Acknowledgements

I thank my supervisor, Prof. Carlo Lucibello, for his guidance, and Filippo Elgorni, for the invaluable collaboration and support. I am also grateful to Brandon Annesi for his insightful ideas and feedback.

Contents

1 Introduction

Motivation

Generative diffusion models have redefined the landscape of modern machine learning, excelling across image, audio, and molecular generation tasks [1, 2]. These models rely on a stochastic process that progressively corrupts data into noise and then learns to invert this process by estimating the score function $S_\theta(x,t) \approx \nabla_x \log p_t(x)$, which governs the backward dynamics [3, 4].

This paradigm builds upon earlier formulations of non-equilibrium thermodynamics for generative learning [5], and has since evolved into one of the most powerful frameworks for high-fidelity sample synthesis across diverse modalities. Yet, a fundamental paradox lies at the core of their success. If the score used during generation were the exact empirical one computed on the training set, the model would merely reproduce data points. In contrast, real diffusion models not only avoid memorization but generate novel, meaningful samples. Understanding why diffusion models generalize so well, despite their immense capacity and expressive power, is therefore a central theoretical question.

Problem Statement

If the backward process were integrated using the exact empirical score, the resulting trajectories would collapse onto the training examples. The generation process would degenerate into a form of data retrieval, not synthesis. However, empirical evidence shows the opposite: diffusion models consistently produce new, coherent data points that reflect the underlying data distribution rather than memorized examples. This discrepancy points to a subtle mechanism whereby the process of *learning* itself, rather than explicit regularization, induces generalization.

From a statistical physics perspective, this observation echoes behaviors seen in disordered systems such as spin glasses, where over-parameterization and noise can lead to smooth macroscopic order rather than chaotic microstates [6, 7, 8]. Such analogies motivate a deeper investigation into how optimization and model structure jointly determine

the generative dynamics of diffusion models.

Research Objective and Approach

The goal of this work is to investigate how learning, through both the architecture of the network and the dynamics of its optimization, shapes the transition between memorization and generalization. We begin by analyzing the limiting case where the score function equals its empirical counterpart, showing that this regime necessarily leads to memorization. We then study how integrating the backward process using a learned score fundamentally alters generative dynamics and gives rise to generalization [9, 10].

To make these ideas precise, we analyze:

1. **Architectural constraints:** the effect of model capacity and expressivity, in particular the transition between under- and over-parameterized regimes. Understanding how this transition depends on the ratio between parameters, data dimension, and sample size is central to identifying the boundary between these two behaviors.
2. **Training dynamics:** how gradient-flow timescales separate generalization and memorization phases, leading to implicit regularization even when the model is large enough to fit the training data. We show that diffusion models can avoid memorization not because of explicit constraints, but because the training trajectory itself halts before slow modes associated with memorization converge.

By connecting these aspects, we aim to uncover the mechanisms that enable diffusion models to learn smooth, globally coherent scores while avoiding collapse to empirical attractors. Our findings highlight that generalization in diffusion models is not a property of the optimum itself, but a consequence of the dynamical path taken to reach it.

Structure of the Thesis

Chapter 2 introduces the formalism of score-based diffusion models and the denoising score-matching loss. Chapter 3 establishes how exact empirical learning leads to memorization through an analogy with Random Energy Models. Chapter 4 analyzes learning dynamics in Random Feature Neural Networks, analyzing how generating through a learned

score impacts the diffusion backward trajectories and identifying distinct generalization-memorization training timescales. Chapter 5 extends this framework to time-dependent architectures, studying the effects of bias modulation and skip connections. Chapter 6 explores structured data, showing how a BaikBen ArousPéché (BBP) transition marks the emergence of meaningful feature directions. Finally, Chapter 7 summarizes our conclusions and discusses open problems in connecting learning dynamics to generalization in generative models.

2 Score-Based Diffusion Models

We study a stochastic process in $\mathbb{R}^d$ defined by

$$\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{F}(\boldsymbol{x}(t), t) + \boldsymbol{\eta}(t), \qquad \langle \eta_i(t) \rangle = 0, \quad \langle \eta_i(t)\eta_j(s) \rangle = 2\,\delta_{ij}\,\delta(t-s), \tag{2.1}$$

i.e., unit isotropic diffusion (noise strength $\sqrt{2}$). The corresponding Fokker–Planck equation (FPE) for the density $P_t(\boldsymbol{x})$ is

$$\partial_t P_t(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\big(-\boldsymbol{F}(\boldsymbol{x},t)P_t(\boldsymbol{x})\big) + \Delta P_t(\boldsymbol{x}) = \boldsymbol{\nabla}\cdot\big(\boldsymbol{F}(\boldsymbol{x},t)P_t(\boldsymbol{x})\big) + \Delta P_t(\boldsymbol{x}). \tag{2.2}$$

When the force derives from a potential, $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{\nabla} U(\boldsymbol{x})$, a steady-state solution (when it exists) is the GibbsBoltzmann measure

$$P(\boldsymbol{x}) \propto e^{-U(\boldsymbol{x})}. \tag{2.3}$$

OrnsteinUhlenbeck (OU) example. Suppose the potential is quadratic:

$$U(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2}. \tag{2.4}$$

Then (2.1) becomes the OU (Langevin) equation

$$\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{\eta}(t). \tag{2.5}$$

Being linear, this SDE integrates explicitly for a given noise trajectory:

$$\boldsymbol{x}(t) = \boldsymbol{a}\, e^{-t} + \int_0^t e^{-(t-\tau)}\, \boldsymbol{\eta}(\tau)\, d\tau, \qquad \boldsymbol{x}(0) = \boldsymbol{a}. \tag{2.6}$$

Equivalently,

$$\boldsymbol{x}(t) = \boldsymbol{a}\, e^{-t} + \sqrt{\Delta_t}\, \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d), \qquad \Delta_t = \int_0^t 2\, e^{-2(t-\tau)} d\tau = 1 - e^{-2t}. \tag{2.7}$$

The deterministic part contracts toward $\boldsymbol{0}$, while the noise variance grows from 0 to 1. As $t \to \infty$, the law becomes standard Gaussian:

$$P_{t\gg 1}(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\!\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right). \tag{2.8}$$

Forward marginal under unknown P_0. Assume $\boldsymbol{a} \sim P_0$ and follow the OU forward process. Conditioning on $\boldsymbol{a}$,

$$\boldsymbol{x}(t) \mid \boldsymbol{a} \;\sim\; \mathcal{N}\big(\boldsymbol{a}\, e^{-t},\, \Delta_t\, \mathbb{I}_d\big). \tag{2.9}$$

Solving the associated FPE for fixed $\boldsymbol{a}$ and averaging over initial conditions yields the marginal density

$$P_t(\boldsymbol{x}_t) = \int d\boldsymbol{a}\, P_0(\boldsymbol{a})\, P(\boldsymbol{x}_t \mid \boldsymbol{a}) = \int d\boldsymbol{a}\, P_0(\boldsymbol{a})\, \frac{\exp\!\Big(-\frac{\|\boldsymbol{x}_t - \boldsymbol{a} e^{-t}\|^2}{2\,\Delta_t}\Big)}{(2\pi\Delta_t)^{d/2}}. \tag{2.10}$$

The challenge is that P_t is unknown because P_0 is unknown.

Reverse-time dynamics and the score. Let $T \gg 1$ and define the reverse time $\tau = T - t$. In reverse time, the FPE reads

$$\partial_\tau \tilde{P}_\tau(\boldsymbol{x}) = -\,\boldsymbol{\nabla}\cdot\big(\boldsymbol{x}\, \tilde{P}_\tau(\boldsymbol{x})\big) + \Delta \tilde{P}_\tau(\boldsymbol{x}), \qquad \tilde{P}_\tau(\boldsymbol{x}) = P_{T-\tau}(\boldsymbol{x}). \tag{2.11}$$

The associated reverse-time SDE is

$$\frac{d\boldsymbol{x}}{d\tau} = \boldsymbol{x}(\tau) \; + \; 2\,\boldsymbol{\nabla} \log \tilde{P}_\tau(\boldsymbol{x}) \; + \; \tilde{\boldsymbol{\eta}}(\tau), \tag{2.12}$$

with fresh Gaussian noise $\tilde{\boldsymbol{\eta}}$ of the same statistics; see Appendix .1 and classical time-reversal results [11]. Using $\tilde{P}_\tau(\boldsymbol{x}) = P_{T-\tau}(\boldsymbol{x})$, the additional drift is the *score*

$$S(\boldsymbol{x}, \tau) \; = \; \boldsymbol{\nabla} \log P_{T-\tau}(\boldsymbol{x}). \tag{2.13}$$

Hence, if we knew $S(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log P_t(\boldsymbol{x})$, we could sample from P_0 by drawing $\bar{\boldsymbol{x}}_T \sim P_T \approx \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d)$ (for $T \gg 1$) and integrating (2.12) backward to $\tau = T$ (i.e., $t = 0$). In particular, $\bar{\boldsymbol{x}}_0 \sim P_0$ [5, 3, 1, 2].

Score expressed via the forward posterior. From (2.10),

$$S(\boldsymbol{x}, t) = \boldsymbol{\nabla} \log P_t(\boldsymbol{x}) = \frac{\boldsymbol{\nabla} P_t(\boldsymbol{x})}{P_t(\boldsymbol{x})} = -\frac{1}{P_t(\boldsymbol{x})} \int d\boldsymbol{a}\; P_0(\boldsymbol{a})\, \frac{\boldsymbol{x} - \boldsymbol{a}e^{-t}}{\Delta_t}\, \frac{\exp\!\left(-\frac{\|\boldsymbol{x} - \boldsymbol{a}e^{-t}\|^2}{2\Delta_t}\right)}{(2\pi\Delta_t)^{d/2}} \tag{2.14}$$

$$= -\frac{\boldsymbol{x}}{\Delta_t} + \frac{e^{-t}}{\Delta_t} \underbrace{\int d\boldsymbol{a}\, \boldsymbol{a}\, \frac{P_0(\boldsymbol{a})\, P(\boldsymbol{x}_t \mid \boldsymbol{a})}{P_t(\boldsymbol{x})}}_{= \mathbb{E}[\boldsymbol{a} \mid \boldsymbol{x}_t = \boldsymbol{x}]} \; = \; -\frac{\boldsymbol{x}}{\Delta_t} + \frac{e^{-t}}{\Delta_t}\, \mathbb{E}[\boldsymbol{a} \mid \boldsymbol{x}_t = \boldsymbol{x}]. \tag{2.15}$$

The conditional mean $\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{x}_t]$ is the *Bayesian denoiser* and is optimal among all denoisers under squared error [4].

Learning the score. Let $\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ be a parametric estimate. A natural objective is the time-weighted MSE between the estimate and the true score,

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \int_0^T \lambda(t)\, \mathbb{E}_{\boldsymbol{x}_t \sim P_t}\Big[\big\|\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\nabla} \log P_t(\boldsymbol{x}_t)\big\|^2\Big]\, dt, \tag{2.16}$$

where $\lambda(t)$ reweights different times. Expanding the square, the $\|\nabla \log P_t\|^2$ term is constant w.r.t. θ. For the cross term, note that

$$\mathbb{E}_{\boldsymbol{x}_t\sim P_t}\Big[\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla\log P_t(\boldsymbol{x}_t)\Big] = \int d\boldsymbol{x}_t\,\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla P_t(\boldsymbol{x}_t) \tag{2.17}$$

$$= -\,\mathbb{E}_{\boldsymbol{a}\sim P_0}\int d\boldsymbol{x}_t\,P(\boldsymbol{x}_t\mid\boldsymbol{a})\;\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\frac{\boldsymbol{x}_t-\boldsymbol{a}e^{-t}}{\Delta_t}. \tag{2.18}$$

Using $\boldsymbol{x}_t=\boldsymbol{a}e^{-t}+\sqrt{\Delta_t}\,\boldsymbol{\xi}$ with $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)$, the resulting objective is

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \int_0^T \lambda(t)\,\mathbb{E}_{\boldsymbol{a}\sim P_0}\,\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)}\left[\left\|\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}\big(\boldsymbol{a}e^{-t}+\sqrt{\Delta_t}\,\boldsymbol{\xi},\,t\big)-\frac{\boldsymbol{x}_t-\boldsymbol{a}e^{-t}}{\Delta_t}\right\|^2\right]dt. \tag{2.19}$$

Since $\boldsymbol{x}_t-\boldsymbol{a}e^{-t}=\sqrt{\Delta_t}\,\boldsymbol{\xi}$, the target can be written in the standard DSM form [4, 3]:

$$\frac{\boldsymbol{x}_t-\boldsymbol{a}e^{-t}}{\Delta_t} \;=\; \frac{\sqrt{\Delta_t}\,\boldsymbol{\xi}}{\Delta_t} \;=\; \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}, \tag{2.20}$$

so that

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \int_0^T \lambda(t)\,\mathbb{E}_{\boldsymbol{a}\sim P_0}\,\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)}\left[\left\|\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}\big(\boldsymbol{a}e^{-t}+\sqrt{\Delta_t}\,\boldsymbol{\xi},\,t\big)+\frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right\|^2\right]dt. \tag{2.21}$$

Replacing the expectation over P_0 with the empirical average over samples a^μ, $\mu=1,\dots,n$, and sampling $t\sim Q(t)$ instead of integrating,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\mu=1}^{n}\mathbb{E}_{t\sim Q(t)}\,\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)}\left[\left\|\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}\big(\boldsymbol{x}_t^\mu(\boldsymbol{\xi}),t\big)+\frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right\|^2\right] \tag{2.22}$$

Remarks and pointers. (i) Eqs. (2.15)(2.21) make explicit the link between the score and the Bayes-optimal denoiser [4]. (ii) The OU noising in (2.5)(2.7) is the continuous-time analogue of variance-preserving diffusion used in DDPM/score models [5, 1, 2]. (iii) Time reversal (2.12) follows the classic constant-diffusion case [11].

3 Memorization vs. Generalization

Our goal is to learn the score at each diffusion time:

$$S_t(\boldsymbol{x}) \;=\; \boldsymbol{\nabla}_{\boldsymbol{x}} \log P_t(\boldsymbol{x}). \tag{3.1}$$

Assume the estimator $\hat{\boldsymbol{S}}^{\boldsymbol{\theta}}$ is sufficiently expressive and trained optimally.

We start from the training data $\{\boldsymbol{a}^\mu\}_{\mu=1}^n$ sampled from the empirical distribution

$$P_0^{\text{emp}}(\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{\mu=1}^{n} \delta\big(\boldsymbol{x}-\boldsymbol{a}^\mu\big). \tag{3.2}$$

At diffusion time t, the model observes the forward-evolved points, whose distribution is

$$P_t^{\text{emp}}(\boldsymbol{x}) \;=\; \int d\boldsymbol{a}\; P_0^{\text{emp}}(\boldsymbol{a})\, P(\boldsymbol{x}\mid \boldsymbol{a}), \qquad \boldsymbol{x} = \boldsymbol{a}e^{-t} + \sqrt{\Delta_t}\,\boldsymbol{\xi}, \tag{3.3}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d)$ and

$$P(\boldsymbol{x}\mid \boldsymbol{a}) \;=\; \frac{\exp\!\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{a}e^{-t}\|^2}{2\,\Delta_t}\Big)}{(2\pi\Delta_t)^{d/2}}, \qquad \Delta_t = 1-e^{-2t}. \tag{3.4}$$

Hence,

$$P_t^{\text{emp}}(\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{\mu=1}^{n} \frac{\exp\!\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\,\Delta_t}\Big)}{(2\pi\Delta_t)^{d/2}}, \tag{3.5}$$

i.e., a mixture of equally weighted Gaussians centered at the forward-evolved data with common variance Δ_t.

An ideal learner that minimizes the score-MSE at each time would learn the corresponding empirical score

$$S_t^{\text{emp}}(\boldsymbol{x}) \;=\; \boldsymbol{\nabla}_{\boldsymbol{x}} \log P_t^{\text{emp}}(\boldsymbol{x}) \;=\; -\frac{\boldsymbol{x}}{\Delta_t} + \frac{\sum_{\mu=1}^{n} \boldsymbol{a}^\mu e^{-t}\, \exp\!\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t}\Big)}{\sum_{\mu=1}^{n} \exp\!\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t}\Big)}. \tag{3.6}$$

This S_t^{emp} is the (unique) minimizer of

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{\mu=1}^{n} \mathbb{E}_{t\sim Q(t)} \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbb{I}_d)} \left[\left\| \hat{\boldsymbol{S}}^{\boldsymbol{\theta}}\big(\boldsymbol{x}_t^{\mu}(\boldsymbol{\xi}), t\big) + \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}} \right\|^2 \right]. \tag{3.7}$$

But is S_t^{emp} what we actually want? [12]

3.1 Random Energy Models

Consider the Ising model with N spins $\sigma_i \in \{-1, 1\}$ and configuration space size 2^N. In spin glasses, couplings J_{ij} are random and

$$E(\boldsymbol{\sigma}) \;=\; -\sum_{\langle i,j\rangle} J_{ij}\,\sigma_i\,\sigma_j, \tag{3.8}$$

so energies of different configurations are correlated. In the Random Energy Model (REM), by contrast, the energies themselves are IID Gaussians

$$E_i \sim \mathcal{N}\!\left(0, \frac{N}{2}\right), \tag{3.9}$$

independent across the 2^N configurations [6].

The partition function is

$$Z \;=\; \sum_{i=1}^{2^N} e^{-E_i/T} \;=\; \int de\, \mathcal{N}(e)\, e^{-N\,e/T}, \qquad e = \frac{E}{N}, \tag{3.10}$$

where $\mathcal{N}(e)$ counts configurations with intensive energy in $[e, e+de]$. Averaging over the disorder,

$$\mathbb{E}\big[\mathcal{N}(e)\big] \;=\; 2^N\, P(e) \;\sim\; \exp\!\big(N\,[\log 2 - e^2]\big), \tag{3.11}$$

so for $S(e) := \log 2 - e^2 > 0$, $\mathcal{N}(e)$ concentrates around its mean (fluctuations are exponentially small). In this regime,

$$Z \;\approx\; \int de\; \exp\!\Big(N\,[S(e) - e/T]\Big) \;\overset{N\to\infty}{\asymp}\; \exp\!\Big(N\,[S(\tilde{e}) - \tilde{e}/T]\Big), \tag{3.12}$$

with $\tilde{e} = \arg\max_e\{S(e) - e/T\}$ determined by $\partial_e S(\tilde{e}) = 1/T$. At high T, $\tilde{e} \approx 0$ and many configurations contribute (entropy near $\log 2$); as T decreases, $\tilde{e} \to -\sqrt{\log 2}$ and the effective number of contributing configurations $\mathcal{N}(\tilde{e})$ shrinks to $O(1)$. This is the classic REM transition: for $T > T_c$ the sum self-averages (law of large numbers), while for $T < T_c$ it is dominated by a few extremal terms.

3.2 The Perfect Machine

Return to generation with drift determined by the empirical score $S_t^{\text{emp}}(\boldsymbol{x})$. Assume data are drawn i.i.d. from an unknown true P_0; we only observe samples $\boldsymbol{a}^\mu \sim P_0$. The perfect machine learns from P_0^{emp} and P_t^{emp} and thus estimates S_t^{emp}, whereas we actually care about

$$P_0(\boldsymbol{x}), \qquad P_t(\boldsymbol{x}) = \int d\boldsymbol{a}\; P_0(\boldsymbol{a})\, \frac{\exp\!\left(-\frac{\|\boldsymbol{x}-\boldsymbol{a}e^{-t}\|^2}{2\Delta_t}\right)}{(2\pi\Delta_t)^{d/2}}, \qquad S_t(\boldsymbol{x}) = \boldsymbol{\nabla}_{\boldsymbol{x}} \log P_t(\boldsymbol{x}). \tag{3.13}$$

The danger is clear: learning from empirical statistics alone tends to reproduce the training data (memorization), rather than generalize.

Operational definitions. We say the model *generalizes* at time t if $P_t^{\text{emp}}(\boldsymbol{x}) \approx P_t(\boldsymbol{x})$ for typical $\boldsymbol{x} \sim P_t$. Conversely, it *memorizes* at time t if $P_t^{\text{emp}}(\boldsymbol{x}) \not\approx P_t(\boldsymbol{x})$ at typical $\boldsymbol{x} \sim P_t$. When $P_t^{\text{emp}} \approx P_t$, then $S_t^{\text{emp}} \approx S_t$ and reverse-time trajectories land on typical points of the true distribution, precisely what we want for sampling from P_0.

