DOUBLE DESCENT AND MALIGN OVERFITTING IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional wisdom in deep learning holds that overparameterization—having more parameters p than training samples n—is benign: larger models generalize better and, even without regularization, interpolating models generalize well, the test error following a double-descent curve. One might expect the same benign overfitting for diffusion models, whose training reduces to regression, i.e. to minimizing a quadratic denoising score-matching loss. Yet the opposite is observed: overfitting here is catastrophic, driving the model into a memorization regime. We resolve this paradox by combining experiments on U-Nets trained on CelebA with a random-features model for which we derive closed-form learning curves. We show that with a fixed number m of noise realizations per training sample, an interpolation peak does occur, but at $p \sim nm$ rather than at $p \sim n$ as in standard regression. The rise of the test loss, however, sets in much earlier, at $p \sim n$, independently of m. This overfitting is *malign* because, although the implicit regularization of training is fully at work, it drives the model toward the empirical score, which memorizes the training set, rather than toward the true score. A bias–variance decomposition pinpoints the mechanism: the bias of the score estimator starts to grow at $p \sim n$; past the peak the variance decays, as in regression, whereas the bias keeps growing and both saturate at a large value. Since diffusion models are trained with $m \gg 1$, the peak is pushed to very large model sizes, and therefore models sit on the rising branch that precedes it, where malign overfitting is already in play. Nevertheless, overparameterization remains beneficial when paired with regularization: in the random-features theory and in U-Net experiments, optimally regularized large models—via a ridge penalty or early stopping, respectively—outperform unregularized models of any size.

1 INTRODUCTION

One of the most surprising discoveries in modern machine learning is that overparameterization is not necessarily harmful, and can in fact be beneficial in classification and regression problems. Contrary to the traditional wisdom that one should "never fit noisy training data exactly", heavily overparameterized models in modern machine learning often interpolate the training data while achieving excellent test performance (Lawrence et al., 1997; Neyshabur et al., 2015; Zhang et al., 2017). The most striking manifestation of this phenomenon is the *double descent* curve of the test error as a function of model capacity—specifically, the number of learnable parameters in the neural network (Belkin et al., 2019; Geiger et al., 2020). Rather than exhibiting the classical U-shaped curve associated to bias–variance tradeoff (Wasserman, 2004), the test error first follows this U-shape, then peaks at the *interpolation threshold*—the point where the model can exactly fit the training data and the variance of the model is large—and then descends again, often reaching a global minimum in the heavily overparameterized regime. This phenomenon, and in particular this last descent, called *benign overfitting*, has proved remarkably ubiquitous across different model classes, underscoring the benefits of overparameterization. It was subsequently demonstrated at scale in deep networks by Nakkiran et al. (2021a) and analyzed theoretically in multiple studies (Bartlett et al., 2020; Hastie et al., 2022; Mei & Montanari, 2019; D'Ascoli et al., 2020).

Our work revisits this central phenomenon for diffusion and score-based models, (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021b) which have become the state of the art for generating images (Esser et al., 2024), sounds (Kong et al., 2021), and videos (Yang

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

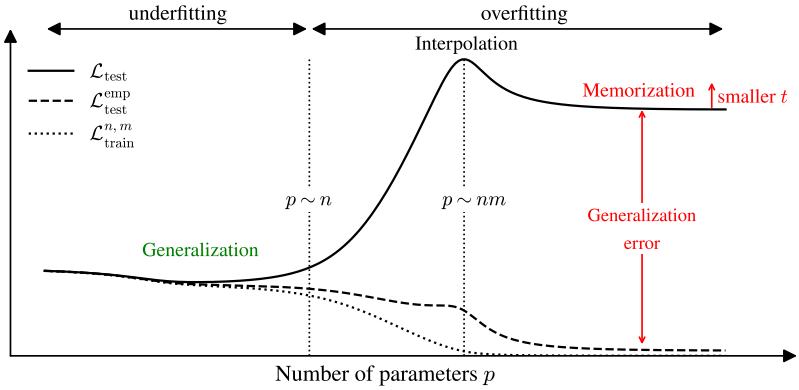

Figure 1: **Qualitative summary of our results.** Sketch of the standard test loss $\mathcal{L}_{\text{test}}$ (solid), the empirical test loss $\mathcal{L}^{\text{emp}}_{\text{test}}$ (dashed), and the training loss $\mathcal{L}^{n,m}_{\text{train}}$ (dotted) versus the number of parameters p, for a diffusion model trained on n samples with $m > 1$ noise realizations each. $\mathcal{L}_{\text{test}}$ reaches its minimum before $p \sim n$, where the model can generalize, then rises to a peak at the *interpolation threshold* $p \sim nm$, where $\mathcal{L}^{n,m}_{\text{train}}$ reaches 0 (shifted from $p \sim n$, its value in the regression setting $m = 1$). For $p \gg nm$, the test loss decreases a second time (*double descent*) but settles on a plateau that lies above its earlier minimum, whereas $\mathcal{L}^{\text{emp}}_{\text{test}}$ decreases toward zero, more so the larger m: since $\mathcal{L}^{\text{emp}}_{\text{test}}$ measures the distance to the *empirical score*, which memorizes the training set, the model is approaching the memorization regime. The generalization error is larger, the smaller is t (the amount of noise). In contrast to the regression setting $m = 1$, where overfitting is benign, here it is malign.

et al., 2025). Diffusion models are trained by progressively noising the data and teaching the model to reverse this process. In the language of nonequilibrium physics, they learn to time-reverse an Ornstein–Uhlenbeck (OU) process. The key ingredient is learning the *score function*, which is the force field that enables time-reversal and denoising (Hyvärinen, 2005; Vincent, 2011). This reduces to *a sequence of regression problems*: at each noise level, one minimizes a quadratic loss between the neural network's predicted score and the ground truth. As a result, in light of the modern machine learning findings discussed above, one might expect benign overfitting to apply here as well. Yet numerical experiments reveal a strikingly different story: increasing model capacity leads to memorization and poor generalization (Yoon et al., 2023; Gu et al., 2025; Kadkhodaie et al., 2024). Overfitting proves highly detrimental ("malign"), as it compromises generative quality by causing the model to reproduce training data rather than capture the underlying distribution.

This leads to a puzzle: diffusion models give rise to regression problems, and therefore one might expect benign overfitting. Yet experiments show the opposite. The aim of our work is to resolve this paradox by revisiting the double descent phenomenon for diffusion models. This is relevant both for the theory of generative models and for applications. In particular, it allows us to clarify a fundamental question about generative models: Does overparameterization help or hurt diffusion models?

Generative Diffusion and Its Different Losses. Standard diffusion models transport a target distribution P_0 on $\mathbb{R}^d$ to Gaussian white noise $\mathcal{N}(0, \boldsymbol{I}_d)$ via an OU *forward process* $\mathrm{d}\mathbf{x} = -\mathbf{x}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{W}(t)$, where $\mathbf{W}(t)$ is a standard Wiener process. Exact time-reversal of this OU process can be implemented (Anderson, 1982; Haussmann & Pardoux, 1986) by using a guiding force field which is the exact score function $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$, where P_t is the marginal density at time t. Following Hyvärinen (2005) and Vincent (2011), generation is performed by using a parametrized score function $\boldsymbol{s}(\mathbf{x}, t)$. Given a database $\mathcal{D} = \{\mathbf{x}^\nu\}_{\nu=1,\dots,n}$ consisting of n i.i.d. samples of P_0, one learns this score function by minimizing the Denoising Score Matching (DSM) loss. At fixed t[1], this DSM loss reads

$$\mathcal{L}^{\text{emp}}_{\text{test}}(\boldsymbol{s}) = \frac{1}{nd} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left[\left\| \sqrt{\Delta_t}\, \boldsymbol{s}(e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\, \boldsymbol{\xi}) + \boldsymbol{\xi} \right\|^2 \right], \tag{1}$$

where $\Delta_t = 1 - e^{-2t}$ and the expectation is over $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. In practice, the average over $\boldsymbol{\xi}$ is not performed exactly: the training is done on batches of the n samples, drawing a fresh noise

[1]The loss at fixed t, rather than averaged over t, reveals how the noise level controls malign overfitting.

$\boldsymbol{\xi}$ each time a sample is visited. Each sample is thus seen with m different noise realizations, m being the number of epochs (typically[2] 10^3). To analyze the role of this finite noise sampling within the classical double-descent framework, which is formulated for the minimizer of an empirical loss (Belkin et al., 2019; Geiger et al., 2020), we fix the noises and consider the loss

$$\mathcal{L}^{n,m}_{\text{train}}(\boldsymbol{s}) = \frac{1}{nmd}\sum_{\nu=1}^{n}\sum_{\mu=1}^{m}\left\|\sqrt{\Delta_t}\,\boldsymbol{s}(e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\mu}) + \boldsymbol{\xi}^{\nu\mu}\right\|^2. \tag{2}$$

Fixing the noises is an idealization of the actual procedure, but it retains its essential feature: only nm pairs $(\mathbf{x}^\nu, \boldsymbol{\xi}^{\nu\mu})$ are presented to the network ($\mathcal{L}^{n,m}_{\text{train}} \to \mathcal{L}^{\text{emp}}_{\text{test}}$ for $m \to \infty$). It is also the standard setting of theoretical analyses (Cui et al., 2024; George et al., 2026). While this resembles a regression problem, it is a non-standard one: the score $\boldsymbol{s}$ is evaluated on points $\boldsymbol{y}^{\nu,\mu} = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\mu}$ that are correlated, forming clusters of m points around each sample $\mathbf{x}^\nu$. Finally, generalization is measured by the *standard* test loss

$$\mathcal{L}_{\text{test}}(\boldsymbol{s}) = \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\left[\left\|\sqrt{\Delta_t}\,\boldsymbol{s}(e^{-t}\mathbf{x} + \sqrt{\Delta_t}\,\boldsymbol{\xi}) + \boldsymbol{\xi}\right\|^2\right]. \tag{3}$$

where the expectation is taken over both fresh noises and fresh samples. We refer to $\mathcal{L}^{\text{emp}}_{\text{test}}(\boldsymbol{s})$ as the *empirical* test loss: it is a test loss since the noise is averaged exactly, but also an empirical one because it depends on the n training samples. In practice it is readily evaluated by averaging over fresh noises. To identify where the test loss comes from, we analyze the bias and variance contributions, $\mathcal{B}^2$ and $\mathcal{V}$, to $\mathcal{L}_{\text{test}} = C_t + \Delta_t\left(\mathcal{B}^2 + \mathcal{V}\right)$, with

$$\mathcal{B}^2 = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\|\nabla_{\boldsymbol{y}}\log P_t(\boldsymbol{y}) - \langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\rangle\|^2\right], \quad \mathcal{V} = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\left\langle\|\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\rangle\|^2\right\rangle\right]. \tag{4}$$

Here $\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}$ denotes the minimizer of the train loss $\mathcal{L}^{n,m}_{\text{train}}$ for a given dataset $\mathcal{D}$ and initialization of the parameters $\boldsymbol{\Theta}$; C_t a term which depends only on the data distribution and is independent of the learned score (see Appendix C.2); $\mathbb{E}_{\boldsymbol{y}}$ is the average over test samples $\boldsymbol{y} \sim P_t$, whereas $\langle\cdot\rangle$ denotes the average over the realizations of the training set, and of other sources of randomness such as $\boldsymbol{\Theta}$. As we will show, distinguishing between $\mathcal{L}^{\text{emp}}_{\text{test}}$ and $\mathcal{L}_{\text{test}}$ is key to resolving the paradox: benign behavior in the former coexists with malign overfitting in the latter.

Contributions and theoretical picture. Figure 1 summarizes the behavior of the different losses, as obtained from both the theoretical analysis and our numerical experiments, each evaluated from the score learned by training on n samples independently noised m times. Importantly, we consider solutions obtained after very large training times τ, representative of the $\tau \to \infty$ limit, and without regularization. The effects of finite τ and of regularization are discussed later. The first important observation is that the test loss $\mathcal{L}_{\text{test}}$ does exhibit a double descent as the number of parameters increases: it first decreases toward a minimum, rises to a peak, then descends again. The peak occurs at the interpolation threshold, where the training loss reaches zero and the test loss is dominated by the variance of the estimator. Compared to the usual double descent, however, the peak is shifted to much larger model sizes: its position now grows with both n and m (for simple models, proportionally to nm). More importantly, while the test loss descends after the peak, it plateaus at a value larger than its first minimum; the smaller t, the higher this plateau is. Whereas in the supervised setting both variance and bias decrease after the interpolation peak (Neal et al., 2019; D'Ascoli et al., 2020), here the variance decays and the bias grows. Both saturate at large values, producing the high plateau observed at small t. Here, overfitting is clearly detrimental. The key to understanding this phenomenon lies in the empirical test loss $\mathcal{L}^{\text{emp}}_{\text{test}}$, which, up to a constant, measures the mean square distance to the *empirical score*, the score associated with the mixture of Gaussians centered on the points $e^{-t}\mathbf{x}^\nu$ with variance Δ_t (see Appendix C.3). Its global minimizer is the empirical score itself (Biroli et al., 2024; Li et al., 2024), which memorizes the training set unless n grows exponentially with d (Biroli et al., 2024). Numerically, $\mathcal{L}^{\text{emp}}_{\text{test}}$ first follows the training loss closely. It then displays a peak at small m, which fades as m grows, and beyond the interpolation threshold it reaches very small values. This is the key to solving the paradox: the benign-overfitting mechanism—whereby the training dynamics selects, among all interpolating solutions, those with the smallest error—is fully operative, but it acts on $\mathcal{L}^{\text{emp}}_{\text{test}}$ and not on the loss we actually care about, $\mathcal{L}_{\text{test}}$. The implicit

[2] With fresh noise at each step, $m = N_{\text{steps}}B/n$ with N_{steps} gradient steps and batch size B; this gives $m \approx 2\cdot 10^3$ for DDPM on CIFAR-10 and EDM on ImageNet-64 (Ho et al., 2020; Karras et al., 2022).

regularization of training drives the model toward the minimizer of the *empirical* test loss, i.e. the empirical score which memorizes the training set, rather than toward the exact score: hence the name *malign* overfitting[3]. This also explains the growth of the bias in the overparameterized regime: it does not reflect a limitation of the model class, but the minimization of the "wrong" loss. None of this implies that overparameterization hurts diffusion models. Overparameterization is still beneficial provided it is paired with a suitable regularization: with a ridge penalty (in the theory) or early stopping (in the experiments), both the peak and the plateau of Fig. 1 disappear, and the test loss and the FID reach their lowest values in the overparameterized regime ($p \gg n$).

The theoretical picture described above rests on two complementary analyses. Analytically, we study Random Feature Neural Networks (Rahimi & Recht, 2007) at arbitrary finite m, in the proportional limit where the dimension d, the data n and the number of parameters p all go to infinity with fixed ratios, extending George et al. (2026). Numerically, we train DDPM-style U-Nets (Ho et al., 2020; Ronneberger et al., 2015) on the CelebA dataset (Liu et al., 2015) across a range of network widths W, tracking how the losses and the Fréchet-Inception Distance (FID, Heusel et al., 2017) evolve. In the following sections, we present these findings in detail, discussing the role of m, the bias and variance of the estimator, and the impact of regularization.

Related works. We present here the closest relevant works, and refer to Appendix A for a more thorough discussion. Diffusion models are known empirically to overfit and memorize the training data, with state-of-the-art image models reproducing a non-negligible fraction of their training set (Carlini et al., 2023; Somepalli et al., 2023a;b). This memorization has been shown to depend heavily on the data distribution, the model, and the training procedure (Gu et al., 2025; Yoon et al., 2023), but also to be significantly reduced by weight decay and early-stopping (Gu et al., 2025; Baptista et al., 2025; Favero et al., 2025). However, it is unavoidable under the empirical score hypothesis (Biroli et al., 2024; Achilli et al., 2024; Ventura et al., 2025) unless n grows exponentially with d (Biroli et al., 2024). Several theoretical analyses account for the network architecture and the high dimensionality of the data, in particular by studying the score learned by a parametric model (Cui et al., 2024; 2025; Merger & Goldt, 2026; Buchanan et al., 2025). Closest to our work, George et al. (2026) compute asymptotic training and test DSM losses for a Random Feature model (Rahimi & Recht, 2007) and link them to memorization for $m = 1$ and $m = \infty$. We extend their result to arbitrary finite m. Bonnaire et al. (2025) study the generalization and memorization timescales of the corresponding $m = \infty$ training dynamics, while, concurrently with our work, Latourelle-Vigeant et al. (2026) analyze the lazy regime ($p \to \infty$) training dynamics, together with the resulting sampling dynamics. Finally, that diffusion models do not overfit benignly is by now well documented empirically (Yoon et al., 2023) and proven in a model-independent way by Farghly et al. (2026): overfitting and good generalization cannot coexist unless the sample size grows exponentially with d. What remains open, and what we clarify in this work, is twofold: the mechanism behind this failure, i.e. what is specific to diffusion models that switches off the benign overfitting at work in supervised learning, and whether double descent carries over at all. On the latter, the evidence is conflicting: in the RFNN, George et al. (2026) find an interpolation peak for $m = 1$ that vanishes for $m = \infty$, while Marion & Wu (2026) observe an epoch-wise double descent in distributional metrics such as the FID, but none in the test loss. Working at finite m is the key ingredient to solve this puzzle: it is what connects the benign overfitting of regression ($m = 1$) to the malign overfitting observed in practice ($m \gg 1$), and allow us to identify the mechanism behind the latter.

2 Double Descent in Diffusion Models

2.1 Numerical Method

We train DDPM-style (Ho et al., 2020) U-Net architectures (Ronneberger et al., 2015) to predict the noise on n (from 512 to 8192) CelebA (Liu et al., 2015) grayscale images downsampled to 32×32. Each image $\mathbf{x}^\nu$ is assigned m fixed noise realizations $\{\boldsymbol{\xi}^{\nu\mu}\}_{\mu=1,\dots,m}$, yielding a total of $n \times m$ frozen training pairs. The architecture of the U-Net follows Bonnaire et al. (2025), except it has four resolution levels with channel multipliers $1, 2, 4, 4$ relative to a base width W that we vary from $W = 2$ ($p \approx 29$K parameters) to 192 ($p \approx 240$M parameters). Note that with such an architecture, p scales

[3] Although this loss vanishes only for $m, p \to \infty$, we find that a very small value is obtained already for moderate m, thus implying a very small distance to the empirical score and hence partial memorization.

as W^2. We discretize diffusion time on a nonuniform grid of $T = 1000$ points, specified in Appendix B.1. Each model is trained to predict the noise $\boldsymbol{\xi}^{\nu\mu}$ from interpolated samples $\mathbf{x}_t^{\nu\mu} = e^{-t}\mathbf{x}^{\nu} + \sqrt{1-e^{-2t}}\,\boldsymbol{\xi}^{\nu\mu}$ by minimizing the standard DDPM MSE loss $(nmd)^{-1}\sum_{\mu\nu}\|\boldsymbol{\xi}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \boldsymbol{\xi}^{\nu\mu}\|^2$ using the Adam optimizer for $\tau_{\max} = 2\text{M}$ steps. At each optimization step, a diffusion time is sampled independently for each pair, uniformly over the eligible grid points.

2.2 Analytical Method

Setting. We model the score function with a Random Features Neural Network (RFNN, Rahimi & Recht, 2007):

$$\boldsymbol{s}_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{A}}{\sqrt{p}}\sigma\left(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\right). \tag{5}$$

An RFNN is a two-layer neural network whose first layer weights ($\mathbf{W} \in \mathbb{R}^{p\times d}$) are drawn from a Gaussian distribution and remain frozen while the second layer weights ($\mathbf{A} \in \mathbb{R}^{d\times p}$) are learned during training. The activation function σ, which is applied element-wise, is assumed to admit a Hermite expansion and to satisfy $\mathbb{E}[\sigma(z)] = 0$. Here, we study the empirical risk minimizer of the DSM loss Eq. 2 with ridge regularization $\frac{\Delta_t\lambda}{pd}\|\mathbf{A}\|_F^2$, trained on a dataset composed of n training data points $\mathbf{x}^{\nu} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ each corrupted m times by $\boldsymbol{\xi}^{\nu\mu} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Most of our results can be extended to any zero-mean sub-Gaussian distribution with extensive trace covariance (see Appendix C.12). We work in the high-dimensional setting $p, n, d \gg 1$ with $n/d \to \psi_n$, $p/d \to \psi_p$ and $\psi_n, \psi_p, m = O(1)$. We derive a tight characterization of the asymptotic learning curves of the losses Eq. 2, Eq. 1, and Eq. 3. This model has already been studied in the context of diffusion (George et al., 2026; Bonnaire et al., 2025) for $m \in \{1, \infty\}$ and recently for linear activation $\sigma(x) = x$ (Farghly et al., 2026). Note that the number of adjustable parameters of the RFNN model (5) is pd.

Gaussian Equivalence Principle. To derive asymptotic closed-form expressions for the losses, we build on the *Gaussian Equivalence Principle* (GEP, Pennington & Worah, 2017; Péché, 2019; Mei & Montanari, 2019; Goldt et al., 2020; Gerace et al., 2020; Goldt et al., 2021; Hu & Lu, 2023), which states that the nonlinear features $\mathbf{F} = \sigma(\mathbf{W}\mathbf{Y}/\sqrt{d})$ can be replaced, in the high-dimensional regime, by a linear surrogate matching their first two moments: $\mathbf{F}^{\text{GEP}} = \mu_1\,\mathbf{W}\mathbf{Y}/\sqrt{d} + \mu_*\,\boldsymbol{\Omega}$, with $\mu_1 = \mathbb{E}[\sigma(z)z]$, $\mu_*^2 = \mathbb{E}[\sigma^2(z)] - \mu_1^2$ and $\boldsymbol{\Omega}$ a Gaussian tensor independent of $\mathbf{W}$ and $\mathbf{Y}$ with covariance $\mathbb{E}[\Omega_\alpha^{\nu\mu}\Omega_\beta^{\nu'\mu'}] = \delta_{\alpha\beta}\,\delta^{\nu\nu'}(\delta^{\mu\mu'} + \kappa(1-\delta^{\mu\mu'}))$. The constant κ captures the correlation between features of different noise copies of the same sample and reads

$$\kappa = \frac{1}{\mu_*^2}\,\mathbb{E}_{u,v,w}\big[\big(\sigma(e^{-t}u + \sqrt{\Delta_t}\,v) - \mu_1 e^{-t}u\big)\big(\sigma(e^{-t}u + \sqrt{\Delta_t}\,w) - \mu_1 e^{-t}u\big)\big], \tag{6}$$

with $u, v, w \sim \mathcal{N}(0,1)$ independent (see Appendix C.5 for more details).

Decomposition of the losses, bias and variance. The GEP allows us to express the losses as rational polynomials of random matrices. Define the resolvent and the two traces

$$\mathcal{G}_{z,\zeta,\epsilon} = \left(\frac{\mathbf{F}\mathbf{F}^T}{nm} - z\boldsymbol{I}_p - \zeta\frac{\mathbf{W}\mathbf{W}^T}{d} + \epsilon\frac{\mathbf{H}\mathbf{H}^T}{n}\right)^{-1}, \quad \mathbf{H} = \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{F}], \tag{7}$$

$$T_1(z,\zeta,\epsilon) = \frac{1}{d(nm)^2}\,\mathrm{Tr}\Big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}_{z,\zeta,\epsilon}\mathbf{F}\boldsymbol{\xi}^T\Big), \qquad T_2 = \frac{1}{dnm}\,\mathrm{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}_{-\lambda,0,0}\frac{\mathbf{W}}{\sqrt{d}}\right). \tag{8}$$

Using the GEP, the three losses and the bias–variance decomposition Eq. 4 read

$$\mathcal{L}_{\text{train}}^{n,m} = 1 - T_1(-\lambda, 0, 0), \tag{9}$$

$$\mathcal{L}_{\text{test}} = 1 - 2\mu_1\sqrt{\Delta_t}\,T_2 + \mu_1^2\,\partial_\zeta T_1\big|_{(-\lambda,0,0)} + \mu_*^2\,\partial_z T_1\big|_{(-\lambda,0,0)}, \tag{10}$$

$$\mathcal{L}_{\text{test}}^{\text{emp}} = 1 - 2\mu_1\sqrt{\Delta_t}\,T_2 - \partial_\epsilon T_1\big|_{(-\lambda,0,0)} + \Delta_t\mu_1^2\,\partial_\zeta T_1\big|_{(-\lambda,0,0)} + \mu_*^2(1-\kappa)\,\partial_z T_1\big|_{(-\lambda,0,0)}, \tag{11}$$

$$\mathcal{B}^2 = \left(1 - \frac{\mu_1 T_2}{\sqrt{\Delta_t}}\right)^2, \qquad \mathcal{V} = \frac{\mu_1^2\,\partial_\zeta T_1\big|_{(-\lambda,0,0)} + \mu_*^2\,\partial_z T_1\big|_{(-\lambda,0,0)}}{\Delta_t} - \frac{\mu_1^2 T_2^2}{\Delta_t}. \tag{12}$$

The traces T_1 and T_2 can be computed asymptotically, via a pair of auxiliary order parameters (q, r).

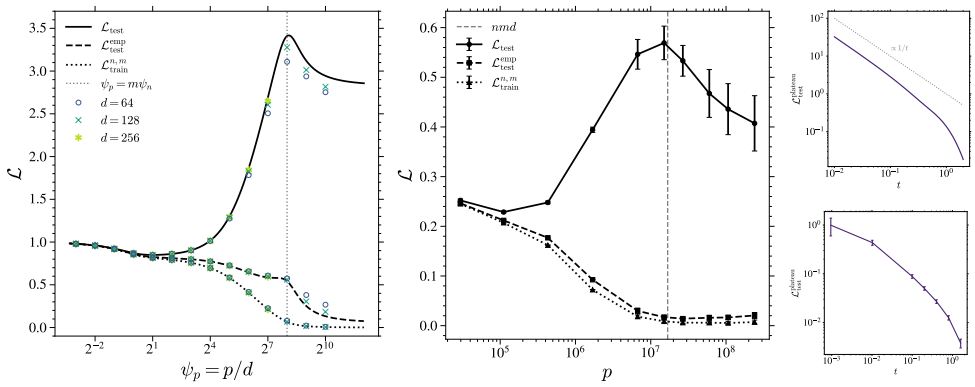

Figure 2: **Double descent in Diffusion Models.** *(Left)* Losses vs. $\psi_p = p/d$ for an RFNN with Gaussian data at $\psi_n = n/d = 16$, $m = 16$, $t = 0.1$, $\lambda = 10^{-3}$; black curves are the predictions of Theorem 2.1, markers the RFNN trained with Adam on Eq. 2 with ridge penalty $\lambda\Delta_t\|\mathbf{A}\|_F^2/(dp)$, averaged over 10 runs. *(Middle)* Losses vs. p for U-Nets trained on CelebA at $n = 2048$, $m = 8$, evaluated at $t \approx 0.01$; error bars are ± 3SE over 6 initial seeds. *(Right)* Large p plateau value of the test loss vs. t for the RFNN *(top)* and the U-Net *(bottom)*, $W = 128$, $p \approx 106$M.

Theorem 2.1 (Asymptotic characterization of the traces)**.** *For $(z, \zeta, \epsilon) \in \mathbb{C}^3$, define the Stieltjes transforms*

$$q(z,\zeta,\epsilon) = \frac{1}{p}\operatorname{Tr}(\mathcal{G}_{z,\zeta,\epsilon})\,, \qquad r(z,\zeta,\epsilon) = \frac{1}{p}\operatorname{Tr}\left(\frac{\mathbf{W}^T}{\sqrt{d}}\,\mathcal{G}_{z,\zeta,\epsilon}\,\frac{\mathbf{W}}{\sqrt{d}}\right). \tag{13}$$

Then, as $d \to \infty$: (i) (q, r) *concentrate and solve a system of algebraic equations (see Appendix C.7);* (ii) $T_1(z,\zeta,\epsilon) = f_1\big(q(z,\zeta,\epsilon),\, r(z,\zeta,\epsilon),\, \epsilon\big)$ *and* $T_2 = f_2\big(q(-\lambda,0,0),\, r(-\lambda,0,0)\big)$*, for explicit functions f_1, f_2 (see Appendix C.8).*

The limiting equations in the overparameterized regime $\psi_p \to \infty$ used in Fig. 2 (top right panel), and in the infinite-noise regime $m \to \infty$, used in Fig. 5, are given in Appendix C.10 and Appendix C.11 respectively. In the following figures, when presenting RFNN asymptotic results, we will also show simulations of the RFNN at finite d to illustrate the convergence toward the asymptotic limit.

2.3 Main Results

Parameter-wise double descent exists in diffusion models. Both the RFNN theory and the U-Net experiments exhibit *double descent*, as shown in Fig. 2. As a function of the model size p, the test loss $\mathcal{L}_{\text{test}}$ first decreases to a minimum and then rises to a peak located where the training loss $\mathcal{L}_{\text{train}}^{n,m}$ vanishes, i.e. at the interpolation threshold, corresponding to the smallest p reaching very small values of the train loss. Past the peak, in the overparameterized regime, $\mathcal{L}_{\text{test}}$ settles at a value larger than at its first minimum; the smaller t, the higher is this plateau, as shown in the right panels of Fig. 2. The empirical test loss $\mathcal{L}_{\text{test}}^{\text{emp}}$, by contrast, tracks the training loss and reaches its minimum in the overparameterized regime. For the RFNN, $\mathcal{L}_{\text{test}}^{\text{emp}}$ exhibits a double descent at moderate m, which fades away as m grows. These results, obtained at fixed t, persist for the test loss integrated over t (Appendix B.7), which also displays a peak and a plateau—as expected, since the integral collects all noise levels and the small-t ones exhibit these features most strongly.

