

---

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

## Supplementary Material

---

**Tony Bonnaire<sup>†</sup>**  
LPENS  
Université PSL, Paris  
tony.bonnaire@phys.ens.fr

**Raphaël Urfin<sup>†</sup>**  
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

This document provides detailed derivations and additional experiments supporting our paper, *Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training*. In Sect. [A](#), we give details about the numerical experiments carried out in Sect. [2](#) of the Main Text (MT). 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](#). Throughout this document, all [teal](#) hyperlinks point back to equations or sections in the MT, while [red](#) hyperlinks refer to items in the current supplement.

---

## Contents

<b>A Numerical experiments on CelebA</b>	<b>2</b>
A.1 Details on the numerical setup . . . . .	2
A.2 Batch-size effect: repetition vs. memorization . . . . .	4
A.3 What about Adam? . . . . .	4
<b>B Generalization–memorization transition in the Gaussian Mixture Model</b>	<b>4</b>
B.1 Settings . . . . .	4
B.2 Scaling of $\tau_{\text{mem}}$ and $\tau_{\text{gen}}$ with $n$ and $W$ . . . . .	6
B.3 Discussion on conditional diffusion models . . . . .	6
<b>C Proofs of the analytical results</b>	<b>7</b>
C.1 Notations . . . . .	7
C.2 Closed form of the learning dynamics . . . . .	8
C.3 Gaussian Equivalence Principle . . . . .	8
C.4 Proof of Theorem 3.1 . . . . .	13

---

<sup>†</sup>Equal contribution.

C.5 Proof of Theorem 3.2. . . . .	18
C.6 Dynamics on the fast timescales . . . . .	21

<b>D Numerical experiments for Random Features</b>	<b>23</b>
--	-----------

---

## 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 [15]. We center-crop each RGB image to  $32 \times 32$  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. 1.

**Architecture.** As commonly done in DDPMs implementations [e.g., 10, 25], the network approximating the score function is a U-Net [20] 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 [26]. 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

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

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

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

**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 [14]. 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  $t = T', \dots, 1$

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}(t-1)} \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}(t)} \boldsymbol{\xi}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}(t)}} + \sqrt{1 - \bar{\alpha}(t-1)} \boldsymbol{\xi}_\theta(\mathbf{x}_t, t), \quad (3)$$

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. 1 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 ( $f_{\text{mem}} = 47.2\%$ ).

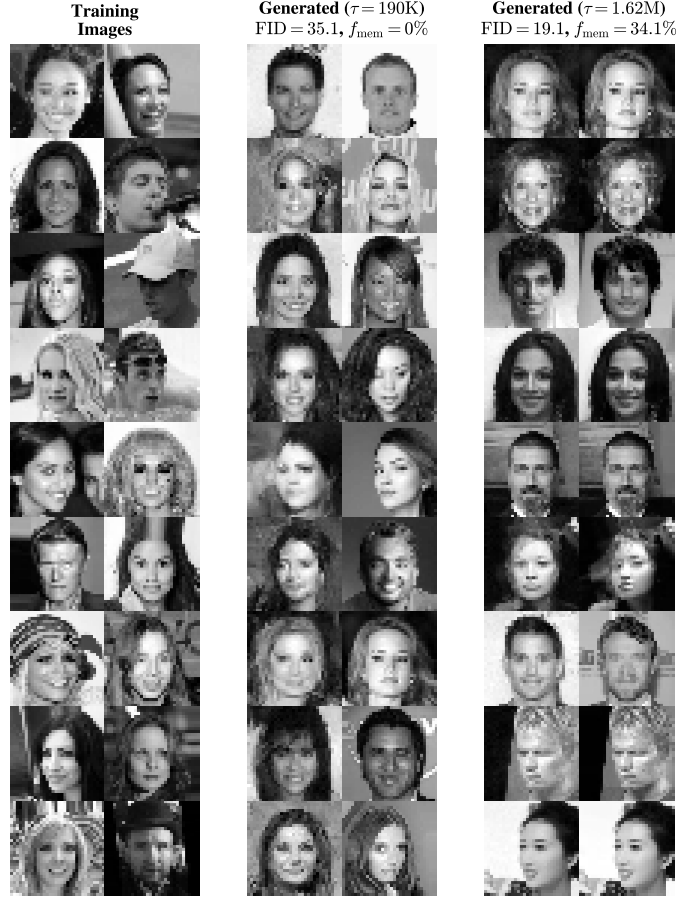


Figure 1: **Training and generation on CelebA.** The left-most block shows random training images. Middle and right blocks show generated samples in the left column (after  $\tau = 190K$  and  $\tau = 1.62M$  SGD updates respectively), alongside each sample’s nearest neighbor in the training set in the right column. All generated images come from model trained on  $n = 1024$  with base width  $W = 32$ .

**Statistical evaluation.** FIDs [9] are computed<sup>†</sup> 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  $\tau_{\text{mem}}$  is insensitive to the choice of the threshold  $k$  used to define  $f_{\text{mem}}$  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.

<sup>†</sup>Using the [pytorch-fid](#) Python package.

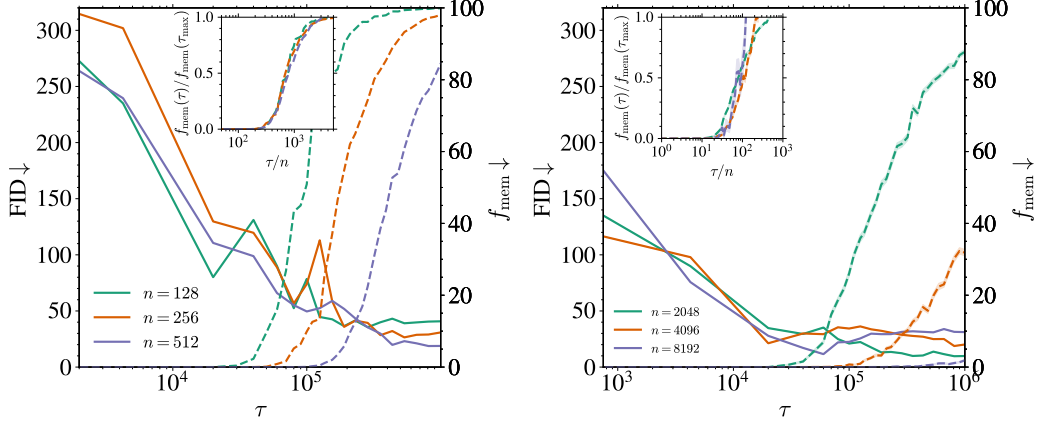


Figure 2: **Impact of batch size and optimizer on the scaling of  $\tau_{\text{mem}}$ .** FID (solid lines, left axis) and memorization fraction  $f_{\text{mem}}$  (in %, dashed lines, right axis) against training time  $\tau$  for various  $n$ . Inset: normalized memorization fraction  $f_{\text{mem}}(\tau)/f_{\text{mem}}(\tau_{\text{max}})$  with the rescaled time  $\tau/n$ . (Left) Memorization scaling for  $B = n$ . (Right) Generalization–Memorization transition with Adam optimizer for  $W = 64$ .

## 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  $\tau_{\text{mem}}$  cannot be explained by repetition over training samples. To validate this statement, the left panel of Fig. 2 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  $\tau$ , every sample has been seen exactly  $\tau$  times. Yet  $\tau_{\text{mem}}$  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 [14] 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. 2, 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  $\tau_{\text{gen}}$  and  $\tau_{\text{mem}}$  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  $\pm \mu$  with unit covariance, i.e.,

$$\mathbb{P}_0 = \frac{1}{2} \mathcal{N}(\mu, \mathbf{I}_d) + \frac{1}{2} \mathcal{N}(-\mu, \mathbf{I}_d). \quad (4)$$

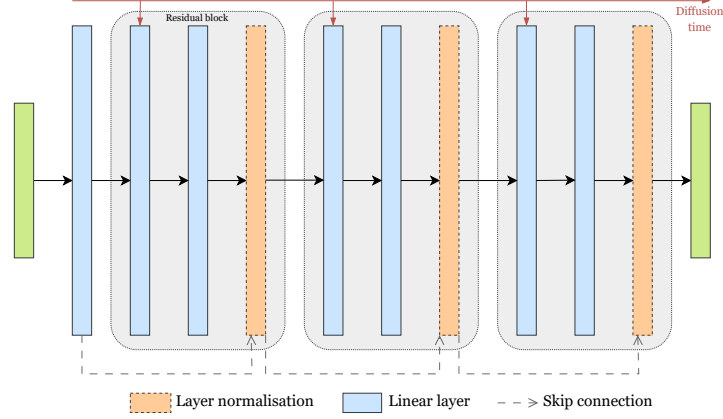


Figure 3: **Basic ResNet architecture of the GMM numerical experiments.** Residual network with three residual blocks, each made of two fully-connected layers followed by a layer normalization. The width of the network is  $W$ , and the input and output sizes are  $d$ .

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

$$\mathbf{s}_{\text{gen}}(\mathbf{x}_t, t) = \boldsymbol{\mu} e^{-t} \tanh(\mathbf{x}_t \cdot \boldsymbol{\mu} e^{-t}) - \mathbf{x}_t. \quad (5)$$

**Score model.** The denoise  $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$  is implemented as a lightweight residual multi-layer neural network (see Fig. 3): 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  $\mathbb{P}_{\boldsymbol{\theta}}$  via the Kullback-Leibler (KL) divergence

$$D_{\text{KL}}(\mathbb{P}_{\boldsymbol{\theta}}|\mathbb{P}_0) = \int d\mathbf{x} \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) \log \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) - \int d\mathbf{x} \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) \log \mathbb{P}_0(\mathbf{x}). \quad (6)$$

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

$$\int d\mathbf{x} \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) \log \mathbb{P}_0(\mathbf{x}) \approx \frac{1}{N} \sum_{\mu=1}^N \log \mathbb{P}_0(\tilde{\mathbf{x}}_{\mu}), \quad (7)$$

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

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}[\log \mathbb{P}_{\boldsymbol{\theta}}] \approx \int_0^T dt I(t) - \frac{d}{2} \log(2\pi e), \quad (8)$$

