
Why Diffusion Models Don’t Memorize: The Role of Implicit Dynamical Regularization in Training

Tony Bonnaire[†]

LPENS

Université PSL, Paris

tony.bonnaire@phys.ens.fr

Raphaël Urfin[†]

LPENS

Université PSL, Paris

raphael.urfin@phys.ens.fr

Giulio Biroli

LPENS

Université PSL, Paris

giulio.biroli@phys.ens.fr

Marc Mézard

Department of Computing Sciences

Bocconi University, Milano

marc.mezard@unibocconi.it

Abstract

Diffusion models have achieved remarkable success across a wide range of generative tasks. A key challenge is understanding the mechanisms that prevent their memorization of training data and allow generalization. In this work, we investigate the role of the training dynamics in the transition from generalization to memorization. Through extensive experiments and theoretical analysis, we identify two distinct timescales: an early time τ_{gen} at which models begin to generate high-quality samples, and a later time τ_{mem} beyond which memorization emerges. Crucially, we find that τ_{mem} increases linearly with the training set size n , while τ_{gen} remains constant. This creates a growing window of training times with n where models generalize effectively, despite showing strong memorization if training continues beyond it. It is only when n becomes larger than a model-dependent threshold that overfitting disappears at infinite training times. These findings reveal a form of implicit dynamical regularization in the training dynamics, which allow to avoid memorization even in highly overparameterized settings. Our results are supported by numerical experiments with standard U-Net architectures on realistic and synthetic datasets, and by a theoretical analysis using a tractable random features model studied in the high-dimensional limit. 

1 Introduction

Diffusion Models [DMs, 49, 20, 54, 55] achieve state-of-the-art performance in a wide variety of AI tasks such as the generation of images [45], audios [64], videos [33], and scientific data [31, 40]. This class of generative models, inspired by out-of-equilibrium thermodynamics [49], corresponds to a two-stage process: the first one, called *forward*, gradually adds noise to a data, whereas the second one, called *backward*, generates new data by denoising Gaussian white noise samples. In DMs, the reverse process typically involves solving a stochastic differential equation (SDE) with a force field called *score*. However, it is also possible to define a deterministic transport through an ordinary differential equation (ODE), treating the score as a velocity field, an approach that is for instance followed in flow matching [32].

Understanding the generalization properties of score-based generative methods is a central issue in machine learning, and a particularly important question is how memorization of the training set

[†]Equal contribution.

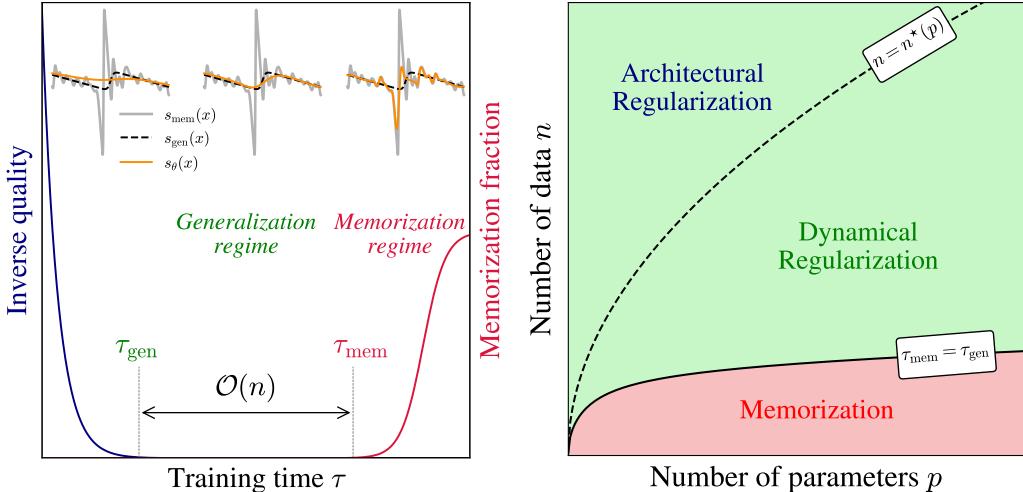


Figure 1: **Qualitative summary of our contributions.** (Left) Illustration of the training dynamics of a diffusion model. Depending on the training time τ , we identify three regimes measured by the inverse quality of the generated samples (blue curve) and their memorization fraction (red curve). The generalization regime extends over a large window of training times which increases with the training set size n . On top, we show a one dimensional example of the learned score function during training (orange). The gray line gives the exact empirical score, at a given noise level, while the black dashed line corresponds to the true (population) score. (Right) Phase diagram in the (n, p) plane illustrating three regimes of diffusion models: **Memorization** when n is sufficiently small at fixed p , **Architectural Regularization** for $n > n^*(p)$ (which is model and dataset dependent, as discussed in [15, 25]), and **Dynamical Regularization**, corresponding to a large intermediate generalization regime obtained when the training dynamics is stopped early, i.e. $\tau \in [\tau_{\text{gen}}, \tau_{\text{mem}}]$.

is avoided in practice. A model without regularization achieving zero training loss only learns the empirical score, and is bound to reproduce samples of the training dataset at the end of the backward process. This memorization regime [30, 6] is empirically observed when the training set is small and disappears when it increases beyond a model-dependent threshold [24]. Understanding the mechanisms controlling this change of regimes from memorization to generalization is a central challenge for both theory and applications. Model regularization and inductive biases imposed by the network architecture were shown to play a role [25, 47], as well as a dynamical regularization due to the finiteness of the learning rate [62]. However, the regime shift described above is consistently observed even in models where all these regularization mechanisms are present. This suggests that the core mechanism behind the transition from memorization to generalization lies elsewhere. In this work, we demonstrate – first through numerical experiments, and then via the theoretical analysis of a simplified model – that this transition is driven by an implicit dynamical bias towards generalizing solutions emerging in the training, which allows to avoid the memorization phase.

Contributions and theoretical picture. We investigate the dynamics of score learning using gradient descent, both numerically and analytically, and study the generation properties of the score depending on the time τ at which the training is stopped. The theoretical picture built from our results and combining several findings from the recent literature is illustrated in Fig. 1. The two main parameters are the size of the training set n and the expressivity of the class of score functions on which one trains the model, characterized by a number of parameters p ; when both n and p are large one can identify three main regimes. Given p , if n is larger than $n^*(p)$ (which depends on the training set and on the class of scores), the score model is not expressive enough to represent the empirical score associated to n data, and instead provides a smooth interpolation, approximately independent of the training set. In this regime, even with a very large training time $\tau \rightarrow \infty$, memorization does not occur because the model is regularized by its architecture and the finite number of parameters. When $n < n^*(p)$ the model is expressive enough to memorize, and two timescales emerge during training: one, τ_{gen} , is the minimum training time required to achieve high-quality data generation; the second, $\tau_{\text{mem}} > \tau_{\text{gen}}$, signals when further training induces memorization, and causes the model to

increasingly reproduce the training samples (left panel). The first timescale, τ_{gen} , is found independent of n , whereas the second, τ_{mem} , grows approximately linearly with n , thus opening a large window of training times during which the model generalizes if early stopped when $\tau \in [\tau_{\text{gen}}, \tau_{\text{mem}}]$. Our results show that implicit dynamical regularization in training plays a crucial role in score-based generative models, substantially enlarging the generalization regime (see right panel of Fig. 1), and hence allowing to avoid memorization even in highly overparameterized settings. We find that the key mechanism behind the widening gap between τ_{gen} and τ_{mem} is the irregularity of the empirical score at low noise level and large n . In this regime the models used to approximate the score provide a smooth interpolation that remains stable for a long period of training times and closely approximates the population score, a behavior likely rooted in the spectral bias of neural networks [42]. Only at very long training times do the dynamics converge to the low lying minimum corresponding to the empirical score, leading to memorization (as illustrated in the one-dimensional examples in the left panel of Fig. 1).

The theoretical picture described above is based on our numerical and analytical results, and builds up on previous works, in particular numerical analysis characterizing the memorization–generalization transition [18, 63], analytical works on memorization of DMs [15, 25, 24], and studies on the spectral bias of deep neural networks [42]. Our numerical experiments use a class of scores based on a realistic U-Net [46] trained on downsampled images of the CelebA dataset [34]. By varying n and p , we measure the evolution of the sample quality (through FID) and the fraction of memorization during learning, which support the theoretical scenario presented in Fig. 1. Additional experimental results on synthetic data are provided in Supplemental Material (SM, Sects. A and B). On the analytical side, we focus on a class of scores constructed from random features and simplified models of data, following [15]. In this setting, the timescales of training dynamics correspond directly to the inverse eigenvalues of the random feature correlation matrix. Leveraging tools from random matrix theory, we compute the spectrum in the limit of large datasets, high-dimensional data, and overparameterized models. This analysis reveals, in a fully tractable way, how the theoretical picture of Fig. 1 emerges within the random feature framework.

Related works.

- The memorization transition in DMs has been the subject of several recent empirical investigations [9, 50, 51] which have demonstrated that state-of-the-art image DMs – including Stable Diffusion and DALL-E – can reproduce a non-negligible portion of their training data, indicating a form of memorization. Several additional works [18, 63] examined how this phenomenon is influenced by factors such as data distribution, model configuration, and training procedure, and provide a strong basis for the numerical part of our work.
- A series of theoretical studies in the high-dimensional regime have analyzed the memorization–generalization transition during the generative dynamics under the empirical score assumption [6, 1, 57], showing how trajectories are attracted to the training samples. Within this high-dimensional framework, [11, 12, 60, 15] study the score learning for various model classes. In particular, [15] uses a Random Feature Neural Network [43]. The authors compute the asymptotic training and test losses for $\tau \rightarrow \infty$ and relate it to memorization. The theoretical part of our work generalizes this approach to study the role of training dynamics and early stopping in the memorization–generalization transition.
- Recent works have also uncovered complementary sources of implicit regularization explaining how DMs avoid memorization. Architectural biases and limited network capacity were for instance shown to constrain memorization in [25, 24], and finiteness of the learning rate prevents the model from learning the empirical score in [62]. Also related to our analysis, [29] provides general bounds showing the beneficial role of early stopping the training dynamics to enhance generalization for finitely supported target distributions, as well as a study of its effect for one-dimensional gaussian mixtures.
- Finally, previous studies on supervised learning [42, 65], and more recently on DMs [59], have shown that deep neural networks display a frequency-dependent learning speed, and hence a learning bias towards low frequency functions. This fact plays an important role in the results we present since the empirical score contains a low frequency part that is close to the population score, and a high-frequency part that is dataset-dependent. To the best of our knowledge, the training time to learn the high-frequency part and hence memorize, that we find to scale with n , has not been studied from this perspective in the context of score-based generative methods.

Setting: generative diffusion and score learning. Standard DMs define a transport from a target distribution P_0 in \mathbb{R}^d to a Gaussian white noise $\mathcal{N}(0, \mathbf{I}_d)$ through a *forward process* defined as an Ornstein-Uhlenbeck (OU) stochastic differential equation (SDE):

$$d\mathbf{x} = -\mathbf{x}(t)dt + dB(t), \quad (1)$$

where $dB(t)$ is square root of two times a Wiener process. Generation is performed by time-reversing the SDE (1) using the score function $s(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log P_t(\mathbf{x})$,

$$-\mathbf{dx} = [\mathbf{x}(t) + 2s(\mathbf{x}, t)] dt + dB(t), \quad (2)$$

where $P_t(\mathbf{x})$ is the probability density at time t along the forward process, and the noise $dB(t)$ is also the square root of two times a Wiener process. As shown in the seminal works [23, 58], $s(\mathbf{x}, t)$ can be obtained by minimizing the score matching loss

$$\hat{s}(\mathbf{x}, t) = \arg \min_s \mathbb{E}_{\mathbf{x} \sim P_0, \xi \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\|\sqrt{\Delta_t} s(\mathbf{x}(t), t) + \xi\|^2 \right], \quad (3)$$

where $\Delta_t = 1 - e^{-2t}$. In practice, the optimization problem is restricted to a parametrized class of functions $s_{\theta}(\mathbf{x}(t), t)$ defined, for example, by a neural network with parameters θ . The expectation over \mathbf{x} is replaced by the empirical average over the training set (n iid samples \mathbf{x}^ν drawn from P_0),

$$\mathcal{L}_t(\theta, \{\mathbf{x}^\nu\}_{\nu=1}^n) = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\|\sqrt{\Delta_t} s_{\theta}(\mathbf{x}^\nu(t)) + \xi\|^2 \right], \quad (4)$$

where $\mathbf{x}_t^\nu(\xi) = e^{-t} \mathbf{x}^\nu + \sqrt{\Delta_t} \xi$. The loss in (4) can be minimized with standard optimizers, such as stochastic gradient descent [SGD, 44] or Adam [28]. In practice, a single model conditioned on the diffusion time t is trained by integrating (4) over time [26]. The solution of the minimization of (4) is the so-called empirical score (e.g. [6, 30]), defined as $s_{\text{emp}}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log P_t^{\text{emp}}(\mathbf{x})$, with

$$P_t^{\text{emp}}(\mathbf{x}) = \frac{1}{n (2\pi\Delta_t)^{d/2}} \sum_{\nu=1}^n e^{-\frac{1}{2\Delta_t} \|\mathbf{x} - \mathbf{x}^\nu e^{-t}\|_2^2}. \quad (5)$$

This solution is known to inevitably recreate samples of the training set at the end of the generative process (i.e., it perfectly memorizes), unless n grows exponentially with the dimension d [6]. However, this is not the case in many practical applications where memorization is only observed for relatively small values of n , and disappears well before n becomes exponentially large in d . The empirical minimization performed in practice, within a given class of models and a given minimization procedure, does *not* drive the optimization to the global minimum of (4), but instead to a smoother estimate of the score that is independent of the training set with good generalization properties [24], as the global minimum of (3) would do. Understanding how it is possible, and in particular the role played by the training dynamics to avoid memorization, is the central aim of the present work.

2 Generalization and memorization during training of diffusion models

Data & architecture. We conduct our experiments on the CelebA face dataset [34], which we convert to grayscale downsampled images of size $d = 32 \times 32$, and vary the training set size n from 128 up to 32768. Our score model has a U-Net architecture [46] with three resolution levels and a base channel width of W with multipliers 1, 2 and 3 respectively. All our networks are DDPMs [20] trained to predict the injected noise at diffusion time t using SGD with momentum at fixed batch size $\min(n, 512)$. The models are all conditioned on t , i.e. a single model approximates the score at all times, and make use of a standard sinusoidal position embedding [56] that is added to the features of each resolution. More details about the numerical setup can be found in SM (Sect. A).

Evaluation metrics. To study the transition from generalization to memorization during training, we monitor the loss (4) during training using a fixed diffusion time $t = 0.01$. At various numbers of SGD updates τ , we compute the loss on all n training examples (training loss) and on a held-out test set of 2048 images (test loss). To characterize the score obtained after a training time τ , we assess the originality and quality of samples by generating 10K samples using a DDIM accelerated sampling [52]. We compute (i) the Fréchet-Inception Distance [FID, 19] against 10K test samples which we use to identify the generalization time τ_{gen} ; and (ii) the fraction of memorized generated

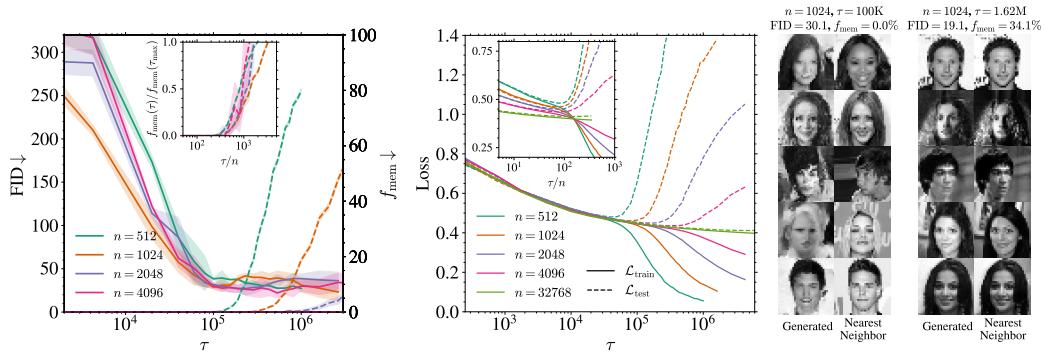


Figure 2: Memorization transition as a function of the training set size n for U-Net score models on CelebA. (Left) FID (solid lines, left axis) and memorization fraction f_{mem} (dashed lines, right axis) against training time τ for various n . Inset: normalized memorization fraction $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\text{max}})$ with the rescaled time τ/n . (Middle) Training (solid lines) and test (dashed lines) loss with τ for several n at fixed $t = 0.01$. Inset: both losses plotted against τ/n . Error bars on the losses are imperceptible. (Right) Generated samples from the model trained with $n = 1024$ for $\tau = 100K$ or $\tau = 1.62M$ steps, along with their nearest neighbors in the training set.

samples $f_{\text{mem}}(\tau)$ granting access to τ_{mem} , the memorization time. Following previous numerical studies [63, 18], a generated sample \mathbf{x}_τ is considered memorized if

$$\mathbb{E}_{\mathbf{x}_\tau} \left[\frac{\|\mathbf{x}_\tau - \mathbf{a}^{\mu_1}\|_2}{\|\mathbf{x}_\tau - \mathbf{a}^{\mu_2}\|_2} \right] < k, \quad (6)$$

where \mathbf{a}^{μ_1} and \mathbf{a}^{μ_2} are the nearest and second nearest neighbors of \mathbf{x}_τ in the training set in the L_2 sense. In what follows, we choose to work with $k = 1/3$ [63, 18], but we checked that varying k to 1/2 or 1/4 does not impact the claims about the scaling. Error bars in the figures correspond to twice the standard deviation over 5 different test sets for FIDs, and 5 noise realizations for $\mathcal{L}_{\text{train}}$ and $\mathcal{L}_{\text{test}}$. For f_{mem} , we report the 95% CIs on the mean evaluated with 1,000 bootstrap samples.

Role of training set size on the learning dynamics. At fixed model capacity ($p = 4 \times 10^6$, base width $W = 32$), we investigate how the training set size n impacts the previous metrics. In the left panel of Fig. 2, we first report the FID (solid lines) and $f_{\text{mem}}(\tau)$ (dashed lines) for various n . All trainings dynamics exhibit two phases. First, the FID quickly decreases to reach a minimum value on a timescale τ_{gen} ($\approx 100K$) that does not depend on n . In the right panel, the generated samples at $\tau = 100K$ clearly differ from their nearest neighbors in the training set, indicating that the model generalizes correctly. Beyond this time, the FID remains flat. $f_{\text{mem}}(\tau)$ is zero until a later time τ_{mem} after which it increases, clearly signaling the entrance into a memorization regime, as illustrated by the generated samples in the right-most panel of Fig. 2, very close to their nearest neighbors. Both the transition time τ_{mem} and the value of the final fraction $f_{\text{mem}}(\tau_{\text{max}})$ (with τ_{max} being one to four million SGD steps) vary with n . The inset plot shows the normalized memorization fraction $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\text{max}})$ against the rescaled time τ/n , making all curves collapse and increase at around $\tau/n \approx 300$, showing that $\tau_{\text{mem}} \propto n$, and demonstrating the existence of a generalization window for $\tau \in [\tau_{\text{gen}}, \tau_{\text{mem}}]$ that widens linearly with n , as illustrated in the left panel of Fig. 1.