Both n and m shift the interpolation peak. As p increases, the RFNN model exhibits three regimes with distinct scalings, as shown in the three leftmost panels of Fig. 3: *(i)* for $p < n$, the test loss decreases and depends on p alone, independently of n and m (left panel); *(ii)* for $n < p < nm$, $\mathcal{L}_{\text{test}}$ increases and depends only on the ratio p/n (top central panel); *(iii)* the peak height, reached at $p = nm$ (lower central panel), and the plateau beyond it are independent of n (left panel). The threshold $p = nm$ is simply where the pd readout parameters match the nmd scalar conditions imposed by $\mathcal{L}_{\text{train}}^{n,m}$, i.e. one d-dimensional equation per noisy point $\boldsymbol{y}^{\nu\mu}$. For the U-Net, the peak shifts to larger p as both n and m grow, as displayed in the right panel of Fig. 3. The RFNN scalings

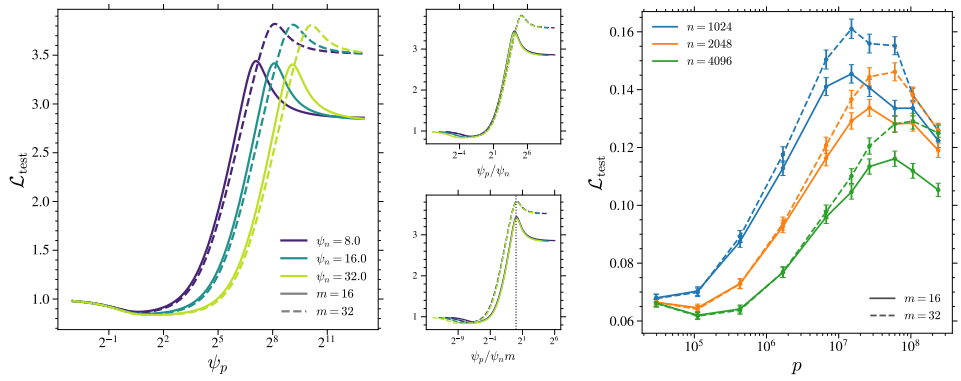

Figure 3: **Effects of n and m on the double descent.** *(Left)* RFNN test loss vs. ψ_p for several ψ_n and m, at $t = 0.1$, $\lambda = 10^{-3}$, $\sigma = \tanh$. *(Middle)* Same, with the x-axis rescaled by ψ_n (top) and $\psi_n m$ (bottom). *(Right)* U-Net test loss vs. p for several m at $t \approx 0.1$; error bars are $\pm$3SE of one seed over test sets.

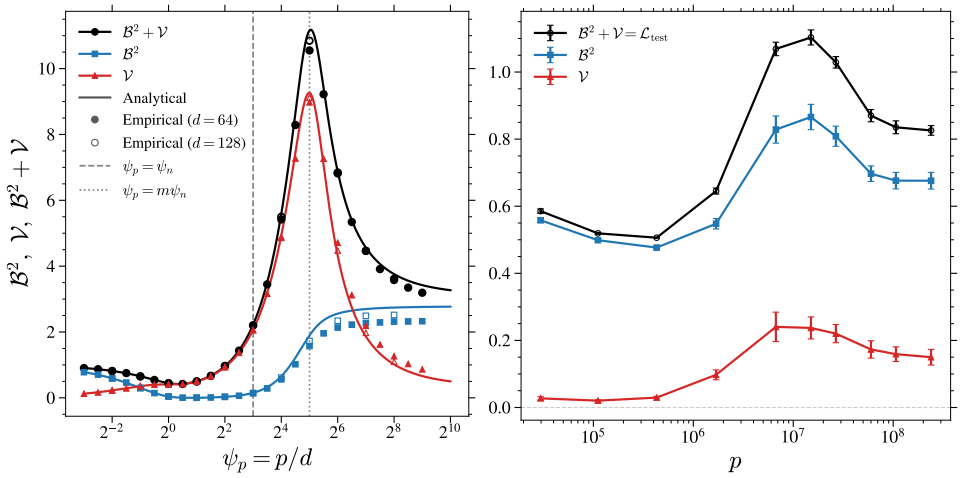

Figure 4: **Bias and variance of the score estimator.** *(Left)* Bias and variance for the RFNN model vs. ψ_p for $\psi_n = 8, t = 0.1, m = 4, \lambda = 10^{-3}, \sigma = \tanh$ and several values of d. Solid curves are the analytical predictions from Theorem 2.1; markers are numerical estimates from 10 independent runs. *(Right)* Bias and variance for U-Net models as a function of p ($n = 2048$, $m = 4$, evaluated at $t \approx 0.01$). Error bars correspond to $\pm$3SE over 10 models trained on disjoint datasets.

carry over to the U-Net across the (n, m, p) range we explore (see Appendix B.2)—the peak is also approximately located where the number of parameters equals nmd—but we do not expect it to hold in general, in particular due to the strong correlations among the nmd targets and to the weight sharing of the convolutional architecture, which decouples p from the effective degrees of freedom.

Malign overfitting and the growth of the bias. A finer picture emerges from the bias–variance decomposition of the test loss from Eq. 4, reported in Fig. 4 for the RFNN (left) and the U-Net (right). For small models the test loss is pure bias; as p grows the bias decreases while the variance builds up which sets the first minimum of the test error, as in the classical bias–variance trade-off (Wasserman, 2004). Approaching the interpolation threshold $p \sim nm$, both terms contribute: the variance diverges, producing the peak, while the bias—rather than continuing to decrease—starts to grow *before* the peak. Past the peak the two terms decouple: the variance decays, as in the classical supervised setting (Neal et al., 2019; D'Ascoli et al., 2020), whereas the bias keeps rising and saturates. The same decomposition for the U-Nets (right panel of Fig. 4) reproduces the key feature: the variance decays after the interpolation peak while the bias increases before peaking. At large

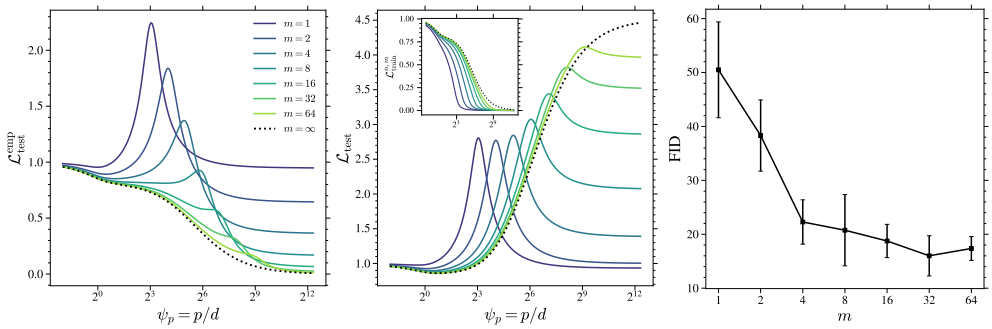

Figure 5: **Effect of m on the double descent.** *(Left)* Empirical test loss $\mathcal{L}_{\text{test}}^{\text{emp}}$ and *(Middle)* test loss $\mathcal{L}_{\text{test}}$ vs ψ_p for an RFNN at several m, with $\psi_n = n/d = 8$, $t = 0.1$, $\lambda = 10^{-3}$, $\sigma = \tanh$; inset: training loss. *(Right)* FID vs m for early-stopped U-Nets trained on CelebA at $n = 16384$, $W = 32$; error bars are $\pm$3SE over 5 seeds.

p, both saturate at a strictly positive plateau.[4] Further supporting experiments on Gaussian mixture data and simpler feedforward network models of the score reproduce this behavior and can be found in Appendix B.6. In all cases, the residual test error at large p is a combination of comparable contributions from the variance and the bias. Both are due to memorization and "benign overfitting" of the empirical score: variance is high because of dependence on the precise drawing of the training set, and the bias is high because the empirical score is different from the population score. The two bias regimes are of different nature. At small p, it is an *approximation error*: the model is too simple to represent the score, so the bias decreases as the capacity grows. At large p, the bias is instead a *property of the objective*: the training drives the estimator towards the minimizers of the empirical test loss, which differ from those of the population one, a gap that no amount of capacity can close.

3 Discussion

From benign to malign: the role of m. We now characterize how m controls the interpolation threshold and the nature of overfitting, focusing on the RFNN setting. Recall that $m = 1$ corresponds to standard regression, where each sample carries a single target, while practice operates at $m \gg 1$ (large number of epochs). For $m = 1$, the middle panel of Fig. 5 shows an interpolation peak in $\mathcal{L}_{\text{test}}$ at $\psi_p = \psi_n$; after the peak, $\mathcal{L}_{\text{test}}$ decreases below its underparameterized minimum: overfitting is here *benign*. Upon increasing m, the interpolation threshold shifts to $\psi_p = m\psi_n$. As m increases, $\mathcal{L}_{\text{test}}$ decreases for $\psi_p < \psi_n m$ but increases for $\psi_p > \psi_n m$: overfitting turns from benign to malign. This effect is already visible at $m = 2$. In the limit $m \to \infty$, the double descent disappears altogether, and $\mathcal{L}_{\text{test}}$ has a single minimum, located in the underparameterized regime (George et al., 2026). The left panel shows the evolution of $\mathcal{L}_{\text{test}}^{\text{emp}}$, which also shows a standard double descent for $m = 1$. Increasing m lowers both the peak height and the loss throughout the $\psi_p > m\psi_n$ regime; for large m the peak fades away and $\mathcal{L}_{\text{test}}^{\text{emp}}$ becomes strictly decreasing, converging to 0 as $\psi_p \to \infty$. Note that already for moderate values of m, $\mathcal{L}_{\text{test}}^{\text{emp}}$, which measures the distance to the empirical score, reaches very small values for large p. Given the harmful effect of increasing m on the test loss, one might be tempted to conclude that it is preferable to work at $m = 1$. In practice, however, one works at $m \gg 1$, i.e. many epochs, and for good reason: once regularized, the model performs better at larger m, as shown in the right panel of Fig. 5: for early-stopped U-Nets trained on $n = 16384$ images with $W = 32$, the FID decreases swiftly as m increases, before saturating. The effect of regularization is discussed further below.

Overparameterize, but regularize. In supervised learning, regularization can drastically alter the double descent picture: optimally tuned ridge regularization is known to flatten the interpolation peak and induce a monotone test error (Nakkiran et al., 2021b; Mei & Montanari, 2019). We now show that the same holds in the diffusion setting, and in particular that overparameterizing is beneficial when paired with regularization. The left panel of Fig. 6 shows that, at small regularization

[4] For the U-Net, the bias dominates the test error at all p. This is expected: the exact score being inaccessible, we estimate it from the target noises, which shifts the bias by a p-independent constant. See Appendix B.6.

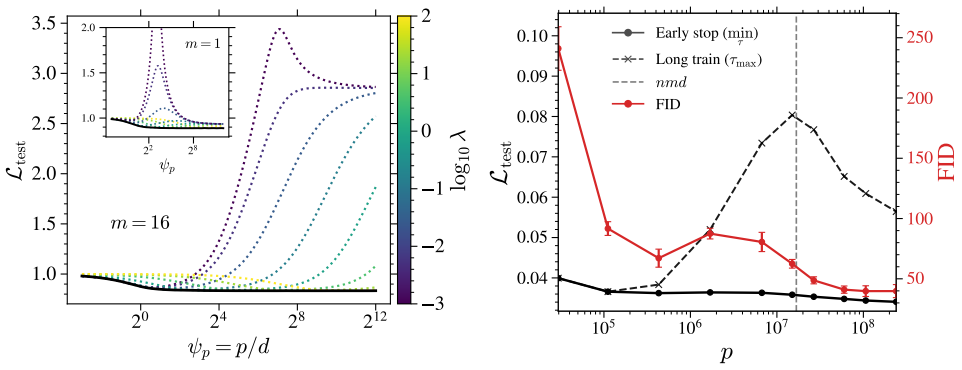

Figure 6: **Benefits of regularization on generalization.** *(Left)* RFNN test loss $\mathcal{L}_{\text{test}}$ vs. ψ_p for several λ, at $\psi_n = 8$, $t = 0.1$, $m = 16$, $\sigma = \tanh$ (inset: $m = 1$); the black curve is the lower envelope. *(Right)* U-Net integrated test loss vs. p at $n = 2048$, $m = 8$, early-stopped (solid) and fully trained (dashed), with the FID of the early-stopped models (red, right axis); the vertical dashed line marks the naive interpolation threshold $p = nmd$.

(dark blue dotted curve), the RFNN still displays malign overfitting. If instead λ is optimized for each p, which yields the lower envelope of the family of curves (solid black line), then $\mathcal{L}_{\text{test}}$ decreases monotonically with ψ_p: an optimally regularized large model outperforms any smaller unregularized one. We discuss the effect of t and the optimal regularization in Appendix C.13.1. For U-Net models trained on CelebA with no explicit regularization, early stopping plays a role analogous to the ridge penalty (Ali et al., 2019). Bonnaire et al. (2025) and Favero et al. (2025) identify two characteristic timescales: a generalization timescale $\tau_{\text{gen}} = O(1)$ and a memorization timescale $\tau_{\text{mem}} = O(n)$, beyond which the model starts to collapse onto individual training samples; stopping within the window $\tau_{\text{gen}} \ll \tau^\star \ll \tau_{\text{mem}}$ yields high-quality samples while avoiding memorization. Our numerical results complement this picture by showing that the test loss of optimally early-stopped models ($\tau^\star = \arg\min_\tau \mathcal{L}_{\text{test}}(\tau)$, right panel of Fig. 6) no longer exhibits the interpolation peak nor the associated double descent. It decreases instead monotonically with the number of parameters. The FID of these early-stopped models does still exhibit a double descent; we conjecture this reflects the test loss being an imperfect proxy for sample quality, and that early-stopping on the FID itself would remove it (see Appendix B.8 for a lengthier discussion). In particular, an overparameterized early-stopped model systematically beats any fully-trained underparameterized one: overparameterization remains beneficial once paired with early stopping.

4 Conclusion, Limitations and Future Work

In this work, we have shown that diffusion models do exhibit *double descent* but with crucial differences from the classical supervised regression setting. The interpolation threshold shifts from $p \sim n$ in regression to $p \sim nm$ in diffusion, where m is the number of noise realizations per training sample. The peak vanishes as $m \to \infty$, but the rise of the test loss beyond its minimum, and the bias mechanism behind it, do not. Since $m \gg 1$ in practice, the peak is pushed to very large model sizes and the overfitting observed in practice, corresponding to the rising branch of a U-shaped curve (Marion & Wu, 2026; Farghly et al., 2026), is the approach to it, where malign overfitting is already at work. A bias–variance decomposition traces this malign overfitting back to a persistent bias of the score estimator, which grows past the interpolation peak and saturates at a plateau. Finally, we establish that an optimally regularized overparameterized model outperforms any unregularized one: overparameterization paired with regularization is beneficial despite malign overfitting.

Limitations & future work. The RFNN and the U-Net differ mainly in the n-dependence of the interpolation peak height and in the bias peak seen for the U-Net, both absent from the RFNN. Clarifying the role of feature learning in these effects is an exciting question we leave for future work. Our empirical validation is moreover restricted to a single dataset, moderate number of samples, and models substantially smaller than the state of the art. Establishing our predictions at scale is certainly an important future step.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion, 2024. URL https://arxiv.org/abs/2410.08727.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/ali19a.html.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and regularization in generative diffusion models, 2025. URL https://arxiv.org/abs/2501.15785.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116:5451–5460, 2019.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907378117.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(9957), 2024. doi: 10.1038/s41467-024-9957-y. URL https://www.nature.com/articles/s41467-024-9957-y.

Antoine Bodin and Nicolas Macris. Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21605–21617. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b4f8e5c5fb53f5ba81072451531d5460-Paper.pdf.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BSZqpqgqM0.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rWW5wdECl8.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborova. Analysis of learning a flow-based generative model from limited sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ndCJeysCPe.

Hugo Cui, Cengiz Pehlevan, and Yue Lu. A solvable model of learning generative diffusion: theory and insights. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 5253–5296. Curran Associates, Inc., 2025. doi: 10.52202/085713-0187. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/082d3d795520c43214da5123e56a3a34-Paper-Conference.pdf.

Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2280–2290. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/d-ascoli20a.html.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where and why do they appear?*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124002, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3909. URL http://dx.doi.org/10.1088/1742-5468/ac3909.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Tyler Farghly, Benjamin Dupuis, Alain Oliviero Durmus, and Umut Simsekli. Benign overfitting does not occur in diffusion models. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://openreview.net/forum?id=QwP5eaXJTj.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models, 2025.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020. URL https://arxiv.org/abs/1909.11572.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=ZnplHm2uRt.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3452–3462. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gerace20a.html.

Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *IEEE Transactions on Information Theory*, 69:1824–1852, 2023.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan S. Hesthaven, and Lenka Zdeborová (eds.), *Mathematical and Scientific Machine Learning, 16-19 August 2021, Virtual Conference / Lausanne, Switzerland*, volume 145 of *Proceedings of Machine Learning Research*, pp. 426–471. PMLR, 2021. URL https://proceedings.mlr.press/v145/goldt22a.html.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=D3DBqvSDbj.

Francesco Guerra and Fabio Lucio Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230:71–79, 2002.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

U.G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4): 1188–1205, 1986. doi: 10.1214/aop/1176992362.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila. Elucidating the design space of diffusion-based generative models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

W. F. Kibble. An extension of a theorem of mehler's on hermite polynomials. *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1):12–15, June 1945. ISSN 0305-0041, 1469-8064. doi: 10.1017/S0305004100022313.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Hugo Latourelle-Vigeant, Sinho Chewi, Aram-Alexandre Pooladian, John Sous, and Theodor Misiakiewicz. Generalization, memorization, and overfitting for diffusion models trained in the lazy high-dimensional regime, 2026. URL https://arxiv.org/abs/2608.23938.

Steve Lawrence, C. Lee Giles, and Ah Chung Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 540–545. AAAI Press, 1997.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model, 2024. URL https://arxiv.org/abs/2401.04856.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Pierre Marion and Yu-Han Wu. Understanding diffusion models requires rethinking (again) generalization, 2026. URL https://arxiv.org/abs/2605.06077.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 2019. URL https://api.semanticscholar.org/CorpusID:199668852.

Claudia Merger and Sebastian Goldt. Generalization dynamics of linear diffusion models, 2026. URL https://arxiv.org/abs/2505.24769.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, volume 9 of *Lecture Notes in Physics*. World Scientific Publishing Company, Singapore, 1987.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, dec 2021a. doi: 10.1088/1742-5468/ac3a74. URL https://doi.org/10.1088/1742-5468/ac3a74.

Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations (ICLR)*, 2021b. URL https://arxiv.org/abs/2003.01897.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. URL https://openreview.net/forum?id=B1guPVr2h4.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (ICLR) Workshop*, 2015. URL https://arxiv.org/abs/1412.6614.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/0f3d014eead934bbdbacb62a01dc4831-Paper.pdf.

S. Péché. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019. doi: 10.1214/19-ECP255. URL https://doi.org/10.1214/19-ECP255.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Michel Talagrand. The parisi formula. *Annals of mathematics*, pp. 221–263, 2006.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion, 2025. URL https://arxiv.org/abs/2410.05898.

M. Vilucchio, Y. Dandi, M. P. Rossignol, C. Gerbelot, and F. Krzakala. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=shciCbSk9h.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Double Descent and Malign Overfitting in Diffusion Models
Appendix

This appendix provides detailed derivations and additional results supporting the main text (MT). Appendix A expands the discussion of closely related works (George et al., 2026; Latourelle-Vigeant et al., 2026; Marion & Wu, 2026; Farghly et al., 2026). Appendix B provides further details about the numerical experiments carried out in Sect. 2.1, as well as additional experiments and discussions. Appendix C gives formal proofs of the main theorems from Sect. 2.2, together with additional results and discussions. We also discuss LLM usage in the conception of this work in Appendix D.

A Discussion of related works

We discuss here in detail four works close to ours: the precise learning curves of George et al. (2026), the lazy regime dynamics of Latourelle-Vigeant et al. (2026), the empirical study of Marion & Wu (2026), and the impossibility result of Farghly et al. (2026). The four tackle a similar question along complementary axes. The common thread of our comparison is that all four works operate either at $m = 1$ or at $m = \infty$, whereas the phenomenology we report—the interpolation peak at $p \sim nm$, the benign-to-malign transition, and the bias mechanism behind it—require to understand the finite-m interpolation between these two limits.

Precise learning curves for the RFNN score. George et al. (2026) derive asymptotically exact train and test errors using the Gaussian Equivalence Principle and the theory of linear pencils for the same framework as ours: nonlinear RFNN and Gaussian data, studied in the proportional regime of finite $\psi_n = n/d$ and $\psi_p = p/d$. Their central message is a crossover at $p \sim n$ between a regime where the model generalizes and a regime where it approaches the empirical score and memorizes, showing that increasing m both improves generalization for $p \leq n$ and intensifies memorization for $p \geq n$. First, our arbitrary-m solutions reveal the intermediate regime with an interpolation threshold sitting at $p \sim nm$. Second, our dissection of the problem into three losses ($\mathcal{L}_{\text{train}}^{n,m}$, $\mathcal{L}_{\text{test}}^{\text{emp}}$, $\mathcal{L}_{\text{test}}$) identifies what is actually being overfit and the origin of the malign overfitting.

Training and generative dynamics in the lazy regime. Latourelle-Vigeant et al. (2026) study the gradient-flow dynamics of neural networks trained on the score-matching objective in the proportional regime $n \asymp d$, in the lazy regime where the network is linearized around its initialization. The infinite-width limit then replaces the network by a kernel operator $\mathcal{K}$ acting on a vector-valued RKHS,

$$(\mathcal{K}f)(\mathbf{x}) = \int \mathrm{d}\boldsymbol{y}\; p_t^{\text{emp}}(\boldsymbol{y})\, K(\mathbf{x}, \boldsymbol{y})\, f(\boldsymbol{y}), \tag{14}$$

with p_t^{emp} the noised empirical distribution. They derive closed-form equations for the training and test losses along the whole trajectory. Their central message is a separation of training timescales: the model generalizes on $\tau = O(d)$, where $\mathcal{L}_{\text{test}}$ decreases; it then overfits the empirical score on $\tau = d^{1+\Theta(1)}$, where $\mathcal{L}_{\text{test}}$ rises again; and it memorizes only on the much later $\tau = d^{\omega(1)}$. Because they also characterize the generated distribution, they can separate these last two stages and conclude that overfitting the empirical score during training does not necessarily entail memorization in the backward dynamics. The differences with our work are twofold. First, they follow the training and generative dynamics, whereas we characterize only the empirical risk minimizer solution. Second, their lazy limit sits at $p \to \infty$ with an infinite number of noise samples $m \to \infty$, whereas our analysis holds at finite p and m.

Empirical study of memorization dynamics. Marion & Wu (2026) perform an extensive numerical study of U-Nets trained on CIFAR-10 subsets at various n and p and track their train and test denoising losses. They never observe a parameter-wise peak or double descent in the test loss. Crucially, a fresh noise realization is drawn at every optimization step, so their setting corresponds to our $m \to \infty$ case. Consistently with our findings, there is therefore no peak to see as p increases

in this case. What they do observe however is a training-time-wise double descent of the distributional distances with a peak located at the onset of memorization, which is accelerated by the model size. This is consistent with our τ-W discussion reported in Appendix B.3. We emphasize that the two double descents are distinct phenomena occurring on different axes and with different mechanisms: theirs is dynamical, observed at $m \to \infty$, and only appears in *distribution space*; ours is static, parameter-wise, exists only at finite m, and appears in the test loss. Their position that the field should rethink generalization in diffusion models is precisely what our two distinct test losses formalize: the benign overfitting mechanism *does exist*, but it targets the minimizers of $\mathcal{L}_{\text{test}}^{\text{emp}}$, the empirical score, rather than those of $\mathcal{L}_{\text{test}}$.

Impossibility of benign overfitting. Farghly et al. (2026) establish, in a model-independent way, that overfitting and generalization cannot coexist in diffusion models unless the sample size grows exponentially with the data dimension. Their first argument is information-theoretic: as $t \to 0$, the relative Fisher information between the empirical and population noised distributions diverges, which is consistent with our picture as it forces the test loss to be large. They also study a linear RFNN and find a U-shaped test loss as p increases. Our analysis refines their conclusions in two ways. First, their setting is again $m \to \infty$: at any finite m, double descent *does* occur, with an interpolation peak at $p \sim nm$ that moves to infinite model size as $m \to \infty$, recovering their U-shape. Along this crossover, we show that overfitting is benign only in the regression case $m = 1$ and increasingly malign as m grows. The impossibility that they derive is thus the $m \to \infty$ endpoint of a benign-to-malign crossover. Second, their finding that time-integrated objective and early-stopping act as implicit regularizers parallels our regularization analysis, to which we add that, once suitably regularized, overparameterization becomes strictly beneficial. Interestingly, they also trace the difference with standard regression to an alignment obstruction: benign overfitting of the minimum-norm interpolator relies on a favorable alignment between the target and the empirical covariances, which the score-matching target cannot reach. This is complementary to our bias mechanism: both analyses locate the malign origin of the overfitting in a *deterministic* distortion of the learned score.

B Additional numerical results

B.1 Architecture & Training

The model is made of an initial 3×3 convolution mapping to W channels. The encoder has 4 resolution levels with channel multipliers $(1, 2, 4, 4)$, meaning the successive width is $W, 2W, 4W, 4W$ at resolutions $L, L/2, L/4, L/8$. Each level is composed of 2 residual blocks with 3×3 stride-2 convolutions for downsampling. The encoder is followed by a bottleneck of 2 residual blocks and then by a symmetric decoder using nearest-neighbor upsampling and traditional skip connections. A final group-norm convolution projects back to a single channel. We apply multi-head self-attention with 4 heads at the two coarsest levels and in the bottleneck. Each residual block uses GroupNorm and SiLU activations. Finally, the diffusion time is encoded with a standard sinusoidal embedding passed through a 2-layer MLP and added to the features in every residual block. At very small $W = 2$, the attention head count and GroupNorm groups are reduced so the layers remain valid, hence the configuration slightly changes at this W compared to the others.

As stated in the main text, each model is trained to predict the noise $\boldsymbol{\xi}^{\nu\mu}$ by minimizing the standard noise-prediction objective. We use Adam with default moment parameters and without weight decay or exponential moving average. The forward pass is discretized using $T = 1000$ steps, indexed by $k = \{0, \dots, T-1\}$, with a linear noise schedule β_k going from $\beta_0 = 10^{-4}$ to $\beta_{T-1} = 2 \times 10^{-2}$ and $\bar{\alpha}_k = \prod_{j=0}^{k}(1-\beta_j)$ (Ho et al., 2020). To connect this schedule to the continuous-time notation of the main text, we define $t_k = -\frac{1}{2}\log\bar{\alpha}_k$. Consequently, the noisy samples can equivalently be written as $\mathbf{x}_{t_k}^{\nu\mu} = \sqrt{\bar{\alpha}_k}\,\mathbf{x}^{\nu} + \sqrt{1-\bar{\alpha}_k}\,\boldsymbol{\xi}^{\nu\mu}$. Throughout training, the learning rate is fixed to 10^{-4} with a batch size $B = 128$. Unlike standard diffusion models training, the noise is here fixed: each of the n images $\mathbf{x}^{\nu}$ has m frozen noise vectors $\boldsymbol{\xi}^{\nu\mu}$, giving a total of $n \times m$ fixed pairs. The training is time-dependent: at every step we pick a diffusion index $k \sim \mathcal{U}[1, \dots, T-1]$ and we compute the interpolated noisy sample at t_k. In the numerical experiments, no explicit regularization is used: the only regularizer we study is the finite training time τ through early-stopping.

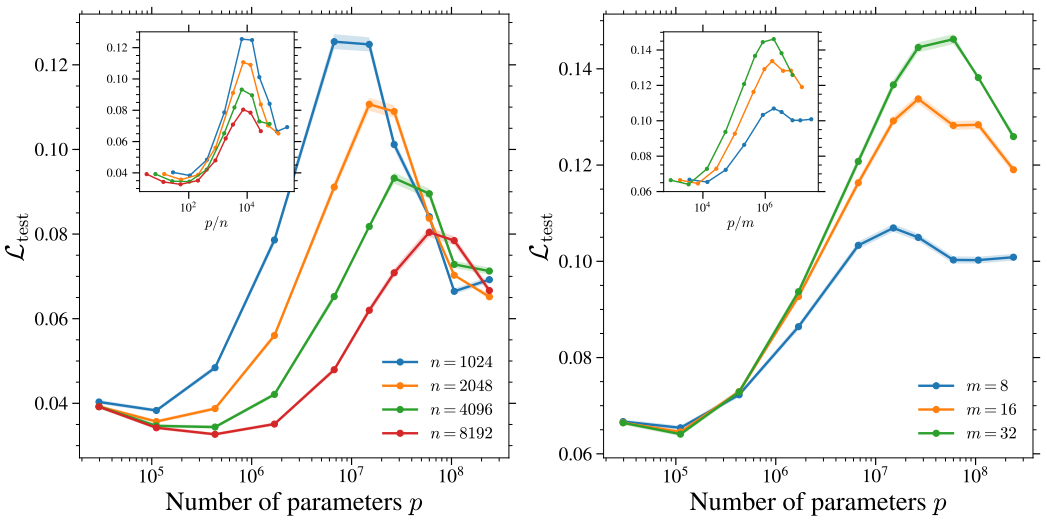

Figure 7: **Additional scaling experiments.** Evolution of *(Left)* the time-integrated test loss with p for several n at $m = 32$, and *(Right)* the $t \approx 0.1$ test loss with p for several m at $n = 2048$. Both plots are obtained at $\tau = \tau_{\max} = 2$M steps.

B.2 More scaling experiments of the double descent with n and m

In the right panel of Fig. 3, we show the evolutions of the test loss for several n and m at once. In Fig. 7 we provide complementary experiments with p for more n and m values, shown individually. In this case, we clearly see the collapse of the (time-integrated) test loss at both small and large p, independently of n in the left panel, with the clear scaling of the peak with n in the inset. The right panel shows three values of m for a fixed $n = 2048$ training. At small p, all curves collapse independently of m while the height of the large p plateau is increasing with it. The inset also shows the p/m evolution, showing the scaling of the peak with m as well. Together, these plots emphasize the scaling of the double descent peak with nm, as in the random features case.