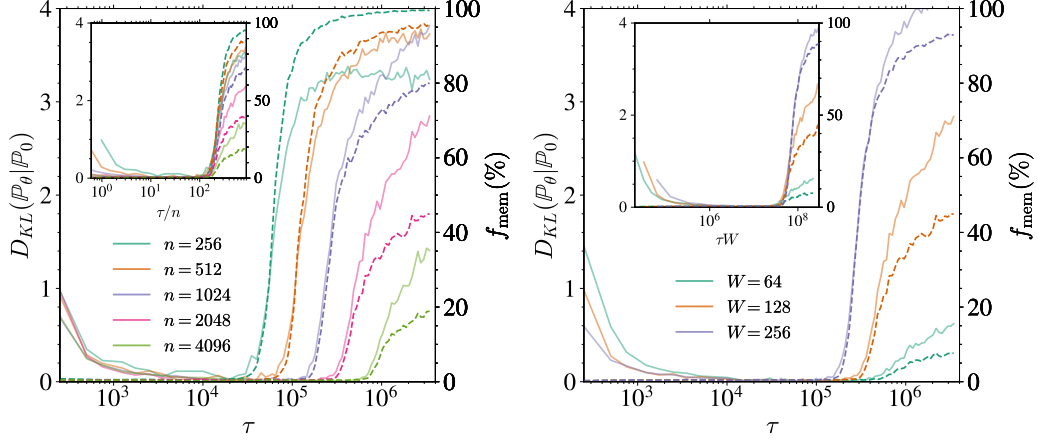


Figure 4: **Generalization–Memorization transition as a function of the training set size  $n$  and width  $W$  for ResNet score models on GMM ( $d = 8$ ).** KL divergences (solid lines, left axis) and memorization fraction  $f_{\text{mem}}$  (in %, dashed lines, right axis) against training time  $\tau$  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  $f_{\text{mem}}$  against the rescaled time  $\tau/n$  (left) and  $\tau W$  (right).

with

$$I(t) = \frac{\beta_t}{2N} \sum_{\mu=1}^N [\tilde{\mathbf{x}}_\mu \mathbf{s}_\theta(\tilde{\mathbf{x}}_\mu, t) + \mathbf{s}_\theta(\tilde{\mathbf{x}}_\mu, t)^2]. \quad (9)$$

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 $\tau_{\text{mem}}$ and $\tau_{\text{gen}}$ with $n$ and $W$

In Fig. 4, the left panel shows how the KL divergence and memorization fraction evolve with training time  $\tau$  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  $\tau_{\text{gen}}$  independent of  $n$  during which the model fully generalizes ( $f_{\text{mem}} = 0$ ). Beyond  $\tau_{\text{gen}}$ , both  $D_{\text{KL}}(\mathbb{P}_\theta|\mathbb{P}_0)$  and  $f_{\text{mem}}$  begin to rise at a time  $\tau_{\text{mem}}$  that scales linearly with  $n$ , as highlighted by the inset of the left panel. In contrast,  $\tau_{\text{mem}}$  scales with  $W^{-1}$ , as shown in the inset of the right panel. While the precise dependence of  $\tau_{\text{gen}}$  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 [11]. Although conditioning often improves sample quality in practice, memorization effects have also been observed in models trained conditionally [22, 27, 5]. 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  $\tau$  that we report in Fig. 5. 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,  $\tau_{\text{mem}}$  and  $\tau_{\text{gen}}$  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.

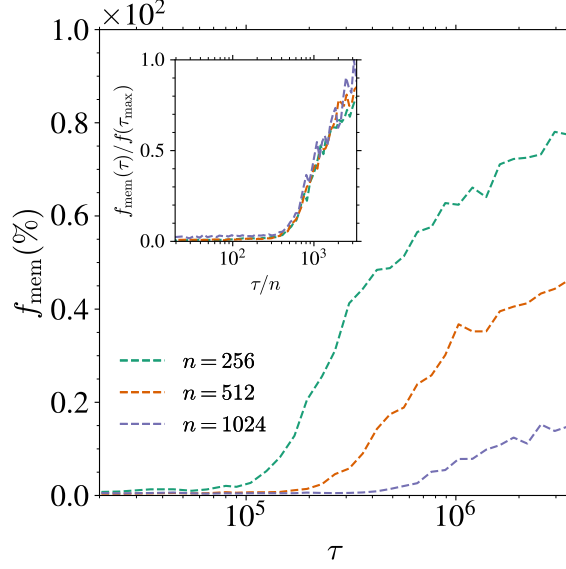


Figure 5: **Effect of guidance on  $\tau_{\text{mem}}$ .** Evolution of  $f_{\text{mem}}$  as a function of  $\tau$  for  $n \in \{256, 512, 1024\}$  at fixed  $W = 64$ .

## 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 [7], 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.

$$d : \text{Data dimension} \quad (10)$$

$$n : \text{Numbers of data points} \quad (11)$$

$$p : \text{Dimension of the hidden layers of the RFNN} \quad (12)$$

$$\mathbf{I}_d : \text{Identity matrix in dimension } d \quad (13)$$

$$\sigma(x) : \text{Activation function of the model} \quad (14)$$

$$P_{\mathbf{x}} : \text{Distribution of the data points} \quad (15)$$

$$P_t : \text{Distribution of the noisy data points at diffusion time } t. \quad (16)$$

$$\psi_n = \frac{n}{d} \quad (17)$$

$$\psi_p = \frac{p}{d} \quad (18)$$

$$\Delta_t = 1 - e^{-2t} \quad (19)$$

$$\Sigma = \mathbb{E}_{\mathbf{x} \sim P_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^T] \quad (20)$$

$$\Sigma_t = e^{-2t}\Sigma + \Delta_t\mathbf{I}_d \quad (21)$$

$$\Gamma_t^2 = \frac{\text{Tr}(\Sigma_t)}{d} \quad (22)$$

$$\sigma_{\mathbf{x}}^2 = \frac{\text{Tr}(\Sigma)}{d} \quad (23)$$

$$\|\sigma\|^2 = \mathbb{E}_z[\sigma(\Gamma_t z)^2] \quad (24)$$

$$b_t^2 = \left( \mathbb{E}_{u,v}[v\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)] \right)^2 \quad (25)$$

$$a_t = \mathbb{E}_{u,v}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v) \frac{u}{e^{-t}\sigma_{\mathbf{x}}}] \quad (26)$$

$$v_t^2 = \mathbb{E}_{u,v,w}[\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}v)\sigma(e^{-t}\sigma_{\mathbf{x}}u + \sqrt{\Delta_t}w)] - a_t^2 e^{-2t}\sigma_{\mathbf{x}}^2 \quad (27)$$

$$s_t^2 = \|\sigma\|^2 - a_t^2 e^{-2t}\sigma_{\mathbf{x}}^2 - v_t^2 - b_t^2 \quad (28)$$

$$\mu_1(t) = \mathbb{E}_u[\sigma(\Gamma_t u)u] = \sqrt{e^{-2t}\sigma_{\mathbf{x}}^2 a_t^2 + b_t^2}. \quad (29)$$

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

$$\mathbf{X} = [\mathbf{x}^1 | \dots | \mathbf{x}^n] \in \mathbb{R}^{d \times n} \quad (30)$$

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

$$\mathbf{W} = \begin{bmatrix} (\mathbf{W}_1)^\top \\ \vdots \\ (\mathbf{W}_p)^\top \end{bmatrix} \in \mathbb{R}^{p \times d}, \quad (31)$$

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  $\mathbf{V}$

$$\mathbf{U} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[ \sigma \left( \frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \sigma \left( \frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right)^\top \right], \quad (32)$$

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[ \sigma \left( \frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \xi^\top \right]. \quad (33)$$

## 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*

$$\frac{\mathbf{A}(\tau)}{\sqrt{p}} = -\frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} + \left( \frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} + \frac{\mathbf{A}_0}{\sqrt{p}} \right) e^{-\frac{2\Delta_t}{\psi_p} \mathbf{U} \tau} \quad (34)$$

with

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[ \sigma \left( \frac{\mathbf{W} \mathbf{x}_t^\nu(\xi)}{\sqrt{d}} \right) \xi^\top \right]. \quad (35)$$

*Proof.* We expand the square in the training loss

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr} \left( \frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U} \right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr} \left( \frac{\mathbf{A}}{\sqrt{p}} \mathbf{V} \right) \quad (36)$$

and compute the gradient

$$\nabla_{\mathbf{A}} \mathcal{L}_{\text{train}}(\mathbf{A}(\tau)) = \frac{2\Delta_t}{d} \frac{\mathbf{A}}{p} \mathbf{U} + \frac{2\sqrt{\Delta_t}}{d\sqrt{p}} \mathbf{V}^T. \quad (37)$$

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 [19, 8, 12], 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,

$$\sigma\left(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\right) \rightarrow \kappa_1 \frac{\mathbf{W}\mathbf{x}'}{\sqrt{d}} + \kappa_* \boldsymbol{\eta}, \quad \mathbf{x}' \sim \mathcal{N}(0, \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T]), \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_p), \quad (38)$$

where  $\kappa_1, \kappa_*$  are constants that depend on the distribution of the data and on the activation function  $\sigma$  whose formula we recall

$$\kappa_1 = \mathbb{E}_z[\sigma(\sigma_{\mathbf{x}} z) \frac{z}{\sigma_{\mathbf{x}}}], \quad (39)$$

$$\kappa_* = \sqrt{\mathbb{E}_z[\sigma(\sigma_{\mathbf{x}} z)^2] - \kappa_1^2 \sigma_{\mathbf{x}}^2}. \quad (40)$$

The expectation function are with respect to  $z \sim \mathcal{N}(0, 1)$  and  $\sigma_{\mathbf{x}}^2 = \text{Tr}(\boldsymbol{\Sigma})/d$ . 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  $\mathbf{x}_i$ ,

$$\mathbb{P}(|\mathbf{x}_i| \geq A) \leq 2e^{-A^2/C}. \quad (41)$$

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

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 [7], which considered only the case of data drawn from  $\mathcal{N}(0, \mathbf{I}_d)$ . A more rigorous approach, which would consist in following [16], is left for future works. We will make use of the Mehler kernel formula [13] which states that for  $f$  a test function defined on  $\mathbb{R}^2$ ,

$$\mathbb{E}_{u, v \sim P_{\gamma}}[f(u, v)] = \sum_{s=1}^{\infty} \frac{\gamma^s}{s!} \mathbb{E}_{u, v \sim \mathcal{N}(0, \mathbf{I}_2)}[He_s(u) He_s(v) f(u, v)], \quad (42)$$

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  $\gamma$ , while on the right-hand side the expectation is taken over independent standard Gaussian variables.  $He_s$  denotes the  $s$ -th Hermite polynomial. We recall some useful properties of the Hermite polynomials [1]:

- They form an orthogonal base of  $L^2(\mathbb{R}, \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx)$ .
- The first Hermite polynomials are  $He_0(x) = 1, He_1(x) = x$ .