REM analogy for self-averaging. Fix $\boldsymbol{x} \sim P_t$. For each μ, the term

$$\frac{1}{n}\frac{\exp\!\left(-\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t}\right)}{(2\pi\Delta_t)^{d/2}} \;=\; e^{-E_\mu/T} \tag{3.14}$$

is random due to the randomness of $\boldsymbol{a}^\mu$, with independent terms across μ. (Here T is simply a placeholder, absorbing constants into E_μ sets $T = 1$; one may also keep the

explicit normalization terms inside E_μ.) Thus

$$P_t^{\text{emp}}(\boldsymbol{x}) \;=\; \sum_{\mu=1}^{n} e^{-E_\mu/T}, \tag{3.15}$$

a sum of independent exponentials, in direct analogy with the REM partition sum Z [6].

Averaging over data disorder,

$$\left\langle P_t^{\text{emp}}(\boldsymbol{x})\right\rangle_{\{\boldsymbol{a}^\mu\}} = \frac{1}{n}\sum_{\mu=1}^{n}\left\langle \frac{\exp\!\left(-\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t}\right)}{(2\pi\Delta_t)^{d/2}}\right\rangle_{\boldsymbol{a}^\mu} = \tag{3.16}$$

$$= \int d\boldsymbol{a}\; P_0(\boldsymbol{a})\frac{\exp\!\left(-\frac{\|\boldsymbol{x}-\boldsymbol{a}e^{-t}\|^2}{2\Delta_t}\right)}{(2\pi\Delta_t)^{d/2}} = P_t(\boldsymbol{x}). \tag{3.17}$$

Therefore $P_t^{\text{emp}}(\boldsymbol{x})$ is *self-averaging* (i.e., concentrates around its mean) exactly when the system is in the generalization regime $P_t^{\text{emp}} \approx P_t$.

Dimensionality and scaling. In REM, Z sums 2^N IID exponentials with extensive energies $E_i = O(N)$. Here, P_t^{emp} sums n terms, while $\boldsymbol{a}^\mu, \boldsymbol{x}_t \in \mathbb{R}^d$. Assuming $\|\boldsymbol{x}_t\| = O(\sqrt{d})$, the energies scale as

$$E_\mu \;\propto\; \frac{\|\boldsymbol{x}_t - \boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t} \;=\; O(d). \tag{3.18}$$

Thus n plays the role of 2^N, and d that of N. Taking $n, d \to \infty$ with

$$\alpha \;:=\; \frac{\log n}{d} \tag{3.19}$$

held fixed yields a direct REM-like thermodynamic limit.

Temperature is diffusion time. In REM, small T favors low-energy (few-term dominated) sums. In our setting, diffusion time t plays the analogous role: small t produces narrow Gaussians and domination by nearest terms (memorization), while large t widens Gaussians so many terms contribute (self-averaging/generalization).

Formally, draw $\boldsymbol{x}_t \sim P_t^{\text{emp}}$ and define the rescaled distances

$$\delta_{\boldsymbol{a}^\mu} \;=\; \frac{\|\boldsymbol{x}_t - \boldsymbol{a}^\mu e^{-t}\|}{\sqrt{d}}. \tag{3.20}$$

Then

$$P_t^{\text{emp}}(\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{\mu=1}^{n} \frac{\exp\!\Big(-\frac{d\,\delta_{\boldsymbol{a}^\mu}^2}{2\Delta_t}\Big)}{(2\pi\Delta_t)^{d/2}} \;=\; \int d\delta\, \mathcal{N}(\delta)\; \exp\!\left(-\frac{d\,\delta^2}{2\Delta_t} - \frac{d}{2}\log(2\pi\Delta_t)\right), \tag{3.21}$$

where $\mathcal{N}(\delta)$ counts the number of samples with rescaled distance in $[\,\delta, \delta + d\delta\,]$. One can show

$$\mathcal{N}(\delta) \;\asymp\; \exp\big(d\,S(\delta)\big), \tag{3.22}$$

and the Laplace method then yields a REM-style competition between an entropy $S(\delta)$ and an energy $\delta^2/(2\Delta_t)$. The sum self-averages when the maximizing δ lies in a region with $S(\delta) > 0$; the *transition time* t_c solves $S(\delta^\star(t_c)) = 0$.

Geometric viewpoint. Empirically, data at each t often concentrate near a low-dimensional manifold. Since $P_t^{\text{emp}} = \frac{1}{n}\sum_\mu \mathcal{N}\big(\boldsymbol{a}^\mu e^{-t}, \Delta_t I_d\big)$, small t yields highly localized bumps: only nearest data contribute at a given $\boldsymbol{x}_t$. At large t, bumps overlap broadly: many data contribute. A natural criterion for the transition compares the volume covered by the n Gaussians to the volume of the true distribution:

$$V^G(t) = n\,\exp\big(d\,S^G(t)\big), \quad S^G(t) = \tfrac{1}{2}\log\big(2\pi e\,\Delta_t\big), \tag{3.23}$$

$$V(t) = \exp\big(d\,S^P(t)\big), \tag{3.24}$$

and we set the transition by

$$V^G(t_c) = V(t_c) \iff \log n + d\,S^G(t_c) = d\,S^P(t_c) \iff \alpha + S^G(t_c) = S^P(t_c). \tag{3.25}$$

Thus t_c depends on $\alpha = \frac{\log n}{d}$: more data $\Rightarrow$ later memorization. Crucially, for finite α, $t_c > 0$. *memorization is inevitable at sufficiently small t*: a diffusion-model curse in the

empirical-score limit.

4 Why Don't Diffusion Models Memorize In Practice?

The previous chapter has shown that, in the perfect-machine framework introduced by [13], memorization is an unavoidable consequence of optimal learning. When the empirical score function is learned exactly, i.e., when the model minimizes the empirical denoising-score-matching (DSM) loss over all possible functions, each training point becomes an attractor of the backward dynamics.

In the high-dimensional limit $n = \mathcal{O}(e^d)$, the empirical distribution $P_0^{\text{emp}}(\boldsymbol{x})$ differs from the population one $P_0(\boldsymbol{x})$ for all finite data densities, and the generated samples collapse onto the training set. This result establishes a clear theoretical baseline: *in the absence of any constraint, generalization is impossible.*

Yet, this picture differs sharply from what is observed phenomenologically. Modern diffusion models, trained with finite data and parameterized networks, routinely generate novel samples while achieving very good generative performances. Empirical evidence shows that memorization appears only under specific conditions (small datasets, excessive training time, or extreme over-parameterization) and can otherwise be avoided without explicit regularization.

The existence of such expressive yet non-memorizing models indicates that mechanisms controlling generalization in real systems cannot be captured by a static analysis of the empirical optimum alone.

From perfect machine to learned models. To move beyond the perfect-machine limit, one must account for the *learning process* itself. In practice, the score function is chosen from a parametric family, and the DSM objective is minimized within this restricted class. The goal from now on is to clarify how *architectural constraints* and *training dynamics* determine the transition between generalization and memorization, following the

program of [9, 10].

4.1 RFNN And Gaussian Data

We adopt a Random Feature Neural Network (RFNN) as an analytically tractable framework to study DSM dynamics and the emergence of memorization. The RFNN is a two-layer neural network amenable to theory and able to capture phenomena observed in richer models (e.g., double-descent-like behavior in over-parameterized regimes).

Model. Let $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ be frozen Gaussian first-layer weights and $\boldsymbol{A} \in \mathbb{R}^{d\times p}$ be trainable second-layer weights. With an element-wise activation σ, the score model maps $\mathbb{R}^d \to \mathbb{R}^d$ as

$$\boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}) \;=\; \frac{\boldsymbol{A}}{\sqrt{p}}\,\sigma\!\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right). \tag{4.1}$$

Data and limit. We consider a training set of n i.i.d. samples $\boldsymbol{x}^\nu \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d)$, and the high-dimensional limit $d, p, n \to \infty$ with ratios $\psi_p = p/d$ and $\psi_n = n/d$ fixed.

Sampling protocol. Following [9, 10], at each time t a distinct RFNN learns the score for that time. We can think at the generative process as follows:

1. Discretize the reverse SDE.
2. For each grid point t_i, train $\boldsymbol{S}_{\boldsymbol{A}}^{t_i}$ to learn the score at t_i.
3. Sample $\bar{\boldsymbol{x}}_T \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d)$.
4. Starting from $t_0 = T$, use $\bar{\boldsymbol{x}}_{t_i}$ and $\boldsymbol{S}_{\boldsymbol{A}}^{t_i}$ to generate $\bar{\boldsymbol{x}}_{t_{i+1}}$ via the discretized reverse SDE.
5. Stop at $\bar{\boldsymbol{x}}_0$.

4.1.1 True Gaussian Score

Since $\boldsymbol{x}^\nu \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d)$, we can derive $P_t(\boldsymbol{x}_t)$ explicitly. Writing the OU forward kernel with variance $\Delta_t = 1 - e^{-2t}$,

$$P_t(\boldsymbol{x}_t) = \int d\boldsymbol{x}\; P_0(\boldsymbol{x})\, P(\boldsymbol{x}_t|\boldsymbol{x}) = \int d\boldsymbol{x}\; \frac{e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma_{\boldsymbol{x}}^2}}}{(2\pi\sigma_{\boldsymbol{x}}^2)^{d/2}} \frac{e^{-\frac{\|\boldsymbol{x}_t - \boldsymbol{x}e^{-t}\|^2}{2\,\Delta_t}}}{(2\pi\,\Delta_t)^{d/2}} \tag{4.2}$$

$$= \frac{\exp\!\big(-\frac{\|\boldsymbol{x}_t\|^2}{2\Gamma_t^2}\big)}{(2\pi\Gamma_t^2)^{d/2}}, \qquad \Gamma_t^2 = \sigma_{\boldsymbol{x}}^2 e^{-2t} + \Delta_t. \tag{4.3}$$

Thus $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \Gamma_t^2 \mathbb{I}_d)$ and the *true* score is

$$\boldsymbol{S}_t(\boldsymbol{x}) = \boldsymbol{\nabla}_{\boldsymbol{x}} \log P_t(\boldsymbol{x}) = -\frac{\boldsymbol{x}}{\Gamma_t^2}. \tag{4.4}$$

4.1.2 Metrics

Figure. 1 shows train (solid) and test (dashed) losses versus ψ_n at $t = 0.1$, illustrating these trends.

Train loss.

$$\mathcal{L}_{\text{train}}(\boldsymbol{A}, \{\boldsymbol{x}^\nu\}, t) = \frac{1}{nd} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \Big\| \sqrt{\Delta_t}\, \boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}_t^\nu(\boldsymbol{\xi})) + \boldsymbol{\xi} \Big\|^2 \tag{4.5}$$

Test loss.

$$\mathcal{L}_{\text{test}}(\boldsymbol{A}, t) = \frac{1}{d}\, \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi}} \Big\| \sqrt{\Delta_t}\, \boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}_t(\boldsymbol{\xi})) + \boldsymbol{\xi} \Big\|^2, \qquad \boldsymbol{x}_t(\boldsymbol{\xi}) = e^{-t}\boldsymbol{x} + \sqrt{\Delta_t}\, \boldsymbol{\xi}. \tag{4.6}$$

Generalization gap.

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}. \tag{4.7}$$

Score error.

$$\epsilon_{\text{score}}(\boldsymbol{A}, t) = \frac{1}{d}\, \mathbb{E}_{\boldsymbol{x} \sim P_t} \Big\| \boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}) + \frac{\boldsymbol{x}}{\Gamma_t^2} \Big\|^2. \tag{4.8}$$

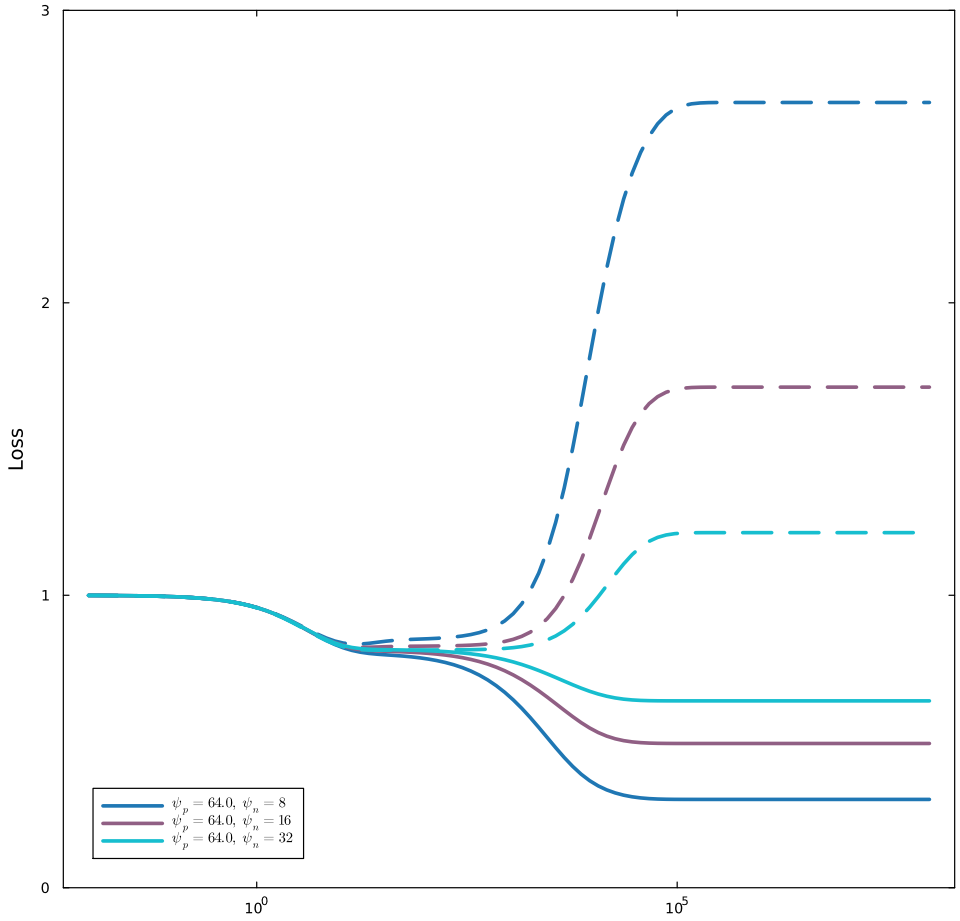


Figure 1: Training (solid) and test (dashed) loss for several values of ψ_n at $t = 0.1$.

This controls generation quality: if $\hat{P}_t$ denotes the reverse-process law under the learned score and $\hat{P}_t = P_t$, then

$$D_{\mathrm{KL}}(P_0 \,\|\, \hat{P}_0) \;\le\; \frac{d}{2} \int_0^T \epsilon_{\mathrm{score}}(\boldsymbol{A}, t)\, dt. \tag{4.9}$$

4.2 Understanding The Learned Score

Theorem 4.1 (Train–test decomposition at fixed t)**.** *Define*

$$\mathcal{M}_t = \frac{1}{nd} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left\| \frac{\boldsymbol{A}}{\sqrt{p}} \sigma\Big(\frac{\boldsymbol{W} \boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} \Big) - \boldsymbol{S}_t^{\mathrm{emp}}\big(\boldsymbol{x}_t^\nu(\boldsymbol{\xi})\big) \right\|^2 \tag{4.10}$$

$$\mathcal{V}_t = \frac{1}{nd} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left\| \sqrt{\Delta_t}\, \boldsymbol{S}_t^{\mathrm{emp}}\big(\boldsymbol{x}_t^\nu(\boldsymbol{\xi})\big) + \boldsymbol{\xi} \right\|^2. \tag{4.11}$$

Then

$$\mathcal{L}_{\text{train}}(\boldsymbol{A}, \{\boldsymbol{x}^\nu\}, t) = \mathcal{V}_t + \Delta_t \, \mathcal{M}_t. \tag{4.12}$$

Proof. See Appendix .2, Theorem .7.

Interpretation. Since $\mathcal{V}_t$ is independent of $\boldsymbol{A}$, any change in $\mathcal{L}_{\text{train}}$ as ψ_p varies must be due to $\mathcal{M}_t$. Small $\mathcal{M}_t$ means the learned score is close to $\boldsymbol{S}_t^{\text{emp}}$; however, for small t where $\Delta_t \approx 0$, P_t^{emp} is dominated, near $e^{-t}\boldsymbol{x}^\nu$, by the ν-th term, so

$$\boldsymbol{S}_t^{\text{emp}}(\boldsymbol{x}) \sim -\frac{\boldsymbol{x} - e^{-t}\boldsymbol{x}^\nu}{\Delta_t},$$

which, in reverse time, adds a drift toward $e^{-t}\boldsymbol{x}^\nu$. Trajectories then move toward training samples*memorization*. Thus, learning $\boldsymbol{S}_t^{\text{emp}}$ too well can *harm* generalization.

Learning curves by t.

- $t = \infty$**.** $e^{-t} = 0$, $\Delta_t = 1$, $\boldsymbol{S}_t^{\text{emp}}(\boldsymbol{x}) \sim -\boldsymbol{x}$. Then $\mathcal{V}_t = 0$ and

  $$\mathcal{L}_{\text{train}}(\boldsymbol{A}, \{\boldsymbol{x}^\nu\}, t = \infty) = \mathcal{M}_\infty = \mathbb{E}_{\boldsymbol{\xi}} \left\| \tfrac{\boldsymbol{A}}{\sqrt{p}} \sigma\left(\tfrac{\boldsymbol{W}\boldsymbol{\xi}}{\sqrt{d}}\right) + \boldsymbol{\xi} \right\|^2.$$

  Train and test coincide, both decreasing with ψ_p as linear fields are better approximated.
- $t \gg 1$**.** Still $\boldsymbol{S}_t^{\text{emp}}(\boldsymbol{x}) \sim -\boldsymbol{x}$; $\mathcal{V}_t \sim e^{-2t}$. Thus $\mathcal{L}_{\text{train}} \sim \Delta_t \mathcal{M}_t + e^{-2t}$. For large ψ_p, $\mathcal{M}_t$ is small and the e^{-2t} term dominates training error, while test $\approx \mathcal{M}_t$ stays flat.
- $t \ll 1$**.** $\boldsymbol{S}_t^{\text{emp}}(e^{-t}\boldsymbol{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}) \sim -\boldsymbol{\xi}/\sqrt{\Delta_t}$, hence $\mathcal{V}_t \sim 0$ and $\mathcal{L}_{\text{train}} \sim \Delta_t \mathcal{M}_t$. As ψ_p increases, train loss decreases (better fit to $\boldsymbol{S}_t^{\text{emp}}$) while test loss increases (worse fit to the population score), indicating memorization onset for small t and large ψ_p.

Crossover by ψ_p/ψ_n. Fix small t and plot versus ψ_p at several ψ_n:

- $\psi_p \ll \psi_n$: train $\approx$ constant and test small (generalization).
- $\psi_p \gg \psi_n$: train small and test large (memorization).
- $\psi_p \approx \psi_n$: rapid test rise (onset of memorization).

Thus $\psi_p = \psi_n$ is a *crossover* where the score transitions from generalization to memorization (gradually, not sharply). Intuitively, at small t, P_t^{emp} is a sum of nearly delta-like peaks; with limited p, the network produces a smoother, regularized field.

4.3 Supervised Vs. Diffusion Learning: Two Paradigms Of Generalization

In supervised learning, test error vs. complexity often shows *double descent*: a first decrease (under-parameterized), a peak near interpolation $p \simeq n$, and a second decrease (over-parameterized) [14]. There, exact interpolation may coexist with generalization (benign overfitting); inductive bias chooses good interpolants.