As highlighted in the introduction, memorization in DMs is ultimately driven by the overfitting of the empirical score $s_{\text{mem}}(\mathbf{x}, t)$. The evolution of $\mathcal{L}_{\text{train}}(\tau)$ and $\mathcal{L}_{\text{test}}(\tau)$ at fixed $t = 0.01$ are shown in the middle panel of Fig. 2 for n ranging from 512 to 32768. Initially, the two losses remain nearly indistinguishable, indicating that the learned score $s_\theta(\mathbf{x}, t)$ does not depend on the training set. Beyond a critical time, $\mathcal{L}_{\text{train}}$ continues to decrease while $\mathcal{L}_{\text{test}}$ increases, leading to a nonzero generalization loss whose magnitude depends on n . As n increases, this critical time also increases and, eventually, the training and test loss gap shrinks: for $n = 32768$, the test loss remains close to the training loss, even after 11 million SGD steps. The inset shows the evolution of both losses with τ/n , demonstrating that the overfitting time scales linearly with the training set size n , just like τ_{mem} identified in the left panel. Moreover, there is a consistent lag between the overfitting time and τ_{mem} at fixed n , reflecting the additional training required for the model to overfit the empirical score sufficiently to reproduce the training samples, and therefore to impact the memorization fraction.

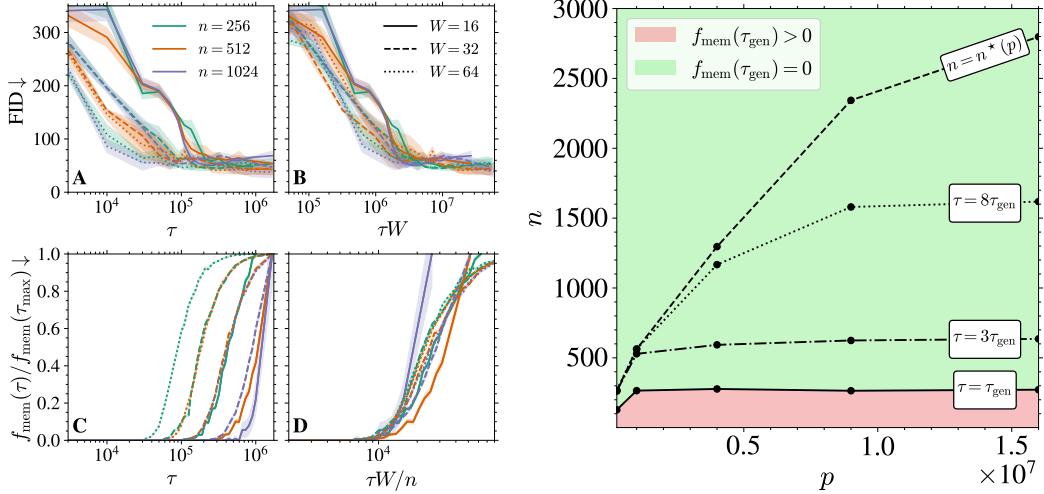


Figure 3: **Effect of the number of parameters in the U-Net architecture on the timescales of the training dynamics.** (Left) FID (panels **A**, **B**) and normalized memorization fraction $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\text{max}})$ (panels **C**, **D**) for various n and W during training. In panels **B** and **D**, time is rescaled such that all curves collapse. (Right) (n, p) phase diagram of generalization vs memorization for U-Nets trained on CelebA. Curves show, for $\tau \in \{\tau_{\text{gen}}, 3\tau_{\text{gen}}, 8\tau_{\text{gen}}\}$, the minimal dataset size $n(p)$ satisfying $f_{\text{mem}}(\tau) = 0$. The shaded background indicates the memorization-generalization boundary for $\tau = \tau_{\text{gen}}$.

Memorization is *not* due to data repetition. We must stress that this delayed memorization with n is *not* due to the mere repetition of training samples, as a first intuition could suggest. In SM Sects. A and B, we show that full-batch updates still yield $\tau_{\text{mem}} \propto n$. In other words, even if at fixed τ all models have processed each sample equally often, larger n consistently postpone memorization. This confirms that memorization in DMs is driven by a fundamental n -dependent change in the loss landscape – not by a sample repetition during training.

Effect of the model capacity. To study more precisely the role of the model capacity on the memorization–generalization transition, we vary the number of parameters p by changing the U-Nets base width $W \in \{8, 16, 32, 48, 64\}$, resulting in a total of $p \in \{0.26, 1, 4, 9, 16\} \times 10^6$ parameters. In the left panel of Fig. 3, we plot both the FID (top row) and the normalized memorization fraction (bottom row) as functions of τ for several width W and training set sizes n . Panels **A** and **C** demonstrate that higher-capacity networks (larger W) achieve high-quality generation and begin to memorize *earlier* than smaller ones. Panels **B** and **D** show that the two characteristic timescales simply scale as $\tau_{\text{gen}} \propto W^{-1}$ and $\tau_{\text{mem}} \propto nW^{-1}$. In particular, this implies that, for $W > 8$, the critical training set size $n_{\text{gm}}(p)$ at which $\tau_{\text{mem}} = \tau_{\text{gen}}$ is approximately independent of p (at least on the limited values of p we focused on). When $n > n_{\text{gm}}(p)$, the interval $[\tau_{\text{gen}}, \tau_{\text{mem}}]$ opens up, so that early stopping within this window yields high quality samples without memorization. In the right panel of Fig. 3, we display this boundary (solid line) in the (n, p) plane by fixing the training time to $\tau = \tau_{\text{gen}}$, that we identify numerically using the collapse of all FIDs at around $W\tau_{\text{gen}} \approx 3 \times 10^6$ (see panel **B**), and computing the smallest n such that $f_{\text{mem}}(\tau) = 0$. The resulting solid curve delineates two regimes: below the curve, memorization already starts at τ_{gen} ; above the curve, the models generalize perfectly under early stopping. We repeat this experiment for $\tau = 3\tau_{\text{gen}}$ and $\tau = 8\tau_{\text{gen}}$, showing saturation to larger and larger p as τ increases. Eventually, for $\tau \rightarrow \infty$, we expect these successive boundaries to converge to the architectural regularization threshold $n^*(p)$, i.e. the point beyond which the network avoids memorization because it is not expressive enough, as found in [15] and highlighted in the right panel of Fig. 1. In order to estimate $n^*(p)$, we measure for a given τ the largest $n(\tau)$ yielding $f_{\text{mem}} \approx 0$. The curve $n(\tau)$ approaches $n^*(p)$ for large τ . We therefore estimate $n^*(p)$ by measuring the asymptotic values of $n(\tau)$, which in practice is reached already at $\tau = \tau_{\text{max}} = 2M$ updates for the values of W we focus on.

3 Training dynamics of a Random Features Network

Notations. We use bold symbols for vectors and matrices. The L^2 norm of a vector \mathbf{x} is denoted by $\|\mathbf{x}\| = (\sum_i \mathbf{x}_i^2)^{1/2}$. We write $f = \mathcal{O}(g)$ to mean that in the limit $n, p \rightarrow \infty$, there exists a constant C such that $|f| \leq C|g|$.

Setting. We study analytically a model introduced in [15], where the data lie in d dimensions. We parametrize the score with a Random Features Neural Network [RFNN, 43]

$$\mathbf{s}_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{A}}{\sqrt{p}} \sigma \left(\frac{\mathbf{W} \mathbf{x}}{\sqrt{d}} \right). \quad (7)$$

An RFNN, illustrated in Fig. 4 (left), is a two-layer neural-network whose first layer weights ($\mathbf{W} \in \mathbb{R}^{p \times d}$) are drawn from a Gaussian distribution and remain frozen while the second layer weights ($\mathbf{A} \in \mathbb{R}^{d \times p}$) are learned during training. This model has already served as theoretical framework for studying several behaviors of deep neural network such as the double descent phenomenon [35, 13]. σ is an element-wise non-linear activation function. We consider a training set of n iid samples $\mathbf{x}^\nu \sim P_{\mathbf{x}}$ for $\nu = 1, \dots, n$ and we focus on the high-dimensional limit $d, p, n \rightarrow \infty$ with the ratios $\psi_p = p/d, \psi_n = n/d$ kept fixed. We study the training dynamics associated to the minimization of the empirical score matching loss defined in (4) at a fixed diffusion time t . This is a simplification compared to practical methods, which use a single model for all t . It has been already studied in previous theoretical works [11, 15]. The loss (4) is rescaled by a factor $1/d$ in order to ensure a finite limit at large d . We also study the evolution of the test loss evaluated on test points and the distance to the exact score $\mathbf{s}(\mathbf{x}) = \nabla \log P_{\mathbf{x}}$,

$$\mathcal{L}_{\text{test}} = \frac{1}{d} \mathbb{E}_{\mathbf{x}, \xi} \left[\|\sqrt{\Delta_t} \mathbf{s}_{\mathbf{A}}(\mathbf{x}_t(\xi)) + \xi\|^2 \right], \quad \mathcal{E}_{\text{score}} = \frac{1}{d} \mathbb{E}_{\mathbf{x}} [\|\mathbf{s}_{\mathbf{A}}(\mathbf{x}) - \nabla \log P_{\mathbf{x}}\|^2], \quad (8)$$

where the expectations $\mathbb{E}_{\mathbf{x}, \xi}$ are computed over $\mathbf{x} \sim P_{\mathbf{x}}$ and $\xi \sim \mathcal{N}(0, \mathbf{I}_d)$. The generalization loss, defined as $\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$, indicates the degree of overfitting in the model while the distance to the exact score $\mathcal{E}_{\text{score}}$ measures the quality of the generation as it is an upper bound on the Kullback–Leibler divergence between the target and generated distributions [53, 8]. The weights \mathbf{A} are updated via gradient descent

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} - \eta \nabla_{\mathbf{A}} \mathcal{L}_{\text{train}}(\mathbf{A}^{(k)}), \quad (9)$$

where η is the learning rate. In the high-dimensional limit, as the learning rate $\eta \rightarrow 0$, and after rescaling time as $\tau = k\eta/d^2$, the discrete-time dynamics converges to the following continuous-time gradient flow:

$$\dot{\mathbf{A}}(\tau) = -d^2 \nabla_{\mathbf{A}} \mathcal{L}_{\text{train}}(\mathbf{A}(\tau)) = -2\Delta_t \frac{d}{p} \mathbf{A} \mathbf{U} - \frac{2d\sqrt{\Delta_t}}{\sqrt{p}} \mathbf{V}^T, \quad (10)$$

with

$$\mathbf{U} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[\sigma \left(\frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \sigma \left(\frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right)^T \right], \quad \mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[\sigma \left(\frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \xi^T \right]. \quad (11)$$

Assumptions. For our analytical results to hold, we make the following mathematical assumptions which are standard when studying Random Features [41, 17, 22] namely (i) the activation function σ admits a Hermite polynomial expansion $\sigma(x) = \sum_{s=0}^{\infty} \frac{\alpha_s}{s!} H e_s(x)$; and (ii) the data distribution $P_{\mathbf{x}}$ has sub-Gaussian tails and a covariance $\Sigma = \mathbb{E}_{P_{\mathbf{x}}}[\mathbf{x} \mathbf{x}^T]$ with bounded spectrum. We assume that the empirical distribution of eigenvalues of Σ converges weakly in the high dimensional limit to a deterministic density $\rho_{\Sigma}(\lambda)$ and that $\text{Tr}(\Sigma)/d$ converges to a finite limit (for a more precise mathematical statement see SM Sect. C.3). Moreover, we make additional assumptions that are not essential to the proofs but which simplify the analysis: (iii) the activation function σ verifies $\mu_0 = \mathbb{E}_z[\sigma(z)] = 0$; and (iv) the second layer \mathbf{A} is initialized with zero weights $\mathbf{A}(\tau = 0) = 0$. In numerical applications, unless specified, we use $\sigma(z) = \tanh(z)$ and $P_{\mathbf{x}} = \mathcal{N}(0, \mathbf{I}_d)$.

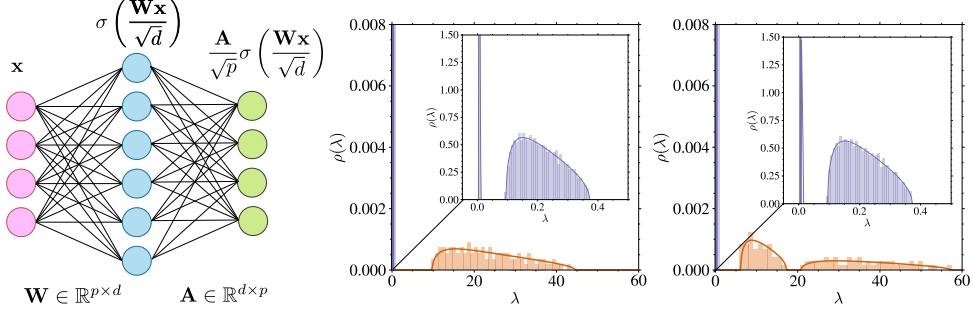


Figure 4: (Left) **Illustration of an RFNN.** (Middle/Right) **Spectrum of \mathbf{U} .** Density $\rho(\lambda)$ from Theorem 3.1 in the overparameterized Regime I described in Theorem 3.2, with $\psi_p = 64$, $\psi_n = 8$, $t = 0.01$, and $\rho_{\Sigma}(\lambda) = \delta(\lambda - 1)$. The bulk of the spectrum (orange) is between $\lambda \approx 10$ and $\lambda \approx 45$. The histogram shows the eigenvalues from a single realization of \mathbf{U} at $d = 100$. Inset: zoom near $\lambda = 0$ (in blue) showing the first bulk ρ_1 and the delta peak at $\lambda = s_t^2$. (Right) Same as (Middle), but with $\rho_{\Sigma}(\lambda) = \frac{1}{2}\delta(\lambda - 0.5) + \frac{1}{2}\delta(\lambda - 1.5)$. The first bulk in blue remains unchanged, as it depends only on $\sigma_x^2 = \text{Tr}(\Sigma)/d = 1$ in both cases, while the second bulk varies with Σ .

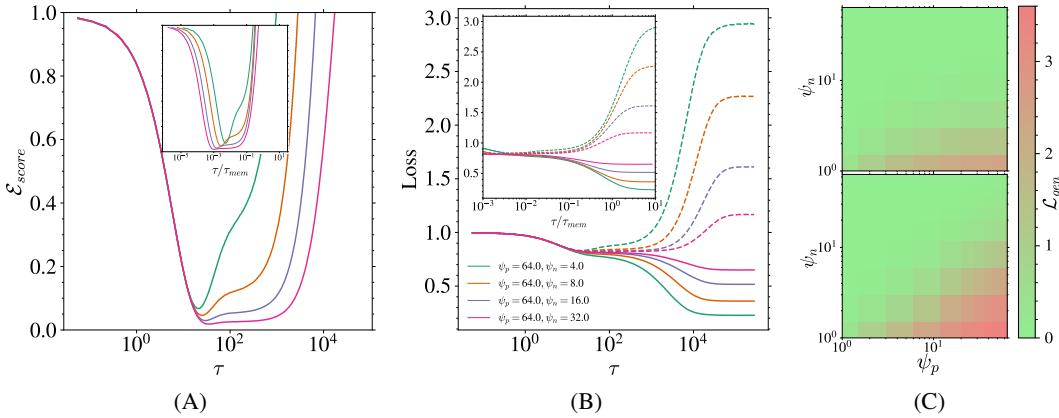


Figure 5: **Evolution of the training and test losses for the RFNN.** (A) Distance to the true score $\mathcal{E}_{\text{score}}$ against training time τ for $\psi_n = 4, 8, 16, 32, \psi_p = 64, t = 0.1$ and $d = 100$. In the inset, the training time is rescaled by $\tau_{\text{mem}} = \psi_p / \Delta_t \lambda_{\min}$. (B) Training (solid) and test (dashed) losses for various ψ_n . The inset shows both losses rescaled by τ_{mem} . (C) Heatmaps of \mathcal{L}_{gen} for $\tau = 10^3$ (top) and $\tau = 10^4$ (bottom) as a function of ψ_n and ψ_p . All the curves use Pytorch [38] gradient descent. More numerical details can be found in SM Sect. D.

Emergence of the two timescales during training. We first show in Fig. 5 that the behavior of training and test losses in the RF model mirrors the one found in realistic cases in Sect. 2, with a separation of timescales τ_{gen} and τ_{mem} which increases with n . Equation (10) is linear in \mathbf{A} and hence it can be solved exactly (see SM). The timescales of the training dynamics are given by the inverse eigenvalues of the $p \times p$ matrix $\Delta_t \mathbf{U} / \psi_p$. Building on the Gaussian Equivalence Principle [GEP, 16, 17, 36] and the theory of linear pencils [7], George et al. (2025) derive a coupled system of equations characterizing the Stieltjes transform of the eigenvalue density $\rho(\lambda)$ of \mathbf{U} for isotropic Gaussian data that lie in a D -dimensional subspace with $D \leq d$ and $D = \mathcal{O}(d)$. We offer an alternative derivation presented in SM for general variance using the replica method [37] – a heuristic method from the statistical physics of disordered systems – yielding the more compact formulation for obtaining the spectrum stated in Theorem 3.1. Before stating the theorem, we introduce

$$b_t = \mathbb{E}_{u,v}[v\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)], \quad a_t = \mathbb{E}_{u,v}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\frac{u}{e^{-t}\sigma_{\mathbf{x}}}], \quad (12)$$

$$v_t^2 = \mathbb{E}_{u,v,w}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}w)] - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2, \quad (13)$$

$$s_t^2 = \mathbb{E}_u[\sigma(\Gamma_t u)^2] - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 - v_t^2 - b_t^2, \quad (14)$$

where $\sigma_{\mathbf{x}}^2 = \frac{\text{Tr}(\Sigma)}{d}$, $\Gamma_t = e^{-2t}\sigma_{\mathbf{x}}^2 + \Delta_t = 1 + e^{-2t}(\sigma_{\mathbf{x}}^2 - 1)$ and the expectation is over the u, v, w random variables which are independent standard Gaussian $\mathcal{N}(0, 1)$.

Theorem 3.1. Let $q(z) = \frac{1}{p} \text{Tr}(\mathbf{U} - z\mathbf{I}_p)^{-1}$, $r(z) = \frac{1}{p} \text{Tr}(\Sigma^{1/2} \mathbf{W}^T (\mathbf{U} - z\mathbf{I}_p)^{-1} \mathbf{W} \Sigma^{1/2})$ and $s(z) = \frac{1}{p} \text{Tr}(\mathbf{W}^T (\mathbf{U} - z\mathbf{I}_p)^{-1} \mathbf{W})$, with $z \in \mathbb{C}$. Let

$$\hat{s}(q) = b_t^2 \psi_p + \frac{1}{q}, \quad (15)$$

$$\hat{r}(r, q) = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q}. \quad (16)$$

Then $q(z), r(z)$ and $s(z)$ satisfy the following set of three equations:

$$s = \int d\rho_{\Sigma}(\lambda) \frac{1}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (17)$$

$$r = \int d\rho_{\Sigma}(\lambda) \frac{\lambda}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (18)$$

$$\psi_p(s_t^2 - z) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q} + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0, \quad (19)$$

The eigenvalue distribution of \mathbf{U} , $\rho(\lambda)$, can then be obtained using the Sokhotski–Plemelj inversion formula $\rho(\lambda) = \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\pi} \text{Im} q(\lambda + i\varepsilon)$.

We now focus on the asymptotic regime $\psi_p, \psi_n \gg 1$, typical for strongly over-parameterized models trained on large data sets. In this limit, the spectrum of \mathbf{U} can be described analytically by the following Theorem 3.2.

Theorem 3.2 (Informal). Let ρ denote the spectral density of \mathbf{U} .

Regime I (overparametrized): $\psi_p > \psi_n \gg 1$.

$$\rho(\lambda) = \left(1 - \frac{1 + \psi_n}{\psi_p}\right) \delta(\lambda - s_t^2) + \frac{\psi_n}{\psi_p} \rho_1(\lambda) + \frac{1}{\psi_p} \rho_2(\lambda).$$

Regime II (underparametrized): $\psi_n > \psi_p \gg 1$.

$$\rho(\lambda) = \left(1 - \frac{1}{\psi_p}\right) \rho_1(\lambda) + \frac{1}{\psi_p} \rho_2(\lambda).$$

where ρ_1 is an atomless measure with support

$$\left[s_t^2 + v_t^2 \left(1 - \sqrt{\psi_p/\psi_n}\right)^2, s_t^2 + v_t^2 \left(1 + \sqrt{\psi_p/\psi_n}\right)^2 \right],$$

and ρ_2 coincides with the asymptotic eigenvalue bulk density of the population covariance $\tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{X}}[\mathbf{U}]$; ρ_2 is independent of ψ_n and its support is on the scale ψ_p . The eigenvectors associated with $\delta(\lambda - s_t^2)$ leave both training and test losses unchanged and are therefore irrelevant. In the limit $\psi_p \gg \psi_n$, the supports of ρ_1 and ρ_2 are respectively on the scales ψ_p/ψ_n and ψ_p , i.e. they are well separated.