**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  $\boldsymbol{\Sigma} = \mathbb{E}_{P_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^T]$ , the matrix*

$$\mathbf{U} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\boldsymbol{\xi}} \left[ \sigma\left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}\right) \sigma\left(\frac{\mathbf{W}\mathbf{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}\right)^T \right] \quad (43)$$

*has the same spectrum as its Gaussian equivalent*

$$\mathbf{U} = \frac{\mathbf{G}}{\sqrt{n}} \frac{\mathbf{G}^T}{\sqrt{n}} + b_t^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + s_t^2 \mathbf{I}_p \quad (44)$$

where

$$\mathbf{G} = e^{-t} a_t \frac{\mathbf{W}}{\sqrt{d}} \mathbf{X}' + v_t \boldsymbol{\Omega}, \quad (45)$$

$\mathbf{X}' \in \mathbb{R}^{d \times n}$  is a matrix whose columns  $\mathbf{x}'^\nu$  are sampled according to  $\mathcal{N}(0, \boldsymbol{\Sigma})$  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  $\Sigma$  appear by writing the data points are written as  $\mathbf{x}^\nu = \Sigma^{1/2} \mathbf{z}^\nu$  where the vectors  $\mathbf{z}^\nu$  have variance 1. Let us focus on the element of  $\mathbf{U}$  in position  $(i, j)$

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[ \sigma \left( \frac{\mathbf{W}_{ik} (e^{-t} (\Sigma^{1/2})_{kl} \mathbf{z}_l^\nu + \sqrt{\Delta_t} \xi_k)}{\sqrt{d}} \right) \sigma \left( \frac{\mathbf{W}_{jk'} (e^{-t} (\Sigma^{1/2})_{k'l'} \mathbf{z}_{l'}^\nu + \sqrt{\Delta_t} \xi_{k'})}{\sqrt{d}} \right) \right], \quad (46)$$

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

$$\mathbf{U}_{ii} = \mathbb{E}_{\chi} [\sigma(\chi)^2] + \mathcal{O}(1/n). \quad (47)$$

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

$$\mathbf{U}_{ii} = \mathbb{E}_{\chi} [\sigma(\chi)^2] + \mathcal{O}(1/n) = \mathbb{E}_{u \sim \mathcal{N}(0,1)} [\sigma(\Gamma_t u)^2] = \|\sigma\|^2. \quad (48)$$

If  $i \neq j$ , we denote  $\eta_i^\nu = e^{-t} \frac{\mathbf{W}_i^T \Sigma^{1/2} \mathbf{z}}{\sqrt{d}}$ . For now we consider  $\mathbf{W}$  and the  $\mathbf{z}^\nu$  fixed and look at  $\xi$ . 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  $\mathbb{E}_{\xi} [uv] = \frac{\mathbf{W}_i^T \mathbf{W}_j}{d}$

$$\mathbb{E}_{u,v} [\sigma(\eta_i^\nu + \sqrt{\Delta_t} u) \sigma(\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (49)$$

$$= \sum_{s=1}^{\infty} \frac{(\mathbf{W}_i^T \mathbf{W}_j / d)^s}{s!} \mathbb{E}_u [H e_s(u) \sigma(\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [H e_s(v) \sigma(\eta_j^\nu + \sqrt{\Delta_t} v)]. \quad (50)$$

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

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_u [\sigma(\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [\sigma(\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (51)$$

$$+ \frac{1}{n} \sum_{\nu=1}^n \frac{\mathbf{W}_i^T \mathbf{W}_j}{d} \mathbb{E}_u [u \sigma(\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [v \sigma(\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (52)$$

$$= \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_u [\sigma(\eta_i^\nu + \sqrt{\Delta_t} u)] \mathbb{E}_v [\sigma(\eta_j^\nu + \sqrt{\Delta_t} v)] \quad (53)$$

$$+ \frac{\mathbf{W}_i^T \mathbf{W}_j}{d} \mathbb{E}_{\eta} [\mathbb{E}_u [u \sigma(\eta_i + \sqrt{\Delta_t} u)] \mathbb{E}_v [v \sigma(\eta_j + \sqrt{\Delta_t} v)]]. \quad (54)$$

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

$$b_t^2 = (\mathbb{E}_{u,v} [v \sigma(e^{-t} \sigma_{\mathbf{x}} u + \sqrt{\Delta_t} v)])^2. \quad (55)$$

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

$$\frac{1}{n} \sum_{\nu} \sigma_0(\eta_i^\nu) \sigma_0(\eta_j^\nu). \quad (56)$$

We use the GEP on  $\sigma_0$

$$\sigma_0 \left( \frac{e^{-t} \mathbf{W}_i^T \mathbf{x}^\nu}{\sqrt{d}} \right) \rightarrow a_t e^{-t} \frac{\mathbf{W}_i^T \mathbf{x}^\nu}{\sqrt{d}} + v_t \Omega_i^\nu, \quad \mathbf{x}^\nu \sim \mathcal{N}(0, \Sigma), \quad \Omega_i^\nu \sim \mathcal{N}(0, \mathbf{I}_p), \quad (57)$$

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  $i \neq j$

$$\mathbf{U}_{ij} = \frac{1}{n} \sum_{\nu=1}^n \left( a_t e^{-t} \frac{\mathbf{W}_i^T \mathbf{x}^{\nu}}{\sqrt{d}} + v_t \Omega_i^{\nu} \right) \left( a_t e^{-t} \frac{\mathbf{W}_j^T \mathbf{x}^{\nu}}{\sqrt{d}} + v_t \Omega_j^{\nu} \right)^T + b_t^2 \frac{\mathbf{W}_i^T \mathbf{W}_j}{d}. \quad (58)$$

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

$$(a_t^2 e^{-2t} \sigma_{\mathbf{x}}^2 + v_t^2 + b_t^2) \mathbf{I}_p. \quad (59)$$

The Gaussian equivalent of  $\mathbf{U}$  reads

$$\mathbf{U} = \frac{\mathbf{G} \mathbf{G}^T}{\sqrt{n} \sqrt{n}} + b_t^2 \frac{\mathbf{W} \mathbf{W}^T}{d} + s_t^2 \mathbf{I}_p, \quad (60)$$

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  $\tilde{\mathbf{U}}$ ). *Let*

$$\tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{y}}[\sigma(\frac{\mathbf{W}\mathbf{y}}{\sqrt{d}})\sigma(\frac{\mathbf{W}\mathbf{y}}{\sqrt{d}})^T], \quad (61)$$

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

$$\mu_1^2(t) \frac{\mathbf{W} \Sigma_t \mathbf{W}^T}{d} + (\|\sigma\|^2 - \mu_1^2(t)) \mathbf{I}_p, \quad (62)$$

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  $\mathbb{E}_{\mathbf{y}}[\frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}} \frac{\mathbf{W}_j^T \mathbf{y}}{\sqrt{d}}] = \frac{\mathbf{W}_i^T \Sigma_t \mathbf{W}_j}{d} \sim \Gamma_t^2$  and correlation  $\mathbb{E}_{\mathbf{y}}[\frac{\mathbf{W}_i^T \mathbf{y}}{\sqrt{d}} \frac{\mathbf{W}_j^T \mathbf{y}}{\sqrt{d}}] = \frac{\mathbf{W}_i^T \Sigma_t \mathbf{W}_j}{d}$ . We apply Mehler Kernel formula to  $\tilde{\mathbf{U}}$

$$\tilde{\mathbf{U}}_{ij} = \sum_s \frac{1}{s!} \left( \frac{\mathbf{W}_{ik}(\Sigma_t)_{kl} \mathbf{W}_{jl}}{\Gamma_t^2 d} \right)^s \mathbb{E}_u[\sigma(\Gamma_t u) He_s(u)] \mathbb{E}_v[\sigma(\Gamma_t v) He_s(v)], \quad (63)$$

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  $s = 1$ .

$$\tilde{\mathbf{U}}_{ij} = \left( \frac{\mathbf{W}_{ik}(\Sigma_t)_{kl} \mathbf{W}_{jl}}{\Gamma_t^2 d} \right) \mathbb{E}_u[\sigma(\Gamma_t u) u]^2. \quad (64)$$

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

$$\tilde{\mathbf{U}}_{ii} = \mathbb{E}_{u \sim \mathcal{N}(0,1)}[\sigma^2(\Gamma_t u)] = \|\sigma\|^2. \quad (65)$$

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

$$\tilde{\mathbf{U}} = \frac{\mu_1^2(t)}{\Gamma_t^2} \frac{\mathbf{W} \Sigma_t \mathbf{W}^T}{d} + (\|\sigma\|^2 - \mu_1^2(t)) \mathbf{I}_p \quad (66)$$

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  $\Sigma$  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*

$$\left(1 - \frac{1}{\psi_p}\right) \delta(\lambda - (\|\sigma\|^2 - \mu_1^2(t))) + \frac{1}{\psi_p} \rho_{\text{bulk}}(\lambda), \quad (67)$$

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

$$\lambda_{\max}\left(\frac{\Sigma_t^{1/2} \mathbf{W} \mathbf{W}^T \Sigma_t^{1/2}}{d}\right) \leq \lambda_{\max}\left(\frac{\mathbf{W} \mathbf{W}^T}{d}\right) \lambda_{\max}(\Sigma_t) \lesssim \mathcal{O}(\psi_p), \quad (68)$$

since we assumed that  $\lambda_{\max}(\Sigma_t) = e^{-2t} \lambda_{\max}(\Sigma_t) + \Delta_t = \mathcal{O}(1)$  and since  $\lambda_{\max}(\frac{\mathbf{W} \mathbf{W}^T}{d}) = \mathcal{O}(\psi_p)$  is given by the Marchenko-Pastur law [18]. To bound the norm from below we use the following inequality

$$\lambda_{\min}(\Sigma_t) \lambda_{\min}\left(\frac{\mathbf{W} \mathbf{W}^T}{d}\right) \leq \lambda_{\min}\left(\frac{\Sigma_t^{1/2} \mathbf{W} \mathbf{W}^T \Sigma_t^{1/2}}{d}\right). \quad (69)$$

Since  $\Sigma_t$  is positive definite, the bound is also of order  $\psi_p$ . This concludes that the support of the bulk is of order  $\psi_p$ .  $\square$

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

$$\mathbf{V} = \frac{1}{n} \sum_{\nu=1}^n \mathbb{E}_{\xi} \left[ \sigma \left( \frac{\mathbf{W} \mathbf{x}_t^{\nu}(\xi)}{\sqrt{d}} \right) \xi^T \right], \quad (70)$$

$$\tilde{\mathbf{V}} = \mathbb{E}_{\mathbf{x}, \xi} \left[ \sigma \left( \frac{\mathbf{W} \mathbf{x}_t(\xi)}{\sqrt{d}} \right) \xi^T \right]. \quad (71)$$

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

$$\tilde{\mathbf{V}} = \mathbf{V} = \frac{\mu_1(t) \sqrt{\Delta_t}}{\Gamma_t} \frac{\mathbf{W}}{\sqrt{d}}. \quad (72)$$

*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  $\eta_i = \frac{\mathbf{W}_{ik}(e^{-t} \mathbf{x}_k + \sqrt{\Delta_t} \xi_k)}{\sqrt{d}}$ . Its has 0 mean, covariance  $\mathbb{E}_{\mathbf{x}, \xi}[\eta_i^2] = \frac{\mathbf{W}_i^T \Sigma_t \mathbf{W}_i}{d} \sim \Gamma_t^2$  and correlation with  $\xi$   $\gamma_{ij} = \mathbb{E}_{\mathbf{x}, \xi}[\eta_i \xi_j] = \frac{\sqrt{\Delta_t} \mathbf{W}_{ij}}{\sqrt{d}}$ . We apply the Mehler Kernel formula

$$\tilde{\mathbf{V}}_{ij} = \mathbb{E}_{\mathbf{x}, \xi} \left[ \sigma \left( \Gamma_t \left( \frac{\mathbf{W}_{ik}(e^{-t}(\Sigma_t)_{kl} \mathbf{z}_l + \sqrt{\Delta_t} \xi_l)}{\Gamma_t \sqrt{d}} \right) \right) \xi_j \right] \quad (73)$$

$$= \sum_s \frac{1}{s!} \left( \frac{\mathbf{W}_{ij} \sqrt{\Delta_t}}{\Gamma_t \sqrt{d}} \right)^s \mathbb{E}_u [\sigma(\Gamma_t u) H e_s(u)] \mathbb{E}_v [v H e_s(v)] \quad (74)$$

$$= 0 + \frac{\sqrt{\Delta_t}}{\Gamma_t} \frac{\mathbf{W}_{ij}}{\sqrt{d}} \mathbb{E}_u [\sigma(\Gamma_t u) u] \mathbb{E}_v [v^2] + \mathcal{O}\left(\frac{1}{d}\right) \quad (75)$$

$$= \frac{\sqrt{\Delta_t} \mu_1(t)}{\Gamma_t} \frac{\mathbf{W}_{ij}}{\sqrt{d}}. \quad (76)$$

$\square$

#### C.4 Proof of Theorem 3.1.

We recall the Theorem 3.1 of the MT.