Diffusion learning is fundamentally different. The DSM objective estimates a *vector field* $S(\boldsymbol{x}, t) = \boldsymbol{\nabla} \log P_t(\boldsymbol{x})$ via averages over infinitely many Gaussian perturbations. Interpolation is ill-defined because the loss at each t averages over uncountably many noisy versions of each point. [9] make this explicit: their Lemma 3.4 yields, asymptotically,

$$\epsilon_{train}^{\infty}(\hat{\boldsymbol{A}}) = \mathcal{V}_t + \Delta_t \mathcal{M}_t,$$

where $\mathcal{V}_t$ depends only on S_t^{emp} (perturbation variance) and not on parameters, while $\mathcal{M}_t$ is the deviation from S_t^{emp}. As p increases, $\mathcal{M}_t$ decreasesbringing the learned score closer to S_t^{emp}but in the small-t regime, this *hurts* generalization: S_t^{emp} is sharply peaked and drives collapse to data. Keeping $\mathcal{M}_t > 0$ (via finite capacity, stochasticity, early stopping) yields smoother fields closer to the population score. At large t, $\mathcal{V}_t$ dominates (with $S_t^{\text{emp}} \sim -\boldsymbol{x}$), explaining train=test coincidence. Thus, unlike supervised learning, the train minimum ($\mathcal{M}_t \to 0$) corresponds to *perfect memorization*, not generalization. This matches the memorizationgeneralization dichotomy reported empirically in diffusion models [15].

4.3.1 Absence Of Double Descent And The Memorization–Generalization Dichotomy

Yoon et al. (2023) [15] show experimentally a striking dichotomy: *generalization and memorization are mutually exclusive; conceptual learning occurs precisely when rote learning fails.* Small datasets or overly expressive networks memorize reproduce training samples; adding unrelated dummy data or noise, consuming capacity, restores generalization. This mirrors [9]: pushing $\mathcal{M}_t \downarrow$ improves train loss but drives memorization; keeping $\mathcal{M}_t > 0$ (finite p/n, finite noise sampling, early stopping) yields the best test loss. The critical line $p \simeq n$ replaces the interpolation peak of double descent; beyond it, increasing capacity destroys generalization.

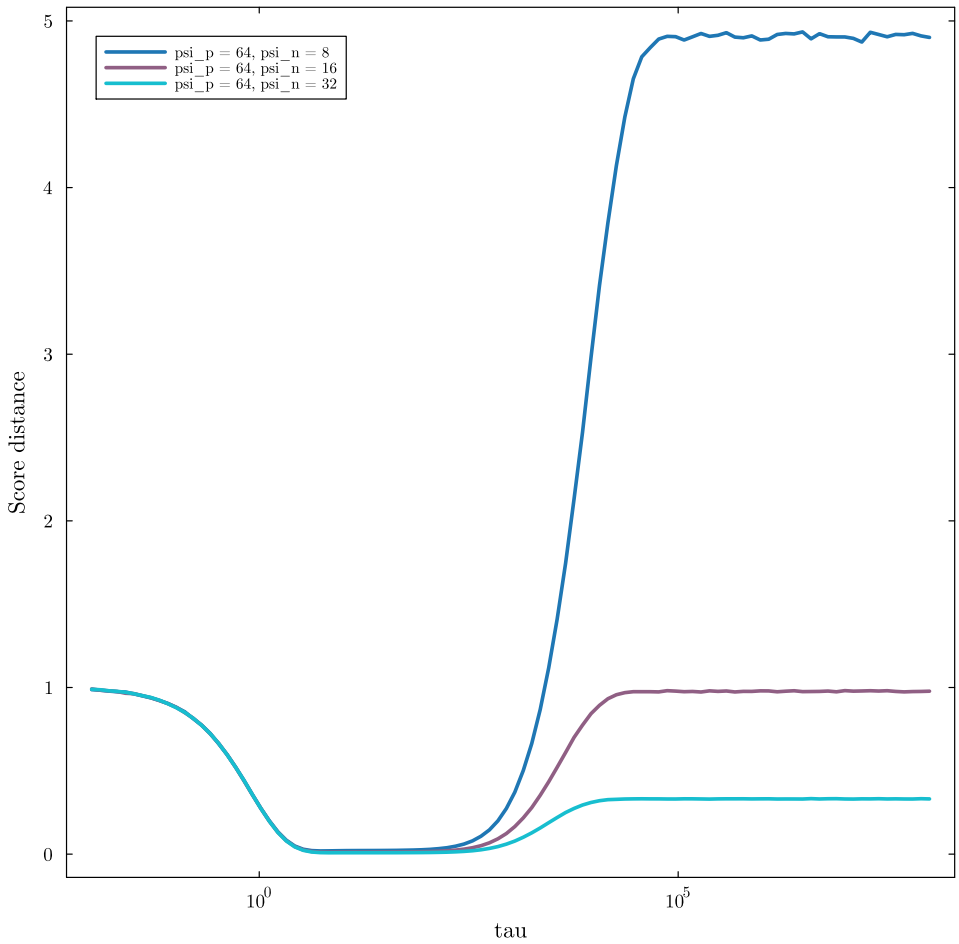


Figure 2: Distance from the true score for several values of ψ_n at $t = 0.1$.

4.4 Training Dynamics

So far we assumed the score is learned by minimizing the empirical DSM loss. We now analyze the *dynamics* of this learning at fixed t, with a $1/d$ rescaling for a finite large-d limit:

$$\mathcal{L}_t(\boldsymbol{A}, \{\boldsymbol{x}^\nu\}) = \frac{1}{nd} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \Big\| \sqrt{\Delta_t} \frac{\boldsymbol{A}}{\sqrt{p}} \sigma\Big(\frac{\boldsymbol{W} \boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} \Big) + \boldsymbol{\xi} \Big\|^2. \tag{4.13}$$

Gradient descent:

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{A}^{(k)} - \eta \, \boldsymbol{\nabla}_{\boldsymbol{A}} \mathcal{L}_t(\boldsymbol{A}^{(k)}), \tag{4.14}$$

with learning rate η. In the high-dimensional limit, as $\eta \to 0$ and with $\tau = \frac{k\eta}{d^2}$, the dynamics converges to

$$\dot{\boldsymbol{A}}(\tau) = -\, d^2 \, \boldsymbol{\nabla}_{\boldsymbol{A}} \mathcal{L}_t(\boldsymbol{A}(\tau)) = -2 \, \Delta_t \, \frac{d}{p} \boldsymbol{A} \boldsymbol{U} - \frac{2 \, d \sqrt{\Delta_t}}{\sqrt{p}} \, \boldsymbol{V}^\top, \tag{4.15}$$

with

$$\boldsymbol{U} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \big[\sigma(\tfrac{\boldsymbol{W} \boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}) \sigma(\tfrac{\boldsymbol{W} \boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}})^\top \big], \qquad \boldsymbol{V} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \big[\sigma(\tfrac{\boldsymbol{W} \boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}) \, \boldsymbol{\xi}^\top \big]. \tag{4.16}$$

This linear ODE solves exactly:

$$\frac{\boldsymbol{A}(\tau)}{\sqrt{p}} = \left[\frac{\boldsymbol{A}_0}{\sqrt{p}} + \frac{1}{\sqrt{\Delta_t}} \boldsymbol{V}^\top \boldsymbol{U}^{-1} \right] \exp\left(-\frac{2\Delta_t}{\psi_p} \boldsymbol{U} \, \tau \right) - \frac{1}{\sqrt{\Delta_t}} \boldsymbol{V}^\top \boldsymbol{U}^{-1}. \tag{4.17}$$

Hence the timescales are set by the inverse eigenvalues of $\frac{\Delta_t}{\psi_p} \boldsymbol{U}$.

4.4.1 Strategy and Key Ideas

The goal is to compute the spectrum of $\boldsymbol{U}$, whose inverse eigenvalues determine the training timescales. Define the resolvent (Stieltjes transform)

$$G_{\boldsymbol{U}}(z) = \frac{1}{p} \operatorname{Tr}\big((\boldsymbol{U} - z \mathbb{I}_p)^{-1} \big), \quad \rho(\lambda) = \lim_{\varepsilon \to 0^+} \frac{1}{\pi} \operatorname{Im} G_{\boldsymbol{U}}(\lambda + i\varepsilon).$$

Here $\boldsymbol{U} = \frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{F}\boldsymbol{F}^\top]$ with $\boldsymbol{F} = \sigma(\frac{\boldsymbol{W}\boldsymbol{X}}{\sqrt{d}})$; nonlinearity makes $\boldsymbol{U}$ untractable as it is. Following [9, 10]:

1. express spectral quantities via $G_{\boldsymbol{U}}$;
2. invoke a Gaussian Equivalence Principle (GEP) to replace $\sigma(\cdot)$-features by a Gaussian surrogate matching second moments (activation-dependent via Hermite coefficients);
3. compute $G_{\boldsymbol{U}}$ by the replica method in the large-dimension limit.

(Adapting the GEP to the DSM-averaged $\boldsymbol{U}$ requires care; see [9, 10].)

4.4.2 Gaussian Equivalence Principle

The nonlinearity makes $\boldsymbol{U}$ untractable as it is. Fortunately the following result holds. As explained in [30], the Gaussian Equivalence Theorem which applies in the high dimensional setting considered here establishes an equivalence regarding its spectral properties to a Gaussian covariate model where the nonlinear activation function is replaced by a linear term and a nonlinear term acting as noise:

$$\sigma\left(\frac{Wx}{\sqrt{d}}\right) \;\to\; \mu_1 \frac{Wx}{\sqrt{d}} + \mu^* \eta, \qquad \eta \sim \mathcal{N}(0, I_p), \tag{4.18}$$

where μ_1, μ^* are defined in Sect. C.1 for random variables x drawn from $P_t = \mathcal{N}(0, \Gamma_t^2 I_d)$.

This allows us to derive a Gaussian Equivalent (spectrum-wise) for $\boldsymbol{U}$:

Lemma 4.2 (Gaussian Equivalence Principle for U)**.** *In the limit $n, p, d \to \infty$ with $\psi_p = p/d$, $\psi_n = n/d$, the matrix*

$$U \;=\; \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\Big[\sigma\Big(\frac{W x_{\nu,t}(\xi)}{\sqrt{d}}\Big)\sigma\Big(\frac{W x_{\nu,t}(\xi)}{\sqrt{d}}\Big)^\top\Big] \tag{48}$$

has the same spectrum as its GEP equivalence

$$U \;=\; \frac{G}{\sqrt{n}}\frac{G^\top}{\sqrt{n}} + \Delta_t \mu_1^2 \frac{WW^\top}{d} + s_t^2 I_p, \tag{49}$$

where

$$G \;=\; e^{-t}\mu_1 \frac{W}{\sqrt{d}} X + v_t\,\Omega, \tag{50}$$

$X \in \mathbb{R}^{d\times n}$ is the matrix whose columns are the x_νs and $\Omega \in \mathbb{R}^{p\times d}$ has Gaussian entries independent of X and W.

The proof of the equivalence is provided in .3.

Exploiting this result we can approach the Stieltjes transform through the replica method.

Theorem 4.3 (Gaussian equivalence resolvent equations)**.** *Let*

$$q(z) = \frac{1}{p}\operatorname{Tr}\big((\boldsymbol{U} - z\mathbb{I}_p)^{-1}\big), \quad r(z) = \frac{1}{p}\operatorname{Tr}\big(\boldsymbol{W}^\top(\boldsymbol{U} - z\mathbb{I}_p)^{-1}\boldsymbol{W}\big), \quad z \in \mathbb{C}.$$

Then $q(z)$ and $r(z)$ satisfy

$$\frac{\psi_p v_t^2}{1 + e^{-2t}\mu_1^2(t)c_t^2\psi_p\psi_n r + \frac{\psi_p(s_t^2 - z)}{q}} + \psi_p(s_t^2 - z) + \frac{1-\psi_p}{q} - \frac{r}{q^2} = 0, \tag{4.19}$$

$$\frac{e^{-2t}\mu_1^2(t)c_t^2\sigma_v^2}{1 + e^{-2t}\mu_1^2(t)c_t^2\psi_p\psi_n r + \frac{\psi_p v_t^2}{q}} + \Delta_t\mu_1^2\psi_p + \frac{1}{q} - \frac{1}{r} = 0, \tag{4.20}$$

and the eigenvalue density is recovered by SokhotskiPlemelj:

$$\rho(\lambda) = \lim_{\epsilon\to 0^+}\frac{1}{\pi}\operatorname{Im}\, q(\lambda + i\epsilon).$$

Proof (replica). See Appendix .4.

In the asymptotics, we are actually able to provide a close expression for the spectral density, as illustrated in the following.

Theorem 4.4 (Spectral decomposition in two regimes)**.** *Let ρ be the spectral density of $\boldsymbol{U}$.*

- *If $\psi_p \gg \psi_n \gg 1$,*

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

- *If* $\psi_n \gg \psi_p \gg 1$,

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

Here ρ_1 *and* ρ_2 *are atomless with disjoint supports. Eigenvectors at* $\lambda = s_t^2$ *leave both train and test losses unchanged. The support of* ρ_1 *is*

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

while ρ_2 *is independent of* ψ_n *with support of order* ψ_p.

Two timescales and phase behavior. These results imply two training timescales: a fast generalization timescale $\tau_{\text{gen}} = O(1)$ for the initial relaxation (sets both train and test), and a slow memorization timescale

$$\tau_{\text{mem}} \ \sim\ \frac{\psi_p}{\Delta_t\, \lambda_{\min}},$$

with $\lambda_{\min}$ the left edge of ρ_1. In the overparameterized regime $p \gg n$, $\tau_{\text{mem}} = O(n)$. On that timescale, train decreases while test increases, and the gap $\mathcal{L}_{\text{gen}}$ grows, in agreement with [9]. As n increases, the asymptotic generalization loss decreases; for $n > n^*(p) = p$, the model is no longer expressive enough to fully memorize (*architectural regularization* phase). The generalization-memorization transition line depends on τ and moves upward for larger τ.

4.5 A Legit Doubt: Does RFNN Exhibit Actual Memorization?

We now test whether RFNNs *actually* memorize on finite data via the *U-turn experiment* of [9].

4.5.1 The U-turn Experiment

Initialize the backward process from a noisy training sample and check whether the trajectory turns back to it. Let a^μ be a training sample and

$$\boldsymbol{x}_t^\mu(\boldsymbol{\xi}) = e^{-t}\boldsymbol{a}^\mu + \sqrt{\Delta_t}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d), \quad \Delta_t = 1 - e^{-2t}. \tag{4.21}$$

Run the reverse SDE from $\boldsymbol{x}_{t_0}^\mu(\boldsymbol{\xi})$ using the learned score:

$$d\boldsymbol{Y}_t = \big(\boldsymbol{Y}_t + 2\,\boldsymbol{S}_{\boldsymbol{A}}^t(\boldsymbol{Y}_t)\big)\,dt + \sqrt{2}\,d\tilde{\boldsymbol{B}}_t, \qquad \boldsymbol{Y}_{t_0} = \boldsymbol{x}_{t_0}^\mu(\boldsymbol{\xi}). \tag{4.22}$$

Declare retrieval if

$$\frac{\|\boldsymbol{Y}_0^{(k)} - \mathrm{NN}_1(\boldsymbol{Y}_0^{(k)})\|}{\|\boldsymbol{Y}_0^{(k)} - \mathrm{NN}_2(\boldsymbol{Y}_0^{(k)})\|} < \delta, \tag{4.23}$$

with $\mathrm{NN}_1, \mathrm{NN}_2$ the nearest and second-nearest training neighbors and $\delta \simeq 1/3$. We measure memorization as the fraction of N runs that retrieve a training sample.

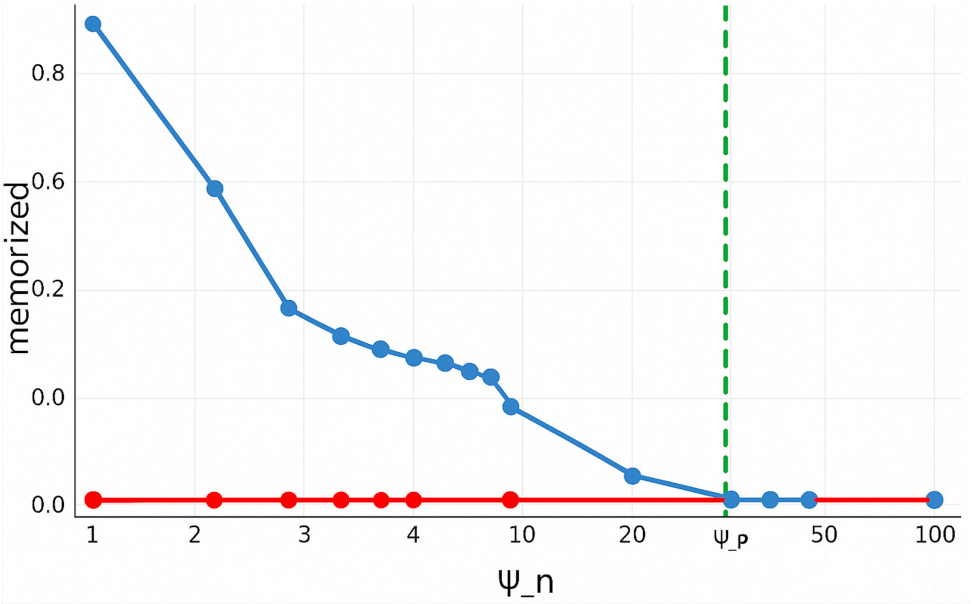


Figure 3: Fraction of memorized samples. Blue line for the learned score, red line for the true score. The vertical line is at $\psi_n = \psi_p$.

4.5.2 Observed Behaviour

Training the RFNN to up to τ_{mem} in $n/p < 1$ (where memorizing S_t^{emp} is feasible), U-turn trajectories started at $\boldsymbol{x}_t^\mu(\boldsymbol{\xi})$ return towards a^μ, confirming local reproduction of S_t^{emp} (Fig. 3). Clearly one could argue that the experiment is in some sense biased: maybe, starting the reverse of the backward process at small times in the vicinity of a data point should fallback to the data itself. That's why we also plot the memorized fraction obtained by integrating the reverse dynamics using the true score: as we see in the figure, no sample can be considered as by-product of memorization. It is also interesting to notice that memorization is zero when $\psi_n > \psi_p$, highlighting once again how under-parametrized regimes favor generalization.

However, starting instead from $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$ (standard generative mode), no memorization is detected: trajectories do not converge to training samples; the memorized fraction drop to zero for every ψ_n. If the learned score were exactly S_t^{emp} *globally*, memorization would appear in both setups. The fact that only the U-turn protocol memorizes shows that RFNNs do not realize S_t^{emp} globally.

We emphatize that the sampling procedure fails to memorize (which is actually a benign failure) for two main reasons: (i) when U-turn is reversed at large t, hence a temporal failure, and (ii) if we sample the initial condition from P_t , without biasing it toward an evolved data point.

4.5.3 Technical Interpretation

Local vs. global fidelity. For small t, each component of P_t^{emp} is a narrow Gaussian at $e^{-t}a^\mu$ and

$$S_t^{\text{emp}}(x) \approx -\frac{x - e^{-t}a^\mu}{\Delta_t}$$

near that mode. RFNNs can reproduce these wells *locally*, explaining U-turns when starting in-basin. But globally representing the entire mixture requires extreme curvature and rapid gradients between modes. Finite capacity, stochastic optimization, and early stopping yield a field that is locally empirical but *globally smooth*, closer to $S_t(x) = -x/\Gamma_t^2$ at larger t, producing regular flows that never re-enter narrow wells.

Hybrid score picture. The learned field behaves like a hybrid: *locally empirical, globally regularized*. Starting near a mode reveals the first (U-turn), while starting from the far Gaussian region reveals the second (smooth sampling). This matches the decomposition $\mathcal{L}_{\text{train}} = \mathcal{V}_t + \Delta_t \mathcal{M}_t$: minimizing $\mathcal{M}_t$ enforces local empirical fidelity, while finite Δ_t and implicit regularization keep $\mathcal{V}_t$ dominant at larger t, ensuring smooth behavior and preventing collapse.

Heuristically, the potential landscape forms shallow basins around training points, enough to trap trajectories that start inside, but too weak to attract those arriving from the Gaussian far field. RFNNs therefore reproduce S_t^{emp} *partially*, enabling generalization precisely because they *fail to memorize globally*.

5 Time-Integrated RFNN with Bias and Skip Connection

In the previous chapters, we explored the solvable framework of diffusion learning through a collection of independently trained Random Feature Neural Networks (RFNNs), each associated with a single diffusion time t. While this setting provides analytical tractability and a clear view of memorization and generalization mechanisms, modern diffusion models (such as those used in image or audio generation) employ a single time-conditioned network that jointly learns the score function across the entire diffusion trajectory.

This chapter bridges this gap by introducing a *time-dependent* RFNN model. We investigate how temporal conditioning, bias terms, and skip connections affect both the learning dynamics and the onset of memorization.