The proofs of both theorems are shown in SM (Sect. C). We recall that training timescales are directly related to eigenvalues λ via the relation $\tau^{-1} = \psi_p/\Delta_t \lambda_{\min}$. Theorem 3.2 therefore demonstrates the emergence of the two training timescales τ_{mem} and τ_{gen} in the overparametrized regime of the RFNN model. They are respectively associated to the measures ρ_1 and ρ_2 , which are well separated in regime I, for $\psi_p \gg \psi_n \gg 1$, as shown in Fig. 4.

Generalization: The timescale τ_{gen} on which the first relaxation takes place is associated to the formation of the generalization regime. It is related to the bulk ρ_2 and is of order $1/\Delta_t$. This regime only depends on the population covariance Σ of the data and is independent of the specific realization

of the dataset. On this timescale, which is of order one, both the training $\mathcal{L}_{\text{train}}$ and test $\mathcal{L}_{\text{test}}$ losses decrease. The generalization loss $\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$ is zero, and $\mathcal{E}_{\text{score}}$ tends to a value that we find to scale as $\mathcal{O}(\psi_n^{-\eta})$ with $\eta \simeq 0.59$ numerically (see Fig. 5).

Memorization: The timescale τ_{mem} , on which the second stage of the dynamics takes place, is associated to overfitting and memorization. It is related to the bulk ρ_1 , and scales as $\psi_p/\Delta_t \lambda_{\min}$, where λ_{\min} is the left edge of ρ_1 . In the overparameterized regime $p \gg n$, τ_{mem} becomes large and of order ψ_n/Δ_t , thus implying a scaling of τ_{mem} with n . On this timescale, the training loss decreases while the test loss increases, converging to their respective asymptotic values as computed in [15]. Fig. 5 indeed shows that all training and test curves separate, correspondingly the generalization loss \mathcal{L}_{gen} increases, at a time that scales with $\psi_p/\Delta_t \lambda_{\min}$, as shown in the inset.

As n increases, the asymptotic ($\tau \rightarrow \infty$) generalization loss \mathcal{L}_{gen} decreases, indicating a reduced overfitting. For $n > n^*(p) = p$, although some overfitting remains (i.e., $\mathcal{L}_{\text{gen}} > 0$), the value of \mathcal{L}_{gen} is sensibly reduced, and the model is no longer expressive enough to memorize the training data, as shown in [15]. This regime corresponds to the *Architectural Regularization* phase in Fig. 1. We show in Fig. 5 (panel C) how the generalization loss \mathcal{L}_{gen} varies in the (n, p) plane depending on the time τ at which training is stopped. In agreement with the above results, we find that the generalization–memorization transition line depends on τ and moves upward for larger values of τ , similarly to the numerical results exposed in Fig. 3 and the illustration in Fig. 1.

4 Conclusions

We have shown that the training dynamics of neural network-based score functions display a form of implicit regularization that prevents memorization even in highly overparameterized diffusion models. Specifically, we have identified two well-separated timescales in the learning: τ_{gen} , at which models begin to generate high-quality, novel samples, and τ_{mem} , beyond which they start to memorize the training data. The gap between these timescales grows with the size of the training set, leading to a broad window where early stopped models generate novel samples of high-quality. We have demonstrated that this phenomenon happens in realistic settings, for controlled synthetic data, and in analytically tractable models. Although our analysis focuses on DMs, the underlying score-learning mechanism we uncover is common to all score-based generative models such as stochastic interpolants [3] or flow matching [32]; we therefore expect our results to generalize to this broader class.

Limitations and future works.

- While we derived our results under SGD optimization, most DMs are trained in practice with Adam [28]. In SM Sects. A and D, we show that the two key timescales still arise using Adam, although with much fewer optimization steps. Studying how different optimizers shift these timescales would be valuable for practical usage.
- All experiments in Sect. 2 are conducted with unconditional DMs. We additionally verify in SM Sect. B, using a toy Gaussian mixture dataset and classifier-free guidance [21], that the same scaling of τ_{mem} with n holds in the conditional settings. Understanding precisely how the absolute timescales τ_{mem} and τ_{gen} depend on the conditioning remains an open question.
- Our numerical experiments cover a range of p between 1M and 16M. Exploring a wider range is essential to map the full (n, p) phase diagram sketched in Fig. 1 and understand the precise effect of expressivity on dynamical regularization.
- Finally, our theoretical analysis rely on well-controlled data and score models that reproduce the core effects. Extending these analytical frameworks to richer data distributions (such as Gaussian mixtures or data from the hidden manifold model) and to structured architectures would be valuable to further characterize the implicit dynamical regularization of training score-functions. In particular investigating how heavy-tailed data distribution [2] affect the picture described here could be valuable.
- Although DMs trained on large and diverse datasets likely avoid the memorization regime we study here, some industrial models were shown to exhibit partial memorization [9, 50]. Our results provide practical guidelines (early-stopping, control the network capacity) to train DMs robustly and hence avoid memorization, which can be especially helpful in data-scarce domains (e.g., physical sciences).

Acknowledgments and Disclosure of Funding

The authors thank Valentin De Bortoli for initial motivating discussions on memorization-generalization transitions. RU thanks Beatrice Achilli, Jérôme Garnier-Brun, Carlo Lucibello and Enrico Ventura for insightful discussions. RU is grateful to Bocconi University for its hospitality during his stay, during which part of this work was conducted. This work was performed using HPC resources from GENCI-IDRIS (Grant 2025-AD011016319). GB acknowledges support from the French government under the management of the Agence Nationale PR[AI]RIE-PSAI (ANR-23-IACL-0008). MM acknowledges the support of the PNRR-PE-AI FAIR project funded by the NextGeneration EU program. After completing this work, we became aware that A. Favero, A. Sclocchi, and M. Wyart [14] had also been investigating the memorization–generalization transition from a similar perspective.

References

- [1] Achilli, B., Ventura, E., Silvestri, G., Pham, B., Raya, G., Krotov, D., Lucibello, C., and Ambrogioni, L. (2024). Losing dimensions: Geometric memorization in generative diffusion.
- [2] Adomaityte, U., Defilippis, L., Loureiro, B., and Sicuro, G. (2024). High-dimensional robust regression under heavy-tailed data: asymptotics and universality. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(11):114002.
- [3] Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. (2023). Stochastic interpolants: A unifying framework for flows and diffusions.
- [4] Bach, F. (2023). Polynomial magic iii: Hermite polynomials. <https://francisbach.com/hermite-polynomials/>. Accessed: 2025-10-09.
- [5] Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442.
- [6] Biroli, G., Bonnaire, T., de Bortoli, V., and Mézard, M. (2024). Dynamical regimes of diffusion models. *Nature Communications*, 15(9957). Open access.
- [7] Bodin, A. P. M. (2024). *Random Matrix Methods for High-Dimensional Machine Learning Models*. Phd thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- [8] Bortoli, V. D. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*. Expert Certification.
- [9] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium, SEC ’23*, USA. USENIX Association.
- [10] Chen, C., Liu, D., and Xu, C. (2024). Towards memorization-free diffusion models.
- [11] Cui, H., Krzakala, F., Vanden-Eijnden, E., and Zdeborova, L. (2024). Analysis of learning a flow-based generative model from limited sample complexity. In *The Twelfth International Conference on Learning Representations*.
- [12] Cui, H., Pehlevan, C., and Lu, Y. M. (2025). A precise asymptotic analysis of learning diffusion models: theory and insights.
- [13] D’Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. (2020). Double trouble in double descent: Bias and variance(s) in the lazy regime. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2280–2290. PMLR.
- [14] Favero, A., Sclocchi, A., and Wyart, M. (2025). Bigger isn’t always memorizing: Early stopping overparameterized diffusion models.
- [15] George, A. J., Veiga, R., and Macris, N. (2025). Denoising score matching with random features: Insights on diffusion models from precise learning curves.

- [16] Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR.
- [17] Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. (2021). The gaussian equivalence of generative models for learning with shallow neural networks.
- [18] Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. (2023). On memorization in diffusion models.
- [19] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium.
- [20] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.
- [21] Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance.
- [22] Hu, H. and Lu, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964.
- [23] Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.
- [24] Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (2024). Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*.
- [25] Kamb, M. and Ganguli, S. (2024). An analytic theory of creativity in convolutional diffusion models.
- [26] Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models.
- [27] Kibble, W. F. (1945). An extension of a theorem of mehler’s on hermite polynomials. *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1):12–15.
- [28] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *ICLR (Poster)*.
- [29] Li, P., Li, Z., Zhang, H., and Bian, J. (2025). On the generalization properties of diffusion models.
- [30] Li, S., Chen, S., and Li, Q. (2024a). A good score does not lead to a good generative model.
- [31] Li, T., Biferale, L., Bonacorso, F., and et al. (2024b). Synthetic lagrangian turbulence by generative diffusion models. *Nat Mach Intell*, 6:393–403.
- [32] Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.
- [33] Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., He, L., and Sun, L. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models.
- [34] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- [35] Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR.
- [36] Mei, S. and Montanari, A. (2020). The generalization error of random features regression: Precise asymptotics and double descent curve.

- [37] Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, volume 9 of *Lecture Notes in Physics*. World Scientific Publishing Company, Singapore.
- [38] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates, Inc.
- [39] Potters, M. and Bouchaud, J.-P. (2020). *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press.
- [40] Price, I., Sanchez-Gonzalez, A., Alet, F., and et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90.
- [41] Péché, S. (2019). A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7.
- [42] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. (2019). On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR.
- [43] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.
- [44] Robbins, H. and Monroe, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.
- [45] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- [46] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.
- [47] Shah, K., Kalavasis, A., Klivans, A. R., and Daras, G. (2025). Does generation require memorization? creative diffusion models using ambient diffusion.
- [48] Silverstein, J. and Bai, Z. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192.
- [49] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.
- [50] Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2023a). Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [51] Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2023b). Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803.
- [52] Song, J., Meng, C., and Ermon, S. (2022). Denoising diffusion implicit models.
- [53] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models.

- [54] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- [55] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021b). Score-based generative modeling through stochastic differential equations.
- [56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [57] Ventura, E., Achilli, B., Silvestri, G., Lucibello, C., and Ambrogioni, L. (2025). Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion.
- [58] Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674.
- [59] Wang, B. (2025). An analytical theory of power law spectral bias in the learning dynamics of diffusion models.
- [60] Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering.
- [61] Wen, Y., Liu, Y., Chen, C., and Lyu, L. (2024). Detecting, explaining, and mitigating memorization in diffusion models.
- [62] Wu, Y.-H., Marion, P., Biau, G., and Boyer, C. (2025). Taking a big step: Large learning rates in denoising score matching prevent memorization.
- [63] Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. (2023). Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.
- [64] Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S.-H., and Kweon, I. S. (2023). A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai.
- [65] Zhi-Qin John Xu, Z.-Q. J. X., Yaoyu Zhang, Y. Z., Tao Luo, T. L., Yanyang Xiao, Y. X., and Zheng Ma, Z. M. (2020). Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767.

Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training

Supplementary Material (SM)

Tony Bonnaire[†], Raphaël Urfin[†], Giulio Biroli, Marc Mézard

This document provides detailed derivations and additional experiments supporting the main text (MT). In Sect. A, we give details about the numerical experiments carried out in Sect. 2. In Sect. B we provide additional numerical experiments on simplified score and data models. Sect. C gives formal proofs of the main theorems of Sect. 3. Finally, Sect. D exposes more details on the numerical experiments of Sect. 3.

A Numerical experiments on CelebA

A.1 Details on the numerical setup

Dataset. All numerical experiments in Sect. 2 of the MT use the CelebA face dataset [34]. We center-crop each RGB image to 32×32 pixels and convert to grayscale images in order to accelerate the training of our Diffusion Models (DMs). To precisely control the samples seen by a model, no data augmentation is applied, and we vary the training set size n in the window $[128, 32768]$. Examples of training samples are shown in the left-most block of Fig. 6.

Architecture. As commonly done in DDPMs implementations [e.g., 20, 55], the network approximating the score function is a U-Net [46] made of three resolution levels, each containing two residual blocks with channel multipliers $\{1, 2, 4\}$ respectively. We apply attention to the two coarsest resolutions, and embed the diffusion time via sinusoidal position embedding [56]. The base channel width W varies from 16 to 64 depending on the experiment, resulting in a total of 1 to 16 million trainable parameters.

Time reparameterization. Compared to the framework presented in the MT, the DDPMs we train make use of a time reparameterization of the forward and backward processes with a variance schedule $\{\beta_{t'}\}_{t'=1}^T$, where T is the time horizon given as a number of steps, fixed to 1000 in our experiments. The variance is evolving linearly from $\beta_1 = 10^{-4}$ to $\beta_{1000} = 2 \times 10^{-2}$. A sample at time t' , denoted $\mathbf{x}(t')$, can be expressed from $\mathbf{x}(0)$ as the following interpolation

$$\mathbf{x}(t') = \sqrt{\bar{\alpha}(t')} \mathbf{x}(0) + \sqrt{1 - \bar{\alpha}(t')} \boldsymbol{\xi}, \quad (20)$$

where $\bar{\alpha}(t') = \prod_{s=1}^{t'} (1 - \beta_s)$, and $\boldsymbol{\xi}$ is a standard and centered Gaussian noise. This is a reparameterization of the Ornstein-Uhlenbeck process from Eq. 1 defined through time t in the MT, with

$$t = -\frac{1}{2} \log(\bar{\alpha}(t')). \quad (21)$$

Training. All DMs are trained with Stochastic Gradient Descent (SGD) at fixed learning rate $\eta = 0.01$, fixed momentum $\beta = 0.95$ and batch size $B = \min(n, 512)$. We focus on SGD to facilitate the analysis of time scaling, avoiding problems that may cause alternative adaptive optimization schemes like Adam [28]. We train each model for at least 2M SGD steps, sometimes more for large values of n displaying memorization only later. We do not employ exponential moving average or learning-rate warm-up.

Generation. To accelerate sampling while preserving FID, we employ the DDIM sampler of Song et al. (2022) which replaces the Markovian reverse SDE with a deterministic, non-Markovian update. Given a trained denoiser $\boldsymbol{\xi}_{\theta}(\mathbf{x}_t, t)$, we iterate for $t = T', \dots, 1$

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}(t-1)} \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}(t)} \boldsymbol{\xi}_{\theta}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}(t)}} + \sqrt{1 - \bar{\alpha}(t-1)} \boldsymbol{\xi}_{\theta}(\mathbf{x}_t, t), \quad (22)$$

with $T' = 200$. During training, we generate at 40 milestones a set of 10,000 samples to assess generalization and memorization. Examples of samples obtained from a model trained on $n = 1024$

samples with base width $W = 32$ are shown in the middle and right blocks from Fig. 6 for two training times, $\tau = 190\text{K}$ and $\tau = 1.62\text{M}$. At $\tau = 190\text{K}$ the model generalizes ($f_{\text{mem}} = 0\%$) and achieve a test FID of 35.1. After too much training, memorization sets in and, by $\tau = 1.62\text{M}$ steps, nearly half the generated samples reproduce training images ($f_{\text{mem}} = 47.2\%$).

Statistical evaluation. FIDs [19] are computed[†] using 10,000 generated samples and 10,000 test samples, averaged over 5 independent runs with disjoint test sets. Error bars in the MT denote twice the standard deviation. Training and test losses are estimated similarly over 5 repeated evaluation on n training samples and 2048 test samples, and give negligible confidence intervals. For the memorization fraction $f_{\text{mem}}(\tau)$, we report the standard error on the mean obtained via bootstrap resampling of the 10,000 generated samples. We also verified that the scaling in the memorization time τ_{mem} is insensitive to the choice of the threshold k used to define f_{mem} in Eq. 6 by testing larger and lower values.

Computing resources. Most trainings were performed on Nvidia H100 GPUs (80GB of memory). A typical run of 2M steps takes approximately 50 hours on two GPUs and vary with the model size (defined through its base width W). In total, we train 18 distinct models for the several n, W configurations of the MT. The longest training ($n = 32768$ and $W = 32$ in Fig. 2) ran for 11M steps. The generation of 10,000 samples over 40 training times takes around an additional hour per model on the same hardware support.

A.2 Batch-size effect: repetition vs. memorization

All the experiments in the MT use a fixed batch size $B = 512$, and in Sect 2 we emphasize that the observed $\mathcal{O}(n)$ scaling of τ_{mem} cannot be explained by repetition over training samples. To validate this statement, the left panel of Fig. 7 shows FID and memorization fraction curves when we train the models with full-batch updates ($B = n$) for $n \in [128, 512]$. At any fixed τ , every sample has been seen exactly τ times. Yet τ_{mem} continues to grow linearly with n , as shown in the inset. This demonstrates that larger datasets reshape the loss landscape – requiring proportionally more updates to overfit – rather than simply increasing memorization through repeated exposure of training samples.

A.3 What about Adam?

We conclude this section by repeating our analysis at fixed $W = 64$ using the Adam optimizer [28] instead of SGD with momentum. The learning rate is $\eta = 1 \times 10^{-4}$, gradient averages take values $(\beta_1, \beta_2) = (0.9, 0.999)$, and batch size $B = \min(512, n)$. We keep all other settings and evaluation metrics as above. As shown in the right panel of Fig. 7, Adam yields the same two-phase training dynamics with first a generalization regime with $f_{\text{mem}} = 0$ and good performances (small FID), and later a memorization phase at $\tau_{\text{mem}} \propto n$, as shown in the inset. The only difference is that both τ_{gen} and τ_{mem} occur after much fewer steps compared to SGD. This also points out that the emergence of the two well-separated timescales and their scaling is a fundamental property of the loss landscape.

B Generalization–memorization transition in the Gaussian Mixture Model

The aim of this section is to show our results hold for other data distributions than natural images, and alternative score model that U-Net architectures.

B.1 Settings

Data distribution. We focus on data iid sampled from a d -dimensional Gaussian Mixture Model (GMM) made of two balanced Gaussians centered on $\pm\boldsymbol{\mu}$ with unit covariance, i.e.,

$$\mathbb{P}_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d). \quad (23)$$

[†]Using the [pytorch-fid](#) Python package.

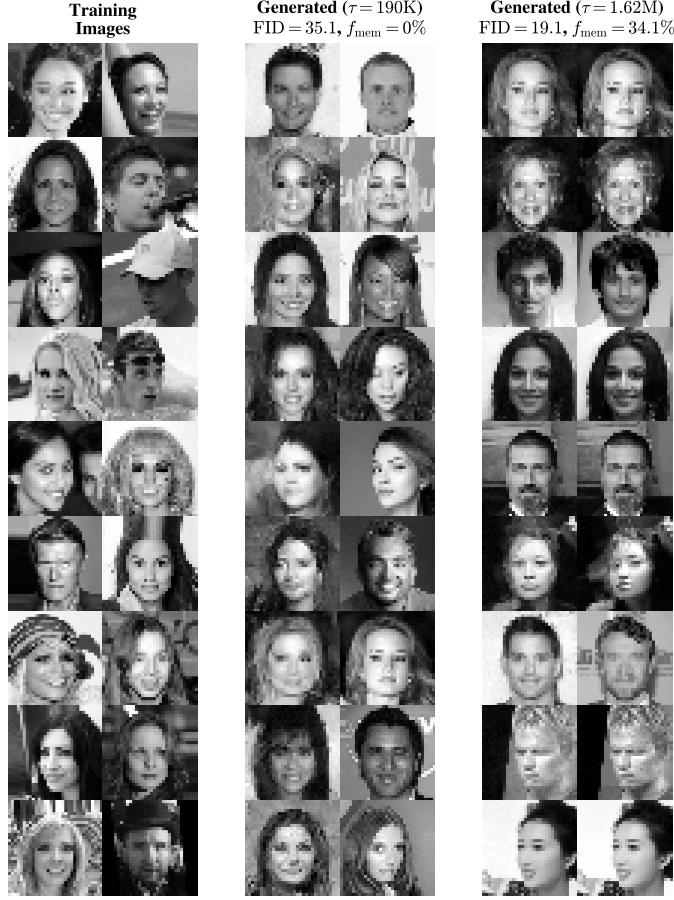


Figure 6: **Training and generation on CelebA.** The left-most block shows random training images. Middle and right blocks show generated samples in the left column (after $\tau = 190K$ and $\tau = 1.62M$ SGD updates respectively), alongside each sample’s nearest neighbor in the training set in the right column. All generated images come from model trained on $n = 1024$ with base width $W = 32$.