**Theorem.** 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

$$\hat{s}(q) = b_t^2 \psi_p + \frac{1}{q}, \quad (77)$$

$$\hat{r}(r, q) = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q}. \quad (78)$$

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

$$s = \int d\rho_\Sigma(\lambda) \frac{1}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (79)$$

$$r = \int d\rho_\Sigma(\lambda) \frac{\lambda}{\hat{s}(q) + \lambda \hat{r}(r, q)}, \quad (80)$$

$$\psi_p(s_t^2 - z) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} \psi_p}{\psi_n} r + \frac{\psi_p v_t^2}{\psi_n} q} + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0, \quad (81)$$

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  $\rho$  found in Ref. [7] with linear pencils [3] in the case  $P_{\mathbf{x}} = \mathcal{N}(0, \mathbf{I}_d)$  i.e.  $\rho_\Sigma(\lambda) = \delta(\lambda - 1)$  can be reduced to the equations of Theorem ?? with our definitions of  $\mu_1(t)$ ,  $s_t$  and  $v_t$ . The equations of [7] read

$$\zeta_1(s_t^2 - z + e^{-2t} \mu_1^2 \zeta_2 \zeta_3 + v_t^2 \zeta_2 + \Delta_t \mu_1^2 \zeta_4) - 1 = 0 \quad (82)$$

$$\zeta_2(\psi_n + v_t^2 \psi_p \zeta_1 - e^{-t} \mu_1 \zeta_3) - \psi_n = 0 \quad (83)$$

$$e^{-t} \mu_1 \psi_p \zeta_1 (1 + e^{-t} \mu_1 \zeta_2 \zeta_3) + (1 + (\Delta_t \mu_1^2 \psi_p \zeta_1) \zeta_3) = 0 \quad (84)$$

$$e^{-2t} \mu_1^2 \psi_p \zeta_1 \zeta_2 \zeta_4 + (1 + \Delta_t \mu_1^2 \psi_p \zeta_1) \zeta_4 - 1 = 0, \quad (85)$$

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  $\zeta_2$  to  $q$  and  $r$

$$\zeta_2 = \frac{1}{1 + \frac{e^{-2t} \mu_1^2 \psi_p}{\psi_n} r + \frac{v_t^2 \psi_p}{\psi_n} q}. \quad (86)$$

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

$$\zeta_4 = \frac{1}{1 + \mu_1^2 \psi_p q (\Delta_t + e^{-2t} \zeta_2)}. \quad (87)$$

Injecting this into the first equation gives

$$q(s_t^2 - z + e^{-2t} \mu_1^2 \zeta_2 r (-e^{-t} \mu_1 \psi_p) + v_t^2 \zeta_2 + \Delta_t \mu_1^2 \frac{1}{1 + \mu_1^2 \psi_p q (\Delta_t + e^{-2t} \zeta_2)}) - 1 = 0. \quad (88)$$

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. [6].

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

$$q = \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\text{Tr}(\mathbf{U} - z\mathbf{I}_p)^{-1}] \quad (89)$$

$$= -\partial_z \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\log \det(\mathbf{U} - z\mathbf{I}_p)] \quad (90)$$

$$= 2\partial_z \lim_{p \rightarrow \infty} \frac{1}{p} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \Omega} [\log \det(\mathbf{U} - z\mathbf{I}_p)^{-1/2}]. \quad (91)$$

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

$$q = 2\partial_z \lim_{s \rightarrow 0} \lim_{p \rightarrow \infty} \frac{1}{ps} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}} [\det(\mathbf{U} - z\mathbf{I}_p)^{-s/2} - 1], \quad (92)$$

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

$$\mathcal{Z} = \det(\mathbf{U} - z\mathbf{I}_p)^{-1/2} = \int \frac{d\phi}{(2\pi)^{p/2}} e^{-\frac{1}{2}\phi^T(\mathbf{U} - z\mathbf{I}_p)\phi}. \quad (93)$$

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

$$\mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}} [\mathcal{Z}^s] = \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}} [e^{-\frac{1}{2}\phi^{aT}(\mathbf{U} - z\mathbf{I}_p)\phi^a}] \quad (94)$$

$$= \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} e^{\frac{1}{2}\phi^{aT}(z - s_t^2)\phi^a} \mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}} [e^{-\frac{1}{2n}\phi^{aT} \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right) \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right)^T \phi^a} e^{-\frac{b_t^2}{2d}\phi^{aT} \mathbf{W}\mathbf{W}^T \phi^a}]. \quad (95)$$

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  $\mathbf{\Omega}$ , we need the following standard result from Gaussian integration

$$\int d\mathbf{x} e^{-\frac{1}{2}\mathbf{x}\mathbf{G}\mathbf{x}^T + \mathbf{J}\mathbf{x}^T} = e^{-\frac{1}{2} \log \det \mathbf{G} + \frac{1}{2} \mathbf{J}\mathbf{G}^{-1}\mathbf{J}^T}, \quad (96)$$

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

**Averaging over the data set.** The dataset dependence enters through

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT} \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right) \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right)^T \phi^a}] \\ &= \mathbb{E}_{\mathbf{X}} [e^{-\frac{a_t^2 e^{-2t}}{2nd}\phi^{aT} \mathbf{W}\mathbf{X}^\nu \mathbf{X}^{\nu T} \mathbf{W}^T \phi^a} e^{-\frac{a_t e^{-t} v_t}{2\sqrt{dn}}\phi^{aT} (\mathbf{W}\mathbf{X}^\nu \mathbf{\Omega}^T + \mathbf{\Omega} \mathbf{X}^{\nu T} \mathbf{W}^T) \phi^a} e^{-\frac{v_t^2}{2n}\phi^{aT} \mathbf{\Omega} \mathbf{\Omega}^T \phi^a}]. \end{aligned} \quad (97)$$

We introduce for each replica  $\phi^a$  a Fourier transform of the delta function by using the auxiliary variables  $\omega^a, \hat{\omega}^a \in \mathbb{R}^d$  as<sup>†</sup>

$$\int d\omega^a d\hat{\omega}^a e^{i\hat{\omega}^a (\sqrt{p}\omega^a - \phi^{aT} \mathbf{W}\Sigma^{1/2})} = 1. \quad (98)$$

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

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT} \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right) \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right)^T \phi^a}] \\ &= \mathbb{E}_{\mathbf{Z}} [e^{-\frac{a_t^2 e^{-2t} p}{2nd}\omega^{aT} \mathbf{Z}^\nu \mathbf{Z}^{\nu T} \omega^a} e^{-\frac{a_t e^{-t} v_t \sqrt{p}}{\sqrt{dn}} \sum_{a,\nu} \mathbf{\Omega}^\nu \phi^a \omega^a \cdot \mathbf{Z}^\nu} e^{-\frac{v_t^2}{2n}\phi^{aT} \mathbf{\Omega} \mathbf{\Omega}^T \phi^a}]. \end{aligned} \quad (99)$$

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

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}} [e^{-\frac{1}{2n}\phi^{aT} \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right) \left( a_t e^{-t} \frac{\mathbf{W}\mathbf{X}^\nu}{\sqrt{d}} + v_t \mathbf{\Omega}^\nu \right)^T \phi^a}] = \\ & e^{-\frac{n}{2} \log \det(1 + \mathbf{G}_{\mathbf{Z}})} e^{\frac{a_t e^{-t} v_t p}{2dn^2} \sum_\nu (\mathbf{\Omega}^\nu \cdot \phi^a) (\mathbf{\Omega}^\nu \cdot \phi^b) \omega^a \cdot \omega^b (1 + \mathbf{G}_{\mathbf{Z}})^{-1}_{k,l} \omega^b_l} e^{-\frac{v_t^2}{2n}\phi^{aT} \mathbf{\Omega} \mathbf{\Omega}^T \phi^a}, \end{aligned} \quad (100)$$

where repeated indices mean that there is an implicit summation.

<sup>†</sup>Throughout the computation, we discard non-exponential prefactors, as they give subleading contributions.