Model extension. We extend the RFNN by including:

1. a time-dependent modulation of the random features, and
2. a linear skip connection from the input $\boldsymbol{x}$ to the output.

The modified score model reads:

$$\boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}, t) = \alpha_t \frac{\boldsymbol{A}}{\sqrt{p}} \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} + t\boldsymbol{b}\right) + \beta_t \boldsymbol{x}. \tag{5.1}$$

We have therefore introduced:

1. a bias term $t\boldsymbol{b}$ with frozen $\boldsymbol{b} \in \mathbb{R}^p$, and
2. a skip connection combining the RFNN output and the raw input $\boldsymbol{x}$.

This construction raises three natural questions:

- How should time dependence be modeled?
- Why are skip connections useful, and how do they act in this context?
- Given the above, how should (α_t, β_t) be chosen?

5.1 Time-Proportional Bias

We choose not to rescale $\boldsymbol{b}$ with the data dimension d.

Heuristic argument. Consider the i-th pre-activation:

$$z_i = \left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)_i + tb_i.$$

Since $\boldsymbol{W}$ and $\boldsymbol{x}$ have i.i.d. zero-mean, unit-variance entries, the normalized dot product

$$\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)_i = \frac{1}{\sqrt{d}} \sum_{j=1}^{d} W_{ij} x_j$$

has variance $\mathcal{O}(1)$ by the central limit theorem. The additive term tb_i also remains $\mathcal{O}(1)$; hence, it is comparable in magnitude and not negligible. Therefore, we keep $\boldsymbol{b}$ unnormalized.

Alternative considerations. One might argue that time-dependence could arise even if $\boldsymbol{b}$ were scaled. For example, if learned output weights $\boldsymbol{A}_i$ align with $\boldsymbol{b}$ but remain orthogonal to the data span $\{\boldsymbol{x}_\nu\}_{\nu=1}^n$, then $\boldsymbol{A}_i^\top \sigma(\cdot)$ could depend primarily on t. However,

this would enforce an effectively linear time dependence and limit expressivity, preventing richer temporal representations. Hence, we adopt the unnormalized parametrization of $\boldsymbol{b}$.

5.1.1 Motivation and Interpretation

From a modeling viewpoint, adding a bias $t\boldsymbol{b}$ is the simplest way to encode time conditioning in a random-feature model. In realistic diffusion architectures, time enters through learned embeddings that modulate either the feature statistics (e.g., FiLM layers) or activation biases. The term $t\boldsymbol{b}$ reproduces the latter mechanism in its minimal analytical form: each hidden unit receives a time-dependent shift in its preactivation, thereby altering the mean and variance of activations along the diffusion trajectory. This allows the random-feature map to interpolate between two regimesone dominated by data at small t, and one dominated by noise at large t without adding any trainable parameters or nonlinearities in t.

5.2 Skip Connection

At small times, $P_t \simeq P_0$: geometrically, the forward trajectory of $\boldsymbol{x}^\nu$ has barely diffused, so $\boldsymbol{x}_t^\nu(\boldsymbol{\xi})$ remains near $\boldsymbol{x}^\nu$ in data space. Since $P_t(\boldsymbol{x}_t^\nu|\boldsymbol{x}^\nu)$ is a Gaussian centered on $\boldsymbol{x}^\nu$, the score (gradient of $\log P_t$) points back toward $\boldsymbol{x}^\nu$. Because P_t is effectively a sum of narrow peaks, moving closer to the training point causes $\log P_t$ to increase sharply. Thus, near $t \to 0$, the score field becomes large and sharply varying, difficult to learn accurately.

Learning perspective. We train by minimizing

$$\left\|\boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}_t^\nu(\boldsymbol{\xi}),t)+\tfrac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right\|^2=\left\|\alpha_t\tfrac{\boldsymbol{A}}{\sqrt{p}}\sigma(\tfrac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b})+\beta_t\boldsymbol{x}_t^\nu(\boldsymbol{\xi})+\tfrac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right\|^2.$$

This is equivalent to minimizing

$$\left\|\tfrac{\boldsymbol{A}}{\sqrt{p}}\sigma\Big(\tfrac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\Big)+\tfrac{\beta_t}{\alpha_t}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})+\tfrac{\boldsymbol{\xi}}{\alpha_t\sqrt{\Delta_t}}\right\|^2. \tag{5.2}$$

Substituting $\boldsymbol{x}_t^\nu(\boldsymbol{\xi}) = e^{-t}\boldsymbol{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}$, we get

$$\left\| \frac{\boldsymbol{A}}{\sqrt{p}} \sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b} \Big) + \frac{\beta_t}{\alpha_t} e^{-t}\boldsymbol{x}^\nu + \frac{(\beta_t\Delta_t + 1)}{\alpha_t\sqrt{\Delta_t}}\boldsymbol{\xi} \right\|^2. \tag{5.3}$$

At large t, $\Delta_t \simeq 1$ and $\boldsymbol{x}_t^\nu \approx \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$, so the empirical score approaches $S_t^{\text{emp}}(\boldsymbol{x}) \simeq -\boldsymbol{x}/\Delta_t$. Hence, we may tune (α_t, β_t) so that at large diffusion times, the model relies mainly on its skip connection to predict the linear score, while at small times, the skip connection compensates the divergence of the target.

Choice and Scaling of α_t and β_t

Roles. α_t controls the nonlinear random-feature contribution, while β_t controls the residual linear term. Their time dependence must satisfy the following:

Asymptotic requirements. At large t, the target score is linear, $S_t(x) \simeq -x$, so the model should emphasize the skip connection (linear term). At small t, the true score diverges as $1/\Delta_t$, and the skip must cancel that divergence.

$$\alpha_t \text{ large for small } t,\ \alpha_t \to 0 \text{ as } t \to \infty; \qquad \beta_t \text{ small for small } t,\ \beta_t \to 1 \text{ as } t \to \infty.$$

In particular,

$$\frac{\beta_t}{\alpha_t} \propto \frac{1}{\Delta_t},$$

ensuring that the skip term neutralizes the $1/\sqrt{\Delta_t}$ divergence, stabilizing the loss for small t and ensuring smoothness across time.

Interpretation. The pair (α_t, β_t) governs the effective kernel

$$U = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\xi,t}\Big[\alpha_t^2 \Delta_t\, \sigma\left(\frac{W x_t^\nu(\xi)}{\sqrt{d}} + tb \right) \sigma\left(\frac{W x_t^\nu(\xi)}{\sqrt{d}} + tb \right)^\top \Big],$$

and therefore influences both spectral properties and learning timescales.

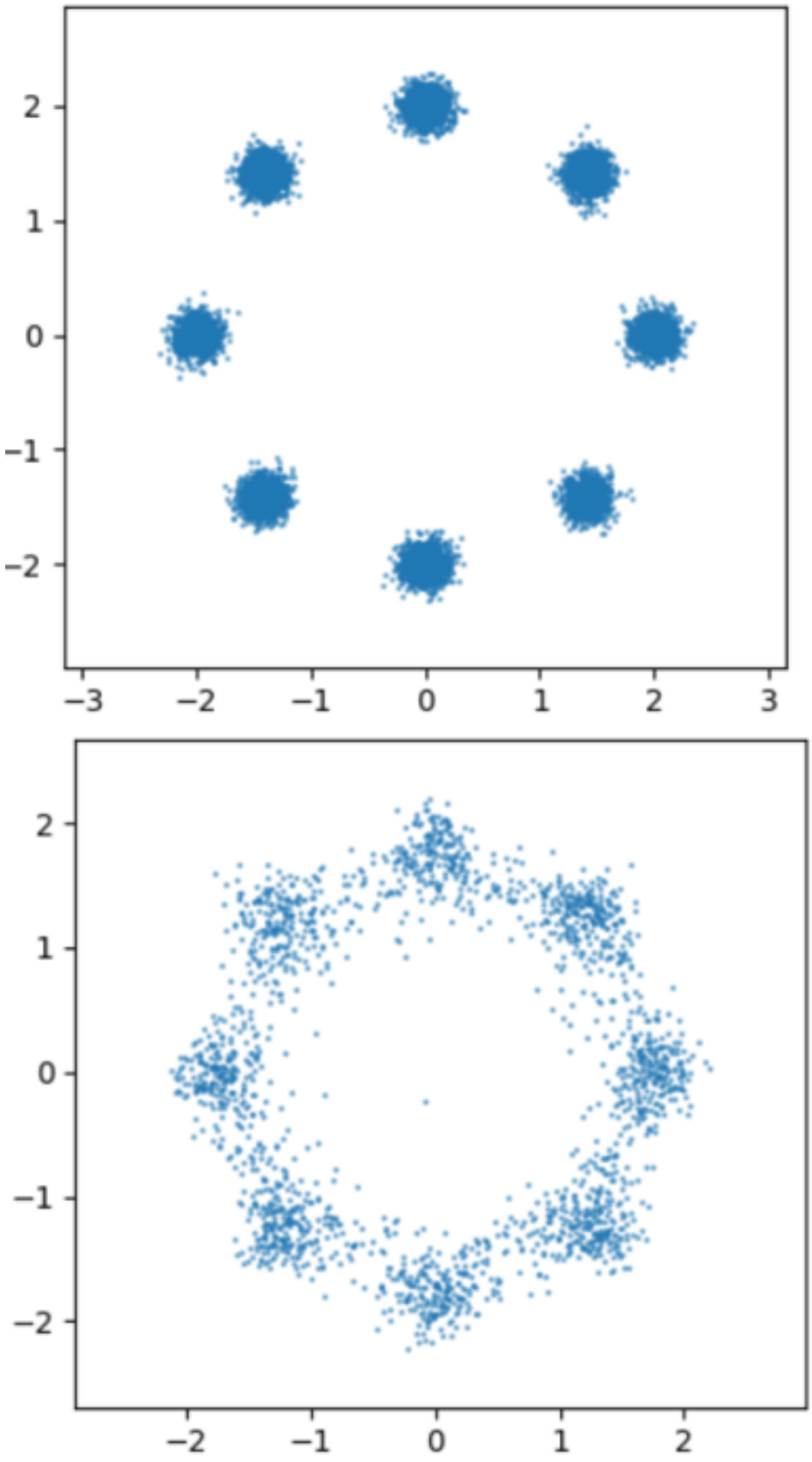


Figure 4: 2d data generation using the time integrated model.

Practical schedules. Common parameterizations include:

Name	α_t	β_t
no_skip	1	0
const	1	1
neutralising	$\frac{1}{\sqrt{\Delta_t}}$	$\frac{1}{\sqrt{\Delta_t}}$
neutralising±	$\frac{1}{\sqrt{\Delta_t}}$	$-\frac{1}{\sqrt{\Delta_t}}$
vanish_neutralising	$\frac{e^{-t}}{\sqrt{\Delta_t}}$	$-\frac{1}{\sqrt{\Delta_t}}$
vanish_neutralising++	$\frac{e^{-t}}{\sqrt{\Delta_t}}$	$\frac{1}{\sqrt{\Delta_t}}$

Among these, the *vanish_neutralising* families best satisfy small-t and large-t constraints: they neutralize the divergence at early times while suppressing nonlinear terms exponentially as $t \to \infty$, allowing the skip connection to recover the Gaussian limit $S_t(x) \simeq -x$.

Fig. 4 shows the generative performance of the model on a toy 2d dataset, using the vanishing netutralising plus schedule.

5.3 Training Dynamics

The training loss is defined as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{A}, \{\boldsymbol{x}^\nu\}) &= \frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{d}\,\mathbb{E}_{\boldsymbol{\xi},t}\left\|\sqrt{\Delta_t}\boldsymbol{S}_{\boldsymbol{A}}(\boldsymbol{x}_t^\nu(\boldsymbol{\xi}),t)+\boldsymbol{\xi}\right\|^2 \\ &= \frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{d}\,\mathbb{E}_{\boldsymbol{\xi},t}\left\|\sqrt{\Delta_t}\alpha_t\frac{\boldsymbol{A}}{\sqrt{p}}\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\right)+\sqrt{\Delta_t}\beta_t\boldsymbol{x}_t^\nu(\boldsymbol{\xi})+\boldsymbol{\xi}\right\|^2,\end{aligned} \tag{5.4}$$

with $\Delta_t = 1 - e^{-2t}$.

5.3.1 Gradient Flow

Inserting $\boldsymbol{x}_t^\nu(\boldsymbol{\xi}) = e^{-t}\boldsymbol{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}$, the gradient with respect to $\boldsymbol{A}$ reads:

$$\nabla_{\boldsymbol{A}}\mathcal{L}(\boldsymbol{A}) = 2\frac{1}{dp}\boldsymbol{A}\boldsymbol{U} + 2\frac{1}{d\sqrt{p}}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top), \tag{5.5}$$

where

$$\boldsymbol{U} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi},t}\left[\alpha_t^2 \Delta_t \sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)^\top\right], \tag{5.6}$$

$$\boldsymbol{V}_1 = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi},t}\left[\alpha_t \beta_t \Delta_t e^{-t} \sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\right]\boldsymbol{x}^{\nu\top}, \tag{5.7}$$

$$\boldsymbol{V}_2 = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi},t}\left[\alpha_t \sqrt{\Delta_t}(1+\beta_t\Delta_t) \sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\boldsymbol{\xi}^\top\right]. \tag{5.8}$$

Here $\boldsymbol{U} \in \mathbb{R}^{p\times p}$, $\boldsymbol{V}_{1,2} \in \mathbb{R}^{p\times d}$.

Continuous limit. Analogously to Sec. 5.3.1,

$$\dot{\boldsymbol{A}} = -2\frac{\boldsymbol{A}\boldsymbol{U}}{\psi_p} - 2\frac{d}{\sqrt{p}}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top). \tag{5.9}$$

This ODE integrates to

$$\boldsymbol{A}(\tau) = \Big[\boldsymbol{A}_0 + \sqrt{p}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}\Big]e^{-\frac{2\tau}{\psi_p}\boldsymbol{U}} - \sqrt{p}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}. \tag{5.10}$$

The stationary optimum is

$$\hat{\boldsymbol{A}} = \lim_{\tau\to\infty} \boldsymbol{A}(\tau) = -\sqrt{p}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}.$$

5.3.2 Discrete Gradient Descent

We can also derive the discrete, full-batch gradient descent evolution of the parameters $\boldsymbol{A}$. Starting from:

$$\nabla_{\boldsymbol{A}}\mathcal{L}(\boldsymbol{A}) = \boldsymbol{A}\,\tilde{\boldsymbol{U}} + \tilde{\boldsymbol{V}}^\top, \tag{5.11}$$

where

$$\tilde{\boldsymbol{U}} = \frac{2}{dp}\,\boldsymbol{U}, \qquad \tilde{\boldsymbol{V}} = \frac{2}{d\sqrt{p}}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top),$$

the gradient update reads

$$\begin{aligned}\boldsymbol{A}^{(k)} &= \boldsymbol{A}^{(k-1)} - \eta \nabla_{\boldsymbol{A}} \mathcal{L}(\boldsymbol{A}^{(k-1)}) \\ &= \boldsymbol{A}^{(k-1)} - \eta(\boldsymbol{A}^{(k-1)}\tilde{\boldsymbol{U}} + \tilde{\boldsymbol{V}}^\top) \\ &= \boldsymbol{A}^{(k-1)}(\mathbb{I}_p - \eta\tilde{\boldsymbol{U}}) - \eta\tilde{\boldsymbol{V}}^\top. \end{aligned} \tag{5.12}$$

Iterating this recurrence telescopically, we obtain

$$\begin{aligned}\boldsymbol{A}^{(k)} &= \boldsymbol{A}^{(0)}(\mathbb{I}_p - \eta\tilde{\boldsymbol{U}})^k - \eta\tilde{\boldsymbol{V}}^\top \sum_{l=0}^{k-1} (\mathbb{I}_p - \eta\tilde{\boldsymbol{U}})^l \\ &= \left[\boldsymbol{A}^{(0)} + \tilde{\boldsymbol{V}}^\top\tilde{\boldsymbol{U}}^{-1}\right](\mathbb{I}_p - \eta\tilde{\boldsymbol{U}})^k - \tilde{\boldsymbol{V}}^\top\tilde{\boldsymbol{U}}^{-1}. \end{aligned} \tag{5.13}$$

Restoring $\boldsymbol{U}$ and $\boldsymbol{V}_{1,2}$ gives:

$$\boldsymbol{A}^{(k)} = \left[\boldsymbol{A}^{(0)} + \sqrt{p}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}\right]\left(\mathbb{I}_p - \eta\tfrac{2}{dp}\boldsymbol{U}\right)^k - \sqrt{p}(\boldsymbol{V}_1^\top + \boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}. \tag{5.14}$$

This expression has the same form as the continuous-time solution, differing only by the discrete decay factor $(\mathbb{I}_p - \eta\frac{2}{dp}\boldsymbol{U})^k$, which becomes an exponential in the limit of small η:

$$\left(\mathbb{I}_p - \eta\tfrac{2}{dp}\boldsymbol{U}\right)^{\frac{d^2\tau}{\eta}} \simeq e^{-\frac{2\tau}{\psi_p}\boldsymbol{U}}.$$

Hence the discrete dynamics converges to the same stationary solution as the continuous gradient flow.

5.4 Metrics: Explicit Derivations

Losses. We now derive explicit analytical expressions for the training and test losses, showing how they depend on $\boldsymbol{U}$, $\boldsymbol{V}_1$, and $\boldsymbol{V}_2$. Once the parameters $\boldsymbol{A}$ are determined (either from the analytical solution of the flow or by discrete optimization), they can be substituted directly into these expressions.

$$\mathcal{L}_{\text{train}} = \frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{d}\mathbb{E}_{\boldsymbol{\xi},t}\Big\|\sqrt{\Delta_t}\alpha_t\frac{\boldsymbol{A}}{\sqrt{p}}\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^{\nu}(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\Big)+\sqrt{\Delta_t}\beta_t e^{-t}\boldsymbol{x}^{\nu}+(1+\beta_t\Delta_t)\boldsymbol{\xi}\Big\|^2$$
$$= \frac{1}{d}\mathrm{Tr}\Big(\tfrac{\boldsymbol{A}^\top}{\sqrt{p}}\tfrac{\boldsymbol{A}}{\sqrt{p}}\boldsymbol{U}\Big)+\frac{2}{d}\mathrm{Tr}\Big(\tfrac{\boldsymbol{A}}{\sqrt{p}}(\boldsymbol{V}_1+\boldsymbol{V}_2)\Big)+\mathbb{E}_t\Big[\Delta_t\beta_t\big(\beta_t\Gamma_t^2+2\big)\Big]+1. \tag{5.15}$$

At $\tau\to\infty$, using $\hat{\boldsymbol{A}}=-\sqrt{p}(\boldsymbol{V}_1^\top+\boldsymbol{V}_2^\top)\boldsymbol{U}^{-1}$, we get:

$$\mathcal{L}_{\text{train}}^{\infty} = 1-\frac{1}{d}\mathrm{Tr}\Big[\boldsymbol{U}^{-1}(\boldsymbol{V}_1+\boldsymbol{V}_2)(\boldsymbol{V}_1^\top+\boldsymbol{V}_2^\top)\Big]+\mathbb{E}_t\Big[\Delta_t\beta_t(\beta_t\Gamma_t^2+2)\Big]. \tag{5.16}$$

Interpretation. The losses depend on the correlations between random features ($\boldsymbol{U}$) and the coupling matrices $\boldsymbol{V}_{1,2}$, which encode data structure and skip effects. Test losses differ only in the estimation of these matrices on an independent test set.

