^{-1} = -\frac{1}{\sqrt{h_t}} \frac{Z}{\sqrt{n}} \frac{F^T}{\sqrt{n}} \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} . \tag{25}$$

Test error:

$$\begin{aligned}
\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t) &= \frac{1}{d} \mathbb{E}_{x\sim P_t} \left[\left\| \Pi_\alpha \frac{\hat{A}_t}{\sqrt{p}} \varrho \left(\frac{W_t}{\sqrt{d}} x \right) - \Pi_\alpha \nabla \log P_t(x) \right\|^2 \right] \\
&= \frac{1}{d} \mathbb{E}_{x\sim P_t} \left[\left\| \frac{F}{\sqrt{n}} \frac{Z^T}{\sqrt{n}} \Pi_\alpha \frac{\hat{A}_t}{\sqrt{p}} \varrho \left(\frac{W_t}{\sqrt{d}} x \right) + \Pi_\alpha \Sigma_t^{-1} x \right\|^2 \right] \\
&= \frac{1}{d} \operatorname{tr}\{ \Pi_\alpha \Sigma_t^{-1} \} + \frac{2}{d} \operatorname{tr} \left\{ \Sigma_t^{-1} \Pi_\alpha \frac{\hat{A}_t}{\sqrt{p}} \underbrace{\mathbb{E}_x \left[\varrho \left(\frac{W_t}{\sqrt{d}} x \right) x^T \right]}_{:=\tilde{V}} \right\} \\
&\quad + \frac{1}{d} \operatorname{tr} \left\{ \frac{\hat{A}_t^T}{\sqrt{p}} \Pi_\alpha \frac{\hat{A}_t}{\sqrt{p}} \underbrace{\mathbb{E}_x \left[\varrho \left(\frac{W_t}{\sqrt{d}} x \right) \varrho \left(\frac{W_t}{\sqrt{d}} x \right)^T \right]}_{:=\tilde{U}} \right\} .
\end{aligned}$$

We have already derived expressions for $\tilde{V}$ and $\tilde{U}$ in Section B. Namely, $\tilde{V} = \frac{\mu_1(\sigma_t)}{\sigma_t} \frac{W_t}{\sqrt{d}} \Sigma_t$, $\tilde{U} = \mu_0(\sigma_t)^2 \mathbf{1}_p \mathbf{1}_p^T + \frac{\mu_1(\sigma_t)^2}{\sigma_t^2} \frac{W_t}{\sqrt{d}} \Sigma_t \frac{W_t^T}{\sqrt{d}} + v(\sigma_t)^2 I_p$. Since we focus on a single time instant, we drop the subscript t henceforth. However, it is important to keep the time-dependence of a and h, and the relation $a^2 + h = 1$. Thus:

$$\begin{aligned}
\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t) &= \frac{1}{d} \operatorname{tr}\{ \Pi_\alpha \Sigma^{-1} \} - \frac{2\mu_1(\sigma)}{\sqrt{h}\sigma d} \operatorname{tr} \left\{ \Pi_\alpha \frac{Z}{\sqrt{n}} \frac{F^T}{\sqrt{n}} \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \frac{W}{\sqrt{d}} \right\} \\
&\quad + \frac{\mu_1(\sigma)^2}{h\sigma^2 d} \operatorname{tr} \left\{ \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \frac{F}{\sqrt{n}} \frac{Z^T}{\sqrt{n}} \Pi_\alpha \frac{Z}{\sqrt{n}} \frac{F^T}{\sqrt{n}} \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \frac{W}{\sqrt{d}} \Sigma \frac{W^T}{\sqrt{d}} \right\} \\
&\quad + \frac{v(\sigma)^2}{hd} \operatorname{tr} \left\{ \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \frac{F}{\sqrt{n}} \frac{Z^T}{\sqrt{n}} \Pi_\alpha \frac{Z}{\sqrt{n}} \frac{F^T}{\sqrt{n}} \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \right\} , \\
&\quad + \frac{\mu_0(\sigma)^2}{hd} \mathbf{1}_p^T \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \frac{F}{\sqrt{n}} \frac{Z^T}{\sqrt{n}} \Pi_\alpha \frac{Z}{\sqrt{n}} \frac{F^T}{\sqrt{n}} \left(\frac{F}{\sqrt{n}} \frac{F^T}{\sqrt{n}} + \lambda I_p \right)^{-1} \mathbf{1}_p .
\end{aligned}$$

To handle the non-linearity in F, we apply the Gaussian equivalence principle. In particular:

$$F = \mu_0(\sigma) \mathbf{1}_p \mathbf{1}_n^T + \frac{\mu_1(\sigma)}{\sigma} \frac{W}{\sqrt{d}} Y + v(\sigma) \Omega .$$

Let

$$
\begin{aligned}
E_{0,\alpha} &= \frac{1}{d}\operatorname{tr}\{\Pi_\alpha \Sigma^{-1}\},\\
E_{1,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\Pi_\alpha \frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\frac{W}{\sqrt{d}}\Pi_\alpha\right\},\\
E_{2,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}}\right\},\\
E_{3,\alpha} &= \frac{1}{d}\operatorname{tr}\left\{\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-2}\right\},\\
E_{4,\alpha} &= \frac{1}{d}\mathbf{1}_p^T\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\Pi_\alpha\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\mathbf{1}_p\,,\\
E_5 &= \frac{1}{d}\operatorname{tr}\left\{\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}}+\lambda I_p\right)^{-1}\right\}.
\end{aligned}
$$

With the above definitions, we can write $\mathcal{E}^1_{\text{test},\alpha}$ as

$$
\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t) = E_{0,\alpha} - \frac{2\mu_1(\sigma)}{\sqrt{h}\sigma}E_{1,\alpha} + \frac{\mu_1(\sigma)^2}{h\sigma^2}E_{2,\alpha} + \frac{v(\sigma)^2}{h}E_{3,\alpha} + \frac{\mu_0(\sigma)^2}{h}E_{4,\alpha}\,.
$$

Train error: To compute the train error, we proceed as follows:

$$
\begin{aligned}
\mathcal{E}^1_{\text{train}}(\hat{A}_t) &= \frac{h}{d}\operatorname{tr}\left\{\frac{\hat{A}_t^T}{\sqrt{p}}\frac{\hat{A}_t}{\sqrt{p}}(U+\lambda I_p)\right\} + \frac{2\sqrt{h}}{d}\operatorname{tr}\left\{\frac{\hat{A}_t}{\sqrt{p}}V\right\} + 1 - \frac{h\lambda}{d}\operatorname{tr}\left\{\frac{\hat{A}_t^T}{\sqrt{p}}\frac{\hat{A}_t}{\sqrt{p}}\right\}\\
&= \frac{1}{d}\operatorname{tr}\{(U+\lambda I_p)^{-1}VV^T\} - \frac{2}{d}\operatorname{tr}\{V^T(U+\lambda I_p)^{-1}V\} + 1 - \frac{\lambda}{d}\operatorname{tr}\{(U+\lambda I_p)^{-2}VV^T\}\\
&= -\frac{1}{d}\operatorname{tr}\left\{(U+\lambda I_p)^{-1}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\right\} + 1 - \frac{\lambda}{d}\operatorname{tr}\left\{(U+\lambda I_p)^{-2}\frac{F}{\sqrt{n}}\frac{Z^T}{\sqrt{n}}\frac{Z}{\sqrt{n}}\frac{F^T}{\sqrt{n}}\right\}\\
&= -E_5 - \lambda(E_{3,\parallel} + E_{3,\perp}) + 1\,.
\end{aligned}
$$

This completes the proof of lemma C.1. □

The following theorem derives the test and train errors when P_0 is a Gaussian supported on a D-dimensional subspace in $\mathbb{R}^d$.