**Averaging over  $\Omega$ .** The terms that depend on  $\Omega$  are

$$\begin{aligned} \mathbb{E}_{\Omega} & \left[ e^{\frac{a_t e^{-t} v_t}{2dn^2} \sum_{\nu} (\Omega^{\nu} \cdot \phi^a)(\Omega^{\nu} \cdot \phi^b) \omega^a_k (1 + \mathbf{G}_{\mathbf{x}})^{-1}_{k,l} \omega^b_l e^{-\frac{v_t^2}{2n} \phi^a T \Omega \Omega^T \phi^a}} \right] \\ &= (\mathbb{E}_{\Omega^{\nu}} [e^{\frac{a_t e^{-t} v_t p}{2dn^2} (\Omega^{\nu} \cdot \phi^a)(\Omega^{\nu} \cdot \phi^b) \omega^a_k (1 + \mathbf{G}_{\mathbf{x}})^{-1}_{k,l} \omega^b_l e^{-\frac{v_t^2}{2n} \phi^a T \Omega^{\nu} \Omega^{\nu T} \phi^a}}])^n \end{aligned} \quad (101)$$

$$= e^{-\frac{n}{2} \log \det(1 + \mathbf{G}_{\Omega})}, \quad (102)$$

with

$$(\mathbf{G}_{\Omega})_{k,l} = \phi^a \left( \frac{v_t^2}{n} \delta_{ab} - \frac{a_t e^{-t} v_t p}{dn^2} \omega^a_k (1 + \mathbf{G}_{\mathbf{x}})^{-1}_{k,l} \omega^b_l \right) \phi^b T. \quad (103)$$

We are left with

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{x}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s \frac{d\phi^a}{(2\pi)^{p/2}} d\omega^a d\hat{\omega}^a e^{\frac{1}{2}(z-s_t^2) \phi^a \phi^a T} e^{-\frac{b_t^2 p}{2d} \omega^a T \Sigma^{-1} \omega^a} \\ \mathbb{E}_{\mathbf{W}} & [e^{i\hat{\omega}^a (\sqrt{p} \omega^a - \phi^a T \mathbf{W} \Sigma^{1/2})} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})}]. \end{aligned} \quad (104)$$

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

$$\mathbb{E}_{\mathbf{W}} [e^{i \sum_a \hat{\omega}^a (\sqrt{p} \omega^a - \mathbf{W}^T \phi^a \Sigma^{1/2})}] = e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} (\mathbb{E}_{\mathbf{W}} [e^{-i\hat{\omega}^a \phi^a_i \mathbf{W}_{li} (\Sigma^{1/2})_{kl}}]) \quad (105)$$

$$= e^{i\sqrt{p} \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \hat{\omega}^a_k (\Sigma)_{kl} \hat{\omega}^b_l \phi^a_i \phi^b_i} \quad (106)$$

$$= e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b}. \quad (107)$$

We end up with

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{x}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\hat{\omega}^a e^{\frac{1}{2}(z-s_t^2) \phi^a \phi^a T} e^{-\frac{b_t^2 p}{2d} \omega^a T \Sigma^{-1} \omega^a} e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} \\ & e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})}. \end{aligned} \quad (108)$$

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

$$\int \prod_a d\hat{\omega}^a e^{i\sqrt{p} \sum_a \hat{\omega}^a \cdot \omega^a} e^{-\frac{1}{2} \sum_{a,b} \hat{\omega}^a \Sigma \hat{\omega}^b \phi^a \cdot \phi^b}. \quad (109)$$

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

$$\int \prod_a d\hat{\omega}^a e^{\sum_{i,a} \mathbf{J}_i^a \hat{\omega}_i^a} e^{-\frac{1}{2} \sum_{i,j,a,b} \hat{\omega}_i^a \mathbf{G}_{ij}^{ab} \hat{\omega}_j^b} = e^{-\frac{1}{2} \log \det(\mathbf{G}) + \frac{1}{2} \mathbf{J}^T \mathbf{G}^{-1} \mathbf{J}}. \quad (110)$$

This gives

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{x}, \Omega} [\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a e^{\frac{1}{2}(z-s_t^2) \phi^a \phi^a T} e^{-\frac{b_t^2 p}{2d} \omega^a T \Sigma^{-1} \omega^a} e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_Z)} \\ & e^{-\frac{n}{2} \log \det(\mathbf{I}_d + \mathbf{G}_{\Omega})} e^{-\frac{1}{2} \log \det(\mathbf{G}) + \frac{1}{2} \mathbf{J}^T \mathbf{G}^{-1} \mathbf{J}}. \end{aligned} \quad (111)$$

**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

$$1 = \int d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} e^{\frac{1}{2} \hat{\mathbf{Q}}^{ab} (p\mathbf{Q}^{ab} - \phi^a \cdot \phi^b)}, \quad (112)$$

$$1 = \int d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} e^{\frac{1}{2} \hat{\mathbf{R}}^{ab} (d\mathbf{R}^{ab} - \omega^a \cdot \omega^b)}, \quad (113)$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, \mathbf{Y}, \mathbf{\Omega}}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} \\
&\quad e^{\frac{1}{2}\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi^a\cdot\phi^b)} e^{\frac{1}{2}\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega^a\cdot\omega^b)} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \text{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2g} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2d} \omega^a T \mathbf{\Sigma}^{-1} \omega^a} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n}(v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(1 + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\mathbf{\Sigma} \otimes \mathbf{Q})} e^{-\frac{1}{2} \omega_k^a \mathbf{\Sigma}_{kl}^{-1} (\mathbf{Q}^{-1})_{ab} \omega_l^b}.
\end{aligned} \tag{114}$$

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

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\phi^a d\omega^a d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} d\mathbf{S}^{ab} d\hat{\mathbf{S}}^{ab} \\
&\quad e^{\frac{1}{2}\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi^a\cdot\phi^b)} e^{\frac{1}{2}\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega^a\cdot\omega^b)} e^{\frac{1}{2}\hat{\mathbf{S}}^{ab}(d\mathbf{S}^{ab}-\omega^a\mathbf{\Sigma}^{-1}\omega^b)} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \text{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2} \text{Tr}(\mathbf{S})} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n}(v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(1 + \frac{a_t^2 v_t^2 p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\mathbf{\Sigma} \otimes \mathbf{Q})} e^{-\frac{d}{2} \text{Tr}(\mathbf{S} \mathbf{Q}^{-1})}.
\end{aligned} \tag{115}$$

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

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}}[\mathcal{Z}^s] &= \int \prod_{a=1}^s d\mathbf{Q}^{ab} d\hat{\mathbf{Q}}^{ab} d\mathbf{R}^{ab} d\hat{\mathbf{R}}^{ab} d\mathbf{S}^{ab} d\hat{\mathbf{S}}^{ab} \\
&\quad e^{\frac{p}{2} \text{Tr}(\hat{\mathbf{Q}} \mathbf{Q})} e^{-\frac{p}{2} \log \det \hat{\mathbf{Q}}} e^{\frac{d}{2} \hat{\mathbf{R}}^{ab} \mathbf{R}^{ab}} e^{\frac{d}{2} \hat{\mathbf{S}}^{ab} \mathbf{S}^{ab}} \\
&\quad e^{-\frac{1}{2} \log \det(\hat{\mathbf{R}} \otimes \mathbf{I}_d + \hat{\mathbf{S}} \otimes \mathbf{\Sigma}^{-1})} \\
&\quad e^{\frac{p}{2}(z-s_t^2) \text{Tr} \mathbf{Q}} e^{-\frac{n}{2} \log \det(\mathbf{I}_m + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})} e^{-\frac{b_t^2 p}{2} \text{Tr}(\mathbf{S})} \\
&\quad e^{-\frac{n}{2} \log(1 + \frac{p}{n}(v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(1 + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1}) \mathbf{Q})} \\
&\quad e^{-\frac{1}{2} \log \det(\mathbf{\Sigma} \otimes \mathbf{Q})} e^{-\frac{d}{2} \text{Tr}(\mathbf{S} \mathbf{Q}^{-1})}.
\end{aligned} \tag{116}$$

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

$$e^{-\frac{1}{2} \log \det(\mathbf{\Sigma} \otimes \mathbf{Q})} e^{-\frac{1}{2} \log \det(\hat{\mathbf{R}} \otimes \mathbf{I}_d + \hat{\mathbf{S}} \otimes \mathbf{\Sigma}^{-1})} = e^{-\frac{1}{2} \log \det(\mathbf{Q} \hat{\mathbf{S}} \otimes \mathbf{I}_d + \mathbf{Q} \hat{\mathbf{R}} \otimes \mathbf{\Sigma})} \tag{117}$$

$$= e^{-\frac{d}{2} \log \det(\mathbf{Q} \hat{\mathbf{S}})} e^{-\frac{1}{2} \log \det(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \mathbf{\Sigma})} \tag{118}$$

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

$$-\frac{1}{2} \log \det(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \mathbf{\Sigma}) = -\frac{1}{2} \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1} \otimes \mathbf{\Sigma}) \tag{119}$$

$$= -\frac{1}{2} \sum_{l \geq 0} \frac{(-1)^l}{l!} (\hat{\mathbf{R}} \hat{\mathbf{S}}^{-1})^l \otimes \mathbf{\Sigma}^l \tag{120}$$

$$= -\frac{d}{2} \int d\lambda \rho_{\mathbf{\Sigma}}(\lambda) \sum_{l \geq 0} \frac{(-1)^l}{l!} \text{Tr}((\hat{\mathbf{R}} \hat{\mathbf{S}}^{-1})^l) \lambda^l \tag{121}$$

$$= -\frac{d}{2} \int d\lambda \rho_{\mathbf{\Sigma}}(\lambda) \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}} \hat{\mathbf{S}}^{-1}). \tag{122}$$

We end up with

$$\mathbb{E}_{\mathbf{W}, \mathbf{X}, \mathbf{\Omega}}[\mathcal{Z}^m] = \int d\mathbf{Q} d\hat{\mathbf{Q}} d\mathbf{R} d\hat{\mathbf{R}} d\mathbf{S} d\hat{\mathbf{S}} e^{-\frac{d}{2} S(\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}})}, \tag{123}$$



where the action reads

$$\begin{aligned}
S(\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}}) &= \psi_p \log \det \hat{\mathbf{Q}} - \psi_p \text{Tr}(\mathbf{Q}\hat{\mathbf{Q}}) - \text{Tr}(\mathbf{R}\hat{\mathbf{R}}) - \text{Tr}(\mathbf{S}\hat{\mathbf{S}}) \\
&- \psi_p(z - s_t^2) \text{Tr} \mathbf{Q} + \psi_n \log \det(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R}) + b_t^2 \psi_p \text{Tr} \mathbf{S} \\
&+ \psi_n \log(\mathbf{I}_s + \frac{p}{n}(v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1})\mathbf{Q}) \\
&+ \log \det(\mathbf{Q}\hat{\mathbf{S}}) + \int d\lambda \rho_{\Sigma}(\lambda) \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}) + \text{Tr}(\mathbf{S}\mathbf{Q}^{-1}).
\end{aligned} \tag{124}$$

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

$$\hat{\mathbf{Q}}^{-1} = \mathbf{Q}, \tag{125}$$

which yields

$$\begin{aligned}
S(\mathbf{Q}, \mathbf{R}, \hat{\mathbf{R}}, \mathbf{S}, \hat{\mathbf{S}}) &= -\psi_p \log \det \mathbf{Q} - \text{Tr}(\mathbf{R}\hat{\mathbf{R}}) - \text{Tr}(\mathbf{S}\hat{\mathbf{S}}) \\
&- \psi_p(z - s_t^2) \text{Tr} \mathbf{Q} + \psi_n \log \det(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R}) + b_t^2 \psi_p \text{Tr} \mathbf{S} \\
&+ \psi_n \log(\mathbf{I}_s + \frac{p}{n}(v_t^2 - \frac{a_t^2 e^{-2t} v_t^2}{n} \mathbf{R}(\mathbf{I}_s + \frac{a_t^2 e^{-2t} p}{n} \mathbf{R})^{-1})\mathbf{Q}) \\
&+ \log \det(\mathbf{Q}\hat{\mathbf{S}}) + \int d\lambda \rho_{\Sigma}(\lambda) \text{Tr} \log(\mathbf{I}_m \otimes \mathbf{I}_d + \lambda \hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}) \\
&+ \text{Tr}(\mathbf{S}\mathbf{Q}^{-1}).
\end{aligned} \tag{126}$$

