BENIGN OVERFITTING DOES NOT OCCUR IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benign overfitting and double descent have shaped our understanding of generalization in deep learning, painting a consistent picture: overfitting is not only compatible with good generalization but can actively benefit it. Since diffusion models share much of the machinery of standard deep learning, it is natural to assume that they also exhibit these properties. In this work, we show that this assumption is largely incorrect. We establish a fundamental impossibility result: for any score network architecture, overfitting and good generalization cannot occur simultaneously unless the sample size grows exponentially with the intrinsic dimension of the data. Through a linear model analysis, we trace this to a key difference between regression and score matching: regression benefits from a kernel–target alignment that is absent in score matching, making overfitting irreparably harmful. We then examine mechanisms that prevent overfitting, identifying implicit regularization from time-smoothness in the score, and early stopping in training. Our findings are then supported by high-dimensional experiments. Together, our results indicate that generalization in diffusion models is governed by mechanisms distinct from those of classical settings.

1 INTRODUCTION

Diffusion models are a popular class of deep generative models (Ho et al., 2020; Song et al., 2021b; Sohl-Dickstein et al., 2015). Their key insight is to reframe generation as the reversal of a noising process, characterized by a score function. Estimation of this score function turns out to be equivalent to a denoising problem that can be solved by training a neural network on a simple regression-like objective (Song & Ermon, 2019; Vincent, 2011; Hyvärinen, 2005). Despite their conceptual simplicity, they can achieve state-of-the-art performance in many applications (Dhariwal & Nichol, 2021; Esser et al., 2024; Rombach et al., 2022; Saharia et al., 2022; Watson et al., 2023), and appear in many of the most widely deployed generative AI systems (Yang et al., 2023).

The widespread success of diffusion models naturally raises questions about the mechanisms underpinning their strengths, motivating a growing body of theoretical work on diffusion models (Oko et al., 2023; Azangulov et al., 2024; Benton et al., 2024; Conforti et al., 2025). Of particular interest is the question of generalization (Li et al., 2023; Bonnaire et al., 2025; Farghly et al., 2026b): understanding when a score network trained on finite data faithfully represents the underlying distribution, rather than memorizing the training set. This question is further sharpened by practical concerns around data privacy (Carlini et al., 2023), making a principled understanding of the generalization-memorization tradeoff both scientifically and practically important.

Alongside these developments, an extensive body of theory has sought to explain how overparameterized neural networks generalize. These models famously challenge the classical bias–variance tradeoff by generalizing well despite perfectly memorizing noisy training data (Zhang et al., 2017; Shalev-Shwartz & Ben-David, 2014), a phenomenon known as *benign overfitting* (Bartlett et al., 2020). This is explained by theoretical analysis showing that generalization is maintained by a form of self-induced regularization that occurs in the overparameterized regime. Furthermore, when taken as a function of model complexity, the test error exhibits a *double descent* (Belkin et al., 2019): it first decreases, peaks near the interpolation threshold, and then decreases again as model size grows (see Figure 1, left). This behavior was first identified empirically in high-dimensional problems (Nakkiran et al., 2019) and then theoretically grounded through the analysis of random feature mod-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

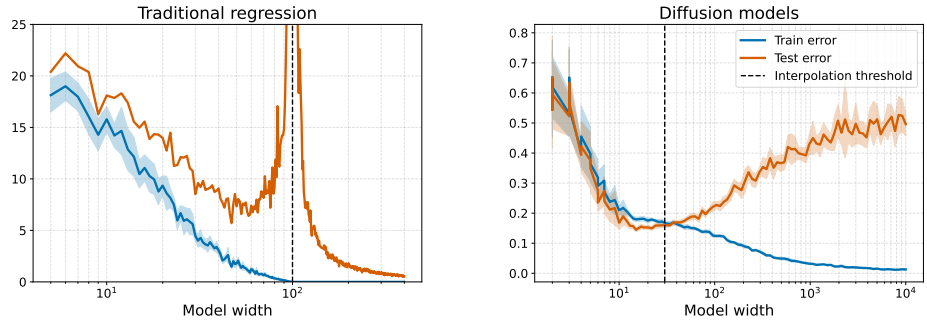

Figure 1: *(Left)* 2-layer random feature network **regression**, displaying the classical double descent curve. *(Right)* 2-layer random feature network **diffusion**, showing the absence of benign overfitting. Details on the setup and hyperparameters for these experiments are available in Section C.1.

els (Mei & Montanari, 2022). Together, these theories form a now widely internalized picture: in highly overparameterized models, fitting the training data perfectly is not only compatible with generalization but, in the right regime, actively beneficial.

One of the benefits diffusion models is that they reuse much of the existing machinery of deep learning, using existing neural network architectures, standard optimization algorithms (e.g., ADAM (Kingma & Ba, 2015)), and fitting by minimizing a regression-like objective (Vincent, 2011; Song et al., 2021b). Thus, it is tempting to conclude that the previously mentioned deep learning phenomena are at work, and benign overfitting and double descent carry over more or less intact. *In this work, we show that such an assumption is largely incorrect.* Several works have hinted that diffusion models may be incapable of benign overfitting (Dupuis et al., 2025; Merger & Goldt, 2025), notably because the empirical score function (the unique minimizer of the empirical loss in function space) leads to the memorization of training data (Farghly et al., 2026b). Other recent works have explored the relationship between over-training and memorization (Pidstrigach, 2022; Bonnaire et al., 2025). Moreover, the idea that memorization and generalization are mutually exclusive was suggested by Yoon et al. (2023) but their analysis is purely empirical and does not explore the presence of the double descent phenomenon. However, the claim that benign overfitting and double descent do not occur in diffusion models is yet to receive rigorous investigation. In particular, a precise quantification of this phenomenon in terms of data dimension, sample size, and noise scale, is lacking.

For diffusion models, we show that, in most practical settings, it is impossible to have both the train and test losses simultaneously small, invalidating benign overfitting. We further show that, instead of a double descent curve, the excess risk forms a U-shaped curve with increasing complexity, akin to the classical bias-variance tradeoff. This is illustrated in the random feature network experiment in Figure 1 where the excess risk forms a U-shape that tapers off as the width of the network grows large. In Figure 2, we see how this phenomenon affects high-dimensional image generation problems with a U-Net architecture (Ho et al., 2020; Ronneberger et al., 2015): as the number of U-Net features is increased, the over-trained diffusion model goes from low quality generation, to generalization, and then towards exactly memorizing the train set (see memorization metrics in Appendix C.2]). Through a combination of fundamental theoretical results, simplified models, and high-dimensional experiments, we explore this phenomenon, the mechanisms underpinning it and how diffusion models generalize in spite of it.

Concretely, we present the following contributions:

- In Section 3, we derive fundamental negative results showing that the population and empirical score matching loss cannot be simultaneously small in practice, unless the sample size grows exponentially with data dimension, regardless of the score network architecture.
- In Section 4, through the analysis of a linear random feature model, we provide a precise understanding of the interplay between model complexity, overfitting and noise scale. Most importantly, by comparing with a classical regression setup, we illustrate the key differences that prevent benign overfitting from occurring in score matching.
- In Section 5, based on the setup of Section 4, we identify key components of score matching that act as implicit regularization and help prevent overfitting. Specifically, we identify the role of the time-smoothness of the score and of early stopping during training.

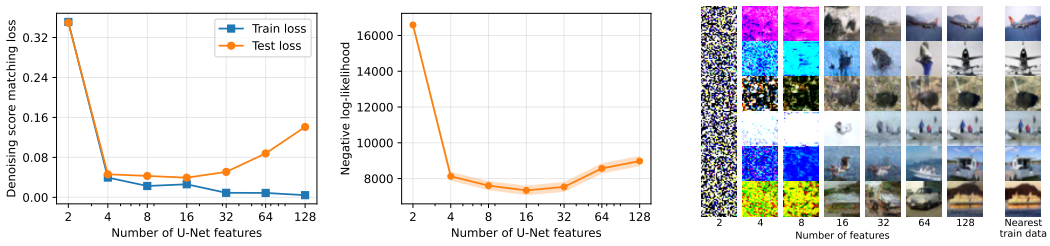

Figure 2: We train a DDPM U-Net model on a subset of CIFAR10 to convergence. We use different configurations of the model, varying the number of features (NF) in the U-Net (and thus, the number of parameters). *(Left)* Test and train error at convergence varying NF. *(Center)* Same setting but for the population negative log-likelihood. *(Right)* Generated samples for different NF values alongside the closest train image for NF $= 8$. Experimental details are given in Appendix C.2.

- We validate our theoretical findings with experiments conducted in high-dimensional image generation settings, as well as more tractable toy settings.

Notations. For probability measures μ and ν, if μ is absolutely continuous *w.r.t.* ν, denoted $\mu \ll \nu$, we denote $\mathrm{d}\mu/\mathrm{d}\nu$ its relative density. If $\mu \ll \nu$, the Kullback-Leibler (KL) divergence is $\mathrm{KL}\left(\mu||\nu\right) = \int \log(\mathrm{d}\mu/\mathrm{d}\nu)\mathrm{d}\mu$, the relative Fisher information is $\mathscr{J}(\mu||\nu) = \int \|\nabla \log(\mathrm{d}\mu/\mathrm{d}\nu)\|^2 \mathrm{d}\mu$, the Fisher information is $\mathscr{J}(\nu) = \int \|\nabla \log(\mathrm{d}\nu/\mathrm{d}x)\|^2 \mathrm{d}\nu$. For $a, b \in \mathbb{R}$, $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, and $a_+ = a \vee 0$.

2 Background

By learning to reverse a noising process, diffusion models approximate a distribution ν on $\mathbb{R}^d$ from a dataset sampled from ν. We briefly review this construction and fix the notation used in the paper.

Noising process. The noising process is given by an SDE of Ornstein–Uhlenbeck-type in $\mathbb{R}^d$,

$$\mathrm{d}X_t = -\kappa X_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathrm{B}_t \;, \qquad X_0 \sim \nu \;, \tag{1}$$

where $\kappa \geq 0$ and B_t is a standard d-dimensional Brownian motion. We assume that $\mathbb{E}_{X\sim\nu}[\|X\|^2] < \infty$. It is well-known that $X_t|X_0 \sim \mathrm{N}(\alpha_t X_0, \sigma_t^2 \mathrm{I}_d)$, where the mean and variance are given by

$$\alpha_t = e^{-\kappa t} \;, \qquad \sigma_t^2 = \kappa^{-1}(1 - \alpha_t^2) \;, \tag{2}$$

with the convention $\sigma_t^2 = 2t$ when $\kappa = 0$. Fixing a time horizon $T > 0$ and writing $Y_t := X_{T-t}$, the classical result of (Haussmann & Pardoux, 1986; Anderson, 1982) identifies the law of the time-reversed process as the solution of

$$\mathrm{d}Y_t = \kappa Y_t \mathrm{d}t + 2\nabla \log p_{T-t}(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathrm{B}_t \;, \qquad Y_0 \sim p_T \;,$$

where p_t denotes the density of X_t. Thus, if we could simulate Y_t starting from $Y_0 \sim p_T$, we would obtain samples $Y_T \sim \nu$. The only unknown quantity is the *score function* $\nabla \log p_t$, which must be learned from data. Once we have an approximate score function $\hat{s} : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$, we generate samples by numerically simulating

$$\mathrm{d}\widehat{Y}_t = \kappa \widehat{Y}_t \mathrm{d}t + 2\hat{s}(\widehat{Y}_t, t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathrm{B}_t \;, \qquad Y_0 \sim q_0 \;.$$

For $\kappa > 0$, we set $q_0 = \mathrm{N}(0, \kappa^{-1}\mathrm{I}_d)$ as it approximates p_T for large T, and for $\kappa = 0$, we set $q_0 = \mathrm{N}(0, 2T\mathrm{I}_d)$. We denote by q_t the probability density of $\widehat{Y}_t$ for $t \in [0, T]$.

Score matching. The score function is approximated using a deep neural network. Ideally, the goal is to minimize the *(population) explicit score matching (ESM) loss*,

$$\mathscr{L}_{\mathrm{ESM}}(s, t) := \mathbb{E}\left[\left\|s(t, X_t) - \nabla \log p_t(X_t)\right\|^2\right] \;, \qquad \mathscr{L}_{\mathrm{ESM}}(s, \varpi) = \int \mathscr{L}_{\mathrm{ESM}}(s, t)\, \mathrm{d}\varpi(t) \;,$$

where ϖ is a non-negative measure on $[0, T]$. While the time-weighting ϖ can be chosen freely, a canonical choice is the *evidence lower bound (ELBO) weighting*, $\varpi^{\epsilon}_{\mathrm{ELBO}}(\mathrm{d}t) = \mathbb{1}_{[\epsilon,T]}(t)$, for $\epsilon > 0$. In this case, the ESM loss upper bounds the marginal KL divergence with $\mathrm{KL}\left(p_\epsilon||q_{T-\epsilon}\right) \leq$

$\mathscr{L}_{\mathrm{ESM}}(s, \varpi^{\epsilon}_{\mathrm{ELBO}}) + \mathrm{KL}\left(p_T || q_0\right)$. Since the ESM loss depends on the intractable score $\nabla \log p_t$, a tractable alternative is considered: the *(population) denoising score matching (DSM) loss*,

$$\mathscr{L}_{\mathrm{DSM}}(s, \varpi) := \int \mathbb{E}[\ell_t(s, X_0)] \mathrm{d}\varpi(t) \ , \quad \ell_t(s, X_0) := \mathbb{E}\big[\big\| s(t, X_t) - \nabla \log p_{t|0}(X_t | X_0) \big\|^2 \big| X_0\big] \ ,$$

where $p_{t|0}(\cdot|\cdot)$ is the density of X_t given X_0. This loss is equal to the ESM loss up to a constant whilst only using the simpler conditional score $\nabla \log p_{t|0}$, turning the learning problem into a denoising problem (Vincent, 2011; Hyvärinen, 2005). It is then estimated using a dataset $S = \{Z_i\}_{i=1}^n \sim \nu^{\otimes n}$, leading to the *empirical DSM loss*,

$$\widehat{\mathscr{L}}_{\mathrm{DSM}}(s, \varpi) := \frac{1}{n} \sum_{1 \leq i \leq n} \int \ell_t(s, Z_i) \, \mathrm{d}\varpi(t) \ , \qquad n \geq 1 \ . \tag{3}$$

We also define the empirical ESM loss, related to the above by an additive data-dependent constant,

$$\widehat{\mathscr{L}}_{\mathrm{ESM}}(s, t) := \mathbb{E}\big[\big\| s(t, \widehat{X}_t) - \nabla \log \hat{p}_t(\widehat{X}_t) \big\|^2 \big| S\big] \ , \quad \widehat{\mathscr{L}}_{\mathrm{ESM}}(s, \varpi) = \int \widehat{\mathscr{L}}_{\mathrm{ESM}}(s, t) \, \mathrm{d}\varpi(t) \ , \tag{4}$$

where $\mathrm{d}\widehat{X}_t = -\kappa \widehat{X}_t \mathrm{d}t + \sqrt{2}\mathrm{dB}_t$, $\widehat{X}_0 \sim n^{-1} \sum_{i=1}^n \delta_{Z_i}$ and for $t > 0$, $\hat{p}_t$ is the density of $\widehat{X}_t$.