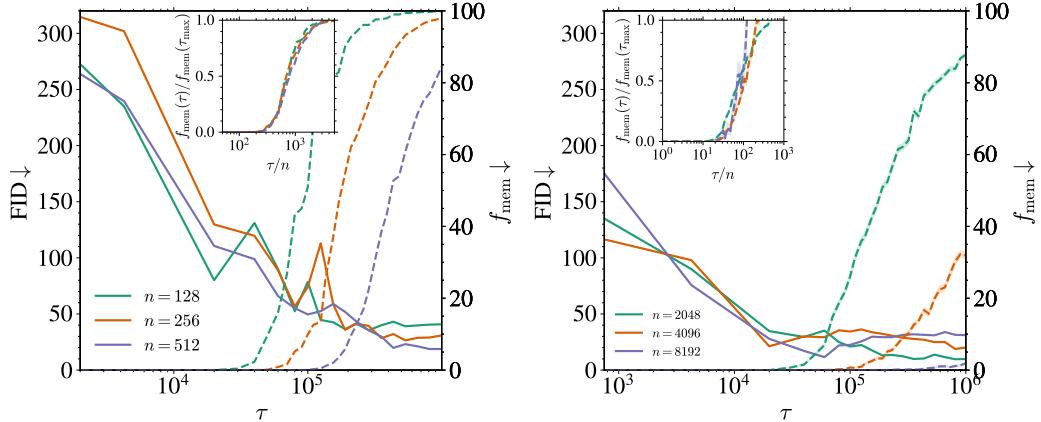


Figure 7: **Impact of batch size and optimizer on the scaling of τ_{mem} .** FID (solid lines, left axis) and memorization fraction f_{mem} (in %, dashed lines, right axis) against training time τ for various n . Inset: normalized memorization fraction $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\text{max}})$ with the rescaled time τ/n . (*Left*) Memorization scaling for $B = n$. (*Right*) Generalization–Memorization transition with Adam optimizer for $W = 64$.

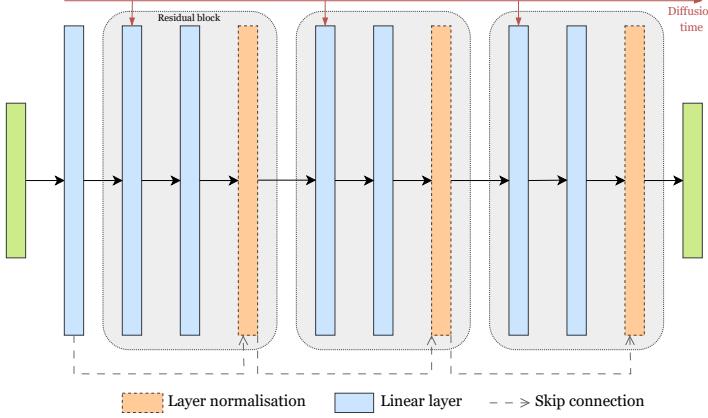


Figure 8: **Basic ResNet architecture of the GMM numerical experiments.** Residual network with three residual blocks, each made of two fully-connected layers followed by a layer normalization. The width of the network is W , and the input and output sizes are d .

In what follows, we choose to work with $\mu = \mathbf{1}_d$, with $\mathbf{1}_d = [1, \dots, 1]^\top \in \mathbb{R}^d$. In this controlled setup, the generalization score can be computed analytically from \mathbb{P}_0 and reads

$$s_{\text{gen}}(\mathbf{x}_t, t) = \mu e^{-t} \tanh(\mathbf{x}_t \cdot \mu e^{-t}) - \mathbf{x}_t. \quad (24)$$

Score model. The denoise $\xi_\theta(\mathbf{x}_t, t)$ is implemented as a lightweight residual multi-layer neural network (see Fig. 8): an input layer projecting $\mathbb{R}^d \rightarrow \mathbb{R}^W$, followed by three identical residual blocks and an output layer projecting back to \mathbb{R}^d . Each block consists of two fully connected layers of width W , a skip connection, and a layer normalization. We encode the diffusion time t via sinusoidal position embedding and add it to the first feature of each block. The total number of parameter in the network is $p(d, W) = W(2d + 13) + d + 6W^2$. For $d = 8$, and $W = 128$, the reference setting of this section, this yields $p = 102,024$ trainable parameters.

Training and computing resources. Unless otherwise specified, we train every model of this section with SGD at fixed learning rate $\eta = 6 \times 10^{-3}$ and momentum $\beta = 0.95$ using full-batch updates $B = n$ for $n \in \{128, 256, 512, 1024, 2048, 4096\}$, running for up to 4M updates. All experiments are executed on an Nvidia RTX 2080 Ti, with the largest $n = 4096$ requiring around 10 hours to complete.

Generalization and memorization metrics. In addition to the memorization fraction $f_{\text{mem}}(\tau)$, we exploit this controlled setting where we know the true data distribution \mathbb{P}_0 to directly measure how closely it matches the generated distribution \mathbb{P}_θ via the Kullback-Leibler (KL) divergence

$$D_{\text{KL}}(\mathbb{P}_\theta \| \mathbb{P}_0) = \int d\mathbf{x} \mathbb{P}_\theta(\mathbf{x}) \log \mathbb{P}_\theta(\mathbf{x}) - \int d\mathbf{x} \mathbb{P}_\theta(\mathbf{x}) \log \mathbb{P}_0(\mathbf{x}). \quad (25)$$

The cross-entropy term $\mathbb{E}_{\mathbb{P}_\theta} [\log \mathbb{P}_0]$ is easy to estimate using Monte Carlo,

$$\int d\mathbf{x} \mathbb{P}_\theta(\mathbf{x}) \log \mathbb{P}_0(\mathbf{x}) \approx \frac{1}{N} \sum_{\mu=1}^N \log \mathbb{P}_0(\tilde{\mathbf{x}}_\mu), \quad (26)$$

where $\{\tilde{\mathbf{x}}_\mu\}_{\mu=1}^N$ are $N = 10,000$ samples drawn from the model with parameters $\theta(\tau)$ at training time τ . Estimating the negative entropy term $\mathbb{E}_{\mathbb{P}_\theta} [\log \mathbb{P}_\theta]$ is more challenging, since DMs only give access to the score function $s_\theta(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log \mathbb{P}_\theta(\mathbf{x})$ and not the underlying probability distribution \mathbb{P}_θ . We can however employ time integration to express it as a function of the score only,

$$\mathbb{E}_{\mathbb{P}_\theta} [\log \mathbb{P}_\theta] \approx \int_0^T dt I(t) - \frac{d}{2} \log(2\pi e), \quad (27)$$

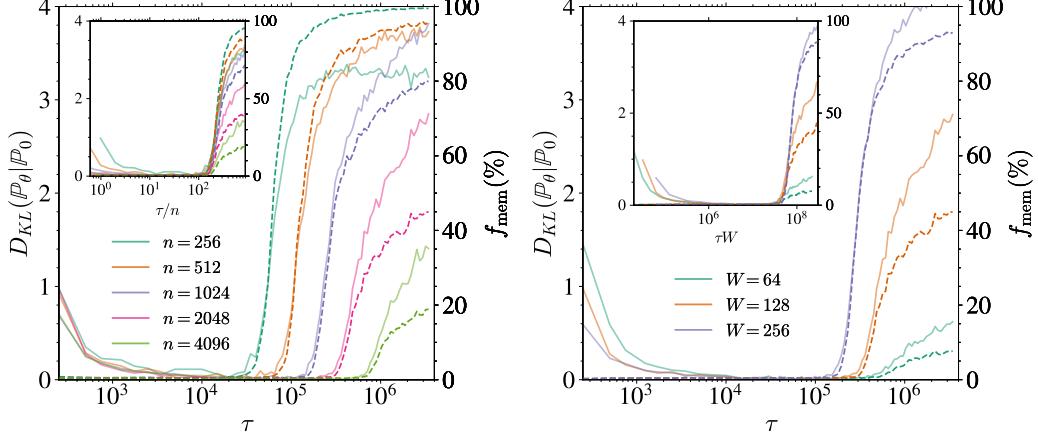


Figure 9: Generalization–Memorization transition as a function of the training set size n and width W for ResNet score models on GMM ($d = 8$). KL divergences (solid lines, left axis) and memorization fraction f_{mem} (in %, dashed lines, right axis) against training time τ for various (Left) $n \in \{256, 512, 1024, 2048, 4096\}$ at fixed $W = 128$. (Right) $W \in \{64, 128, 256\}$ at fixed $n = 2048$. Insets: $D_{\text{KL}}(\mathbb{P}_{\theta} \|\mathbb{P}_0)$ and f_{mem} against the rescaled time τ/n (left) and τW (right).

with

$$I(t) = \frac{\beta_t}{2N} \sum_{\mu=1}^N [\tilde{\mathbf{x}}_\mu \mathbf{s}_\theta(\tilde{\mathbf{x}}_\mu, t) + \mathbf{s}_\theta(\tilde{\mathbf{x}}_\mu, t)^2]. \quad (28)$$

This expression assumes that the model learns an accurate representation of the score function. It is noteworthy to mention that samples are generated using standard Euler-Maruyama discretization of the backward process 2 of the MT over $T = 1000$ timesteps.

B.2 Scaling of τ_{mem} and τ_{gen} with n and W

In Fig. 9, the left panel shows how the KL divergence and memorization fraction evolve with training time τ for different training set sizes n at fixed width $W = 128$, while the right panel fixes $n = 2048$ and varies W . In both cases, we observe two distinct phases. First, the KL divergence decreases to near zero on a timescale τ_{gen} independent of n during which the model fully generalizes ($f_{\text{mem}} = 0$). Beyond τ_{gen} , both $D_{\text{KL}}(\mathbb{P}_{\theta} \|\mathbb{P}_0)$ and f_{mem} begin to rise at a time τ_{mem} that scales linearly with n , as highlighted by the inset of the left panel. In contrast, τ_{mem} scales with W^{-1} , as shown in the inset of the right panel. While the precise dependence of τ_{gen} with W remains inconclusive in this setting and require a more careful analysis, these results on the GMM mirror the main findings of the MT: the training dynamics of DMs unfolds first in a generalization phase and only later – at $\tau_{\text{mem}} \propto n/W$ – memorization begins.

B.3 Discussion on conditional diffusion models

Conditional generation aims to sample from distributions of the form $p(\mathbf{x}|\mathbf{y})$, where \mathbf{y} denotes a conditioning variable such as a class label, a text embedding, or any other contextual information. DMs can naturally be extended to this setting using for instance classifier-free guidance [21]. Although conditioning often improves sample quality in practice, memorization effects have also been observed in models trained conditionally [51, 61, 10]. Our analysis does not rely on the model being unconditional since these variables typically enter the model as additional inputs and we expect our result to hold in this setting as well. To investigate it, we train a classifier-free guidance model to generate sample from our Gaussian mixture conditionally on the class label, and compute the memorized fraction as a function of τ that we report in Fig. 10. In the inset, when rescaling the training time by n , the curves for $n \in \{256, 512, 1024\}$ all collapse perfectly, confirming that the phenomenon persists in the conditional setting. For more complex datasets, τ_{mem} and τ_{gen} may in fact depend on the conditioning variable and intermediate regimes could exist where certain classes have already entered the generalization (or memorization) phase while others have not yet.

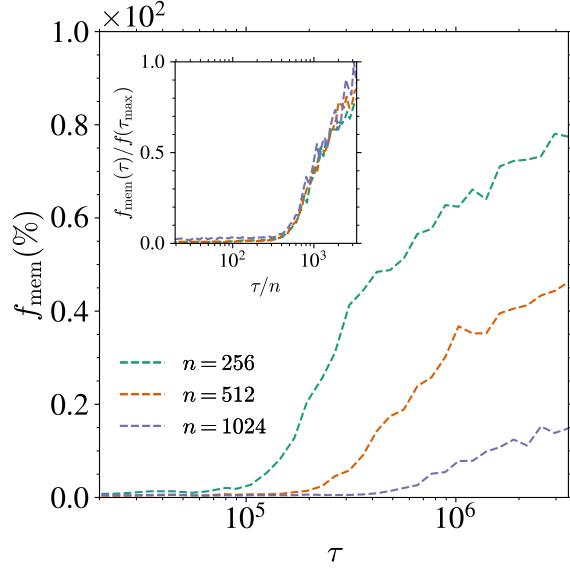


Figure 10: **Effect of guidance on τ_{mem} .** Evolution of f_{mem} as a function of τ for $n \in \{256, 512, 1024\}$ at fixed $W = 64$.

C Proofs of the analytical results

In the following we provide the mathematical arguments and the proofs for the statement in the MT. The section using the replica method is not mathematically rigorous but uses a well established method of theoretical physics, which has been already shown to provide correct results in several cases. The final result is rigorous, since it can alternatively be obtained from the rigorous free random matrix approach of [15], as shown in Sect. C.4.

C.1 Notations

We recall here the notations used throughout Sect. 3 of the MT and Sect. C of the SM.

$$d : \text{Data dimension} \quad (29)$$

$$n : \text{Numbers of data points} \quad (30)$$

$$p : \text{Dimension of the hidden layers of the RFNN} \quad (31)$$

$$\mathbf{I}_d : \text{Identity matrix in dimension } d \quad (32)$$

$$\sigma(x) : \text{Activation function of the model} \quad (33)$$

$$P_{\mathbf{x}} : \text{Distribution of the data points} \quad (34)$$

$$P_t : \text{Distribution of the noisy data points at diffusion time } t. \quad (35)$$

$$\psi_n = \frac{n}{d} \quad (36)$$

$$\psi_p = \frac{p}{d} \quad (37)$$

$$\Delta_t = 1 - e^{-2t} \quad (38)$$

$$\Sigma = \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{x}}} [\mathbf{x}\mathbf{x}^T] \quad (39)$$

$$\Sigma_t = e^{-2t} \Sigma + \Delta_t \mathbf{I}_d \quad (40)$$

$$\Gamma_t^2 = \frac{\text{Tr}(\Sigma_t)}{d} \quad (41)$$

$$\sigma_{\mathbf{x}}^2 = \frac{\text{Tr}(\Sigma)}{d} \quad (42)$$

$$\|\sigma\|^2 = \mathbb{E}_z[\sigma(\Gamma_t z)^2] \quad (43)$$

$$b_t^2 = \left(\mathbb{E}_{u,v}[v\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)] \right)^2 \quad (44)$$

$$a_t = \mathbb{E}_{u,v}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v) \frac{u}{e^{-t}\sigma_{\mathbf{x}}}] \quad (45)$$

$$v_t^2 = \mathbb{E}_{u,v,w}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}w)] - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 \quad (46)$$

$$s_t^2 = \|\sigma\|^2 - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 - v_t^2 - b_t^2 \quad (47)$$

$$\mu_1(t) = \mathbb{E}_u[\sigma(\Gamma_t u)u] = \sqrt{e^{-2t}\sigma_{\mathbf{x}}^2 a_t^2 + b_t^2}. \quad (48)$$

Unless specified, all the expectation values are taken for standard Gaussian variables. We will denote

$$\mathbf{X} = [\mathbf{x}^1 | \dots | \mathbf{x}^n] \in \mathbb{R}^{d \times n} \quad (49)$$

the matrix whose columns are the data point vectors and likewise we decompose \mathbf{W} as

$$\mathbf{W} = \begin{bmatrix} (\mathbf{W}_1)^\top \\ \vdots \\ (\mathbf{W}_p)^\top \end{bmatrix} \in \mathbb{R}^{p \times d}, \quad (50)$$

where $\mathbf{W}_i \in \mathbb{R}^d$ denotes the i th row of \mathbf{W} . We recall the definitions of the matrices \mathbf{U} and \mathbf{V}

$$\mathbf{U} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[\sigma \left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \sigma \left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right)^T \right], \quad (51)$$

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[\sigma \left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \xi^T \right]. \quad (52)$$

C.2 Closed form of the learning dynamics

Proposition C.1. *Let $\mathbf{A}(\tau)$ be the solution of the gradient flow (10) defined in the MT with initial conditions $\mathbf{A}(\tau=0) = \mathbf{A}_0$, then*

$$\frac{\mathbf{A}(\tau)}{\sqrt{p}} = -\frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} + \left(\frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} + \frac{\mathbf{A}_0}{\sqrt{p}} \right) e^{-\frac{2\Delta_t}{\psi_p} \mathbf{U}\tau} \quad (53)$$

with

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[\sigma \left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \xi^T \right]. \quad (54)$$

Proof. We expand the square in the training loss

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr} \left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U} \right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr} \left(\frac{\mathbf{A}}{\sqrt{p}} \mathbf{V} \right) \quad (55)$$

and compute the gradient

$$\nabla_{\mathbf{A}} \mathcal{L}_{\text{train}}(\mathbf{A}(\tau)) = \frac{2\Delta_t}{d} \frac{\mathbf{A}}{p} \mathbf{U} + \frac{2\sqrt{\Delta_t}}{d\sqrt{p}} \mathbf{V}^T. \quad (56)$$

Solving this ordinary differential equation yields the desired result. Consequently, the timescales of the dynamics of $\mathbf{A}(\tau)$ is determined by the inverse of the eigenvalues of $\Delta_t \mathbf{U}/\psi_p$. \square

C.3 Gaussian Equivalence Principle

As explained in [41, 17, 22], the Gaussian Equivalence Principle which applies in the high dimensional setting considered here establishes an equivalence between the spectral properties of the original

model and those of a Gaussian covariate model in which the nonlinear activation function is replaced by a linear term and a nonlinear term that acting as noise,

$$\sigma\left(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\right) \rightarrow \kappa_1 \frac{\mathbf{W}\mathbf{x}'}{\sqrt{d}} + \kappa_* \boldsymbol{\eta}, \quad \mathbf{x}' \sim \mathcal{N}(0, \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T]), \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_p), \quad (57)$$

where κ_1, κ_* are constants that depend on the distribution of the data and on the activation function σ whose formula we recall

$$\kappa_1 = \mathbb{E}_z[\sigma(\sigma_{\mathbf{x}} z) \frac{z}{\sigma_{\mathbf{x}}}], \quad (58)$$

$$\kappa_* = \sqrt{\mathbb{E}_z[\sigma(\sigma_{\mathbf{x}} z)^2] - \kappa_1^2 \sigma_{\mathbf{x}}^2}. \quad (59)$$

The expectation function are with respect to $z \sim \mathcal{N}(0, 1)$ and $\sigma_{\mathbf{x}}^2 = \text{Tr}(\Sigma)/d$. The Gaussian Equivalence Principle (GEP) holds if the distribution $P_{\mathbf{x}}$ of the vector \mathbf{x} verifies

- (i) $P_{\mathbf{x}}(\mathbf{x})$ has sub-Gaussian tails: there exists a constant $C > 0$ such that for all $A \geq 0$ and each entry \mathbf{x}_i ,

$$\mathbb{P}(|\mathbf{x}_i| \geq A) \leq 2 e^{-A^2/C}. \quad (60)$$

- (ii) The data covariance matrix $\Sigma = \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^T]$ is bounded: there exists a constant $K > 0$ independent of d such that $\lambda_{\max}(\Sigma) < K$ and $\frac{\text{Tr} \Sigma}{d} < K$ where $\lambda_{\max}(\Sigma)$ denotes the spectral norm of Σ .