Theorem C.2. *Let $\mathcal{M}_\parallel$ be a D-dimensional subspace in $\mathbb{R}^D$, $\Pi_\parallel$ a projection matrix onto it, and $P_0 \equiv \mathcal{N}\left(0, \Pi_\parallel\right)$ with ϱ satisfying Assumption 3.1. Define $\sigma_0^2 = \frac{1}{d}\operatorname{tr}\{\Pi_\parallel\}$, $\sigma_t^2 = a_t^2\sigma_0^2 + h_t$, and $\mu_{1,t} = \mu_1(\sigma_t)/\sigma_t$. Set $v^2 = \|\varrho(\sigma_t\cdot)\|^2 - \mu_1(\sigma_t)^2$. Let $\psi_D = \frac{D}{d}, \psi_n = \frac{n}{d}$, $\psi_p = \frac{p}{d}$ and $\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6$ be the solution of the following system of algebraic equations in q and z:*

$$
\begin{aligned}
\zeta_1(-z + (1-\psi_D)(q+\mu_{1,t}^2\zeta_4)h_t\zeta_6 + \psi_D(q+\mu_{1,t}^2\zeta_4)\zeta_3 + v^2\zeta_4) - 1 &= 0\,,\\
\zeta_2(1+q\psi_p\zeta_1) + \mu_{1,t}^2\psi_p\zeta_1\zeta_2\zeta_4 + \psi_p a_t\mu_{1,t}\zeta_1 &= 0\,,\\
\zeta_5(1+qh_t\psi_p\zeta_1) + \mu_{1,t}\psi_p\sqrt{h_t}\zeta_1(1+\mu_{1,t}\sqrt{h_t}\zeta_4\zeta_5) &= 0\,,\\
\zeta_3(1+q\psi_p\zeta_1) + \mu_{1,t}^2\psi_p\zeta_1\zeta_3\zeta_4 - 1 &= 0\,,\\
\zeta_4(\psi_n + \psi_p v^2\zeta_1 - (1-\psi_D)\mu_{1,t}\sqrt{h_t}\zeta_5 - \psi_D\mu_{1,t}\zeta_2/a_t) - \psi_n &= 0\,,\\
\zeta_6(1+qh_t\psi_p\zeta_1) + \mu_{1,t}^2\psi_p h_t\zeta_1\zeta_6\zeta_4 - 1 &= 0\,.
\end{aligned}
$$

Let $e_{0,\|} = \psi_D$ *and* $e_{0,\perp} = \frac{1-\psi_D}{h_t}$. *Define the functions*

$$e_{1,\alpha} = \begin{cases} -\frac{\psi_D\sqrt{h_t}}{a_t}\zeta_4\zeta_2 \ , & \alpha = \| \\ -(1-\psi_D)\zeta_4\zeta_5 \ , & \alpha = \perp \end{cases} , \tag{26}$$

$$K_\alpha(q,z) = \begin{cases} \psi_D(1-\zeta_4(1+\mu_{1,t}h_t\zeta_4\zeta_2/a_t)) \ , & \alpha = \| \\ 1-\psi_D - \frac{1}{\mu_{1,t}\sqrt{h_t}}(qh_te_{1,\perp} + (1-\psi_D)\mu_{1,t}\sqrt{h_t}\zeta_4\zeta_6) \ , & \alpha = \perp \end{cases} . \tag{27}$$

For $\alpha \in \{\|, \perp\}$, *let* $\varepsilon^1_{test,\alpha} = e_{0,\alpha} - \frac{2\mu_{1,t}}{\sqrt{h_t}}e_{1,\alpha} - \frac{\mu_{1,t}^2}{h_t}\frac{\partial \mathcal{K}_\alpha}{\partial q}(0,-\lambda) + \frac{v^2}{h_t}\frac{\partial \mathcal{K}_\alpha}{\partial z}(0,-\lambda)$ *and* $\varepsilon^1_{train} = 1 - (\mathcal{K}_\|(0,-\lambda) + \mathcal{K}_\perp(0,-\lambda)) - \lambda(\frac{\partial \mathcal{K}_\|}{\partial z}(0,-\lambda) + \frac{\partial \mathcal{K}_\perp}{\partial z}(0,-\lambda))$. *Then, for the minimizer of (10)* $\hat{A}_t$, *as* $d,n,p \to \infty$:

$$\lim_{d,n,p\to\infty} \mathbb{E}\left[\mathcal{E}^1_{test,\|}(\hat{A}_t)\right] = \varepsilon^1_{test,\|} \ , \quad \lim_{d,n,p\to\infty} \mathbb{E}\left[\mathcal{E}^1_{test,\perp}(\hat{A}_t)\right] = \varepsilon^1_{test,\perp} \ , \quad \lim_{d,n,p\to\infty} \mathbb{E}\left[\mathcal{E}^1_{train}(\hat{A}_t)\right] = \varepsilon^1_{train} \ .$$

Remark C.3. Theorem 3.6 can be obtained by substituting $\psi_D = 1$.

Proof. To prove Theorem 3.6, we specialize Lemma C.1 to $\mathcal{C} = \Pi_\|$. As in the proof of Lemma C.1, since we focus on a single time instant, we drop the subscript t in the above expressions. For the $\mathcal{C}$ used here, we note that $\sigma^2 = a^2\psi_D + h, \sigma_0^2 = \psi_D$. Also, let $\mu_0 = \mu_0(\sigma), \mu_1 = \mu_1(\sigma)/\sigma, \|\varrho(\sigma\,\cdot)\| = \|\varrho\|$. Let $W = [W_\|, W_\perp]$, $Z = [Z_\|^T, Z_\perp^T]^T$, $X = [X_\|^T, X_\perp^T]^T$, and $Y = [Y_\|^T, Y_\perp^T]^T$

For simplicity in presentation, we assume that $\mu_0 = 0$ and write

$$\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t) = E_{0,\alpha} - \frac{2\mu_1}{\sqrt{h}}E_{1,\alpha} + \frac{\mu_1^2}{h}E_{2,\alpha} + \frac{v^2}{h}E_{3,\alpha} \ ,$$
$$\mathcal{E}^1_{\text{train}}(\hat{A}_t) = -E_5 - \lambda(E_{3,\|} + E_{3,\perp}) + 1 \ .$$

Letting

$$R(q,z) = \left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}} + q\frac{W}{\sqrt{d}}\Sigma\frac{W^T}{\sqrt{d}} - zI_p\right)^{-1} ,$$
$$= \left(\frac{F}{\sqrt{n}}\frac{F^T}{\sqrt{n}} + q\frac{W_\|}{\sqrt{d}}\frac{W_\|^T}{\sqrt{d}} + qh\frac{W_\perp}{\sqrt{d}}\frac{W_\perp^T}{\sqrt{d}} - zI_p\right)^{-1} ,$$

and

$$K_\alpha(q,z) = \frac{1}{d}\operatorname{tr}\left\{\frac{Z_\alpha}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{F}{\sqrt{n}}\frac{Z_\alpha^T}{\sqrt{n}}\right\} ,$$

we have

$$E_{2,\alpha} = -\frac{\mathrm{d}K_\alpha}{\mathrm{d}q}(0,-\lambda) \ ,$$
$$E_{3,\alpha} = \frac{\mathrm{d}K_\alpha}{\mathrm{d}z}(0,-\lambda) \ ,$$
$$E_5 = K_\|(0,-\lambda) + K_\perp(0,-\lambda) \ .$$

Therefore, assuming we can interchange limit, derivatives and expectations, it suffices to have $e_{1,\alpha} = \lim_{d\to\infty}\mathbb{E}[E_{1,\alpha}]$ and the function $\mathcal{K}_\alpha(q,z) := \lim_{d\to\infty}\mathbb{E}[K_\alpha(q,z)]$ to have an expression for $\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t)\right]$. We have

$$\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t)\right] = \lim_{d\to\infty}E_{0,\alpha} - \frac{2\mu_1}{\sqrt{h}}\lim_{d\to\infty}\mathbb{E}[E_{1,\alpha}] + \frac{\mu_1^2}{h}\lim_{d\to\infty}\mathbb{E}[E_{2,\alpha}] + \frac{v^2}{h}\lim_{d\to\infty}\mathbb{E}[E_{3,\alpha}]$$
$$= e_{0,\alpha} - \frac{2\mu_1}{\sqrt{h}}e_{1,\alpha} + \frac{\mu_1^2}{h}e_{2,\alpha} + \frac{v^2}{h}e_{3,\alpha} \ , \tag{28}$$

where $e_{2,\alpha} = -\frac{\partial \mathcal{K}_\alpha}{\partial q}(0,-\lambda)$, $e_{3,\alpha} = \frac{\partial \mathcal{K}_\alpha}{\partial z}(0,-\lambda)$. As in the previous theorem, one could avoid interchanging limit of expectation and derivatives by differentiating first $K_\alpha(q,z)$ and using a larger linear pencil.