3 Fundamental limitations of overfitting in diffusion models

Benign overfitting occurs when good generalization performance is achieved despite interpolating training data. In this section, we derive impossibility results that show that this cannot happen in diffusion models without an exceedingly large sample size. Unlike Sections 4 and 5, the results presented here hold *regardless of the score network architecture*. Thus, they identify fundamental limitations of the score matching framework itself, rather than of any particular class of models.

3.1 Warm-up: A Fisher information lower bound

We begin with the following simple lemma, which is a generic lower bound on the sum of the population and empirical ESM losses. The proof can be found in Section B.1.1.

Lemma 3.1. *Let $t \in (0, T]$, for any measurable score function $s : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$, we have*

$$2\mathbb{E}[\widehat{\mathscr{L}}_{\mathrm{ESM}}(s, t) + \mathscr{L}_{\mathrm{ESM}}(s, t)] \geq \mathbb{E}\left[\mathscr{J}\left(\hat{p}_t || p_t\right)\right] \ . \tag{5}$$

In particular, if $\mathscr{J}(\nu) < +\infty$, we have that $\inf_s \mathbb{E}[\widehat{\mathscr{L}}_{\mathrm{ESM}}(s, t) + \mathscr{L}_{\mathrm{ESM}}(s, t)] \to +\infty$ as $t \to 0^+$.

Lemma 3.1 shows that, for a fixed score network (independent of the data), the empirical and population score matching can only be simultaneously small when the empirical and population measures $\hat{p}_t$ and p_t are close also. The measure of closeness here is the relative Fisher information, which under additional concentration assumptions, can be shown to upper bound the KL divergence (see Section B.1.2). As $\hat{p}_t$ is initialized at the empirical distribution $n^{-1} \sum_{i=1}^n \delta_{Z_i}$ where $(Z_1, \ldots, Z_n) \sim \nu^{\otimes n}$, this suggests that benign overfitting only happens when the sample size is large enough. We make this argument precise in the next subsection.

3.2 A quantitative lower bound under time-integration

We now refine this result by obtaining lower bounds on the amount of data required for the empirical and population losses to be small simultaneously and by allowing the score function to depend on data. The lower bound is a function of the effective dimension of the data distribution, which we capture through the *lower Rényi dimension*, defined in Section A.1. We also consider the full time-integrated loss as opposed to an individual time-slice, requiring only that the weighting ϖ has positive non-decreasing density. The bound is generic and holds for *any random score function $\hat{s}$* that can depend on training data. Given a real-valued function w, we define $w^-(r) := \inf\{t : w(t) \geq r\}$.

Theorem 3.2. *Suppose that ϖ is a positive measure supported in $[\epsilon, T]$ with $\epsilon > 0$ and non-decreasing density $w : [\epsilon, T] \to \mathbb{R}_+$, that ν has finite second moments, and that both T is sufficiently large. Suppose that the random score function $\hat{s}$ satisfies*

$$\mathbb{E}_S[\widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi)] \leq \varepsilon \ ,$$

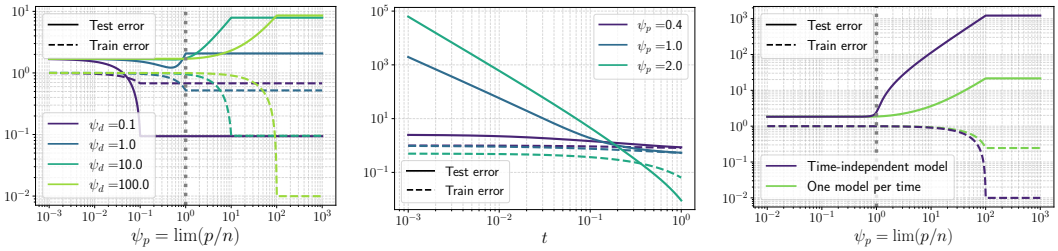

Figure 3: Linear random features under Assumption 4.1. *(Left)* time is fixed at $t = 10^{-1}$ and ψ_p varies. *(Center)* ψ_d is fixed at 10^2 and t varies. *(Right)* time integrated losses (Section 5.1) with π the uniform distribution on $[0, T]$. `Train error` denotes the empirical DSM loss with the weighting used in training in practice. `Test error` is the population ESM loss with ELBO weighting.

for some $\varepsilon \leq w(T)/32$. Then, for any $c \in (0,1)$, if any of the following conditions hold:

$$\mathbb{E}_S[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi)] \leq \varepsilon, \quad \textit{or} \quad \mathbb{E}_S[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi^{\epsilon}_{\mathrm{ELBO}})] \leq \tfrac{1}{32}, \quad \textit{or} \quad \mathrm{KL}\left(p_\delta || q_{T-\delta}\right) \leq \tfrac{1}{32}, \tag{6}$$

and $\delta := w^{-}(32\varepsilon)$ is sufficiently small, it follows that,

$$n \geq \left(2\left(\sqrt{d} + \sqrt{\log(64)}\right)\alpha_\delta^{-1}\sigma_\delta\right)^{-cd'},$$

where d' is the lower Rényi dimension of ν, defined in Section A.1.

The assumption on the measure ϖ, encompasses several standard choices of time-weighting within the *continuous training* framework (Song et al., 2021b). In the following remark, we consider several of these, obtaining closed-form bounds.

Remark 3.3. *Under the ELBO weighting $\varpi = \varpi^{\epsilon}_{\mathrm{ELBO}}$, the sample size lower bound becomes $n \geq (2\left(\sqrt{d} + \sqrt{\log(64)}\right)\alpha_\epsilon^{-1}\sigma_\epsilon)^{-cd'}$ for ϵ sufficiently small. Under the variance weighting $w(t) = \sigma_t^2$, commonly used for training (Ho et al., 2020), the analogous bound is $n \geq (4(32\varepsilon \vee \sigma_\epsilon)\sqrt{d})^{-cd'}$ for ε, ϵ sufficiently small. Most strikingly, when $\epsilon = 0$ and $\varepsilon \leq \frac{1}{32}\limsup_{t\to 0^+} w(t)$, there is no sample size at which benign overfitting can occur: $\mathbb{E}[\widehat{\mathscr{L}_{\mathrm{ESM}}}(\hat{s}, \varpi)] \leq \varepsilon$ necessarily implies $\mathbb{E}[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi)] > \varepsilon$, $\mathbb{E}[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi^{0}_{\mathrm{ELBO}})] > 1/32$, and $\mathrm{KL}\left(p_\delta || q_{T-\delta}\right) > 1/32$.*

We highlight that the lower bound is increasing exponentially in the intrinsic dimension and depends on ϵ, supporting both the idea that manifold distributed data is beneficial to diffusion models and that early stopping in the reverse process can benefit generalization (Farghly et al., 2026a; Bortoli, 2022; Farghly et al., 2026b; Azangulov et al., 2024). Since many popular implementations take $\epsilon \in \{10^{-3}, 10^{-5}\}$, the expectation that the amount of data scales with $\epsilon^{-d'/2}$ is unreasonably strong. To make this precise, we provide in Section B.1 a numerical evaluation of our bounds in relevant settings, including our experiments. Thus, in most practical setups, diffusion models will not exhibit the property of benign overfitting.

The proof, presented in Section B.1.3, uses a similar approach as Lemma 3.1, showing that for empirical and population losses to be small, $\hat{p}_t$ and p_t must be close. Instead of the Fisher information, we derive a connection to the total variation via an argument based on Girsanov's theorem.

4 Comparison with traditional regression on linear RFNNs

Section 3 establishes at a general level that benign overfitting cannot occur in diffusion models in practical settings. However, it does not reveal how overfitting manifests, nor how it depends on the interplay between model size, data dimension, and noise scale. To get a finer understanding, we now analyze a tractable model class: two-layer linear random feature networks (Rahimi & Recht, 2007), similar to the linear activation case of the models considered in (George & Macris, 2026; George et al., 2025; Bonnaire et al., 2025). Despite the simplicity of the linear setting, it has been central to the theoretical analysis of benign overfitting and double descent in traditional regression (Belkin et al., 2019; Hastie et al., 2022). Therefore, we rely on this linear model to understand how diffusion models differ from traditional regression and what is preventing benign overfitting to occur.

Setup. We study 2-layer linear networks $s_{W,A} : x \mapsto AWx$, where $W \in \mathbb{R}^{p\times d}$ is fixed and $A \in \mathbb{R}^{d\times p}$ is a trainable parameter. We fix $t \in (0, T]$ and consider the minimization of the empirical risk $A \mapsto \sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)$. Let $\mathbf{\Sigma}$ be the covariance matrix of the data distribution ν and $\widehat{\mathbf{\Sigma}} := n^{-1}\sum_{i=1}^n Z_i Z_i^T$ be the empirical covariance matrix. Let $\mathbf{\Sigma}_t := \alpha_t^2\mathbf{\Sigma}+\sigma_t^2\mathrm{I}_d$ and $\widehat{\mathbf{\Sigma}}_t := \alpha_t^2\widehat{\mathbf{\Sigma}}+\sigma_t^2\mathrm{I}_d$.

In the following lemma, we characterize the DSM and ESM loss associated with the empirical risk minimization problem. The proof can be found in Section B.2.

Lemma 4.1. *Assume that W has full rank and take $\widehat{A} \in \arg\min \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)$. Let $\Omega_t := \widehat{\mathbf{\Sigma}}_t^{-1}$ if $p \geq d$ and $\Omega_t := W^T(W\widehat{\mathbf{\Sigma}}_t W^T)^{-1}W$ if $p < d$. Then, $\widehat{A}W = -\Omega_t$ and,*

$$\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, t) = d - \sigma_t^2 \mathrm{Tr}(\Omega_t)\ , \qquad \mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) = \|(\Omega_t - \mathbf{\Sigma}_t^{-1})\mathbf{\Sigma}_t^{1/2}\|_{\mathrm{F}}^2\ .$$

For $p \geq d$, the ESM and DSM loss at $\widehat{A}$ are independent of p. By the proof of Lemma 4.1, this is because the empirical risk minimization (ERM) problem becomes equivalent to the ERM associated with a d-dimensional linear model (independent of p), similar to (Merger & Goldt, 2025). See Section B.2 for details. We specify our setup further in the next assumption.

Assumption 4.1. *$W \in \mathbb{R}^{p\times d}$ has $\mathrm{N}(0,1)$ i.i.d. entries, $A \in \mathbb{R}^{d\times p}$ and $\nu = \mathrm{N}(0, \beta\mathrm{I}_d)$, with $\beta > 0$.*

The next proposition is a precise characterization of overfitting in a proportional asymptotic regime.

Proposition 4.2 (Overfitting in linear networks)**.** *Under Assumption 4.1, we consider the limit $d, n, p \to \infty$ such that $p/n \to \psi_p$ and $d/n \to \psi_d$. When $\psi_d, \psi_p > 1$, we have, almost-surely,*

$$\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) \to \frac{1}{\alpha_t^2\beta + \sigma_t^2}\left(\frac{\psi_p \wedge \psi_d - 1}{\psi_d}\frac{\alpha_t^4\beta^2}{\sigma_t^4} + \left(1 - \frac{\psi_p}{\psi_d}\right)_+\right) + \mathcal{O}\left(\frac{1}{\psi_d}\right)\ , \tag{7}$$

where $\widehat{A} \in \arg\min \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)$. For the empirical risk, we have

$$\frac{\sigma_t^2}{d}\widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, t) \to \frac{1}{\psi_d} + \left(1 - \frac{\psi_p}{\psi_d}\right)_+ + \mathcal{O}\left(\frac{1}{\psi_d(\psi_p \wedge \psi_d)}\right)\ . \tag{8}$$

Proposition 4.2 demonstrates how the population loss is a "U-shaped" function of model complexity, captured by the relative model width, ψ_p. When $\psi_p > \psi_d$, we observe that the empirical risk of (8) is of order $\mathcal{O}\left(\psi_d^{-1}\right)$ and can become arbitrarily small and interpolate training data when the data dimension is large. In the same regime, we observe that the population score matching loss is of order σ_t^{-4} and explodes as $t \to 0^+$. This can also be observed in Figure 3, displaying the exact asymptotics of the risks, obtained in the proofs in Section B.2. It demonstrates that interpolation and generalization are mutually incompatible. We also study the dependence on t in Figure 3, showing that the rate of explosion of the excess risk as $t \to 0^+$ depends on the model complexity.

Comparison with regression. The results above differ strikingly from traditional regression. To better understand what is the fundamental difference between the two settings, we consider a multi-output regression problem with the same data and architecture. Given a target $B_\star \in \mathbb{R}^{d\times d}$ and noiseless labels $Y_i := B_\star Z_i$, we minimize the empirical risk $A \mapsto n^{-1}\sum_{i=1}^n \|s_{W,A}(Z_i) - Y_i\|^2 + \lambda\|AW\|_{\mathrm{F}}^2$. We focus on the overparameterized regime $p > d > n$ and $\lambda = 0$, where minimizers A satisfy $(AW - B_\star)\widehat{\mathbf{\Sigma}} = 0$. Since $\widehat{\mathbf{\Sigma}}$ has rank $n < d$ a.s., this gives a $d(d-n)$-dimensional affine subspace of solutions for AW, contrasting sharply with Lemma 4.1, where score matching produces a unique value of $AW = -\widehat{\mathbf{\Sigma}}_t^{-1}$ even without regularization.

The excess risk for the regression problem is given by $\|(AW - B_\star)\mathbf{\Sigma}^{1/2}\|_{\mathrm{F}}^2$. It has a similar form as the ESM loss in Lemma 4.1 once replacing $B_\star$ with $-\mathbf{\Sigma}_t^{-1}$, and $\mathbf{\Sigma}$ with $\mathbf{\Sigma}_t$ in the weighting. To resolve the under-determination, we consider the ridgeless limit $\lambda \to 0^+$, recovering the minimum-norm interpolator. This projects AW on the orthogonal of $\ker(\widehat{\mathbf{\Sigma}})$, leading to the excess risk $\|B_\star \Pi_{\ker(\widehat{\mathbf{\Sigma}})}\mathbf{\Sigma}^{1/2}\|_{\mathrm{F}}^2$, where $\Pi_{\ker(\widehat{\mathbf{\Sigma}})}$ is the column-wise projection onto the kernel of $\widehat{\mathbf{\Sigma}}$. Thus, generalization can occur in the overparameterized regime due to two factors: (i) the minimum-norm bias projects AW on the orthogonal of $\ker(\widehat{\mathbf{\Sigma}})$; (ii) the residual is small when $B_\star$ *aligns* with $\widehat{\mathbf{\Sigma}}^{1/2}$ within the kernel of $\widehat{\mathbf{\Sigma}}$. This is made precise in the benign overfitting literature that identifies a form of *self-induced regularization* resulting from alignment (Bartlett et al., 2020; Hastie et al., 2022).