In this section, we outline the derivation of the Gaussian Equivalence Principle (GEP) for the matrices $\mathbf{U}, \tilde{\mathbf{U}}, \mathbf{V}$ and $\tilde{\mathbf{V}}$ under arbitrary input covariance. This generalizes the approach developed in [15], which considered only the case of data drawn from $\mathcal{N}(0, \mathbf{I}_d)$. A more rigorous approach, which would consist in following [36], is left for future works. We will make use of the Mehler kernel formula [27] which states that for f a test function defined on \mathbb{R}^2 ,

$$\mathbb{E}_{u,v \sim P^{\gamma}}[f(u,v)] = \sum_{s=1}^{\infty} \frac{\gamma^s}{s!} \mathbb{E}_{u,v \sim \mathcal{N}(0, \mathbf{I}_2)}[H e_s(u) H e_s(v) f(u,v)], \quad (61)$$

where the expectation on the left-hand side is taken over jointly Gaussian random variables u and v with zero mean, unit variance, and correlation γ , while on the right-hand side the expectation is taken over independent standard Gaussian variables. $H e_s$ denotes the s -th Hermite polynomial. We recall some useful properties of the Hermite polynomials [4]:

- They form an orthogonal base of $L^2(\mathbb{R}, \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx)$.
- The first Hermite polynomials are $H e_0(x) = 1, H e_1(x) = x$.

Lemma C.1 (Gaussian Equivalence Principle for \mathbf{U}). *In the limit $n, p, d \rightarrow \infty$ with $\psi_p = p/d, \psi_n = n/d$ and with a dataset $\{\mathbf{x}^{\nu}\}_{\nu=1}^n$ sampled from a distribution $P_{\mathbf{x}}$ which verifies assumptions (i) and (ii) with $\Sigma = \mathbb{E}_{P_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^T]$, the matrix*

$$\mathbf{U} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi}[\sigma(\frac{\mathbf{W}\mathbf{x}_t^{\nu}(\xi)}{\sqrt{d}}) \sigma(\frac{\mathbf{W}\mathbf{x}_t^{\nu}(\xi)}{\sqrt{d}})^T] \quad (62)$$

has the same spectrum as its Gaussian equivalent

$$\mathbf{U} = \frac{\mathbf{G}}{\sqrt{n}} \frac{\mathbf{G}^T}{\sqrt{n}} + b_t^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + s_t^2 \mathbf{I}_p \quad (63)$$

where

$$\mathbf{G} = e^{-t} a_t \frac{\mathbf{W}}{\sqrt{d}} \mathbf{X}' + v_t \boldsymbol{\Omega}, \quad (64)$$

$\mathbf{X}' \in \mathbb{R}^{d \times n}$ is a matrix whose columns \mathbf{x}'^{ν} are sampled according to $\mathcal{N}(0, \Sigma)$ and $\boldsymbol{\Omega} \in \mathbb{R}^{p \times d}$ has gaussian entries independent of \mathbf{X} and \mathbf{W} .

Proof. For the sake of clarity, in this proof we explicitly make the covariance of the data Σ appear by writing the data points are written as $\mathbf{x}^\nu = \Sigma^{1/2} \mathbf{z}^\nu$ where the vectors \mathbf{z}^ν have variance 1. Let us focus on the element of \mathbf{U} in position (i, j)

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_\xi [\sigma \left(\frac{\mathbf{W}_{ik}(e^{-t}(\Sigma^{1/2})_{kl}\mathbf{z}_l^\nu + \sqrt{\Delta_t}\xi_k)}{\sqrt{d}} \right) \sigma \left(\frac{\mathbf{W}_{jk'}(e^{-t}(\Sigma^{1/2})_{k'l'}\mathbf{z}_{l'}^\nu + \sqrt{\Delta_t}\xi_{k'}}{\sqrt{d}} \right)], \quad (65)$$

where repeated indices mean that there is a hidden sum. We introduce the random variable $\chi_i^\nu = \frac{\mathbf{W}_{ik}(e^{-t}(\Sigma^{1/2})_{kl}\mathbf{z}_l^\nu + \sqrt{\Delta_t}\xi_k)}{\sqrt{d}}$. In the high dimensional limit it converges to a Gaussian random variable by the Central Limit Theorem (since the tails of the data distribution are sub-Gaussian). If $i = j$, the diagonal terms concentrate with respect to the data points and we can thus replace the sum by an average

$$\mathbf{U}_{ii} = \mathbb{E}_\chi [\sigma(\chi)^2] + \mathcal{O}(1/n). \quad (66)$$

The finite n corrections can be discarded because they cannot change the spectrum of \mathbf{U} . χ can be taken Gaussian with mean 0 and covariance $\mathbb{E}_{\mathbf{W}_i, \mathbf{z}, \xi} [\chi^2] = \mathbb{E}_{\mathbf{W}_i, \mathbf{z}, \xi} [\frac{\mathbf{W}_i^T \Sigma_t \mathbf{W}_i}{d}] = \frac{\text{Tr}(\Sigma_t)}{d} = \Gamma_t^2$ hence

$$\mathbf{U}_{ii} = \mathbb{E}_\chi [\sigma(\chi)^2] + \mathcal{O}(1/n) = \mathbb{E}_{u \sim \mathcal{N}(0, 1)} [\sigma(\Gamma_t u)^2] = \|\sigma\|^2. \quad (67)$$

If $i \neq j$, we denote $\eta_i^\nu = e^{-t} \frac{\mathbf{W}_i^T \Sigma^{1/2} \mathbf{z}^\nu}{\sqrt{d}}$. For now we consider \mathbf{W} and the \mathbf{z}^ν fixed and look at ξ . We use the Mehler Kernel formula for the random variables $u = \mathbf{W}_i^T \xi / \sqrt{d}$ and $v = \mathbf{W}_j^T \xi / \sqrt{d}$ that have correlation $\mathbb{E}_\xi [uv] = \frac{\mathbf{W}_i^T \mathbf{W}_j}{d}$

$$\mathbb{E}_{u,v} [\sigma (\eta_i^\nu + \sqrt{\Delta_t} u) \sigma (\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (68)$$

$$= \sum_{s=1}^{\infty} \frac{(\mathbf{W}_i^T \mathbf{W}_j / d)^s}{s!} \mathbb{E}_u [H e_s(u) \sigma (\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [H e_s(v) \sigma (\eta_j^\nu + \sqrt{\Delta_t} v)]. \quad (69)$$

We truncate at order $s = 1$ since the corrections are order $\mathcal{O}(1/d)$.

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_u [\sigma (\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [\sigma (\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (70)$$

$$+ \frac{1}{n} \sum_{\nu=1}^n \frac{\mathbf{W}_i^T \mathbf{W}_j}{d} \mathbb{E}_u [u \sigma (\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [v \sigma (\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (71)$$

$$= \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_u [\sigma (\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [\sigma (\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (72)$$

$$+ \frac{\mathbf{W}_i^T \mathbf{W}_j}{d} \mathbb{E}_\eta [\mathbb{E}_u [u \sigma (\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [v \sigma (\eta_j^\nu + \sqrt{\Delta_t} v)]]. \quad (73)$$

by neglecting $\mathcal{O}(1/d)$ corrections and where the law of η can be considered Gaussian with zero mean correlation $\mathbb{E}[\eta_i^\nu \eta_j^\nu] = \frac{e^{-2t} \text{Tr}(\Sigma)}{d} \delta_{ij} = e^{-2t} \sigma_x^2 \delta_{ij}$. The coefficient in front of $\frac{\mathbf{W}_i^T \mathbf{W}_j}{d}$ is therefore

$$b_t^2 = (\mathbb{E}_{u,v} [v \sigma (e^{-t} \sigma_x u + \sqrt{\Delta_t} v)])^2. \quad (74)$$

Denote $\sigma_0(\eta) = \mathbb{E}_u [\sigma (\eta + \sqrt{\Delta_t} u)]$. We now focus on

$$\frac{1}{n} \sum_\nu \sigma_0(\eta_i^\nu) \sigma_0(\eta_j^\nu). \quad (75)$$

We use the GEP on σ_0

$$\sigma_0 \left(\frac{e^{-t} \mathbf{W}_i^T \mathbf{x}^\nu}{\sqrt{d}} \right) \rightarrow a_t e^{-t} \frac{\mathbf{W}_i^T \mathbf{x}'^\nu}{\sqrt{d}} + v_t \boldsymbol{\Omega}_i^\nu, \quad \mathbf{x}'^\nu \sim \mathcal{N}(0, \Sigma), \quad \boldsymbol{\Omega}_i^\nu \sim \mathcal{N}(0, \mathbf{I}_p), \quad (76)$$

with $a_t = \mathbb{E}_u[\sigma_0(e^{-t}\sigma_{\mathbf{x}}u)\frac{u}{e^{-t}\sigma_{\mathbf{x}}}] = \mathbb{E}_{u,v}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\frac{u}{e^{-t}\sigma_{\mathbf{x}}}]$ and $v_t^2 = \mathbb{E}_u[\sigma_0(e^{-t}\sigma_{\mathbf{x}}u)^2] - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 = \mathbb{E}_{u,v,w}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}w)] - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2$. Hence the truncated expansion yields for $i \neq j$

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \left(a_t e^{-t} \frac{\mathbf{W}_i^T \mathbf{x}'^\nu}{\sqrt{d}} + v_t \boldsymbol{\Omega}_i^\nu \right) \left(a_t e^{-t} \frac{\mathbf{W}_j^T \mathbf{x}'^\nu}{\sqrt{d}} + v_t \boldsymbol{\Omega}_j^\nu \right)^T + b_t^2 \frac{\mathbf{W}_i^T \mathbf{W}_j}{d}. \quad (77)$$

Now we need to deal with the diagonal term. We need to subtract

$$(a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 + v_t^2 + b_t^2) \mathbf{I}_p. \quad (78)$$

The Gaussian equivalent of \mathbf{U} reads

$$\mathbf{U} = \frac{\mathbf{G}}{\sqrt{n}} \frac{\mathbf{G}^T}{\sqrt{n}} + b_t^2 \frac{\mathbf{W} \mathbf{W}^T}{d} + s_t^2 \mathbf{I}_p, \quad (79)$$

with $s_t^2 = \|\sigma\|^2 - a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 - v_t^2 - b_t^2$.

□

Lemma C.2 (GEP for \tilde{U}). *Let*

$$\tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{y}}[\sigma(\frac{\mathbf{W}\mathbf{y}}{\sqrt{d}})\sigma(\frac{\mathbf{W}\mathbf{y}}{\sqrt{d}})^T], \quad (80)$$

where the expectation value is taken $\mathbf{y} \sim P_t$. Then the GEP of \tilde{U} reads

$$\mu_1^2(t) \frac{\mathbf{W} \boldsymbol{\Sigma}_t \mathbf{W}^T}{d} + (\|\sigma\|^2 - \mu_1^2(t)) \mathbf{I}_p, \quad (81)$$

where $\mu_1^2(t)$ and $\|\sigma\|^2$ are defined in Sect. C.1.

Proof. For a vector \mathbf{y} sampled from P_t , the $\frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}}$ are asymptotically Gaussian with 0 mean, variance $\mathbb{E}_{\mathbf{y}}[\frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}} \frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}}] = \frac{\mathbf{W}_i^T \boldsymbol{\Sigma}_t \mathbf{W}_i}{d} \sim \Gamma_t^2$ and correlation $\mathbb{E}_{\mathbf{y}}[\frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}} \frac{\mathbf{W}_j^T \mathbf{y}}{\sqrt{d}}] = \frac{\mathbf{W}_i^T \boldsymbol{\Sigma}_t \mathbf{W}_j}{d}$. We apply Mehler Kernel formula to $\tilde{\mathbf{U}}$

$$\tilde{\mathbf{U}}_{ij} = \sum_s \frac{1}{s!} \left(\frac{\mathbf{W}_{ik}(\boldsymbol{\Sigma}_t)_{kl} \mathbf{W}_{jl}}{\Gamma_t^2 d} \right)^s \mathbb{E}_u[\sigma(\Gamma_t u) H e_s(u)] \mathbb{E}_v[\sigma(\Gamma_t v) H e_s(v)], \quad (82)$$

where the expectation on u and v is standard Gaussian. We keep only terms at order $\mathcal{O}(1/\sqrt{d})$. If $i \neq j$ we keep the terms up to order $s = 1$.

$$\tilde{\mathbf{U}}_{ij} = \left(\frac{\mathbf{W}_{ik}(\boldsymbol{\Sigma}_t)_{kl} \mathbf{W}_{jl}}{\Gamma_t^2 d} \right) \mathbb{E}_u[\sigma(\Gamma_t u) u]^2. \quad (83)$$

For $i = j$ we cannot truncate because all terms are $\mathcal{O}_d(1)$. Hence the diagonal terms are asymptotically

$$\tilde{\mathbf{U}}_{ii} = \mathbb{E}_{u \sim \mathcal{N}(0,1)}[\sigma^2(\Gamma_t z)] = \|\sigma\|^2. \quad (84)$$

Taking care of the diagonal terms, the Gaussian Equivalent matrix reads

$$\tilde{\mathbf{U}} = \frac{\mu_1^2(t) \mathbf{W} \boldsymbol{\Sigma}_t \mathbf{W}^T}{\Gamma_t^2 d} + (\|\sigma\|^2 - \mu_1^2(t)) \mathbf{I}_p \quad (85)$$

where $\mu_1(t) = \mathbb{E}_u[\sigma(\Gamma_t u) u]$.

□

Building on the GEP of $\tilde{\mathbf{U}}$, we prove the following lemma on the scaling of the eigenvalues in the bulk.

Lemma C.3 (Scaling of the bulk of $\tilde{\mathbf{U}}$). *We assume that Σ is positive definite and that the spectral norm $\lambda_{\max}(\Sigma)$ stays $\mathcal{O}_d(1)$. In the high dimensional limit $p > d \gg 1$, the spectrum of $\tilde{\mathbf{U}}$ is asymptotically equal to*

$$\left(1 - \frac{1}{\psi_p}\right) \delta(\lambda - (\|\sigma\|^2 - \mu_1^2(t))) + \frac{1}{\psi_p} \rho_{\text{bulk}}(\lambda), \quad (86)$$

where $\rho_{\text{bulk}}(\lambda)$ is an atomless measure whose support is of order $\mathcal{O}(\psi_p)$.

Proof. Since $p > d$ and $\mathbf{W} \in \mathbb{R}^{p \times d}$ and $\Sigma \in \mathbb{R}^{d \times d}$, the spectrum admits a Dirac mass at $\lambda = \|\sigma\|^2 - \mu_1^2(t)$ with weight $(p-d)/p$. For the order of magnitude of the eigenvalues in the bulk, let us first observe that the bulk of $\frac{\mathbf{W}^T \Sigma_t \mathbf{W}}{d}$ is the same as the one of $\frac{\Sigma_t^{1/2} \mathbf{W} \mathbf{W}^T \Sigma_t^{1/2}}{d}$. We can bound the spectral norm of the product by the product of the spectral norms

$$\lambda_{\max}\left(\frac{\Sigma_t^{1/2} \mathbf{W} \mathbf{W}^T \Sigma_t^{1/2}}{d}\right) \leq \lambda_{\max}\left(\frac{\mathbf{W} \mathbf{W}^T}{d}\right) \lambda_{\max}(\Sigma_t) \lesssim \mathcal{O}(\psi_p), \quad (87)$$

since we assumed that $\lambda_{\max}(\Sigma_t) = e^{-2t} \lambda_{\max}(\Sigma_t) + \Delta_t = \mathcal{O}(1)$ and since $\lambda_{\max}\left(\frac{\mathbf{W} \mathbf{W}^T}{d}\right) = \mathcal{O}(\psi_p)$ is given by the Marchenko-Pastur law [39]. To bound the norm from below we use the following inequality

$$\lambda_{\min}(\Sigma_t) \lambda_{\min}\left(\frac{\mathbf{W} \mathbf{W}^T}{d}\right) \leq \lambda_{\min}\left(\frac{\Sigma_t^{1/2} \mathbf{W} \mathbf{W}^T \Sigma_t^{1/2}}{d}\right). \quad (88)$$

Since Σ_t is positive definite, the bound is also of order ψ_p . This concludes that the support of the bulk is of order ψ_p . \square

Lemma C.4 (GEP for \mathbf{V} and $\tilde{\mathbf{V}}$). *Let*

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi}[\sigma\left(\frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}}\right) \xi^T], \quad (89)$$

$$\tilde{\mathbf{V}} = \mathbb{E}_{\mathbf{x}, \xi}[\sigma\left(\frac{\mathbf{W} \mathbf{x}_t(\xi)}{\sqrt{d}}\right) \xi^T]. \quad (90)$$

They can be replaced by their Gaussian Equivalence Principle in the train and test losses.

$$\tilde{\mathbf{V}} = \mathbf{V} = \frac{\mu_1(t)\sqrt{\Delta_t}}{\Gamma_t} \frac{\mathbf{W}}{\sqrt{d}}. \quad (91)$$

Proof. The two matrices only differ element-wise by quantity of order $\mathcal{O}(1/n)$ and therefore have the same Gaussian Equivalent matrix. We focus on $\tilde{\mathbf{V}}$. Introduce the random variable $\boldsymbol{\eta}_i = \frac{\mathbf{W}_{ik}(e^{-t}\mathbf{x}_k + \sqrt{\Delta_t}\xi_k)}{\sqrt{d}}$. Its has 0 mean, covariance $\mathbb{E}_{\mathbf{x}, \xi}[\boldsymbol{\eta}_i^2] = \frac{\mathbf{W}_i^T \Sigma_t \mathbf{W}_i}{d} \sim \Gamma_t^2$ and correlation with ξ $\gamma_{ij} = \mathbb{E}_{\mathbf{x}, \xi}[\boldsymbol{\eta}_i \boldsymbol{\xi}_j] = \frac{\sqrt{\Delta_t} \mathbf{W}_{ij}}{\sqrt{d}}$. We apply the Mehler Kernel formula

$$\tilde{\mathbf{V}}_{ij} = \mathbb{E}_{\mathbf{x}, \xi}[\sigma\left(\Gamma_t\left(\frac{\mathbf{W}_{ik}(e^{-t}(\Sigma_t)_{kl}\mathbf{z}_l + \sqrt{\Delta_t}\xi_l)}{\Gamma_t \sqrt{d}}\right)\right) \xi_j] \quad (92)$$

$$= \sum_s \frac{1}{s!} \left(\frac{\mathbf{W}_{ij}\sqrt{\Delta_t}}{\Gamma_t \sqrt{d}}\right)^s \mathbb{E}_u[\sigma(\Gamma_t u) H e_s(u)] \mathbb{E}_v[v H e_s(v)] \quad (93)$$

$$= 0 + \frac{\sqrt{\Delta_t}}{\Gamma_t} \frac{\mathbf{W}_{ij}}{\sqrt{d}} \mathbb{E}_u[\sigma(\Gamma_t u) u] \mathbb{E}_v[v^2] + \mathcal{O}\left(\frac{1}{d}\right) \quad (94)$$

$$= \frac{\sqrt{\Delta_t} \mu_1(t)}{\Gamma_t} \frac{\mathbf{W}_{ij}}{\sqrt{d}}. \quad (95)$$

\square

C.4 Proof of Theorem 3.1

We recall the Theorem 3.1 of the MT.