As in Theorem 3.2, proved in Appendix B, we use linear pencils to obtain the desired terms. Explicitly, the following 6×6 linear pencil matrix:

$$L = \left[\begin{array}{c|ccccccc} -zI_p & qh\frac{W_\perp}{\sqrt{d}} & q\frac{W_\parallel}{\sqrt{d}} & v\frac{\Omega}{\sqrt{n}} & \mu_1\sqrt{h}\frac{W_\perp}{\sqrt{d}} & \mu_1\sqrt{h}\frac{W_\parallel}{\sqrt{d}} & \mu_1 a\frac{W_\parallel}{\sqrt{d}} & 0 \\ \hline -\frac{W_\perp^T}{\sqrt{d}} & I_d & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{W_\parallel^T}{\sqrt{d}} & 0 & I_d & 0 & 0 & 0 & 0 & 0 \\ -v\frac{\Omega^T}{\sqrt{n}} & -\mu_1\sqrt{h}\frac{Z_\perp^T}{\sqrt{n}} & -\mu_1\frac{Y_\parallel^T}{\sqrt{n}} & I_n & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{Z_\perp}{\sqrt{n}} & I_d & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{Z_\parallel}{\sqrt{n}} & 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & -\frac{X_\parallel^T}{\sqrt{n}} & 0 & 0 & I_n & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{Z_\parallel^T}{\sqrt{n}} & 0 & I_n \end{array}\right].$$

By computing L^{-1}, we obtain:

$$(L^{-1})^{8,4}(q,z) = \frac{Z_\parallel^T}{\sqrt{n}}\frac{Z_\parallel}{\sqrt{n}} - \frac{Z_\parallel^T}{\sqrt{n}}\frac{Z_\parallel}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{F}{\sqrt{n}}\,, \tag{29}$$

$$(L^{-1})^{6,7}(q,z) = -\mu_1 a\frac{Z_\parallel}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{W_\parallel}{\sqrt{d}}\,, \tag{30}$$

$$(L^{-1})^{5,2}(q,z) = \mu_1\sqrt{h}\frac{Z_\perp}{\sqrt{n}}\frac{Z_\perp^T}{\sqrt{n}} - \mu_1\sqrt{h}\frac{Z_\perp}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{F}{\sqrt{n}}\frac{Z_\perp^T}{\sqrt{n}} - qh\frac{Z_\perp}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{W_\perp^T}{\sqrt{d}}\,, \tag{31}$$

$$(L^{-1})^{5,5}(q,z) = I - \mu_1\sqrt{h}\frac{Z_\perp}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{W_\perp^T}{\sqrt{d}}\,. \tag{32}$$

Hence,

$$E_{1,\alpha} = \frac{1}{d}\operatorname{tr}\left\{\frac{Z_\alpha}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(0,-\lambda)\frac{W_\alpha}{\sqrt{d}}\right\} = \begin{cases} -\frac{1}{\mu_1 ad}\operatorname{tr}\{(L^{-1})^{6,7}(0,-\lambda)\}, & \alpha = \parallel, \\ \frac{1-\psi_D}{\mu_1\sqrt{h}}(1-\frac{1}{d-D}\operatorname{tr}\{(L^{-1})^{5,5}(0,-\lambda)\}), & \alpha = \perp, \end{cases}\,, \tag{33}$$

$$K_\alpha(q,z) = \frac{1}{d}\operatorname{tr}\left\{\frac{Z_\alpha}{\sqrt{n}}\frac{F^T}{\sqrt{n}}R(q,z)\frac{F}{\sqrt{n}}\frac{Z_\alpha^T}{\sqrt{n}}\right\} = \begin{cases} \psi_D - \frac{1}{d}\operatorname{tr}\{(L^{-1})^{8,4}(0,-\lambda)\}, & \alpha = \parallel, \\ 1-\psi_D - \frac{1}{\mu_1\sqrt{h}}(qhE_{1,\perp} + \frac{1}{d}\operatorname{tr}\{(L^{-1})^{5,2}(0,-\lambda)\}), & \alpha = \perp, \end{cases}\,. \tag{34}$$

The traces of blocks in L^{-1} are computed as in the Appendix B. We then have:

$$g = \begin{bmatrix} g_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & g_{22} & 0 & 0 & g_{25} & 0 & 0 & 0 \\ 0 & 0 & g_{33} & 0 & 0 & g_{36} & g_{37} & 0 \\ 0 & 0 & 0 & g_{44} & 0 & 0 & 0 & 0 \\ 0 & g_{52} & 0 & 0 & g_{55} & 0 & 0 & 0 \\ 0 & 0 & g_{63} & 0 & 0 & g_{66} & g_{67} & 0 \\ 0 & 0 & g_{73} & 0 & 0 & g_{76} & g_{77} & 0 \\ 0 & 0 & 0 & g_{84} & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -z & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\eta_L(g) = \begin{bmatrix} \sigma_{12}^{21}g_{22}+\sigma_{15}^{21}g_{52}+\sigma_{13}^{31}g_{33}+\sigma_{16}^{31}g_{63}+\sigma_{17}^{31}g_{73}+\sigma_{14}^{41}g_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{21}^{12}g_{11} & 0 & 0 & \sigma_{21}^{15}g_{11} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{31}^{13}g_{11} & 0 & 0 & \sigma_{31}^{16}g_{11} & \sigma_{31}^{17}g_{11} & 0 \\ 0 & 0 & 0 & \sigma_{41}^{14}g_{11}+\sigma_{42}^{54}g_{25}+\sigma_{43}^{64}g_{36}+\sigma_{43}^{74}g_{37} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{54}^{42}g_{44} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{64}^{43}g_{44} & 0 & 0 & \sigma_{64}^{86}g_{48} & 0 & 0 \\ 0 & 0 & \sigma_{74}^{43}g_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{86}^{64}g_{66} & 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$\sigma_{12}^{21} = -(1-\psi_D)qh\,, \quad \sigma_{21}^{12} = -\psi_p qh\,, \quad \sigma_{15}^{21} = -(1-\psi_D)\mu_1\sqrt{h}\,, \quad \sigma_{21}^{15} = -\psi_p\mu_1\sqrt{h} \tag{35}$$

$$\sigma_{13}^{31} = -\psi_D q\,, \quad \sigma_{31}^{13} = -\psi_p q\,, \quad \sigma_{16}^{31} = -\psi_D\mu_1\sqrt{h}\,, \quad \sigma_{31}^{16} = -\psi_p\mu_1\sqrt{h} \tag{36}$$

$$\sigma_{17}^{31} = -\psi_D a\mu_1\,, \quad \sigma_{31}^{17} = -\psi_p a\mu_1\,, \quad \sigma_{14}^{41} = -v^2\,, \quad \sigma_{41}^{14} = -\psi_p v^2/\psi_n \tag{37}$$

$$\sigma_{42}^{54} = (1-\psi_D)\mu_1\sqrt{h}/\psi_n\,, \quad \sigma_{54}^{42} = \mu_1\sqrt{h}\,, \quad \sigma_{43}^{64} = \psi_D\mu_1\sqrt{h}/\psi_n\,, \quad \sigma_{64}^{43} = \mu_1\sqrt{h} \tag{38}$$

$$\sigma_{43}^{74} = \psi_D a\mu_1/\psi_n\,, \quad \sigma_{74}^{43} = a\mu_1\,, \quad \sigma_{64}^{86} = 1\,, \quad \sigma_{86}^{64} = \psi_D/\psi_n. \tag{39}$$