B.3 Training-time-wise double descent in U-Net diffusion models

τ–W coupling of the double descent. The left panel of Fig. 8 traces the interaction between model size (through the width W) and training time in the double descent phenomenon: each curve shows $\mathcal{L}_{\text{test}}$ as a function of W at fixed number of updates τ. A double descent emerges only for large enough τ, and its peak is not static, slipping towards smaller W as τ grows. This couples W and τ as two substitutable axes of the double descent phenomenon, echoing the view of the supervised learning setting from Nakkiran et al. (2021a) in which training time and model size feed a single effective complexity measure governing the peak location, although our analysis does not provide any insight on the nature of the origin of these two peaks.

Training-time double descent. In the middle panel of Fig. 8, we show the evolution of the test loss with the number of Adam updates τ for different W. It reveals three regimes. For small-size models, the test loss decreases monotonically to a minimum value reached at the end of training. For intermediate-size models ($16 \leq W \leq 64$) the test loss reaches a minimum and then rises again at a moderate number of updates, the signature of overfitting and memorization at large training times. Larger models ($W \geq 96$), however, exhibit a *training-time-wise double descent*: the test loss quickly decreases to a minimum ($\tau \approx 10^4$), rises, and then decreases again at even larger τ. Strikingly, the first two regimes with the descent and the subsequent rise are largely robust to W with all the curves collapsing, whereas the second descent clearly depends on the width, with larger models exhibiting a sharper late-time decrease. The right panel actually informs us about the mechanism behind this rise at large W: the gap between $\mathcal{L}_{\text{test}}^{\text{emp}}$ and $\mathcal{L}_{\text{train}}^{n,m}$ grows with both τ (at fixed W) and W (at fixed τ), showing that the model progressively fits the specific training noise realizations rather than the underlying score, meaning it is heading towards memorization and the overfitting of $\mathcal{L}_{\text{test}}^{\text{emp}}$

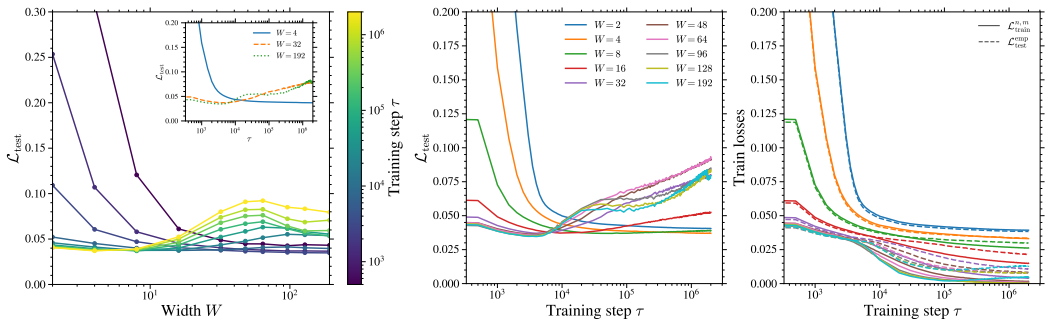

Figure 8: **Training-time double descent in diffusion models.** *(Left)* Evolution of the integrated test loss with W for several values of training times τ. The middle and right panels show the evolution of the time-integrated *(Middle)* test and *(Right)* train losses against τ for several W. All the models were trained with $n = 2048$ and $m = 8$.

is detrimental. These three regimes were also observed in the supervised setting (Nakkiran et al., 2021a) but with a crucial difference: there, the large-model second descent is persistent so that longer trainings decrease the test loss, whereas in diffusion models, every sufficiently expressive model eventually turns back towards memorization (Bonnaire et al., 2025). This marks a clear departure from the epoch-wise double descent in supervised learning.

B.4 Visual inspection of the generated samples

Sample generation. For each configuration (n, m, W), we generate from the early-stopped model at $\tau^\star = \arg\min_\tau \mathcal{L}_{\text{test}}(\tau)$ that minimizes the time-integrated test loss. From this checkpoint, we draw 10^4 samples with the deterministic DDIM sampler from Song et al. (2021a) ($\eta = 0$, $T' = 100$ steps). All models are initialized with the same Gaussian noise seed so that a given position in the grid of Fig. 9 corresponds to the same starting point for all models. Samples are then de-normalized to the original range and the quality is measured using the Fréchet-Inception Distance (FID, Heusel et al., 2017) against 10^4 CelebA held-out test images that no models have seen during training, and using the `PyTorch-FID` package[5].

Results. In Fig. 9, we show for $n = 2048$ a 3×3 set of generated samples from the early-stopped models at different m and W, with increasing m over the rows and W across the columns. The FID of the full generated set is reported in each panel. Since all panels share the same initial noise, it is easy to visually appreciate the quality variations with m and W. Two trends are visible. First, increasing the number of noise realizations m improves the sample quality up to a certain point: the $m = 1$ row is clearly the blurriest (FID$\approx$ 100), whereas $m = 8$ and $m = 32$ produce sharper faces at fixed W, although the trend is not consistent for all widths. Second, at fixed m (say 8), the FID decreases with the width, reflecting the benefit of additional capacity coupled with regularization, as emphasized in the main text. The only notable exception is $m = 1$ where the largest model is *not* the best one. Note that all models shown here are trained with only $n = 2048$ images with no data augmentation, hence the modest sample quality. This is intentional to keep n moderate so that the generalization–memorization transition during training (Bonnaire et al., 2025) and the interpolation peak in the test loss remain visible in a reasonable amount of training time and number of parameters.

B.5 More details on the effect of m on U-Nets

In Fig. 10 we display the evolution of the time-integrated test loss with the training step τ for several m at fixed $n = 16384$ and $W = 64$. We deliberately take n large so that the models reach good generalization and operate in a regime closer to practical settings. Every curve follows the dynamics discussed previously: the test loss first decreases to a minimum and then rises again, making early stopping necessary, even at this n. The effect of m is clear and monotonic: as m increases, the

[5] Available at https://github.com/mseitzer/pytorch-fid.

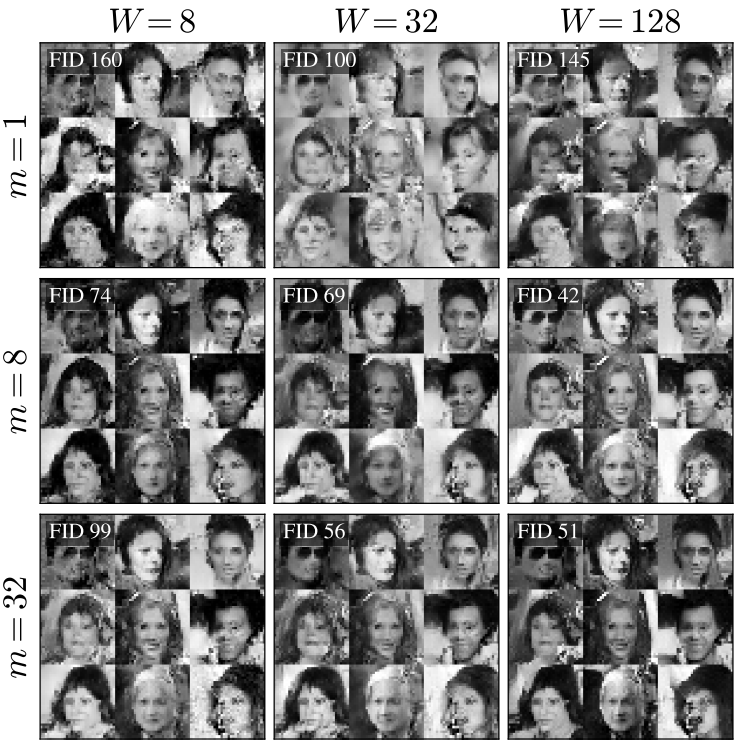

Figure 9: **Non-curated generated samples.** Example of CelebA generated samples for various combinations of m (across rows) and W (across columns) along with the test FID on the full set of generated images. All the models were trained with the same $n = 2048$ data.

minimum test loss decreases and is reached at progressively larger τ. This improvement saturates for $m \gtrsim 16$, where the curves essentially collapse onto one another. Together with the FID measurements in the right panel of Fig. 5, this confirms that a larger m is beneficial for generalization, both at the level of the test loss and of the quality of the generated samples.

B.6 Bias–Variance decomposition in the Gaussian Mixture Model

Data and exact score. We use here a symmetric two-component mixture model such that

$$\mathbf{x}_0 \sim \tfrac{1}{2}\,\mathcal{N}(+\boldsymbol{\mu}, \boldsymbol{I}_d) + \tfrac{1}{2}\,\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{I}_d), \qquad \boldsymbol{\mu} = \mu\,\mathbf{1}_d, \quad \mu = 1, \;\; \boldsymbol{\Sigma} = \boldsymbol{I}_d. \tag{15}$$

In this case, the analytical (population) score is known at all t and corresponds to

$$\boldsymbol{s}^\star(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = \sqrt{\bar{\alpha}_t}\,\boldsymbol{\mu}\,\tanh\!\big(\sqrt{\bar{\alpha}_t}\,\boldsymbol{\mu}^\top \mathbf{x}\big) - \mathbf{x}. \tag{16}$$

Model and training. We learn the score at a single, fixed diffusion time t with a fully-connected network of L hidden layers of width W and $\tanh$ activations. With $\boldsymbol{h}_0 = \mathbf{x}$ and

$$\boldsymbol{h}_\ell = \tanh\!\big(\mathbf{W}_\ell\,\boldsymbol{h}_{\ell-1} + \boldsymbol{b}_\ell\big), \quad \ell = 1, \dots, L, \qquad \mathbf{W}_1 \in \mathbb{R}^{W\times d}, \;\; \mathbf{W}_\ell \in \mathbb{R}^{W\times W}\;(\ell \geq 2), \tag{17}$$

the noise prediction is the linear final read-out

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}_{L+1}\,\boldsymbol{h}_L + \boldsymbol{b}_{L+1}, \qquad \mathbf{W}_{L+1} \in \mathbb{R}^{d\times W}, \tag{18}$$

so that the model has a total of $p = 2dW + d + LW + (L-1)W^2$ trainable parameters. The network predicts the noise, equivalently the score $\boldsymbol{s}_{\boldsymbol{\theta}} = -\boldsymbol{\epsilon}_{\boldsymbol{\theta}}/\sqrt{1-\bar{\alpha}_t}$, and is time-agnostic since t is fixed to the discrete step $t = 137$, which corresponds to $t \approx 0.1$ in the continuous-time notation of the main text. Following the benign-overfitting protocol of the main text, we draw n clean points $\{\mathbf{x}_0^i\}_{i=1}^n$ from P_0 and, for each, m *fixed* noise realizations $\{\boldsymbol{\epsilon}^{ij}\}_{j=1}^m$, yielding nm anchors $\mathbf{x}_t^{ij} = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0^i + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}^{ij}$, and minimize the empirical denoising loss

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\big\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t^{ij}) - \boldsymbol{\epsilon}^{ij}\big\|^2 \tag{19}$$

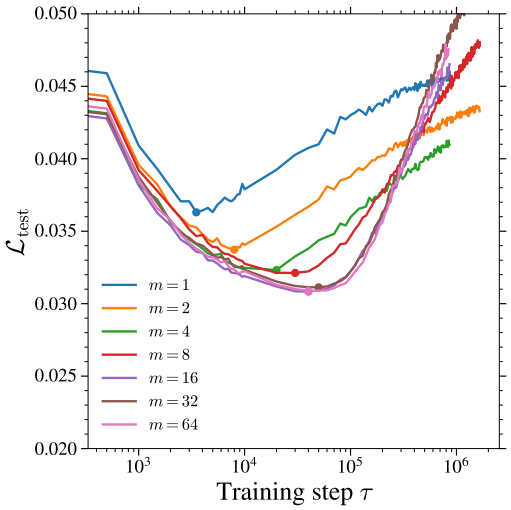

Figure 10: **Effect of m on the test loss dynamics.** Time-integrated test loss as a function of τ at fixed $n = 16384$ and $W = 64$ for several values of m. The colored dots mark the per-curve minima, i.e. the early-stopped checkpoints $\tau^\star$ used in the main text.

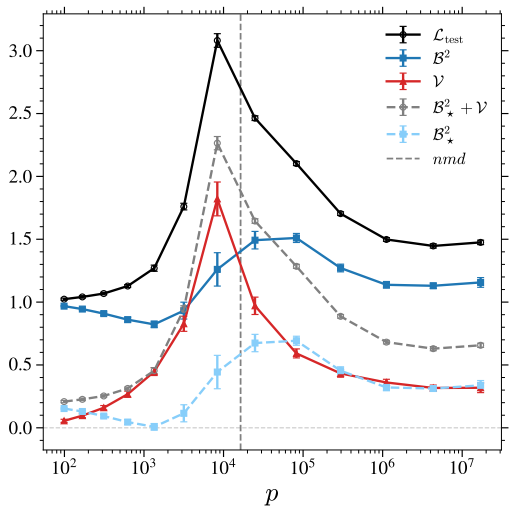

Figure 11: **Bias–variance decomposition for Gaussian Mixture Model.** Evolution of the bias and variance for a 2-hidden-layer neural network trained on data from a 2-component Gaussian Mixture Model. $\mathcal{B}_\star$ is the population bias computed from the analytically-known score while $\mathcal{B}$ is the one accessible in practice. Here, $n = 64$, $m = 8$, $d = 32$, and $t \approx 0.1$.

with Adam for $\tau_{\max} = 100$K steps.

Estimating $\mathcal{B}^2$ and $\mathcal{V}$. We train an ensemble of $K = 20$ independent models that differ only in their random seed (data, noises, and initialization), and evaluate all of them on a single frozen test set of nm fresh points $\{\mathbf{x}_t^\nu\}$ from the same mixture at the same fixed t. We fix $L = 2$ and vary only the width W to increase the number of parameters in the model. We then estimate empirically the bias, variance and test loss. The bias is computed against two targets: the empirical labels $\boldsymbol{\epsilon}$ (i.e. the target noises) giving the quantities $\mathcal{B}^2$ and $\mathcal{L}_{\text{test}} = \mathcal{B}^2 + \mathcal{V}$ as measured in main text; and against the population target $\boldsymbol{\epsilon}^\star$ from Eq. 16, granting access to population estimates of $\mathcal{B}_\star^2$ and $\mathcal{B}_\star^2 + \mathcal{V}$.

Empirical vs. population targets. Fig. 11 reports both decompositions. The variance is target-independent, and the two biases differ only by the *irreducible* denoising error

$$\mathcal{B}^2 - \mathcal{B}_\star^2 = \mathbb{E}\left[\left\|\boldsymbol{\epsilon}^\star - \boldsymbol{\epsilon}\right\|^2\right], \tag{20}$$

i.e., a p-independent constant. Consequently, the population curves $\mathcal{B}_\star^2$ and $\mathcal{B}_\star^2 + \mathcal{V}$ (dashed lines in Fig. 11) exactly correspond to the empirical $\mathcal{B}^2$ and $\mathcal{L}_{\text{test}}$ (solid lines) shifted downward by $\mathcal{B}^2 - \mathcal{B}_\star^2$. Beyond this effect, we observe that the plot reproduces exactly the non-monotonic behaviors of the bias and the variance that were depicted in the main text for the U-Net architecture trained on CelebA, both saturating to a non-vanishing $O(1)$ value at large p. The interpolation threshold, corresponding to where the train loss becomes small and the test loss peaks, is also consistent with $p \approx nmd$.

B.7 Integrated test loss

The main text reports losses at a fixed diffusion time t, which isolates the phenomenon at the noise levels where it is most pronounced. However, the double descent and the peak-and-plateau structure appear across the whole small-t range. Figure 12 displays the integrated test loss over all diffusion times[6] as a function of the model size, for the same model and settings as the middle panel of

[6]Note that many figures from the Appendix (e.g., Fig. 7 (Left), Fig. 8, Fig. 10 and Fig. 13) in fact display the integrated loss.

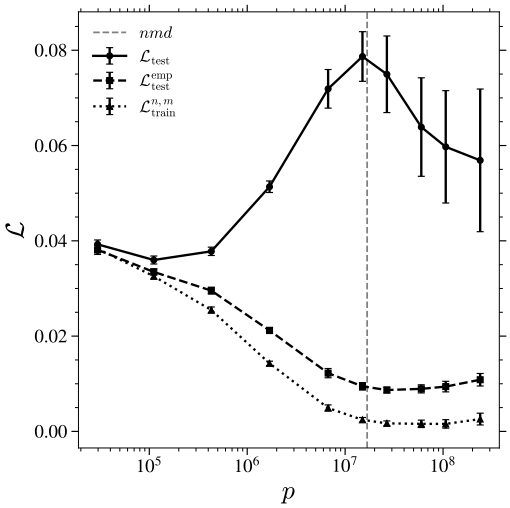

Figure 12: **Integrated losses behavior.** Losses versus p for U-Net models trained on CelebA at fixed $n = 2048$, $m = 8$. Error bars correspond to ± 3SE obtained from 6 random initial seeds. The vertical dashed line corresponds to the naive interpolation threshold nmd.

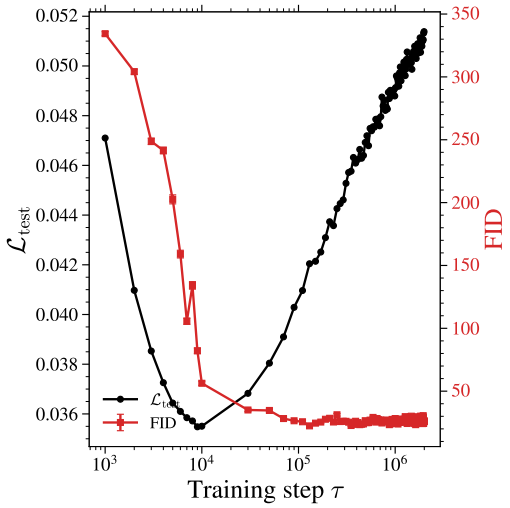

Figure 13: **Test loss and FID with τ.** Time-integrated test loss (black, left axis) and FID (red, right axis) as a function of τ at fixed $n = 2048$, $m = 8$ and $W = 16$. Throughout the training, the evaluation of f_{mem} stays below 10^{-3}.

Fig. 2. The peak is found at around the same value of p and then plateaus above its minimum. The qualitative picture we draw in the main text is therefore robust to aggregating over diffusion times, although the peak amplitude and the plateau may depend on the precise loss parameterization and the weights associated to diffusion times.

B.8 Test loss and FID during training

Figure 13 shows the evolution of the time-integrated test loss and of the FID as a function of the training time τ for one of the models used in the right panel of Fig. 6 at $n = 2048$, $m = 8$ and $W = 16$ ($p \approx 1.8$M). We select this model in particular because it corresponds to the first point at which the FID of the early-stopped models increases in Fig. 6. The two early-stopping criteria clearly disagree: while the test loss selects a finite $\tau^\star = \text{argmin}_\tau \mathcal{L}_{\text{test}}(\tau)$ (used in the main text), the FID decreases monotonically during training and its own optimum lies at long training times. This supports the conjecture of Sect. 3 that early-stopping on the FID itself would remove the FID double descent as the peak value ($W = 16$) would now sit below the FID measure of $W = 8$. The two criteria (test loss and FID) do not optimize the same quantities as the FID only measures similarities in the features of the training and generated samples and is therefore insensitive to memorization: a model copying exactly its training set would get a good FID. The test loss, however, penalizes the deviation of the learned score from the population one, and therefore increases as the estimator starts to correlate with the training set. Stopping at $\tau^\star$ therefore minimizes this precise criterion and remains far cheaper to monitor.[7] In the precise configuration of Fig. 13 (small m and W), the memorization fraction (as measured by the ratio of first to second nearest neighbor in the L_2 sense being below $1/3$, see Yoon et al., 2023; Bonnaire et al., 2025) remains below 10^{-3}: starting from random noise does not lead to memorized copies of training samples.

C Proof of the analytical results and additional results

In this section of the Appendix we first present some preliminary notions on random matrix theory, then we present the derivation of the results of the main text, namely the closed-form losses and

[7] Evaluating the FID requires generating many samples at different training times, whereas the test loss only computes a few forward passes on a test set.

bias–variance decomposition of Theorem 2.1, together with the overparameterized ($\psi_p \to \infty$) and infinite noise ($m \to \infty$) limits, which we state and prove here (Proposition C.3 and Proposition C.5).

C.1 Preliminaries

In this subsection, we present some useful results and formulae of Random Matrix Theory and Gaussian calculus.

C.1.1 Resolvent and Stieltjes transform

For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ and z outside its spectrum, we define its *resolvent* as the matrix $\mathcal{G}_N = (\mathbf{M} - z\boldsymbol{I})^{-1}$ and its *Stieltjes transform* as

$$g_N(z) = \frac{1}{N}\operatorname{Tr}\big(\mathbf{M} - z\boldsymbol{I}\big)^{-1} = \int \frac{\mathrm{d}\rho_N(\lambda)}{\lambda - z}, \tag{21}$$

where ρ_N is the empirical spectral distribution of $\mathbf{M}$. Although $g_N(z)$ is a random variable, typically, in the high-dimensional limit $N \gg 1$ it converges almost surely to a deterministic limit $\bar{g}(z)$.

C.1.2 Gaussian Toolkit

When computing the traces, we will use usual formulae of Gaussian calculus,

1. For symmetric positive-definite $\mathbf{M} \in \mathbb{R}^{N\times N}$ and $\mathbf{J} \in \mathbb{R}^N$,
$$\int_{\mathbb{R}^N} \mathrm{d}\phi\, e^{-\frac{1}{2}\phi^T\mathbf{M}\phi + \mathbf{J}^T\phi} = (2\pi)^{N/2}(\det\mathbf{M})^{-1/2}\, e^{\frac{1}{2}\mathbf{J}^T\mathbf{M}^{-1}\mathbf{J}}. \tag{22}$$
2. Let $\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ in $\mathbb{R}^N$ and let $\mathbf{K} \in \mathbb{R}^{N\times N}$ be symmetric with $\boldsymbol{I} + \boldsymbol{\Sigma}\mathbf{K} \succ 0$. Then
$$\mathbb{E}\big[e^{-\frac{1}{2}\mathbf{X}^T\mathbf{K}\mathbf{X}}\big] = \det\big(\boldsymbol{I} + \boldsymbol{\Sigma}\mathbf{K}\big)^{-1/2}. \tag{23}$$
3. (Wick's theorem) For jointly centered Gaussian variables $X_1, \dots, X_{2k}$,
$$\mathbb{E}[X_1 \cdots X_{2k}] = \sum_{\text{pairings } \pi} \prod_{(i,j)\in\pi} \mathbb{E}[X_i X_j], \tag{24}$$
the sum running over all perfect matchings. In particular
$$\mathbb{E}[X_1X_2X_3X_4] = \mathbb{E}[X_1X_2]\mathbb{E}[X_3X_4] + \mathbb{E}[X_1X_3]\mathbb{E}[X_2X_4] + \mathbb{E}[X_1X_4]\mathbb{E}[X_2X_3]. \tag{25}$$

C.1.3 The Replica Method

In our context, we frequently need to compute the expected logarithm of a partition function $\mathcal{Z}$ that arises from the Gaussian integral representation of the resolvent, e.g., $\mathcal{Z} \propto \int \mathrm{d}\phi\, e^{-\frac{1}{2}\phi^T\mathbf{M}\phi}$, where $\mathbf{M}$ is a random matrix of interest. To compute such expectations of logarithms of random variables, we employ the replica method (Mézard et al., 1987). The core idea is based on the algebraic identity:

$$\log\mathcal{Z} = \lim_{s\to 0}\frac{\mathcal{Z}^s - 1}{s}. \tag{26}$$

By evaluating the integer moments $\mathbb{E}[\mathcal{Z}^s]$ for $s \in \mathbb{N}$, one can analytically continue the result to $s \to 0$. This procedure is a cornerstone of the replica method from statistical physics, a theoretical physics technique that has had a wide range of applications, notably in statistical learning and the theory of neural networks (D'Ascoli et al., 2020). The replica method is widely believed to yield exact asymptotic predictions, as has been established rigorously in a wide range of problems (Guerra & Toninelli, 2002; Talagrand, 2006; Barbier et al., 2019; Gerbelot et al., 2023; Vilucchio et al., 2025). More precisely, our results rely on a so-called replica-symmetry (RS) assumption. The RS ansatz is known to hold for the trace of rational polynomials of random matrices, where it has been shown to yield the exact same results as rigorous methods such as linear pencils (Bodin & Macris, 2021; George et al., 2026; Bonnaire et al., 2025).

C.1.4 REPLICA REPRESENTATION OF AN INVERSE MATRIX

To obtain Gaussian integrals we use the "replica" representation of the element $(\alpha\beta)$ of a $p \times p$ matrix $\mathbf{M}$:

$$\left(\mathbf{M}^{-1}\right)_{\alpha\beta} = \lim_{s\to 0} \int \left(\prod_{a=1}^{s} \prod_{\gamma=1}^{p} \mathrm{d}\phi_\gamma^a \right) \phi_\alpha^1 \, \phi_\beta^1 \, \exp\!\left(-\tfrac{1}{2}\, \phi_\gamma^a \mathbf{M}_{\gamma\delta} \phi_\delta^a\right). \tag{27}$$

Indeed, using the Gaussian integral representation of the inverse of $\mathbf{M}$,

$$\left(\mathbf{M}^{-1}\right)_{\alpha\beta} = \mathcal{Z}^{-1} \int \left(\prod_{\gamma=1}^{p} \mathrm{d}\phi_\gamma \right) \phi_\alpha \phi_\beta \, \exp\!\left(-\tfrac{1}{2}\, \phi_\gamma \mathbf{M}_{\gamma\delta} \phi_\delta\right),$$

$$\mathcal{Z} = \sqrt{\frac{(2\pi)^p}{\det \mathbf{M}}} = \int \left(\prod_{\gamma=1}^{p} \mathrm{d}\phi_\gamma \right) \exp\!\left(-\tfrac{1}{2}\, \phi_\gamma \mathbf{M}_{\gamma\delta} \phi_\delta\right).$$

Using the replica identity, we rewrite the normalization as $\mathcal{Z}^{-1} = \lim_{s\to 0} \mathcal{Z}^{\,s-1}$, obtaining

$$\left(\mathbf{M}^{-1}\right)_{\alpha\beta} = \lim_{s\to 0} \mathcal{Z}^{\,s-1} \int \left(\prod_{\gamma=1}^{p} \mathrm{d}\phi_\gamma \right) \phi_\alpha \phi_\beta \, \exp\!\left(-\tfrac{1}{2}\, \phi_\gamma \mathbf{M}_{\gamma\delta} \phi_\delta\right).$$

Renaming the integration variable of this integral as ϕ^1 and writing the factor $\mathcal{Z}^{\,s-1}$ as $s-1$ further Gaussian integrals over replicas ϕ^a, $a \in \{2, \dots, s\}$, we obtain expression Eq. 27.

C.1.5 SCHUR COMPLEMENT

For a block matrix $\left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right)$: if $\mathbf{A}$ is invertible,

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det(\mathbf{A}) \, \det\!\big(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\big), \tag{28}$$

and if $\mathbf{D}$ is invertible, the top-left block of the inverse is

$$\left[\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} \right]_{11} = \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}. \tag{29}$$

C.1.6 SADDLE-POINT METHOD

In the computation of the traces we will obtain expressions of the form

$$\int \mathrm{d}\boldsymbol{\Theta} \; P(\boldsymbol{\Theta}) \, e^{-\frac{d}{2}\, s\, S(\boldsymbol{\Theta})}, \tag{30}$$

where $\boldsymbol{\Theta}$ collects the order parameters, S is the effective action, P a non-exponential prefactor, and s the number of replicas. As $d \to \infty$, the Laplace method applies and the measure $\propto e^{-\frac{d}{2} s S(\boldsymbol{\Theta})}$ concentrates on the stationary point $\boldsymbol{\Theta}^*$ defined by $\partial_{\boldsymbol{\Theta}} S(\boldsymbol{\Theta}^*) = 0$, giving

$$\int \mathrm{d}\boldsymbol{\Theta} \; P(\boldsymbol{\Theta}) \, e^{-\frac{d}{2} s S(\boldsymbol{\Theta})} \;\sim\; P(\boldsymbol{\Theta}^*). \tag{31}$$

C.2 BIAS–VARIANCE DECOMPOSITION OF THE TEST LOSS IN THE GENERAL CASE

In this section we derive the bias–variance decomposition of the denoising score matching (DSM) test loss, for a general data distribution.

We recall that $\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}$ denotes the minimizer of the train loss $\mathcal{L}_{\text{train}}^{n,m}$ obtained from a training set $\mathcal{D}$ and an initialization $\boldsymbol{\Theta}$ of the parameters. Throughout, $\mathbb{E}_{\boldsymbol{y}}$ denotes the average over a fresh test sample $\boldsymbol{y} \sim P_t$, drawn independently of $\mathcal{D}$, while $\langle\cdot\rangle$ denotes the average over the realizations of the training set and over the remaining sources of randomness, such as $\boldsymbol{\Theta}$. Note that $\mathcal{L}_{\text{test}}(\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}})$ is itself a random variable: the decomposition below concerns its average $\langle \mathcal{L}_{\text{test}} \rangle$, which is the quantity we look at. During the computation, we shall also use Tweedie's identity

$$\mathbb{E}[\boldsymbol{\xi} \,|\, \boldsymbol{y}] = -\sqrt{\Delta_t}\, \nabla \log P_t(\boldsymbol{y}). \tag{32}$$

Adding and subtracting $\sqrt{\Delta_t}\,\nabla \log P_t(\boldsymbol{y})$ inside the norm, at fixed $\mathcal{D}$ and $\boldsymbol{\Theta}$, yields

$$\mathcal{L}_{\text{test}}(\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}) = \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\left[\left\|\sqrt{\Delta_t}\,\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) + \boldsymbol{\xi}\right\|^2\right] \tag{33}$$

$$= \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\left[\left\|\sqrt{\Delta_t}\Big(\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \nabla \log P_t(\boldsymbol{y})\Big) + \Big(\boldsymbol{\xi} + \sqrt{\Delta_t}\,\nabla \log P_t(\boldsymbol{y})\Big)\right\|^2\right] \tag{34}$$

$$= \frac{\Delta_t}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\left\|\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \nabla \log P_t(\boldsymbol{y})\right\|^2\right] + C_t. \tag{35}$$

The cross term vanishes: since the test sample is drawn independently of $(\mathcal{D}, \boldsymbol{\Theta})$, the factor $\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \nabla \log P_t(\boldsymbol{y})$ is a deterministic function of $\boldsymbol{y}$ and can be pulled out of the conditional expectation, while Eq. 32 gives $\mathbb{E}[\boldsymbol{\xi} + \sqrt{\Delta_t}\,\nabla \log P_t(\boldsymbol{y}) \mid \boldsymbol{y}] = \mathbf{0}$. The remaining term,

$$C_t = \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\left[\left\|\boldsymbol{\xi} + \sqrt{\Delta_t}\,\nabla \log P_t(\boldsymbol{y})\right\|^2\right] = 1 - \frac{\Delta_t}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\left\|\nabla \log P_t(\boldsymbol{y})\right\|^2\right], \tag{36}$$

depends only on the data distribution and on t, and is independent of the learned score. It is the irreducible part of the test loss, namely the residual uncertainty on $\boldsymbol{\xi}$ left after observing $\boldsymbol{y}$, and is reached by the exact score $\nabla \log P_t$.