**Replica Symmetric Ansatz.** We introduce a replica symmetric (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

$$\begin{aligned}
S(q, r, \hat{r}, s, \hat{s})/s &= -\psi_p \log q - r\hat{r} - s\hat{s} \\
&- \psi_p(z - s_t^2)q + \psi_n \log(1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q) + b_t^2 \psi_p s \\
&+ \log(q) + \int d\lambda \rho_{\Sigma}(\lambda) \log(\hat{s} + \lambda \hat{r}) + \frac{s}{q}.
\end{aligned} \tag{127}$$

Let us differentiate with respect to the 5 variables

$$\frac{\partial S}{\partial s} = -\hat{s} + b_t^2 \psi_p + \frac{1}{q}, \tag{128}$$

$$\frac{\partial S}{\partial r} = -\hat{r} + \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q}, \tag{129}$$

$$\frac{\partial S}{\partial \hat{s}} = -s + \int d\lambda \rho_{\Sigma}(\lambda) \frac{1}{\hat{s} + \lambda \hat{r}}, \tag{130}$$

$$\frac{\partial S}{\partial \hat{r}} = -r + \int d\lambda \rho_{\Sigma}(\lambda) \frac{\lambda}{\hat{s} + \lambda \hat{r}}, \tag{131}$$

$$\frac{\partial S}{\partial q} = -\frac{\psi_p}{q} - \psi_p(z - s_t^2) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q} + \frac{1}{q} - \frac{s}{q^2}. \tag{132}$$

Hence the saddle point equations read

$$\hat{s} = b_t^2 \psi_p + \frac{1}{q}, \quad (133)$$

$$\hat{r} = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q}, \quad (134)$$

$$s = \int d\rho_\Sigma(\lambda) \frac{1}{\hat{s} + \lambda \hat{r}}, \quad (135)$$

$$r = \int d\rho_\Sigma(\lambda) \frac{\lambda}{\hat{s} + \lambda \hat{r}}, \quad (136)$$

$$\psi_p(s_t^2 - z) + \frac{\psi_p v_t^2}{1 + \frac{a_t^2 e^{-2t} p}{n} r + \frac{p v_t^2}{n} q} + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0. \quad (137)$$

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

$$2\partial_z \frac{1}{p} \frac{\mathbb{E}[\mathcal{Z}^s] - 1}{s} = 2\partial_z \frac{1}{p} \frac{e^{-\frac{d}{2} S(q^*, r^*)} - 1}{m} \xrightarrow{m \rightarrow 0} -2\partial_z \frac{1}{p} \frac{d}{2} S(q^*, r^*) = q^*. \quad (138)$$

□

### C.5 Proof of Theorem 3.2.

We recall Theorem 3.2 of the MT.

**Theorem** (Informal). *Let  $\rho$  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  $\rho_1$  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  $\rho_2$  coincides with the asymptotic eigenvalue bulk density of the population covariance  $\tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{X}}[\mathbf{U}]$ ;  $\rho_2$  is independent of  $\psi_n$  and its support is on the scale  $\psi_p$ . 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  $\rho_1$  and  $\rho_2$  are respectively on the scales  $\psi_p/\psi_n$  and  $\psi_p$ , 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  $\nu$ th column of  $\Omega$  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

$$\forall \nu = 1, \dots, n, \quad \sum_{i=1}^p \Omega_i^\nu \mathbf{v}_i = 0, \quad (139)$$

$$\forall k = 1, \dots, d, \quad \sum_{i=1}^p \mathbf{W}_{ik} \mathbf{v}_i = 0. \quad (140)$$

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

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

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

$$\psi_n \frac{\frac{pv_t^2}{n}}{1 + \frac{e^{-2t}\mu_1^2 p \sigma_x^2}{n} r + \frac{pv_t^2}{n} q} + \psi_p (s_t^2 - z) + \frac{1 - \psi_p}{q} - \frac{s}{q^2} = 0, \quad (141)$$

and simplifies to

$$\frac{\psi_n \varepsilon}{f} + \psi_p \varepsilon + \frac{(1 - \psi_p) \varepsilon}{f} = 0. \quad (142)$$

It readily gives

$$f = 1 - \frac{1}{\psi_p} - \frac{\psi_n}{\psi_p}. \quad (143)$$

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

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \mathbf{V}\right) \quad (144)$$

$$\mathcal{L}_{\text{test}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}\left(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{U}}\right) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{V}}\right) \quad (145)$$

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  $r(z) = \frac{1}{p} \text{Tr}(\Sigma^{1/2} \mathbf{W}^T (\mathbf{U} - z\mathbf{I})^{-1} \mathbf{W} \Sigma^{1/2})$  we make the following Ansatz in the large  $\psi_p$  limit:

$$q = \mathcal{O}_{\psi_p}(1), \quad r = \mathcal{O}_{\psi_p}\left(\frac{1}{\psi_p}\right). \quad (146)$$

In this limit the saddle point equations becomes at leading order in  $\psi_p$

$$\hat{s} = b_t^2 \psi_p \quad (147)$$

$$\hat{r} = \frac{\psi_p a_t^2 e^{-2t}}{1 + \frac{v_t^2 p}{n} r} \quad (148)$$

$$s = \mathcal{O}(1/\psi_p) \quad (149)$$

$$r = \mathcal{O}(1/\psi_p) \quad (150)$$

$$(s_t^2 - z) + \frac{v_t^2}{1 + \frac{pv_t^2}{n} q} - \frac{1}{q} = 0. \quad (151)$$

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

$$\Delta = (s_t^2 - \lambda(1 - \frac{p}{n})v_t^2)^2 + 4(s_t^2 - \lambda)\frac{p}{n}v_t^2 = 0. \quad (152)$$

It vanishes for

$$\lambda_{\pm} = s_t^2 + v_t^2 \left(1 \pm \sqrt{\frac{p}{n}}\right)^2 \quad (153)$$

which are the edges of the first bulk  $\rho_1$ . 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  $\rho_{\Sigma}$ .

**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

$$\hat{s} = \psi_p b_t^2 + \frac{1}{q} \quad (154)$$

$$\hat{r} = \psi_p a_t^2 e^{-2t}. \quad (155)$$

This yields the following equation on  $q$

$$\psi_p(s_t^2 - z) + \psi_p v_t^2 + \frac{1 - \psi_p}{q} - \frac{1}{q} \int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = 0. \quad (156)$$

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

$$-\psi_p z' + \frac{1 - \psi_p}{q} - \frac{1}{q} \int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = 0. \quad (157)$$

We decompose the integral

$$\int \frac{d\rho_{\Sigma}(\lambda)}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} = \int \frac{d\rho_{\Sigma}(\lambda)(1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t}) - q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t}))}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} \quad (158)$$

$$= 1 - q\psi_p \int \frac{d\rho_{\Sigma}(\lambda)(b_t^2 + \lambda a_t^2 e^{-2t})}{1 + q\psi_p(b_t^2 + \lambda a_t^2 e^{-2t})} \quad (159)$$

By plugging this back in the equation we find

$$q = - \left( z' - \int \frac{d\rho_{\Sigma}(\lambda)(b_t^2 + \lambda a_t^2 e^{-2t})}{1 + \psi_p q(b_t^2 + \lambda a_t^2 e^{-2t})} \right)^{-1}. \quad (160)$$

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

$$q = - \left( z' - \frac{1}{a_t^2 e^{-2t}} \int \frac{d\mu \rho_{\Sigma}(\frac{\mu - b_t^2}{a_t^2 e^{-2t}}) \mu}{1 + \psi_p q \mu} \right)^{-1}. \quad (161)$$

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

$$q = - \left( z' - \int \frac{d\mu \rho_{\mu_1^2(t)\Sigma_t}(\mu) \mu}{1 + \psi_p q \mu} \right)^{-1}. \quad (162)$$

We recognize the Bai-Silverstein equations [21, 2] for the eigenvalue density of the matrix

$$\tilde{\mathbf{U}} = \mu_1^2(t) \frac{\mathbf{W}\Sigma_t\mathbf{W}^T}{d} + (s_t^2 + v_t^2)\mathbf{I}_p = \mathbb{E}_{\mathbf{x}}[\mathbf{U}] \quad (163)$$

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  $\rho_2$ .

□

## C.6 Dynamics on the fast timescales

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

$$\|\mathbf{A}\|_{\text{op}} = \sup_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\| \quad (164)$$

the operator norm and

$$\|\mathbf{A}\|_{\text{F}} = \left( \sum_{i,j=1}^p \mathbf{A}_{ij}^2 \right)^{1/2} \quad (165)$$

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*

$$\|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} = \mathcal{O}\left(\frac{\psi_p}{\sqrt{\psi_n}}\right), \quad (166)$$

when  $p \gg n \gg d$ .

*Proof.* On the one hand,

$$\mathbf{U} = e^{-2t} a_t^2 \frac{\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T}{d} + v_t^2 \frac{\mathbf{\Omega}\mathbf{\Omega}^T}{n} + \frac{e^{-t} a_t v_t}{n\sqrt{d}} (\mathbf{W}\mathbf{X}\mathbf{\Omega}^T + \mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T) + (s_t^2 + v_t^2) \mathbf{I}_p \quad (167)$$

and on the other hand,

$$\tilde{\mathbf{U}} = \mu_1^2 e^{-2t} \frac{\mathbf{W}\mathbf{\Sigma}\mathbf{W}^T}{d} + \Delta_t \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + (s_t^2 + v_t^2) \mathbf{I}_p. \quad (168)$$

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

$$\mathbf{U} - \tilde{\mathbf{U}} = a_t^2 e^{-2t} \frac{\mathbf{W}}{\sqrt{d}} \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \frac{\mathbf{W}^T}{\sqrt{d}} + v_t^2 \left( \frac{\mathbf{\Omega}\mathbf{\Omega}^T}{n} - \mathbf{I}_p \right) + \frac{a_t v_t e^{-t}}{n\sqrt{d}} (\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T). \quad (169)$$

We can bound its operator norm

$$\begin{aligned} \|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} &\leq C_1 \left\| \frac{\mathbf{W}}{\sqrt{d}} \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} + C_2 \left\| \left( \frac{\mathbf{\Omega}\mathbf{\Omega}^T}{n} - \mathbf{I}_p \right) \right\|_{\text{op}} \\ &\quad + \frac{C_3}{n\sqrt{d}} \left\| \mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T \right\|_{\text{op}}, \end{aligned} \quad (170)$$

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.**

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}}. \quad (171)$$

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

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} \leq \left\| \frac{\mathbf{W}^T}{\sqrt{d}} \frac{\mathbf{W}}{\sqrt{d}} \right\|_{\text{op}} \left\| \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \right\|_{\text{op}}. \quad (172)$$