Theorem C.1. Let $q(z) = \frac{1}{p} \text{Tr}(\mathbf{U} - z\mathbf{I}_p)^{-1}$, $r(z) = \frac{1}{p} \text{Tr}(\boldsymbol{\Sigma}^{1/2} \mathbf{W}^T (\mathbf{U} - z\mathbf{I}_p)^{-1} \mathbf{W} \boldsymbol{\Sigma}^{1/2})$ and $s(z) = \frac{1}{p} \text{Tr}(\mathbf{W}^T (\mathbf{U} - z\mathbf{I}_p)^{-1} \mathbf{W})$, with $z \in \mathbb{C}$. Let

$$\hat{s}(q) = b_t^2 \psi_p + \frac{1}{q}, \quad (96)$$

$$\hat{r}(r, q) = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q}. \quad (97)$$

Then $q(z)$, $r(z)$ and $s(z)$ satisfy the following set of three equations:

$$s = \int d\rho_{\boldsymbol{\Sigma}}(\lambda) \frac{1}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (98)$$

$$r = \int d\rho_{\boldsymbol{\Sigma}}(\lambda) \frac{\lambda}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (99)$$

$$\psi_p(s_t^2 - z) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q} + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0, \quad (100)$$

The eigenvalue distribution of \mathbf{U} , $\rho(\lambda)$, can then be obtained using the Sokhotski–Plemelj inversion formula $\rho(\lambda) = \lim_{\varepsilon \rightarrow 0^+} \frac{1}{\pi} \text{Im } q(\lambda + i\varepsilon)$.

We first show that the equations of the Stieltjes transform of ρ found in Ref. [15] with linear pencils [7] in the case $P_{\mathbf{x}} = \mathcal{N}(0, \mathbf{I}_d)$ i.e. $\rho_{\boldsymbol{\Sigma}}(\lambda) = \delta(\lambda - 1)$ can be reduced to the equations of Theorem 3.1 with our definitions of $\mu_1(t)$, s_t and v_t . The equations of [15] read

$$\zeta_1(s_t^2 - z + e^{-2t} \mu_1^2 \zeta_2 \zeta_3 + v_t^2 \zeta_2 + \Delta_t \mu_1^2 \zeta_4) - 1 = 0 \quad (101)$$

$$\zeta_2(\psi_n + v_t^2 \psi_p \zeta_1 - e^{-t} \mu_1 \zeta_3) - \psi_n = 0 \quad (102)$$

$$e^{-t} \mu_1 \psi_p \zeta_1 (1 + e^{-t} \mu_1 \zeta_2 \zeta_3) + (1 + (\Delta_t \mu_1^2 \psi_p \zeta_1) \zeta_3) = 0 \quad (103)$$

$$e^{-2t} \mu_1^2 \psi_p \zeta_1 \zeta_2 \zeta_4 + (1 + \Delta_t \mu_1^2 \psi_p \zeta_1) \zeta_4 - 1 = 0, \quad (104)$$

with $\zeta_1 = q$ and $\zeta_{2,3,4}$ auxiliary variables. We make the following change of variables $r = -\frac{\zeta_3}{e^{-t} \mu_1 \psi_p}$.

The second equations relates ζ_2 to q and r

$$\zeta_2 = \frac{1}{1 + \frac{e^{-2t} \mu_1^2 \psi_p}{\psi_n} r + \frac{v_t^2 \psi_p}{\psi_n} q}. \quad (105)$$

Injecting this into the second equations gives the second equation of Theorem 3.1. The fourth equation gives

$$\zeta_4 = \frac{1}{1 + \mu_1^2 \psi_p q (\Delta_t + e^{-2t} \zeta_2)}. \quad (106)$$

Injecting this into the first equation gives

$$q(s_t^2 - z + e^{-2t} \mu_1^2 \zeta_2 r (-e^{-t} \mu_1 \psi_p) + v_t^2 \zeta_2 + \Delta_t \mu_1^2 \frac{1}{1 + \mu_1^2 \psi_p q (\Delta_t + e^{-2t} \zeta_2)}) - 1 = 0. \quad (107)$$

After some massaging we find back the first equation of Theorem 3.1.

We now prove Theorem 3.1 using a replica computation, inspired by the calculation done in Ref. [13].

Proof. Our goal is to compute the Stieltjes transform of the matrix \mathbf{U} .

$$q = \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \boldsymbol{\Omega}} [\text{Tr}(\mathbf{U} - z\mathbf{I}_p)^{-1}] \quad (108)$$

$$= -\partial_z \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \boldsymbol{\Omega}} [\log \det(\mathbf{U} - z\mathbf{I}_p)] \quad (109)$$

$$= 2\partial_z \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \boldsymbol{\Omega}} [\log \det(\mathbf{U} - z\mathbf{I}_p)^{-1/2}]. \quad (110)$$

The so-called *replica trick* consists of replacing the $\log x$ by $\lim_{s \rightarrow \infty} \frac{x^s - 1}{s}$. Applying this identity, we obtain

$$q = 2\partial_z \lim_{s \rightarrow 0} \lim_{p \rightarrow \infty} \frac{1}{ps} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\det(\mathbf{U} - z\mathbf{I}_p)^{-s/2} - 1], \quad (111)$$

where as usual with replica computations we have inverted the order of the limits $p \rightarrow \infty$ and $s \rightarrow 0$. We define the partition function \mathcal{Z} as

$$\mathcal{Z} = \det(\mathbf{U} - z\mathbf{I}_p)^{-1/2} = \int \frac{d\phi}{(2\pi)^{p/2}} e^{-\frac{1}{2}\phi^T(\mathbf{U}-z\mathbf{I}_p)\phi}. \quad (112)$$

We replace \mathbf{U} by its Gaussian equivalent proved in Lemma C.1 and write the partition function for an arbitrary integer s

$$\mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\mathcal{Z}^s] = \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [e^{-\frac{1}{2}\phi^{aT}(\mathbf{U}-z\mathbf{I}_p)\phi^a}] \quad (113)$$

$$\begin{aligned} &= \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} e^{\frac{1}{2}\phi^{aT}(z-s_t^2)\phi^a} \\ &\quad \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [e^{-\frac{1}{2n}\phi^{aT}(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)^T \phi^a} e^{-\frac{b_t^2}{2d}\phi^{aT}\mathbf{W}\mathbf{W}^T\phi^a}]. \end{aligned} \quad (114)$$

We first perform the computation for integer values of s , and then analytically continue the result to the limit $s \rightarrow 0$. To compute the expectation over \mathbf{X} , \mathbf{W} , and Ω , we need the following standard result from Gaussian integration

$$\int d\mathbf{x} e^{-\frac{1}{2}\mathbf{x}\mathbf{G}\mathbf{x}^T + \mathbf{J}\mathbf{x}^T} = e^{-\frac{1}{2}\log \det \mathbf{G} + \frac{1}{2}\mathbf{J}\mathbf{G}^{-1}\mathbf{J}^T}, \quad (115)$$

where \mathbf{G} is a square matrix and \mathbf{J} a vector.

Averaging over the data set. The dataset dependence enters through

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT}(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)^T \phi^a}] \\ &= \mathbb{E}_{\mathbf{X}} [e^{-\frac{a_t^2 e^{-2t}}{2nd}\phi^{aT}\mathbf{W}\mathbf{X}^\nu\mathbf{X}^{\nu T}\mathbf{W}^T\phi^a} e^{-\frac{a_t e^{-t}v_t}{2\sqrt{dn}}\phi^{aT}(\mathbf{W}\mathbf{X}^\nu\Omega^T + \Omega\mathbf{X}^{\nu T}\mathbf{W}^T)\phi^a}] e^{-\frac{v_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}. \end{aligned} \quad (116)$$

We introduce for each replica ϕ^a a Fourier transform of the delta function by using the auxiliary variables $\omega^a, \hat{\omega}^a \in \mathbb{R}^d$ as[†]

$$\int d\hat{\omega}^a e^{i\hat{\omega}^a(\sqrt{p}\omega^a - \phi^{aT}\mathbf{W}\Sigma^{1/2})} = 1. \quad (117)$$

In the following, we do the change of variable $\mathbf{X}^\nu = \Sigma^{1/2}\mathbf{Z}^\nu$ with \mathbf{Z}^ν a d dimensional Gaussian random variable with unit variance.

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT}(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)^T \phi^a}] \\ &= \mathbb{E}_{\mathbf{Z}} [e^{-\frac{a_t^2 e^{-2t}}{2nd}\omega^{aT}\mathbf{Z}^\nu\mathbf{Z}^{\nu T}\omega^a} e^{-\frac{a_t e^{-t}v_t\sqrt{p}}{\sqrt{dn}}\sum_{a,\nu} \Omega^\nu \phi^a \omega^a \cdot \mathbf{Z}^\nu}] e^{-\frac{v_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}. \end{aligned} \quad (118)$$

Denote $\mathbf{G}_Z = \frac{a_t^2 p}{dn} \sum_a \omega^a \omega^{aT}$ and $(\mathbf{J}_Z)^\nu = \frac{a_t e^{-t} v_t \sqrt{p}}{\sqrt{dn}} \sum_a (\Omega^\nu \cdot \phi^a) \omega^a$, then

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT}(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)(a_t e^{-t}\frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \Omega^\nu)^T \phi^a}] = \\ &e^{-\frac{n}{2}\log \det(1+\mathbf{G}_Z)} e^{\frac{a_t e^{-t} v_t p}{2dn^2} \sum_\nu (\Omega^\nu \cdot \phi^a) (\Omega^\nu \cdot \phi^b) \omega^a_h (1+\mathbf{G}_Z)^{-1} \omega^b_l} e^{-\frac{v_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}, \end{aligned} \quad (119)$$

where repeated indices mean that there is an implicit summation.

[†]Throughout the computation, we discard non-exponential prefactors, as they give subleading contributions.

Averaging over Ω . The terms that depend on Ω are

$$\begin{aligned} & \mathbb{E}_{\Omega} [e^{\frac{a_t e^{-t} v_t}{2dn^2} \sum_{\nu} (\Omega^{\nu} \cdot \phi^a)(\Omega^{\nu} \cdot \phi^b) \omega^a_k (1 + \mathbf{G}_X)^{-1}_{k,l} \omega^b_l} e^{-\frac{v_t^2}{2n} \phi^{aT} \Omega \Omega^T \phi^a}] \\ &= (\mathbb{E}_{\Omega^{\nu}} [e^{\frac{a_t e^{-t} v_t p}{2dn^2} (\Omega^{\nu} \cdot \phi^a)(\Omega^{\nu} \cdot \phi^b) \omega^a_k (1 + \mathbf{G}_X)^{-1}_{k,l} \omega^b_l} e^{-\frac{v_t^2}{2n} \phi^{aT} \Omega^{\nu} \Omega^{\nu T} \phi^a}])^n \end{aligned} \quad (120)$$

$$= e^{-\frac{n}{2} \log \det(1 + \mathbf{G}_{\Omega})}, \quad (121)$$

with

$$(\mathbf{G}_{\Omega})_{k,l} = \phi^a \left(\frac{v_t^2}{n} \delta_{ab} - \frac{a_t e^{-t} v_t p}{dn^2} \omega^a_k (1 + \mathbf{G}_Z)^{-1}_{k,l} \omega^b_l \right) \phi^{bT}. \quad (122)$$

We are left with

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} d\omega^a d\hat{\omega}^a e^{\frac{1}{2}(z-s_t^2)\phi^a \phi^{aT}} e^{-\frac{b_t^2 p}{2d} \omega^{aT} \Sigma^{-1} \omega^a} \\ &\mathbb{E}_{\mathbf{W}} [e^{i\hat{\omega}^a (\sqrt{p}\omega^a - \phi^{aT} \mathbf{W} \Sigma^{1/2})} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})}]. \end{aligned} \quad (123)$$

Averaging over the random features \mathbf{W} . \mathbf{W} only appears through $e^{-i\hat{\omega}^a \mathbf{W}^T \phi^a \Sigma^{1/2}}$.

$$\mathbb{E}_{\mathbf{W}} [e^{i \sum_a \hat{\omega}^a (\sqrt{p}\omega^a - \mathbf{W}^T \phi^a \Sigma_t^{1/2})}] = e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} (\mathbb{E}_{\mathbf{W}} [e^{-i\hat{\omega}_k^a \phi^a_i \mathbf{W}_{li} (\Sigma_t^{1/2})_{kl}}]) \quad (124)$$

$$= e^{i\sqrt{p} \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \hat{\omega}_k^a (\Sigma)_{kl} \hat{\omega}_l^b \phi_i^a \phi_i^b} \quad (125)$$

$$= e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b}. \quad (126)$$

We end up with

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\hat{\omega}^a e^{\frac{1}{2}(z-s_t^2)\phi^a \phi^{aT}} e^{-\frac{b_t^2 p}{2d} \omega^{aT} \Sigma^{-1} \omega^a} e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} \\ &\quad e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})}. \end{aligned} \quad (127)$$

Averaging over the $\hat{\omega}^a$. We can integrate with respect to $\hat{\omega}$. The only terms that appear with it are

$$\int \prod_a d\hat{\omega}^a e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b}. \quad (128)$$

Denote $\mathbf{J}_i^a = i\sqrt{p}\omega_i^a$ and $\mathbf{G}_{kl}^{ab} = \Sigma_{kl} \phi^a \cdot \phi^b$, then the integral is of the form

$$\int \prod_a d\hat{\omega}^a e^{\sum_{i,a} \mathbf{J}_i^a \hat{\omega}_i^a} e^{-\frac{1}{2} \sum_{i,j,a,b} \hat{\omega}_i^a \mathbf{G}_{ij}^{ab} \hat{\omega}_j^b} = e^{-\frac{1}{2} \log \det(\mathbf{G}) + \frac{1}{2} \mathbf{J}^T \mathbf{G}^{-1} \mathbf{J}}. \quad (129)$$

This gives

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a e^{\frac{1}{2}(z-s_t^2)\phi^a \phi^{aT}} e^{-\frac{b_t^2 p}{2d} \omega^{aT} \Sigma^{-1} \omega^a} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} \\ &\quad e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})} e^{-\frac{1}{2} \log \det(\mathbf{G}) + \frac{1}{2} \mathbf{J}^T \mathbf{G}^{-1} \mathbf{J}}. \end{aligned} \quad (130)$$

Introducing the order parameters. We define the order parameters as $\mathbf{Q}^{ab} = \frac{1}{p} \phi^a \cdot \phi^b$ and $\mathbf{R}^{ab} = \frac{1}{d} \omega^a \cdot \omega^b$. To enforce these constraints, we use the following delta function representations

$$1 = \int d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} e^{\frac{1}{2} \hat{\mathbf{Q}}^{ab} (p\mathbf{Q}^{ab} - \phi^a \cdot \phi^b)}, \quad (131)$$

$$1 = \int d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} e^{\frac{1}{2} \hat{\mathbf{R}}^{ab} (d\mathbf{R}^{ab} - \omega^a \cdot \omega^b)}, \quad (132)$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, Y, \Omega}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} \\
&\quad e^{\frac{1}{2}\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi^a\cdot\phi^b)} e^{\frac{1}{2}\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega^a\cdot\omega^b)} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \operatorname{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2d} \omega^{aT} \Sigma^{-1} \omega^a} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n} (v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R} (1 + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\Sigma \otimes \mathbf{Q})} e^{-\frac{1}{2} \omega_k^a \Sigma_{kl}^{-1} (\mathbf{Q}^{-1})_{ab} \omega_l^b}. \tag{133}
\end{aligned}$$

We also introduce $\mathbf{S}^{ab} = \omega_k^a \Sigma^{-1} \omega_l^b / d$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} d\mathbf{S}^{ab} d\hat{\mathbf{S}}^{ab} \\
&\quad e^{\frac{1}{2}\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi^a\cdot\phi^b)} e^{\frac{1}{2}\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega^a\cdot\omega^b)} e^{\frac{1}{2}\hat{\mathbf{S}}^{ab}(d\mathbf{S}^{ab}-\omega^a \Sigma^{-1} \omega^b)} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \operatorname{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2} \operatorname{Tr}(\mathbf{S})} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n} (v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R} (1 + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\Sigma \otimes \mathbf{Q})} e^{-\frac{d}{2} \operatorname{Tr}(\mathbf{S} \mathbf{Q}^{-1})}. \tag{134}
\end{aligned}$$

The integration over $d\phi^a$ and $d\omega^a$ gives

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} d\mathbf{S}^{ab} d\hat{\mathbf{S}}^{ab} \\
&\quad e^{\frac{p}{2} \operatorname{Tr}(\hat{\mathbf{Q}} \mathbf{Q})} e^{-\frac{p}{2} \log \det \hat{\mathbf{Q}}} e^{\frac{d}{2} \hat{\mathbf{R}}^{ab} \mathbf{R}^{ab}} e^{\frac{d}{2} \hat{\mathbf{S}}^{ab} \mathbf{S}^{ab}} \\
&\quad e^{-\frac{1}{2} \log \det(\hat{\mathbf{R}} \otimes \mathbf{I}_d + \hat{\mathbf{S}} \otimes \Sigma^{-1})} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \operatorname{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2} \operatorname{Tr}(\mathbf{S})} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n} (v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R} (1 + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\Sigma \otimes \mathbf{Q})} e^{-\frac{d}{2} \operatorname{Tr}(\mathbf{S} \mathbf{Q}^{-1})}. \tag{135}
\end{aligned}$$

We need to combine $e^{-\frac{1}{2} \log \det(\Sigma \otimes \mathbf{Q})}$ and $e^{-\frac{1}{2} \log \det(\hat{\mathbf{R}} \otimes \mathbf{I}_d + \hat{\mathbf{S}} \otimes \Sigma^{-1})}$,

$$e^{-\frac{1}{2} \log \det(\Sigma \otimes \mathbf{Q})} e^{-\frac{1}{2} \log \det(\hat{\mathbf{R}} \otimes \mathbf{I}_d + \hat{\mathbf{S}} \otimes \Sigma^{-1})} = e^{-\frac{1}{2} \log \det(\mathbf{Q} \hat{\mathbf{S}} \otimes \mathbf{I}_d + \mathbf{Q} \hat{\mathbf{R}} \otimes \Sigma)} \tag{136}$$

$$= e^{-\frac{d}{2} \log \det(\mathbf{Q} \hat{\mathbf{S}})} e^{-\frac{1}{2} \log \det(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \Sigma)} \tag{137}$$

Then for $e^{-\frac{1}{2} \log \det(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \Sigma)}$, we can introduce $\rho_\Sigma(\lambda)$ the density of eigenvalues of Σ

$$-\frac{1}{2} \log \det(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \Sigma) = -\frac{1}{2} \operatorname{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \Sigma) \tag{138}$$

$$= -\frac{1}{2} \sum_{l \geq 0} \frac{(-1)^l}{l!} (\hat{\mathbf{R}} \hat{\mathbf{S}}^{-1})^l \otimes \Sigma^l \tag{139}$$

$$= -\frac{d}{2} \int d\lambda \rho_\Sigma(\lambda) \sum_{l \geq 0} \frac{(-1)^l}{l!} \operatorname{Tr}((\hat{\mathbf{R}} \hat{\mathbf{S}}^{-1})^l) \lambda^l \tag{140}$$

$$= -\frac{d}{2} \int d\lambda \rho_\Sigma(\lambda) \operatorname{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1}). \tag{141}$$

We end up with

$$\mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega}[\mathcal{Z}^m] = \int d\mathbf{Q} d\hat{\mathbf{Q}} d\mathbf{R} d\hat{\mathbf{R}} d\mathbf{S} d\hat{\mathbf{S}} e^{-\frac{d}{2} S(\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}})}, \tag{142}$$

where the action reads

$$\begin{aligned}
S(\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}}) = & \psi_p \log \det \hat{\mathbf{Q}} - \psi_p \text{Tr}(\mathbf{Q}\hat{\mathbf{Q}}) - \text{Tr}(\mathbf{R}\hat{\mathbf{R}}) - \text{Tr}(\mathbf{S}\hat{\mathbf{S}}) \\
& - \psi_p(z - s_t^2) \text{Tr} \mathbf{Q} + \psi_n \log \det(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R}) + b_t^2 \psi_p \text{Tr} \mathbf{S} \\
& + \psi_n \log(\mathbf{I}_s + \frac{p}{n} (v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q}) \\
& + \log \det(\mathbf{Q}\hat{\mathbf{S}}) + \int d\lambda \rho_{\Sigma}(\lambda) \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}) + \text{Tr}(\mathbf{S}\mathbf{Q}^{-1}).
\end{aligned} \tag{143}$$

In the high dimensional limit, the partition function is dominated by the saddle point. By deriving with respect to $\hat{\mathbf{Q}}$ we get

$$\hat{\mathbf{Q}}^{-1} = \mathbf{Q}, \tag{144}$$

which yields

$$\begin{aligned}
S(\mathbf{Q}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}}) = & -\psi_p \log \det \mathbf{Q} - \text{Tr}(\mathbf{R}\hat{\mathbf{R}}) - \text{Tr}(\mathbf{S}\hat{\mathbf{S}}) \\
& - \psi_p(z - s_t^2) \text{Tr} \mathbf{Q} + \psi_n \log \det(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R}) + b_t^2 \psi_p \text{Tr} \mathbf{S} \\
& + \psi_n \log(\mathbf{I}_s + \frac{p}{n} (v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q}) \\
& + \log \det(\mathbf{Q}\hat{\mathbf{S}}) + \int d\lambda \rho_{\Sigma}(\lambda) \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}) \\
& + \text{Tr}(\mathbf{S}\mathbf{Q}^{-1}).
\end{aligned} \tag{145}$$

As a sanity check, if $\Sigma = \mathbf{I}_d$, differentiation with respect to $\hat{\mathbf{R}}$ and $\hat{\mathbf{S}}$ yields

$$\mathbf{R} = \mathbf{S} = (\hat{\mathbf{S}} + \hat{\mathbf{R}})^{-1}, \tag{146}$$

and we find back the same action as before.