The picture is fundamentally different for score matching. First, since the empirical problem exactly determines the solution $AW = -\widehat{\mathbf{\Sigma}}_t^{-1}$, there can be no minimum-norm bias that projects AW orthogonally to $\ker(\widehat{\mathbf{\Sigma}})$. Moreover, whereas $B_\star$ could align with $\mathbf{\Sigma}^{1/2}$ on the null space in such a way that makes their product small, it is less likely that $-\mathbf{\Sigma}_t^{-1}$ aligns in the same way with $\mathbf{\Sigma}_t^{1/2}$. Indeed, if $\mathbf{\Sigma}$ is not full rank, we have $\|\mathbf{\Sigma}_t^{-1}\Pi_{\ker(\widehat{\mathbf{\Sigma}})}\mathbf{\Sigma}_t^{1/2}\|_{\mathrm{F}}^2 \geq (d-n)\sigma_t^{-2}$, which explodes as $t \to 0$. This is made precise in the proof of Proposition 4.2, leading to the first term of (7), corresponding to the contribution of $\ker(\widehat{\mathbf{\Sigma}})$ to the ESM loss. These conclusions easily extend in the presence of label noise, even though score matching was linked to regression with label noise (Han et al., 2024).

5 Preventing overfitting in diffusion models

The absence of benign overfitting in diffusion models suggests that regularization is required to achieve good generalization in the overparameterized regime. In this section, we show that the score matching framework contains forms of *implicit regularization* that help prevent overfitting.

5.1 Implicit regularization due to time smoothness

In most practical applications of diffusion models, a single network is trained across multiple noise levels. Whilst primarily done for efficiency, the resulting time-smoothness of the score function has been shown empirically to benefit generalization (Li et al., 2025). Here, we study its impact in the setup of Section 4, in the extreme case where the score network is time-independent.

Setup. We parameterize the score network as in Section 4 with $p > d$ and we consider a time weighting ϖ which is a probability distribution on $[0, T]$ such that we can define the probability measure $\pi(\mathrm{d}t) \propto \sigma_t^{-2}\varpi(\mathrm{d}t)$. For any distribution $\omega \in \{\pi, \varpi\}$ on $[0, T]$, we write

$$\alpha_\omega^2 := \int \alpha_t^2 \mathrm{d}\omega(t)\ , \quad \sigma_\omega^2 := \int \sigma_t^2 \mathrm{d}\omega(t)\ , \quad \mathbf{\Sigma}_\omega := \int \mathbf{\Sigma}_t \mathrm{d}\omega(t)\ , \quad \widehat{\mathbf{\Sigma}}_\omega := \int \widehat{\mathbf{\Sigma}}_t \mathrm{d}\omega(t)\ .$$

In the next lemma, we characterize the empirical risk minimization in this setting.

Lemma 5.1. *Assume that $p > d$ and W is full-rank. Then, any $\widehat{A} \in \arg\min_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W}, \varpi)$ satisfies $\widehat{A}W = -\widehat{\mathbf{\Sigma}}_\varpi^{-1}$ and we have*

$$\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W}, \pi) = \|(\widehat{\mathbf{\Sigma}}_\varpi^{-1} - \mathbf{\Sigma}_\pi^{-1})\mathbf{\Sigma}_\pi^{1/2}\|_{\mathrm{F}}^2 + \mathrm{C}_\pi\ , \quad \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{\widehat{A},W}, \varpi) = d\sigma_\pi^{-2} - \mathrm{Tr}(\widehat{\mathbf{\Sigma}}_\varpi^{-1})\ ,$$

where $\mathrm{C}_\pi \geq 0$ is given in the proof in Equation (14). C_π is 0 only when π is a Dirac distribution.

Remark 5.2 (Equivalence with ridge)**.** *By Lemma 5.1, we have $\widehat{A}W = -\widehat{\mathbf{\Sigma}}_\varpi^{-1}$. As ϖ is a probability distribution, there exists $t_\varpi > 0$ such that $\alpha_{t_\varpi}^2 = \alpha_\varpi^2$. Thus, for any $t < t_\varpi$, we have*

$$\widehat{\mathbf{\Sigma}}_\varpi = \eta_{t,\varpi}^{-1}(\widehat{\mathbf{\Sigma}}_t + 2\lambda_{t,\varpi}\mathrm{I}_d)\ , \quad 2\lambda_{t,\varpi} := \eta_{t,\varpi}(\sigma_\varpi^2 - \sigma_t^2) > 0\ , \quad \eta_{t,\varpi} := \alpha_t^2/\alpha_\varpi^2\ ,$$

and we see that $\eta_{t,\varpi}^{-1}\widehat{A} \in \arg\min_A(\widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W}, t) + \lambda_{t,\varpi}\|AW\|^2)$. Therefore, in this case, considering the time-integrated loss is formally similar to adding ridge regularization for times $t < t_\varpi$.

Next, we make this observation precise by characterizing the asymptotic score matching losses.

Proposition 5.3. *Under Assumption 4.1, we consider the limit $d, n, p \to \infty$ such that $p/n \to \psi_p$ and $d/n \to \psi_d$. Let $\widehat{A} \in \arg\min_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, \varpi)$. When $\psi_p > \psi_d > 1$, we have, almost-surely,*

$$\frac{1}{d}\widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, \varpi) \to \frac{1}{\sigma_\pi^2} - \frac{1}{\sigma_\varpi^2} + \mathcal{O}(\psi_d^{-1}), \quad \frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, \pi) \to \mathrm{L}_\pi + \mathcal{O}(\psi_d^{-1}),$$

where $\mathrm{L}_\pi \geq 0$ is a constant whose expression is given in Equation (17), and $\sigma_\pi^{-2} - \sigma_\varpi^{-2} \geq 0$.

Proposition 5.3 shows that the time smoothness prevents the limiting empirical score matching loss from being arbitrarily small as soon as ϖ is not a Dirac distribution. This contrasts with the fixed noise scale case of Proposition 4.2, even when ψ_d is large, showing that the time smoothness prevents overfitting. Moreover, while in Proposition 4.2 the asymptotic ESM loss could explode as $t \to 0^+$,

in Proposition 5.3 the ESM loss stays bounded, showing that the time smoothness acts as a form of implicit regularization, as hinted in Remark 5.2.

In the proof of Proposition 5.3 (see Section B.3), we derive exact asymptotics for the population and empirical losses, which we report in the right figure in Figure 3 (green curves). Using the asymptotics of the previous section, we compare it against an overfit linear score network that has independent weights at each time, corresponding to the completely non-time-smooth case (purple curves). We observe that the time smoothness increases the empirical risk while decreasing the population loss, showing that it is indeed acting as an implicit regularization mechanism.

We verify this in a practical image generation setting similar to Figure 2, but considering a time-independent DDPM U-Net model trained across a range $[t, t+r]$ for varying r. We evaluate train and test error at t to see how changing r affects the error. We present the results in Figure 4 where we find that increasing r increases the train error at t and therefore is indeed preventing overfitting. Interestingly, the test error first decreases and then increases for large r, suggesting a tradeoff on r.

5.2 Implicit regularization via early stopping

Next, we study the impact of early stopping during the training of the score network. To this end, we fix a time $t \in [0, T]$ and study the dynamics of the parameter A under gradient flow,

$$\dot{A}(\tau) = -p^{-1}\nabla_A(\sigma_t^2 \widehat{\mathscr{L}_{\mathrm{DSM}}}(s_{W,A(\tau)}, t)) \;, \quad \tau \geq 0 \;, \tag{9}$$

where $W \in \mathbb{R}^{p\times d}$ is as in Section 4, and the p^{-1} factor is ensuring a finite limit as $p \to \infty$ (Bonnaire et al., 2025). Without loss of generality, we assume that $A(0) = 0$. We show in Section B.4 that in the limit $p \to \infty$ with (n, d) fixed, the matrix $A(\tau)W$ converges to a finite limit following the gradient flow of a linear model (see Section B.4). Moreover, when $\tau \to \infty$, $A(\tau)W \to -\widehat{\mathbf{\Sigma}}_t^{-1}$, which leads to the same score matching bounds as in Proposition 4.2 for $p > d$. In the next proposition, we show that for small enough values of τ, we avoid the explosion of the test loss in Proposition 4.2.

Proposition 5.4. *Suppose that Assumption 4.1 holds and define $\mathcal{E}_t(\tau, n, d, p) := \mathscr{L}_{\mathrm{ESM}}(s_{W,A(\tau)}, t)$. Then, $\mathcal{E}_t(\tau, n, d, p)$ has an almost-sure limit $\overline{\mathcal{E}}_t(\tau, n, d) := \lim_{p\to\infty} \mathscr{L}_{\mathrm{ESM}}(s_{W,A(\tau)}, t)$. Moreover, in the limit as $n, d \to \infty$ with $\lim(d/n) = \psi_d > 1$, we have, almost-surely*

$$\frac{1}{d}\overline{\mathcal{E}}_t(\tau, n, d) \longrightarrow \left(1 - \frac{1}{\psi_d}\right) \frac{\left((\alpha_t^2\beta + \sigma_t^2)\sigma_t^{-2}(1 - e^{-2\tau\sigma_t^4}) - 1\right)^2}{\alpha_t^2\beta + \sigma_t^2} + \mathcal{O}\left(\frac{1}{\psi_d}\right) \;,$$

In particular, if $\tau = \mathcal{O}\left(\sigma_t^{-2}\right)$, then the limit above is bounded as $t \to 0^+$.

This result shows that early stopping prevents the explosion of the test loss as $t \to 0$ given that the early stopping τ is of order σ_t^{-2}. This suggests that, despite not being benign, overfitting appears later in training, allowing early stopping to prevent it, which aligns with recent results (Bonnaire et al., 2025; Merger & Goldt, 2025). The advantages of early stopping can also be observed in high-dimensional experiments: in Figure 5, we report the evolution of the test and train errors for a U-Net across iterations, observing that there is a unique optimal stopping time where the ESM loss is controlled as the number of parameters grows.

6 Related works & Conclusion

Random features analysis & linear models. While we focus on quantifying overfitting and comparing with traditional regression, we also derive exact asymptotics of the losses in the proof of Proposition 4.2, complementing prior analyses (George & Macris, 2026; George et al., 2025; Merger & Goldt, 2025). As discussed in Section 4, the two-layer linear RFNNs analysis recovers linear diffusion models as a particular case, extending the regularized linear model analysis of (Merger & Goldt, 2025). In particular, we recover the conclusion that overfitting is driven by the null space. Recent work also studied the learning dynamics of diffusion models (Biroli et al., 2024; Merger & Goldt, 2025), including RFNN settings (Bonnaire et al., 2025; Li et al., 2023). Section 5 supplements these approaches by quantifying the role of early stopping and time smoothness in mitigating overfitting in linear networks. Overall, we provide a fine-grained characterization of overfitting in

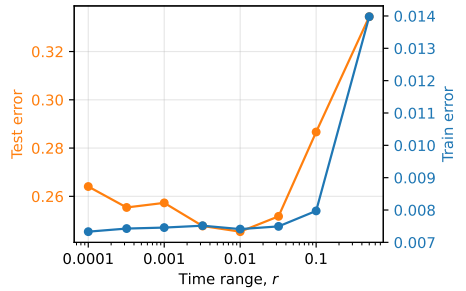

Figure 4: Time-independent DDPM model evaluated at $t = 0.1$. We train it on the range $[t, t+r]$ and observe how increasing r can benefit generalization at time t. Details in Appendix C.2.

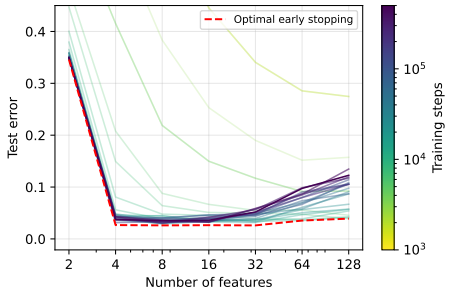

Figure 5: The same setting as Figure 2 but we observe how the test error curve changes with training steps. We find that early stopping produces a test error that does not grow with NF.

linear RFNNs and, crucially, *(i)* provide a precise analysis of how diffusion models **differ from traditional regression** (Section 4), and *(ii)* incorporate the **time-smoothness** of the models with time-integrated losses (Section 5.1), which is beyond the scope of prior work.

Comparison with Bonnaire et al. (2025). These authors show that diffusion models trained by gradient methods typically generalize before memorizing the training set. Our results differ in both scope and focus. Their theoretical analysis is restricted to two-layer RFNNs and characterizes when memorization and generalization occur during gradient-flow training at a fixed noise scale. In contrast, our results in Section 3 apply to generic score estimators and show, independently of the architecture and training procedure, that overfitting and generalization cannot occur simultaneously. The closest connection is our analysis in the short subsection Section 5.2: whereas Bonnaire et al. (2025) identify time scales in the learning dynamics, we derive exact asymptotics for the score-matching error, providing a complementary characterization.

Diffusion model generalization. In addition to the aforementioned works, there has been a lot of recent interest in the generalization properties of diffusion models (Kadkhodaie et al., 2024; Li et al., 2023; Chakraborty et al., 2026). In particular, minimax statistical rates have been obtained (Oko et al., 2023; Zhang et al., 2024; Azangulov et al., 2024), the relationship with data geometry has been explored (Farghly et al., 2026a; Pidstrigach, 2022; He et al., 2026) and algorithm-dependent generalization bounds were proposed for diffusion models (Dupuis et al., 2025; Farghly et al., 2026b). Our work offers a new but complementary perspective, by suggesting that the mechanisms causing generalization in diffusion models are fundamentally different from classical settings.

Statistical lower bounds. Recent work analyzed memorization through lower bounds (Buchanan et al., 2025). In particular, Lemma 3.1 can be compared with (Ye et al., 2026), who showed that $\mathscr{J}(\hat{p}_t||p_t) = \widehat{\mathscr{L}}_{\mathrm{DSM}}(\nabla \log p_t, t) - \widehat{\mathscr{L}}_{\mathrm{DSM}}(\nabla \log \hat{p}_t, t)$. This is different from the left-hand side of Equation (5), which is the sum of the empirical and population ESM losses for a *generic score network*. The lower bounds on $\mathscr{J}(\hat{p}_t||p_t)$ derived by Ye et al. (2026) for mixture data could still easily be combined with Lemma 3.1. Our results in Section 3.2 differ in that they consider a very generic setting and exploit the intrinsic dimension of the data in statistical guarantees. Achilli et al. (2025, Section 6) also compare the true score and the empirical score, while our setting in Section 3.2 is more general and applies to a broad class of score estimators. Note that (5) does not require any restrictive assumptions on ν, such as the log-Sobolev inequality used in (Koehler et al., 2023).

Conclusion. In this work, we showed that the phenomenon of benign overfitting, known to occur in many deep learning settings, does not occur in diffusion models. We first derive a fundamental limitation of score matching, showing that the empirical and population losses cannot be small without an exceedingly large training set. Then, through a combination of high-dimensional experiments and a fine-grained analysis in a simplified setting, we showed that the double descent phenomenon of deep learning does not occur and that the test error follows a U-shaped curve with respect to model complexity. Finally, by incorporating the time-smoothness of the score and early-stopped gradient dynamics into our experiments and linear network analysis, we demonstrated that these key components of score estimation can help prevent overfitting.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in checking our proofs. We have *not* used generative AI tools for developing theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, writing proofs, designing the research methodology, or implementing methods, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing software code, summarizing existing literature, and improving the readability of the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work is largely theoretical and we do not believe that it has any negative ethical or societal impact.