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

$$\left\| \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \right\|_{\text{op}} = \left\| \mathbf{\Sigma}^{1/2} \left( \frac{\mathbf{Z}\mathbf{Z}^T}{n} - \mathbf{I}_d \right) \mathbf{\Sigma}^{1/2} \right\|_{\text{op}} \leq \left\| \left( \frac{\mathbf{Z}\mathbf{Z}^T}{n} - \mathbf{I}_d \right) \right\|_{\text{op}} \left\| \mathbf{\Sigma} \right\|_{\text{op}}. \quad (173)$$

Among our assumptions, we had  $\|\mathbf{\Sigma}\|_{\text{op}} < \mathcal{O}(1)$ . The spectrum of  $\left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right)$  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

$$\left\| \frac{\mathbf{W}}{\sqrt{d}} \left( \frac{\mathbf{X}\mathbf{X}^T}{n} - \mathbf{\Sigma} \right) \frac{\mathbf{W}^T}{\sqrt{d}} \right\|_{\text{op}} \leq \mathcal{O}\left(\frac{p}{\sqrt{nd}}\right). \quad (174)$$

---

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

**Second term.**

$$\|(\frac{\mathbf{\Omega}\mathbf{\Omega}^T}{n} - \mathbf{I}_p)\|_{\text{op}}. \quad (175)$$

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

$$\|(\frac{\mathbf{\Omega}\mathbf{\Omega}^T}{n} - \mathbf{I}_p)\|_{\text{op}} \leq \mathcal{O}(p/n). \quad (176)$$

**Third term.**

$$\|\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T\|_{\text{op}}. \quad (177)$$

We first bound the operator norm by the Frobenius norm.

$$\|\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T\|_{\text{op}} \leq 2\|\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T\|_{\text{F}}. \quad (178)$$

We expand the square

$$\|\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T\|_{\text{F}}^2 = C \sum_{k=1}^d \sum_{i=1}^p (\sum_{\nu=1}^n \mathbf{\Omega}_i^\nu \mathbf{X}_k^\nu \mathbf{W}_{kl})^2. \quad (179)$$

The Central Limit Theorem yields

$$\sum_{\nu=1}^n \mathbf{\Omega}_i^\nu \mathbf{X}_k^\nu \mathbf{W}_{kl} = \mathcal{O}(\sqrt{n}) \mathbf{W}_{kl}, \quad (180)$$

hence

$$\frac{1}{n\sqrt{d}} \|\mathbf{\Omega}\mathbf{X}^T\mathbf{W}^T + \mathbf{W}\mathbf{X}\mathbf{\Omega}^T\|_{\text{op}} = \mathcal{O}(\frac{\sqrt{ndp}}{n\sqrt{d}}) = \mathcal{O}(\sqrt{\frac{p}{n}}) \quad (181)$$

Putting all the contributions together yields

$$\|(\mathbf{U} - \tilde{\mathbf{U}})\|_{\text{op}} \leq \mathcal{O}(\frac{p}{\sqrt{dn}}) = \mathcal{O}(\frac{\psi_p}{\sqrt{\psi_n}}). \quad (182)$$

□

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

$$\mathcal{L}_{\text{train}} \simeq \mathcal{L}_{\text{test}} \simeq 1 - \mathcal{O}(\Delta_t). \quad (183)$$

*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  $\psi_p$  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  $\lambda$  is in the second bulk and is  $e^{-\lambda \frac{\Delta_t \tau}{\psi_p}} \sim 1$  if  $\lambda$  is in the first bulk. We remind the expressions of the train and test loss

$$\mathcal{L}_{\text{train}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \mathbf{U}) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}(\frac{\mathbf{A}}{\sqrt{p}} \mathbf{V}) \quad (184)$$

$$\mathcal{L}_{\text{test}}(\mathbf{A}) = 1 + \frac{\Delta_t}{d} \text{Tr}(\frac{\mathbf{A}^T}{\sqrt{p}} \frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{U}}) + \frac{2\sqrt{\Delta_t}}{d} \text{Tr}(\frac{\mathbf{A}}{\sqrt{p}} \tilde{\mathbf{V}}) \quad (185)$$

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

$$\frac{\mathbf{A}(\tau)}{\sqrt{p}} = \frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} (e^{-\frac{2\Delta_t}{d} \mathbf{U} \tau} - \mathbf{I}_p) \quad (186)$$

$$= \frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} \sum_{\lambda} (e^{-\frac{2\Delta_t}{d} \lambda \tau} - 1) \mathbf{v}_\lambda \mathbf{v}_\lambda^T \quad (187)$$

$$\sim -\frac{1}{\sqrt{\Delta_t}} \mathbf{V}^T \mathbf{U}^{-1} \sum_{\lambda \in \rho_2} \mathbf{v}_\lambda \mathbf{v}_\lambda^T, \quad (188)$$

where  $\lambda \in \rho_2$  means that the eigenvalue  $\lambda$  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

$$\mathcal{L}_{\text{gen}} = -\frac{\mu_1^2(t)\Delta_t}{\Gamma_t^2 d} \text{Tr} \left( \sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'} \mathbf{v}_{\lambda'}^T \mathbf{U}^{-1} \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \right) \quad (189)$$

$$= -\frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left( \sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \mathbf{U}^{-1} \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \quad (190)$$

$$= -\frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left( \sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \quad (191)$$

$$(192)$$

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  $\mathcal{L}_{\text{gen}}$ ,

$$|\mathcal{L}_{\text{gen}}| \leq \left\| \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left( \sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T (\mathbf{U} - \tilde{\mathbf{U}}) \mathbf{v}_{\lambda'} \right) \right\|_{\text{op}} \quad (193)$$

$$\leq \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} d \frac{1}{\psi_p^2} \left\| \frac{\mathbf{W} \mathbf{W}^T}{d} \right\|_{\text{op}} \frac{\psi_p}{\sqrt{\psi_n}} \leq \mathcal{O}\left(\frac{d\psi_p^2}{d\psi_p^2 \sqrt{\psi_n}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{\psi_n}}\right). \quad (194)$$

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  $\psi_p$ . The bound yield that  $\mathcal{L}_{\text{gen}}$  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  $\mathcal{L}_{\text{train}}$

$$\mathcal{L}_{\text{train}} = 1 + \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \left( \sum_{\lambda, \lambda' \in \rho_2} \mathbf{v}_{\lambda'}^T \frac{1}{\lambda'} \frac{\mathbf{W} \mathbf{W}^T}{d} \frac{1}{\lambda} \mathbf{v}_{\lambda} \mathbf{v}_{\lambda}^T \mathbf{U} \mathbf{v}_{\lambda'} \right) - \frac{2\Delta_t \mu_1^2}{\Gamma_t^2 d} \sum_{\lambda \in \rho_2} \mathbf{v}_{\lambda}^T \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{U}^{-1} \mathbf{v}_{\lambda} \quad (195)$$

$$= 1 - \frac{\mu_1^2 \Delta_t}{\Gamma_t^2 d} \sum_{\lambda \in \rho_2} \frac{1}{\lambda} \mathbf{v}_{\lambda}^T \frac{\mathbf{W} \mathbf{W}^T}{d} \mathbf{v}_{\lambda}. \quad (196)$$

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:

$$\mathcal{L}_{\text{train}} \sim \mathcal{L}_{\text{test}} = 1 - \mathcal{O}(\Delta_t). \quad (197)$$

□

## D Numerical experiments for Random Features

**Details on the numerical experiments.** All the numerical experiments for the RFNN were conducted using  $\sigma = \tanh$  and  $\sigma_x = 1$  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. 6 and show that the rescaling of the training times  $\tau$  by  $\tau_{\text{mem}} = \psi_p / \Delta_t \lambda_{\min}$  also makes the loss curves collapse. Of particular interest is the behavior of  $\tau_{\text{mem}}$ , and more specifically the ratio  $\tau_{\text{mem}} / \tau_{\text{gen}}$ , at small  $t$ . Recall that

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

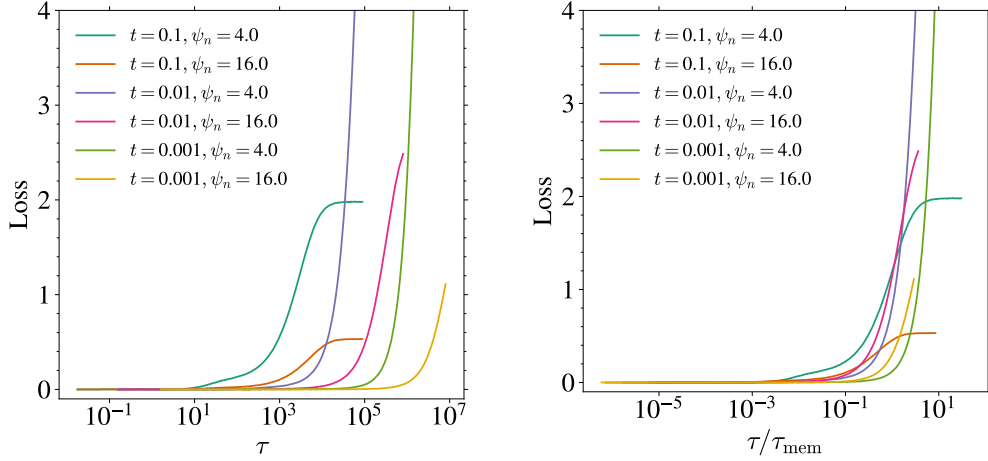


Figure 6: **Generalization loss for different diffusion times  $t$ .** Generalization loss  $\mathcal{L}_{\text{gen}}$  against (Left) training time  $\tau$  and (Right) rescaled training time  $\tau/\tau_{\text{gen}}$  for different  $\psi_p = 32$ ,  $d = 100$  and different  $\psi_n$  and  $t$ .

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. 7, 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,

$$\mathcal{E}_{\text{Score}} = \frac{1}{d} \mathbb{E}_{\mathbf{y} \sim \mathcal{N}(0, \Gamma_t^2 I_p)} \left[ \left\| \mathbf{s}_{\mathbf{A}(\tau)}(\mathbf{y}) + \frac{\mathbf{y}}{\Gamma_t^2} \right\|^2 \right], \quad (198)$$

serves as a measure of the generalization capability of the generative process. As shown in [24], the Kullback–Leibler divergence between the true data distribution  $P_{\mathbf{x}}$  and the generated distribution  $\hat{P}$  can be upper bounded

$$\mathcal{D}_{\text{KL}}(P_{\mathbf{x}} \parallel \hat{P}) \leq \frac{d}{2} \int dt \mathcal{E}_{\text{Score}}(\mathbf{A}_t), \quad (199)$$

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 [4]. We have numerically investigate the behaviour of  $\mathcal{E}_{\text{score}}$  on Fig. 8. On the fast timescale  $\tau_{\text{gen}}$ , it decreases until a minimal value  $\mathcal{E}_{\text{score}}^*$  that depends only on  $\psi_n$  with a power-law  $\psi_n^{-\eta}$  with  $\eta \simeq 0.59$ . We leave for future work performing an accurate numerical estimate of  $\eta$  and a developing a theory for it.