Distance from the true score. The true score is $\boldsymbol{S}(\boldsymbol{x},t)=-\boldsymbol{x}/\Gamma_t^2$, so the score error reads

$$\varepsilon_{\text{score}} = \frac{1}{d}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big\|\alpha_t\tfrac{\boldsymbol{A}}{\sqrt{p}}\sigma\Big(\tfrac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\Big)+\Big(\beta_t+\tfrac{1}{\Gamma_t^2}\Big)\boldsymbol{x}_t(\boldsymbol{\xi})\Big\|^2. \tag{5.17}$$

Expanding $\boldsymbol{x}_t(\boldsymbol{\xi})=e^{-t}\boldsymbol{x}+\sqrt{\Delta_t}\boldsymbol{\xi}$ gives:

$$\varepsilon_{\text{score}} = \frac{1}{d}\mathrm{Tr}\Big(\tfrac{\boldsymbol{A}^\top}{\sqrt{p}}\tfrac{\boldsymbol{A}}{\sqrt{p}}\tilde{\boldsymbol{U}}\Big)+\frac{2}{d}\mathrm{Tr}\Big(\tfrac{\boldsymbol{A}}{\sqrt{p}}(\tilde{\boldsymbol{V}}_3+\tilde{\boldsymbol{V}}_4)\Big)+\mathbb{E}_t\Big[\big(\beta_t+\tfrac{1}{\Gamma_t^2}\big)^2\Gamma_t^2\Big], \tag{5.18}$$

where

$$\tilde{\boldsymbol{V}}_3 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big[\big(\beta_t+\tfrac{1}{\Gamma_t^2}\big)e^{-t}\alpha_t\sigma\Big(\tfrac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\Big)\boldsymbol{x}^\top\Big], \tag{5.19}$$

$$\tilde{\boldsymbol{V}}_4 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big[\big(\beta_t+\tfrac{1}{\Gamma_t^2}\big)\sqrt{\Delta_t}\alpha_t\sigma\Big(\tfrac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}}+t\boldsymbol{b}\Big)\boldsymbol{\xi}^\top\Big]. \tag{5.20}$$

Distance from the empirical score. Recalling

$$\boldsymbol{S}^{\text{emp}}(\boldsymbol{x}_t,t) = -\frac{\boldsymbol{x}_t}{\Delta_t}+\frac{\sum_\nu \boldsymbol{x}_\nu e^{-t}\exp\big(-\frac{\|\boldsymbol{x}_t-\boldsymbol{x}_\nu e^{-t}\|^2}{2\Delta_t}\big)}{\Delta_t\sum_\nu\exp\big(-\frac{\|\boldsymbol{x}_t-\boldsymbol{x}_\nu e^{-t}\|^2}{2\Delta_t}\big)},$$

we find

$$\varepsilon_{\text{score}}^{\text{emp}} = \frac{1}{d}\text{Tr}\Big(\frac{\boldsymbol{A}^\top}{\sqrt{p}}\frac{\boldsymbol{A}}{\sqrt{p}}\tilde{\boldsymbol{U}}\Big) + \frac{2}{d}\text{Tr}\Big(\frac{\boldsymbol{A}}{\sqrt{p}}(\tilde{\boldsymbol{V}}_5 + \tilde{\boldsymbol{V}}_6 + \boldsymbol{V}^{\text{emp}})\Big) + \mathcal{O}_\tau(1), \tag{5.21}$$

where

$$\boldsymbol{V}^{\text{emp}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big[\frac{e^{-t}\alpha_t \exp\big(-\frac{\|\boldsymbol{x}_t(\boldsymbol{\xi})-\boldsymbol{x}_\nu e^{-t}\|^2}{2\Delta_t}\big)}{\Delta_t \sum_\mu \exp\big(-\frac{\|\boldsymbol{x}_t(\boldsymbol{\xi})-\boldsymbol{x}_\mu e^{-t}\|^2}{2\Delta_t}\big)}\sum_\nu \sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\boldsymbol{x}_\nu^\top\Big], \tag{5.22}$$

$$\tilde{\boldsymbol{V}}_5 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big[\Big(\beta_t + \tfrac{1}{\Delta_t}\Big)e^{-t}\alpha_t\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\boldsymbol{x}^\top\Big], \tag{5.23}$$

$$\tilde{\boldsymbol{V}}_6 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi},t}\Big[\Big(\beta_t + \tfrac{1}{\Delta_t}\Big)\sqrt{\Delta_t}\alpha_t\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t(\boldsymbol{\xi})}{\sqrt{d}} + t\boldsymbol{b}\Big)\boldsymbol{\xi}^\top\Big]. \tag{5.24}$$

Monitoring $\varepsilon_{\text{score}}^{\text{emp}}(\tau)$ thus provides a measure of how close the learned score is to the empirical optimum along training.

5.5 Breakdown of Gaussian Equivalence under Time Integration

In the previous chapter, the Gaussian Equivalence Principle (GEP) [7, 8, 9, 13] was central to the analytical treatment of the RFNN: linearizing the nonlinear random feature map via a Gaussian surrogate enabled spectral computation for

$$\boldsymbol{U}(t) = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{E}_{\boldsymbol{\xi}}\Big[\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}\Big)\sigma\Big(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}}\Big)^\top\Big],$$

and hence determination of the spectral density $\rho(\lambda)$ from its Stieltjes transform.

Conditions for validity. The GEP relies on:

1. independence between $\boldsymbol{W}$ and $\boldsymbol{x}$, and
2. stationarity of feature statistics, ensuring that preactivations entering $\sigma(\cdot)$ are Gaussian with fixed mean and variance in the large-d limit.

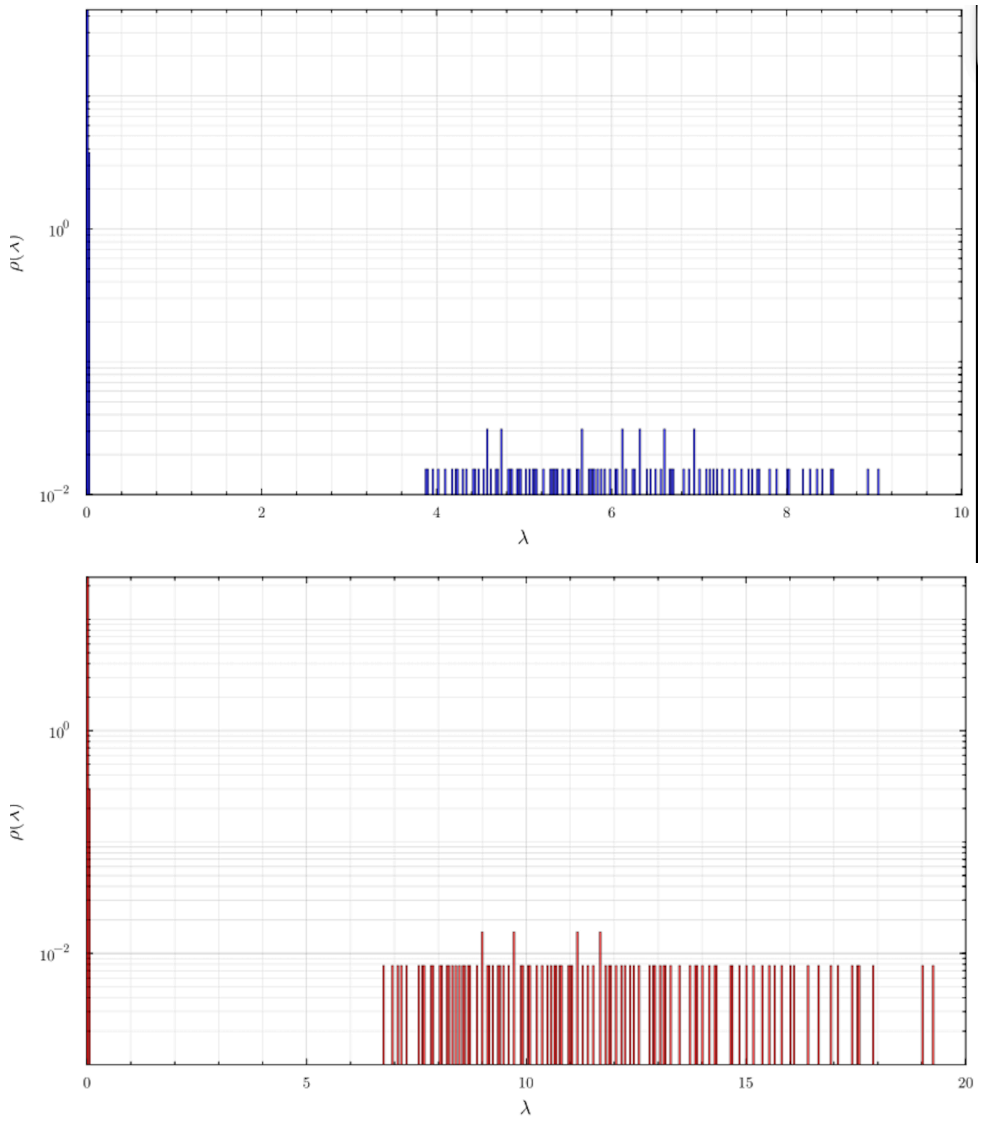


Figure 5: Generalization bulk. Blue histogram for U, red histogram for the integrated GEP.

5.5.1 Time Integration and the Loss of Stationarity

Let us simplify to the static RFNN (without bias or skip), but with time-integrated training:

$$\overline{\boldsymbol{U}} = \mathbb{E}_{t\sim\mathcal{U}[0,T]}[\boldsymbol{U}(t)].$$

Each $\boldsymbol{U}(t)$ separately satisfies the GEP, admitting a Gaussian equivalent ensemble with spectrum $\rho_{\text{GEP}}(\lambda;t)$. However, the time-averaged kernel $\overline{\boldsymbol{U}}$ does not correspond to the covariance of a single Gaussian ensemble. The preactivations

$$z_i(t) = \frac{(\boldsymbol{W}\boldsymbol{x}_t)_i}{\sqrt{d}}$$

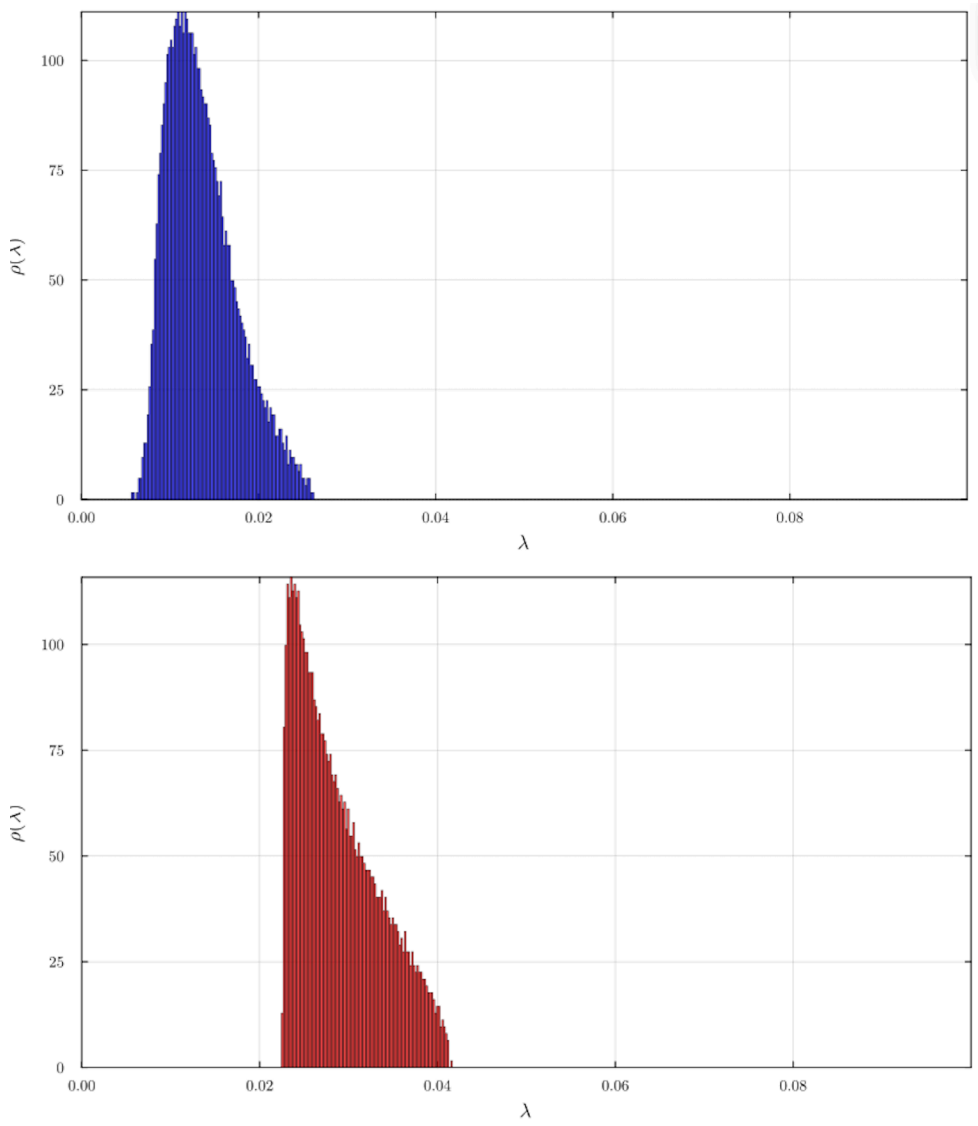


Figure 6: Memorization bulk. Blue histogram for U, red histogram for the integrated GEP.

are Gaussian at fixed t, but form a mixture once integrated over t. This mixture breaks the core GEP assumption of stationary Gaussian statistics.

Hence, even without explicit architectural time dependence, the equivalence fails:

$$\rho_{\text{direct}}(\lambda) \neq \mathbb{E}_t[\rho_{\text{GEP}}(\lambda; t)].$$

This non-commutativity between the GEP and temporal averaging reflects that the GEP only holds for fixed feature distributions, whereas here the features evolve with t. Integrating over time couples non-Gaussian ensembles with distinct covariances, invalidating

the single-Gaussian surrogate approximation.

Implications. The breakdown of GEP under time integration renders the spectrum of $\overline{\boldsymbol{U}}$ analytically intractable. Thus, the random-matrix tools used in Chapter 4 no longer apply, and we must resort to numerical spectral analysis to study learning dynamics in the time-integrated RFNN.

5.6 Numerical Spectral Analysis

Once the GEP breaks down, the spectral properties of $\overline{\boldsymbol{U}}$ must be accessed numerically. We now explore these spectra to identify qualitative phenomenology in the time-dependent RFNN.

5.6.1 Parameter Dependence

The spectrum of $\overline{\boldsymbol{U}}$ depends on many control parameters: the architectural ratios ψ_n, ψ_p, the final diffusion time T, and the time schedules (α_t, β_t). Each affects the scale and structure of the effective feature covariance, precluding a universal spectral law. Our numerical exploration therefore focuses on robust qualitative trends rather than exhaustive mapping.

5.6.2 Emergent Phenomenology

Numerically, the eigenvalue distribution of $\overline{\boldsymbol{U}}$ consistently exhibits two distinct bulks, reminiscent of the spectral separation analytically derived in Chapter 4 for fixed t. With reasonable (α_t, β_t) choices, this bimodal structure persists even for wide integration intervals.

The relative positions and scales of the bulks, however, depend strongly on (α_t, β_t) and T. This reflects how temporal rescaling of nonlinear and linear components modulates the effective signal-to-noise ratio of features along diffusion.

Although the lack of a closed-form theory prevents quantitative predictions, the persistence of the two-bulk structure indicates that the fundamental mechanisms identified

previouslycoexisting generalization and memorization modesremain active even in the time-integrated regime. The scaling of bulks with (α_t, β_t) further suggests that temporal weighting of nonlinear and skip paths controls the balance between these two modes.

In the following, we visualize representative spectral densities obtained for different parameter regimes, highlighting how architectural and temporal choices shape the learning spectrum.

6 Gaussian Mixture Data

In this chapter, we return to the analytically solvable framework of Chapter 4, in which an independent Random Feature Neural Network (RFNN) is trained at each diffusion time t. Our goal is to investigate how the learning dynamics and spectral properties are modified when the data distribution departs from a single isotropic Gaussian and instead contains a latent low-rank structure.

Specifically, we consider data drawn from a symmetric mixture of two Gaussians with opposite means:

$$p(\boldsymbol{x}^\mu) = \frac{1}{2}\mathcal{N}(\boldsymbol{m}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{m}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d), \tag{6.1}$$

or equivalently,

$$\boldsymbol{x}^\mu = c^\mu \boldsymbol{m} + \sigma_{\boldsymbol{x}} \boldsymbol{z}^\mu = \sigma_{\boldsymbol{x}} \left(\frac{c^\mu \boldsymbol{m}}{\sigma_{\boldsymbol{x}}} + \boldsymbol{z}^\mu \right), \qquad c^\mu \in \{-1, 1\}, \quad \boldsymbol{z}^\mu \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d). \tag{6.2}$$

6.1 Setting and Goals

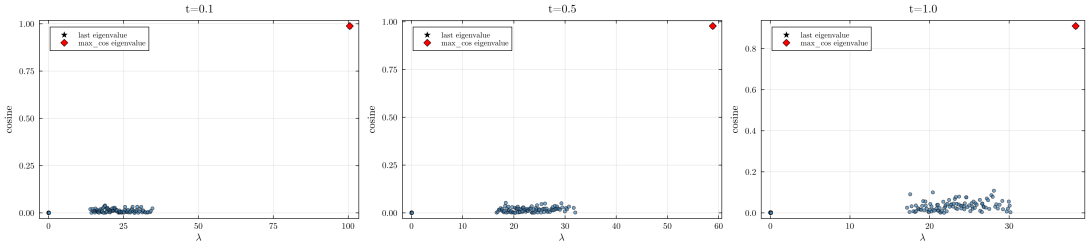
The vector $\boldsymbol{m}$ introduces a preferred direction in data space, breaking the rotational symmetry of the isotropic Gaussian case. Its norm $\|\boldsymbol{m}\|$ controls the separation between the mixture components:

- For $\|\boldsymbol{m}\|^2 \ll d\sigma_{\boldsymbol{x}}^2$, the two Gaussians strongly overlap;
- For $\|\boldsymbol{m}\|^2 \gg d\sigma_{\boldsymbol{x}}^2$, they are well separated, forming two nearly disjoint clusters.

We focus on the intermediate, weak-signal regime, where the components overlap but retain a detectable mean difference.

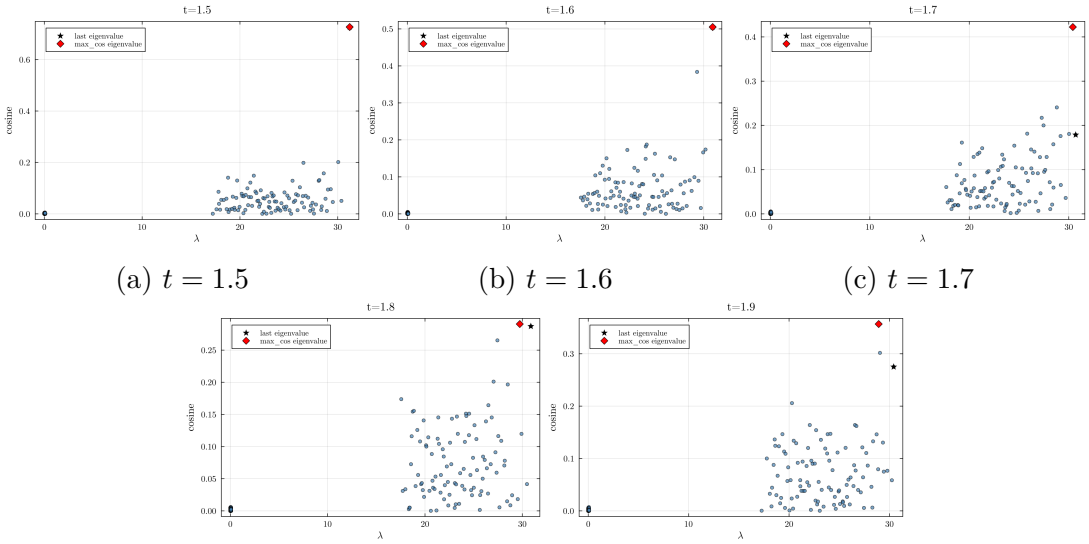
Our central question is: *Do the mechanisms underlying the generalization–memorization transition identified in Chapter 4 persist when the data contain low-rank structure?* If so, how does this structure couple to diffusion time t?

6.2 Emergence of a BBP Transition



(a) $t = 0.1$ (b) $t = 0.5$ (c) $t = 1.0$

Figure 7: At small diffusion times t, the signal is clearly visible: the outlier eigenvalue is well separated from the generalization bulk and its eigenvector aligns almost perfectly with the signal direction.