REPRODUCIBILITY STATEMENT

To ensure reproductbility of our experiments, we have included in appendix C a detailed discussion of the experimental setting and all the hyperparameters that were used to obtain the plots presented in the paper. This covers both our low- and high-dimensional experiments. All the models and dataets used in our experiments have also been disclosed. Moreover, the source code for our experiments has been submitted as part of the supplementary material.

REFERENCES

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and Generalization in Generative Diffusion under the Manifold Hypothesis. *ArXiv preprint 2502.09578*, 2025.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions, 2024.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, December 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, August 2019.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15, 2024.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models, 2025.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023.

Djalil Chafai. Entropies, convexity, and functional inequalities. *Kyoto Journal of Mathematics*, 44 (2), January 2004.

Saptarshi Chakraborty, Quentin Berthet, and Peter L. Bartlett. Generalization properties of score-matching diffusion models for intrinsically low-dimensional data, 2026.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Kl convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 1 edition, July 2022. doi: 10.1017/9781009128490.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Benjamin Dupuis, Dario Shariatian, Maxime Haddouche, Alain Durmus, and Umut Simsekli. Algorithm- and data-dependent generalization bounds for score-based generative models. In *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Kenneth Falconer. *Fractal Geometry - Mathematical Foundations and Applications*. Wiley, 2014.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a.

Tyler Farghly, Patrick Rebeschini, George Deligiannidis, and Arnaud Doucet. Implicit regularisation in diffusion models: An algorithm-dependent generalisation analysis. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Anand Jerry George and Nicolas Macris. Asymptotic Learning Curves for Diffusion Models with Random Features Score and Manifold Data, March 2026.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising Score Matching with Random Features: Insights on Diffusion Models from Precise Learning Curves. *arXiv preprint at arXiv:2502.00336*, 2025.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=h8GeqOxtd4.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4): 1188–1205, 1986.

Ye He, Yitong Qiu, and Molei Tao. Diffusion model's generalization can be characterized by inductive biases toward a data-dependent ridge manifold. *arXiv [stat.ML]*, February 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiao Li, Zekai Zhang, Xiang Li, Siyi Chen, Zhihui Zhu, Peng Wang, and Qing Qu. Understanding representation dynamics of diffusion models via low-dimensional modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107365–107404, 2025.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457–483, 1967.

Pertti Mattila, Manuel Moran, and Jose-manual Rey. Dimension of a measure. *Studia mathematica 142 (3)*, 2000.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A Three Sample Hypothesis Test for Evaluating Generative Models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 3546–3556. PMLR, 2020.

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

Claudia Merger and Sebastian Goldt. Generalization dynamics of linear diffusion models. In *EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM)*, 2025.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. *ICLR 2020*, December 2019.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 26517–26582. PMLR, 23–29 Jul 2023.

Felix Otto and Cédric Villani. Comment on: "Hypercontractivity of Hamilton–Jacobi equations", by S. Bobkov, I. Gentil and M. Ledoux. *Journal de Mathématiques Pures et Appliquées*, 80(7): 697–700, 2001.

Ya. B. Pesin. On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions. *Journal of Statistical Physics*, 71(3):529–547, May 1993. ISSN 1572-9613. doi: 10.1007/BF01058436.

Yakov B. Pesin. *Dimension Theory in Dynamical Systems - Contemporary Views and Applications*. Chicago Lectures in Mathematics. The University of Chicago Press, 1997.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35852–35865. Curran Associates, Inc., 2022.

I. Procaccia, P. Grassberger, and H. G. E. Hentschel. On the characterization of chaotic motions. In *Dynamical Systems and Chaos*, volume 179 of *Lecture Notes in Physics*, pp. 212–222, Berlin, 1983. Springer. Proceedings of the conference held in Sitges/Barcelona, 1982.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. ISSN 0899-7667, 1530-888X. doi: 10.1162/NECO_a_00142.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), 2023.

Zeqi Ye, Qijie Zhu, Molei Tao, and Minshuo Chen. Provable separations between memorization and generalization in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion Probabilistic Models Generalize when They Fail to Memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60134–60178. PMLR, 21–27 Jul 2024.

A Additional Technical Background

A.1 Intrinsic dimensions

Our main results in Section 3 are expressed in terms of the *Rényi* dimension of a Borel measure μ on $\mathbb{R}^d$. This notion was first developed by (Procaccia et al., 1983) and then rigorously treated in (Pesin, 1997; 1993). It is a standard notion in the theory of chaotic dynamical systems. We provide below the definition.

Definition A.1 (Rényi dimension of a measure)**.** *Let μ be a Borel probability measure on $\mathbb{R}^d$. The upper Rényi dimension of μ is defined by*

$$\overline{\gamma}(\mu) := \limsup_{\delta \to 0} \left(\frac{1}{\log \delta} \log \left(\int \mu(B(x,\delta)) \mathrm{d}\mu(x) \right) \right) .$$

Similarly, the lower Rényi dimension of μ is defined by

$$\underline{\gamma}(\mu) := \liminf_{\delta \to 0} \left(\frac{1}{\log \delta} \log \left(\int \mu(B(x,\delta)) \mathrm{d}\mu(x) \right) \right) .$$

Remark A.2. *The definition above is actually a particular case of the larger family of Rényi dimension, where it actually corresponds to the Rényi dimension of order 2. Note that Definition A.1 is also sometimes called the* correlation dimension *in the literature (Pesin, 1993).*

Remark A.3. *In particular, if a measure is supported on a smooth manifold and is Ahlfors regular (see (Mattila et al., 2000; Falconer, 2014)) on this manifold, then the Rényi dimension of Definition A.1 is equal to the manifold dimension.*

A.2 Random covariance matrices

We recall here some basic random matrix theory results related to random covariance matrices and the Marchenko-Pastur theorem. We invite the reader to consult (Couillet & Liao, 2022) for additional details.

Given a symmetric matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, we denote its spectrum by $\mathrm{Spec}(\mathbf{\Sigma})$ (where the eigenvalues are counted with their multiplicity). The spectral measure of Σ is defined by

$$\widehat{\mu}_{\mathbf{\Sigma}} := \frac{1}{d} \sum_{\lambda \in \mathrm{Spec}(\mathbf{\Sigma})} \delta_\lambda \, . \tag{10}$$

We recall below the celebrated Marchenko-Pastur theorem (Marchenko & Pastur, 1967). To this end, we first define the Marchenko-Pastur distribution below.

Definition A.4 (Marchenko-Pastur distribution)**.** *Let $\psi > 0$. The Marchenko-Pastur distribution with shape parameter ψ, denoted here $\mu_{\mathrm{MP}}^{(\psi)}$, is the probability distribution on $\mathbb{R}_+$ defined by*

$$\mu_{\mathrm{MP}}^{(\psi)}(\mathrm{d}x) := \left(1 - \frac{1}{\psi}\right)_+ \delta_0 + \frac{\sqrt{(\psi^+ - x)(x - \psi^-)}}{2\pi\psi x} \mathbb{1}_{[\psi^-, \psi^+]}(x) \mathrm{d}x \, ,$$

where $y_+ := \max(0, y)$, $\psi^- = |1 - \sqrt{\psi}|^2$, and $\psi^+ := (1 + \sqrt{\psi})^2$.

The Marchenko-Pastur's theorem is given below.

Theorem A.5 (Marchenko-Pastur's theorem)**.** *Let X be random vector $\mathbb{R}^d$ with i.i.d. components with zero mean, unit variance, and uniformly bounded moments[1] of order $4 + \epsilon$ for some $\epsilon > 0$. Define the empirical covariance matrix,*

$$\widehat{\mathbf{\Sigma}} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \, ,$$

with $X_1, \ldots, X_n$ i.i.d. copies of X. Consider that $d = d_n$ with $\lim_{n \to \infty}(d_n/n) = \psi > 0$. Then, with probability one, the empirical measure $\widehat{\mu}_{\widehat{\mathbf{\Sigma}}}$ converges weakly to the Marchenko-Pastur distribution with shape parameter ψ.

[1] We refer to Theorem 2.4 of (Couillet & Liao, 2022) for a discussion on the different conditions required by the Marchenko-Pastur theorem.

B OMITTED PROOFS

In this section, we present the omitted proofs of the results from Sections 3 to 5.

B.1 OMITTED PROOFS OF SECTION 3

B.1.1 PROOF OF LEMMA 3.1.

We present below the proof of Lemma 3.1.

Proof. Let $S := (Z_1, \dots, Z_n) \sim \nu^{\otimes n}$ be a dataset sampled from the data distribution. Let $(\widehat{X}_t)_{t\geq 0}$ the Ornstein-Uhlenbeck process $\mathrm{d}\widehat{X}_t = -\kappa \widehat{X}_t \mathrm{d}t + \sqrt{2}\mathrm{dB}_t$ initialized at the empirical data distribution $\widehat{X}_0 \sim \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$. We denote by $\hat{p}_t$ the probability density of $\widehat{X}_t$, conditioned on $S := (Z_1, \dots, Z_n)$. Let $s : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary measurable score function. By definition of the Fisher information and the triangle inequality (note that the expectation is over both the noise and the dataset), we have, for any $t > 0$,

$$\begin{aligned}
\mathbb{E}\left[\mathscr{J}\left(\hat{p}_t || p_t\right)\right] &= \mathbb{E}\left[\left\|\nabla \log \hat{p}_t(\widehat{X}_t) - \nabla \log p_t(\widehat{X}_t)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla \log \hat{p}_t(\widehat{X}_t) - s(t, \widehat{X}_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|s(t, \widehat{X}_t) - \nabla \log p_t(\widehat{X}_t)\right\|^2\right] \\
&= 2\mathbb{E}\left[\left\|\nabla \log \hat{p}_t(\widehat{X}_t) - s(t, \widehat{X}_t)\right\|^2\right] + 2\mathbb{E}\left[\left\|s(t, X_t) - \nabla \log p_t(X_t)\right\|^2\right] ,
\end{aligned}$$

where the last equality follows as $\mathrm{law}(X_t)$ is the marginal distribution of $\widehat{X}_t$ under $S \sim \nu^{\otimes n}$. We immediately deduce that for any $t > 0$, we have

$$\mathbb{E}\left[\mathscr{J}\left(\hat{p}_t || p_t\right)\right] \leq 2\mathbb{E}\left[\mathscr{L}_{\mathrm{ESM}}(s,t) + \widehat{\mathscr{L}}_{\mathrm{ESM}}(s,t)\right] .$$

For the second part of the statement, by the triangle inequality, we have that

$$\mathbb{E}\left[\mathscr{J}\left(\hat{p}_t || p_t\right)\right] = \int_{\mathbb{R}^d} \mathbb{E}\left[\left\|\nabla \log \frac{\hat{p}_t}{p_t}(x)\right\|^2 \hat{p}_t(x)\right] \mathrm{d}x \geq \frac{1}{2}\mathbb{E}\left[\mathscr{J}\left(\hat{p}_t\right)\right] - \mathscr{J}\left(p_t\right) .$$

Note that $\mathscr{J}\left(p_t\right) \leq \alpha_t^{-2}\mathscr{J}\left(\nu\right) < +\infty$ under our assumptions. Then, the fact that $\mathbb{E}\left[\mathscr{J}\left(\hat{p}_t\right)\right] \to +\infty$ as $t \to 0^+$ follows from standard arguments. For the sake of completeness, let us provide a short proof in the case $\kappa = 0$ (without loss of generality, the other cases follow by the same computation by using the invariant distribution of the forward process instead of the Lebesgue measure as a reference measure). Let $h(f) := -\int f(x)\log f(x)\mathrm{d}x$ denote the entropy functional, as soon as it is well-defined. As a consequence of the De-Bruijn's identity and the fact that the Fisher information is non-increasing along the semigroup (Bakry et al., 2014), we have, for $t > s > 0$, that

$$h(\hat{p}_t) - h(\hat{p}_s) = \int_s^t \mathscr{J}\left(\hat{p}_u\right)\mathrm{d}u \leq (t-s)\,\mathscr{J}\left(\hat{p}_s\right) .$$

Given $(Z_1, \dots, Z_n) \sim \nu^{\otimes n}$, we easily see that

$$-h(\hat{p}_s) \geq -\frac{1}{n}\sum_{i=1}^n h(f_{Z_i,\sigma_s}) - \log(n) = -\frac{d}{2}\log(2\pi e) + \frac{d}{2}\log(\sigma_s^{-2}) - \log(n) ,$$

where $f_{\mu,\sigma}$ is the density of $\mathrm{N}(\mu, \sigma \mathrm{I}_d)$. Thus, if we fix t and let $s \to 0^+$, $\mathbb{E}\left[\mathscr{J}\left(\hat{p}_s\right)\right] \to +\infty$.

This concludes the proof. □

Remark B.1. *As an immediate consequence, for any positive Borel measure ϖ on $[0,T]$, we have*

$$\mathbb{E}\left[\mathscr{L}_{\mathrm{ESM}}(s,\varpi) + \widehat{\mathscr{L}}_{\mathrm{ESM}}(s,\varpi)\right] \geq \frac{1}{2}\int_{[0,T]} \mathbb{E}\left[\mathscr{J}\left(\hat{p}_t || p_t\right)\right] \mathrm{d}\varpi(t) .$$

B.1.2 Lower bound under log-Sobolev inequalities

In this subsection, we present some additional results to complement Lemma 3.1. Our goal is to show that, under the assumption that the data distribution satisfies a logarithmic Sobolev inequality, we can obtain a similar lower bound to Lemma 3.1, where the relative Fisher information term can be replaced by other divergences, such as the total variation distance.

Conventions for probabilistic inequalities. We define the total variation distance as

$$\mathrm{TV}(\mu, \nu) := \sup_A |\mu(A) - \nu(A)| \, .$$

We say that a Borel probability distribution ν satisfies the log-Sobolev inequality with constant ρ if for all differentiable $f \in \mathrm{L}^1(\nu)$, we have

$$\mathrm{Ent}_\nu(f) \leq \frac{\rho}{2} \int \frac{\|\nabla f\|^2}{f} \mathrm{d}\nu \, .$$

Equipped with these definitions, we have the following proposition.