**Spectrum of  $\mathbf{U}$ .** In Fig. 9, we compare the solutions of the equations of Theorem 3.1 to the histogram of finite size realizations of  $\mathbf{U}$ .



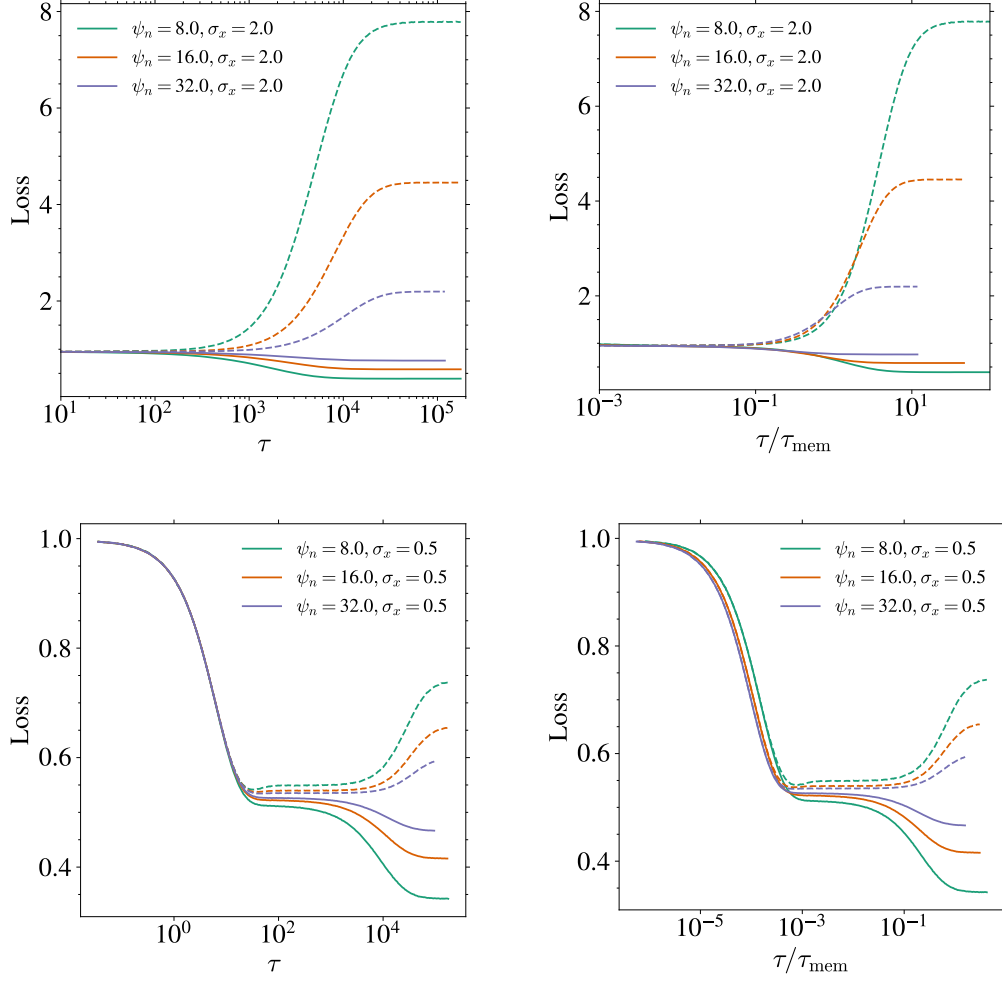


Figure 7: **Different  $\sigma_{\mathbf{x}}^2$ .** Train loss (solid line) and test loss (dotted line) for  $\psi_p = 64$ ,  $t = 0.1$ ,  $d = 100$ , different  $\psi_n$  and  $\sigma_{\mathbf{x}} = 2.0$  (top) and  $\sigma_{\mathbf{x}} = 0.5$  (bottom) against the training time  $\tau$  and the rescaled training time  $\tau/\tau_{\text{mem}}$ .

**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.10.

## References

- [1] Bach, F. (2023). Polynomial magic iii: Hermite polynomials. <https://francisbach.com/hermite-polynomials/>. Accessed: 2025-10-09.
- [2] Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442.
- [3] Bodin, A. P. M. (2024). *Random Matrix Methods for High-Dimensional Machine Learning Models*. Phd thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- [4] Bortoli, V. D. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*. Expert Certification.
- [5] Chen, C., Liu, D., and Xu, C. (2024). Towards memorization-free diffusion models.

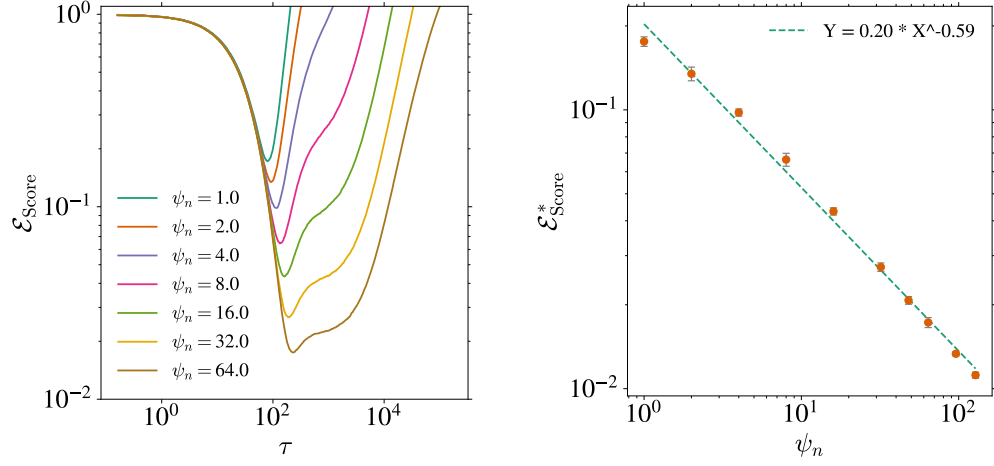


Figure 8: **Effect of  $\psi_n$  on  $\mathcal{E}_{\text{Score}}^*$ .** (Left) Error between the learned score and the true score  $\mathcal{E}_{\text{Score}}$  for  $\psi_p = 32$ ,  $t = 0.01$ , and various values of  $\psi_n$ . (Right) Minimum score error  $\mathcal{E}_{\text{Score}}^* = \min_{\tau} [\mathcal{E}_{\text{Score}}(\tau)]$  as a function of  $\psi_n$ , 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.

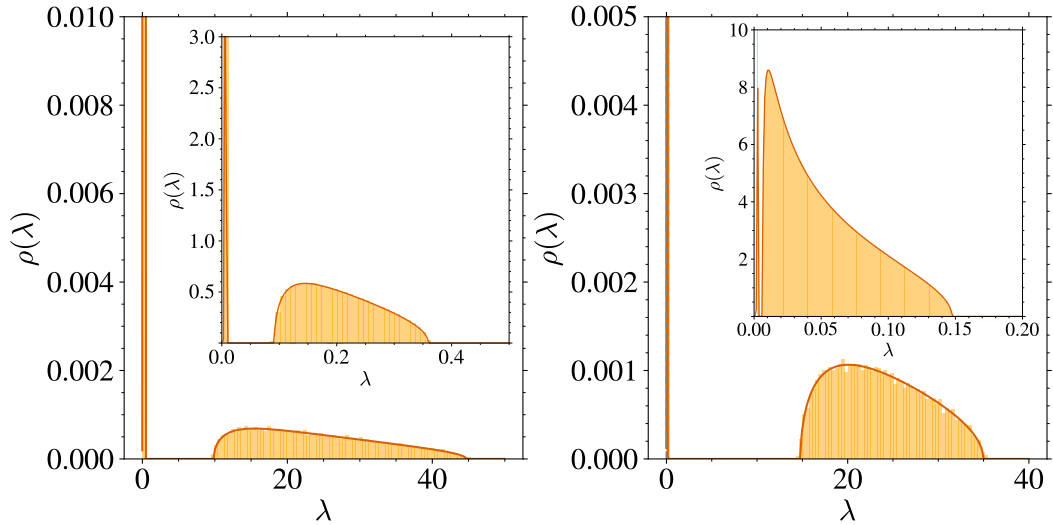


Figure 9: **Spectrum of  $U$ .** Solutions of the equations in Theorem 3.1. (solid lines) and empirical spectrum for  $\rho_{\Sigma}(\lambda) = \delta(\lambda - 1)$  and  $d = 100$  (histogram). (Left)  $\psi_p = 64$ ,  $\psi_n = 8$ ,  $t = 0.01$ . (Right)  $\psi_p = 64$ ,  $\psi_n = 32$ ,  $t = 0.01$ .

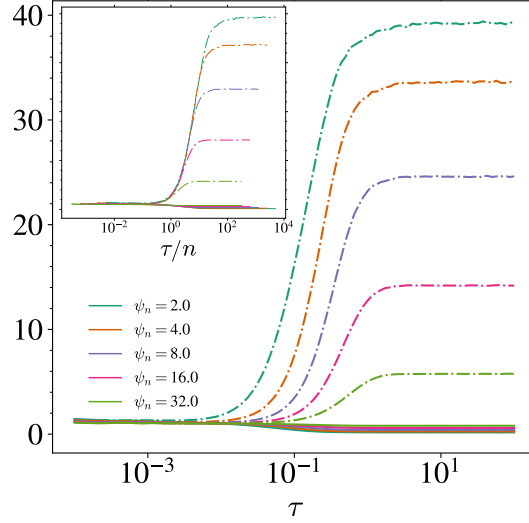


Figure 10: **Adam**. Train loss (solid line) and test loss (dotted line) at  $t = 0.01$ ,  $d = 100$ ,  $\psi_p = 64$  for several  $\psi_n$  with the Pytorch [17] implementation of Adam. The inset shows the effect of a rescaling of the training time by  $n$ .

- [6] 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.
- [7] George, A. J., Veiga, R., and Macris, N. (2025). Denoising score matching with random features: Insights on diffusion models from precise learning curves.
- [8] 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.
- [9] 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.
- [10] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.
- [11] Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance.
- [12] 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.
- [13] Kibble, W. F. (1945). An extension of a theorem of mehlér’s on hermite polynomials. *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1):12–15.
- [14] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *ICLR (Poster)*.
- [15] 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)*.
- [16] Mei, S. and Montanari, A. (2020). The generalization error of random features regression: Precise asymptotics and double descent curve.
- [17] 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.

- [18] Potters, M. and Bouchaud, J.-P. (2020). *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press.
- [19] Péché, S. (2019). A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7.
- [20] 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.
- [21] 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.
- [22] Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803.
- [23] Song, J., Meng, C., and Ermon, S. (2022). Denoising diffusion implicit models.
- [24] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021a). Maximum likelihood training of score-based diffusion models.
- [25] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021b). Score-based generative modeling through stochastic differential equations.
- [26] 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.
- [27] Wen, Y., Liu, Y., Chen, C., and Lyu, L. (2024). Detecting, explaining, and mitigating memorization in diffusion models.