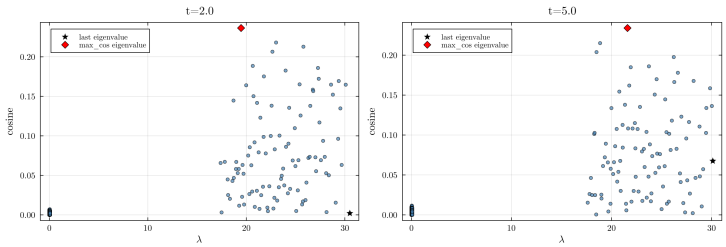
(a) $t = 1.5$ (b) $t = 1.6$ (c) $t = 1.7$

(d) $t = 1.8$ (e) $t = 1.9$

Figure 8: The BBP transition: as t increases, the outlier approaches the bulk and its alignment with the signal direction decreases continuously.

Spectral structure. For each diffusion time t, we compute the empirical feature covariance:

$$U(t) = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}_t^{\nu}(\boldsymbol{\xi})}{\sqrt{d}}\right)\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}_t^{\nu}(\boldsymbol{\xi})}{\sqrt{d}}\right)^{\top}\right], \tag{6.3}$$



(a) $t = 2$ (b) $t = 5$

Figure 9: At large diffusion times, the signal is lost and the spectrum becomes isotropic: no outlier eigenvalue remains.

and analyze its eigenvalue spectrum.

As in the single-Gaussian case, two main spectral bulks are present, corresponding respectively to the *generalization* and *memorization* modes of the model. The Gaussian mixture, however, introduces a rank-one perturbation aligned with the projection of $\boldsymbol{m}$ into feature space. This perturbation leads to a Baik–Ben Arous–Péché (BBP) transition in the largest eigenvalue of $U(t)$.

Phenomenology across diffusion times. At large t, the diffusion dynamics has effectively erased the signal: the data distribution is nearly isotropic, and the eigenvalue spectrum exhibits two clean bulks with no discernible outlier. (Fig. 9). As t decreases, the diffusion noise weakens and the latent mean difference $\boldsymbol{m}$ gradually emerges as a detectable direction. An outlier eigenvalue separates from the generalization bulk, and its associated eigenvector starts to align with the $\boldsymbol{m}$ features projection. (Fig. 8). This alignment increases as t decreases, reaching nearly perfect correspondence in the low-t regime (Fig. 7).

This phenomenon is usually referred to as BBP transition: the top eigenvalue detaches from the bulk once the effective signal-to-noise ratio exceeds a critical threshold. Here, t acts as the control parameter governing the onset of detectability. Thus, the BBP transition in t serves as a clear *spectral marker* for the emergence of signal sensitivity in the feature space.

Alignment structure and zero-sum property. An interesting feature of this spectrum is that, regardless of whether the system is below or above the BBP threshold, the

memorization eigenmodes always exhibit zero alignment with the signal. The signal is carried exclusively by the generalization modes. When an outlier is present, it alone captures the alignment with the signal, while the remaining generalization modes contribute negligibly. Conversely, when no outlier is present (above the BBP threshold in t), all generalization eigenvectors share the alignment equally, each having a small but finite correlation with the signal.

Numerically, we find that the total alignmentdefined as the sum of the cosine similarities between all eigenvectors and the signal directionis conserved and equals one:

$$\sum_{i=1}^{p} \langle \boldsymbol{u}_i, Signal \rangle^2 = 1. \tag{6.4}$$

This "zero-sum" interplay indicates a redistribution of alignment between generalization modes depending on whether the signal is detectable. The signal-bearing outlier effectively "absorbs" all the correlation when present, while in its absence, the weak signal is evenly spread across the generalization bulk.

Dependence on signal strength and sample ratio. Although the present work focuses on the dependence on diffusion time t, the same BBP phenomenology can be explored as a function of signal strength $\|\boldsymbol{m}\|$ and sample ratio $\alpha = p/n$. In both cases, the transition follows the same universal pattern: below a critical value of $\|\boldsymbol{m}\|$ or α, the spectrum is isotropic and uninformative; above it, a rank-one outlier appears and the corresponding eigenvector recovers the latent structure.

In the numerical examples shown here, the signal vector $\boldsymbol{m}$ has norm of order unity, $\|\boldsymbol{m}\| = \mathcal{O}(1)$. Other scaling regimessuch as $\|\boldsymbol{m}\| = \mathcal{O}(\sqrt{d})$ will be investigated in future work.

In summary, the Gaussian mixture introduces a controlled, low-rank perturbation in the data covariance that interacts non-trivially with diffusion dynamics. The resulting BBP transition in diffusion time t demarcates the onset of structure detection at the feature level, without altering the overall two-bulk spectral topology of the generalization–memorization landscape.

6.3 Gaussian Equivalence for Gaussian Mixture

To analyze the spectrum of $U(t)$, we extend the Gaussian Equivalence Principle (GEP) to inputs drawn from the Gaussian mixture defined above. Recall that in the isotropic Gaussian case, the nonlinearity $\boldsymbol{F} = \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)$ can be replaced by an equivalent Gaussian field with the same first two moments, due to the rotational invariance of $\boldsymbol{x}$. Here, the mixture structure breaks this invariance and introduces a weak rank-one perturbation aligned with the mean direction $\boldsymbol{m}$.

We parameterize the data as

$$\boldsymbol{x}^\nu = c^\nu \boldsymbol{m} + \sigma_{\boldsymbol{x}} \boldsymbol{u}^\nu, \qquad c^\nu \in \{-1, +1\}, \quad \boldsymbol{u}^\nu \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d), \tag{6.5}$$

and, without loss of generality, fix $\boldsymbol{m} = m\,\boldsymbol{e}_1$. This yields

$$\boldsymbol{x}^\nu = \sigma_{\boldsymbol{x}} \left(\frac{m}{\sigma_{\boldsymbol{x}}} c^\nu \boldsymbol{e}_1 + \boldsymbol{u}^\nu \right). \tag{6.6}$$

Let's assume that the activation function σ admits a Hermite polynomial expansion $\sigma(x) = \sum_{s=0}^{\infty} \alpha_s \,\mathrm{He}_s(x)$ and is odd, hence $\mu_0 = \mathbb{E}_z[\sigma(z)] = 0$. The feature vector $\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)$ can be replaced by a Gaussian surrogate with matched statistics:

$$\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}^\nu}{\sqrt{d}}\right) \;\xrightarrow{d}\; \mu_1 \frac{m}{\sigma_{\boldsymbol{x}}} c^\nu \frac{\boldsymbol{W}\boldsymbol{e}_1}{\sqrt{d}} + \mu_1 \frac{\boldsymbol{W}\boldsymbol{u}^\nu}{\sqrt{d}} + \mu_* \boldsymbol{\eta}^\nu, \tag{6.7}$$

where $\boldsymbol{\eta}^\nu \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_p)$, and μ_1, μ_* are scalar coefficients determined by the activation function and normalization of the pre-activations. The first term captures the coherent alignment with $\boldsymbol{m}$, the second corresponds to random isotropic fluctuations, and the last adds an independent Gaussian residual ensuring the correct variance.

Recall that we are interested in the spectrum of

$$U(t) = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}} \left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} \right) \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}} \right)^\top \right]. \tag{6.8}$$

Similarly to the isotropic Gaussian case, we can exploit the GEP for the features map in

order to derive a spectral equivalent for U(t). We indeed obtain:

$$U \;=\; \frac{G}{\sqrt{n}}\frac{G^\top}{\sqrt{n}} + \Delta_t \mu_1^2 \frac{WW^\top}{d} + s_t^2 I_p, \tag{6.9}$$

with

$$\left(\frac{1}{n}\boldsymbol{G}\boldsymbol{G}^\top\right)_{ij} = \frac{1}{n}\sum_\nu \left(e^{-t}\mu_1 \frac{m}{\sigma_{\boldsymbol{x}}} c^\nu \frac{w_i}{\sqrt{d}} + e^{-t}\mu_1 \frac{W_{ik}}{\sqrt{d}} u_k^\nu + v_t \eta_i^\nu\right) \times \tag{6.10}$$

$$\left(e^{-t}\mu_1 \frac{m}{\sigma_{\boldsymbol{x}}} c^\nu \frac{w_j}{\sqrt{d}} + e^{-t}\mu_1 \frac{W_{jl}}{\sqrt{d}} u_l^\nu + v_t \eta_l^\nu\right), \tag{6.11}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots)^\top$ denotes the first column of $\boldsymbol{W}$, and v_t absorbs the diffusion-dependent variance renormalization.

The derivation of such gaussian equivalent follows precisely the one in .3: it's enough to use the GEP in 6.7 instead of 4.18 when needed.

6.4 Diagrammatic approach to the Features Covariance Spectrum

While in 3 we exploited the Replica Method in order to derive the features covariance spectrum, here we approach the computations exploiting Wick Theorem and Feynman diagrams.

6.4.1 Setting

Define the resolvent and its Stieltjes transform:

$$\boldsymbol{G}(z) = \frac{1}{z\mathbb{I}_p - \boldsymbol{U}}, \qquad g(z) = \lim_{p\to\infty} \frac{1}{p}\operatorname{Tr}\big(\overline{\boldsymbol{G}}(z)\big), \qquad \overline{\boldsymbol{G}}(z) = \mathbb{E}_{\{\boldsymbol{x}^\mu\}}\left[\frac{1}{z\mathbb{I}_p - \boldsymbol{U}}\right]. \tag{6.12}$$

The spectral density follows by Stieltjes inversion:

$$\rho(\lambda) = \lim_{\varepsilon\to 0^+} \frac{1}{\pi} \operatorname{Im}\, g(\lambda - i\varepsilon). \tag{6.13}$$

Concentration implies $g(z)$ equals its disorder average in the limit $p, n, d \to \infty$ at fixed ratios.

A Gaussian integral identity rewrites the matrix inverse as a quadratic functional integral over an auxiliary p-vector field $\boldsymbol{\psi}$:

$$\overline{G}_{ij}(z) = \mathbb{E}_{\boldsymbol{x}} \frac{\int d\boldsymbol{\psi}\ \exp\{-\frac{1}{2}\boldsymbol{\psi}^\top [z\mathbb{I}_p - \boldsymbol{U}]\,\boldsymbol{\psi}\}\,\psi_i\psi_j}{\int d\boldsymbol{\psi}\ \exp\{-\frac{1}{2}\boldsymbol{\psi}^\top [z\mathbb{I}_p - \boldsymbol{U}]\,\boldsymbol{\psi}\}}. \tag{6.14}$$

We eliminate the denominator via replicas:

$$\overline{G}_{ij}(z) = \lim_{m\to 0} \mathbb{E}_{\boldsymbol{x}} \int \prod_{a=1}^{m} d\boldsymbol{\psi}^{\,a}\ \exp\left\{-\frac{z}{2}\sum_a (\boldsymbol{\psi}^{\,a})^\top \boldsymbol{\psi}^{\,a}\right\} \times \tag{6.15}$$

$$\exp\left\{\frac{1}{2}\sum_a (\boldsymbol{\psi}^{\,a})^\top \boldsymbol{U}\,\boldsymbol{\psi}^{\,a}\right\} \psi_i^1 \psi_j^1. \tag{6.16}$$

We call

$$\langle\cdot\rangle_0 := \int \prod_{a=1}^{m} d\boldsymbol{\psi}^{\,a}\ \exp\left\{-\frac{z}{2}\sum_a (\boldsymbol{\psi}^{\,a})^\top \boldsymbol{\psi}^{\,a}\right\} (\cdot), \qquad \langle \psi_i^a \psi_j^b \rangle_0 = z^{-1}\delta_{ij}\delta_{ab}, \tag{6.17}$$

the *bare measure* and propagator.

Let $\boldsymbol{X} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n] \in \mathbb{R}^{d\times n}$ and introduce the *interaction*

$$V := \frac{1}{2}\sum_a (\boldsymbol{\psi}^{\,a})^\top \boldsymbol{U}\,\boldsymbol{\psi}^{\,a}. \tag{6.18}$$

Expanding $e^V = \sum_{k\geq 0} V^k/k!$ we obtain the formal series

$$\overline{G}_{ij} = \lim_{m\to 0} \sum_{k=0}^{\infty} \frac{1}{k!} \left\langle V^k\, \psi_i^1\, \psi_j^1 \right\rangle_{0,\boldsymbol{x}}. \tag{6.19}$$

6.4.2 Wicks theorem for Gaussian fields

In our framework, all random fields are jointly Gaussian. Hence, all their higher-order moments can be expressed in terms of pairwise covariances through *Wicks theorem*.

Statement. For any centered jointly Gaussian variables $\{\Phi_\alpha\}$ with covariances $C_{\alpha\beta} = \mathbb{E}[\Phi_\alpha \Phi_\beta]$, all odd moments vanish, and even moments factorize as

$$\mathbb{E}[\Phi_{\alpha_1} \Phi_{\alpha_2} \cdots \Phi_{\alpha_{2k}}] = \sum_{\pi \in \Pi_{2k}} \prod_{(r,s)\in\pi} C_{\alpha_r \alpha_s}, \tag{6.20}$$

where Π_{2k} denotes the set of all pairwise contractions (pairings) of $\{1, \ldots, 2k\}$.

In our context this means that all higher-order expectations in (6.19) are therefore completely determined by pairwise pairings between the involved fields. Each pair contributes to the average through the respective covariance under the specific distribution. For example for ψs and xs we have:

$$\langle \psi_i^a \psi_j^b \rangle_0 = z^{-1} \delta_{ij} \delta_{ab}, \qquad \mathbb{E}[x_\ell^\mu x_m^\nu] = (\sigma_{\boldsymbol{x}}^2 \delta_{\ell m}) \delta_{\mu\nu}.$$

Different fields contribute to the average through V. Hence we have to take into account how pairs from different fields interact with each other. The diagrammatic approach consist in rappresenting each interaction as a so-called vertex. Since fields can interact in various ways, different kind of verteces might exist, each describing a particular interaction. Each kind of vertex is associated to a "weights", which can be read off V itself. Given a vertex, which associated to two or more pairs of different fields, we must understand how different pairing choices affect the resulting diagram, obtained by collapsing indeces in according to the pairings. The power k indicates how many verteces will the diagram built by.

6.4.3 Self-energy and Dyson equations

Every entry $\bar{\mathbf{G}}_{ij}$ is obtained by summing the weights of all diagrams with any number of vertices that connect a ψ_i^1 and a ψ_j^1 field. If we call self-energy the matrix $\mathbf{\Sigma}^b \in \mathbb{R}^{N\times N}$ whose elements Σ_{ij} are the sum of all such 1-Particle Irreducible Diagrams that connect fields with block indices i and j, where we have removed every incoming and outgoing bare propagator (so-called *amputated diagrams*), then we can write the matrix Dyson equation

$$\mathbf{G}^b = \left[\left((g^0_{\text{quark}})^{-1} I_N - \mathbf{\Sigma}^b \right)^{-1} \right], \tag{6.21}$$

This gives a self-consistent equation for the matrix $\bar{\mathbf{G}}^b$, since $\mathbf{\Sigma}^b$ depends on $\bar{\mathbf{G}}^b$.

Bulks. A fundamental fact in random-matrix theory is that adding or subtracting a matrix of rank $o(p)$ does not affect the limiting spectral distribution (LSD). Hence, *bulk* eigenvalues are unchanged by finite-rank perturbations; only a finite number of eigenvalues may detach as *outliers*. Operationally, we thus:

1. Remove the low-rank (signal-correlated) terms and compute the averaged resolvent of the resulting *bulk* matrix.
2. Re-introduce the low-rank spike as a rank-r selfenergy $\widetilde{\Sigma}(z)$, and apply Woodbury/ShermanMorrison to the resolvent: $G(z) = (G_b(z)^{-1} - \widetilde{\Sigma}(z))^{-1}$.

The resolvent is defined as

$$G = \frac{1}{z\mathbb{I}_p - \boldsymbol{U}}, \tag{6.22}$$

and can be written leveraging the Dyson equation as

$$G = \left(z\mathbb{I}_p - \Sigma_{\text{det}} - \Sigma \right)^{-1}. \tag{6.23}$$

The deterministic self-energy term Σ_{det} includes only the first two terms of $\boldsymbol{U}$, namely

$$\Sigma_{\text{det}} = s_t^2 \mathbb{I}_p + \Delta_t \mu_1^2 \frac{\boldsymbol{W}\boldsymbol{W}^T}{d}. \tag{6.24}$$

This can be seen either by including Σ_{det} in the bare propagator, or by noticing that these terms are one-particle irreducible (1PI) and do not couple with other diagrams except in series.

We are therefore interested in the self-energy computed from the 1PI diagrams of

$$\frac{1}{n}\boldsymbol{G}\boldsymbol{G}^T.$$

This computation was already performed in [16] for a similar matrix:

$$\mathcal{H} = \frac{1}{p}\sum_{\nu} f(y^{\nu})\left(\kappa_1 \boldsymbol{F}\boldsymbol{z}^{\nu} + \kappa_* \boldsymbol{\eta}^{\nu}\right)\left(\kappa_1 \boldsymbol{F}\boldsymbol{z}^{\nu} + \kappa_* \boldsymbol{\eta}^{\nu}\right)^T . \tag{6.25}$$

We have already mapped the scales as

$$D \longleftrightarrow d, \qquad N \longleftrightarrow p, \qquad M \longleftrightarrow n,$$

so that in our case $\alpha = n/p$.

To find a precise mapping between $\mathcal{H}$ and $\frac{1}{n}\boldsymbol{G}\boldsymbol{G}^T$ we impose:

$$f(y^{\nu}) = \frac{p}{n}, \tag{6.26}$$

$$\boldsymbol{F} = \boldsymbol{W}, \tag{6.27}$$

$$\kappa_1 = e^{-t}\mu_1, \tag{6.28}$$

$$\kappa_* = v_t, \tag{6.29}$$

$$\boldsymbol{z}^{\nu} = \frac{\boldsymbol{x}^{\nu}}{\sqrt{d}}. \tag{6.30}$$

Notice that the $\frac{1}{\sqrt{d}}$ factor has been absorbed into $\boldsymbol{x}^{\nu}$, so that

$$\mathbb{E}_{\boldsymbol{z}}\left[(z_i^{\nu})^2\right] = \mathbb{E}_{\boldsymbol{x}}\left[\left(\frac{x_i^{\nu}}{\sqrt{d}}\right)^2\right] = \frac{1}{d},$$

as in the reference paper.

Following the same computation, one obtains the coupled fixed-point equations:

$$g(z) = \int dh(\ell)\, \frac{1}{z - s_t^2 - \Delta_t \mu_t^2 \ell - \dfrac{v_t^2 + e^{-2t}\mu_1^2 \ell}{1 - \dfrac{p}{n}\left(v_t^2 g(z) + e^{-2t}\mu_1^2 \tau(z)\right)}}, \tag{6.31}$$

$$\tau(z) = \int dh(\ell)\, \frac{\ell}{z - s_t^2 - \Delta_t \mu_t^2 \ell - \dfrac{v_t^2 + e^{-2t}\mu_1^2 \ell}{1 - \dfrac{p}{n}\left(v_t^2 g(z) + e^{-2t}\mu_1^2 \tau(z)\right)}}. \tag{6.32}$$

Here, $h(\ell)$ denotes the MarchenkoPastur distribution with aspect ratio

$$\psi_p = \frac{p}{d},$$

and

$$\tau = \frac{1}{p}\mathrm{Tr}\left(\frac{\boldsymbol{W}\boldsymbol{W}^T}{d}G\right).$$

Outlier Again, this maps to a matrix studied in [16], namely

$$\begin{aligned}(\mathcal{H})_{ij} = \frac{1}{p}\sum_{\nu} f(y^\nu)\Big(\kappa_1(f_i z_1^\nu + F_{ik}^{(1)} z_k^\nu) + \kappa_* \eta_i^\nu\Big) \\ \times \Big(\kappa_1(f_j z_1^\nu + F_{jl}^{(1)} z_l^\nu) + \kappa_* \eta_j^\nu\Big).\end{aligned} \tag{6.33}$$

Alongside the substitutions from above (with the only difference that we now use $\boldsymbol{u}$ instead of $\boldsymbol{x}$), we also replace

$$x_1^\nu = \frac{m}{\sigma_{\boldsymbol{x}}}\frac{c^\nu}{\sqrt{d}} \;\longrightarrow\; y^\nu = \frac{m}{\sigma_{\boldsymbol{x}}}c^\nu. \tag{6.34}$$

We thus compute the self-energy correction given by the outlier as

$$G = \left(G_b^{-1} - \tilde{\Sigma}\right)^{-1}, \tag{6.35}$$

where G_b is the resolvent computed in the previous section. By substitution, we obtain

$$\tilde{\Sigma} = e^{-2t}\mu_1^2\left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \frac{1}{1 - \frac{p}{n}\left(e^{-2t}\mu_1^2\tau(z) + v_t^2 g(z)\right)}\frac{\boldsymbol{w}\boldsymbol{w}^T}{d}. \tag{6.36}$$

Let

$$\gamma = \frac{1}{1 - \frac{p}{n}\left(e^{-2t}\mu_1^2\tau(z) + v_t^2 g(z)\right)}. \tag{6.37}$$

Then

$$G = \left(G_b^{-1} - e^{-2t}\mu_1^2\left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \gamma\frac{\boldsymbol{w}\boldsymbol{w}^T}{d}\right)^{-1}. \tag{6.38}$$

We compute the inverse using the Sherman–Morrison formula:

$$G = G_b + \gamma e^{-2t}\mu_1^2 \left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \frac{G_b \frac{\boldsymbol{w}\boldsymbol{w}^T}{d} G_b}{1 - \gamma e^{-2t}\mu_1^2 \left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \frac{1}{d}\boldsymbol{w}^T G_b \boldsymbol{w}}. \tag{6.39}$$

The denominator of the outlier contribution vanishes when

$$1 = \gamma e^{-2t}\mu_1^2 \left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \frac{1}{d}\boldsymbol{w}^T G_b \boldsymbol{w}. \tag{6.40}$$

Taking the normalized trace of both sides yields

$$1 = \gamma e^{-2t}\mu_1^2 \left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 \tau(z). \tag{6.41}$$

Substituting the original expression for γ, we obtain the self-consistent equation for the outlier eigenvalue λ^*:

$$\mu_1^2 e^{-2t}\tau(\lambda^*) = \frac{1 - \frac{p}{n} v_t^2 g(\lambda^*)}{\left(\frac{m}{\sigma_{\boldsymbol{x}}}\right)^2 + \frac{p}{n}}. \tag{6.42}$$

Solving numerically the equations derived in 6.4.3 and in 6.4.3 gives both the bulk spectral density and the value of the outlier. This is actually a work-in-progress, but first simulations, comparing these results with the numerical spectrum, seems to confirm the calculations.