Averaging over $\mathcal{D}$ and $\boldsymbol{\Theta}$ and inserting $\pm\langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\rangle$ inside the norm yields

$$\left\langle \mathcal{L}_{\text{test}} \right\rangle = C_t + \Delta_t\left(\mathcal{B}^2 + \mathcal{V}\right), \tag{37}$$

with

$$\mathcal{B}^2 = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\left\|\nabla \log P_t(\boldsymbol{y}) - \left\langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\right\rangle\right\|^2\right], \qquad \mathcal{V} = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{y}}\left[\left\langle \left\|\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \left\langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\right\rangle\right\|^2 \right\rangle\right]. \tag{38}$$

Here the cross term also vanishes. The factor $\langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\rangle - \nabla \log P_t(\boldsymbol{y})$ is deterministic once $\boldsymbol{y}$ is fixed, while $\boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y}) - \langle \boldsymbol{s}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{y})\rangle$ has zero mean under $\langle\cdot\rangle$ by construction.

C.3 Loss decomposition

In this section we use the Gaussian Equivalence Principle to decompose the three losses into traces of rational functions of random matrices.

The three losses are all of the form

$$\mathcal{L}(\boldsymbol{s}) = \frac{1}{d}\,\mathbb{E}_{(\mathbf{x},\boldsymbol{\xi})\sim p}\left[\left\|\sqrt{\Delta_t}\boldsymbol{s}(e^{-t}\mathbf{x} + \sqrt{\Delta_t}\boldsymbol{\xi}) + \boldsymbol{\xi}\right\|^2\right], \tag{39}$$

where the joint law $p(\mathbf{x}, \boldsymbol{\xi})$ of the noisy data and the noise is one of

$$p_{\text{data}} \otimes \mathcal{N}(0,1), \qquad \tfrac{1}{n}\sum_{\nu=1}^{n} \delta(\mathbf{x} - \mathbf{x}^\nu) \otimes \mathcal{N}(0,1), \qquad \tfrac{1}{nm}\sum_{\nu=1}^{n}\sum_{\mu=1}^{m} \delta(\mathbf{x} - \mathbf{x}^\nu) \otimes \delta(\boldsymbol{\xi} - \boldsymbol{\xi}^{\nu\mu}), \tag{40}$$

yielding respectively $\mathcal{L}_{\text{test}}$, $\mathcal{L}_{\text{test}}^{\text{emp}}$ and $\mathcal{L}_{\text{train}}^{n,m}$. We derive the following result on $\mathcal{L}_{\text{test}}^{\text{emp}}$ which relates to the distance to the empirical score,

Proposition C.1. *Let $\hat{P}_t(\boldsymbol{y}) = \frac{1}{n}\sum_{\nu=1}^{n} \mathcal{N}(\boldsymbol{y};\, e^{-t}\mathbf{x}^\nu, \Delta_t \boldsymbol{I}_d)$ be the empirical noised distribution and $\boldsymbol{s}_{\text{emp}}(\boldsymbol{y}) = \nabla_{\boldsymbol{y}} \log \hat{P}_t(\boldsymbol{y})$ the associated empirical score. Then, for any score function $\boldsymbol{s}$,*

$$\mathcal{L}_{\text{test}}^{\text{emp}}(\boldsymbol{s}) = \frac{\Delta_t}{d}\,\mathbb{E}_{\boldsymbol{y}\sim\hat{P}_t}\left[\left\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}_{\text{emp}}(\boldsymbol{y})\right\|^2\right] + C, \tag{41}$$

where C depends only on t and the empirical score $\boldsymbol{s}_{\text{emp}}$, but not on $\boldsymbol{s}$. Moreover, assume that n grows at most polynomially in d, that $t = O_d(1)$, and that $\|\mathbf{x}^\mu - \mathbf{x}^\nu\|^2 = O(d)$ for all $\mu \neq \nu$. Then $C \to 0$ as $d \to \infty$.

Proof. Write $\boldsymbol{y}^\nu = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Since each $\boldsymbol{y}^\nu$ is drawn from $\hat{P}_t$, we can express $\mathcal{L}_{\text{test}}^{\text{emp}}$ as an expectation over $\hat{P}_t$:

$$\mathcal{L}_{\text{test}}^{\text{emp}}(\boldsymbol{s}) = \frac{1}{d}\,\mathbb{E}_{\boldsymbol{y}\sim\hat{P}_t}\left[\Delta_t\|\boldsymbol{s}(\boldsymbol{y})\|^2 + 2\sqrt{\Delta_t}\,\boldsymbol{s}(\boldsymbol{y}) \cdot \mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{y}]\right] + C_1, \tag{42}$$

where $C_1 = \frac{1}{d}\mathbb{E}\|\boldsymbol{\xi}\|^2$ does not depend on $\boldsymbol{s}$. To evaluate $\mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{y}]$, note that for each component ν the conditional score of the Gaussian kernel satisfies $\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y} \mid \mathbf{x}^\nu) = -(\boldsymbol{y} - e^{-t}\mathbf{x}^\nu)/\Delta_t = -\boldsymbol{\xi}^\nu/\sqrt{\Delta_t}$, so $\boldsymbol{\xi}^\nu = -\sqrt{\Delta_t}\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y} \mid \mathbf{x}^\nu)$. Taking the posterior expectation over $\mathbf{x}^\nu$ given $\boldsymbol{y}$ under $\hat{P}_t$,

$$\mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{y}] = -\sqrt{\Delta_t}\,\mathbb{E}_{\mathbf{x}^\nu|\boldsymbol{y}}[\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y} \mid \mathbf{x}^\nu)] = -\sqrt{\Delta_t}\,\nabla_{\boldsymbol{y}} \log \hat{P}_t(\boldsymbol{y}) = -\sqrt{\Delta_t}\,\boldsymbol{s}_{\mathrm{emp}}(\boldsymbol{y}), \quad (43)$$

where the second equality uses $\nabla_{\boldsymbol{y}} \log \hat{P}_t(\boldsymbol{y}) = \mathbb{E}_{\mathbf{x}^\nu|\boldsymbol{y}}[\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y} \mid \mathbf{x}^\nu)]$. Substituting back and completing the square gives the result, with $C = C_1 - \frac{\Delta_t}{d}\mathbb{E}_{\boldsymbol{y}}\|\boldsymbol{s}_{\mathrm{emp}}(\boldsymbol{y})\|^2$. The vanishing of C as $d \to \infty$ with $t = O(1)$ follows from $\boldsymbol{s}_{\mathrm{emp}}(e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}) \to -\boldsymbol{\xi}/\sqrt{\Delta_t}$, which gives $\mathcal{L}^{\mathrm{emp}}_{\mathrm{test}}(\boldsymbol{s}_{\mathrm{emp}}) \to 0$, i.e. $C \to 0$. □

Hence, in our regime, the smaller $\mathcal{L}^{\mathrm{emp}}_{\mathrm{test}}$ is, the closer the learned score is to the empirical score and hence to memorization.

For the random-feature score $\boldsymbol{s}_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{A}}{\sqrt{p}}\sigma(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}})$, writing $\boldsymbol{s}(\mathbf{x}_t)$ for the model at the noised input $\mathbf{x}_t = e^{-t}\mathbf{x} + \sqrt{\Delta_t}\boldsymbol{\xi}$ and expanding the square,

$$\mathcal{L}(\boldsymbol{s}) = \frac{1}{d}\mathbb{E}[\|\boldsymbol{\xi}\|^2] + \frac{2\sqrt{\Delta_t}}{d}\,\mathrm{Tr}\,\frac{\mathbf{A}}{\sqrt{p}}\mathbb{E}\Big[\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})\,\boldsymbol{\xi}^T\Big] + \frac{\Delta_t}{d}\,\mathrm{Tr}\left(\frac{\mathbf{A}}{\sqrt{p}}\mathbb{E}\Big[\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})^T\Big]\,\frac{\mathbf{A}^T}{\sqrt{p}}\right) + \frac{\Delta_t\lambda}{pd}\|\mathbf{A}\|_F^2, \quad (44)$$

where we added the ridge penalty $\frac{\Delta_t\lambda}{pd}\|\mathbf{A}\|_F^2$. In all three cases $\frac{1}{d}\mathbb{E}[\|\boldsymbol{\xi}\|^2]$ concentrates to 1. Introducing the feature–noise and feature–feature correlations

$$\tilde{\mathbf{V}} = \tfrac{1}{\sqrt{\Delta_t}}\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}[\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})\boldsymbol{\xi}^T], \qquad \tilde{\mathbf{U}} = \mathbb{E}_{\mathbf{x}_t\sim p_t}[\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})\sigma(\tfrac{\mathbf{W}\mathbf{x}_t}{\sqrt{d}})^T], \quad (45)$$

$$\mathbf{V}_n^m = \tfrac{1}{\sqrt{\Delta_t}nm}\sum_{\nu,\mu}\sigma(\tfrac{\mathbf{W}(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\mu})}{\sqrt{d}})(\boldsymbol{\xi}^{\nu\mu})^T, \quad \mathbf{U}_n^m = \tfrac{1}{nm}\sum_{\nu,\mu}\sigma(\tfrac{\mathbf{W}\mathbf{x}_t^{\nu\mu}}{\sqrt{d}})\sigma(\tfrac{\mathbf{W}\mathbf{x}_t^{\nu\mu}}{\sqrt{d}})^T, \quad (46)$$

$$\mathbf{V}_n^\infty = \tfrac{1}{\sqrt{\Delta_t}n}\sum_{\nu}\mathbb{E}_{\boldsymbol{\xi}}[\sigma(\tfrac{\mathbf{W}(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi})}{\sqrt{d}})\boldsymbol{\xi}^T], \quad \mathbf{U}_n^\infty = \tfrac{1}{n}\sum_{\nu}\mathbb{E}_{\boldsymbol{\xi}}[\sigma(\tfrac{\mathbf{W}\mathbf{x}_t^\nu}{\sqrt{d}})\sigma(\tfrac{\mathbf{W}\mathbf{x}_t^\nu}{\sqrt{d}})^T], \quad (47)$$

the regularized empirical risk minimizer is the ridge estimator

$$\frac{\mathbf{A}_{\mathrm{ERM}}}{\sqrt{p}} = -(\mathbf{V}_n^m)^T\,(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}. \quad (48)$$

Substituting it back, the three losses read

$$\mathcal{L}^{n,m}_{\mathrm{train}} = 1 - \frac{2\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{V}_n^m\big) \quad (49)$$

$$+ \frac{\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{U}_n^m(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{V}_n^m\big)$$

$$+ \frac{\Delta_t\lambda}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-2}\mathbf{V}_n^m\big), \quad (50)$$

$$\mathcal{L}^{\mathrm{emp}}_{\mathrm{test}} = 1 - \frac{2\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{V}_n^\infty\big) \quad (51)$$

$$+ \frac{\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{U}_n^\infty(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{V}_n^m\big), \quad (52)$$

$$\mathcal{L}_{\mathrm{test}} = 1 - \frac{2\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\tilde{\mathbf{V}}\big) \quad (53)$$

$$+ \frac{\Delta_t}{d}\,\mathrm{Tr}\big((\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\tilde{\mathbf{U}}(\mathbf{U}_n^m + \lambda\boldsymbol{I}_p)^{-1}\mathbf{V}_n^m\big). \quad (54)$$

C.4 Assumptions and notation

In this section we recall the assumptions made in the analytical part as well as the notations used.

The assumptions are

(A1) Data distribution. The training samples $\mathbf{x}^\nu$ are i.i.d. samples from P_0 and the frozen noises are i.i.d. from $\mathcal{N}(0, \boldsymbol{I}_d)$. We present the result with $P_0 = \mathcal{N}(0, \boldsymbol{I}_d)$ but in the appendix we extend the results to P_0 with zero mean, sub Gaussian and with a covariance $\boldsymbol{\Sigma}$ that admits a limiting spectral distribution $\rho_{\boldsymbol{\Sigma}}$ and verifies $\mathrm{Tr}(\boldsymbol{\Sigma})/d \to \sigma_{\mathbf{x}}^2 = O_d(1)$.

(A2) RFNN Architecture. The score is modeled by the RFNN $\boldsymbol{s}_{\mathbf{A}}(\boldsymbol{y}) = \frac{\mathbf{A}}{\sqrt{p}}\sigma\big(\frac{\mathbf{W}\boldsymbol{y}}{\sqrt{d}}\big)$, where the first-layer weights $\mathbf{W} \in \mathbb{R}^{p\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries and are frozen and independent of the data and of the noises, while only the read-out $\mathbf{A} \in \mathbb{R}^{d\times p}$ is trained.

(A3) Activation Function. The activation σ admits the Hermite expansion $\sigma(z) = \sum_{s\geq 0} \frac{c_s}{s!} He_s(z)$ with $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)^2] = \sum_{s\geq 0} \frac{c_s^2}{s!} < \infty$, and is centered, $\mu_0 = c_0 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)] = 0$. We write $\mu_1 = c_1 = \mathbb{E}[z\,\sigma(z)]$ and $\mu_*^2 = \mathbb{E}[\sigma(z)^2] - \mu_1^2 = \sum_{s\geq 2} \frac{c_s^2}{s!}$.

(A4) Asymptotic limit. We consider the proportional regime $n, p, d \to \infty$ with $n/d \to \psi_n \in (0,\infty)$ and $p/d \to \psi_p \in (0,\infty)$, the remaining parameters $m \in \mathbb{N}^*$, $t > 0$ being held fixed and $O(1)$. The two further limits we consider, $\psi_p \to \infty$ (Appendix C.10) and $m \to \infty$ (Appendix C.11), are always taken after $d \to \infty$.

Index conventions. Unless stated otherwise we use the Einstein summation convention, repeated indices being implicitly summed over. Latin indices $i, j, k \in \{1, \dots, d\}$ label input coordinates, Greek indices $\alpha, \beta \in \{1, \dots, p\}$ label features (hidden units), $\nu, \nu' \in \{1, \dots, n\}$ label training samples, $\mu, \mu' \in \{1, \dots, m\}$ label noise realizations, and $a, b \in \{1, \dots, s\}$ label replicas. Bold upright symbols denote vectors, matrices and tensors, and $\boldsymbol{I}_k$ is the $k \times k$ identity.

$d,\ n,\ p$	input dimension, number of training samples, number of features
m	number of frozen noise realizations per training sample
$\psi_n = n/d, \psi_p = p/d$	sample and parameter ratios
$t, \Delta_t = 1 - e^{-2t}$	diffusion time and variance of the forward kernel
$\lambda, \tilde{\lambda} = \lambda/\psi_p$	ridge strength, in the $O(1)$ and $O(\psi_p)$ scalings of Appendix C.10
$\boldsymbol{\Sigma}$, $\rho_{\boldsymbol{\Sigma}}$	data covariance ($\boldsymbol{\Sigma} = \boldsymbol{I}_d$ in the main text) and its limiting spectral density
$\mathbf{x}^\nu \in \mathbb{R}^d$	training sample ν; $\mathcal{D} = \{\mathbf{x}^\nu\}_{\nu\leq n}$ is the training set
$\boldsymbol{\xi} \in \mathbb{R}^{d\times nm}$	frozen training noises
$\mathbf{Y} \in \mathbb{R}^{d\times nm}$	noised inputs, $\mathbf{Y}_i^{\nu\mu} = e^{-t}\mathbf{x}_i^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}_i^{\nu\mu}$
$\mathbf{W} \in \mathbb{R}^{p\times d}$, $\mathbf{A} \in \mathbb{R}^{d\times p}$	frozen first layer and trained read-out
$\mathbf{F} \in \mathbb{R}^{p\times nm}$	feature matrix $\sigma(\mathbf{W}\mathbf{Y}/\sqrt{d})$
$\boldsymbol{v} \in \mathbb{R}^m$	all-ones vector, $\boldsymbol{v}^\mu = 1$
$\mathbf{1}_m = \boldsymbol{v}\boldsymbol{v}^T$	all-ones matrix, $\mathbf{1}_m \in \mathbb{R}^{m\times m}$
μ_1	$\mathbb{E}[z\sigma(z)]$
μ_*^2	$\mathbb{E}[\sigma(z)^2] - \mu_1^2$
κ	$\frac{1}{\mu_*^2}\,\mathbb{E}_{u,v,w}\big[\big(\sigma(e^{-t}u + \sqrt{\Delta_t}\,v) - \mu_1 e^{-t}u\big)$ $\times\big(\sigma(e^{-t}u + \sqrt{\Delta_t}\,w) - \mu_1 e^{-t}u\big)\big]$, $u, v, w \sim \mathcal{N}(0,1)$ independent
$\mathcal{L}_{\text{train}}^{n,m}$, $\mathcal{L}_{\text{test}}^{\text{emp}}$, $\mathcal{L}_{\text{test}}$	train loss at finite m, empirical test loss ($m = \infty$ at fixed $\mathcal{D}$), and population test loss, Eqs. 2, 1 and 3
$\mathcal{B}^2, \mathcal{V}$	bias, variance

C.5 Gaussian Equivalence Principle

According to the Gaussian Equivalence Principle (GEP) (Pennington & Worah, 2017; Péché, 2019; Goldt et al., 2020; 2021; Mei & Montanari, 2019; Hu & Lu, 2023), nonlinear random matrices of the form $\mathbf{F} = \sigma(\frac{\mathbf{W}\mathbf{Y}}{\sqrt{d}})$ have asymptotically the same limiting spectral distribution and resolvent traces as linear matrices with the same first two moments,

$$\mathbf{F} = \sigma\Big(\frac{\mathbf{W}\mathbf{Y}}{\sqrt{d}}\Big) \;\longrightarrow\; \mu_0\,\mathbf{1}\mathbf{1}^T + \mu_1\frac{\mathbf{W}\mathbf{Y}}{\sqrt{d}} + \mu_*\boldsymbol{\Omega}, \tag{55}$$

with scalar constants μ_0, μ_1, μ_* and $\boldsymbol{\Omega}$ a Gaussian tensor independent of $\mathbf{W}$ and $\mathbf{Y}$. Since the data are sampled from $\mathcal{N}(0, \boldsymbol{I}_d)$ and we assumed that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)] = 0$, we have $\mu_0 = 0$. Matching the first two moments of the tensor $\mathbf{F}_\alpha^{\nu\mu}$ requires $\mu_1 = \mathbb{E}[\sigma(z)z]$, $\mu_*^2 = \mathbb{E}[\sigma^2(z)] - \mu_1^2$, and

$$\mathbb{E}[\boldsymbol{\Omega}_\alpha^{\nu\mu}\boldsymbol{\Omega}_\beta^{\nu'\mu'}] = \delta_{\alpha\beta}\delta^{\nu\nu'}\big(\delta^{\mu\mu'} + \kappa(1 - \delta^{\mu\mu'})\big), \tag{56}$$

with

$$\kappa = \frac{1}{\mu_*^2}\,\mathbb{E}_{u,v,w}\Big[\big(\sigma(e^{-t}u+\sqrt{\Delta_t}v)-\mu_1 e^{-t}u\big)\big(\sigma(e^{-t}u+\sqrt{\Delta_t}w)-\mu_1 e^{-t}u\big)\Big], \tag{57}$$

$u, v, w \sim \mathcal{N}(0,1)$ independent. We now replace $\mathbf{F}$ by its Gaussian equivalent in the matrices $\mathbf{U}$ and $\mathbf{V}$.[8] Using $\mathbb{E}[\mathbf{Y}\boldsymbol{\xi}^T] = \sqrt{\Delta_t}\,\boldsymbol{I}_d$ and $\mathbb{E}[\boldsymbol{\Omega}\boldsymbol{\xi}^T] = 0$,

$$\tilde{\mathbf{V}} = \frac{1}{\sqrt{\Delta_t}}\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\Big[\Big(\mu_1\frac{\mathbf{W}\mathbf{Y}}{\sqrt{d}}+\mu_*\boldsymbol{\Omega}\Big)\boldsymbol{\xi}^T\Big] = \mu_1\frac{\mathbf{W}}{\sqrt{d}} = \mathbf{V}_n^\infty, \qquad \tilde{\mathbf{U}} = \mu_1^2\frac{\mathbf{W}\mathbf{W}^T}{d}+\mu_*^2\boldsymbol{I}_p. \tag{58}$$

For $\mathbf{V}_n^m$ and $\mathbf{U}_n^m$ there is no expectation to compute and one simply substitutes the surrogate,

$$\mathbf{V}_n^m = \frac{1}{\sqrt{\Delta_t}nm}\sum_{\nu,\mu}\Big(\mu_1\frac{\mathbf{W}(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\mu})}{\sqrt{d}}+\mu_*\boldsymbol{\Omega}^{\nu\mu}\Big)(\boldsymbol{\xi}^{\nu\mu})^T, \tag{59}$$

$$\mathbf{U}_n^m = \frac{1}{nm}\sum_{\nu,\mu}\Big(\mu_1\frac{\mathbf{W}(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\mu})}{\sqrt{d}}+\mu_*\boldsymbol{\Omega}^{\nu\mu}\Big)\Big(\mu_1\frac{\mathbf{W}(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\mu})}{\sqrt{d}}+\mu_*\boldsymbol{\Omega}^{\nu\mu}\Big)^T. \tag{60}$$

Gaussian equivalent of $\mathbf{U}_n^\infty$. The matrix $\mathbf{U}_n^\infty = \frac{1}{n}\sum_\nu \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{F}\mathbf{F}^T]$ is more involved because the expectation is taken on the noise only. Its GEP has been derived in George et al. (2026); Bonnaire et al. (2025).[9] We introduce the scalar constants

$$v_t^2 = \mathbb{E}_{u,v,w}[\sigma(e^{-t}u+\sqrt{\Delta_t}v)\sigma(e^{-t}u+\sqrt{\Delta_t}w)] - \big(\mathbb{E}_{u,v}[\sigma(e^{-t}u+\sqrt{\Delta_t}v)\,u]\big)^2, \tag{61}$$

$$s_t^2 = \mathbb{E}_u[\sigma(u)^2] - \mathbb{E}_{u,v,w}[\sigma(e^{-t}u+\sqrt{\Delta_t}v)\sigma(e^{-t}u+\sqrt{\Delta_t}w)] - \big(\mathbb{E}_{u,v}[v\,\sigma(e^{-t}u+\sqrt{\Delta_t}v)]\big)^2, \tag{62}$$

which, as we show below, coincide with $v_t^2 = \mu_*^2\kappa$ and $s_t^2 = \mu_*^2(1-\kappa)$. Recall the Hermite expansion $\sigma(z) = \sum_{s\geq 0}\frac{c_s}{s!}He_s(z)$, with $\mathbb{E}[He_s He_{s'}] = s!\,\delta_{ss'}$, $c_0 = 0$, $c_1 = \mu_1$ and $\mu_*^2 = \sum_{s\geq 2}\frac{c_s^2}{s!}$. Write $\boldsymbol{h}_\alpha^\nu(\boldsymbol{\xi}) = \frac{\mathbf{W}_\alpha\cdot(e^{-t}\mathbf{x}^\nu+\sqrt{\Delta_t}\boldsymbol{\xi})}{\sqrt{d}}$ for the preactivation, and

$$R(h) = \sigma(h) - \mu_1 h = \sum_{s\geq 2}\frac{c_s}{s!}He_s(h), \qquad \sigma_0(g) = \mathbb{E}_{w\sim\mathcal{N}(0,1)}[\sigma(g+\sqrt{\Delta_t}w)]. \tag{63}$$

Diagonal terms. For $\alpha = \beta$, $\boldsymbol{h}_\alpha^\nu(\boldsymbol{\xi}) \sim \mathcal{N}(0,1)$ asymptotically, and the data average concentrates, so up to $O(1/n)$[10] we have,

$$(\mathbf{U}_n^\infty)_{\alpha\alpha} = \mathbb{E}_{\boldsymbol{\xi}}[\sigma(\boldsymbol{h}_\alpha^\nu)^2] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)^2] = \mu_1^2+\mu_*^2 = \|\sigma\|^2. \tag{64}$$

Off-diagonal terms. Fix $\alpha \neq \beta$, freeze $\mathbf{W}, \mathbf{x}^\nu$, and write $\boldsymbol{h}_\alpha^\nu = \boldsymbol{g}_\alpha^\nu + \sqrt{\Delta_t}u_\alpha$ with $\boldsymbol{g}_\alpha^\nu = \frac{e^{-t}\mathbf{W}_\alpha\cdot\mathbf{x}^\nu}{\sqrt{d}} \sim \mathcal{N}(0, e^{-2t})$ and $u_\alpha = \frac{\mathbf{W}_\alpha\cdot\boldsymbol{\xi}}{\sqrt{d}}$, so $\mathbb{E}_{\boldsymbol{\xi}}[u_\alpha u_\beta] = \rho_{\alpha\beta} = \frac{\mathbf{W}_\alpha\cdot\mathbf{W}_\beta}{d} = O(d^{-1/2})$. The Mehler–Kibble formula (Kibble, 1945) gives

$$\mathbb{E}_{\boldsymbol{\xi}}[\sigma(\boldsymbol{h}_\alpha^\nu)\sigma(\boldsymbol{h}_\beta^\nu)] = \sum_{s\geq 0}\frac{\rho_{\alpha\beta}^s}{s!}\,\mathbb{E}_u[He_s(u)\sigma(\boldsymbol{g}_\alpha^\nu+\sqrt{\Delta_t}u)]\,\mathbb{E}_u[He_s(u)\sigma(\boldsymbol{g}_\beta^\nu+\sqrt{\Delta_t}u)]. \tag{65}$$

Terms $s \geq 2$ are $O(\rho^2) = O(1/d)$ and are dropped. The $s = 0$ term is $\sigma_0(\boldsymbol{g}_\alpha^\nu)\sigma_0(\boldsymbol{g}_\beta^\nu)$, while averaging the $s = 1$ coefficient over the data gives $\rho_{\alpha\beta}(\mathbb{E}_{g,u}[u\,\sigma(g+\sqrt{\Delta_t}u)])^2 = \rho_{\alpha\beta}\,\Delta_t\mu_1^2$, where Stein's lemma yields $\mathbb{E}[u\,\sigma(h)] = \sqrt{\Delta_t}\mu_1$ with $h = g+\sqrt{\Delta_t}u$. Hence

$$(\mathbf{U}_n^\infty)_{\alpha\beta} = \frac{1}{n}\sum_\nu \sigma_0(\boldsymbol{g}_\alpha^\nu)\sigma_0(\boldsymbol{g}_\beta^\nu) + \Delta_t\mu_1^2\frac{\mathbf{W}_\alpha\cdot\mathbf{W}_\beta}{d} + O(1/d). \tag{66}$$

[8] With a slight abuse of notation we use the equal sign $=$ for two matrices that have asymptotically the same spectral densities.

[9] See Lemma B.1 of George et al. (2026) and Lemma C.1 of Bonnaire et al. (2025).

[10] Terms of order $O(1/d)$ are irrelevant to the asymptotic spectrum.

The field $\boldsymbol{g}_\alpha^\nu$ is Gaussian and σ_0 smooth, so a second Gaussian equivalence applies, $\sigma_0(\boldsymbol{g}_\alpha^\nu) \to \mu_1 \boldsymbol{g}_\alpha^\nu + v_t \boldsymbol{\eta}_\alpha^\nu$ with $\boldsymbol{\eta}_\alpha^\nu \sim \mathcal{N}(0,1)$. The linear coefficient is μ_1 (Stein's lemma) and the residual variance is

$$v_t^2 = \mathbb{E}_g[\sigma_0(g)^2] - \mu_1^2 e^{-2t}, \qquad \mathbb{E}_g[\sigma_0(g)^2] = \sum_{s\geq 0} \frac{c_s^2}{s!} e^{-2st}, \tag{67}$$

using Mehler with shared $g \sim \mathcal{N}(0, e^{-2t})$. Since $\sum_{s\geq 2} \frac{c_s^2}{s!} e^{-2st} = \mu_*^2 \kappa$, we obtain $v_t^2 = \mu_*^2 \kappa$, i.e. $v_t = \mu_* \sqrt{\kappa}$. On the diagonal the same surrogate gives $\mu_1^2 + \mu_*^2 \kappa$; the deficit $\mu_*^2(1-\kappa)$ relative to $\|\sigma\|^2 = \mu_1^2 + \mu_*^2$ is restored by an isotropic term, giving $s_t^2 = \mu_*^2(1-\kappa)$ and

$$\mathbf{U}_n^\infty = \frac{\mathbf{G}}{\sqrt{n}} \frac{\mathbf{G}^T}{\sqrt{n}} + \Delta_t \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + \mu_*^2(1-\kappa)\boldsymbol{I}_p, \qquad \mathbf{G}_\alpha^\nu = e^{-t}\mu_1 \frac{\mathbf{W}_\alpha \cdot \mathbf{x}^\nu}{\sqrt{d}} + v_t \boldsymbol{\eta}_\alpha^\nu. \tag{68}$$

The surrogate $\boldsymbol{\eta}$ is the noise-average of the residual $\boldsymbol{\Omega}$ and is correlated with it. Writing $R^{\nu\mu} = \mu_* \boldsymbol{\Omega}^{\nu\mu}$ and $R_0^\nu = v_t \boldsymbol{\eta}^\nu = \mathbb{E}_w[R(\boldsymbol{g}^\nu + \sqrt{\Delta_t} w)] = \lim_{m\to\infty} \frac{1}{m} \sum_{\mu'} R^{\nu\mu'}$, conditioning on $\boldsymbol{g}^\nu$ gives $\mathbb{E}[R^{\nu\mu} R_0^\nu] = \mathbb{E}_g[(R_0^\nu)^2] = v_t^2 = \mu_*^2 \kappa$, hence

$$\mathbb{E}[\boldsymbol{\Omega}_\alpha^{\nu\mu} \boldsymbol{\eta}_\beta^{\nu'}] = \frac{\mathbb{E}[R^{\nu\mu} R_0^\nu]}{\mu_* \, v_t} \, \delta_{\alpha\beta} \delta^{\nu\nu'} = \sqrt{\kappa} \, \delta_{\alpha\beta} \delta^{\nu\nu'}. \tag{69}$$

Summary. The GEP yields $\tilde{\mathbf{V}} = \mathbf{V}_n^\infty = \mu_1 \frac{\mathbf{W}}{\sqrt{d}}$, $\tilde{\mathbf{U}} = \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + \mu_*^2 \boldsymbol{I}_p$, $\mathbf{V}_n^m, \mathbf{U}_n^m$ as above, and

$$\mathbf{U}_n^\infty = \frac{1}{n} \sum_\nu \left(e^{-t}\mu_1 \frac{\mathbf{W}\mathbf{x}^\nu}{\sqrt{d}} + v_t \boldsymbol{\eta}^\nu\right)\left(e^{-t}\mu_1 \frac{\mathbf{W}\mathbf{x}^\nu}{\sqrt{d}} + v_t \boldsymbol{\eta}^\nu\right)^T + \Delta_t \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + s_t^2 \boldsymbol{I}_p, \tag{70}$$

with $v_t^2 = \mu_*^2 \kappa$, $s_t^2 = \mu_*^2(1-\kappa)$. Here $\boldsymbol{\eta}$ is a Gaussian noise with $\boldsymbol{\eta}_\alpha^\nu \sim \mathcal{N}(0,1)$, independent of $\mathbf{W}$ and $\mathbf{Y}$ but correlated with the residual $\boldsymbol{\Omega}$ through $\mathbb{E}[\boldsymbol{\Omega}_\alpha^{\nu\mu} \boldsymbol{\eta}_\beta^{\nu'}] = \sqrt{\kappa} \, \delta_{\alpha\beta} \delta^{\nu\nu'}$.