Finally, the matrix g satisfies the fixed point equation:

$$(B - \eta_L(g))g = I\,.$$

This leads, after simplification, the following set of algebraic equations:

$$\begin{aligned}
\zeta_1(-z + (1-\psi_D)(q+\mu_1^2\zeta_4)h\zeta_6 + \psi_D(q+\mu_1^2\zeta_4)\zeta_3 + v^2\zeta_4) - 1 &= 0\,,\\
\zeta_2(1+q\psi_p\zeta_1) + \mu_1^2\psi_p\zeta_1\zeta_2\zeta_4 + \psi_p a\mu_1\zeta_1 &= 0\,,\\
\zeta_5(1+qh\psi_p\zeta_1) + \mu_1\psi_p\sqrt{h}\zeta_1(1+\mu_1\sqrt{h}\zeta_4\zeta_5) &= 0\,,\\
\zeta_3(1+q\psi_p\zeta_1) + \mu_1^2\psi_p\zeta_1\zeta_3\zeta_4 - 1 &= 0\,,\\
\zeta_4(\psi_n + \psi_p v^2\zeta_1 - (1-\psi_D)\mu_1\sqrt{h}\zeta_5 - \psi_D\mu_1\zeta_2/a) - \psi_n &= 0\,,\\
\zeta_6(1+qh\psi_p\zeta_1) + \mu_1^2\psi_p h\zeta_1\zeta_6\zeta_4 - 1 &= 0\,.
\end{aligned}$$

where $\zeta_1 = g_{11}$, $\zeta_2 = g_{37}$, $\zeta_3 = g_{33}$, $\zeta_4 = g_{44}$, $\zeta_5 = g_{25}$, $\zeta_6 = g_{22}$. As before, we solve the system numerically to obtain $\zeta_1, \zeta_2, \zeta_3, \zeta_4$, ζ_5, and ζ_6. We then obtain $e_{1,\alpha}$ and $\mathcal{K}_\alpha$ through the expressions:

$$e_{1,\alpha} = \begin{cases} -\dfrac{\psi_D\sqrt{h}}{a}\zeta_4\zeta_2\,, & \alpha = \| \\ -(1-\psi_D)\zeta_4\zeta_5\,, & \alpha = \perp \end{cases}, \tag{40}$$

$$K_\alpha(q,z) = \begin{cases} \psi_D(1-\zeta_4(1+\mu_1 h\zeta_4\zeta_2/a))\,, & \alpha = \| \\ 1-\psi_D - \dfrac{1}{\mu_1\sqrt{h}}(qhe_{1,\perp} + (1-\psi_D)\mu_1\sqrt{h}\zeta_4\zeta_6)\,, & \alpha = \perp \end{cases}. \tag{41}$$

Finally, we use (28) to compute $\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^1_{\text{test},\alpha}(\hat{A}_t)\right]$.

Train error: To compute the train error, we proceed as follows:

$$\begin{aligned}
\mathcal{E}^1_{\text{train}}(\hat{A}_t) &= -E_5 - \lambda(E_{3,\|} + E_{3,\perp}) + 1\,,\\
&= -(K_\|(0,-\lambda) + K_\perp(0,-\lambda)) - \lambda(\frac{\partial K_\|}{\partial z}(0,-\lambda) + \frac{\partial K_\perp}{\partial z}(0,-\lambda)) + 1\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
\lim_{d\to\infty}\mathbb{E}\left[\mathcal{E}^1_{\text{train}}(\hat{A}_t)\right] &= 1 - (\mathcal{K}_\|(0,-\lambda) + \mathcal{K}_\perp(0,-\lambda)) - \lambda(\frac{\partial \mathcal{K}_\|}{\partial z}(0,-\lambda) + \frac{\partial \mathcal{K}_\perp}{\partial z}(0,-\lambda))\,,\\
&= 1 - (\mathcal{K}_\|(0,-\lambda) + \mathcal{K}_\perp(0,-\lambda)) - \lambda(e_{3,\perp} + e_{3,\perp}).\,,
\end{aligned} \tag{42}$$

where $e_{3,\alpha} = \frac{\partial \mathcal{K}_\alpha}{\partial z}(0,-\lambda)$. This concludes the proof of Theorem 3.6.

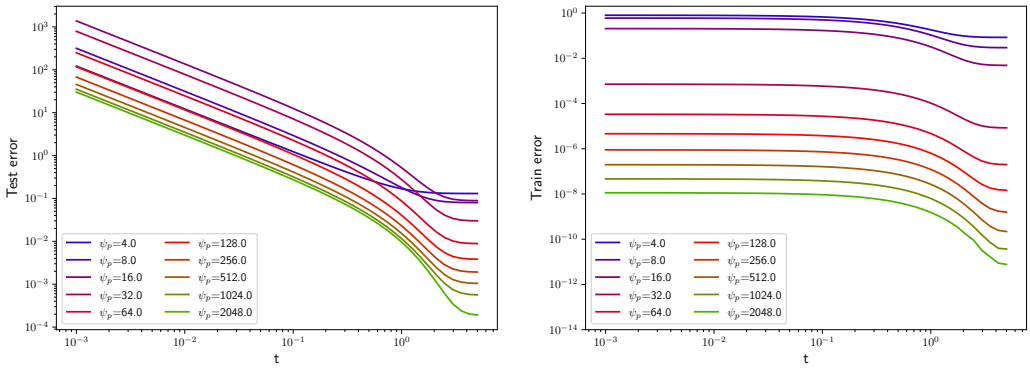
□

D ILLUSTRATIONS OF ANALYTICALLY COMPUTED TEST AND TRAIN ERRORS

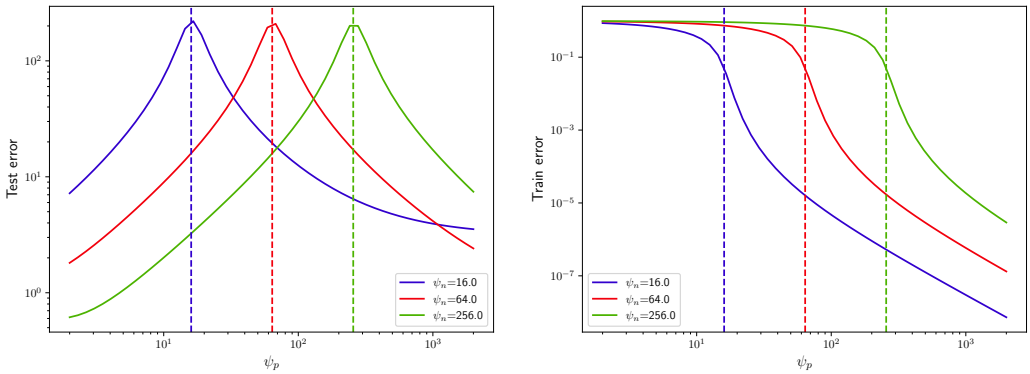
In this section, we provide additional plots to further illustrate the analytical predictions of test and train errors derived from Theorems B.2 and C.2. First, we present the learning curves for the case of $m = 1$, which were

omitted from the main text due to space constraints. Next we show the learning curves when the data lies on a D-dimensional subspace in $\mathbb{R}^d$. Thereafter, we include plots demonstrating the impact of the regularization strength λ on the learning curves. Finally, we provide learning curves for cases where different activation functions are employed, highlighting their influence on the model's performance.