RS Ansatz. As before we introduce a RS ansatz for all the the matrices and moreover suppose that only the diagonal terms are non vanishing i.e. they are of the form $\mathbf{Q} = q\mathbf{I}_s$. This ansatz yields

$$\begin{aligned}
S(q, r, \hat{r}, s, \hat{s})/s = & -\psi_p \log q - r\hat{r} - s\hat{s} \\
& - \psi_p(z - s_t^2)q + \psi_n \log(1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q) + b_t^2 \psi_p s \\
& + \log(q) + \int d\lambda \rho_{\Sigma}(\lambda) \log(\hat{s} + \lambda \hat{r}) + \frac{s}{q}.
\end{aligned} \tag{147}$$

Let us differentiate with respect to the 5 variables

$$\frac{\partial S}{\partial s} = -\hat{s} + b_t^2 \psi_p + \frac{1}{q}, \tag{148}$$

$$\frac{\partial S}{\partial r} = -\hat{r} + \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q}, \tag{149}$$

$$\frac{\partial S}{\partial \hat{s}} = -s + \int d\lambda \rho_{\Sigma}(\lambda) \frac{1}{\hat{s} + \lambda \hat{r}}, \tag{150}$$

$$\frac{\partial S}{\partial \hat{r}} = -r + \int d\lambda \rho_{\Sigma}(\lambda) \frac{\lambda}{\hat{s} + \lambda \hat{r}}, \tag{151}$$

$$\frac{\partial S}{\partial q} = -\frac{\psi_p}{q} - \psi_p(z - s_t^2) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q} + \frac{1}{q} - \frac{s}{q^2}. \tag{152}$$

Hence the saddle point equations read

$$\hat{s} = b_t^2 \psi_p + \frac{1}{q}, \quad (153)$$

$$\hat{r} = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q}, \quad (154)$$

$$s = \int d\rho_{\Sigma}(\lambda) \frac{1}{\hat{s} + \lambda \hat{r}}, \quad (155)$$

$$r = \int d\rho_{\Sigma}(\lambda) \frac{\lambda}{\hat{s} + \lambda \hat{r}}, \quad (156)$$

$$\psi_p(s_t^2 - z) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q} + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0. \quad (157)$$

Finally, we observe that the solution q^* to the saddle point equations corresponds to the Stieltjes transform of ρ .

$$2\partial_z \frac{1}{p} \frac{\mathbb{E}[\mathcal{Z}^s] - 1}{s} = 2\partial_z \frac{1}{p} \frac{e^{-\frac{d}{2}S(q^*, r^*)} - 1}{m} \xrightarrow{m \rightarrow 0} -2\partial_z \frac{1}{p} \frac{d}{2} S(q^*, r^*) = q^*. \quad (158)$$

□

C.5 Proof of Theorem 3.2

We recall Theorem 3.2 of the MT.

Theorem C.2 (Informal). *Let ρ denote the spectral density of \mathbf{U} .*

Regime I (overparametrized): $\psi_p > \psi_n \gg 1$.

$$\rho(\lambda) = \left(1 - \frac{1 + \psi_n}{\psi_p}\right) \delta(\lambda - s_t^2) + \frac{\psi_n}{\psi_p} \rho_1(\lambda) + \frac{1}{\psi_p} \rho_2(\lambda).$$

Regime II (underparametrized): $\psi_n > \psi_p \gg 1$.

$$\rho(\lambda) = \left(1 - \frac{1}{\psi_p}\right) \rho_1(\lambda) + \frac{1}{\psi_p} \rho_2(\lambda).$$

where ρ_1 is a atomless measure with support

$$\left[s_t^2 + v_t^2 \left(1 - \sqrt{\psi_p/\psi_n}\right)^2, s_t^2 + v_t^2 \left(1 + \sqrt{\psi_p/\psi_n}\right)^2 \right],$$

and ρ_2 coincides with the asymptotic eigenvalue bulk density of the population covariance $\tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{X}}[\mathbf{U}]$; ρ_2 is independent of ψ_n and its support is on the scale ψ_p . The eigenvectors associated with $\delta(\lambda - s_t^2)$ leave both training and test losses unchanged and are therefore irrelevant. In the limit $\psi_p \gg \psi_n$, the supports of ρ_1 and ρ_2 are respectively on the scales ψ_p/ψ_n and ψ_p , i.e. they are well separated.

We now proceed to prove Theorem 3.2.

Proof. Delta peak. We first account for the delta peak in the spectrum. We use the Gaussian equivalence for \mathbf{U} computed in Lemma C.1. Let $\Omega^\nu \in \mathbb{R}^p$ be the ν th column of Ω and $\mathbf{W}_i \in \mathbb{R}^p$ the i th row of \mathbf{W} . Suppose a vector $\mathbf{v} \in \mathbb{R}^p$ lies in the kernel of all these

$$\forall \nu = 1, \dots, n, \quad \sum_{i=1}^p \Omega_i^\nu \mathbf{v}_i = 0, \quad (159)$$

$$\forall k = 1, \dots, d, \quad \sum_{i=1}^p \mathbf{W}_{ik} \mathbf{v}_i = 0. \quad (160)$$

then $\mathbf{U}\mathbf{v} = s_t^2 \mathbf{v}$. These are $n + d$ linear constraints on a vector of size p hence there are non trivial solutions for $n + d \leq p$. Hence a delta-peak at s_t^2 appears as soon as $\psi_p \geq \psi_n + 1$. Next, we extract its weight. Recall that the Stieltjes transform satisfies

$$q(z) = \int \frac{\rho(\lambda)}{\lambda - z} d\lambda,$$

and a point mass of weight f at $\lambda = s_t^2$ contributes $\frac{-f}{z-s_t^2} \approx \frac{f}{\varepsilon}$ as $z \rightarrow s_t^2 - \varepsilon$. Meanwhile

$$s(z) = \frac{1}{p} \text{Tr}[\mathbf{W}^T (\mathbf{U} - z\mathbf{I})^{-1} \mathbf{W}], \quad r(z) = \frac{1}{p} \text{Tr}[\boldsymbol{\Sigma}^{1/2} \mathbf{W}^T (\mathbf{U} - z\mathbf{I})^{-1} \mathbf{W} \boldsymbol{\Sigma}^{1/2}]$$

remain finite in that limit, since the corresponding eigenvectors satisfy $\mathbf{W}\mathbf{v} = 0$. We substitute this Ansatz into the equations of Theorem 3.1. The first equation reads

$$\psi_n \frac{\frac{pv_t^2}{n}}{1 + \frac{e^{-2t} \mu_1^2 p \sigma_x^2}{n} r + \frac{p}{n} v_t^2 q} + \psi_p(s_t^2 - z) + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0, \quad (161)$$

and simplifies to

$$\frac{\psi_n \varepsilon}{f} + \psi_p \varepsilon + \frac{(1 - \psi_p) \varepsilon}{f} = 0. \quad (162)$$

It readily gives

$$f = 1 - \frac{1}{\psi_p} - \frac{\psi_n}{\psi_p}. \quad (163)$$

Thus the point mass at s_t^2 has weight $1 - \frac{1}{\psi_p} - \frac{\psi_n}{\psi_p}$, in agreement with the counting of degrees of freedom presented above.

Finally, one checks that these isolated eigenvalues do not contribute to the train and test losses. After expanding the square they read

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \mathbf{V}\right) \quad (164)$$

$$\mathcal{L}_{\text{test}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{U}}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{V}}\right) \quad (165)$$

The terms that appear in the loss are of the form $\text{Tr}(\mathbf{A}^T \mathbf{A} \dots)$ and $\text{Tr}(\mathbf{A} \mathbf{W})$. The trace can be decomposed on the basis of eigenvectors of \mathbf{U} . The eigenvectors associated with the delta peak satisfy $\mathbf{W}^T \mathbf{v} = 0$. Looking at the expression of the matrix $\mathbf{A} = \mathbf{W}^T \dots + \mathbf{A}_0$, one can easily see that, for initial conditions $\mathbf{A}_0 = 0$, one has $\mathbf{v}^T \mathbf{A}^T = 0$ and the subspace corresponding to these isolated eigenvalues does not contribute to the loss.

First bulk. Using the expression for $q = \frac{1}{p} \text{Tr} \frac{1}{\mathbf{U} - z\mathbf{I}_p}$ and $r(z) = \frac{1}{p} \text{Tr}(\boldsymbol{\Sigma}^{1/2} \mathbf{W}^T (\mathbf{U} - z\mathbf{I})^{-1} \mathbf{W} \boldsymbol{\Sigma}^{1/2})$ we make the following Ansatz in the large ψ_p limit:

$$q = \mathcal{O}_{\psi_p}(1), \quad r = \mathcal{O}_{\psi_p}\left(\frac{1}{\psi_p}\right). \quad (166)$$

In this limit the saddle point equations becomes at leading order in ψ_p

$$\hat{s} = b_t^2 \psi_p \quad (167)$$

$$\hat{r} = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{v_t^2 p}{n} r} \quad (168)$$

$$s = \mathcal{O}(1/\psi_p) \quad (169)$$

$$r = \mathcal{O}(1/\psi_p) \quad (170)$$

$$(s_t^2 - z) + \frac{v_t^2}{1 + \frac{pv_t^2}{n} q} - \frac{1}{q} = 0. \quad (171)$$

We can focus only on the last equation on q only. This is a quadratic polynomial in q . If its discriminant is negative then the solutions are imaginary and thus the density of eigenvalues is non-zero. The edge of the bulk are where the discriminant vanishes

$$\Delta = (s_t^2 - \lambda(1 - \frac{p}{n})v_t^2)^2 + 4(s_t^2 - \lambda)\frac{p}{n}v_t^2 = 0. \quad (172)$$

It vanishes for

$$\lambda_{\pm} = s_t^2 + v_t^2 \left(1 \pm \sqrt{\frac{p}{n}} \right)^2 \quad (173)$$

which are the edges of the first bulk ρ_1 . We have checked this result, and hence validated the Ansatz solving numerically the equations on r, q . Interestingly at leading order the expression of the first bulk is independent of ρ_{Σ} .

Second Bulk. We scale $q = \mathcal{O}_{\psi_p}(1/\psi_p)$ and $r = \mathcal{O}_{\psi_p}(1/\psi_p)$. The equations on \hat{s} and \hat{r} lead to

$$\hat{s} = \psi_p b_t^2 + \frac{1}{q} \quad (174)$$

$$\hat{r} = \psi_p a_t^2 e^{-2t}. \quad (175)$$

This yields the following equation on q

$$\psi_p(s_t^2 - z) + \psi_p v_t^2 + \frac{1 - \psi_p}{q} - \frac{1}{q} \int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = 0. \quad (176)$$

We denote the shifted variable $z' = z - s_t^2 - v_t^2$. This yields

$$-\psi_p z' + \frac{1 - \psi_p}{q} - \frac{1}{q} \int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = 0. \quad (177)$$

We decompose the integral

$$\int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = \int \frac{d\rho_{\Sigma}(\lambda)(1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t}) - q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t}))}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} \quad (178)$$

$$= 1 - q\psi_p \int \frac{d\rho_{\Sigma}(\lambda)(b_t^2 + \lambda a_t^2 e^{-2t})}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} \quad (179)$$

By plugging this back in the equation we find

$$q = - \left(z' - \int \frac{d\rho_{\Sigma}(\lambda)(b_t^2 + \lambda a_t^2 e^{-2t})}{1 + \psi_p q(b_t^2 + \lambda a_t^2 e^{-2t})} \right)^{-1}. \quad (180)$$

We do the change of variable $\mu = b_t^2 + \lambda a_t^2 e^{-2t}$. This yields

$$q = - \left(z' - \frac{1}{a_t^2 e^{-2t}} \int \frac{d\mu \rho_{\Sigma}(\frac{\mu - b_t^2}{a_t^2 e^{-2t}}) \mu}{1 + \psi_p q \mu} \right)^{-1}. \quad (181)$$

An integration by parts give that $b_t^2 = \Delta_t \mu_1^2(t)$ $a_t^2 = \mu_1^2(t)/\sigma_x^2$. We thus realize that the integral is over the eigenvalue distribution of $\mu_1^2(t)(e^{-2t}\Sigma + \Delta_t \mathbf{I}_d)$,

$$q = - \left(z' - \int \frac{d\mu \rho_{\mu_1^2(t)\Sigma_t}(\mu) \mu}{1 + \psi_p q \mu} \right)^{-1}. \quad (182)$$

We recognize the Bai-Silverstein equations [48, 5] for the eigenvalue density of the matrix

$$\tilde{\mathbf{U}} = \mu_1^2(t) \frac{\mathbf{W} \Sigma_t \mathbf{W}^T}{d} + (s_t^2 + v_t^2) \mathbf{I}_p = \mathbb{E}_{\mathbf{x}}[\mathbf{U}] \quad (183)$$

which is the population version of \mathbf{U} and is thus independent of n . Lemma C.3 concludes on the order of the eigenvalues in the bulk of ρ_2 .

□

C.6 Dynamics on the fast timescales

In the following we denote for a matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$,

$$\|\mathbf{A}\|_{\text{op}} = \sup_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \|\mathbf{Av}\| \quad (184)$$

the operator norm and

$$\|\mathbf{A}\|_{\text{F}} = \left(\sum_{i,j=1}^p \mathbf{A}_{ij}^2 \right)^{1/2} \quad (185)$$

the Frobenius norm. Before deriving the fast-time behavior, we need the following lemma.

Lemma C.5. *The operator norm of $\mathbf{U} - \tilde{\mathbf{U}}$ satisfies*

$$\|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} = \mathcal{O}\left(\frac{\psi_p}{\sqrt{\psi_n}}\right), \quad (186)$$

when $p \gg n \gg d$.

Proof. On the one hand,

$$\mathbf{U} = e^{-2t} a_t^2 \frac{\mathbf{W} \mathbf{X} \mathbf{X}^T \mathbf{W}^T}{d} + v_t^2 \frac{\Omega \Omega^T}{n} + \frac{e^{-t} a_t v_t}{n \sqrt{d}} \left(\mathbf{W} \mathbf{X} \Omega^T + \Omega \mathbf{X}^T \mathbf{W}^T \right) + (s_t^2 + v_t^2) \mathbf{I}_p \quad (187)$$

and on the other hand,

$$\tilde{\mathbf{U}} = \mu_1^2 e^{-2t} \frac{\mathbf{W} \Sigma \mathbf{W}^T}{d} + \Delta_t \mu_1^2 \frac{\mathbf{W} \mathbf{W}^T}{d} + (s_t^2 + v_t^2) \mathbf{I}_p. \quad (188)$$

We also note the identities $b_t^2 = \Delta_t \mu_1^2(t)$ and $a_t^2 = \mu_1^2(t)$.

$$\mathbf{U} - \tilde{\mathbf{U}} = a_t^2 e^{-2t} \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}} + v_t^2 \left(\frac{\Omega \Omega^T}{n} - \mathbf{I}_p \right) + \frac{a_t v_t e^{-t}}{n \sqrt{d}} (\Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T). \quad (189)$$

We can bound its operator norm

$$\begin{aligned} \|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} &\leq C_1 \left\| \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} + C_2 \left\| \left(\frac{\Omega \Omega^T}{n} - \mathbf{I}_p \right) \right\|_{\text{op}} \\ &\quad + \frac{C_3}{n \sqrt{d}} \|\Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T\|_{\text{op}}, \end{aligned} \quad (190)$$

where C_1, C_2, C_3 are constants independent of p, n, d . We bound each of the three terms on the right hand side. We will use the fact that for a symmetric matrix, the operator norm $\|\cdot\|_{\text{op}}$ is equal to its largest eigenvalue.

First term.

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}}. \quad (191)$$

We observe that $\frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}}$ and $\frac{\mathbf{W}^T}{\sqrt{d}} \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right)$ have the same eigenvalues up to the multiplicity of 0[†]. We then use the sub-multiplicativity of the operator norm

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} \leq \left\| \frac{\mathbf{W}^T}{\sqrt{d}} \frac{\mathbf{W}}{\sqrt{d}} \right\|_{\text{op}} \left\| \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \right\|_{\text{op}}. \quad (192)$$

We can do the same operation by introducing $\mathbf{X} = \Sigma \mathbf{Z}$ with $\mathbf{Z} \in \mathbb{R}^{d \times n}$ with standard Gaussian entries,

$$\left\| \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \right\|_{\text{op}} = \left\| \Sigma^{1/2} \left(\frac{\mathbf{Z} \mathbf{Z}^T}{n} - \mathbf{I}_d \right) \Sigma^{1/2} \right\|_{\text{op}} \leq \left\| \left(\frac{\mathbf{Z} \mathbf{Z}^T}{n} - \mathbf{I}_d \right) \right\|_{\text{op}} \|\Sigma\|_{\text{op}}. \quad (193)$$

Among our assumptions, we had $\|\Sigma\|_{\text{op}} < \mathcal{O}(1)$. The spectrum of $\left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \mathbf{I}_d \right)$ is the Marchenko-Pastur law whose largest eigenvalue is of order $\sqrt{d/n}$ while for $\frac{\mathbf{W}^T \mathbf{W}}{d}$ it is order $\frac{p}{d}$. The bound reads

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left(\frac{\mathbf{X} \mathbf{X}^T}{n} - \Sigma \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} \leq \mathcal{O}\left(\frac{p}{\sqrt{nd}}\right). \quad (194)$$

[†]They both have the same moments $\text{Tr}(\cdot)^k$ owing to the cyclicity of the trace.