Proposition B.2. *Assume that the data distribution satisfies the log-Sobolev inequality with constant ρ_0. Let ϖ be a positive Borel measure on $[0, T]$, we have*

$$\mathbb{E}\left[\mathscr{L}_{\mathrm{ESM}}(s, \varpi) + \widehat{\mathscr{L}}_{\mathrm{ESM}}(s, \varpi)\right] \geq \int_{[0,T]} \frac{2}{\alpha_t^2 \rho_0 + \sigma_t^2} \mathbb{E}\left[\mathrm{TV}(p_t, \hat{p}_t)^2\right] \mathrm{d}\varpi(t) \, .$$

Proof. By stability of the log-Sobolev inequality under Lipschitz mappings and convolutions (Chafai, 2004), we have that p_t satisfies the log-Sobolev inequality with constant $\rho_t := \alpha_t^2 \rho_0 + \sigma_t^2$. Therefore, for any $t > 0$, we have

$$\begin{aligned}
\frac{1}{2} \mathscr{J}\left(\hat{p}_t || p_t\right) &\geq \frac{1}{\alpha_t^2 \rho_0 + \sigma_t^2} \mathrm{KL}\left(\hat{p}_t || p_t\right) \quad \text{(log-Sobolev inequality)} \\
&\geq \frac{2}{\alpha_t^2 \rho_0 + \sigma_t^2} \mathrm{TV}(\hat{p}_t, p_t)^2 \quad \text{(Pinsker's inequality)} \, .
\end{aligned}$$

The result immediately follows by integrating over ϖ and applying Lemma 3.1. □

Remark B.3. *We immediately see that, if $\kappa > 0$, we have,*

$$\mathbb{E}\left[\mathscr{L}_{\mathrm{ESM}}(s, \varpi) + \widehat{\mathscr{L}}_{\mathrm{ESM}}(s, \varpi)\right] \geq \frac{2}{\max(\kappa^{-1}, \rho_0)} \int_{[0,T]} \mathbb{E}\left[\mathrm{TV}(p_t, \hat{p}_t)^2\right] \mathrm{d}\varpi(t) \, .$$

B.1.3 Proof of Theorem 3.2

Before giving the proof of Theorem 3.2, we provide some technical lemmas below. Recall that we denote by $p_{t|0}(\cdot|x)$ the conditional density of X_t given $X_0 = x$, defined by Equation (2).

Lemma B.4. *Given any $x \in \mathbb{R}^d$ and $Y \sim p_{\epsilon|0}(\cdot|x)$, we have that for any $\Delta \geq 0$,*

$$\begin{aligned}
\mathbb{P}(\|Y - \alpha_\epsilon x\| \geq \sigma_\epsilon m + \Delta) &\leq \exp\left(-\frac{\Delta^2}{2\sigma_\epsilon^2}\right) , \\
\mathbb{P}(\|Y - \alpha_\epsilon x\| \leq \sigma_\epsilon m - \Delta) &\leq \exp\left(-\frac{\Delta^2}{2\sigma_\epsilon^2}\right) ,
\end{aligned}$$

where $m = \mathbb{E}_{Z \sim \mathrm{N}(0, \mathrm{I}_d)}[\|Z\|]$.

Proof. By definition, we have that the random variable $Z = \sigma_\epsilon^{-1}(Y - \alpha_\epsilon x)$ is a standard multivariate Gaussian. Thus, by Gaussian concentration of measure (see e.g., Section 5.4.2 of (Bakry et al., 2014)), we have that for any 1-Lipschitz function f,

$$\mathbb{P}(f(Z) - \mathbb{E}[f(Z)] \geq t) \leq \exp(-t^2/2) \, .$$

Choosing $f(z) = \|z\|$ and $f(z) = -\|z\|$ along with $t = \Delta/\sigma_\epsilon$. □

The proof of Theorem 3.2 is based on the following proposition, providing a lower bound on the sample size, provided that two KL divergences associated to the population and empirical forward processes are both smaller than an absolute constant. Below, we use the following convention and notation for the total variation distance:

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_A |\mu(A) - \nu(A)| \,, \tag{11}$$

for probability measure μ and ν. We also recall that we denote by p_t the probability density of X_t following Equation (1) and by $\hat{p}_t$ the probability density of the process $\mathrm{d}\widehat{X}_t = -\kappa \widehat{X}_t \mathrm{d}t + \sqrt{2}\mathrm{dB}_t$ initialized at $\widehat{X}_0 \sim \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, for $t > 0$, where $S := (Z_1, \dots, Z_n) \sim \nu^{\otimes n}$ is the dataset.

Proposition B.5. *Let $\epsilon > 0$ and $c \in (0,1)$. For any distribution $\hat{\nu}$ on $\mathbb{R}^d$, which might depend on the dataset S, if $\mathbb{E}\left[\mathrm{KL}\left(\hat{p}_\epsilon || \hat{\nu}\right)\right] \leq 1/16$ and $\mathbb{E}\left[\mathrm{KL}\left(p_\epsilon || \hat{\nu}\right)\right] \leq 1/16$, then we have $n \geq n_{\min}(\epsilon)$, with, for ϵ small enough*

$$\log n_{\min}(\epsilon) \geq -c \log\left(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d})\right) \underline{\gamma}(\nu) \,,$$

where $\underline{\gamma}(\nu)$ is the lower Rényi dimension defined in Definition A.1.

Proof of Proposition B.5. We begin with Pinsker's inequality and the reverse triangle inequality for the total variation distance, which are used to obtain

$$\begin{aligned}
\mathbb{E}\left[\mathrm{KL}\left(p_\epsilon || \hat{\nu}\right)\right] &\geq 2\mathbb{E}[\|p_\epsilon - \hat{\nu}\|_{\mathrm{TV}}^2] \\
&\geq 2\mathbb{E}\Big[\left(\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}} - \|\hat{p}_\epsilon - \hat{\nu}\|_{\mathrm{TV}}\right)^2\Big] \\
&\geq 2\left(\sqrt{\mathbb{E}\left[\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}}^2\right]} - \sqrt{\mathbb{E}\left[\|\hat{p}_\epsilon - \hat{\nu}\|_{\mathrm{TV}}^2\right]}\right)^2 \,,
\end{aligned}$$

where the final inequality follows from the triangle inequality in weighted L^2-norm. Since, we have $2\mathbb{E}[\|\hat{p}_\epsilon - \hat{\nu}\|_{\mathrm{TV}}^2] \leq \mathbb{E}\left[\mathrm{KL}\left(\hat{p}_\epsilon || \hat{\nu}\right)\right] \leq 1/16$, we conclude from this that if $\mathbb{E}\left[\mathrm{KL}\left(p_\epsilon || \hat{\nu}\right)\right] \leq 1/16$ then $\mathbb{E}[\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}}^2] \leq \frac{1}{8}$.

We now identify a test set to lower bound the total variation using that for any Borel set A,

$$\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}} \geq |p_\epsilon(A) - \hat{p}_\epsilon(A)| \,.$$

As in Lemma B.4, we define $m = \mathbb{E}_{Z\sim \mathrm{N}(0,\mathrm{I}_d)}[\|Z\|]$. We define the set $A_R := \{y \in \mathbb{R}^d : y \notin B_{\sigma_\epsilon m + R}(\alpha_\epsilon Z_i), \forall i = 1, \dots, n\}$ so that the empirical density is upper bounded by

$$\hat{p}_\epsilon(A_R) \leq \frac{1}{N}\sum_{i=1}^N p_{\epsilon|0}(B_{m\sigma_\epsilon + R}(\alpha_\epsilon Z_i)^{\complement} | Z_i) = 1 - p_{\epsilon,R} \,,$$

where $p_{\epsilon,R} := p_{\epsilon|0}(B_{m\sigma_\epsilon+R}(\mathbf{0})|\mathbf{0})$. Define the set

$$\tilde{A}_R := \{y \in \mathbb{R}^d : \|y - Z_i\| \geq 2(R + m\sigma_\epsilon)/\alpha_\epsilon, \forall i = 1, \dots, n\} \,.$$

The population density is lower bounded using that for any $x \in \tilde{A}_R$, we have $B_{m\sigma_\epsilon+R}(\alpha_\epsilon x) \subset A_R$ and hence,

$$\begin{aligned}
p_\epsilon(A_R) &\geq \int_{\tilde{A}_R} p_{\epsilon|0}(A_R|x)\nu(\mathrm{d}x) \\
&\geq \int_{\tilde{A}_R} p_{\epsilon|0}(B_{m\sigma_\epsilon+R}(\alpha_\epsilon x)|x)\nu(\mathrm{d}x) \\
&= \nu(\tilde{A}_R) p_{\epsilon,R} \,.
\end{aligned}$$

We further bound this using the union bound,

$$\begin{aligned}
\mathbb{E}_S[\nu(\tilde{A}_R)] &= 1 - \mathbb{E}_S\left[\mathbb{P}_{X\sim\nu}\left(\bigcup_{i=1}^n \{\|X - x_i\| < 2(R + m\sigma_\epsilon)/\alpha_\epsilon\}\right)\right] \\
&\geq \left(1 - \sum_{i=1}^n \mathbb{E}[\mathbb{P}_{X\sim\nu}(\|X - x_i\| < 2(R + m\sigma_\epsilon)/\alpha_\epsilon)]\right)_+ \\
&= (1 - n\delta(2(R + m\sigma_\epsilon)/\alpha_\epsilon))_+ \,,
\end{aligned}$$

where we define $\delta(r) = \mathbb{E}_{X\sim\nu}[\nu(B_r(X))]$. The total variation is then lower bounded using A_R as a test set:

$$\mathbb{E}[\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}}] \geq \mathbb{E}[p_\epsilon(A_R) - \hat{p}_\epsilon(A_R)] \geq \left(\left(1 - n\delta\left(\frac{2(R+m\sigma_\epsilon)}{\alpha_\epsilon}\right)\right)_+ p_{\epsilon,R} - (1-p_{\epsilon,R})\right)_+ .$$

To guarantee that this bound is non-trivial, we now choose R using Lemma B.4. Setting $R^2 = 2\log(8)\sigma_\epsilon^2$, leads to the bound,

$$p_{\epsilon,R} \geq 1 - \exp\left(-\frac{R^2}{2\sigma_\epsilon^2}\right) \geq \frac{3}{4} .$$

We also observe that

$$\frac{2(R+m\sigma_\epsilon)}{\alpha_\epsilon} \leq 2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d}),$$

and we choose ϵ so that $2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d}), 1$.

Thus, by Jensen's inequality, we have the lower bound,

$$\mathbb{E}[\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}}^2]^{1/2} \geq \frac{1}{4}\Big(3(1 - n\delta(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d})))_+ - 1\Big)_+ .$$

From this bound we deduce that,

$$\begin{aligned}\mathbb{E}\left[\|p_\epsilon - \hat{p}_\epsilon\|_{\mathrm{TV}}^2\right] \leq \frac{1}{8} &\implies (3(1 - n\delta(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d})))_+ \leq 1 + \sqrt{2} \\ &\implies n \geq \frac{2-\sqrt{2}}{3\,\delta(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d}))} =: n_{\min}(\epsilon) .\end{aligned}$$

By the definition of the lower Rényi dimension in Definition A.1, we have that

$$\underline{\gamma}(\nu) = \liminf_{\epsilon\to 0^+} \frac{\log\delta(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d}))}{\log(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d}))} .$$

For any fixed $\eta > 0$, by definition, we have that there exists ϵ_0 such that, for all $\epsilon < \epsilon_0$, we have

$$\log n_{\min}(\epsilon) \geq \log\left(\frac{2-\sqrt{2}}{3}\right) - \log\left(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d})\right)(\underline{\gamma}(\nu) - \eta) .$$

In particular, by setting $\eta = \frac{1-c}{2}\underline{\gamma}(\nu)$, if $\underline{\gamma}(\nu) > 0$ and ϵ is sufficiently small, we have that

$$\log n_{\min}(\epsilon) \geq -c\log\left(2\alpha_\epsilon^{-1}\sigma_\epsilon(\sqrt{\log(64)} + \sqrt{d})\right)\underline{\gamma}(\nu) . \tag{12}$$

Moreover, if $\underline{\gamma}(\nu) = 0$, then the above inequality is always verified. This concludes the proof. □

We can now present the proof of Theorem 3.2.

Proof of Theorem 3.2. Suppose that $\mathbb{E}_{X\sim\nu}[\|X\|^2] = \sigma^2 < \infty$, $T \geq 1 + \frac{1}{2\kappa}\log(32(\sigma^2 + \kappa^{-1}d))$, and $d \geq 5$. We begin by expressing the weighted ESM loss in terms of the standard ESM loss by using the fact that the density w of ϖ is non-decreasing on its support. We have, for $\delta \geq \epsilon$,

$$\widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi) = \int_\epsilon^T \widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, t)w(t)\mathrm{d}t \geq w(\delta)\widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}}) ,$$

where we recall $\delta = w^{-1}(32\varepsilon)$. By the assumption that $\varepsilon \leq \sup(w)/32$, we have,

$$\mathbb{E}[\widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}})] \leq \frac{\varepsilon}{w(\delta)} \leq \frac{1}{32} .$$

Via the same argument, we also obtain that $\mathbb{E}[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi)] \leq \varepsilon$ implies that $\mathbb{E}[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}})] \leq 1/32$. Therefore it is sufficient to show that $\mathbb{E}[\widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}})] \leq 1/32$, $\mathbb{E}[\mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}})] \leq 1/32$ leads to the lower bound on n.

It follows from standard bounds based on Girsanov's theorem that

$$\mathrm{KL}\left(p_\delta||q_{T-\delta}\right) \leq \mathscr{L}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}}) + \mathrm{KL}\left(p_T||q_0\right) \ ,$$
$$\mathrm{KL}\left(\hat{p}_\delta||q_{T-\delta}\right) \leq \widehat{\mathscr{L}}_{\mathrm{ESM}}(\hat{s}, \varpi^\delta_{\mathrm{ELBO}}) + \mathrm{KL}\left(\hat{p}_T||q_0\right) \ .$$

Using that q_0 satisfies a logarithmic Sobolev inequality with constant κ^{-1}, in combination with Corollary 2 of (Otto & Villani, 2001), we obtain the upper bound,

$$\begin{aligned}\mathrm{KL}\left(p_T||q_0\right) &\leq \exp(-2\kappa(T-s))\mathrm{KL}\left(p_s||q_0\right) \\ &\leq \frac{\exp(-2\kappa(T-s))}{4s}\,\mathrm{W}_2(p_0,q_0)^2 \ ,\end{aligned}$$

any $s \in (0, T]$. Setting $s = 1$, we obtain,

$$\mathrm{KL}\left(p_T||q_0\right) \leq \frac{\exp(-2\kappa(T-1))}{2}(\sigma^2 + \kappa^{-1}d) \ .$$

By the lower bound assumption on T, we have that $\mathrm{KL}\left(p_T||q_0\right) \leq 1/32$ and, via a similar argument, $\mathbb{E}[\mathrm{KL}\left(\hat{p}_T||q_0\right)] \leq 1/32$ also. Thus, it follows that

$$\mathbb{E}\left[\mathrm{KL}\left(p_\delta||q_{T-\delta}\right)\right], \ \mathbb{E}\left[\mathrm{KL}\left(\hat{p}_\delta||q_{T-\delta}\right)\right] \leq 1/16 \ .$$

The result then follows immediately from Proposition B.5. □

Numerical evaluation. A legitimate concern is whether our bounds indeed lead to the conclusion that benign overfitting is impossible in practically relevant settings. Indeed, our bound only grows exponentially fast in the intrisic dimension of the data when

$$2\frac{\sigma_\epsilon}{\alpha_\epsilon}(\sqrt{\log(64)} + \sqrt{d}) < 1.$$

We address this question in this paragraph by numerically evaluatin our bound in this setting. Here are the values of the ambient dimension and early stopped noise scale in widely used benchmark datasets:

- *Dimension* d, for CIFAR10, we have $d = 3072$, whereas for MNIST we have $d = 784$. In latent diffusion models, it typically ranges from approcimately $4 \cdot 10^3$ to 6×10^4 (Rombach et al., 2022).
- *Noise scale* σ_ϵ. Much of the literature uses noise scales of order 10^{-2} to 10^{-3} (Song et al., 2021a; Karras et al., 2022), sometimes even ranging to 10^{-5}.