D.1 Test and train errors for $m = 1$



(a) Test and train error as a function of t for different values of ψ_p and a fixed $\psi_n = 20.0$.



(b) Test and train error as a function of ψ_p for different values of ψ_n and a fixed $t = 0.01$. The dashed vertical lines indicate $\psi_p = \psi_n$.

Figure 5: Learning curves for $m = 1$.We used $\lambda = 0.001$ and the activation function used is ReLU.

Fig. 5 shows test and train errors for $m = 1$ case. Fig. 5a presents the learning curves as a function of t for various ψ_p with ψ_n fixed. In Fig. 5b we plot the learning curves as a function of ψ_p for different values of ψ_n and keeping t fixed and small.

As already discussed in the main text, the curves reveal several trends. The train error decreases monotonically with increasing ψ_p for all t, reflecting the model's capacity to interpolate the training data. However, the test error shows a non-monotonic behavior with ψ_p. The dependence of test error on t is also evident. The test error increases as t decreases, but for small t, it remains at least two orders of magnitude lower than in the $m = \infty$ case. Furthermore, the test error decreases as ψ_p increases beyond ψ_n, suggesting the model does not display memorization behavior under these conditions. It contrasts with the $m = \infty$ case, where memorization significantly impacts the test error. These findings indicate that larger m increases the tendency of diffusion models to memorize the initial dataset.

Lastly, as observed in earlier works such as Mei and Montanari (2022); Bodin and Macris (2021); Hu et al. (2024), we also detect the double descent phenomenon, characterized by a peak in the test error at $\psi_p = \psi_n$ (see Fig. 5b).

D.2 Plots when $D < d$

In this section, we present the learning curves for the case $\psi_D = \frac{D}{d} = 0.2$. When the data lies on a subspace, we need to investigate separately the components of the score function that are parallel and orthogonal to subspace, since they exhibit different behavior. As t decreases, the magnitude of the score in the orthogonal direction becomes large. This would dominate the test error and we would not be able to observe the memorization-to-generalization transition along the subspace. This is the reason why we define directional generalization in (13). Our Theorems B.2 and C.2 compute the parallel and orthogonal contributions to the generalization separately. Figure 6 shows the resulting learning curves for $m = \infty$ and $m = 1$. It is important to note that, as ψ_p increases, the parallel and orthogonal test errors behave differently for small t. This happens because the memorization-to-generalization transition occurs only in the parallel component of the score.

D.3 Plots for different values of λ

In this section, we illustrate the behavior of the test and train errors for different values of λ. Fig. 7 shows the learning errors for $m = \infty$. Note that for small t, the train error increases and the test error decreases as λ increases. This can be explained as follows: In this regime, we have $s^e(t, a_t x_i + \sqrt{h_t} z) \approx -\frac{z}{\sqrt{h_t}}$. However, as regularization increases, $\hat{A}_t$ cannot take very large values. Therefore, $\mathcal{M}_t$ scales as $\frac{1}{h_t}$, making $\mathcal{E}^{\infty}_{\text{train}}(\hat{A}_t) \approx 1$ for very small values of t. This leads to lower memorization as well, as evidenced by the small test error.

Fig. 8 displays the $m = 1$ case. The peak at $\psi_p = \psi_n$ is attenuated as λ increases, which is expected as regularization helps reducing overfitting.

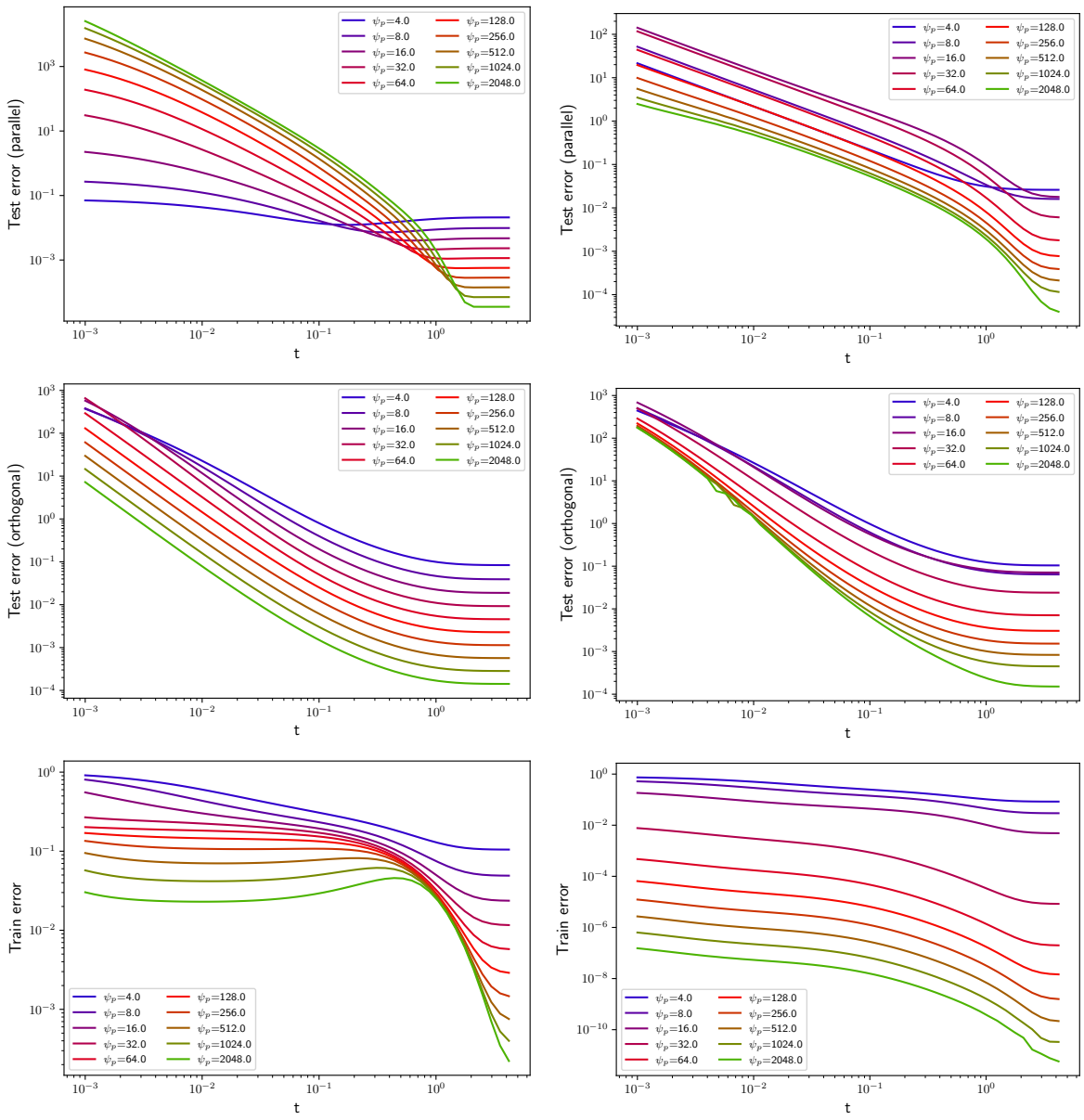
D.4 Plots for different activation functions

We now exemplify the behavior of the learning curves for different activation functions: ReLU, tanh, and sigmoid. To make them have the same L_2 norm and $\mu_0 = 0$, we introduce proper shiftings and rescalings. Concretely:

$$\begin{aligned}
\text{ReLU}(x) &= x\mathbb{1}\{x \geq 0\} - \frac{1}{\sqrt{2\pi}} \,, \\
\tanh(x) &= 0.93 \left(\frac{e^x - e^{-x}}{e^x + e^{-x}} \right) \,, \\
\text{sigmoid}(x) &= \frac{2.8}{1 + e^{-x}} - 1.4 \,.
\end{aligned}$$

Fig. 9 displays the results for $m = \infty$. The activation function enters the analysis through the coefficients μ_1, v, v_0, and s.