C.6 Gaussian-equivalent losses and trace decomposition

From now on we denote by $\mathbf{F}_\alpha^{\nu\mu}$ its Gaussian equivalent $\mu_1 \frac{\mathbf{W}\mathbf{Y}}{\sqrt{d}} + \mu_* \boldsymbol{\Omega}$, and by $\mathbf{H} = \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{F}] \in \mathbb{R}^{p\times n}$ the noise-averaged feature matrix, with entries $\mathbf{H}_\alpha^\nu = e^{-t}\mu_1 \frac{\mathbf{W}_\alpha \cdot \mathbf{x}^\nu}{\sqrt{d}} + v_t \boldsymbol{\eta}_\alpha^\nu$ and $v_t = \mu_* \sqrt{\kappa}$. Substituting the surrogates into the losses,

$$\mathcal{L}_{\text{train}}^{n,m} = 1 - \frac{1}{d(nm)^2} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \mathbf{F}\boldsymbol{\xi}^T\right), \tag{71}$$

$$\begin{aligned}\mathcal{L}_{\text{test}} = 1 &- \frac{2\mu_1\sqrt{\Delta_t}}{dnm} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \frac{\mathbf{W}}{\sqrt{d}}\right) \\ &+ \frac{1}{d(nm)^2} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \, \tilde{\mathbf{U}} \, \mathcal{G}_{-\lambda,0,0} \mathbf{F}\boldsymbol{\xi}^T\right),\end{aligned} \tag{72}$$

$$\begin{aligned}\mathcal{L}_{\text{test}}^{\text{emp}} = 1 &- \frac{2\mu_1\sqrt{\Delta_t}}{dnm} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \frac{\mathbf{W}}{\sqrt{d}}\right) \\ &+ \frac{1}{d(nm)^2} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \, \mathbf{U}_n^\infty \, \mathcal{G}_{-\lambda,0,0} \mathbf{F}\boldsymbol{\xi}^T\right),\end{aligned} \tag{73}$$

with $\tilde{\mathbf{U}} = \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + \mu_*^2 \boldsymbol{I}_p$ and $\mathbf{U}_n^\infty = \frac{\mathbf{H}\mathbf{H}^T}{n} + \Delta_t \mu_1^2 \frac{\mathbf{W}\mathbf{W}^T}{d} + s_t^2 \boldsymbol{I}_p$ the Gaussian-equivalent second-moment matrices of the summary above, $s_t^2 = \mu_*^2(1-\kappa)$, and $\mathcal{G}_{z,\zeta,\epsilon}$ the resolvent Eq. 7 of the main text,

$$\mathcal{G}_{z,\zeta,\epsilon} = \left(\frac{\mathbf{F}\mathbf{F}^T}{nm} - z\boldsymbol{I}_p - \zeta \frac{\mathbf{W}\mathbf{W}^T}{d} + \epsilon \frac{\mathbf{H}\mathbf{H}^T}{n}\right)^{-1}, \qquad \mathcal{G}_{-\lambda,0,0} = \left(\frac{\mathbf{F}\mathbf{F}^T}{nm} + \lambda \boldsymbol{I}_p\right)^{-1}. \tag{74}$$

Every quadratic form appearing above is a derivative of the single generating trace T_1 of Eq. 8,

$$T_1(z,\zeta,\epsilon) = \frac{1}{d(nm)^2} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{z,\zeta,\epsilon} \mathbf{F}\boldsymbol{\xi}^T\right), \qquad T_2 = \frac{1}{dnm} \operatorname{Tr}\left(\boldsymbol{\xi}\mathbf{F}^T \mathcal{G}_{-\lambda,0,0} \frac{\mathbf{W}}{\sqrt{d}}\right). \tag{75}$$

Indeed, differentiating Eq.74 gives $\partial_z \mathcal{G} = \mathcal{G}^2$, $\partial_\zeta \mathcal{G} = \mathcal{G}\frac{\mathbf{W}\mathbf{W}^T}{d}\mathcal{G}$ and $\partial_\epsilon \mathcal{G} = -\mathcal{G}\frac{\mathbf{H}\mathbf{H}^T}{n}\mathcal{G}$, so that, writing $\mathcal{G} = \mathcal{G}_{-\lambda,0,0}$ and evaluating all derivatives at $(z,\zeta,\epsilon) = (-\lambda,0,0)$,

$$T_3 = -\partial_\epsilon T_1 = \tfrac{1}{d(nm)^2}\,\mathrm{Tr}\Big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}\tfrac{\mathbf{H}\mathbf{H}^T}{n}\mathcal{G}\mathbf{F}\boldsymbol{\xi}^T\Big)\,,$$
$$T_4 = \partial_\zeta T_1 = \tfrac{1}{d(nm)^2}\,\mathrm{Tr}\Big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}\tfrac{\mathbf{W}\mathbf{W}^T}{d}\mathcal{G}\mathbf{F}\boldsymbol{\xi}^T\Big)\,, \qquad T_5 = \partial_z T_1 = \tfrac{1}{d(nm)^2}\,\mathrm{Tr}\Big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}^2\mathbf{F}\boldsymbol{\xi}^T\Big)\,. \tag{76}$$

Collecting terms, the three losses read as announced in Eqs. 9–11 of the main text,

$$\mathcal{L}_{\text{train}}^{n,m} = 1 - T_1(-\lambda,0,0), \tag{77}$$
$$\mathcal{L}_{\text{test}} = 1 - 2\mu_1\sqrt{\Delta_t}\,T_2 + \mu_1^2\,\partial_\zeta T_1\big|_{(-\lambda,0,0)} + \mu_*^2\,\partial_z T_1\big|_{(-\lambda,0,0)}, \tag{78}$$
$$\mathcal{L}_{\text{test}}^{\text{emp}} = 1 - 2\mu_1\sqrt{\Delta_t}\,T_2 - \partial_\epsilon T_1\big|_{(-\lambda,0,0)} + \Delta_t\mu_1^2\,\partial_\zeta T_1\big|_{(-\lambda,0,0)} + \mu_*^2(1-\kappa)\,\partial_z T_1\big|_{(-\lambda,0,0)}, \tag{79}$$

for which we use throughout the shorthands T_3, T_4, T_5 of Eq.76. It remains to compute T_1 and T_2. Both are controlled by the order parameters of Eq. 13,

$$q(z,\zeta,\epsilon) = \frac{1}{p}\,\mathrm{Tr}\;\mathcal{G}_{z,\zeta,\epsilon}, \qquad r(z,\zeta,\epsilon) = \frac{1}{p}\,\mathrm{Tr}\Big(\tfrac{\mathbf{W}^T}{\sqrt{d}}\,\mathcal{G}_{z,\zeta,\epsilon}\,\tfrac{\mathbf{W}}{\sqrt{d}}\Big), \tag{80}$$

which, as shown next, obey the self-consistent equations of Theorem 2.1. Proposition C.2 derives them at $\epsilon = 0$, which suffices for T_1, T_2, T_4, T_5; the ϵ-deformed saddle, needed only for T_3, is obtained in Appendix C.8.3.

C.7 Self-consistent equations for the Stieltjes transform

Proposition C.2 (Self-consistent equations)**.** *Let $\rho_{\boldsymbol{\Sigma}}$ denote the limiting spectral distribution of the data covariance $\boldsymbol{\Sigma}$. In the proportional limit the order parameters (q,r,h) and the conjugates $(\hat{r},\hat{h})$ satisfy the fixed-point system*

$$z\psi_p = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p\mu_*^2\frac{1+(m-1)\kappa}{m}}{L_1} + \frac{(m-1)\,\psi_p\mu_*^2\frac{1-\kappa}{m}}{L_2}, \tag{81}$$
$$\hat{r} = \frac{1}{q} - \psi_p\zeta + \frac{\psi_p\mu_1^2\frac{\Delta_t}{m}}{L_1} + \frac{(m-1)\,\psi_p\mu_1^2\frac{\Delta_t}{m}}{L_2}, \qquad \hat{h} = \frac{\psi_p\mu_1^2 e^{-2t}}{L_1}, \tag{82}$$
$$r = \int \frac{\mathrm{d}\rho_{\boldsymbol{\Sigma}}(\lambda)}{\hat{r}+\lambda\hat{h}}, \qquad h = \int \frac{\lambda\,\mathrm{d}\rho_{\boldsymbol{\Sigma}}(\lambda)}{\hat{r}+\lambda\hat{h}}, \tag{83}$$

with $L_1 = L_1(q,r,h) = 1+\frac{\psi_p}{\psi_n}\mu_1^2e^{-2t}h+\frac{\psi_p}{\psi_n m}\big[\mu_1^2\Delta_t r+\mu_*^2 q(1+(m-1)\kappa)\big]$ and $L_2 = L_2(q,r) = 1+\frac{\psi_p}{\psi_n m}\big[\mu_1^2\Delta_t r+\mu_*^2 q(1-\kappa)\big]$.

Remark. For isotropic data $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ one has $\rho_{\boldsymbol{\Sigma}} = \delta(\lambda-1)$, so Eq.83 gives $r = h = 1/(\hat{r}+\hat{h})$, i.e. $\hat{r}+\hat{h} = 1/r$ and $L_1(q,r,r) = L_1(q,r)$. Adding the two equations of Eq.82 eliminates the conjugates and the system collapses to

$$\frac{1}{r} + \psi_p\zeta = \frac{1}{q} + \frac{\mu_1^2\psi_p\big(e^{-2t}+\frac{\Delta_t}{m}\big)}{L_1(q,r)} + \frac{(m-1)\,\psi_p\mu_1^2\frac{\Delta_t}{m}}{L_2(q,r)}, \tag{84}$$
$$z\psi_p = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p\mu_*^2\frac{1+(m-1)\kappa}{m}}{L_1(q,r)} + \frac{(m-1)\,\psi_p\mu_*^2\frac{1-\kappa}{m}}{L_2(q,r)}, \tag{85}$$

with $L_1(q,r) = 1+\frac{\psi_p}{\psi_n m}\big[\mu_1^2\,r(e^{-2t}m+\Delta_t)+\mu_*^2\,q(1+(m-1)\kappa)\big]$ and $L_2(q,r) = 1+\frac{\psi_p}{\psi_n m}\big[\mu_1^2\,r\,\Delta_t+\mu_*^2\,q(1-\kappa)\big]$. This is the self-consistent system verified by (q,r) of Theorem 2.1 of the main text.

We prove the system for general covariance matrix $\boldsymbol{\Sigma}$ and then specialize to the case $\boldsymbol{\Sigma} = \boldsymbol{I}_d$.

Proof. Throughout the replica computations we use the *Einstein summation convention* (repeated indices mean implicit summation), with feature indices $\alpha, \beta \in \{1, \dots, p\}$, Latin coordinate indices $i, j, k \in \{1, \dots, d\}$, sample indices $\nu, \nu' \in \{1, \dots, n\}$, noise indices $\mu, \mu' \in \{1, \dots, m\}$, and replica indices $a, b \in \{1, \dots, s\}$. We compute $q(z, \zeta) = \frac{1}{p} \operatorname{Tr}(\mathbf{U} - z\boldsymbol{I}_p - \zeta \frac{\mathbf{W}\mathbf{W}^T}{d})^{-1}$ with $\mathbf{U}_{\alpha\beta} = \frac{(\mathbf{F}\mathbf{F}^T)_{\alpha\beta}}{nm}$ and $\mathbf{U}$ replaced by its Gaussian equivalent,

$$\mathbf{U}_{\alpha\beta} = \frac{1}{nm}\big(\mu_1 \mathbf{W}_{\alpha i} \mathbf{Y}_i^{\nu\mu} + \mu_* \mathbf{\Omega}_\alpha^{\nu\mu}\big)\big(\mu_1 \mathbf{W}_{\beta j} \mathbf{Y}_j^{\nu\mu} + \mu_* \mathbf{\Omega}_\beta^{\nu\mu}\big), \quad \mathbf{Y}_i^{\nu\mu} = e^{-t} \mathbf{\Sigma}_{ij}^{1/2} \mathbf{x}_j^\nu \boldsymbol{v}^\mu + \sqrt{\Delta_t}\, \boldsymbol{\xi}_i^{\nu\mu}, \tag{86}$$

$\mathbb{E}[\mathbf{x}_i^\nu \mathbf{x}_j^{\nu'}] = \delta^{\nu\nu'} \mathbf{\Sigma}_{ij}$, $\mathbb{E}[\boldsymbol{\xi}_i^{\nu\mu} \boldsymbol{\xi}_j^{\nu'\mu'}] = \delta_{ij} \delta^{\nu\nu'} \delta^{\mu\mu'}$, $\boldsymbol{v}^\mu = 1$. Write $\mathbf{M} = \mathbf{M}(z, \zeta) = \mathbf{U} - z\boldsymbol{I}_p - \zeta \frac{\mathbf{W}\mathbf{W}^T}{d}$, so that $q(z, \zeta) = \frac{1}{p} \operatorname{Tr} \mathbf{M}^{-1}$. The trace of the inverse is the derivative of a log-determinant: since $\partial_z \mathbf{M} = -\boldsymbol{I}_p$ and $\partial_z \log\det \mathbf{M} = \operatorname{Tr}(\mathbf{M}^{-1} \partial_z \mathbf{M})$,

$$q(z, \zeta) = \frac{1}{p} \operatorname{Tr} \mathbf{M}^{-1} = -\frac{1}{p}\, \partial_z \log\det \mathbf{M} = \frac{2}{p}\, \partial_z \log\det \mathbf{M}^{-1/2}. \tag{87}$$

We introduce the partition function

$$\mathcal{Z} = \det \mathbf{M}^{-1/2} = \int \frac{\mathrm{d}\phi}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}\phi^T \mathbf{M} \phi}, \tag{88}$$

Combining Eq.87–Eq.88, $q = \frac{2}{p}\, \partial_z \log \mathcal{Z}$; as q concentrates, we replace it by its quenched average $\frac{2}{p}\, \partial_z\, \mathbb{E} \log \mathcal{Z}$.

The remaining obstacle is the average of a logarithm, which we handle with the replica method (Mézard et al., 1987): the identity $\log x = \lim_{s \to 0} (x^s - 1)/s$ trades $\mathbb{E} \log \mathcal{Z}$ for the integer moments $\mathbb{E}\, \mathcal{Z}^s$,[11]

$$q(z, \zeta) = 2\, \partial_z \lim_{s \to 0} \lim_{p \to \infty} \frac{1}{ps}\, \mathbb{E}[\mathcal{Z}^s - 1], \qquad \mathcal{Z}^s = \det\Big(\mathbf{U} - z\boldsymbol{I}_p - \zeta \frac{\mathbf{W}\mathbf{W}^T}{d}\Big)^{-s/2}. \tag{89}$$

For integer s the s-th moment is a product of s identical Gaussian integrals over independent copies (*replicas*) ϕ^a, $a = 1, \dots, s$, namely $\mathcal{Z}^s = \int \prod_{a=1}^s \frac{\mathrm{d}\phi^a}{(2\pi)^{p/2}}\, e^{-\frac{1}{2}\sum_a \phi^{aT} \mathbf{M} \phi^a}$, and we average over the disorder before continuing to real $s \to 0$. Carrying out this disorder average,

$$\begin{aligned} \mathbb{E}[\mathcal{Z}^s] = \int \prod_a \mathrm{d}\phi^a\; & e^{\frac{z}{2}\phi^a \cdot \phi^a + \frac{\zeta}{2d} \phi^{aT} \mathbf{W}\mathbf{W}^T \phi^a} \\ & \times \mathbb{E}_{\mathbf{W}, \mathbf{x}, \boldsymbol{\xi}, \mathbf{\Omega}} \Big[\exp\Big(-\tfrac{1}{2nm} \phi_\alpha^a (\mu_1 \mathbf{W}_{\alpha i} \mathbf{Y}_i^{\nu\mu} + \mu_* \mathbf{\Omega}_\alpha^{\nu\mu})(\mu_1 \mathbf{W}_{\beta j} \mathbf{Y}_j^{\nu\mu} + \mu_* \mathbf{\Omega}_\beta^{\nu\mu}) \phi_\beta^a \Big) \Big]. \end{aligned} \tag{90}$$

We decouple the $\mathbf{W}$-dependence from the data with the auxiliary field

$$1 = \int \mathrm{d}\omega^a \mathrm{d}\hat{\omega}^a\; e^{i\hat{\omega}_i^a (\sqrt{p}\, \omega_i^a - \phi_\alpha^a \mathbf{W}_{\alpha i})}. \tag{91}$$

Average over $\mathbf{\Omega}$. Collecting all $\mathbf{\Omega}$-dependent terms and using the Gaussian toolkit,

$$\mathbb{E}_{\mathbf{\Omega}} \Big[\exp\Big(-\tfrac{\mu_*^2}{2nm} \phi_\alpha^a \phi_\beta^a \mathbf{\Omega}_\alpha^{\nu\mu} \mathbf{\Omega}_\beta^{\nu\mu} - \tfrac{\mu_1 \mu_*}{nm} \phi_\alpha^a \phi_\beta^a \mathbf{W}_{\alpha i} \mathbf{Y}_i^{\nu\mu} \mathbf{\Omega}_\beta^{\nu\mu} \Big) \Big] \tag{92}$$

$$= e^{-\frac{1}{2}\log\det \mathbf{C}_\Omega}\, e^{\frac{1}{2} \mathbf{J}_\Omega^T \mathbf{C}_\Omega^{-1} \mathbf{J}_\Omega}, \tag{93}$$

with, using $\mathbf{W}_{\alpha i} \phi_\alpha^a = \sqrt{p}\, \omega_i^a$,

$$(\mathbf{C}_\Omega)_{\alpha\beta}^{\nu\mu\nu'\mu'} = \boldsymbol{I}_p \otimes \boldsymbol{I}_n \otimes [\kappa\, \mathbf{1}_m + (1 - \kappa) \boldsymbol{I}_m]^{-1} + \frac{\mu_*^2}{nm} (\phi_\alpha^a \phi_\beta^a) \otimes \boldsymbol{I}_n \otimes \boldsymbol{I}_m, \tag{94}$$

$$(\mathbf{J}_\Omega)_\beta^{\nu\mu} = \frac{\mu_1 \mu_* \sqrt{p}}{nm}\, \omega_i^a\, \mathbf{Y}_i^{\nu\mu}\, \phi_\beta^a. \tag{95}$$

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

Average over $\mathbf{W}$. The only remaining $\mathbf{W}$-dependent factor is

$$\mathbb{E}_{\mathbf{W}}\Big[e^{-i\hat{\omega}_i^a\phi_\alpha^a\mathbf{W}_{\alpha i}}\Big] = e^{-\frac{1}{2}\hat{\omega}_i^a\phi_\alpha^a\hat{\omega}_i^b\phi_\alpha^b}, \tag{96}$$

so that

$$\mathbb{E}[\mathcal{Z}^s] = \int\prod_a \mathrm{d}\phi^a\mathrm{d}\omega^a\mathrm{d}\hat{\omega}^a\; e^{\frac{z}{2}\phi^a\cdot\phi^a+\frac{p\zeta}{2d}\omega^a\cdot\omega^a-\frac{1}{2}\hat{\omega}_i^a\phi_\alpha^a\hat{\omega}_i^b\phi_\alpha^b+i\sqrt{p}\,\hat{\omega}_i^a\omega_i^a}$$
$$\times\, e^{-\frac{1}{2}\log\det\mathbf{C}_\Omega}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\Big[e^{-\frac{\mu_1^2 p}{2nm}\omega_i^a\omega_j^a\mathbf{Y}_i^{\nu\mu}\mathbf{Y}_j^{\nu\mu}+\frac{1}{2}\mathbf{J}_\Omega^T\mathbf{C}_\Omega^{-1}\mathbf{J}_\Omega}\Big]. \tag{97}$$

Average over $\mathbf{x}$ and $\boldsymbol{\xi}$. With $\mathbf{Y}_i^{\nu\mu} = e^{-t}\boldsymbol{\Sigma}_{ij}^{1/2}\mathbf{x}_j^\nu\boldsymbol{v}^\mu + \sqrt{\Delta_t}\boldsymbol{\xi}_i^{\nu\mu}$, the data covariance and the quadratic form in the exponent read

$$\mathbf{K}_{ij}^{\nu\mu\nu'\mu'} = \mathbb{E}[\mathbf{Y}_i^{\nu\mu}\mathbf{Y}_j^{\nu'\mu'}] = \boldsymbol{I}_n\otimes\big(e^{-2t}\boldsymbol{\Sigma}\otimes\mathbf{1}_m + \Delta_t\boldsymbol{I}_d\otimes\boldsymbol{I}_m\big), \tag{98}$$

$$(\mathbf{C}_y)_{ij}^{\nu\mu\nu'\mu'} = \frac{\mu_1^2 p}{nm}\omega_i^a\omega_j^a\,\boldsymbol{I}_n\otimes\boldsymbol{I}_m - \frac{\mu_1^2\mu_*^2 p}{(nm)^2}\omega_i^a\phi_\alpha^a\,\omega_j^b\phi_\beta^b\,(\mathbf{C}_\Omega^{-1})_{\alpha\beta}^{\nu\mu\nu'\mu'}. \tag{99}$$

Integrating out $\mathbf{x},\boldsymbol{\xi}$ produces a determinant,

$$\mathbb{E}[\mathcal{Z}^s] = \int\prod_a \mathrm{d}\phi^a\mathrm{d}\omega^a\mathrm{d}\hat{\omega}^a\; e^{\frac{z}{2}\phi^a\cdot\phi^a+\frac{p\zeta}{2d}\omega^a\cdot\omega^a-\frac{1}{2}\hat{\omega}_i^a\phi_\alpha^a\hat{\omega}_i^b\phi_\alpha^b+i\sqrt{p}\,\hat{\omega}_i^a\omega_i^a}$$
$$\times\, e^{-\frac{1}{2}\log\det\mathbf{C}_\Omega}\,e^{-\frac{1}{2}\log\det(\boldsymbol{I}+\mathbf{K}\mathbf{C}_y)}. \tag{100}$$

Integrating over $\hat{\omega}^a$. The $\hat{\omega}$-integral is Gaussian,

$$\int\mathrm{d}\hat{\omega}^a\; e^{-\frac{1}{2}\hat{\omega}_i^a\phi_\alpha^a\hat{\omega}_i^b\phi_\alpha^b+i\sqrt{p}\,\hat{\omega}_i^a\omega_i^a} = e^{-\frac{d}{2}\log\det(\phi^a\cdot\phi^b)}\,e^{-\frac{p}{2}\omega_i^a\omega_i^b(\phi^a\cdot\phi^b)^{-1}}. \tag{101}$$

Introducing the overlaps between replicas. We insert

$$1 = \int\prod_{ab}\frac{\mathrm{d}\mathbf{Q}^{ab}\mathrm{d}\hat{\mathbf{Q}}^{ab}}{2\pi i/p}\; e^{\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi_\alpha^a\phi_\alpha^b)}, \tag{102}$$

$$1 = \int\prod_{ab}\frac{\mathrm{d}\mathbf{R}^{ab}\mathrm{d}\hat{\mathbf{R}}^{ab}}{2\pi i/d}\; e^{\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega_i^a\omega_i^b)}, \qquad 1 = \int\prod_{ab}\frac{\mathrm{d}\mathbf{H}^{ab}\mathrm{d}\hat{\mathbf{H}}^{ab}}{2\pi i/d}\; e^{\hat{\mathbf{H}}^{ab}(d\mathbf{H}^{ab}-\omega_i^a\boldsymbol{\Sigma}_{ij}\omega_j^b)}, \tag{103}$$

so that

$$\mathbb{E}[\mathcal{Z}^s] = \int\prod_a \mathrm{d}\phi^a\mathrm{d}\omega^a\,\mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\mathrm{d}\mathbf{R}\mathrm{d}\hat{\mathbf{R}}\mathrm{d}\mathbf{H}\mathrm{d}\hat{\mathbf{H}}$$
$$\times\, e^{\hat{\mathbf{Q}}^{ab}(p\mathbf{Q}^{ab}-\phi_\alpha^a\phi_\alpha^b)+\hat{\mathbf{R}}^{ab}(d\mathbf{R}^{ab}-\omega_i^a\omega_i^b)+\hat{\mathbf{H}}^{ab}(d\mathbf{H}^{ab}-\omega_i^a\boldsymbol{\Sigma}_{ij}\omega_j^b)}$$
$$\times\, e^{\frac{pz}{2}\operatorname{Tr}\mathbf{Q}+\frac{p\zeta}{2}\operatorname{Tr}\mathbf{R}}\,e^{-\frac{d}{2}\log\det\mathbf{Q}}\,e^{-\frac{d}{2}\operatorname{Tr}(\mathbf{R}\mathbf{Q}^{-1})}\,e^{-\frac{1}{2}\log\det(\cdot)}, \tag{104}$$

where the remaining determinant, after simplification, is

$$\log\det(\cdot) = n\log\det\Big(\delta^{ab}[\kappa\mathbf{1}_m+(1-\kappa)\boldsymbol{I}_m]^{-1} + \frac{\mu_*^2\psi_p}{\psi_n m}\mathbf{Q}^{ab}\boldsymbol{I}_m$$
$$-\frac{\mu_1^2\mu_*^2\psi_p^2}{\psi_n^2 m^2}\big(e^{-2t}\mathbf{H}^{ac}\mathbf{1}_m+\Delta_t\mathbf{R}^{ac}\boldsymbol{I}_m\big)\mathbf{Q}^{cb}$$
$$+\frac{\mu_1^2\psi_p}{\psi_n m}\big[e^{-2t}\mathbf{H}\mathbf{1}_m+\Delta_t\mathbf{R}\boldsymbol{I}_m\big]\big[\delta^{ab}(\kappa\mathbf{1}_m+(1-\kappa)\boldsymbol{I}_m)^{-1}+\frac{\mu_*^2\psi_p}{\psi_n m}\mathbf{Q}^{ab}\boldsymbol{I}_m\big]\Big). \tag{105}$$

Integrating out ϕ^a,ω^a. These are now Gaussian,

$$\int\prod_a\mathrm{d}\phi^a\mathrm{d}\omega^a\; e^{-\hat{\mathbf{Q}}^{ab}\phi_\alpha^a\phi_\alpha^b-\hat{\mathbf{R}}^{ab}\omega_i^a\omega_i^b-\hat{\mathbf{H}}^{ab}\omega_i^a\boldsymbol{\Sigma}_{ij}\omega_j^b} = e^{-\frac{p}{2}\log\det\hat{\mathbf{Q}}}\,e^{-\frac{1}{2}\log\det(\hat{\mathbf{R}}\otimes\boldsymbol{I}_d+\hat{\mathbf{H}}\otimes\boldsymbol{\Sigma})}, \tag{106}$$

leaving a single integral over $(\mathbf{Q},\hat{\mathbf{Q}},\mathbf{R},\hat{\mathbf{R}},\mathbf{H},\hat{\mathbf{H}})$.