Summary

In this chapter, we extended the analytical framework of independent Random Feature Neural Networks (RFNNs) to the case of data drawn from a *symmetric Gaussian mixture*, thereby introducing a controlled low-rank structure in the input distribution. This modification breaks the rotational symmetry characteristic of the single-Gaussian case and gives rise to a clear *Baik–Ben Arous–Péché (BBP) transition* in the spectrum of the feature covariance matrix $U(t)$.

From both theoretical construction and numerical analysis, we observed that the diffusion

time t plays the role of a control parameter for signal detectability:

- For small t, the latent mean direction $\boldsymbol{m}$ is strongly expressed. The feature covariance spectrum exhibits an isolated outlier eigenvalue whose associated eigenvector aligns almost perfectly with the signal.
- As t increases, diffusion gradually erases the signal, the outlier approaches the bulk, and the alignment with $\boldsymbol{m}$ decreases continuously—marking the onset of the BBP transition.
- For large t, the spectrum becomes isotropic again, recovering the two-bulk topology observed in the single-Gaussian setting.

An additional structural property concerns the alignment distribution among eigenmodes: the *memorization* bulk remains orthogonal to the signal, while the *generalization* modes carry all of it. When the outlier is present, it absorbs the full alignment; when it disappears, the total alignment redistributes uniformly across the bulk, preserving a total "zero-sum" alignment equal to one.

At the analytical level, we generalized the *Gaussian Equivalence Principle (GEP)* to Gaussian mixture inputs, enabling the replacement of nonlinear random features by statistically equivalent Gaussian surrogates. We then outlined a *diagrammatic expansion* based on Wicks theorem and Dyson equations to compute the spectrum of $U(t)$.

Finally, by combining the bulk analysis with a rank-one perturbative correction, we derived a self-consistent condition for the outlier eigenvalue λ^*, confirming the emergence of a BBP-type transition as a function of diffusion time, signal strength, and sample ratio.

In summary, the Gaussian mixture setting introduces a minimal yet non-trivial low-rank perturbation that interacts with diffusion dynamics in a controlled way. The resulting BBP transition in t marks the spectral onset of structure detectability in the feature space, while preserving the generalization–memorization spectral topology identified in Chapter 4. **The results presented in this chapter are still a work in progress.**

Fig. 10 shows projected samples onto $\boldsymbol{m}$. The plots show that RFNN is able to learn a bimodal distribution, although we can see that the generated distribution expresses some degree of degeneration. Clearly the sampling procedure is not ideal, and interpolating

different models at different diffusion times, as discussed above, might lead to numerical issues. Also, sampling from scratch do not produce evident memorization phenomena even for gaussian distributed data, as discussed in Chapter 4. Still, Fig. 10 supports the idea the RFNN, although somewhat simple in its architecture, is expressive enough to be adopted as meaningful theoretical framework.

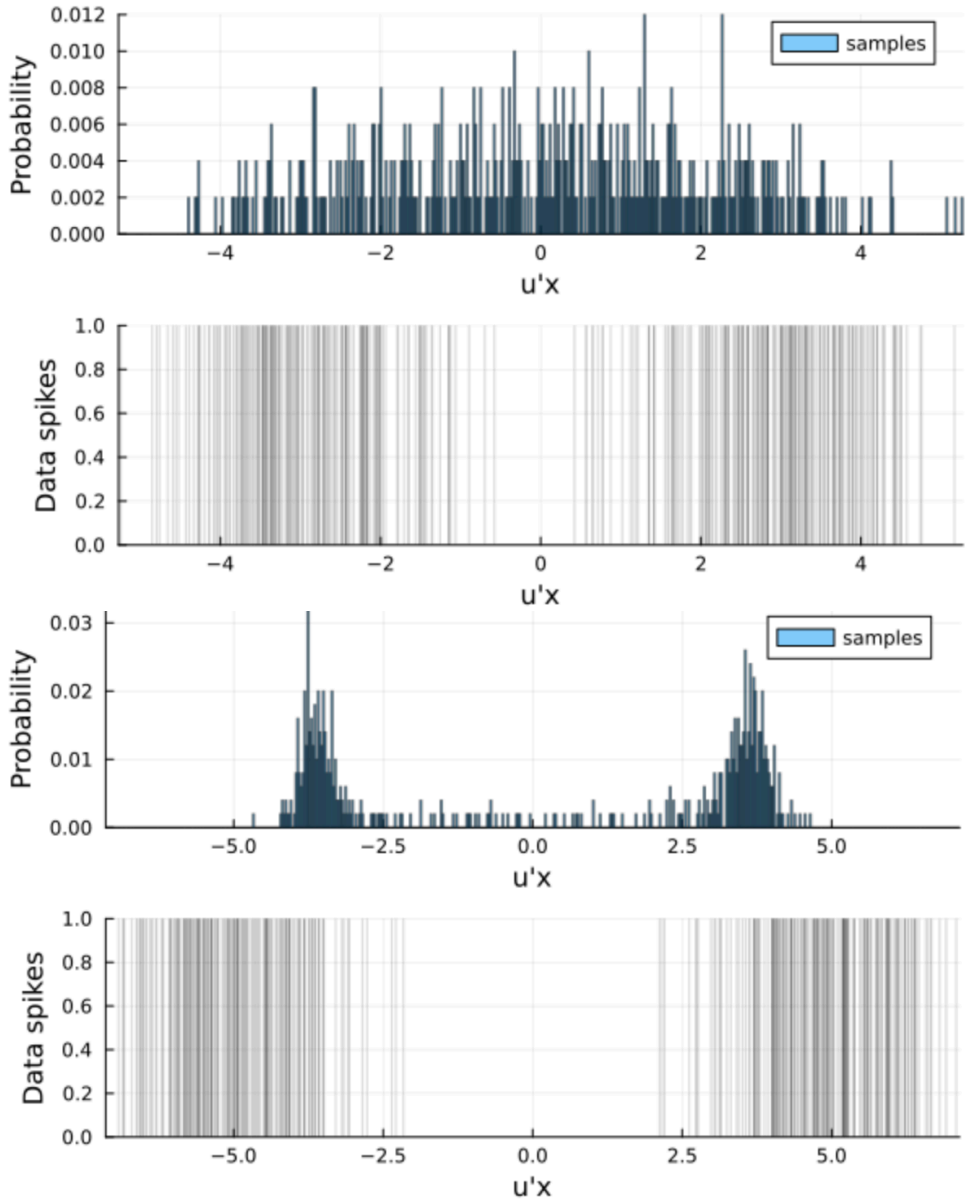


Figure 10: 1d projections of samples against data points at different diffusion times. Here we show $t = 0.5$ and $t = 0$.

7 Conclusions

Summary of Findings

This work examined the mechanisms governing memorization and generalization in diffusion models. Starting from the limiting case where the model learns the exact empirical score, we showed that such a system inevitably memorizes: each training point becomes an attractor of the reverse process, and generated samples collapse onto the training set. In contrast, real diffusion models generate new, meaningful data revealing that generalization arises not from explicit regularization, but from the learning process itself.

Architectural Scaling: Role of n and p

A key insight is that the transition between memorization and generalization is controlled by the relative scales of data size n and model size p. In the *under-parameterized* regime ($p \ll n$), the model cannot fit the empirical score exactly and instead learns a smoothed approximation, enabling generalization. As p grows toward n, the model enters a crossover where test loss rises and memorization begins. Beyond this point ($p \gg n$), the model starts to overfit the empirical score field, causing generated samples to collapse onto the training data.

This dependence defines an *architectural phase diagram* of diffusion learning, parameterized by the aspect ratios $\psi_p = p/d$ and $\psi_n = n/d$, separating generalization and memorization phases.

Learning as a Dynamic Regularizer

Generalization in diffusion models is inherently *dynamical*. During training, the model first captures coarse, global features of the score field (early *generalization phase*, at timescale τ_{gen}), then slowly converges toward the empirical optimum (*memorization phase*, at τ_{mem}). In practice, appropriate early stopping halts training before the second phase, leaving the model in the dynamically regularized generalization regime.

Extensions: Integrating Time-Dependent Architectures

We introduced a time-integrated RFNN with bias and skip connections, bridging the gap between analytically solvable models and realistic diffusion networks. Although time integration breaks analytical solvability, numerical evidence shows that the same two dynamical modes - generalization and memorization- persist. Temporal modulation through (α_t, β_t) effectively reweights these modes along the diffusion trajectory, offering a controllable means to stabilize learning across times.

Structured Data and Emerging Timescales

When the data contain latent structure, as in the Gaussian mixture model, a third, *signal timescale* appears. As diffusion time decreases, a BaikBen ArousPéché (BBP) transition emerges in the feature-spectrum, marking the point at which the model begins to encode latent data geometry. This enriches the phase diagram with an additional regime: signal emergence complements generalization and memorization within the same dynamical picture.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

[3] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[5] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning (ICML)*, 2015.

[6] Bernard Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79–82, 1980.

[7] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020.

[8] Claudia Merger and Sebastian Goldt. Generalization dynamics of linear diffusion models. *arXiv preprint*, arXiv:2505.24769, May 2025.

[9] Nicolas Macris et al. Theoretical analysis of generalization in score-based diffusion models. *arXiv preprint arXiv:2502.06789*, 2025.

[10] Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *arXiv preprint arXiv:2505.17638*, 2025.

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

[12] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, nov 2024.

[13] Giulio Biroli et al. Dynamical regimes of diffusion models. *arXiv preprint arXiv:2405.12345*, 2024.

[14] 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.

[15] Joonho Yoon et al. Generalization and memorization are mutually exclusive in diffusion models. *arXiv preprint arXiv:2310.11223*, 2023.

[16] Brandon Livio Annesi, Dario Bocchi, and Chiara Cammarota. Spectral initialization with random features. Unpublished manuscript, oct 2025.

[17] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, 2nd edition, 1996.

[18] Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Giorgio Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv preprint arXiv:2410.08727*, oct 2024.

[19] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 74–84. PMLR, nov 2020.

[20] Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, apr 2024.

[21] Brandon Livio Annesi, Dario Bocchi, and Chiara Cammarota. Overparametrization bends the landscape: Bbp transitions at initialization in simple neural networks. *arXiv preprint arXiv:2510.18435*, 2025. Submitted on 21 Oct 2025.

[22] Joshua Benton, Valentin de Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, oct 2023.

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

[24] Antoine P. M. Bodin. *Random matrix methods for high-dimensional machine learning models*. PhD thesis, EPFL, 2024.

[25] Valentin de Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, aug 2022.

[26] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[27] 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 Security Symposium*, pages 5253–5270. USENIX Association, 2023.

[28] Hao Chen, Han Lee, and Ji Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4735–4763, jul 2023.

[29] Ming Chen, Kai Huang, Tianyu Zhao, and Ming Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4672–4712, 2023.

[30] Shuo Chen, Sinho Chewi, Jikai Li, Yifei Li, Ahmed Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, sep 2022.

[31] 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*, oct 2023.

[32] Hugo Cui, Cengiz Pehlevan, and Yue M. Lu. A precise asymptotic analysis of learning diffusion models: theory and insights. *arXiv preprint arXiv:2501.03937*, jan 2025.

[33] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[34] A. Zee. Law of addition in random matrix theory. *Nuclear Physics B*, 474(3):726–744, 1996.

[35] Jack W. Silverstein and Zhidong D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

[36] Camille Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[37] Soobin Yoon, Minsu Kwon, and Gunhee Kim. When diffusion models learn the training set: A memorization–generalization dichotomy. In *ICLR Workshop on Understanding Generalization in Deep Learning*, 2023.

[38] 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 *USENIX Security Symposium*, 2023.

[39] Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. *arXiv preprint arXiv:2502.00336*, 2025.

[40] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning (ICML)*, pages 3452–3462, 2020.

[41] 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. *arXiv preprint arXiv:2006.14709*, 2022.

[42] Antoine P. M. Bodin. *Random Matrix Methods for High-Dimensional Machine Learning Models*. PhD thesis, EPFL, 2024.

[43] Stéphane D'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In *International Conference on Machine Learning (ICML)*, pages 2280–2290, 2020.

[44] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[45] Shu Wang, Qi Zheng, Rui Zhao, and Han Chen. On the generalization of diffusion models. *arXiv preprint arXiv:2401.05291*, 2024.

[46] Jinho Baik, Gerard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

[47] Crispin W. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, 4th edition, 2009.

[48] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[49] Giorgio Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

[50] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2022.

Appendix

.1 Appendix A

We present detailed derivations and proofs of: (i) the Fokker–Planck equation for Itô SDEs in one and multiple dimensions; (ii) the formula for the time-reversed SDE (identification of the reverse drift containing the score term); and (iii) Andersons theorem stating that the time-reversed SDE has marginals equal to those of the original forward SDE, provided appropriate uniqueness holds. Throughout, we state assumptions, make all integration-by-parts steps explicit, and indicate where decay and well-posedness are used.

Notation and assumptions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting a standard d-dimensional Brownian motion $(W_t)_{t\in[0,T]}$. We consider the Itô SDE

$$dX_t \;=\; f(X_t, t)\, dt \;+\; g(t)\, dW_t, \qquad X_0 \sim p_0, \tag{.1}$$

where:

- $f : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ is sufficiently smooth (e.g. C^2 in x, continuous in t), with at most linear growth to ensure non-explosion.
- $g(t) \in \mathbb{R}^{d\times d}$ is continuous in t and (for simplicity) uniformly nondegenerate. The diffusion matrix is $a(t) := g(t)g(t)^\top$.
- All densities decay sufficiently fast at infinity (or have compact support) so that boundary terms vanish when integrating by parts.

For test functions φ, the (time-dependent) generator acts as

$$\mathcal{L}_t\varphi(x) := f(x,t) \cdot \nabla\varphi(x) \;+\; \tfrac{1}{2}\,\mathrm{tr}\big(a(t)\nabla^2\varphi(x)\big),$$

where $\nabla^2\varphi$ denotes the Hessian.

Fokker–Planck equation (1D)

Assume $d = 1$ and write $a(t) = g(t)^2$. The SDE is

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t, \qquad a(t) = g(t)^2.$$

Theorem .1 (Fokker–Planck in 1D)**.** *Under the stated regularity and decay assumptions, the density $p_t(x)$ of X_t satisfies*

$$\partial_t p_t(x) \;=\; -\partial_x\big(f(x,t)\,p_t(x)\big) \;+\; \tfrac{1}{2}\,\partial_x^2\big(a(t)\,p_t(x)\big). \tag{.2}$$

Proof. Let $\varphi \in C_c^\infty(\mathbb{R})$. Itôs formula yields

$$d\varphi(X_t) \;=\; \big[f(X_t,t)\varphi'(X_t) + \tfrac{1}{2}a(t)\varphi''(X_t)\big]\,dt \;+\; (\text{martingale}).$$

Taking expectations, using $\frac{d}{dt}\mathbb{E}[\varphi(X_t)] = \int \varphi\,\partial_t p_t$, and integrating by parts twice (boundary terms vanish),

$$\int \varphi\,\partial_t p_t = \int \big(f\varphi' + \tfrac{1}{2}a\varphi''\big)p_t = -\int \varphi\,\partial_x(fp_t) + \int \varphi\,\tfrac{1}{2}\partial_x^2(ap_t).$$

Since this holds for all φ, (.2) follows. □

Remark .2 (State-dependent diffusion). If $g = g(x,t)$, then $a = a(x,t)$ and the diffusion term becomes $\frac{1}{2}\partial_x^2\big(a(x,t)p_t(x)\big)$.

Fokker–Planck in multiple dimensions

Now $x \in \mathbb{R}^d$, $f(x,t) \in \mathbb{R}^d$, and $a(t) = g(t)g(t)^\top \in \mathbb{R}^{d\times d}$.

Theorem .3 (Fokker–Planck in $\mathbb{R}^d$)**.** *Under the same assumptions,*

$$\partial_t p_t(x) \;=\; -\nabla\cdot\big(f(x,t)\,p_t(x)\big) \;+\; \tfrac{1}{2}\nabla\cdot\big(a(t)\,\nabla p_t(x)\big), \tag{.3}$$

i.e.,

$$\partial_t p_t = -\sum_{i=1}^{d} \partial_{x_i}\big(f_i p_t\big) + \tfrac{1}{2}\sum_{i,j=1}^{d} \partial_{x_i}\partial_{x_j}\big(a_{ij}(t)\, p_t\big).$$

Sketch. Apply Itôs formula to $\varphi(X_t)$ with $\varphi \in C_c^\infty(\mathbb{R}^d)$, take expectations, and perform componentwise integration by parts using decay/compact support. □

Convention bridge to Section 2 (main text). In Section 2 the isotropic noise had covariance $\langle \eta_i(t)\eta_j(s)\rangle = 2\,\delta_{ij}\,\delta(t-s)$, which corresponds to $g(t) \equiv \sqrt{2}\, I_d$ and hence $a(t) \equiv 2\, I_d$. Plugging $a = 2I_d$ into (.3) gives the Laplacian form

$$\partial_t p_t = -\nabla\cdot(f\, p_t) + \Delta p_t,$$

exactly matching the convention used there.

Time reversal and the score term

We derive the drift of the time-reversed process $Y_t := X_{T-t}$, which will contain the score $\nabla \log p_t$.

Let $q_t(x)$ be the density of Y_t, i.e. $q_t(x) = p_{T-t}(x)$. Consider the ansatz

$$dY_t \;=\; \widetilde{f}(Y_t, t)\, dt \;+\; g(T-t)\, d\widetilde{W}_t, \tag{.4}$$

with $\widetilde{W}_t$ a Brownian motion in the reversed filtration. The Fokker–Planck for q_t associated with (.4) is

$$\partial_t q_t \;=\; -\nabla\cdot(\widetilde{f}\, q_t) + \tfrac{1}{2}\nabla\cdot\big(a(T-t)\nabla q_t\big).$$

But $q_t(x) = p_{T-t}(x)$ implies $\partial_t q_t(x) = -\partial_s p_s(x)\big|_{s=T-t}$. Using the forward Fokker–Planck for p_s at $s = T-t$,

$$-\Big[-\nabla\cdot(f\, p_{T-t}) + \tfrac{1}{2}\nabla\cdot(a\,\nabla p_{T-t})\Big] = -\nabla\cdot(\widetilde{f}\, q_t) + \tfrac{1}{2}\nabla\cdot(a\,\nabla q_t).$$

Cancel the diffusion terms, getting $\nabla\cdot(\widetilde{f}\, q_t) = \nabla\cdot(f\, q_t)$. Under decay/uniqueness, we take

$\widetilde{f}\, q_t = f\, q_t$ pointwise and obtain

$$\boxed{\widetilde{f}(x,t) \;=\; f(x,T-t) \;-\; a(T-t)\,\nabla \log p_{T-t}(x)}$$

where the minus sign appears because $\nabla\cdot\big((f - a\nabla \log p)\, p\big) = \nabla\cdot(f\, p) - \nabla\cdot(a\nabla p)$ reproduces the required PDE.