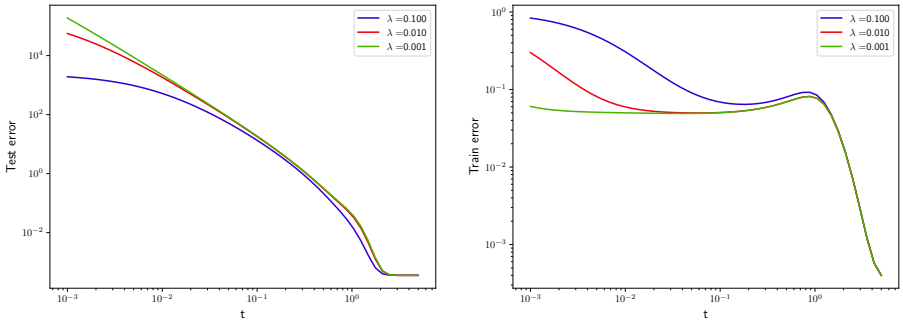
Fig. 10 presents the case $m = 1$.



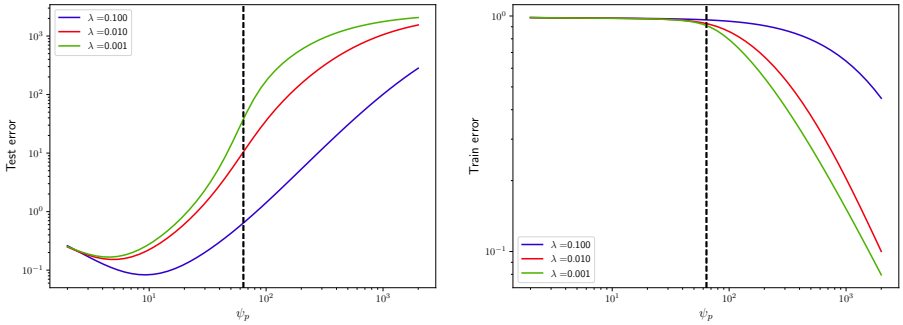
(a) Test and train error for $m = \infty$ as a function of t for different values of ψ_p and a fixed $\psi_n = 20.0$.

(b) Test and train error for $m = 1$ as a function of t for different values of ψ_p and a fixed $\psi_n = 20.0$.

Figure 6: Learning curves for $\psi_D = \frac{D}{d} = 0.2$. We used $\lambda = 0.001$ and the activation function used is ReLU.

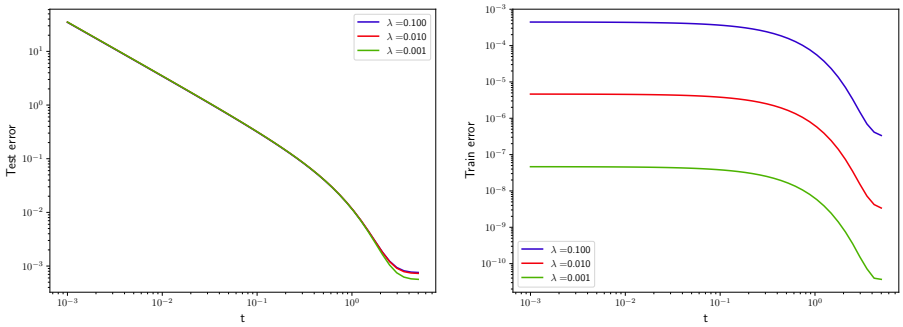


(a) Test and train error as a function of t for different values of λ for a fixed $\psi_n = 20.0$ and $\psi_p = 1024.0$.

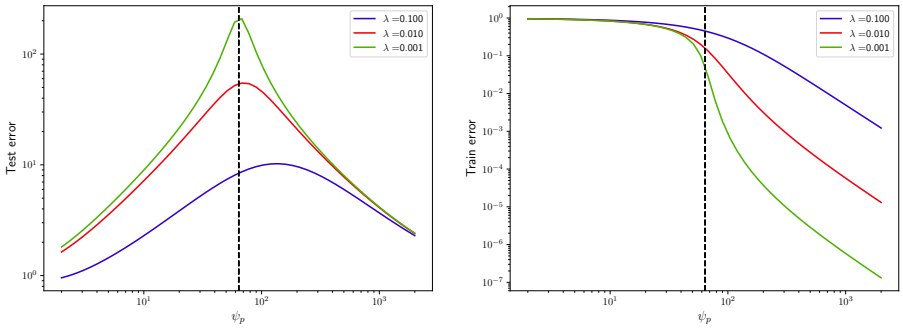


(b) Test and train error as a function of ψ_p for different values of λ for a fixed $\psi_n = 64.0$ and $t = 0.01$. The dashed vertical line indicates $\psi_p = \psi_n$.

Figure 7: Learning curves for different values of λ for $m = \infty$ and with ReLU activation.

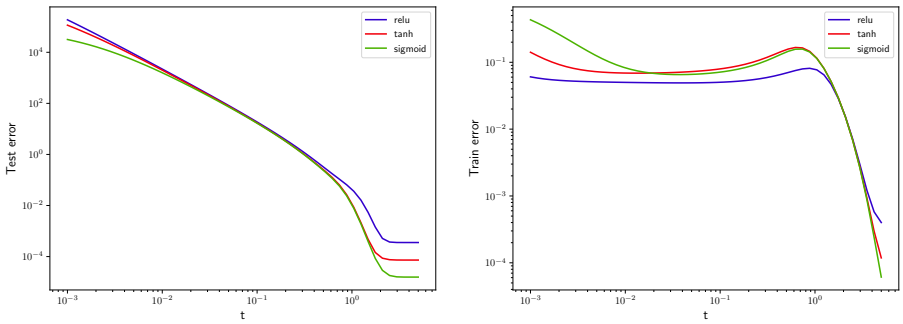


(a) Test and train error as a function of t for different values of λ for a fixed $\psi_n = 20.0$ and $\psi_p = 1024.0$.

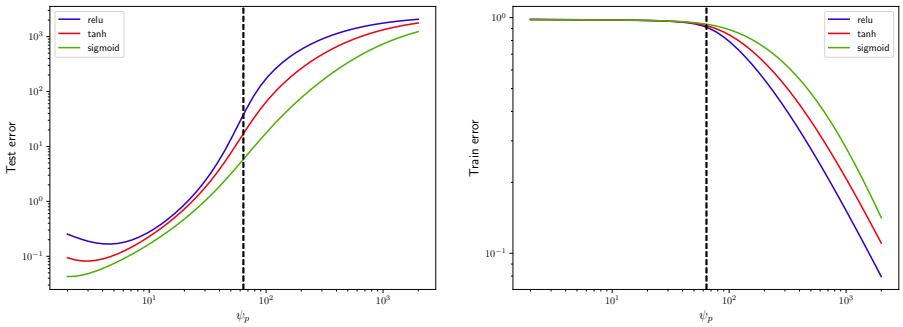


(b) Test and train error as a function of ψ_p for different values of λ for a fixed $\psi_n = 64.0$ and $t = 0.01$. The dashed vertical line indicates $\psi_p = \psi_n$.

Figure 8: Learning curves for different values of λ for $m = 1$. The activation function used is ReLU.