Replica-symmetric diagonal Ansatz. We adopt the replica-symmetric diagonal ansatz (Mézard et al., 1987) $\mathbf{Q}^{ab} = q\delta^{ab}$, $\mathbf{R}^{ab} = r\delta^{ab}$, $\mathbf{H}^{ab} = h\delta^{ab}$ (and likewise for the conjugates).[12] Using $\log\det(a\boldsymbol{I}_m + b\mathbf{1}_m) = \log(a+mb) + (m-1)\log a$, the saddle on $\hat{q}$ enforces $\hat{q} = 1/(2q)$, and the effective action (divided by $2/sd$) reads

$$\begin{aligned} S(q,r,\hat{r},h,\hat{h}) = &-\psi_p \log q - r\hat{r} - h\hat{h} + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\log(\hat{r}+\lambda\hat{h}) - \psi_p z q - \psi_p \zeta r + \log q + \frac{r}{q} \\ &+ \psi_n \log\left[1 + \frac{\mu_1^2 e^{-2t}\psi_p h}{\psi_n} + \frac{\mu_1^2 \Delta_t \psi_p r}{\psi_n m} + \big(1+(m-1)\kappa\big)\frac{\mu_*^2\psi_p q}{\psi_n m}\right] \\ &+ \psi_n(m-1)\log\left[1 + \frac{\mu_1^2\Delta_t\psi_p r}{\psi_n m} + (1-\kappa)\frac{\mu_*^2\psi_p q}{\psi_n m}\right]. \end{aligned} \tag{107}$$

Saddle-point equations. Define

$$L_1(q,r,h) = 1 + \frac{\mu_1^2 e^{-2t}\psi_p h}{\psi_n} + \frac{\mu_1^2\Delta_t\psi_p r}{\psi_n m} + \big(1+(m-1)\kappa\big)\frac{\mu_*^2\psi_p q}{\psi_n m}, \tag{108}$$

$$L_2(q,r) = 1 + \frac{\mu_1^2\Delta_t\psi_p r}{\psi_n m} + (1-\kappa)\frac{\mu_*^2\psi_p q}{\psi_n m}. \tag{109}$$

Stationarity of Eq.107 gives

$$\partial_q S = 0: \quad \psi_p z = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{(1+(m-1)\kappa)\mu_*^2\psi_p/m}{L_1} + \frac{(m-1)(1-\kappa)\mu_*^2\psi_p/m}{L_2}, \tag{110}$$

$$\partial_r S = 0: \quad \hat{r} = \frac{1}{q} - \psi_p\zeta + \frac{\mu_1^2\psi_p\Delta_t/m}{L_1} + \frac{(m-1)\mu_1^2\psi_p\Delta_t/m}{L_2}, \tag{111}$$

$$\partial_h S = 0: \quad \hat{h} = \frac{\mu_1^2\psi_p e^{-2t}}{L_1}, \tag{112}$$

$$\partial_{\hat{r}} S = 0: \quad r = \int \frac{\mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)}{\hat{r}+\lambda\hat{h}}, \qquad \partial_{\hat{h}} S = 0: \quad h = \int \frac{\lambda\,\mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)}{\hat{r}+\lambda\hat{h}}. \tag{113}$$

For the isotropic case $\rho_{\mathbf{\Sigma}} = \delta(\lambda-1)$ one has $r = h = 1/(\hat{r}+\hat{h})$, i.e. $\hat{r}+\hat{h} = 1/r$, and $h = r$ so $L_1(q,r,r) = L_1(q,r)$. Adding the $\hat{r}$ and $\hat{h}$ equations,

$$\frac{1}{r} + \psi_p\zeta = \frac{1}{q} + \frac{\mu_1^2\psi_p(e^{-2t} + \frac{\Delta_t}{m})}{L_1(q,r)} + \frac{(m-1)\mu_1^2\psi_p\Delta_t/m}{L_2(q,r)}, \tag{114}$$

$$\psi_p z = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{(1+(m-1)\kappa)\mu_*^2\psi_p/m}{L_1(q,r)} + \frac{(m-1)(1-\kappa)\mu_*^2\psi_p/m}{L_2(q,r)}, \tag{115}$$

with $L_1(q,r) = 1 + \frac{\psi_p}{\psi_n m}[\mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1+(m-1)\kappa)]$ and $L_2(q,r) = 1 + \frac{\psi_p}{\psi_n m}[\mu_1^2 r\Delta_t + \mu_*^2 q(1-\kappa)]$. □

C.8 Computation of the traces

C.8.1 Computation of T_1, T_4 and T_5

We compute here the generating trace Eq.75 at $\epsilon = 0$, which by Eq.76 also yields $T_4 = \partial_\zeta T_1$ and $T_5 = \partial_z T_1$,

$$T_1(z,\zeta,0) = \frac{1}{d(nm)^2}\,\mathbb{E}\Big[\mathrm{Tr}\big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}_{z,\zeta,0}\,\mathbf{F}\boldsymbol{\xi}^T\big)\Big], \quad \mathcal{G}_{z,\zeta,0} = \Big(\frac{\mathbf{F}\mathbf{F}^T}{nm} - z\boldsymbol{I}_p - \zeta\frac{\mathbf{W}\mathbf{W}^T}{d}\Big)^{-1}. \tag{116}$$

Using the replica representation of an inverse matrix Eq.27,

$$(\mathcal{G}_{z,\zeta,0})_{\alpha\beta} = \int \prod_a \mathrm{d}\phi^a\; \phi^1_\alpha\phi^1_\beta\, e^{-\frac{1}{2}\phi^a_{\alpha'}(\mathcal{G}^{-1}_{z,\zeta,0})_{\alpha'\beta'}\phi^a_{\beta'}},$$

[12] The off-diagonal overlaps $\mathbf{Q}^{ab}$ for $a \neq b$ vanish at the saddle point; see e.g. D'Ascoli et al. (2020).

we obtain

$$T_1(z,\zeta,0) = \frac{1}{dnm}\,\boldsymbol{\xi}_i^{\nu\mu}\,\frac{\mathbf{F}_\alpha^{\nu\mu}}{\sqrt{nm}} \times \int \prod_a \mathrm{d}\phi^a\ \phi_\alpha^1\phi_\beta^1\ e^{-\frac{1}{2}\phi_{\alpha'}^a\left(\frac{\mathbf{F}_{\alpha'}^{\nu'\mu'}\mathbf{F}_{\beta'}^{\nu'\mu'}}{nm} - z\delta_{\alpha'\beta'} - \zeta\frac{\mathbf{W}_{\alpha' k}\mathbf{W}_{\beta' k}}{d}\right)\phi_{\beta'}^a}\,\frac{\mathbf{F}_\beta^{\nu''\mu''}}{\sqrt{nm}}\boldsymbol{\xi}_i^{\nu''\mu''}. \quad (117)$$

We decouple the dependence on $\mathbf{W}$ with the auxiliary field $1 = \int \mathrm{d}\omega^a \mathrm{d}\hat{\omega}^a\, e^{i\hat{\omega}_i^a(\sqrt{p}\,\omega_i^a - \phi_\alpha^a \mathbf{W}_{\alpha i})}$ and replace the contraction $\mathbf{F}_\alpha^{\nu\mu}\phi_\alpha^a$ by the scalar Gaussian field

$$\mathbf{G}_a^{\nu\mu} = \mu_1\sqrt{p}\,\frac{\omega_i^a \mathbf{Y}_i^{\nu\mu}}{\sqrt{d}} + \mu_*\,\phi_\alpha^a \boldsymbol{\Omega}_\alpha^{\nu\mu}. \quad (118)$$

As in the proof of Proposition C.2, we introduce the replica overlaps $\mathbf{Q}^{ab} = \frac{1}{p}\phi_\alpha^a\phi_\alpha^b$ and $\mathbf{R}^{ab} = \frac{1}{d}\omega_i^a\omega_i^b$. The covariances of $\mathbf{G}$ follow from $\mathbb{E}[\mathbf{Y}_i^{\nu\mu}\mathbf{Y}_j^{\nu'\mu'}]$ and the GEP covariance of $\boldsymbol{\Omega}$,

$$\mathbb{E}[\mathbf{G}_a^{\nu\mu}\mathbf{G}_b^{\nu'\mu'}] = \mu_1^2 p\,\mathbf{R}^{ab}\delta^{\nu\nu'}\left(e^{-2t}\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + \Delta_t\delta^{\mu\mu'}\right) + \mu_*^2 p\,\mathbf{Q}^{ab}\delta^{\nu\nu'}\left(\kappa\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + (1-\kappa)\delta^{\mu\mu'}\right), \quad (119)$$

$$\mathbb{E}[\mathbf{G}_a^{\nu\mu}\boldsymbol{\xi}_i^{\nu'\mu'}] = \mu_1\sqrt{\tfrac{p}{d}}\,\omega_i^a\sqrt{\Delta_t}\,\delta^{\nu\nu'}\delta^{\mu\mu'}. \quad (120)$$

Evaluating at the replica-symmetric saddle $(\mathbf{Q}^{11}, \mathbf{R}^{11}) = (q, r)$ of Proposition C.2, we only need to compute the non-exponential prefactor: the action and the saddle-point equations are identical to those of Proposition C.2. Writing $\langle\cdot\rangle_*$ for the Gaussian average over $(\boldsymbol{\xi}, \mathbf{G})$ *tilted* by the replica weight $e^{-\frac{1}{2nm}\mathbf{G}_a\mathbf{G}_a^T}$,

$$\langle\mathcal{O}\rangle_* := \frac{\mathbb{E}\left[\mathcal{O}\,e^{-\frac{1}{2nm}\mathbf{G}_a\mathbf{G}_a^T}\right]}{\mathbb{E}\left[e^{-\frac{1}{2nm}\mathbf{G}_a\mathbf{G}_a^T}\right]}, \quad (121)$$

the prefactor reads

$$T_1(z,\zeta,0) = \frac{1}{nm}\left\langle\boldsymbol{\xi}_i^{\nu\mu}\boldsymbol{\xi}_i^{\nu'\mu'}\mathbf{G}_1^{\nu\mu}\mathbf{G}_1^{\nu'\mu'}\right\rangle_*. \quad (122)$$

The covariance of $\mathbf{G}_1$ over the noise index μ is the $m\times m$ matrix, where $\boldsymbol{v}^\mu = 1$ denotes that $\boldsymbol{v}\in\mathbb{R}^m$ is the all-ones vector (since each data point $\mathbf{x}^\nu$ is shared identically across all m noise copies),

$$(\boldsymbol{\Sigma}_G)^{\mu\mu'} = \mu_1^2 r\left(e^{-2t}\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + \Delta_t\delta^{\mu\mu'}\right) + \mu_*^2 q\left(\kappa\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + (1-\kappa)\delta^{\mu\mu'}\right). \quad (123)$$

Since $\boldsymbol{\Sigma}_G$ has rank-1 plus diagonal structure, the matrix $\boldsymbol{\Psi}_m = (\boldsymbol{I}_m + \frac{p}{nm}\boldsymbol{\Sigma}_G)^{-1}$ has exactly two distinct eigenvalues whose inverses are

$$L_1 = 1 + \tfrac{\psi_p}{\psi_n m}\left[\mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1 + (m-1)\kappa)\right] \text{ (eigenvector } \boldsymbol{v}), \quad (124)$$

$$L_2 = 1 + \tfrac{\psi_p}{\psi_n m}\left[\mu_1^2 r\Delta_t + \mu_*^2 q(1-\kappa)\right] \text{ (mult. } m-1), \quad (125)$$

so $\operatorname{Tr}\boldsymbol{\Psi}_m = \frac{1}{L_1} + \frac{m-1}{L_2}$. The measure $\langle\cdot\rangle_*$ is Gaussian. At the RS saddle, $\boldsymbol{\xi}^{\nu\mu}$ is a standard Gaussian and $\mathbf{G}_1^{\nu\mu} = \mu_1\sqrt{p}\,\omega_i^1\mathbf{Y}_i^{\nu\mu}/\sqrt{d} + \mu_*\phi_\alpha^1\boldsymbol{\Omega}_\alpha^{\nu\mu}$ is a linear function of the Gaussian variables $(\boldsymbol{\xi}, \mathbf{x}^\nu, \boldsymbol{\Omega})$, so $(\boldsymbol{\xi}, \mathbf{G}_1)$ are jointly Gaussian. The tilt $e^{-\frac{1}{2nm}\sum_{\nu,\mu}(\mathbf{G}_1^{\nu\mu})^2}$ is a quadratic exponential in $\mathbf{G}_1$; tilting a Gaussian by a Gaussian weight yields another Gaussian (it shifts the precision matrix of $\mathbf{G}_1$ by $\frac{1}{nm}\boldsymbol{I}$), so $\langle\cdot\rangle_*$ is a Gaussian measure and Wick's theorem applies.

According to Wick's theorem, for jointly Gaussian fields, $\langle A_1A_2B_1B_2\rangle = \langle A_1A_2\rangle\langle B_1B_2\rangle + \langle A_1B_1\rangle\langle A_2B_2\rangle + \langle A_1B_2\rangle\langle A_2B_1\rangle$. With $A_k = \boldsymbol{\xi}_i^{\nu_k\mu_k}$ and $B_k = \mathbf{G}_1^{\nu_k\mu_k}$,

$$\left\langle\boldsymbol{\xi}_i^{\nu\mu}\boldsymbol{\xi}_i^{\nu'\mu'}\mathbf{G}_1^{\nu\mu}\mathbf{G}_1^{\nu'\mu'}\right\rangle_* = \underbrace{\langle\boldsymbol{\xi}^{\nu\mu}\boldsymbol{\xi}^{\nu'\mu'}\rangle_*\langle\mathbf{G}^{\nu\mu}\mathbf{G}^{\nu'\mu'}\rangle_*}_{\text{P1}} + \underbrace{\langle\boldsymbol{\xi}^{\nu\mu}\mathbf{G}^{\nu\mu}\rangle_*\langle\boldsymbol{\xi}^{\nu'\mu'}\mathbf{G}^{\nu'\mu'}\rangle_*}_{\text{P2}} \quad (126)$$

$$+ \underbrace{\langle\boldsymbol{\xi}^{\nu'\mu'}\mathbf{G}^{\nu\mu}\rangle_*\langle\boldsymbol{\xi}^{\nu\mu}\mathbf{G}^{\nu'\mu'}\rangle_*}_{\text{P3}}. \quad (127)$$

The tilt Eq.121 shifts the two-point functions. For jointly Gaussian $(\boldsymbol{\xi}, \mathbf{G})$, tilting by $e^{-\frac{1}{2nm}\mathbf{G}\mathbf{G}^T}$ shifts the precision of $\mathbf{G}$ by $\frac{1}{nm}\boldsymbol{I}$: $\boldsymbol{\Sigma}_{GG,*} = (\boldsymbol{\Sigma}_{GG}^{-1} + \frac{1}{nm}\boldsymbol{I})^{-1}$, $\boldsymbol{\Sigma}_{\xi G,*} = \boldsymbol{\Sigma}_{\xi G}(\boldsymbol{I} + \frac{1}{nm}\boldsymbol{\Sigma}_{GG})^{-1}$, and $\boldsymbol{\Sigma}_{\xi\xi}$ receives an $O(1/nm)$ correction. Since $\boldsymbol{\Sigma}_{GG} = p\boldsymbol{\Sigma}_G$, one has $(\boldsymbol{I} + \frac{p}{nm}\boldsymbol{\Sigma}_G)^{-1} = \boldsymbol{\Psi}_m$; the noise therefore stays standard, $\langle \boldsymbol{\xi}_i^{\nu\mu}\boldsymbol{\xi}_j^{\nu'\mu'}\rangle_* = \delta_{ij}\delta^{\nu\nu'}\delta^{\mu\mu'} + O(1/nm)$, while

$$\langle \mathbf{G}^{\nu\mu}\mathbf{G}^{\nu'\mu'}\rangle_* = \delta^{\nu\nu'}\, p\,(\boldsymbol{\Sigma}_G\boldsymbol{\Psi}_m)^{\mu\mu'}, \qquad \langle \boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_1^{\nu'\mu'}\rangle_* = \mu_1\sqrt{\psi_p\Delta_t}\,\omega_i^1\,\delta^{\nu\nu'}(\boldsymbol{\Psi}_m)^{\mu\mu'}. \tag{128}$$

Since $\boldsymbol{\Psi}_m^{-1} = \boldsymbol{I}_m + \frac{p}{nm}\boldsymbol{\Sigma}_G$, multiplying on the right by $\boldsymbol{\Psi}_m$ gives $\boldsymbol{I}_m = \boldsymbol{\Psi}_m + \frac{p}{nm}\boldsymbol{\Sigma}_G\boldsymbol{\Psi}_m$, i.e. $p\,\boldsymbol{\Sigma}_G\boldsymbol{\Psi}_m = nm(\boldsymbol{I}_m - \boldsymbol{\Psi}_m)$. The diagonal pairing then collapses (Einstein summation, $\delta_{ii} = d$):

$$\mathrm{P1} = \frac{1}{d(nm)^2}\,\delta_{ii}\delta^{\nu\nu'}\delta^{\mu\mu'}\,\delta^{\nu\nu'}p(\boldsymbol{\Sigma}_G\boldsymbol{\Psi}_m)^{\mu\mu'} = \frac{dnp}{d(nm)^2}\,\mathrm{Tr}(\boldsymbol{\Sigma}_G\boldsymbol{\Psi}_m) \tag{129}$$

$$= \frac{1}{m}\big[m - \mathrm{Tr}\,\boldsymbol{\Psi}_m\big] = 1 - \frac{\mathrm{Tr}\,\boldsymbol{\Psi}_m}{m}. \tag{130}$$

The crossed pairings use $\frac{1}{d}\sum_i(\omega_i^1)^2 = r$:

$$\mathrm{P2} = \frac{1}{d(nm)^2}\langle \boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_1^{\nu\mu}\rangle_*\langle \boldsymbol{\xi}_i^{\nu'\mu'}\mathbf{G}_1^{\nu'\mu'}\rangle_* = \frac{\mu_1^2\psi_p\Delta_t}{d(nm)^2}(dr)\,n^2(\mathrm{Tr}\,\boldsymbol{\Psi}_m)^2 = \frac{\mu_1^2\psi_p\Delta_t r}{m^2}(\mathrm{Tr}\,\boldsymbol{\Psi}_m)^2, \tag{131}$$

$$\begin{aligned}\mathrm{P3} &= \frac{1}{d(nm)^2}\langle \boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_1^{\nu'\mu'}\rangle_*\langle \boldsymbol{\xi}_i^{\nu'\mu'}\mathbf{G}_1^{\nu\mu}\rangle_* \\ &= \frac{\mu_1^2\psi_p\Delta_t}{d(nm)^2}\,\omega_i^1\omega_i^1\,\delta^{\nu\nu'}\delta^{\nu'\nu}\,(\boldsymbol{\Psi}_m)^{\mu\mu'}(\boldsymbol{\Psi}_m)^{\mu'\mu} \\ &= \frac{\mu_1^2\psi_p\Delta_t}{d(nm)^2}\cdot dr\cdot n\cdot\mathrm{Tr}(\boldsymbol{\Psi}_m^2) = \frac{\mu_1^2\psi_p\Delta_t}{nm^2}\,r\;\mathrm{Tr}(\boldsymbol{\Psi}_m^2) = O(1/n),\end{aligned} \tag{132}$$

In the second line of P3, the two Kronecker deltas $\delta^{\nu\nu'}\delta^{\nu'\nu} = \delta^{\nu\nu'}$ force $\nu = \nu'$, collapsing the double sum over samples to n terms; then $\frac{1}{d}\sum_i(\omega_i^1)^2 = r$ and $\sum_{\mu,\mu'}(\boldsymbol{\Psi}_m)^{\mu\mu'}(\boldsymbol{\Psi}_m)^{\mu'\mu} = \mathrm{Tr}(\boldsymbol{\Psi}_m^2) = \frac{1}{L_1^2} + \frac{m-1}{L_2^2} = O(1)$. The resulting factor $1/n$ makes P3 subleading. Collecting P1 + P2 (and dropping the subleading P3),

$$T_1(z,\zeta,0) = f_1\big(q,r,0\big), \quad f_1(q,r,0) = 1 - \frac{\mathcal{T}}{m} + \frac{\mu_1^2\psi_p\Delta_t r}{m^2}\,\mathcal{T}^2, \quad \mathcal{T} = \mathrm{Tr}\,\boldsymbol{\Psi}_m = \frac{1}{L_1} + \frac{m-1}{L_2}, \tag{133}$$

which is claim *(ii)* of Theorem 2.1 for T_1 at $\epsilon = 0$; the ϵ-dependence of f_1 is obtained in Appendix C.8.3 below. Here $(q,r) = (q(z,\zeta,0), r(z,\zeta,0))$ is the saddle of Proposition C.2, and $T_4 = \partial_\zeta T_1|_{(-\lambda,0,0)}$, $T_5 = \partial_z T_1|_{(-\lambda,0,0)}$ follow by differentiating f_1 through that saddle.

C.8.2 Computation of T_2

We wish to compute $T_2 = \frac{1}{dnm}\,\mathrm{Tr}\big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}_{-\lambda,0,0}\frac{\mathbf{W}}{\sqrt{d}}\big)$ of Eq.75, with $\mathcal{G}_{-\lambda,0,0} = (\frac{\mathbf{F}\mathbf{F}^T}{nm} + \lambda\boldsymbol{I}_p)^{-1}$. Applying the replica representation Eq.27 to $(\mathcal{G}_{-\lambda,0,0})_{\alpha\beta}$:

$$T_2 = \frac{1}{dnm}\,\mathbb{E}\left[\boldsymbol{\xi}_i^{\nu\mu}\mathbf{F}_\alpha^{\nu\mu}\int\prod_a \mathrm{d}\phi^a\;\phi_\alpha^1\phi_\beta^1\;e^{-\frac{1}{2}\phi_{\alpha'}^a\left(\frac{\mathbf{F}_{\alpha'}^{\nu'\mu'}\mathbf{F}_{\beta'}^{\nu'\mu'}}{nm}+\lambda\delta_{\alpha'\beta'}\right)\phi_{\beta'}^a}\,\frac{\mathbf{W}_{i\beta}}{\sqrt{d}}\right]. \tag{134}$$

We introduce the auxiliary field $1 = \int \mathrm{d}\omega^a\mathrm{d}\hat{\omega}^a\,e^{i\hat{\omega}_i^a(\sqrt{p}\,\omega_i^a - \phi_\alpha^a\mathbf{W}_{\alpha i})}$ and the replica overlaps $\mathbf{Q}^{ab} = \frac{1}{p}\phi_\alpha^a\phi_\alpha^b$, $\mathbf{R}^{ab} = \frac{1}{d}\omega_i^a\omega_i^b$, and replace $\mathbf{F}_\alpha^{\nu\mu}\phi_\alpha^a$ by

$$\mathbf{G}_a^{\nu\mu} = \mu_1\sqrt{p}\,\frac{\omega_i^a\mathbf{Y}_i^{\nu\mu}}{\sqrt{d}} + \mu_*\,\phi_\alpha^a\boldsymbol{\Omega}_\alpha^{\nu\mu}. \tag{135}$$

The contraction $\phi_\beta^1\frac{\mathbf{W}_{i\beta}}{\sqrt{d}} = \sqrt{\psi_p}\,\omega_i^1$ then follows from the decoupling. The exponential action is identical to that in the T_1 computation above, so evaluating at the saddle (q,r) we only need the

prefactor. Since $\sqrt{\psi_p}\,\omega_i^1$ is constant with respect to the $\langle\cdot\rangle_*$ average, it factors out, and the two-point function $\langle \boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_1^{\nu\mu}\rangle_* = \mu_1\sqrt{\psi_p\Delta_t}\,\omega_i^1\,(\boldsymbol{\Psi}_m)^{\mu\mu}$ gives

$$T_2 = \frac{\sqrt{\psi_p}}{dnm}\sum_{i,\nu,\mu}\omega_i^1\,\langle \boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_1^{\nu\mu}\rangle_* = \frac{\psi_p\mu_1\sqrt{\Delta_t}}{dnm}\sum_{i,\nu,\mu}(\omega_i^1)^2\,(\boldsymbol{\Psi}_m)^{\mu\mu}. \tag{136}$$

Summing the indices: $\frac{1}{d}\sum_i(\omega_i^1)^2 = r$, the n sample indices contribute a factor n, and $\sum_\mu(\boldsymbol{\Psi}_m)^{\mu\mu} = \operatorname{Tr}\boldsymbol{\Psi}_m$, so

$$T_2 = f_2(q,r), \qquad f_2(q,r) = \frac{\psi_p}{m}\mu_1\sqrt{\Delta_t}\,r\,\operatorname{Tr}\boldsymbol{\Psi}_m = \frac{\psi_p}{m}\mu_1\sqrt{\Delta_t}\,r\left(\frac{1}{L_1(q,r)} + \frac{m-1}{L_2(q,r)}\right), \tag{137}$$

with $(q,r) = (q(-\lambda,0,0), r(-\lambda,0,0))$. This is claim *(ii)* of Theorem 2.1 for T_2.

C.8.3 Computation of T_3

Recall from Eq.76 that $T_3 = -\partial_\epsilon T_1|_{(-\lambda,0,0)}$, so that computing it amounts to switching on the third deformation of the resolvent Eq.74. Recall also that $\mathbf{H} = \mathbb{E}_{\boldsymbol{\xi}}[\mathbf{F}]$ has entries $\mathbf{H}_\alpha^\nu = e^{-t}\mu_1\frac{\mathbf{W}_\alpha\cdot\mathbf{x}^\nu}{\sqrt{d}} + \mu_*\sqrt{\kappa}\,\boldsymbol{\eta}_\alpha^\nu$, where $\boldsymbol{\eta}_\alpha^\nu \sim \mathcal{N}(0,1)$ is independent of $\mathbf{W}$, $\mathbf{x}^\nu$, and $\boldsymbol{\xi}^{\nu\mu}$ (see Appendix C.5). The correlations of $\boldsymbol{\eta}$ with the other random variables are

$$\mathbb{E}[\boldsymbol{\eta}_\alpha^\nu\boldsymbol{\eta}_\beta^{\nu'}] = \delta_{\alpha\beta}\delta^{\nu\nu'}, \quad \mathbb{E}[\boldsymbol{\Omega}_\alpha^{\nu\mu}\boldsymbol{\eta}_\beta^{\nu'}] = \sqrt{\kappa}\,\delta_{\alpha\beta}\delta^{\nu\nu'}, \quad \mathbb{E}[\boldsymbol{\xi}_i^{\nu\mu}\boldsymbol{\eta}_\alpha^{\nu'}] = 0. \tag{138}$$

Writing the full resolvent $\mathcal{G}_{z,\zeta,\epsilon}$ of Eq.74 with the replica representation, introducing $\omega = \mathbf{W}^T\phi/\sqrt{p}$ as before, and denoting $\mathbf{G}_a^{\nu\mu} = \mu_1\sqrt{p}\,\frac{\omega_i^a\mathbf{Y}_i^{\nu\mu}}{\sqrt{d}} + \mu_*\phi_\alpha^a\boldsymbol{\Omega}_\alpha^{\nu\mu}$ as in the computation of $T_1(z,\zeta,0)$,

$$T_1(z,\zeta,\epsilon) = \frac{1}{d(nm)^2}\mathbb{E}\int\prod_a \mathrm{d}\phi^a\mathrm{d}\omega^a\mathrm{d}\hat{\omega}^a\; e^{i\hat{\omega}^a(\sqrt{p}\,\omega^a - \phi^a\mathbf{W})}\;\operatorname{Tr}(\boldsymbol{\xi}\mathbf{G}_1\mathbf{G}_1^T\boldsymbol{\xi}^T)$$
$$\times\, e^{\frac{z}{2}\phi^a\cdot\phi^a + \frac{\zeta}{2d}\phi^{aT}\mathbf{W}\mathbf{W}^T\phi^a - \frac{1}{2nm}\mathbf{G}_a\mathbf{G}_a^T - \frac{\epsilon}{2n}\mathbf{H}_\alpha^\nu\phi_\alpha^a\mathbf{H}_\beta^\nu\phi_\beta^a}. \tag{139}$$

We introduce the replica overlaps $\mathbf{Q}^{ab} = \frac{1}{p}\phi_\alpha^a\phi_\alpha^b$, $\mathbf{R}^{ab} = \frac{1}{d}\omega_i^a\omega_i^b$, and denote by $\mathbf{L}_a^\nu = \mathbf{H}_\alpha^\nu\phi_\alpha^a = e^{-t}\mu_1\sqrt{\psi_p}\,\omega_i^a\mathbf{x}_i^\nu + \mu_*\sqrt{\kappa}\,\phi_\alpha^a\boldsymbol{\eta}_\alpha^\nu$ the Gaussian field associated with the term $\epsilon\frac{\mathbf{H}\mathbf{H}^T}{n}$ in $\mathcal{G}_{z,\zeta,\epsilon}$, with covariances

$$\begin{aligned}
\mathbb{E}[\mathbf{L}^\nu\mathbf{L}^{\nu'}] &= \delta^{\nu\nu'}p\,\sigma_H^2, & \mathbb{E}[\mathbf{G}^{\nu\mu}\mathbf{L}^{\nu'}] &= \delta^{\nu\nu'}p\,\sigma_H^2\,\boldsymbol{v}^\mu,\\
\mathbb{E}[\boldsymbol{\xi}_i^{\nu\mu}\mathbf{G}_a^{\nu'\mu'}] &= \mu_1\sqrt{\psi_p\Delta_t}\,\omega_i^a\delta^{\nu\nu'}\delta^{\mu\mu'}, & \mathbb{E}[\boldsymbol{\xi}^{\nu\mu}\mathbf{L}^{\nu'}] &= 0,
\end{aligned} \tag{140}$$

where $r = \mathbf{R}^{11}$ and $q = \mathbf{Q}^{11}$ are the RS saddle values and $\sigma_H^2 = \mu_1^2e^{-2t}r + \mu_*^2\kappa q$. The joint covariance matrix of $(\mathbf{G},\mathbf{L})$ is accordingly

$$\boldsymbol{\Sigma}_{(\mathbf{G},\mathbf{L})} = \begin{pmatrix} p\boldsymbol{\Sigma}_G & p\sigma_H^2\,\boldsymbol{v} \\ p\sigma_H^2\,\boldsymbol{v}^T & p\sigma_H^2 \end{pmatrix}. \tag{141}$$