We can thus estimate the noise scale that is necessary for the exponentially growing condition of our bounds to be satisfied.

- For $d = 3072$, we need approximately $\sigma_\epsilon < 0.0087$.
- For $d = 3072$, we need approximately $\sigma_\epsilon < 0.017$.
- For $d = 3072$, we need approximately $\sigma_\epsilon < 0.002$.

Thus, in each case, the noise scales required for our lower bound to be informative are consistent with common practice.

To make our analysis more precise, we need to estimate the intrinsic dimension d' of the data. While the exact notion of dimension used in our paper is not always the same as the ones considered in the literature, there exist many intrinsic dimension estimators, such as the ML dimension considered by (Nakada & Imaizumi, 2020), providing results varying from approximately 10 to 30 for CIFAR10, and from 13 to 40 for MNIST. Using the ML estimator in (Nakada & Imaizumi, 2020) ($d' = 26$) and taking $\sigma_\epsilon = 0.002$ as in (Karras et al., 2022), our lower bound gives a large sample complexity: $n \geq 3 \cdot 10^8$ (with $c = 1/2$, assuming the assumptions of our theorem are satisfied). This setting is close to the one used in our experiments, suggesting that our bounds are relevant in realistic settings.

B.2 Omitted proofs of Section 4

In this section, we present the proofs of Lemma 4.1 and Proposition 4.2.

B.2.1 PROOF OF LEMMA 4.1

We present below the proof of Lemma 4.1.

Proof. Let $(Z_1, \dots, Z_n) \in (\mathbb{R}^d)^N$ be a dataset and $W \in \mathbb{R}^{p\times d}$ be fixed. By assumption, W has full rank. Let $\Xi \sim \mathrm{N}(0, \mathrm{I}_d)$. We can write the empirical risk as

$$A \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\sigma_t AW(\alpha_t Z_i + \sigma_t \Xi) + \Xi\|^2\right] .$$

As $\mathbb{E}[\Xi] = 0$, we have

$$\begin{aligned}
\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t) &= d + \sigma_t^2 \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Tr}\left(W^T A^T AW(\alpha_t Z_i + \sigma_t \Xi)(\alpha_t Z_i + \sigma_t \Xi)^T\right)\right] \\
&\quad + 2\sigma_t^2 \mathbb{E}\left[\mathrm{Tr}\left(AW\Xi\Xi^T\right)\right] \\
&= d + \sigma_t^2 \mathrm{Tr}\left(A^T AW(\alpha_t^2 \widehat{\mathbf{\Sigma}} + \sigma_t^2 \mathrm{I}_d)W^T\right) + 2\sigma_t^2 \mathrm{Tr}(AW) \\
&= d + \sigma_t^2 \mathrm{Tr}\left(A^T AW\widehat{\mathbf{\Sigma}}_t W^T\right) + 2\sigma_t^2 \mathrm{Tr}(AW) .
\end{aligned}$$

Let $A \in \arg\min \widehat{\mathcal{R}}_{W,t}(\widehat{A})$, by taking the gradient above, we have

$$AW\widehat{\mathbf{\Sigma}}_t W^T + W^T = 0 . \tag{13}$$

Case 1 ($p < d$). By our assumptions, the matrix W has full rank, so its rank is p in this case. The matrix $\widehat{\mathbf{\Sigma}}_t$ is always invertible, as soon as $t > 0$; we also know that $\mathrm{rank}(W) = \mathrm{rank}(W\widehat{\mathbf{\Sigma}}_t^{1/2})$, therefore, we conclude that $\mathrm{rank}(W\widehat{\mathbf{\Sigma}}_t W^T) = p$, so the matrix $W\widehat{\mathbf{\Sigma}}_t W^T \in \mathbb{R}^{p\times p}$ is invertible.

Therefore, the empirical risk minimizer is given by

$$\widehat{A} := -W^T(W\widehat{\mathbf{\Sigma}}_t W^T)^{-1} .$$

Using Equation (13), we have that

$$\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, t) = d - \sigma_t^2 \mathrm{Tr}\left(W^T(W\widehat{\mathbf{\Sigma}}_t W^T)^{-1} W\right) .$$

To compute the score matching loss, we note that when $\overrightarrow{X}_0 \sim \nu = \mathrm{N}(0, \mathbf{\Sigma})$, then $X_t \sim \mathrm{N}(0, \alpha_t^2\Sigma + \sigma_t^2 \mathrm{I}_d) = \mathrm{N}(0, \mathbf{\Sigma}_t)$. Therefore, the score can be expressed as

$$\nabla \log p_t(x) = -\mathbf{\Sigma}_t^{-1} x .$$

The score matching loss is then equal to

$$\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) = \mathbb{E}\left[\left\|\widehat{A}W(\alpha_t Z + \sigma_t \Xi) + \mathbf{\Sigma}_t^{-1}(\alpha_t Z + \sigma_t \Xi)\right\|^2\right] ,$$

with $(Z, \Xi) \sim \nu \otimes \mathrm{N}(0, \mathrm{I}_d)$. By symmetry of $\widehat{A}W$, we obtain that

$$\begin{aligned}
\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) &= \mathbb{E}\left[\left\|\widehat{A}W(\alpha_t Z + \sigma_t \Xi) + \mathbf{\Sigma}_t^{-1}(\alpha_t Z + \sigma_t \Xi)\right\|^2\right] \\
&= \mathrm{Tr}\left(W^T \widehat{A}^T \widehat{A} W \mathbf{\Sigma}_t + 2\widehat{A}W + \mathbf{\Sigma}_t^{-1}\right) \\
&= \left\|\left(W^T(W\widehat{\mathbf{\Sigma}}_t W^T)^{-1} W - \mathbf{\Sigma}_t^{-1}\right)\mathbf{\Sigma}_t^{1/2}\right\|_{\mathrm{F}}^2 .
\end{aligned}$$

Case 2. ($p \geq d$) In this case, by similar argument as in the first case, we note that the matrix $W \in \mathbb{R}^{p\times d}$ has full rank by assumption. Consider the linear map $\Phi_W : \mathbb{R}^{d\times p} \to \mathbb{R}^{d\times d}$ defined by $\Phi_W(A) = AW$. Then, we easily show that this map is almost-surely surjective, because the matrix W admits a left inverse W^+ such that $W^+W = \mathrm{I}_d$ (which can be obtained from the singular value decomposition of W).

As both the empirical risk and the score matching loss depend only on A through the product AW, we have that all the empirical risk minimizers $\widehat{A}$ satisfy $\widehat{A}W = \widehat{H} \in \mathbb{R}^{d\times d}$, with $\widehat{H}$ the empirical minimizer of the following linear model:

$$\widehat{H} \in \arg\min_{H} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\sigma_t H(\alpha_t Z_i + \sigma_t \Xi) + \Xi\|^2 \right] \right\} , \quad (Z, \Xi) \sim \nu \otimes \mathrm{N}(0, \mathrm{I}_d) .$$

By similar computations as before, we easily show that $\widehat{H} = -\widehat{\mathbf{\Sigma}}_t^{-1}$. Therefore, the minimum empirical risk is

$$\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, t) = d - \sigma_t^2 \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}_t^{-1} \right) .$$

Finally, the score matching loss can be expressed as

$$\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) = \mathrm{Tr}\left((\widehat{A}W + \mathbf{\Sigma}_t^{-1})^2 \mathbf{\Sigma}_t \right) = \left\| \left(\widehat{\mathbf{\Sigma}}_t^{-1} - \mathbf{\Sigma}_t^{-1} \right) \mathbf{\Sigma}_t^{1/2} \right\|_{\mathrm{F}}^2 .$$

This concludes the proof. □

B.2.2 Proof of Proposition 4.2

We prove Proposition 4.2 as a corollary of the following more general result, which gives the exact asymptotic behavior of the score matching loss and the empirical risk.

Proposition B.6. *Consider the same setup as in Proposition 4.2. Let $\widehat{A} \in \arg\min_A \widehat{\mathcal{R}}_{W,t}(A)$. If $\psi_p > \psi_d$, we have*

$$\lim_{n\to\infty} \frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t) \right) = 1 - \sigma_t^2 \int \frac{\mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_t^2 \beta \lambda + \sigma_t^2} ,$$

and

$$\lim_{n\to\infty} \mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) = (\alpha_t^2\beta + \sigma_t^2) \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2} \right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda) .$$

If $\psi_p < \psi_d$, we have

$$\frac{1}{d}\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,\widehat{A}}, t) \underset{n\to\infty}{\longrightarrow} 1 - \sigma_t^2 \frac{\psi_p}{\psi_d} \int \frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_p)}(\lambda) ,$$

and

$$\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) \underset{n\to\infty}{\longrightarrow} (\alpha_t^2\beta + \sigma_t^2)\frac{\psi_p}{\psi_d} \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2} \right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_p)}(\lambda) + \frac{1 - \frac{\psi_p}{\psi_d}}{\alpha_t^2\beta + \sigma_t^2} .$$

Proof. First, we note that in that case we have $\mathbf{\Sigma}_t = (\alpha_t^2\beta + \sigma_t^2)\mathrm{I}_d$. Moreover, we observe that, as the matrix $W \in \mathbb{R}^{p\times d}$ has i.i.d. standard Gaussian entries, it has almost-surely full rank.

We can therefore apply Lemma 4.1, we distinguish two cases.

Case 1. ($\psi_p > \psi_d$) For n large enough, we have $p > d$. In this case, by Lemma 4.1, the normalized empirical risk can be written as

$$\frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t) \right) = 1 - \sigma_t^2 \int \frac{\mathrm{d}\mu_{\beta^{-1}\widehat{\mathbf{\Sigma}}}(\lambda)}{\alpha_t^2\beta\lambda + \sigma_t^2} ,$$

where, for a positive definite matrix M, $\widehat{\mu}_M$ denotes the empirical spectral measure of M, see Section A.2.

By the Marchenko-Pastur theorem (e.g. Theorem 2.4 of (Couillet & Liao, 2022)), we know that, almost-surely, the measure $\mu_{\beta^{-1}\widehat{\mathbf{\Sigma}}}$ converges weakly to the Marchenko-Pastur distribution with shape parameter ψ_d, denoted $\mu_{\mathrm{MP}}^{(\psi_d)}$, see Section A.2. The map $\lambda \mapsto (\alpha_t^2\beta\lambda + \sigma_t^2)$ is bounded and continuous, therefore, we have

$$\lim_{n\to\infty} \frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t) \right) = 1 - \sigma_t^2 \int \frac{\mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_t^2\beta\lambda + \sigma_t^2} .$$

Similarly, we have, with $\widehat{A} \in \arg\min_A \widehat{\mathcal{R}}_{W,t}(A)$,

$$\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) = \frac{1}{d}(\alpha_t^2\beta + \sigma_t^2) \left\| \left(\widehat{\mathbf{\Sigma}}_t^{-1} - \mathbf{\Sigma}_t^{-1} \right) \mathbf{\Sigma}_t^{1/2} \right\|_{\mathrm{F}}^2 \underset{n\to\infty}{\longrightarrow} (\alpha_t^2\beta + \sigma_t^2) \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2} \right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda) \, .$$

Case 2. ($\psi_p < \psi_d$) For n large enough, we have $d > p$ In this case, we note (see the proof of Lemma 4.1) that the matrix W is almost surely of rank p. Therefore, we can write the singular value decomposition of W in the form $W = U\Lambda V^T$, where $U \in \mathbb{R}^{p\times p}$ is orthonormal, $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive diagonal coefficients, and $V \in \mathbb{R}^{d\times p}$ is column orthonormal ($V^TV = \mathrm{I}_p$).

With these notations and noting that $V^TV = \mathrm{I}_p$, we have, almost-surely, that

$$\begin{aligned} \frac{1}{d}\sigma_t^2 \widehat{\mathscr{L}_{\mathrm{DSM}}}(s_{W,\widehat{A}}, t) &= 1 - \frac{\sigma_t^2}{d} \mathrm{Tr}\left(V\Lambda U^T (U\Lambda V^T \widehat{\mathbf{\Sigma}}_t V \Lambda U^T)^{-1} U\Lambda V^T \right) \\ &= 1 - \frac{\sigma_t^2}{d} \mathrm{Tr}\left(V (V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} V^T \right) \\ &= 1 - \frac{\sigma_t^2}{d} \mathrm{Tr}\left((V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} \right) \, . \end{aligned}$$

Now we observe that

$$V^T \widehat{\mathbf{\Sigma}}_t V = \alpha_t^2 V^T \widehat{\mathbf{\Sigma}} V + \sigma_t^2 \mathrm{I}_p = \frac{\alpha_t^2}{n} \sum_{i=1}^n V^T Z_i (V^T Z_i)^T + \sigma_t^2 \mathrm{I}_p \, .$$

Recall that $(Z_1, \dots, Z_n) \sim \mathrm{N}(0, \beta \mathrm{I}_d)^{\otimes n}$ and that V is column orthonormal (actually, it is even uniformly distributed) and independent of $(Z_1, \dots, Z_n)$. By independence and invariance by rotation of $\mathrm{N}(0, \beta\mathrm{I}_d)$, we have that $V^TZ_i \sim \mathrm{N}(0, \beta\mathrm{I}_p)$. Moreover, we have that $(V^TZ_1, \dots, V^TZ_n) \sim \mathrm{N}(0, \beta\mathrm{I}_p)^{\otimes n}$. Thus, we can apply the p-dimensional version of the Marchenko-Pastur theorem, which gives that

$$\frac{1}{d}\sigma_t^2 \widehat{\mathscr{L}_{\mathrm{DSM}}}(s_{W,\widehat{A}}, t) \underset{n\to\infty}{\longrightarrow} 1 - \sigma_t^2 \frac{\psi_p}{\psi_d} \int \frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_p)}(\lambda) \, .$$

By Lemma 4.1, we have

$$\begin{aligned} \frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) &= \frac{1}{d} \left\| \left(W^T (W \widehat{\mathbf{\Sigma}}_t W^T)^{-1} W - \mathbf{\Sigma}_t^{-1} \right) \mathbf{\Sigma}_t^{1/2} \right\|_{\mathrm{F}}^2 \\ &= \frac{\alpha_t^2\beta + \sigma_t^2}{d} \left\| \left(V (V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} V^T - \mathbf{\Sigma}_t^{-1} \right) \right\|_{\mathrm{F}}^2 \, . \end{aligned}$$

Recall that $\mathbf{\Sigma}_t = (\alpha_t^2\beta + \sigma_t^2)\mathrm{I}_d$. By the Pythagorean theorem, we have

$$\begin{aligned} \frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) &= \frac{\alpha_t^2\beta + \sigma_t^2}{d} \left(\left\| V (V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} V^T - \frac{VV^T}{\alpha_t^2\beta + \sigma_t^2} \right\|_{\mathrm{F}}^2 + \frac{\left\| \mathrm{I}_d - VV^T \right\|_{\mathrm{F}}^2}{(\alpha_t^2\beta + \sigma_t^2)^2} \right) \\ &= \frac{\alpha_t^2\beta + \sigma_t^2}{d} \left(\left\| (V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} - \frac{\mathrm{I}_p}{\alpha_t^2\beta + \sigma_t^2} \right\|_{\mathrm{F}}^2 + \frac{\mathrm{Tr}\left(\mathrm{I}_d - VV^T \right)}{(\alpha_t^2\beta + \sigma_t^2)^2} \right) \\ &= \frac{\alpha_t^2\beta + \sigma_t^2}{d} \left(\left\| (V^T \widehat{\mathbf{\Sigma}}_t V)^{-1} - \frac{\mathrm{I}_p}{\alpha_t^2\beta + \sigma_t^2} \right\|_{\mathrm{F}}^2 + \frac{d-p}{(\alpha_t^2\beta + \sigma_t^2)^2} \right) \, . \end{aligned}$$

By the arguments above (for the empirical risk), we observe that we can apply the p dimensional version of the Marchenko-Pastur theorem to the first term, this gives, almost-surely,

$$\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t) \underset{n\to\infty}{\longrightarrow} (\alpha_t^2\beta + \sigma_t^2) \frac{\psi_p}{\psi_d} \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2} \right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_p)}(\lambda) + \frac{1 - \frac{\psi_p}{\psi_d}}{\alpha_t^2\beta + \sigma_t^2} \, .$$

This concludes the proof. □

We present below the proof of Proposition 4.2.