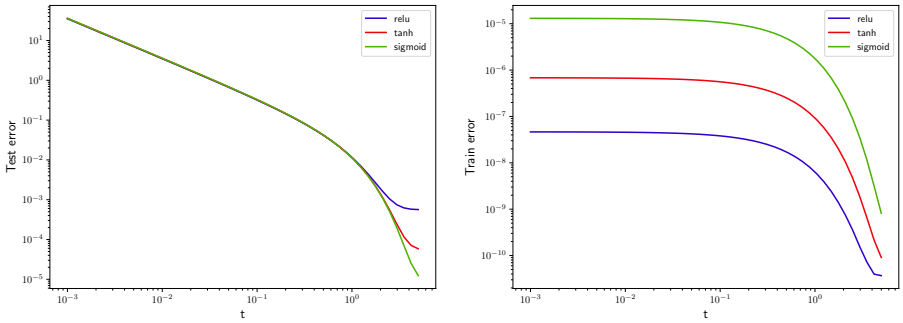


(a) Test and train error as a function of t for different activation functions for a fixed $\psi_n = 20.0$ and $\psi_p = 1024.0$.

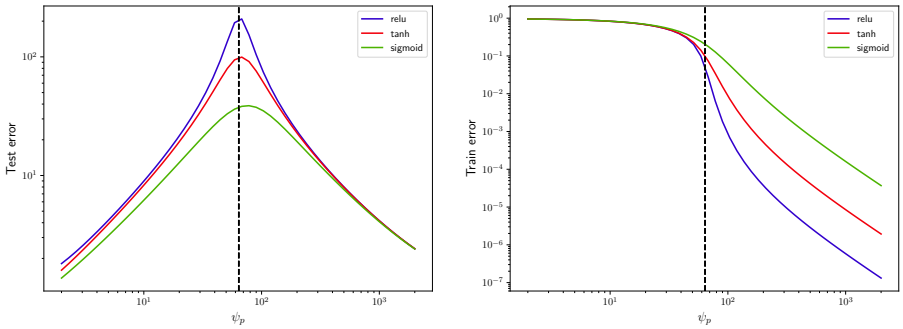


(b) Test and train error as a function of ψ_p for different activation functions for a fixed $\psi_n = 64.0$ and $t = 0.01$. The dashed vertical line indicates $\psi_p = \psi_n$.

Figure 9: Learning curves for different activation functions for $m = \infty$. We used $\lambda = 0.001$.



(a) Test and train error as a function of t for different activation functions for a fixed $\psi_n = 20.0$ and $\psi_p = 1024.0$.



(b) Test and train error as a function of ψ_p for different activation functions for a fixed $\psi_n = 64.0$ and $t = 0.01$. The dashed vertical line indicates $\psi_p = \psi_n$.

Figure 10: Learning curves for different activation functions for $m = 1$. We used $\lambda = 0.001$.

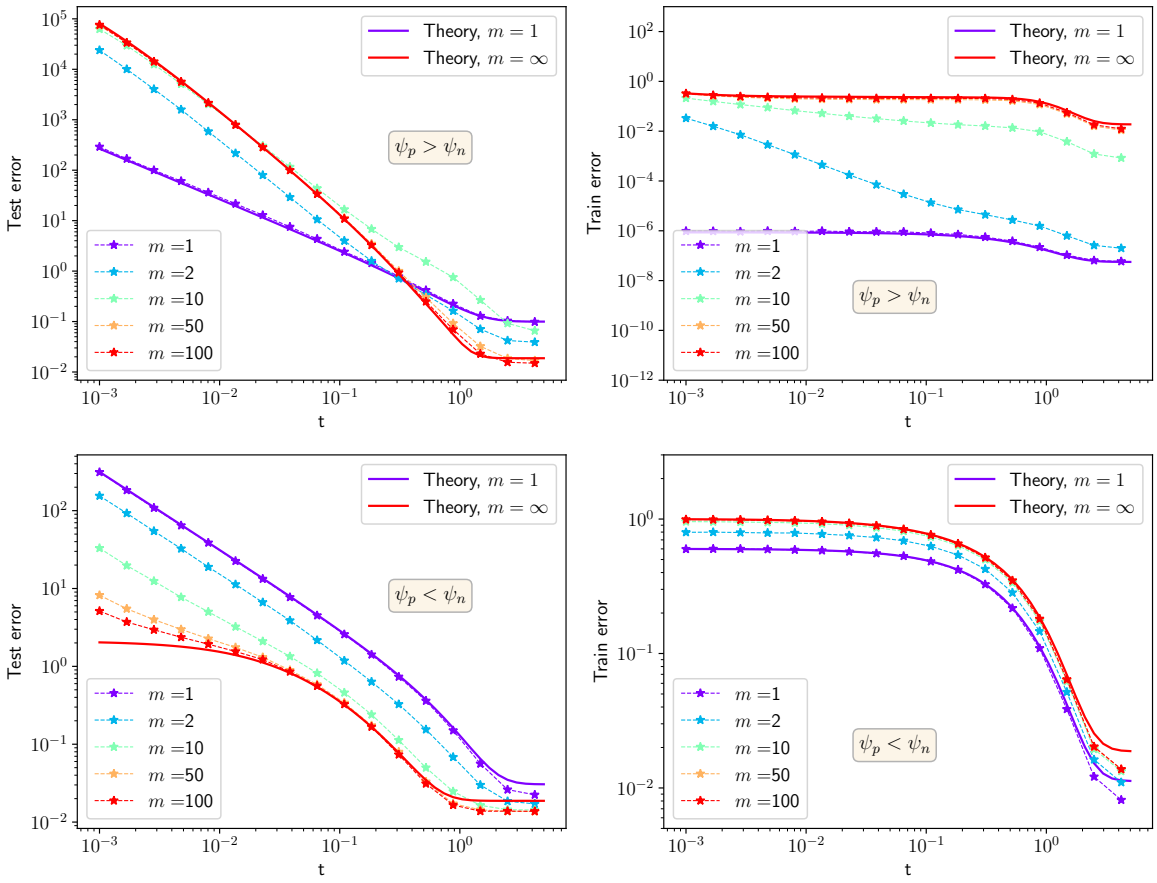


Figure 11: Simulation results ($d = 100$) for different values of m and fixed $\psi_p = 20$; with $\psi_n = 2$ (upper plots) and $\psi_n = 50$ (lower plots). Theoretical results for $m = 1$ and $m = \infty$ are depicted as continuous lines.

E COMPARISON WITH NUMERICALLY OBTAINED LEARNING CURVES

In this section, we discuss the learning curves obtained numerically for different values of m in $\psi_p > \psi_n$ and $\psi_p < \psi_n$ regimes. Fig. 11 presents the plots of test and train errors thus obtained. The upper plot displays the learning errors for the case $\psi_p > \psi_n$, while the lower plot corresponds to $\psi_p < \psi_n$. Based on the previous discussions, memorization is expected when $\psi_p > \psi_n$, and this behavior is evident in Fig. 11 when t is small. Additionally, the extent of memorization increases with m, indicating that large m is detrimental to generalization (small test error) in this regime. This is in contrast to the behavior of test and train errors in the $\psi_p < \psi_n$ regime, where the generalization improves as m increases, evidenced by a decrease in the test error.

The solid lines in Fig. 11 plots the analytical predictions derived in Section 3 for $m = 1$ and $m = \infty$. These predictions align closely with the numerical results for $m = 1$ and $m = 100$ respectively, validating its consistency with the observed data. Minor mismatches between theoretical and numerical curves in the small t and large t regimes can be attributed to finite size effects.