Computation of the action. We denote $\boldsymbol{\Sigma}_G$ the covariance matrix of $\mathbf{G}$, $\mathbb{E}[\mathbf{G}_1^{\nu\mu}\mathbf{G}_1^{\nu'\mu'}] = \delta^{\nu\nu'}(\boldsymbol{\Sigma}_G)^{\mu\mu'}$,

$$(\boldsymbol{\Sigma}_G)^{\mu\mu'} = \mu_1^2 r\big(e^{-2t}\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + \Delta_t\delta^{\mu\mu'}\big) + \mu_*^2 q\big(\kappa\boldsymbol{v}^\mu\boldsymbol{v}^{\mu'} + (1-\kappa)\delta^{\mu\mu'}\big), \tag{142}$$

and $\boldsymbol{\Psi}_m = (\boldsymbol{I}_m + \frac{p}{nm}\boldsymbol{\Sigma}_G)^{-1}$ with eigenvalues $L_1(q,r)$ (eigenvector $\boldsymbol{v}$) and $L_2(q,r)$ (multiplicity $m-1$), as derived in the computation of $T_1(z,\zeta,0)$ above. The precision matrix of $(\mathbf{G},\mathbf{L})$ is shifted by $\mathbf{K}_{(\mathbf{G},\mathbf{L})} = \operatorname{diag}(\frac{1}{nm}\boldsymbol{I}_m, \frac{\epsilon}{n})$, so $\mathbb{E}[e^{-\frac{1}{2nm}\mathbf{G}\mathbf{G}^T - \frac{\epsilon}{2n}\mathbf{L}\mathbf{L}}] = e^{-\frac{1}{2}\log\det(\boldsymbol{I}+\boldsymbol{\Sigma}_{(\mathbf{G},\mathbf{L})}\mathbf{K}_{(\mathbf{G},\mathbf{L})}) - \frac{1}{2}\log\det\boldsymbol{\Sigma}_{(\mathbf{G},\mathbf{L})}^{-1}}$. With

$$\boldsymbol{I} + \boldsymbol{\Sigma}_{(\mathbf{G},\mathbf{L})}\mathbf{K}_{(\mathbf{G},\mathbf{L})} = \begin{pmatrix} \boldsymbol{\Psi}_m^{-1} & \frac{\epsilon p\sigma_H^2}{n}\boldsymbol{v} \\ \frac{p\sigma_H^2}{nm}\boldsymbol{v}^T & 1 + \frac{\epsilon p\sigma_H^2}{n} \end{pmatrix}, \tag{143}$$

Define $b(q,r) = \frac{\psi_p \sigma_H^2}{\psi_n}\Big(1 - \frac{\psi_p \sigma_H^2}{\psi_n L_1(q,r)}\Big)$. We apply the Schur determinant identity $\det\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \det(A)\cdot\det(D - CA^{-1}B)$ with $A = \mathbf{\Psi}_m^{-1}$, $B = \frac{\epsilon p \sigma_H^2}{n}\boldsymbol{v}$, $C = \frac{p\sigma_H^2}{nm}\boldsymbol{v}^T$, $D = 1 + \frac{\epsilon p \sigma_H^2}{n}$. Using $\boldsymbol{v}^T \mathbf{\Psi}_m \boldsymbol{v} = m/L_1(q,r)$, the Schur complement evaluates to

$$D - CA^{-1}B = 1 + \frac{\epsilon p \sigma_H^2}{n} - \frac{\epsilon p^2 \sigma_H^4}{n^2 m}\cdot\frac{m}{L_1(q,r)} = 1 + \frac{\epsilon p \sigma_H^2}{n}\left(1 - \frac{p\sigma_H^2}{nL_1(q,r)}\right) = 1 + \epsilon\, b(q,r), \quad (144)$$

so $\det(\boldsymbol{I} + \mathbf{\Sigma}_{(\mathbf{G},\mathbf{L})}\mathbf{K}_{(\mathbf{G},\mathbf{L})}) = \det(\mathbf{\Psi}_m^{-1})\cdot(1+\epsilon b) = (1+\epsilon b)/\det \mathbf{\Psi}_m$, giving

$$-\log\det(\boldsymbol{I} + \mathbf{\Sigma}_{(\mathbf{G},\mathbf{L})}\mathbf{K}_{(\mathbf{G},\mathbf{L})}) = \log\det \mathbf{\Psi}_m - \log(1+\epsilon b). \quad (145)$$

Hence the new action reads, with S_0 denoting the action at $\epsilon = 0$,

$$S_\epsilon(q,r) = S_0(q,r) + \psi_n \log[1 + \epsilon\, b(q,r)]. \quad (146)$$

Since the ζ-deformation only adds the term $-\psi_p \zeta r$ to the action, which is ϵ-independent and does not touch the prefactor, it carries through unchanged and the new saddle point equations read

$$\frac{1}{r} + \psi_p \zeta = \frac{1}{q} + \frac{\psi_p \mu_1^2 (e^{-2t} + \frac{\Delta_t}{m})}{L_1(q,r)} + \frac{(m-1)\psi_p \mu_1^2 \frac{\Delta_t}{m}}{L_2(q,r)} + \frac{\epsilon \psi_p [\mu_1^2 e^{-2t}\gamma + \sigma_H^2 \partial_r \gamma]}{1+\epsilon b}, \quad (147)$$

$$z\psi_p = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p \mu_*^2 \frac{1+(m-1)\kappa}{m}}{L_1(q,r)} + \frac{(m-1)\psi_p \mu_*^2 \frac{1-\kappa}{m}}{L_2(q,r)} + \frac{\epsilon \psi_p [\mu_*^2 \kappa \gamma + \sigma_H^2 \partial_q \gamma]}{1+\epsilon b}, \quad (148)$$

with $\gamma(q,r) = 1 - \frac{\psi_p \sigma_H^2}{\psi_n L_1(q,r)}$ and

$$\partial_r \gamma = -\frac{\psi_p \mu_1^2 e^{-2t}}{\psi_n L_1(q,r)} + \frac{\psi_p^2 \mu_1^2 (e^{-2t} m + \Delta_t)\sigma_H^2}{\psi_n^2 m L_1(q,r)^2}, \quad (149)$$

$$\partial_q \gamma = -\frac{\psi_p \mu_*^2 \kappa}{\psi_n L_1(q,r)} + \frac{\psi_p^2 \mu_*^2 (1 + (m-1)\kappa)\sigma_H^2}{\psi_n^2 m L_1(q,r)^2}. \quad (150)$$

Computation of the prefactor. The tilted two-point functions are modified by the GG block of $(\boldsymbol{I} + \mathbf{\Sigma}\mathbf{K})^{-1}$. Applying the block matrix inversion formula $(M^{-1})_{11} = (A - BD^{-1}C)^{-1}$ with $A = \mathbf{\Psi}_m^{-1}$, $B = \frac{\epsilon p \sigma_H^2}{n}\boldsymbol{v}$, $C = \frac{p\sigma_H^2}{nm}\boldsymbol{v}^T$, $D = 1 + \frac{\epsilon p \sigma_H^2}{n}$, and using $\mathbf{\Psi}_m \boldsymbol{v} = \boldsymbol{v}/L_1(q,r)$, the GG block of $(\boldsymbol{I} + \mathbf{\Sigma}_{(\mathbf{G},\mathbf{L})}\mathbf{K}_{(\mathbf{G},\mathbf{L})})^{-1}$ is

$$\mathbf{\Psi}_{GG}^{(\epsilon)} = \big(A - BD^{-1}C\big)^{-1} = \mathbf{\Psi}_m + \frac{\epsilon\chi}{1+\epsilon b}\,\boldsymbol{v}\boldsymbol{v}^T, \qquad \chi = \frac{\psi_p^2 \sigma_H^4}{\psi_n^2 m L_1^2}, \quad (151)$$

so that $\operatorname{Tr} \mathbf{\Psi}_{GG}^{(\epsilon)} = \frac{1}{L_1} + m\,\alpha(\epsilon) + \frac{m-1}{L_2}$ with $\alpha(\epsilon) = \frac{\epsilon\chi}{1+\epsilon b}$. Replacing $\mathbf{\Psi}_m$ by $\mathbf{\Psi}_{GG}^{(\epsilon)}$ in the formula for $T_1(z,\zeta,0)$ gives, evaluating at the new saddle point (q_ϵ, r_ϵ) by the same argument as in the computation of $T_1(z,\zeta,0)$, the function f_1 of Theorem 2.1 in full,

$$\begin{aligned} T_1(z,\zeta,\epsilon) &= f_1(q,r,\epsilon), \qquad f_1(q,r,\epsilon) = 1 - \frac{\mathcal{T}(\epsilon)}{m} + \frac{\mu_1^2 \psi_p \Delta_t\, r}{m^2}\,\mathcal{T}(\epsilon)^2, \\ \mathcal{T}(\epsilon) &= \operatorname{Tr} \mathbf{\Psi}_{GG}^{(\epsilon)} = \frac{1}{L_1} + \frac{m-1}{L_2} + \frac{m\,\epsilon\,\chi}{1+\epsilon b}, \end{aligned} \quad (152)$$

with $(q,r) = (q(z,\zeta,\epsilon), r(z,\zeta,\epsilon))$ the solution of Eq.147–Eq.148. Setting $\epsilon = 0$ recovers the expression of the previous subsection. Hence

$$T_3 = -\partial_\epsilon T_1\big|_{(-\lambda,0,0)} = -\frac{\mathrm{d} f_1(q_\epsilon, r_\epsilon, \epsilon)}{\mathrm{d}\epsilon}\Big|_{\epsilon=0}, \quad (153)$$

the ϵ-dependence entering both explicitly (through α) and implicitly (through q_ϵ, r_ϵ).

C.9 Bias and variance of the score estimator

We compute here the bias and variance of the decomposition Eq.4 of the main text as a function of T_2, T_4 and T_5 defined above. With the shorthands $T_4 = \partial_\zeta T_1|_{(-\lambda,0,0)}$ and $T_5 = \partial_z T_1|_{(-\lambda,0,0)}$ of Eq.76, Eq.158 below is exactly Eq. 12 of the main text. We emphasize that the proof relies on an isotropy argument and thus does not apply for anisotropic data.

The exact score is linear. Since $\mathbf{x}^\nu \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $e^{-2t} + \Delta_t = 1$, the noised marginal is exactly $P_t = \mathcal{N}(0, \boldsymbol{I}_d)$, so

$$\boldsymbol{s}_{\text{exact}}(\boldsymbol{y}) = \nabla_{\boldsymbol{y}} \log P_t(\boldsymbol{y}) = -\boldsymbol{y}, \tag{154}$$

hence

$$\mathcal{L}_{\text{test}}(\boldsymbol{s}) = e^{-2t} + \frac{\Delta_t}{d}\, \mathbb{E}_{\boldsymbol{y}} \|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}_{\text{exact}}(\boldsymbol{y})\|^2. \tag{155}$$

The mean estimator is linear. The joint law of $(\mathbf{W}, \{\mathbf{x}^\nu\}, \{\boldsymbol{\xi}^{\nu\mu}\})$ is invariant under the simultaneous rotation $\mathbf{x} \mapsto \mathbf{R}\mathbf{x}$, $\boldsymbol{\xi} \mapsto \mathbf{R}\boldsymbol{\xi}$, $\mathbf{W} \mapsto \mathbf{W}\mathbf{R}^T$ for any $\mathbf{R} \in O(d)$, and the ridge ERM is equivariant, $\boldsymbol{s}_{\mathbf{R}\mathcal{D}, \mathbf{W}\mathbf{R}^T}(\mathbf{R}\boldsymbol{y}) = \mathbf{R}\, \boldsymbol{s}_{\mathcal{D},\mathbf{W}}(\boldsymbol{y})$, since loss and penalty are invariant under $(\mathbf{A}, \mathcal{D}, \mathbf{W}) \mapsto (\mathbf{R}\mathbf{A}, \mathbf{R}\mathcal{D}, \mathbf{W}\mathbf{R}^T)$ and the minimizer is unique for $\lambda > 0$. Averaging, $\langle \boldsymbol{s}_{\mathcal{D}}(\mathbf{R}\boldsymbol{y})\rangle = \mathbf{R}\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle$ for all $\mathbf{R}$. Writing $\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle = a\, \hat{\boldsymbol{y}} + \boldsymbol{v}$ with $\boldsymbol{v} \perp \boldsymbol{y}$ and applying the element of the stabilizer of $\boldsymbol{y}$ acting as $-\mathrm{id}$ on $\boldsymbol{y}^\perp$ gives $\boldsymbol{v} = -\boldsymbol{v} = 0$; the amplitude then depends on $\boldsymbol{y}$ only through $\|\boldsymbol{y}\|$. Hence $\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle = c(\|\boldsymbol{y}\|)\, \boldsymbol{y}$ at any finite d, and since $\|\boldsymbol{y}\|^2/d \to 1$ concentrates in the proportional limit,

$$\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle = c\, \boldsymbol{y}, \tag{156}$$

with c a deterministic constant. Projecting Eq.156 on $\boldsymbol{y}$ and using $\mathbb{E}\|\boldsymbol{y}\|^2 = d$ gives $c = \frac{1}{d}\mathbb{E}_{\boldsymbol{y}}[\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle \cdot \boldsymbol{y}]$. Because $\mathbb{E}[\boldsymbol{\xi} \mid \boldsymbol{y}] = \sqrt{\Delta_t}\boldsymbol{y}$, the feature–noise correlation of Appendix C.5 obeys, identically at finite d, $\hat{\mathbf{V}} = \frac{1}{\sqrt{\Delta_t}}\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}[\sigma(\cdot)\boldsymbol{\xi}^T] = \mathbb{E}_{\boldsymbol{y}}[\sigma(\mathbf{W}\boldsymbol{y}/\sqrt{d})\boldsymbol{y}^T]$, so computing c needs no approximation beyond the substitution $\hat{\mathbf{V}} \to \mu_1 \mathbf{W}/\sqrt{d}$ already used for $\mathcal{L}_{\text{test}}$. With the explicit minimizer $\mathbf{A}^{\text{ERM}}/\sqrt{p} = -(\mathbf{V}_n^m)^T(\mathbf{U}_n^m + \lambda \boldsymbol{I}_p)^{-1}$,

$$c = \frac{\mu_1}{d}\left\langle \mathrm{Tr}\Big(\frac{\mathbf{A}^{\text{ERM}}}{\sqrt{p}}\frac{\mathbf{W}}{\sqrt{d}}\Big)\right\rangle = -\frac{\mu_1}{\sqrt{\Delta_t}}\frac{1}{dnm}\left\langle \mathrm{Tr}\Big(\boldsymbol{\xi}\mathbf{F}^T\mathcal{G}_{-\lambda,0,0}\frac{\mathbf{W}}{\sqrt{d}}\Big)\right\rangle = -\frac{\mu_1}{\sqrt{\Delta_t}}\, T_2. \tag{157}$$

Closed-form bias and variance. Since $\boldsymbol{s}_{\text{exact}}(\boldsymbol{y}) = -\boldsymbol{y}$ and $\langle \boldsymbol{s}_{\mathcal{D}}(\boldsymbol{y})\rangle = c\boldsymbol{y}$, the bias of Eq.4 is immediate. For the variance, expand $\mathcal{V} = \frac{1}{d}\mathbb{E}_{\boldsymbol{y}}\langle\|\boldsymbol{s}\|^2\rangle - \frac{1}{d}\mathbb{E}_{\boldsymbol{y}}\|\langle \boldsymbol{s}_{\mathcal{D}}\rangle\|^2$ and note that the second moment is the object already producing the quadratic part of $\mathcal{L}_{\text{test}}$: with the same substitution $\tilde{\mathbf{U}} = \mu_1^2\mathbf{W}\mathbf{W}^T/d + \mu_*^2\boldsymbol{I}_p$, one has $\frac{\Delta_t}{d}\mathbb{E}_{\boldsymbol{y}}\langle\|\boldsymbol{s}(\boldsymbol{y})\|^2\rangle = \langle \mathrm{Tr}(\frac{\mathbf{A}}{\sqrt{p}}\tilde{\mathbf{U}}\frac{\mathbf{A}^T}{\sqrt{p}})\rangle\frac{\Delta_t}{d} = \mu_1^2 T_4 + \mu_*^2 T_5$. Hence

$$\mathcal{B}^2 = (1+c)^2, \qquad \mathcal{V} = \frac{\mu_1^2 T_4 + \mu_*^2 T_5}{\Delta_t} - c^2, \qquad c = -\frac{\mu_1 T_2}{\sqrt{\Delta_t}}. \tag{158}$$

Using $\Delta_t \cdot 2c = -2\mu_1\sqrt{\Delta_t}T_2$ and $e^{-2t} + \Delta_t = 1$,

$$e^{-2t} + \Delta_t\big(\mathcal{B}^2 + \mathcal{V}\big) = e^{-2t} + \Delta_t(1+2c) + \mu_1^2 T_4 + \mu_*^2 T_5 = 1 - 2\mu_1\sqrt{\Delta_t}T_2 + \mu_1^2 T_4 + \mu_*^2 T_5 = \mathcal{L}_{\text{test}}. \tag{159}$$

Substituting the asymptotic value of T_2 obtained above yields the explicit slope

$$c = -\frac{\psi_p}{m}\,\mu_1^2\, r\left(\frac{1}{L_1(q,r)} + \frac{m-1}{L_2(q,r)}\right), \tag{160}$$

at the saddle (q, r) of Proposition C.2 evaluated at $(z, \zeta) = (-\lambda, 0)$.

C.10 The overparameterized limit $\psi_p \to \infty$

We consider two scalings of the ridge strength as $\psi_p \to \infty$: regime (i), fixed $\lambda = O(1)$ (same scaling as George et al. (2026)) and regime (ii), $\lambda = \tilde{\lambda}\,\psi_p$ with $\tilde{\lambda} > 0$ fixed (same scaling as Mei & Montanari (2019)).

C.10.1 Regime (i): fixed regularization $\lambda = O(1)$

Proposition C.3 (Overparameterized limit, fixed regularization)**.** *Define*

$$\ell_1^\star(r) = \mu_1^2(e^{-2t}m + \Delta_t)r + \frac{\mu_*^2(1 + (m-1)\kappa)}{\lambda}, \qquad \ell_2^\star(r) = \mu_1^2\Delta_t r + \frac{\mu_*^2(1-\kappa)}{\lambda},$$

and

$$g(r) = \frac{1}{\ell_1^\star(r)} + \frac{m-1}{\ell_2^\star(r)}, \quad R(r) = \mu_1^2\psi_n\left[\frac{e^{-2t}m + \Delta_t}{\ell_1^\star(r)} + \frac{(m-1)\Delta_t}{\ell_2^\star(r)}\right],$$
$$\Phi(r) = \psi_n g(r) - \mu_1^2\Delta_t\psi_n^2 r g(r)^2. \tag{161}$$

For the trace T_3, let $r(\tilde\epsilon)$ solve the deformed fixed point $\frac{1}{r} = \lambda + R(r) + \frac{\tilde\epsilon K(r)}{1+\tilde\epsilon\beta(r)}$ with $r(0) = r_\infty$, where

$$\sigma_\star^2(r) = \mu_1^2 e^{-2t} r + \frac{\mu_*^2\kappa}{\lambda}, \;\; \gamma_\infty = 1 - \frac{m\sigma_\star^2}{\ell_1^\star}, \;\; K = \mu_1^2 e^{-2t}\gamma_\infty + \sigma_\star^2\gamma_\infty', \;\; \beta = \frac{\sigma_\star^2\gamma_\infty}{\psi_n}, \;\; \chi_\infty = \frac{m\sigma_\star^4}{(\ell_1^\star)^2},$$

and set $G(\tilde\epsilon) = g(r) + \frac{1}{\psi_n}\frac{\tilde\epsilon\chi_\infty}{1+\tilde\epsilon\beta}$ and $\hat\Phi(\tilde\epsilon) = \psi_n G - \mu_1^2\Delta_t\psi_n^2 r G^2$. Let r_∞ be the unique root in $(0, 1/\lambda)$ of $\frac{1}{r} = \lambda + R(r)$, and set $D = \frac{1}{r_\infty^2} + R'(r_\infty)$. Then, as $\psi_p \to \infty$ at fixed $(\lambda > 0, \psi_n, m, t)$, the solutions of Proposition C.2 converge to $q_\infty = 1/\lambda$, $r \to r_\infty$, the traces converge to

$$T_1 \to 1, \quad T_5 \to 0, \quad T_2 \to \mu_1\sqrt{\Delta_t}\,\psi_n r_\infty g(r_\infty), \quad T_4 \to -\frac{\Phi'(r_\infty)}{D}, \quad T_3 \to T_3^\infty = \hat\Phi'(0),$$

and the losses converge to

$$\mathcal{L}_{\text{train}}^{n,m}(\psi_p = \infty) = 0, \qquad \mathcal{L}_{\text{test}}(\psi_p = \infty) = 1 - 2\mu_1^2\Delta_t\psi_n r_\infty g(r_\infty) - \mu_1^2\frac{\Phi'(r_\infty)}{D},$$
$$\mathcal{L}_{\text{test}}^{\text{emp}}(\psi_p = \infty) = 1 - 2\mu_1^2\Delta_t\psi_n r_\infty g(r_\infty) + \hat\Phi'(0) - \Delta_t\mu_1^2\frac{\Phi'(r_\infty)}{D},$$

where $\hat\Phi'(0)$ is the derivative of $\hat\Phi$ at $\tilde\epsilon = 0$.

Proof. Recall the finite-ψ_p equations of Proposition C.2 at $z = -\lambda$, $\zeta = 0$,

$$\frac{1}{r} = \frac{1}{q} + \frac{\mu_1^2\psi_p(e^{-2t} + \Delta_t/m)}{L_1} + \frac{(m-1)\mu_1^2\psi_p\Delta_t/m}{L_2}, \tag{162}$$
$$-\lambda\psi_p = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p\mu_*^2(1+(m-1)\kappa)/m}{L_1} + \frac{(m-1)\psi_p\mu_*^2(1-\kappa)/m}{L_2}, \tag{163}$$

with $L_i = 1 + \frac{\psi_p}{\psi_n m}\ell_i$, $\ell_1 = \mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1+(m-1)\kappa)$ and $\ell_2 = \mu_1^2 r\Delta_t + \mu_*^2 q(1-\kappa)$. Here $\ell_1, \ell_2 = O(1)$, so

$$\frac{1}{L_i} = \frac{\psi_n m}{\psi_p \ell_i}\big(1 + O(\psi_p^{-1})\big), \qquad \frac{\psi_p}{L_i} \to \frac{\psi_n m}{\ell_i}. \tag{164}$$

Dividing the q-equation by ψ_p, the $1/L_i$ terms drop and $z = -1/q + O(\psi_p^{-1})$, i.e. $q_\infty = 1/\lambda$. Inserting $1/q \to \lambda$ and Eq.164 into the r-equation at $\zeta = 0$ gives the scalar fixed point

$$\frac{1}{r_\infty} = \lambda + R(r_\infty), \qquad R(r) = \mu_1^2\psi_n\left[\frac{e^{-2t}m + \Delta_t}{\ell_1^\star(r)} + \frac{(m-1)\Delta_t}{\ell_2^\star(r)}\right], \tag{165}$$

with $\ell_i^\star(r) = \ell_i(1/\lambda, r)$, i.e. $\ell_1^\star(r) = \mu_1^2(e^{-2t}m + \Delta_t)r + \frac{\mu_*^2(1+(m-1)\kappa)}{\lambda}$ and $\ell_2^\star(r) = \mu_1^2\Delta_t r + \frac{\mu_*^2(1-\kappa)}{\lambda}$. The root is unique in $(0, 1/\lambda)$. Writing $g(r) = \frac{1}{\ell_1^\star} + \frac{m-1}{\ell_2^\star}$, Eq.164 gives $\mathcal{T} = \frac{\psi_n m}{\psi_p}g(r) + O(\psi_p^{-2})$, so

$$T_1 = 1 - \frac{1}{\psi_p}\Phi(r) + O(\psi_p^{-2}), \qquad \Phi(r) = \psi_n g(r) - \mu_1^2\Delta_t\psi_n^2 r g(r)^2, \tag{166}$$

hence $T_1 \to 1$ (exact interpolation, $\mathcal{L}_{\text{train}}^{n,m}(\psi_p = \infty) = 0$) and $T_2 \to \mu_1\sqrt{\Delta_t}\psi_n r_\infty g(r_\infty)$.

For the derivatives, the term $\psi_p z$ in the q-equation is balanced by $-\psi_p/q$, so a perturbation $\delta z = O(1)$ is absorbed by $\delta q = O(1)$ and $\partial_z q, \partial_z r = O(1)$, giving $T_5 = \partial_z T_1 = -\frac{1}{\psi_p}(\partial_q\Phi\,\partial_z q + \partial_r\Phi\,\partial_z r) = O(\psi_p^{-1}) \to 0$. In the r-equation, $\psi_p\zeta$ competes with $O(1)$ terms, so the natural

coordinate is $\tilde{\zeta} = \psi_p \zeta$ and the reduced equation is $\frac{1}{r} = -\tilde{\zeta} + \lambda + R(r)$, giving $\frac{dr}{d\tilde{\zeta}} = 1/D$ with $D = \frac{1}{r_\infty^2} + R'(r_\infty)$. Then $\partial_\zeta = \psi_p \partial_{\tilde{\zeta}}$ yields

$$T_4 = -\Phi'(r)\frac{dr}{d\tilde{\zeta}} \;\longrightarrow\; T_4^\infty = -\frac{\Phi'(r_\infty)}{D}, \tag{167}$$

with $g'(r) = -\mu_1^2[\frac{e^{-2t}m+\Delta_t}{(\ell_1^\star)^2} + \frac{(m-1)\Delta_t}{(\ell_2^\star)^2}]$, $R'(r) = -\mu_1^4\psi_n[\frac{(e^{-2t}m+\Delta_t)^2}{(\ell_1^\star)^2} + \frac{(m-1)\Delta_t^2}{(\ell_2^\star)^2}]$ and $\Phi'(r) = \psi_n g'(r) - \mu_1^2\Delta_t\psi_n^2(g^2 + 2rgg')$. Collecting the limits,

$$\mathcal{L}_{\text{test}}(\psi_p = \infty) = 1 - 2\mu_1^2\Delta_t\psi_n r_\infty g(r_\infty) - \mu_1^2\frac{\Phi'(r_\infty)}{D}. \tag{168}$$

The trace T_3 comes from the ϵ-deformed prefactor of the T_3 subsection. In the present limit the deformation parameter is large, $b = O(\psi_p)$, so we rescale $\tilde{\epsilon} = \psi_p\epsilon$; the deformed saddle then reads $\frac{1}{r} = \lambda + R(r) + \frac{\tilde{\epsilon}K(r)}{1+\tilde{\epsilon}\beta(r)}$ with

$$\sigma_\star^2(r) = \mu_1^2 e^{-2t} r + \tfrac{\mu_*^2\kappa}{\lambda}, \quad \gamma_\infty = 1 - \tfrac{m\sigma_\star^2}{\ell_1^\star}, \quad K = \mu_1^2 e^{-2t}\gamma_\infty + \sigma_\star^2\gamma_\infty',$$
$$\beta = \tfrac{\sigma_\star^2\gamma_\infty}{\psi_n}, \quad \chi_\infty = \tfrac{m\,\sigma_\star^4}{(\ell_1^\star)^2}, \tag{169}$$

and $\gamma_\infty' = \partial_r\gamma_\infty$. Setting $G(\tilde{\epsilon}) = g(r) + \frac{1}{\psi_n}\frac{\tilde{\epsilon}\chi_\infty}{1+\tilde{\epsilon}\beta}$ and $\hat{\Phi}(\tilde{\epsilon}) = \psi_n G - \mu_1^2\Delta_t\psi_n^2 r G^2$, the deformed prefactor is $T_1(\tilde{\epsilon}) = 1 - \frac{1}{\psi_p}\hat{\Phi}(\tilde{\epsilon}) + O(\psi_p^{-2})$, so $T_3^\infty = \hat{\Phi}'(0)$ with $r'(0) = -K(r_\infty)/D$ and $G'(0) = g'(r_\infty)r'(0) + \chi_\infty(r_\infty)/\psi_n$, i.e.

$$T_3^\infty = \psi_n G'(0) - \mu_1^2\Delta_t\psi_n^2\big[r'(0)\,g(r_\infty)^2 + 2r_\infty g(r_\infty)\,G'(0)\big]. \tag{170}$$

Inserting T_2, T_3^∞ and T_4, and using $T_5 \to 0$, into $\mathcal{L}_{\text{test}}^{\text{emp}} = 1 - 2\mu_1\sqrt{\Delta_t}T_2 + T_3 + \Delta_t\mu_1^2 T_4 + \mu_*^2(1-\kappa)T_5$ gives the stated $\mathcal{L}_{\text{test}}^{\text{emp}}(\psi_p = \infty)$. □

C.10.2 Regime (II): strong regularization $\lambda = \tilde{\lambda}\,\psi_p$

Proposition C.4 (Overparameterized limit, strong regularization)**.** *Let $\lambda = \tilde{\lambda}\,\psi_p$ with $\tilde{\lambda} > 0$ fixed, and define the scaled order parameters $\bar{q} = \psi_p q$, $\bar{r} = \psi_p r$, the $O(1)$ resolvent factors*

$$\bar{L}_1(\bar{q},\bar{r}) = 1 + \tfrac{1}{\psi_n m}\big[\mu_1^2(e^{-2t}m + \Delta_t)\bar{r} + \mu_*^2(1 + (m-1)\kappa)\bar{q}\big], \tag{171}$$
$$\bar{L}_2(\bar{q},\bar{r}) = 1 + \tfrac{1}{\psi_n m}\big[\mu_1^2\Delta_t\bar{r} + \mu_*^2(1-\kappa)\bar{q}\big], \tag{172}$$

and, writing $\bar{L}_i^\star(\bar{r}) = \bar{L}_i(1/\tilde{\lambda},\bar{r})$,

$$\bar{R}(\bar{r}) = \tfrac{1}{m}\Big(\tfrac{\mu_1^2(e^{-2t}m+\Delta_t)}{\bar{L}_1^\star(\bar{r})} + \tfrac{(m-1)\mu_1^2\Delta_t}{\bar{L}_2^\star(\bar{r})}\Big), \quad \bar{g}(\bar{r}) = \tfrac{1}{\bar{L}_1^\star(\bar{r})} + \tfrac{m-1}{\bar{L}_2^\star(\bar{r})}, \quad \bar{\Phi}(\bar{r}) = \tfrac{\bar{g}(\bar{r})}{m} - \tfrac{\mu_1^2\Delta_t\bar{r}}{m^2}\bar{g}(\bar{r})^2.$$

For the trace T_3, let $\bar{r}(\epsilon)$ solve the deformed fixed point $\frac{1}{\bar{r}} = \tilde{\lambda} + \bar{R}(\bar{r}) + \frac{\epsilon\bar{K}(\bar{r})}{1+\epsilon\bar{b}(\bar{r})}$ with $\bar{r}(0) = \bar{r}_\infty$, where

$$\bar{\sigma}_\star^2(\bar{r}) = \mu_1^2 e^{-2t}\bar{r} + \tfrac{\mu_*^2\kappa}{\tilde{\lambda}}, \;\; \bar{\gamma} = 1 - \tfrac{\bar{\sigma}_\star^2}{\psi_n\bar{L}_1^\star}, \;\; \bar{K} = \mu_1^2 e^{-2t}\bar{\gamma} + \bar{\sigma}_\star^2\bar{\gamma}', \tag{173}$$
$$\bar{b} = \tfrac{\bar{\sigma}_\star^2}{\psi_n}\bar{\gamma}, \;\; \bar{\chi} = \tfrac{\bar{\sigma}_\star^4}{\psi_n^2 m(\bar{L}_1^\star)^2}, \tag{174}$$

and set $\bar{G}(\epsilon) = \bar{g}(\bar{r}(\epsilon)) + m\frac{\epsilon\bar{\chi}}{1+\epsilon\bar{b}}$, $\bar{\Phi}_\epsilon = \frac{\bar{G}(\epsilon)}{m} - \frac{\mu_1^2\Delta_t\bar{r}(\epsilon)}{m^2}\bar{G}(\epsilon)^2$ (so $\bar{\Phi}_0 = \bar{\Phi}(\bar{r}_\infty)$). Let $\bar{r}_\infty$ be the root of $\frac{1}{\bar{r}} = \tilde{\lambda} + \bar{R}(\bar{r})$ and $\bar{D} = \frac{1}{\bar{r}_\infty^2} + \bar{R}'(\bar{r}_\infty)$. Then, as $\psi_p \to \infty$, $\bar{q} \to 1/\tilde{\lambda}$, $\bar{r} \to \bar{r}_\infty$,

$$T_1 \to 1 - \bar{\Phi}(\bar{r}_\infty), \quad T_5 \to 0, \quad T_2 \to \tfrac{1}{m}\mu_1\sqrt{\Delta_t}\,\bar{r}_\infty\bar{g}(\bar{r}_\infty), \tag{175}$$
$$T_4 \to -\tfrac{\bar{\Phi}'(\bar{r}_\infty)}{\bar{D}}, \quad T_3 \to \bar{T}_3^\infty = \tfrac{\mathrm{d}\bar{\Phi}_\epsilon}{\mathrm{d}\epsilon}\big|_0, \tag{176}$$

and the losses converge to

$$\mathcal{L}_{\text{train}}^{n,m}(\psi_p = \infty) = \bar{\Phi}(\bar{r}_\infty), \qquad \mathcal{L}_{\text{test}}(\psi_p = \infty) = 1 - \tfrac{2}{m}\mu_1^2\Delta_t\,\bar{r}_\infty\bar{g}(\bar{r}_\infty) - \mu_1^2\tfrac{\bar{\Phi}'(\bar{r}_\infty)}{\bar{D}},$$
$$\mathcal{L}_{\text{test}}^{\text{emp}}(\psi_p = \infty) = 1 - \tfrac{2}{m}\mu_1^2\Delta_t\,\bar{r}_\infty\bar{g}(\bar{r}_\infty) + \bar{T}_3^\infty - \Delta_t\mu_1^2\tfrac{\bar{\Phi}'(\bar{r}_\infty)}{\bar{D}},$$

where $\bar{T}_3^\infty = \frac{\mathrm{d}\bar{\Phi}_\epsilon}{\mathrm{d}\epsilon}\big|_0$ is the derivative at $\epsilon = 0$.