Proof. (of Proposition 4.2) We apply Proposition B.6 and focus on the case where $\psi_p \wedge \psi_d > 1$. Then, in both cases below, the Marchenko-Pastur distributions appearing in Proposition B.6 have a mass at zero, which we exploit to obtain our estimates. Let $f_{\mathrm{MP}}^{(\psi)}$ denote the bulk density of the Marchenko-Pastur distribution with shape parameter $\psi > 0$. Note that this is only a probability density when $\psi \leq 1$, see Section A.2.

Case 1. ($\psi_p > \psi_d$). Then, we have, almost-surely,

$$\lim_{n\to\infty} \frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)\right) = 1 - \sigma_t^2 \left(1 - \frac{1}{\psi_d}\right) \frac{1}{\sigma_t^2} - \sigma_t^2 \int \frac{f_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_t^2 \beta \lambda + \sigma_t^2} \mathrm{d}\lambda \, .$$

The bulk density $f_{\mathrm{MP}}^{(\psi_d)}$ is supported on $[(\sqrt{\psi_d}-1)^2, (\sqrt{\psi_d}+1)^2]$ and we have $\int f_{\mathrm{MP}}^{(\psi_d)}(\lambda)\mathrm{d}\lambda = \psi_d^{-1}$. Therefore, we have, almost-surely

$$\lim_{n\to\infty} \frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)\right) = \frac{1}{\psi_d} + \mathcal{O}\left(\frac{1}{\psi_d^2}\right) \, ,$$

as $\psi_d \to \infty$. For the score matching loss, we have similarly that, almost-surely,

$$\begin{aligned}
\lim_{n\to\infty} \left(\frac{1}{d} \mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t)\right) &= \left(1 - \frac{1}{\psi_d}\right) \frac{\alpha_t^4 \beta^2}{\sigma_t^4(\alpha_t^2\beta + \sigma_t^2)} \\
&\quad + (\alpha_t^2\beta + \sigma_t^2) \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2}\right)^2 f_{\mathrm{MP}}^{(\psi_d)}(\lambda)\mathrm{d}\lambda \\
&= \left(1 - \frac{1}{\psi_d}\right) \frac{\alpha_t^4 \beta^2}{\sigma_t^4(\alpha_t^2\beta + \sigma_t^2)} + \mathcal{O}\left(\frac{1}{\psi_d}\right) \, .
\end{aligned}$$

Case 2. ($\psi_p < \psi_d$) In that case, by very similar computations, we obtain

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{d} \inf_A \left(\sigma_t^2 \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{W,A}, t)\right) &= 1 - \frac{\psi_p}{\psi_d}\left(1 - \frac{1}{\psi_p}\right) - \sigma_t^2 \frac{\psi_p}{\psi_d} \int \frac{f_{\mathrm{MP}}^{(\psi_p)}(\lambda)}{\alpha_t^2\beta\lambda + \sigma_t^2} \mathrm{d}\lambda \\
&= 1 + \frac{1}{\psi_d} - \frac{\psi_p}{\psi_d} + \mathcal{O}_t\left(\frac{1}{\psi_p \psi_d}\right) \, .
\end{aligned}$$

Finally, for the score matching loss, we obtain

$$\begin{aligned}
\lim_{n\to\infty} \left(\frac{1}{d} \mathscr{L}_{\mathrm{ESM}}(s_{W,\widehat{A}}, t)\right) &= \frac{\psi_p}{\psi_d}\left(1 - \frac{1}{\psi_p}\right) \frac{\alpha_t^4 \beta^2}{\sigma_t^4(\alpha_t^2\beta + \sigma_t^2)} + \frac{1 - \frac{\psi_p}{\psi_d}}{\alpha_t^2\beta + \sigma_t^2} \\
&\quad + (\alpha_t^2\beta + \sigma_t^2) \frac{\psi_p}{\psi_d} \int \left(\frac{1}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2}\right)^2 f_{\mathrm{MP}}^{(\psi_p)}(\lambda)\mathrm{d}\lambda \\
&= \frac{\psi_p}{\psi_d}\left(1 - \frac{1}{\psi_p}\right) \frac{\alpha_t^4 \beta^2}{\sigma_t^4(\alpha_t^2\beta + \sigma_t^2)} + \frac{1 - \frac{\psi_p}{\psi_d}}{\alpha_t^2\beta + \sigma_t^2} + \mathcal{O}\left(\frac{1}{\psi_d}\right) \, .
\end{aligned}$$

The result follows by rearranging the terms. □

B.3 Omitted proofs of Section 5.1

In this section, we present the proofs of Section 5.1, related to the impact of the time integration in the score matching loss.

B.3.1 Proof of Lemma 5.1

Proof. (of Lemma 5.1) By calculations similar to the proof of Lemma 4.1, we can write the denoising score matching loss as

$$\begin{aligned}
\widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W}, \varpi) &= \sigma_\pi^{-2} \int_{[0,T]} \left(d + \sigma_t^2 \mathrm{Tr}(A^T A W \widehat{\mathbf{\Sigma}}_t W^T) + 2\sigma_t^2 \mathrm{Tr}(AW)\right) \mathrm{d}\pi(t) \\
&= \frac{d}{\sigma_\pi^2} + \mathrm{Tr}(W^T A^T A W \widehat{\mathbf{\Sigma}}_\varpi) + 2\mathrm{Tr}(AW) \, .
\end{aligned}$$

By reasoning as in the proof Lemma 4.1, we observe that (because $p > d$) all empirical risk minimizers $\widehat{A}$ satisfy

$$\widehat{A}W = -\widehat{\mathbf{\Sigma}}_{\varpi}^{-1} .$$

Therefore, the minimum empirical risk is equal to

$$\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W}, \varpi) = \frac{d}{\sigma_\pi^2} - \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}_{\varpi}^{-1}\right) .$$

Again, by similar calculations as in the proof of Lemma 4.1, we can express the population score matching loss as

$$\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W}, \pi) = \int_{[0,T]} \mathbb{E}\left[\left\|\widehat{A}W(\alpha_t Z + \sigma_t \Xi) + \mathbf{\Sigma}_t^{-1}(\alpha_t Z + \sigma_t \Xi)\right\|^2\right] \mathrm{d}\pi(t) .$$

with $(Z, \Xi) \sim \nu \otimes \mathrm{N}(0, I_d)$. Therefore, we have

$$\begin{aligned}
\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W}, \pi) &= \int_{[0,T]} \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}_{\varpi}^{-2}\mathbf{\Sigma}_t - 2\widehat{\mathbf{\Sigma}}_{\varpi}^{-1} + \mathbf{\Sigma}_t^{-1}\right) \mathrm{d}\pi(t) \\
&= \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}_{\varpi}^{-2}\mathbf{\Sigma}_\pi - 2\widehat{\mathbf{\Sigma}}_{\varpi}^{-1} + \int_{[0,T]} \mathbf{\Sigma}_t^{-1} \mathrm{d}\pi(t)\right) \\
&= \left\|\left(\widehat{\mathbf{\Sigma}}_{\varpi}^{-1} - \mathbf{\Sigma}_\pi^{-1}\right)\mathbf{\Sigma}_\pi^{1/2}\right\|_{\mathrm{F}}^2 + \mathrm{C}_\pi ,
\end{aligned}$$

with

$$\mathrm{C}_\pi := \mathrm{Tr}\left(\int_{[0,T]} \mathbf{\Sigma}_t^{-1} \mathrm{d}\pi(t) - \mathbf{\Sigma}_\pi^{-1}\right) . \tag{14}$$

By diagonalizing Σ_t and using Jensen's inequality, we see that $\mathrm{C}_\pi \geq 0$. □

Remark B.7 (Case where $p \leq d$.)**.** *By proceeding like in the proof of Lemma 4.1, we have that*

$$\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W}, \varpi) = \frac{d}{\sigma_\pi^2} - \mathrm{Tr}\left(W^T(W\widehat{\mathbf{\Sigma}}_{\varpi}W^T)^{-1}W\right) .$$

Similarly, for the explicit score matching loss, we have

$$\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W}, \pi) \left\|\left(W^T(W\widehat{\mathbf{\Sigma}}_{\varpi}W^T)^{-1}W - \mathbf{\Sigma}_\pi^{-1}\right)\mathbf{\Sigma}_\pi^{1/2}\right\|_{\mathrm{F}}^2 + \mathrm{C}_\pi ,$$

where $\widehat{A} = -(W\widehat{\mathbf{\Sigma}}_{\varpi}W^T)^{-1}$ is the empirical risk minimizer. These formulas are used to obtain the exact asymptotics presented in Figure 3, using very similar computations as in the proofs of Section 4.

Remark B.8 (Case where π is the uniform distribution and $\kappa = 1$.)**.** *In that case, we have*

$$\frac{1}{d}\mathrm{C}_\pi = \frac{1}{T}\int_0^T \frac{\mathrm{d}t}{\alpha_t^2\beta + \sigma_t^2} - \frac{1}{\alpha_\pi^2\beta + \sigma_\pi^2} .$$

We easily see that

$$\alpha_\pi^2 = \frac{1}{T}\int_0^T e^{-2t}\mathrm{d}t = \frac{1 - e^{-2T}}{2T} , \quad \sigma_\pi^2 = 1 - \frac{1 - e^{-2T}}{2T} .$$

Therefore, we have

$$\begin{aligned}
\frac{1}{d}\mathrm{C}_\pi &= \frac{1}{T}\int_0^T \frac{\mathrm{d}t}{e^{-2t}\beta + 1 - e^{-2t}} - \frac{2T}{\beta(1 - e^{-2T}) + 2T - 1 + e^{-2T}} \\
&= \frac{1}{2T}\log\left(\frac{e^{2T} + \beta - 1}{\beta}\right) - \frac{2T}{\beta(1 - e^{-2T}) + 2T - 1 + e^{-2T}} .
\end{aligned}$$

Therefore, we have

$$\frac{1}{d}\mathrm{C}_\pi = \frac{1}{2T}\log\left(\frac{1 + e^{-2T}(\beta - 1)}{\beta}\right) + \frac{\beta(1 - e^{-2T}) - 1 + e^{-2T}}{\beta(1 - e^{-2T}) + 2T - 1 + e^{-2T}} = \mathcal{O}\left(\frac{1}{T}\right) . \tag{15}$$

B.3.2 Proof of Proposition 5.3

We provide below the proof of Proposition 5.3. We first provide, in the next proposition, the exact asymptotics of the score matching losses in the case of time-integrated losses.

Proposition B.9. *Consider the same setup as in Proposition 5.3, we have*

$$\lim_{n\to\infty}\left(\frac{1}{d}\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W},\varpi)\right) = \frac{1}{\sigma_\pi^2} - \int \frac{\mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}\ ,$$

and

$$\lim_{n\to\infty}\left(\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W},\pi)\right) = (\alpha_\pi^2\beta+\sigma_\pi^2)\int\left(\frac{1}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}-\frac{1}{\alpha_\pi^2\beta+\sigma_\pi^2}\right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda) + \overline{\mathrm{C}}_\pi\ ,$$

where, as before, $\mu_{\mathrm{MP}}^{(\psi_d)}$ is the Marchenko-Pastur distribution with shape parameter ψ_d, and

$$\overline{\mathrm{C}}_\pi := \int_{[0,T]} \frac{\pi(\mathrm{d}t)}{\alpha_t^2\beta+\sigma_t^2} - \frac{1}{\alpha_\pi^2\beta+\sigma_\pi^2} \geq 0\ . \tag{16}$$

Proof. Assume that $\psi_p > \psi_d$. Then, by Lemma 4.1 (using the same notations), we have

$$\frac{1}{d}\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W},\varpi) = \frac{1}{\sigma_\pi^2} - \int \frac{\mathrm{d}\mu_{\beta^{-1}\widehat{\mathbf{\Sigma}}}(\lambda)}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}\ .$$

By the Marchenko-Pastur's theorem, we have

$$\lim_{n\to\infty}\left(\frac{1}{d}\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W},\varpi)\right) = \frac{1}{\sigma_\pi^2} - \int \frac{\mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}\ .$$

Similarly, the score matching loss can be expressed as

$$\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W},\pi) = \frac{\alpha_\pi^2\beta+\sigma_\pi^2}{d}\left\|\widehat{\mathbf{\Sigma}}_\varpi^{-1} - \frac{\mathrm{I}_d}{\alpha_\pi^2\beta+\sigma_\pi^2}\right\|_{\mathrm{F}}^2 + \frac{\mathrm{C}_\pi}{d}\ ,$$

where C_π is defined in the proof of Lemma 5.1. Under Assumption 4.1, we observe that C_π/d is independent of d, and therefore we can define

$$\overline{\mathrm{C}}_\pi := \frac{\mathrm{C}_\pi}{d} = \int_{[0,T]} \frac{\pi(\mathrm{d}t)}{\alpha_t^2\beta+\sigma_t^2} - \frac{1}{\alpha_\pi^2\beta+\sigma_\pi^2} \geq 0\ ,$$

where the inequality follows from Jensen's inequality. By applying the Marchenko-Pastur theorem, we obtain

$$\lim_{n\to\infty}\left(\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W},\pi)\right) = (\alpha_\pi^2\beta+\sigma_\pi^2)\int\left(\frac{1}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}-\frac{1}{\alpha_\pi^2\beta+\sigma_\pi^2}\right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda) + \overline{\mathrm{C}}_\pi\ .$$

This concludes the proof. □

Remark B.10. *Based on Remark B.7, we can easily extend these derivations to the case $\psi_p < \psi_d$.*

We present below the proof of Proposition 5.3.