F DETAILS OF REAL DATA EXPERIMENT

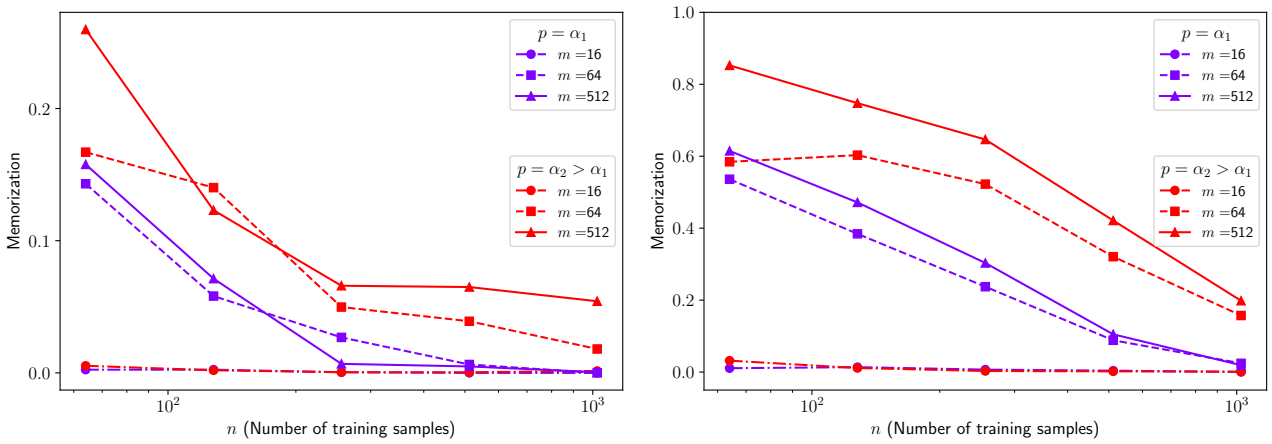
In this section, we provide details of the experiments with MNIST and Fashion-MNIST datasets and a U-Net based score function. This setting is used to obtain the results in Figures 4b, 12 and 13. The U-Net based neural network implementation was adapted from the one by Yang Song available here. We used ADAM Kingma and Ba (2017) for optimizing the U-Net parameters. The hyperparameters used during training are: Epochs = 4000, Batch size = $\min\{64, n\}$, and Learning rate = 0.001. For a fixed m, we draw noise samples $\{z_1, z_2, \cdots, z_m\}$ apriori, and randomly sample from this set to do training step for one batch.

In Fig. 12, we plot the measure of memorization for MNIST and Fashion-MNIST datasets as a function of n for different m, and for two U-Net neural networks with different number of parameters (p). As mentioned in Section 4, we draw the following conclusions from this experiment: 1) Memorization increases as $\frac{n}{p}$ decreases and 2) Memorization increases as m increases in the overparametrized regime. Fig. 13 exhibits samples generated by the U-Net based diffusion model trained on Fashion-MNIST dataset in the above experiment. For each pair n and m, we display samples generated by the diffusion model (top row) alongside their first nearest neighbors in the training dataset (bottom row).

We emphasize that Fig. 13 is presented as an illustration, as the phenomenology is not obvious directly from the images.

G DISCUSSION ON Bonnaire et al. (2025)

Here, we briefly discuss the related work Bonnaire et al. (2025). The authors study how early stopping during training affects memorization. In their theoretical analysis, the score is also parameterized by a random features model that minimizes the loss (10) under an isotropic Gaussian data distribution. Memorization and generalization performances are defined as we do here, with an infinite number of noise samples ($m = +\infty$). Unlike our setting, which seeks the least square estimator, Bonnaire et al. (2025) focuses on gradient flow optimization and analyze generalization and memorization as a function of training time. Their approach is based on studying the spectrum of the same feature covariance matrix (U in Lemma B.1). In particular, two spectral bulks are identified, that control two different time scales, leading to a training window where generalization occurs instead of memorization. Moreover, the width of this window grows with the number of data samples. Our present work corresponds to the regime of infinite training time and is consistent with their findings. From a technical perspective, they use the replica method to derive spectral properties of U whereas we use linear pencil techniques, and a sanity check shows that both methods lead to equivalent sets of algebraic equations determining the resolvent of U.



(a) Results for Fashion-MNIST dataset.

(b) Results for MNIST dataset.

Figure 12: Results of real data experiment on memorization with U-Net score function. We use $N = 2000$ and $\delta = 1/2$.

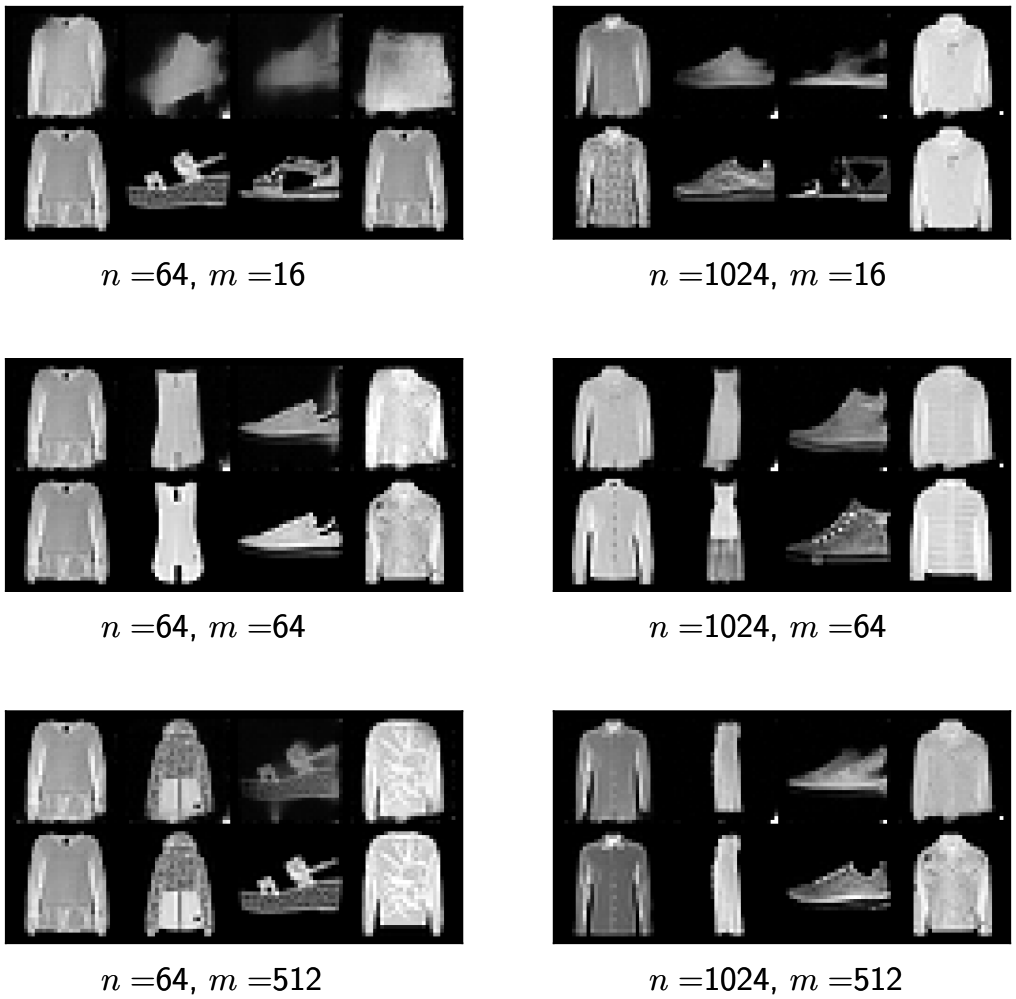


Figure 13: Images generated and their nearest neighbors in the training data set for different values of n, m. For each pair n and m, we display the samples generated by diffusion model (top row) alongside their first nearest neighbors in the training dataset (bottom row).