Second term.

$$\left\| \left(\frac{\Omega \Omega^T}{n} - \mathbf{I}_p \right) \right\|_{\text{op}}. \quad (195)$$

We observe that the spectrum of $\Omega \Omega^T / n - \mathbf{I}_p$ is Marchenko-Pastur and thus its largest eigenvalue is order $\mathcal{O}(p/n)$ yielding

$$\left\| \left(\frac{\Omega \Omega^T}{n} - \mathbf{I}_p \right) \right\|_{\text{op}} \leq \mathcal{O}(p/n). \quad (196)$$

Third term.

$$\left\| \Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T \right\|_{\text{op}}. \quad (197)$$

We first bound the operator norm by the Frobenius norm.

$$\left\| \Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T \right\|_{\text{op}} \leq 2 \left\| \Omega \mathbf{X}^T \mathbf{W}^T \right\|_{\text{F}}. \quad (198)$$

We expand the square

$$\left\| \Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T \right\|_{\text{F}}^2 = C \sum_{k=1}^d \sum_{i=1}^p \left(\sum_{\nu=1}^n \Omega_i^\nu \mathbf{X}_k^\nu \mathbf{W}_{kl} \right)^2. \quad (199)$$

The Central Limit Theorem yields

$$\sum_{\nu=1}^n \Omega_i^\nu \mathbf{X}_k^\nu \mathbf{W}_{kl} = \mathcal{O}(\sqrt{n}) \mathbf{W}_{kl}, \quad (200)$$

hence

$$\frac{1}{n\sqrt{d}} \left\| \Omega \mathbf{X}^T \mathbf{W}^T + \mathbf{W} \mathbf{X} \Omega^T \right\|_{\text{op}} = \mathcal{O}\left(\frac{\sqrt{ndp}}{n\sqrt{d}}\right) = \mathcal{O}\left(\sqrt{\frac{p}{n}}\right) \quad (201)$$

Putting all the contributions together yields

$$\|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} \leq \mathcal{O}\left(\frac{p}{\sqrt{dn}}\right) = \mathcal{O}\left(\frac{\psi_p}{\sqrt{\psi_n}}\right). \quad (202)$$

□

Proposition C.2 (Informal). *On timescales $1 \ll \tau \ll \psi_n$, both the train and test losses satisfy*

$$\mathcal{L}_{\text{train}} \simeq \mathcal{L}_{\text{test}} \simeq 1 - \mathcal{O}(\Delta_t). \quad (203)$$

Proof. According to the spectral analysis of \mathbf{U} conducted previously, there are two bulks in the spectrum that contribute to the dynamics: a first bulk with eigenvalues of order $\frac{\psi_p}{\psi_n}$ and a second bulk with eigenvalues of order ψ_p in the $\psi_p, \psi_n \gg 1$ limit. Hence, in the regime $1 \ll \tau \ll \psi_n$, $e^{-\lambda \frac{\Delta_t \tau}{\psi_p}} \sim 0$ if λ is in the second bulk and is $e^{-\lambda \frac{\Delta_t \tau}{\psi_p}} \sim 1$ if λ is in the first bulk. We remind the expressions of the train and test loss

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \mathbf{V}\right) \quad (204)$$

$$\mathcal{L}_{\text{test}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{U}}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{V}}\right) \quad (205)$$

and use the expression of $\mathbf{A}(\tau)$ in Proposition C.1 that we expand on the basis of eigenvectors $\{\mathbf{v}_\lambda\}_{\lambda \in Sp(\mathbf{U})}$ of \mathbf{U} .

$$\frac{\mathbf{A}(\tau)}{\sqrt{p}} = \frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} (e^{-\frac{2\Delta_t}{d} \mathbf{U}\tau} - \mathbf{I}_p) \quad (206)$$

$$= \frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} \sum_{\lambda} (e^{-\frac{2\Delta_t}{d} \lambda \tau} - 1) \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T \quad (207)$$

$$\sim -\frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} \sum_{\lambda \in \rho_2} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T, \quad (208)$$

where $\lambda \in \rho_2$ means that the eigenvalue λ belongs to the second bulk. We also have that \mathbf{V} and $\tilde{\mathbf{V}}$ have the same GEP $\frac{\mu_1(t)\sqrt{\Delta_t}}{\Gamma_t} \frac{\mathbf{W}}{\sqrt{d}}$ and they thus cancel each other when computing the generalization loss $\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$. It reads

$$\mathcal{L}_{\text{gen}} = -\frac{\mu_1^2(t)\Delta_t}{\Gamma_t^2 d} \text{Tr}\left(\sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'} \mathbf{v}_{\lambda'}^T \mathbf{U}^{-1} \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}})\right) \quad (209)$$

$$= -\frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left(\sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \mathbf{U}^{-1} \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \quad (210)$$

$$= -\frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left(\sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \quad (211)$$

(212)

We then use Lemma C.5 — which states that the operator norm of $\mathbf{U} - \tilde{\mathbf{U}}$ in the subspace spanned by the eigenvectors of the second bulk is bounded by $\mathcal{O}(\frac{\psi_p}{\sqrt{\psi_n}})$ — to bound \mathcal{L}_{gen} ,

$$|\mathcal{L}_{\text{gen}}| \leq \left\| \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left(\sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \right\|_{\text{op}} \quad (213)$$

$$\leq \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} d \frac{1}{\psi_p^2} \left\| \frac{\mathbf{W} \mathbf{W}^T}{d} \right\|_{\text{op}} \frac{\psi_p}{\sqrt{\psi_n}} \leq \mathcal{O}\left(\frac{d \psi_p^2}{d \psi_p^2 \sqrt{\psi_n}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{\psi_n}}\right). \quad (214)$$

We also used the fact that the sums contain d terms — the only terms that matter are the diagonal ones — and that the eigenvalues scale as ψ_p . The bound yield that \mathcal{L}_{gen} vanishes asymptotically in the large number of data and large number of parameters regime. Therefore, on the fast timescale we find $\mathcal{L}_{\text{train}} \simeq \mathcal{L}_{\text{test}}$. Let us now focus on $\mathcal{L}_{\text{train}}$

$$\mathcal{L}_{\text{train}} = 1 + \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left(\sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T \mathbf{U} \mathbf{v}_{\lambda'} \right) - \frac{2 \Delta_t \mu_1^2}{\Gamma_t^2 d} \sum_{\lambda \in \rho_2} \mathbf{v}_{\lambda}^T \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \quad (215)$$

$$= 1 - \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \sum_{\lambda \in \rho_2} \frac{1}{\lambda} \mathbf{v}_{\lambda}^T \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{v}_{\lambda}. \quad (216)$$

There are d values in the sum and the eigenvalues of \mathbf{U} and $\frac{\mathbf{W} \mathbf{W}^T}{d}$ are both order $\mathcal{O}(\psi_p)$ hence the sum divided by d is a positive $\mathcal{O}(1)$ quantity thus in this training time regime, $1 \ll \tau \ll \psi_n$, we obtain:

$$\mathcal{L}_{\text{train}} \sim \mathcal{L}_{\text{test}} = 1 - \mathcal{O}(\Delta_t). \quad (217)$$

□

D Numerical experiments for Random Features

Details on the numerical experiments. All the numerical experiments for the RFNN were conducted using $\sigma = \tanh$ and $\sigma_{\mathbf{x}} = 1$ unless specified. At each step, the gradient of the loss was computed using the full batch of data points. The train loss was estimated by adding noise to each data point $N = 100$ times. The test loss was computed by drawing n new points from the data distribution and noising each one N times. The error on the score was evaluated by drawing 10,000 points from the noisy distribution $P_t = \mathcal{N}(0, \Gamma_t^2 \mathbf{I}_d)$.

Effect of t . We present plots for different diffusion times t in Fig. 11 and show that the rescaling of the training times τ by $\tau_{\text{mem}} = \psi_p / \Delta_t \lambda_{\min}$ also makes the loss curves collapse. Of particular interest is the behavior of τ_{mem} , and more specifically the ratio $\tau_{\text{mem}} / \tau_{\text{gen}}$, at small t . Recall that

$$\lambda_{\min} = s_t^2 + v_t^2 \left(1 - \sqrt{\frac{\psi_p}{\psi_n}} \right)^2.$$

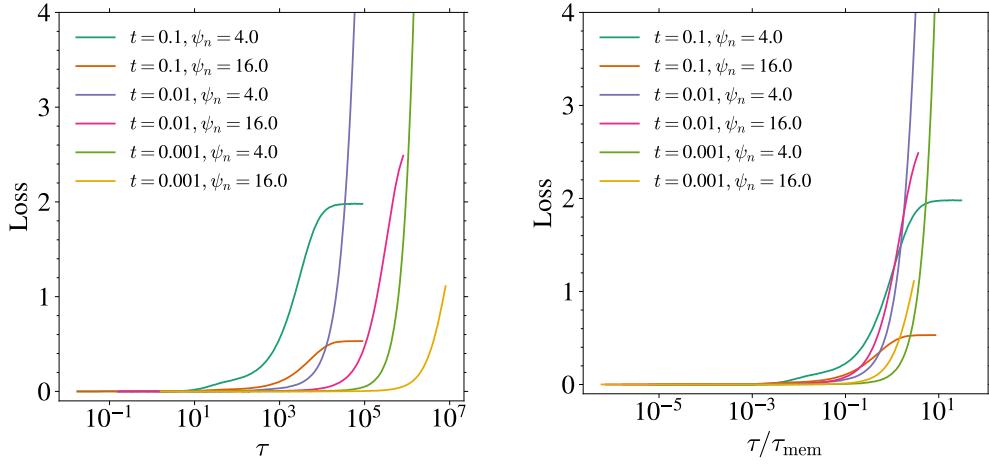


Figure 11: Generalization loss for different diffusion times t . Generalization loss \mathcal{L}_{gen} against (Left) training time τ and (Right) rescaled training time τ/τ_{gen} for different $\psi_p = 32$, $d = 100$ and different ψ_n and t .

In the overparameterized regime $p \gg n$, this ratio is independent of t since $v_t^2 \sim \mu_*^2$ and $s_t^2 \sim t$. However, when $p \sim n$, a nontrivial scaling emerges: since $\lambda_{\min} \sim s_t^2 \sim t$, it follows that

$$\frac{\tau_{\text{mem}}}{\tau_{\text{gen}}} \sim \frac{1}{t},$$

implying that the two timescales become increasingly separated. It is unclear whether this behavior is related to specific properties of the learned score function, and is related to the approach of the interpolation threshold. We leave this question for future investigation.

Experiments with $\sigma_x^2 \neq 1$. In Fig. 12, we present train and test loss curves for $\sigma_x \neq 1$. We see that our prediction of the timescale of memorization computed in the MT holds for general data variance.

Scaling of $\mathcal{E}_{\text{score}}$ with n . In the RF model, the error with respect to the true score, as defined in the main text,

$$\mathcal{E}_{\text{Score}} = \frac{1}{d} \mathbb{E}_{\mathbf{y} \sim \mathcal{N}(0, \Gamma_t^2 \mathbf{I}_p)} \left[\left\| \mathbf{s}_{\mathbf{A}(\tau)}(\mathbf{y}) + \frac{\mathbf{y}}{\Gamma_t^2} \right\|^2 \right], \quad (218)$$

serves as a measure of the generalization capability of the generative process. As shown in [53], the Kullback–Leibler divergence between the true data distribution P_x and the generated distribution \hat{P} can be upper bounded

$$\mathcal{D}_{\text{KL}}(P_x \parallel \hat{P}) \leq \frac{d}{2} \int dt \mathcal{E}_{\text{Score}}(\mathbf{A}_t), \quad (219)$$

where the integral is taken over all estimations of the parameter matrix \mathbf{A} at all diffusion times t . This bound assumes that the reverse dynamics are integrated exactly, starting from infinite time. In practical settings, however, one typically relies on an approximate scheme and initiates the reverse process at a large but finite time T . A generalization of this bound under such conditions can be found in [8]. We have numerically investigate the behaviour of $\mathcal{E}_{\text{score}}$ on Fig. 13. On the fast timescale τ_{gen} , it decreases until a minimal value $\mathcal{E}_{\text{score}}^*$ that depends only on ψ_n with a power-law $\psi_n^{-\eta}$ with $\eta \simeq 0.59$. We leave for future work performing an accurate numerical estimate of η and a developing a theory for it.

Spectrum of \mathbf{U} . In Fig. 14, we compare the solutions of the equations of Theorem 3.1 to the histogram of finite size realizations of \mathbf{U} .

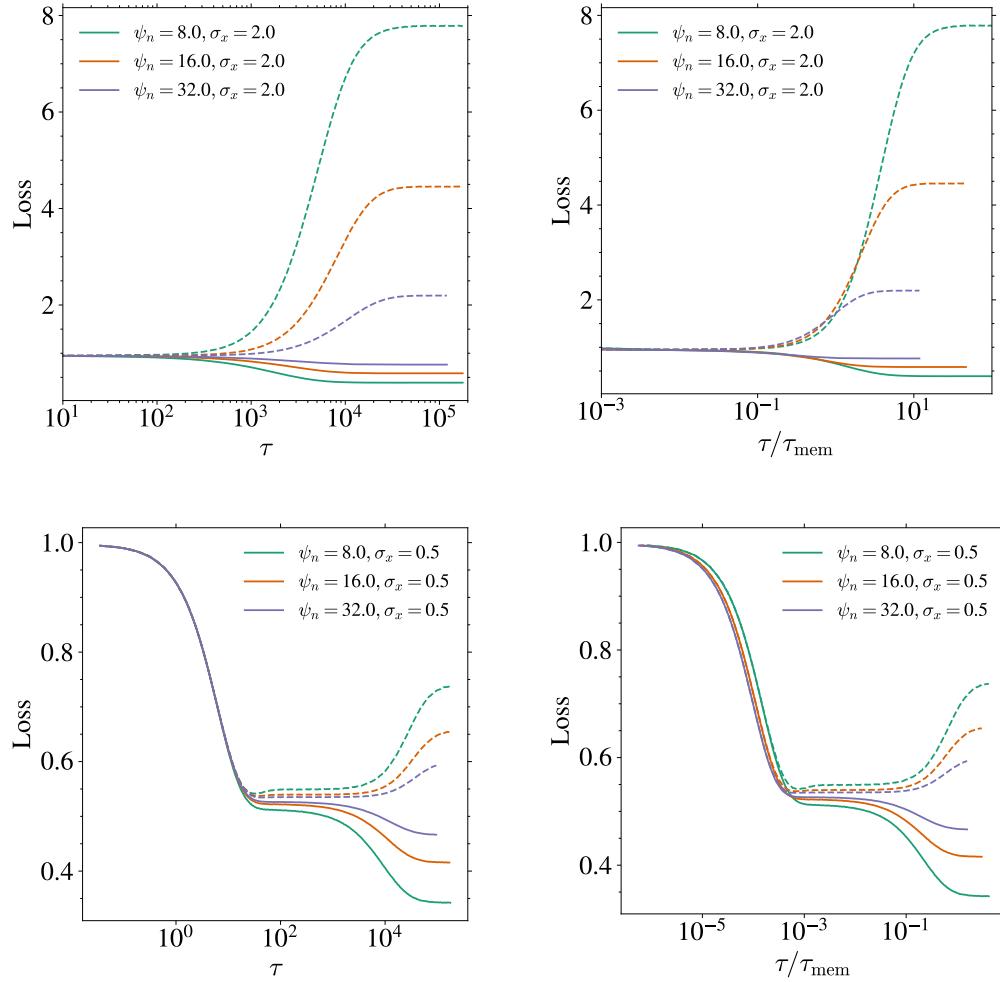


Figure 12: **Different σ_x^2 .** Train loss (solid line) and test loss (dotted line) for $\psi_p = 64, t = 0.1, d = 100$, different ψ_n and $\sigma_x = 2$.(top) and $\sigma_x = 0.5$ (bottom) against the training time τ and the rescaled training time τ/τ_{mem} .

Effect of Adam optimization. Numerical experiments with RFNN on Gaussian data show that the linear scaling of the memorization time with n holds also for the Adam optimizer as shown in Fig.15.

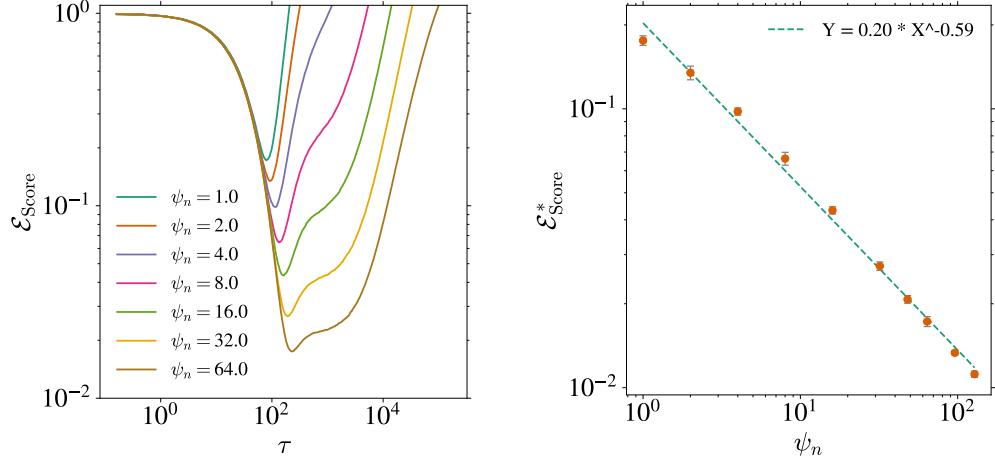


Figure 13: **Effect of ψ_n on $\mathcal{E}_{\text{Score}}^*$.** (Left) Error between the learned score and the true score $\mathcal{E}_{\text{Score}}$ for $\psi_p = 32$, $t = 0.01$, and various values of ψ_n . (Right) Minimum score error $\mathcal{E}_{\text{Score}}^* = \min_{\tau} [\mathcal{E}_{\text{Score}}(\tau)]$ as a function of ψ_n , showing a power-law decay with exponent approximately -0.59 . The error bars correspond to thrice the standard deviation over 10 runs with new initial conditions.

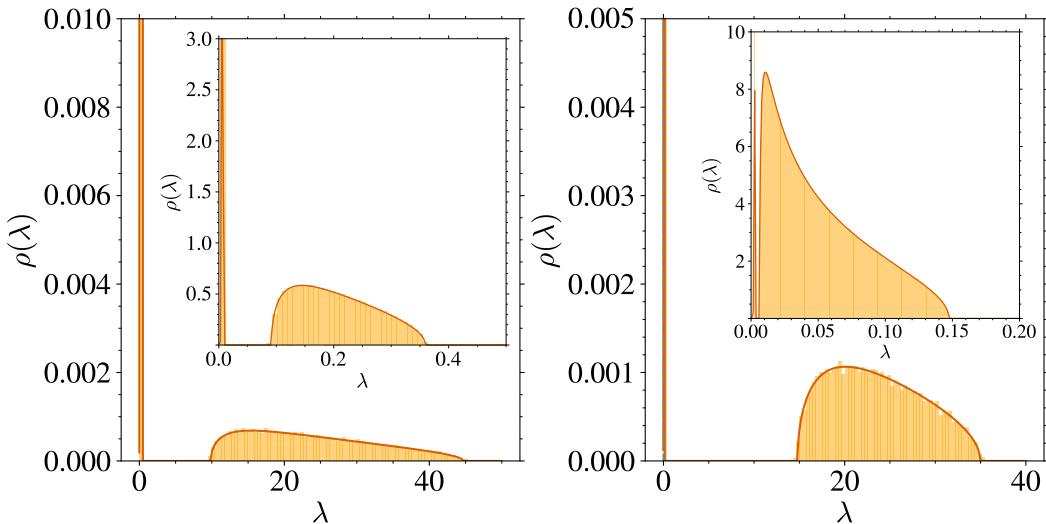


Figure 14: **Spectrum of \mathbf{U} .** Solutions of the equations in Theorem 3.1. (solid lines) and empirical spectrum for $\rho_{\Sigma}(\lambda) = \delta(\lambda - 1)$ and $d = 100$ (histogram). (Left) $\psi_p = 64$, $\psi_n = 8$, $t = 0.01$. (Right) $\psi_p = 64$, $\psi_n = 32$, $t = 0.01$.

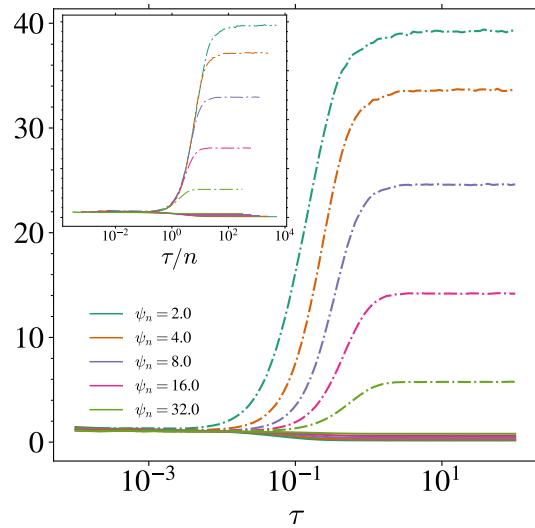


Figure 15: **Adam.** Train loss (solid line) and test loss (dotted line) at $t = 0.01$, $d = 100$, $\psi_p = 64$ for several ψ_n with the Pytorch [38] implementation of Adam. The inset shows the effect of a rescaling of the training time by n .