Proof. (of Proposition 5.3) Assume that $\psi_p > \psi_d > 1$, then the Marchenko-Pastur has a mass at zero in all the cases below. By Proposition B.9, we immediately have

$$\begin{aligned}\lim_{n\to\infty}\left(\frac{1}{d}\inf_A \widehat{\mathscr{L}}_{\mathrm{DSM}}(s_{A,W},\varpi)\right) &= \frac{1}{\sigma_\pi^2} - \frac{1}{\sigma_\varpi^2}\left(1-\frac{1}{\psi_d}\right) - \int \frac{f_{\mathrm{MP}}^{(\psi_d)}(\lambda)}{\alpha_\varpi^2\beta\lambda+\sigma_\varpi^2}\mathrm{d}\lambda \\ &= \frac{1}{\sigma_\pi^2} - \frac{1}{\sigma_\varpi^2} + \frac{1}{\psi_d\sigma_\varpi^2} + \mathcal{O}\left(\frac{1}{\psi_d^2}\right)\ .\end{aligned}$$

Note that, by the Cauchy-Schwarz inequality, we always have $\sigma_\pi^{-2} \geq \sigma_\varpi^{-2}$. Similarly, the asymptotic score matching loss satisfies

$$\lim_{n\to\infty}\left(\frac{1}{d}\mathscr{L}_{\mathrm{ESM}}(s_{\widehat{A},W},\pi)\right) = \left(1-\frac{1}{\psi_d}\right)\frac{(\alpha_\pi^2\beta+\sigma_\pi^2-\sigma_\varpi^2)^2}{\sigma_\varpi^4(\alpha_\pi^2\beta+\sigma_\pi^2)} + \overline{\mathrm{C}}_\pi + \mathcal{O}\left(\frac{1}{\psi_d}\right)\ .$$

The proof follows by defining

$$\mathrm{L}_\pi := \frac{(\alpha_\pi^2\beta+\sigma_\pi^2-\sigma_\varpi^2)^2}{\sigma_\varpi^4(\alpha_\pi^2\beta+\sigma_\pi^2)} + \overline{\mathrm{C}}_\pi \tag{17}$$

This concludes the proof. □

B.4 Proof of Section 5.2

We present below the proof of Proposition 5.4.

Proof. (of Proposition 5.4) **Step 1.** We start by solving the gradient flow dynamics of Equation (9). We can expand the gradient flow dynamics as

$$\frac{\mathrm{d}}{\mathrm{d}\tau}A(\tau) = -2p^{-1}\sigma_t^2 A(\tau)W\widehat{\mathbf{\Sigma}}_t W^T - 2p^{-1}\sigma_t^2 W^T \ .$$

We can assume, without loss of generality, that $p > d$. Then, by reasoning as in the proof of Proposition 4.2, we observe that the matrix $W^T W \in \mathbb{R}^{d\times d}$ is almost-surely invertible. Therefore, we can define $\tilde{A} := (W^T W)^{-1}AW$ and we have almost-surely that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde{A}(\tau) = -2p^{-1}\sigma_t^2 \tilde{A}(\tau)\widehat{\mathbf{\Sigma}}_t W^T W - 2p^{-1}\sigma_t^2 \mathrm{I}_d \ .$$

Taking into account that the initialization satisfies $A(0) = 0$, we deduce that

$$\tilde{A}(\tau) = -(\widehat{\mathbf{\Sigma}}_t W^T W)^{-1}\left(\mathrm{I}_d - e^{-2\tau p^{-1}\sigma_t^2 \widehat{\mathbf{\Sigma}}_t W^T W}\right) \ .$$

Therefore, we have

$$A(\tau)W = -\widehat{\mathbf{\Sigma}}_t^{-1}\left(\mathrm{I}_d - e^{-2\tau p^{-1}\sigma_t^2 \widehat{\mathbf{\Sigma}}_t W^T W}\right) \in \mathbb{R}^{d\times d} \ .$$

By the strong law of large numbers, we have, almost-surely (for fixed n, d, and τ), that $p^{-1}W^T W \to \mathrm{I}_d$ as $p \to \infty$. Thus, almost-surely, we have

$$\lim_{p\to\infty}(A(\tau)W) = -\widehat{\mathbf{\Sigma}}_t^{-1}\left(\mathrm{I}_d - e^{-2\tau\sigma_t^2\widehat{\mathbf{\Sigma}}_t}\right) \in \mathbb{R}^{d\times d} \ . \tag{18}$$

Step 2. By computations similar to the proof of Lemma 4.1 and Proposition 4.2, we have (for fixed (p, d, n))

$$\mathcal{E}_t(\tau, n, d, p) := \mathscr{L}_{\mathrm{ESM}}(s_{W,A(\tau)}, t) = \left\|\left(\widehat{\mathbf{\Sigma}}_t^{-1}\left(\mathrm{I}_d - e^{-2\tau p^{-1}\sigma_t^2\widehat{\mathbf{\Sigma}}_t W^T W}\right) - \mathbf{\Sigma}_t^{-1}\right)\mathbf{\Sigma}_t^{1/2}\right\|_{\mathrm{F}}^2 \ .$$

Therefore, by the law of large numbers, we have almost surely that

$$\begin{aligned}\lim_{p\to\infty}\mathcal{E}(\tau, n, d, p) &= \left\|\left(\widehat{\mathbf{\Sigma}}_t^{-1}\left(\mathrm{I}_d - e^{-2\tau\sigma_t^2\widehat{\mathbf{\Sigma}}_t}\right) - \mathbf{\Sigma}_t^{-1}\right)\mathbf{\Sigma}_t^{1/2}\right\|_{\mathrm{F}}^2 \\ &= (\alpha_t^2\beta + \sigma_t^2)\left\|\widehat{\mathbf{\Sigma}}_t^{-1}\left(\mathrm{I}_d - e^{-2\tau\sigma_t^2\widehat{\mathbf{\Sigma}}_t}\right) - \mathbf{\Sigma}_t^{-1}\right\|_{\mathrm{F}}^2 =: \overline{\mathcal{E}}_t(\tau, n, d) \ .\end{aligned}$$

Recall that we take $n, d \to \infty$ with $\psi_d := \lim(d/n) > 1$. Therefore, we can apply the Marchenko-Pastur theorem and obtain, through similar computations as in the proof of Proposition 4.2, that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{d}\overline{\mathcal{E}}_t(\tau, n, d) &= (\alpha_t^2\beta + \sigma_t^2)\int\left(\frac{1 - e^{-2\tau\sigma_t^2(\alpha_t^2\beta\lambda + \sigma_t^2)}}{\alpha_t^2\beta\lambda + \sigma_t^2} - \frac{1}{\alpha_t^2\beta + \sigma_t^2}\right)^2 \mathrm{d}\mu_{\mathrm{MP}}^{(\psi_d)}(\lambda) \\ &= \left(1 - \frac{1}{\psi_d}\right)\frac{\left((\alpha_t^2\beta + \sigma_t^2)\left(1 - e^{-2\tau\sigma_t^4}\right) - \sigma_t^2\right)^2}{\sigma_t^4(\alpha_t^2\beta + \sigma_t^2)} + \mathcal{O}\left(\frac{1}{\psi_d}\right) \ .\end{aligned}$$

This concludes the proof. □

C Experimental details

In this section, we provide some additional details on the experiments presented in the main text.

C.1 Random features experiments

In this short subsection, we quickly describe the empirical setup for the diffusion random feature experiment presented in Figure 1.

Regression experiments. In the regression experiment presented in Figure 1, we use a classical random feature regression setting. We consider a data distribution of the form

$$y_i = \langle \beta_\star, x_i \rangle + \sigma \epsilon_i \ , \quad x_i \sim \mathrm{N}(0, \mathrm{I}_d) \ , \quad \epsilon \sim \mathrm{N}(0,1) \ ,$$

where ϵ_i and x_i are independent and we sample n i.i.d. pairs (x_i, y_i) for $i \in \{1, \dots, n\}$. We consider a model of the form

$$f_{A,W}(x) = \frac{A}{\sqrt{p}} \varrho\left(\frac{W}{\sqrt{d}} x\right) \ ,$$

where $A \in \mathbb{R}^{d \times p}$, $W \in \mathbb{R}^{p \times d}$ is a random matrix with i.i.d. standard Gaussian entries, and $\varrho(x) := \max(x, 0)$ is the ReLU activation.

We consider the minimization of the empirical risk

$$\widehat{A} \in \arg\min_A \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f_{A,W}(x_i))^2 + \frac{\lambda}{\sqrt{p}} \|A\|_{\mathrm{F}}^2 \right\} \ ,$$

with λ a ridge regularization parameter.

Then, we evaluate the test (using a validation set) and train risk evaluated at the empirical risk minimizer $\widehat{A}$ and report it in Figure 1.

Hyperparameters details. In Figure 1, we use $n = 10^2$, $d = 20$, $N_g = 10$, $\lambda = 10^{-4}$, and $p \in [5, 4 \cdot 10^2]$.

Diffusion experiments. We consider the following family of random features score networks, which is similar to existing works (Bonnaire et al., 2025; George & Macris, 2026), for $x \in \mathbb{R}^d$

$$f_{A,W}(x) = \frac{A}{\sqrt{p}} \varrho\left(\frac{W}{\sqrt{d}} x\right) \ ,$$

where $A \in \mathbb{R}^{d \times p}$, $W \in \mathbb{R}^{p \times d}$ is a random matrix with i.i.d. standard Gaussian entries, and $\varrho(x) := \max(x, 0)$ is the ReLU activation. For the data distribution, we take the data distribution to by $\nu = \mathrm{N}(0, d^{-1} \mathrm{I}_d)$.

Given a fixed $t > 0$, a random W and a dataset $S^{(n)} := (Z_1, \dots, Z_n) \sim \nu^{\otimes n}$, we consider the minimization of the regularized empirical risk, defined by

$$A \mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\Xi \left[\left\| \sigma_t f_{A,W}(\alpha_t Z_i + \sigma_t \Xi) + \Xi \right\|^2 \right] + \frac{\lambda}{\sqrt{p}} \|A\|_{\mathrm{F}}^2 \ ,$$

with $\Xi \sim \mathrm{N}(0, \mathrm{I}_d)$ and λ a ridge regularization parameter. An important difference with the regression case above is that, in order to make the empirical risk minimization tractable, we use a Monte Carlo approximation of the expectation over $\Xi \sim \mathrm{N}(0, \mathrm{I}_d)$. Given an integer $N_g \in \mathbb{N}^\star$ and $(\xi_{ij})_{1 \le i \le n,\ 1 \le j \le N_g} \sim \mathrm{N}(0, \mathrm{I}_d)^{\otimes (n N_g)}$, we therefore compute

$$\widehat{A} \in \arg\min_A \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{N_g} \left\| \sigma_t f_{A,W}(\alpha_t Z_i + \sigma_t \xi_{ij}) + \xi_{ij} \right\|^2 + \frac{\lambda}{\sqrt{p}} \|A\|_{\mathrm{F}}^2 \right\}$$

by noting that it can be written as the solution of a linear regression problem. This procedure is similar to (George et al., 2025).

After estimating A, we compute the empirical and population denoising score matching loss, where the latter is estimated using a validation set. The hyperparameters details are provided below.

Hyperparameters details. In Figure 1, we use $\kappa = 1$, $n = 2 \cdot 10^4$, $t = 10^{-1}$, $d = 3$, $N_g = 10$, $\lambda = 10^{-4}$, and $p \in [2, 10^4]$.

Table 1: Metrics as a function of N_F.

N_F	ρ	M_r	M_{ex}
2	2.669	0.000	0.000
4	1.842	0.000	0.000
8	1.602	0.000	0.000
16	0.930	0.000	0.007
32	0.787	0.001	0.037
64	0.299	0.641	0.530
128	0.098	0.993	0.622

C.2 High-dimensional experiments

Our experiments are performed using an implementation of the DDPM model from (Song et al., 2021b). We use the configuration titled`vp.ddpm.cifar10_continuous` which implements the DDPM model of (Ho et al., 2020) but for the continuous-time variance preserving setting. The architecture is a U-Net (Ronneberger et al., 2015) with the encoder and decoder each consisting of four resolution levels $(32 \times 32, 16 \times 16, 8 \times 8, 4 \times 4)$, with two residual blocks per level and utilizes self-attention at a resolution of 16×16. The model is conditioned on time, with the timestep encoded via a sinusoidal embedding and injected into each residual block.

For figures 2 and 5 we modify the number of features (NF) parameter (denoted `nf` in the configuration file and taking a default value of 128). This parameter defines the base channel width used throughout the model which is multiplied by $m = 1, 2, 2, 2$ to produce the actual channel widths for each of the four resolution levels. The number of parameters in the model scales quadratically with channel width. We consider NF$\in \{2, 4, 8, 16, 32, 64, 128\}$ to see how the number of parameters affects overparameterization. Some technical modifications to the code were required to support smaller values of NF.

For all high-dimensional experiments we use CIFAR-10 but restrict to the first $1,000$ examples to make overfitting easier to achieve. Evaluation is performed on the full $10,000$ held-out examples. We train the model with the default setup in the codebase (ADAM with 10% dropout and EMA decay 0.9999) and we train for $1,000,000$ iterations, which is larger than typical to better guarantee overfitting. In the experiments of Figure 2 we choose the checkpoint with smallest train loss. The samples generated on the right-hand side of Figure 2 uses the DDIM implementation in the codebase to generate samples with a fixed seed across NF values. We perform log-likelihood calculations using the `bpd` implementation in the codebase.

Quantitative memorization measures. To complement the qualitative evidence of Figure 2, we compute three quantitative memorization measures on the same checkpoints and samples (Table 1). First, following Meehan et al. (2020), we report ρ, the mean distance of generated samples to their nearest training point divided by the mean distance of test points to their nearest training point; $\rho < 1$ indicates that samples lie closer to the training set than fresh data from the same distribution. Second, we report the memorization fraction M_r of Yoon et al. (2023), i.e. the fraction of samples whose nearest training point is much closer than the second nearest. Third, we report the extraction fraction M_{ex} of Carlini et al. (2023), the fraction of samples that are near-copies of a training image.

For the experiment in Figure 4, we consider the DDPM network but we restrict the access of the neural network to the time input. However, we do still scale the network properly according to the time input. We then train the model as usual but restrict the time variable to the range $[t, t + r]$ for $t = 0.1$ and $r \in [0.0001, 0.00032, 0.001, 0.0032, 0.01, 0.032, 0.1, 0.5]$. We then evaluate the train and test error at t to obtain the plot.

C.3 Compute resources

Our experiments were run on an Amazon EC2 G6e (g6e.xlarge) instance which has a single NVIDIA L40S Tensor Core GPU.

C.4 LICENSES

Codebases:

- Score-Based Generative Modeling through Stochastic Differential Equations (Song et al., 2021b): Apache License 2.0.

Datasets:

- CIFAR-10 (Krizhevsky, 2009): MIT license.