Proof. *Recall the finite-ψ_p equations* of Proposition C.2 at $\zeta = 0$,

$$\frac{1}{r} = \frac{1}{q} + \frac{\mu_1^2\psi_p(e^{-2t} + \Delta_t/m)}{L_1} + \frac{(m-1)\mu_1^2\psi_p\Delta_t/m}{L_2}, \tag{177}$$

$$z\psi_p = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p\mu_*^2(1+(m-1)\kappa)/m}{L_1} + \frac{(m-1)\psi_p\mu_*^2(1-\kappa)/m}{L_2}, \tag{178}$$

with $L_i = 1 + \frac{\psi_p}{\psi_n m}\ell_i$, $\ell_1 = \mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1+(m-1)\kappa)$ and $\ell_2 = \mu_1^2 r\Delta_t + \mu_*^2 q(1-\kappa)$. *Take $\psi_p \to \infty$ with $z = -\lambda = -\tilde{\lambda}\psi_p$.* With $\bar{q} = \psi_p q$, $\bar{r} = \psi_p r$, the resolvent factors become $L_i = \bar{L}_i(\bar{q}, \bar{r}) + O(\psi_p^{-1})$; crucially they remain $O(1)$, so—unlike regime (i)—the $+1$ cannot be dropped. Dividing the q-equation by ψ_p^2, all terms vanish except the leading one, leaving $-\tilde{\lambda} = -1/\bar{q}$, i.e. $\bar{q}_\infty = 1/\tilde{\lambda}$. Dividing the r-equation by ψ_p gives $\frac{1}{\bar{r}} + \zeta = \frac{1}{\bar{q}} + \frac{\mu_1^2(e^{-2t}+\Delta_t/m)}{\bar{L}_1} + \frac{(m-1)\mu_1^2\Delta_t/m}{\bar{L}_2}$; at $\zeta = 0$, $\bar{q} = 1/\tilde{\lambda}$ this is the scalar fixed point $\frac{1}{\bar{r}} = \tilde{\lambda} + \bar{R}(\bar{r})$. Substituting $r = \bar{r}/\psi_p$ into $T_1 = 1 - \frac{\mathcal{T}}{m} + \frac{\mu_1^2\psi_p\Delta_t r}{m^2}\mathcal{T}^2$, with $\mathcal{T} \to \bar{g}(\bar{r}_\infty)$ and $\psi_p r \to \bar{r}_\infty$, gives $T_1 \to 1 - \bar{\Phi}(\bar{r}_\infty)$, bounded away from 1: the training error stays positive. Likewise $T_2 \to \frac{1}{m}\mu_1\sqrt{\Delta_t}\bar{r}_\infty\bar{g}(\bar{r}_\infty)$. For $T_4 = \partial_\zeta T_1$, implicit differentiation of $\frac{1}{\bar{r}} + \zeta = \tilde{\lambda} + \bar{R}(\bar{r})$ gives $-\frac{1}{\bar{r}^2}\partial_\zeta\bar{r} + 1 = \bar{R}'(\bar{r})\partial_\zeta\bar{r}$, hence $\partial_\zeta\bar{r} = 1/\bar{D}$ and $T_4 = -\bar{\Phi}'(\bar{r}_\infty)/\bar{D}$. For $T_5 = \partial_z T_1$: since $\bar{q} \approx -\psi_p/z$, $\partial_z\bar{q} = \psi_p/z^2 = 1/(\tilde{\lambda}^2\psi_p) \to 0$, and as the z-dependence of $\bar{r}$ is mediated by $\bar{q}$, $T_5 \to 0$. Assembling $\mathcal{L}_{\text{test}} = 1 - 2\mu_1\sqrt{\Delta_t}T_2 + \mu_1^2 T_4 + \mu_*^2 T_5$ yields the stated $\mathcal{L}_{\text{train}}^{n,m}$ and $\mathcal{L}_{\text{test}}$.

Here the field variance of the T_3 deformation is small, $\sigma_H^2 = \mu_1^2 e^{-2t}r + \mu_*^2\kappa q = \bar{\sigma}_\star^2/\psi_p$ with $\bar{\sigma}_\star^2(\bar{r}) = \mu_1^2 e^{-2t}\bar{r} + \frac{\mu_*^2\kappa}{\tilde{\lambda}}$, so the deformation parameters stay $O(1)$ and *no* rescaling of ϵ is needed (contrast regime (i), where $b = O(\psi_p)$):

$$\bar{\gamma} = 1 - \frac{\bar{\sigma}_\star^2}{\psi_n\bar{L}_1^\star}, \qquad \bar{b} = \frac{\bar{\sigma}_\star^2}{\psi_n}\bar{\gamma}, \qquad \bar{\chi} = \frac{\bar{\sigma}_\star^4}{\psi_n^2 m(\bar{L}_1^\star)^2}, \qquad \bar{K} = \mu_1^2 e^{-2t}\bar{\gamma} + \bar{\sigma}_\star^2\bar{\gamma}', \quad \bar{\gamma}' = \partial_{\bar{r}}\bar{\gamma}. \tag{179}$$

The deformed saddle becomes $\frac{1}{\bar{r}} = \tilde{\lambda} + \bar{R}(\bar{r}) + \frac{\epsilon\bar{K}(\bar{r})}{1+\epsilon\bar{b}}$, giving $\bar{r}'(0) = -\bar{K}(\bar{r}_\infty)/\bar{D}$. With $\bar{G}(\epsilon) = \bar{g}(\bar{r}(\epsilon)) + m\frac{\epsilon\bar{\chi}}{1+\epsilon\bar{b}}$ (so $\bar{G}(0) = \bar{g}(\bar{r}_\infty)$, $\bar{G}'(0) = \bar{g}'(\bar{r}_\infty)\bar{r}'(0) + m\bar{\chi}(\bar{r}_\infty)$) the deformed prefactor is $T_1(\epsilon) \to 1 - \bar{\Phi}_\epsilon$, $\bar{\Phi}_\epsilon = \frac{\bar{G}(\epsilon)}{m} - \frac{\mu_1^2\Delta_t\bar{r}(\epsilon)}{m^2}\bar{G}(\epsilon)^2$, and since $T_3 = -\frac{\mathrm{d}T_1(\epsilon)}{\mathrm{d}\epsilon}|_0$,

$$\bar{T}_3^\infty = \frac{\mathrm{d}\bar{\Phi}_\epsilon}{\mathrm{d}\epsilon}\Big|_0 = \frac{\bar{G}'(0)}{m} - \frac{\mu_1^2\Delta_t}{m^2}\left[\bar{r}'(0)\,\bar{g}(\bar{r}_\infty)^2 + 2\bar{r}_\infty\bar{g}(\bar{r}_\infty)\,\bar{G}'(0)\right]. \tag{180}$$

Inserting T_2, $\bar{T}_3^\infty$ and T_4, and using $T_5 \to 0$, into $\mathcal{L}_{\text{test}}^{\text{emp}} = 1 - 2\mu_1\sqrt{\Delta_t}T_2 + T_3 + \Delta_t\mu_1^2 T_4 + \mu_*^2(1-\kappa)T_5$ gives the stated $\mathcal{L}_{\text{test}}^{\text{emp}}(\psi_p = \infty)$. □

C.11 The $m \to \infty$ limit of the equations

Proposition C.5 ($m \to \infty$ limit)**.** *As $m \to \infty$ at fixed $(\psi_p, \psi_n, t, \lambda)$, the order parameters $(q, r) = (q(z,\zeta), r(z,\zeta))$ satisfy*

$$\frac{1}{r} + \psi_p\zeta = \frac{1}{q} + \frac{\mu_1^2\psi_p e^{-2t}}{L_1^\infty} + \psi_p\mu_1^2\Delta_t, \tag{181}$$

$$\psi_p z = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p\mu_*^2\kappa}{L_1^\infty} + \psi_p\mu_*^2(1-\kappa). \tag{182}$$

The training loss and the empirical test loss coincide, $\mathcal{L}_{\text{train}}^{n,m} = \mathcal{L}_{\text{test}}^{\text{emp}}$, with $\mathcal{L}_{\text{test}}^{\text{emp}} = 1 - \mu_1^2\psi_p\Delta_t r$, and

$$\mathcal{L}_{\text{test}}(m=\infty) = 1 - 2\mu_1^2\Delta_t\psi_p r + \mu_1^2\psi_p\Delta_t\big(\mu_1^2\partial_\zeta r + \mu_*^2\partial_z r\big), \tag{183}$$

with $r = r(-\lambda, 0)$ and the derivatives evaluated at $(z,\zeta) = (-\lambda, 0)$.

We recover the results of George et al. (2026); Bonnaire et al. (2025)[13].

[13] In Theorem 3.2 of George et al. (2026), they denote r by $\mathcal{K}$.

Proof. By Proposition C.2 (isotropic case), at finite m the order parameters obey

$$\frac{1}{r} + \psi_p \zeta = \frac{1}{q} + \frac{\mu_1^2 \psi_p \big(e^{-2t} + \frac{\Delta_t}{m}\big)}{L_1} + \frac{(m-1)\mu_1^2 \psi_p \frac{\Delta_t}{m}}{L_2}, \tag{184}$$

$$\psi_p z = \frac{1-\psi_p}{q} - \frac{r}{q^2} + \frac{\psi_p \mu_*^2 \frac{1+(m-1)\kappa}{m}}{L_1} + \frac{(m-1)\psi_p \mu_*^2 \frac{1-\kappa}{m}}{L_2}, \tag{185}$$

with $L_1 = 1 + \frac{\psi_p}{\psi_n m}[\mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1 + (m-1)\kappa)]$ and $L_2 = 1 + \frac{\psi_p}{\psi_n m}[\mu_1^2 r \Delta_t + \mu_*^2 q(1-\kappa)]$. In L_1 the bracket grows linearly in m, since $\frac{1}{m}[\mu_1^2 r(e^{-2t}m + \Delta_t) + \mu_*^2 q(1 + (m-1)\kappa)] \to \mu_1^2 e^{-2t} r + \mu_*^2 \kappa q$; hence $L_1 \to L_1^\infty = 1 + \frac{\psi_p}{\psi_n}(\mu_1^2 e^{-2t} r + \mu_*^2 \kappa q)$, whereas the L_2-bracket stays $O(1)$ so $L_2 \to 1$. Using in addition $e^{-2t} + \frac{\Delta_t}{m} \to e^{-2t}$, $\frac{m-1}{m} \to 1$, $\frac{1+(m-1)\kappa}{m} \to \kappa$ and $\frac{(m-1)(1-\kappa)}{m} \to 1-\kappa$, the four fractions tend to $\frac{\mu_1^2 \psi_p e^{-2t}}{L_1^\infty}$, $\psi_p \mu_1^2 \Delta_t$, $\frac{\psi_p \mu_*^2 \kappa}{L_1^\infty}$ and $\psi_p \mu_*^2 (1-\kappa)$, giving the stated equations. For the losses, $\tau = \mathcal{T}/m \to 1$ with $\tau - 1 = O(1/m)$, so $\partial_z \tau, \partial_\zeta \tau \to 0$ and $f_1(q, r, 0) \to \mu_1^2 \psi_p \Delta_t r$, whence

$$T_1 \to \mu_1^2 \psi_p \Delta_t r, \quad T_2 \to \mu_1 \sqrt{\Delta_t} \psi_p r, \quad T_4 \to \mu_1^2 \psi_p \Delta_t \, \partial_\zeta r, \quad T_5 \to \mu_1^2 \psi_p \Delta_t \, \partial_z r. \tag{186}$$

At $m = \infty$ the empirical noise average in $\mathcal{L}_{\text{train}}^{n,m}$ coincides with the exact one, so $\mathcal{L}_{\text{train}}^{n,m} = \mathcal{L}_{\text{test}}^{\text{emp}}$ by definition and the common value follows from $\mathcal{L}_{\text{train}}^{n,m} = 1 - T_1$. Substituting into the loss decomposition gives the stated expressions, the derivatives $\partial_\zeta r, \partial_z r$ being obtained by implicit differentiation of the two self-consistent equations. □

C.12 Extension to sub-Gaussian data

Throughout the main text we assumed $\mathbf{x}^\nu \sim \mathcal{N}(0, \boldsymbol{I}_d)$. We show here that the results on the losses extend to any zero-mean distribution with covariance $\boldsymbol{\Sigma}$ and sub-Gaussian tails, provided that $\operatorname{Tr}(\boldsymbol{\Sigma})/d \to \sigma_x^2 = O(1)$ as $d \to \infty$, and that the activation satisfies $\mu_0 := \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(\sqrt{e^{-2t}\sigma_x^2 + \Delta_t}\, z)] = 0$. We emphasize that the expressions for the bias-variance decomposition relies on an isotropy argument and thus cannot be extended to the anisotropy case. The argument has two steps.

The preactivation is asymptotically Gaussian. The preactivation of neuron α on sample ν at noise level t is $h_\alpha^{\nu\mu} = e^{-t}\mathbf{W}_\alpha \cdot \mathbf{x}^\nu/\sqrt{d} + \sqrt{\Delta_t}\mathbf{W}_\alpha \cdot \boldsymbol{\xi}^{\nu\mu}/\sqrt{d}$. The noise term is exactly Gaussian. For the data term $g_\alpha^\nu = e^{-t}\mathbf{W}_\alpha \cdot \mathbf{x}^\nu/\sqrt{d}$, conditioning on $\mathbf{W}$, the CLT for linear forms of sub-Gaussian vectors yields that $\boldsymbol{g}^\nu = (g_\alpha^\nu)_\alpha$ is asymptotically Gaussian with covariance $e^{-2t}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^T/d$. This covariance matrix concentrates to a diagonal one: its diagonal entries converge to $e^{-2t}\operatorname{Tr}(\boldsymbol{\Sigma})/d =: e^{-2t}\sigma_x^2$ while its off-diagonal entries vanish, since $\mathbf{W}_\alpha$ and $\mathbf{W}_\beta$ are independent for $\alpha \neq \beta$. Setting $\Gamma_t = e^{-2t}\sigma_x^2 + \Delta_t$, the full preactivation is therefore asymptotically $h_\alpha^{\nu\mu} \sim \mathcal{N}(0, \Gamma_t)$ with independent components across neurons, and the GEP can be applied.

Redefinition of the GEP coefficients. Since the marginal distribution of $h_\alpha^{\nu\mu}$ is $\mathcal{N}(0, \Gamma_t)$, the GEP coefficients must be recomputed accordingly. Writing $h = \sqrt{\Gamma_t}\, z$ with $z \sim \mathcal{N}(0,1)$ and the data preactivation as $e^{-t}\sigma_x u$ with $u \sim \mathcal{N}(0,1)$,

$$\mu_1 = \mathbb{E}\big[\sigma'(\sqrt{\Gamma_t}\, z)\big], \tag{187}$$

$$\mu_*^2 = \mathbb{E}\big[\sigma(\sqrt{\Gamma_t}\, z)^2\big] - \mu_1^2 \Gamma_t, \tag{188}$$

$$\kappa = \frac{1}{\mu_*^2}\, \mathbb{E}_u\Big[\big(\mathbb{E}_v\big[\sigma(e^{-t}\sigma_x u + \sqrt{\Delta_t}\, v)\big] - \mu_1 e^{-t}\sigma_x u\big)^2\Big], \tag{189}$$

where $z, u, v \sim \mathcal{N}(0,1)$ are independent.

For $\sigma_x = 1$ one recovers $\Gamma_t = 1$ and the original definitions of the main text. The GEP itself—the equivalence of resolvent traces between $\mathbf{F} = \sigma(\mathbf{W}\mathbf{Y}/\sqrt{d})$ and its linear Gaussian surrogate—extends to sub-Gaussian data by universality (Péché, 2019; Goldt et al., 2021; Hu & Lu, 2023). The data covariance $\boldsymbol{\Sigma}$ enters the self-consistent equations only through the spectral measure $\rho_{\boldsymbol{\Sigma}}$ in Proposition C.2.

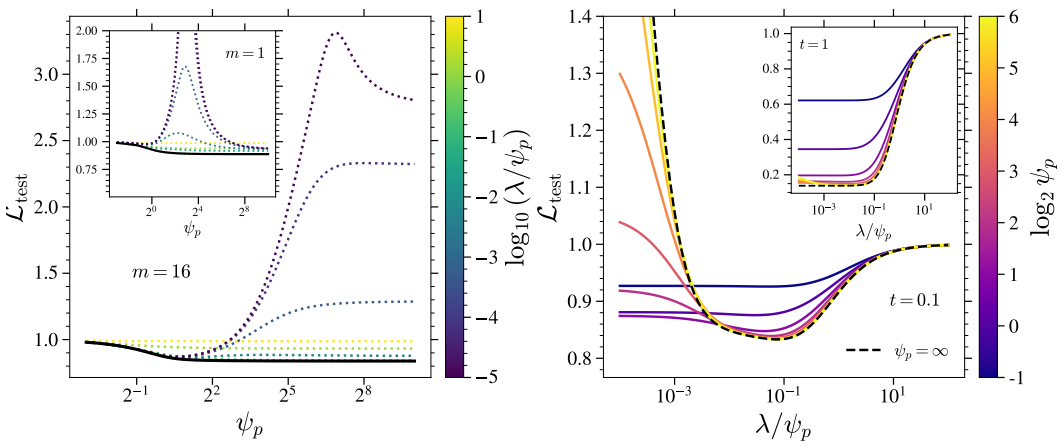

Figure 14: **Benefits of regularization on generalization.** *(Left)* Test loss $\mathcal{L}_{\text{test}}$ of the RFNN as a function of $\psi_p = p/d$ for several values of $\lambda_p = \psi_p\tilde{\lambda}$ with $\psi_n = n/d = 8$, $\sigma = \tanh$, $t = 0.1$ and $m = 16$; inset: same at $m = 1$. The full black line is the lower envelope. *(Right)* Test loss as a function of λ/ψ_p for several values of ψ_p at $m = 16$, $t = 0.1$, together with the $\psi_p = \infty$ limit (black dashed); inset: same at $t = 1.0$.

Conclusion. All results of the main text of the three losses extend to zero-mean sub-Gaussian data with covariance $\mathbf{\Sigma}$ and $\text{Tr}(\mathbf{\Sigma})/d \to \sigma_x^2 = O(1)$, with μ_1, μ_*, κ redefined via $\Gamma_t = e^{-2t}\sigma_x^2 + \Delta_t$, and $\rho_{\mathbf{\Sigma}}$ replacing $\delta(\lambda - 1)$ in the resolvent integrals.

C.13 Additional Results and Figures

In this section we present additional results and figures on the analytical part.

C.13.1 Extended discussion on the effect of regularization.

In the RFNN model, the optimal training strategy depends critically on the diffusion time t. At large t, we recover the benign overfitting of the classical regression setting (Mei & Montanari, 2019) (inset of the right panel of Fig. 14): for every fixed ψ_p, the test loss is a strictly increasing function of λ, so zero regularization is always optimal. The global minimum over (ψ_p, λ) is attained at $\psi_p \to \infty$ and $\lambda = 0$: overparameterization is unconditionally beneficial, and any regularization strictly hurts. On the other hand, at small t, malign overfitting fundamentally alters this picture. Without regularization ($\lambda \to 0^+$), the optimal model size remains $\psi_p \to \infty$ for $m = 1$ (inset of the left panel of Fig. 14), but shifts to the underparameterized regime ($\psi_p^* < \psi_n$) for $m > 1$ (left panel of Fig. 14): overparameterizing an unregularized model is detrimental. However, jointly optimizing over (ψ_p, λ) restores the benefit of large models: the global optimum is always $\psi_p \to \infty$ with $\lambda/\psi_p = O(1)$ (right panel of Fig. 14), so a large well-regularized model always outperforms a small unregularized one.

C.13.2 Additional results on the bias and variance decomposition

In Fig. 15 we show the bias and the variance as functions of the model size ψ_p, comparing the analytical curves to numerical experiments at finite d, for $m = 1$ (*left*) and $m = 4$ (*middle*). At $m = 1$ the bias decreases monotonically with ψ_p, whereas at $m = 4$ it passes through a minimum and increases again, before saturating at a finite value. In Fig. 15 (*right*) we plot the $\psi_p \to \infty$ limit of both quantities as a function of ψ_n: the bias saturates at a value that remains $O_{\psi_n}(1)$ for $m > 1$, while it decays as a power law in ψ_n for $m = 1$; the variance decays as a power law in ψ_n for both $m = 1$ and $m > 1$.

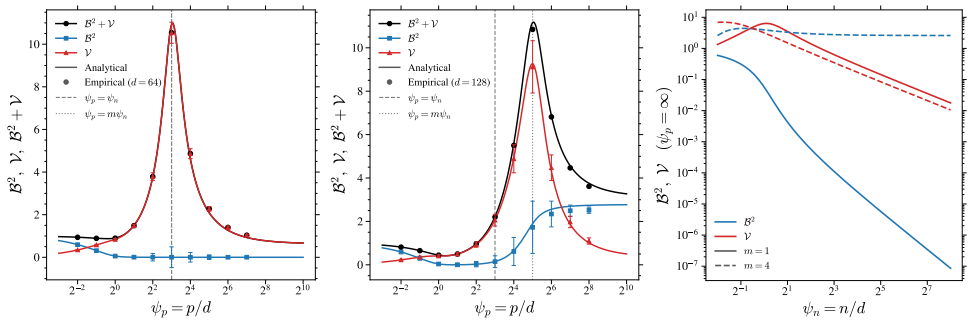

Figure 15: (*Left*) Bias and variance for the RFNN model as a function of $\psi_p = p/d$ for $\psi_n = 8,\ t = 0.1,\ m = 1,\ \lambda = 10^{-3}$ and finite dimension $d = 64$. Solid curves are the analytical predictions; markers are numerical estimates averaged over 20 independent runs. Error bars correspond to ± 3SE. (*Middle*) Bias and variance for the RFNN model as a function of $\psi_p = p/d$ for $\psi_n = 8,\ t = 0.1,\ m = 4,\ \lambda = 10^{-3}$ and finite dimension $d = 128$. Solid curves are the analytical predictions obtained from Theorem 2.1 through the bias–variance expressions of Appendix C.9; markers are numerical estimates from 10 independent runs. Error bars correspond to ± 3SE. *(Right)* Bias (full line) and variance (dotted line) as a function of the sample complexity ψ_n for $m = 1$ and $m = 4$ obtained by solving the analytical equations at $\psi_p = \infty$ for $t = 0.1$.

C.13.3 Additional Figures

In this section, we present some additional figures on the RFNN model. The activation function taken for the figures is always $\sigma = \tanh$.

Colormaps of the losses. Fig.16 shows colormaps of the three losses in the (ψ_n, ψ_p) space. We observe that the line $\psi_p = m\psi_n$ delimits a region with low $\mathcal{L}_{\text{test}}^{\text{emp}}$ and $\mathcal{L}_{\text{train}}^{n,m}$ but large $\mathcal{L}_{\text{test}}$. We also observe that the test loss presents two peaks: one at $\psi_p = m\psi_n$ as discussed in the main text as well as another peak at $m\psi_n = 1$ which is related to the two peaks found in d'Ascoli et al. (2021) that were located at $\psi_n = 1$ for supervised learning.

Scaling of the generalization and noise gaps. Fig.17 plots the generalization gap $\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}^{n,m}$ and the noise gap $\mathcal{L}_{\text{test}}^{\text{emp}} - \mathcal{L}_{\text{train}}^{n,m}$ as a function of ψ_p and shows that they are actually functions of ψ_p/ψ_n for a fixed m.

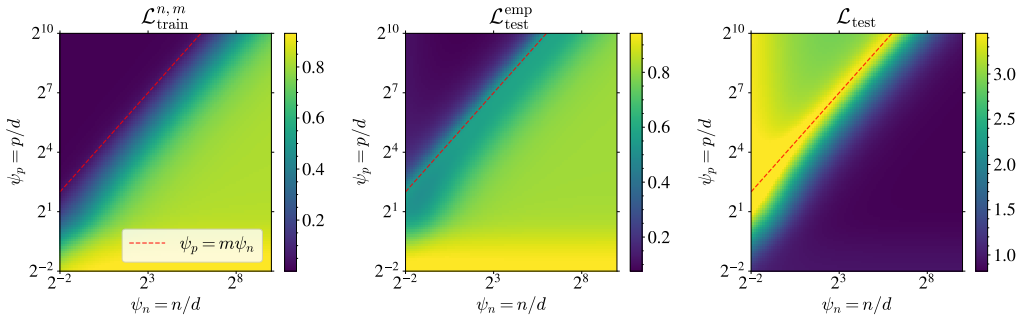

Figure 16: Colormap of the analytical solutions of the three losses for $m = 16,\ t = 0.1$ as a function of ψ_n and ψ_p.

D LLM Usage

We describe here the role played by large language models (LLMs) in the preparation of this work.

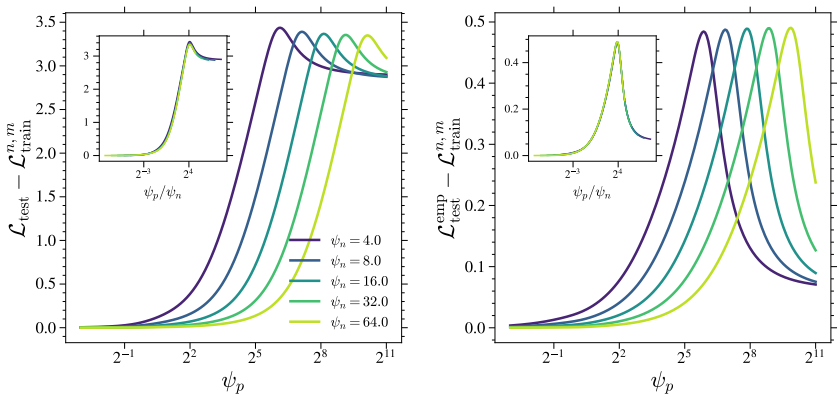

Figure 17: Evolution of the *(Left)* generalization and *(Right)* noise gaps as a function of ψ_p for the RFNN model at $t = 0.1$, $\lambda = 10^{-3}$, $\sigma = \tanh$ and $m = 16$. We observe that both gaps scale as ψ_p/ψ_n.

Code. LLMs were used to write and debug parts of the code, both for the numerical experiments and for the scripts producing the figures of this paper.

Writing and presentation. LLMs were used for copy-editing throughout: correcting grammar, finding typographical errors, and uniformizing notation, style and cross-referencing across sections.

Checking the derivations. We used LLMs to re-check the algebra of our analytical results and to audit the manuscript for internal inconsistencies. In particular, an LLM located a typo in an intermediate step of our computation of the trace $T_1(z, \zeta, 0)$ (Appendix C.8), which had been blocking our progress on the remainder of the derivation.

Responsibility. Every LLM-assisted derivation was checked independently and thoroughly by the authors, and every suggested edit was reviewed before inclusion. The authors take full responsibility for the content, the originality and the scientific integrity of this work, including any remaining errors.