Rewriting in the original (forward) time variable, the reverse SDE (run backward in the original time) is

$$d\widetilde{X}_t \;=\; \big(f(\widetilde{X}_t,t) \;-\; a(t)\,\nabla \log p_t(\widetilde{X}_t)\big)\,dt \;+\; g(t)\,d\widetilde{W}_t, \tag{.5}$$

which is the standard reverse-diffusion formula used in score-based generative modeling.

Remark .4 (Sign conventions and parameterizations). Different textbooks write time reversal either by introducing $Y_t = X_{T-t}$ as above, or by evolving a process $\widetilde{X}_\tau$ with τ counting backward time. Both are equivalent after the chain rule $\partial_\tau = -\partial_t$. The form (2.12) matches the convention in the main text where the reverse drift is $f - a\nabla \log p$ at the same (forward) time index.

Andersons theorem (equality of marginals)

We now formalize that the reverse SDE with drift $f - a\nabla \log p$ reproduces the forward marginals in reverse time.

Theorem .5 (Andersons reverse-time SDE)**.** *Let X_t solve (.1) with marginals p_t satisfying the Fokker–Planck (.3). Consider*

$$d\widetilde{X}_t \;=\; \big(f(\widetilde{X}_t,t) - a(t)\nabla \log p_t(\widetilde{X}_t)\big)\,dt \;+\; g(t)\,d\widetilde{W}_t, \quad \widetilde{X}_0 \sim p_T. \tag{.6}$$

Assume well-posedness and uniqueness of distributional solutions to the associated Fokker–Planck with initial data p_T. Then for all $t \in [0,T]$,

$$\widetilde{p}_t(x) \;=\; p_{T-t}(x),$$

i.e. the reversed process reproduces the forward marginals in reverse order.

Proof. Set $q_t(x) := p_{T-t}(x)$. Then $q_0 = p_T$ and

$$\partial_t q_t = -\partial_s p_s\big|_{s=T-t} = \nabla\cdot(f\,q_t) - \tfrac{1}{2}\nabla\cdot(a\nabla q_t) = -\nabla\cdot\big((f - a\nabla\log q_t)\,q_t\big) + \tfrac{1}{2}\nabla\cdot(a\nabla q_t).$$

Thus q_t solves the Fokker–Planck associated to the drift $\widetilde{f} = f - a\nabla\log q_t$ with initial condition $q_0 = p_T$. By uniqueness, any solution $\widetilde{p}_t$ must equal q_t, hence $\widetilde{p}_t = p_{T-t}$. □

Consistency check: Ornstein–Uhlenbeck (OU). In Section 2, with $f(x,t) = -x$ and isotropic $a(t) \equiv 2I_d$, the forward Fokker–Planck is

$$\partial_t p_t = -\nabla\cdot((-x)p_t) + \Delta p_t = \nabla\cdot(xp_t) + \Delta p_t,$$

and the stationary Gaussian $p_\infty \propto \exp(-\|x\|^2/2)$ solves it. The reverse drift (2.12) is

$$\widetilde{f}(x,t) = -x \;-\; 2\,\nabla\log p_t(x),$$

exactly the score-corrected drift used to sample from p_0 by integrating backward from $p_T \approx \mathcal{N}(0, I_d)$, in agreement with the constructions in the main text.

Link to the score-based chapter. Equations (.3) and (2.12) are precisely the PDE/SDE identities used in Section 2 to introduce the score $S(x,t) = \nabla\log p_t(x)$ and to justify the reverse-time sampling dynamics. The OU check above recovers the explicit Gaussian formulas employed there (e.g. Γ_t^2 and the linear score $-x/\Gamma_t^2$ in the isotropic case), ensuring full consistency between the Appendix and the main derivations.

.2 Appendix B

Throughout, we consider the OU forward process with

$$a_t := e^{-t}, \qquad \Delta_t := 1 - e^{-2t},$$

so $x_t = a_t x_0 + \sqrt{\Delta_t}\, z$, $z \sim \mathcal{N}(0, I_d)$.

Let $\{x_\mu\}_{\mu=1}^n$ be the training set, and let the empirical (mixture) density at time t be

$$P_t^{\text{emp}}(x) = \frac{1}{n(2\pi\Delta_t)^{d/2}} \sum_{\mu=1}^{n} \exp\!\left(-\frac{\|x - a_t x_\mu\|^2}{2\Delta_t}\right), \qquad S_t^{\text{emp}}(x) := \nabla \log P_t^{\text{emp}}(x).$$

Let $S_A(\cdot, t)$ denote the learned score at time t (e.g. an RFNN). The DSM training error at fixed t and $m = \infty$ is

$$E_{\text{train}}^{\infty}(A; t) := \frac{1}{dn} \sum_{\mu=1}^{n} \mathbb{E}_z \Big\| \sqrt{\Delta_t}\, S_A\big(a_t x_\mu + \sqrt{\Delta_t}\, z,\, t\big) + z \Big\|^2.$$

Lemma .6 (Conditional-score identity)**.** *Let $y = a_t x_\mu + \sqrt{\Delta_t}\, z$ with μ uniform on $\{1, \dots, n\}$ and $z \sim \mathcal{N}(0, I_d)$ independent of the data. Then*

$$\mathbb{E}\big[z \,\big|\, y\big] = -\sqrt{\Delta_t}\, \nabla \log P_t^{\text{emp}}(y) = -\sqrt{\Delta_t}\, S_t^{\text{emp}}(y).$$

Proof. For each index μ, $p(y \mid \mu) = \mathcal{N}(y; a_t x_\mu, \Delta_t I_d)$, hence

$$\nabla \log P_t^{\text{emp}}(y) = \frac{\sum_\mu \nabla_y p(y \mid \mu)}{\sum_\mu p(y \mid \mu)} = \frac{\sum_\mu \big(-(y - a_t x_\mu)/\Delta_t\big)\, p(y \mid \mu)}{\sum_\mu p(y \mid \mu)} = -\frac{1}{\Delta_t}\Big(y - a_t\, \mathbb{E}[x_\mu \mid y]\Big).$$

Since $y = a_t x_\mu + \sqrt{\Delta_t}\, z$,

$$\mathbb{E}[z \mid y] = \frac{1}{\sqrt{\Delta_t}}\Big(y - a_t\, \mathbb{E}[x_\mu \mid y]\Big) = -\sqrt{\Delta_t}\, \nabla \log P_t^{\text{emp}}(y),$$

as claimed. □

Theorem .7 (Loss decomposition at fixed t)**.** *Define*

$$M_t \;:=\; \frac{1}{dn} \sum_{\mu=1}^{n} \mathbb{E}_z \Big\| S_A\big(a_t x_\mu + \sqrt{\Delta_t}\, z,\, t\big) - S_t^{\text{emp}}\big(a_t x_\mu + \sqrt{\Delta_t}\, z\big) \Big\|^2,$$

$$V_t \;:=\; \frac{1}{dn} \sum_{\mu=1}^{n} \mathbb{E}_z \Big\| \sqrt{\Delta_t}\, S_t^{\text{emp}}\big(a_t x_\mu + \sqrt{\Delta_t}\, z\big) + z \Big\|^2.$$

Then the DSM training error decomposes as

$$E^{\infty}_{\text{train}}(A;t) \;=\; V_t \;+\; \Delta_t\, M_t.$$

Proof. Abbreviate $y = a_t x_\mu + \sqrt{\Delta_t}\, z$. Then

$$E^{\infty}_{\text{train}}(A;t) = \frac{1}{d}\,\mathbb{E}\Big\|\sqrt{\Delta_t}\, S_A(y,t) + z\Big\|^2.$$

Insert and subtract $\sqrt{\Delta_t}\, S^{\text{emp}}_t(y)$ and expand:

$$\begin{aligned}\big\|\sqrt{\Delta_t}\, S_A(y,t) + z\big\|^2 = \big\|\sqrt{\Delta_t}\,(S_A - S^{\text{emp}}_t)(y,t)\big\|^2 + \big\|\sqrt{\Delta_t}\, S^{\text{emp}}_t(y) + z\big\|^2 \\ + 2\Big\langle \sqrt{\Delta_t}\,(S_A - S^{\text{emp}}_t)(y,t)\,,\; \sqrt{\Delta_t}\, S^{\text{emp}}_t(y) + z\Big\rangle.\end{aligned}$$

Taking expectations and dividing by d gives

$$E^{\infty}_{\text{train}}(A;t) = \Delta_t\, M_t + V_t + \frac{2}{d}\,\mathbb{E}\Big\langle S_A(y,t) - S^{\text{emp}}_t(y),\; \sqrt{\Delta_t}\, S^{\text{emp}}_t(y) + z\Big\rangle.$$

By the tower property and Lemma .6,

$$\mathbb{E}\Big[\big\langle S_A(y,t) - S^{\text{emp}}_t(y),\; \sqrt{\Delta_t}\, S^{\text{emp}}_t(y) + z\big\rangle\Big] = \mathbb{E}\Big[\big\langle S_A(y,t) - S^{\text{emp}}_t(y),\; \sqrt{\Delta_t}\, S^{\text{emp}}_t(y) + \mathbb{E}[z \mid y]\big\rangle\Big] = 0,$$

so the cross term vanishes and the identity follows. □

.3 Appendix C

Gaussian Equivalence Principle As explained in [30], the Gaussian Equivalence Theorem which applies in the high dimensional setting considered here establishes an equivalence regarding its spectral properties to a Gaussian covariate model where the nonlinear activation function is replaced by a linear term and a nonlinear term acting as noise:

$$\sigma\left(\frac{Wx}{\sqrt{d}}\right) \;\to\; \mu_1\,\frac{Wx}{\sqrt{d}} + \mu^*\eta, \qquad \eta \sim \mathcal{N}(0, I_p),$$

where μ_1, μ^* are defined in Sect. C.1 for random variables x drawn from $P_t = \mathcal{N}(0, \Gamma_t^2 I_d)$. In this section, we outline the derivation of the Gaussian Equivalence Principle (GEP) for the matrices $U, \tilde{U}, V$ and $\tilde{V}$ under arbitrary input variance. This is a simple generalization of the result of [9], which considered only the case of data drawn from $\mathcal{N}(0, I_d)$.

Lemma .8 (Gaussian Equivalence Principle for U)**.** *In the limit $n, p, d \to \infty$ with $\psi_p = p/d$, $\psi_n = n/d$, the matrix*

$$U \;=\; \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\Big[\sigma\Big(\frac{W x_{\nu,t}(\xi)}{\sqrt{d}}\Big)\sigma\Big(\frac{W x_{\nu,t}(\xi)}{\sqrt{d}}\Big)^\top\Big] \tag{.7}$$

has the same spectrum as its GEP equivalence

$$U \;=\; \frac{G}{\sqrt{n}}\frac{G^\top}{\sqrt{n}} + \Delta_t \mu_1^2 \frac{WW^\top}{d} + s_t^2 I_p, \tag{.8}$$

where

$$G \;=\; e^{-t}\mu_1 \frac{W}{\sqrt{d}} X + v_t\, \Omega, \tag{.9}$$

$X \in \mathbb{R}^{d\times n}$ is the matrix whose columns are the x_νs and $\Omega \in \mathbb{R}^{p\times d}$ has Gaussian entries independent of X and W.

Proof. For the sake of clarity, in this proof the vector x_ν have variance 1 and we explicitly make the variance of the data σ_x^2 appear. Let us focus on the element of U in position (i, j)

$$U_{ij} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_\xi\Big[\sigma\Big(\frac{W_{ik}(e^{-t}\sigma_x x_{\nu k} + \sqrt{\Delta_t}\xi_k)}{\sqrt{d}}\Big)\sigma\Big(\frac{W_{jl}(e^{-t}\sigma_x x_{\nu l} + \sqrt{\Delta_t}\xi_l)}{\sqrt{d}}\Big)\Big]. \tag{.10}$$

where repeated indices mean that there is a hidden sum. If $i = j$, the diagonal terms concentrate with respect to the datapoints to

$$U_{ii} = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(\Gamma_t z)^2] + \mathcal{O}(1/n) = \|\sigma\|^2 + \mathcal{O}(1/n), \tag{.11}$$

since the noisy data points have variance Γ_t^2. The finite n corrections can be discarded because they cannot change the spectrum of U. Now if $i \neq j$, the expectation value on ξ

can also be written as the expectation value on $u = W_{ik}\xi_k/\sqrt{d}$, $v = W_{jl}\xi_l/\sqrt{d}$ standard Gaussian random variables with correlation $\langle uv\rangle = W_{ik}W_{jk}/d$. Denote this expectation value as

$$\mathbb{E}_{u,v\sim P_{W_{ik}W_{jk}/d}}[\cdot]. \tag{.12}$$

We use Mehler Kernel formula [19] to decouple u and v

$$\mathbb{E}_{u,v\sim P_\gamma}[\cdot] \;=\; \sum_{s=1}^{\infty}\frac{\gamma^s}{s!}\,\mathbb{E}_{u,v}\big[\mathrm{He}_s(u)\,\mathrm{He}_s(v)\;\cdot\,\big], \tag{.13}$$

where now $\mathbb{E}_{u,v}[\cdot]$ is over u and v independent and He_s are the Hermite polynomials.

$$U_{ij} = \sum_{\nu=1}^{n}\sum_{s=1}^{\infty}\frac{(W_{ik}W_{jk}/d)^s}{s!}\mathbb{E}_u\Big[\mathrm{He}_s(u)\,\sigma\Big(\frac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}}+\sqrt{\Delta_t}\,u\Big)\Big] \tag{.14}$$

$$\times\,\mathbb{E}_v\Big[\mathrm{He}_s(v)\,\sigma\Big(\frac{W_{jl}\,e^{-t}\sigma_x\,x_{\nu l}}{\sqrt{d}}+\sqrt{\Delta_t}\,v\Big)\Big] \tag{.15}$$

$$=\;\frac{1}{n}\sum_{\nu=1}^{n}\Big(\mathbb{E}_u\big[\sigma\Big(\tfrac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}}+\sqrt{\Delta_t}\,u\Big)\big]\Big)\Big(\mathbb{E}_v\big[\sigma\Big(\tfrac{W_{jl}\,e^{-t}\sigma_x\,x_{\nu l}}{\sqrt{d}}+\sqrt{\Delta_t}\,v\Big)\big]\Big) \tag{.16}$$

$$+\,\frac{1}{n}\sum_{\nu=1}^{n}\frac{W_{ik}W_{jk}}{d}\,\mathbb{E}_u\Big[u\,\sigma\Big(\frac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}}+\sqrt{\Delta_t}\,u\Big)\Big]\,\mathbb{E}_v\Big[v\,\sigma\Big(\frac{W_{jl}\,e^{-t}\sigma_x\,x_{\nu l}}{\sqrt{d}}+\sqrt{\Delta_t}\,v\Big)\Big] \tag{.17}$$

$$=\;\frac{1}{n}\sum_{\nu=1}^{n}\Big(\mathbb{E}_u\big[\sigma\Big(\tfrac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}}+\sqrt{\Delta_t}\,u\Big)\big]\Big)\Big(\mathbb{E}_v\big[\sigma\Big(\tfrac{W_{jl}\,e^{-t}\sigma_x\,x_{\nu l}}{\sqrt{d}}+\sqrt{\Delta_t}\,v\Big)\big]\Big) \tag{.18}$$

$$+\,\frac{W_{ik}W_{jk}}{d}\,\mathbb{E}_x\Big[\mathbb{E}_u\Big[u\,\sigma\Big(\frac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}}+\sqrt{\Delta_t}\,u\Big)\Big]\,\mathbb{E}_v\Big[v\,\sigma\Big(\frac{W_{jl}\,e^{-t}\sigma_x\,x_{\nu l}}{\sqrt{d}}+\sqrt{\Delta_t}\,v\Big)\Big]\Big], \tag{.19}$$

by neglecting order $\mathcal{O}(1/d)$. Let us first focus on the second term. We expand $\sigma(\Gamma_t\cdot)$ on

the Hermite polynomials base:

$$\sigma(\Gamma_t x) \;=\; \sum_{s=1}^{\infty} \frac{\alpha_s}{s!}\,\mathrm{He}_s(x), \qquad \alpha_s \;=\; \mathbb{E}_z\big[\sigma(\Gamma_t z)\,\mathrm{He}_s(z)\big]. \tag{.20}$$

$$\mathbb{E}_u\Big[u\,\sigma\Big(\frac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}} + \sqrt{\Delta_t}\,u\Big)\Big] \;= \tag{.21}$$

$$\sum_{s=1}^{\infty} \frac{\alpha_s}{s!\,\Gamma_t}\,\mathbb{E}_u\Big[u\,\mathrm{He}_s\Big(\frac{1}{\Gamma_t}\Big(\frac{W_{ik}\,e^{-t}\sigma_x\,x_{\nu k}}{\sqrt{d}} + \sqrt{\Delta_t}\,u\Big)\Big)\Big] \;=\; \sqrt{\Delta_t}\,\frac{\alpha_1}{\Gamma_t}, \tag{.22}$$

hence the second term gives

$$\frac{WW^\top}{d}\,\Delta_t\,\frac{\alpha_1^2}{\Gamma_t^2} \;=\; \frac{WW^\top}{d}\,\Delta_t\,\mu_1^2. \tag{.23}$$

For the first term we write

$$\frac{1}{n}\sum_{\nu=1}^{n}\Big(\mathbb{E}_u\big[\sigma\Big(\frac{W_i\,e^{-t}\sigma_x\,x_\nu}{\sqrt{d}} + \sqrt{\Delta_t}\,u\Big)\big]\Big)\Big(\mathbb{E}_v\big[\sigma\Big(\frac{W_j\,e^{-t}\sigma_x\,x_\nu}{\sqrt{d}} + \sqrt{\Delta_t}\,v\Big)\big]\Big), \tag{.24}$$

and define

$$\sigma_0(x) = \mathbb{E}_z\big[\sigma(e^{-t}\sigma_x x + \sqrt{\Delta_t}\,z)\big]. \tag{.25}$$

It follows that

$$U_{ij} = \frac{1}{n}\sum_{\nu=1}^{n}\sigma_0\Big(\frac{W_i x_\nu}{\sqrt{d}}\Big)\sigma_0\Big(\frac{W_j x_\nu}{\sqrt{d}}\Big). \tag{.26}$$

We use the GEP on σ_0 and its argument $W_{ik}x_{\nu k}/\sqrt{d}$ of variance 1

$$\sigma_0\Big(\frac{WX}{\sqrt{d}}\Big) = \mathbb{E}_z[\sigma_0(z)z]\frac{WX}{\sqrt{d}} + \sqrt{\mathbb{E}_z[\sigma_0(z)^2] - \mathbb{E}_z[\sigma_0(z)z]^2}\,\Omega, \tag{.27}$$

$$\mathbb{E}_z[\sigma_0(z)z] = \mathbb{E}_{z,u}\big[\sigma(e^{-t}\sigma_x z + \sqrt{\Delta_t}\,u)\,z\big] = e^{-t}\sigma_x\,\mu_1/\Gamma_t, \tag{.28}$$

$$\mathbb{E}_z[\sigma_0(z)^2] = \mathbb{E}_{z,u,v}\big[\sigma(e^{-t}\sigma_x z + \sqrt{\Delta_t}\,u)\,\sigma(e^{-t}\sigma_x z + \sqrt{\Delta_t}\,v)\big], \tag{.29}$$

$$= \mathbb{E}_{u,v\sim P_{e^{-2t}\sigma_x^2/\Gamma_t^2}}[\sigma(\Gamma_t u)\sigma(\Gamma_t v)] = c_t. \tag{.30}$$

Hence the zero order term can be written as

$$\frac{1}{n}\Big(e^{-t}\mu_1\frac{\sigma_x}{\Gamma_t}\frac{WX}{\sqrt{d}}+\sqrt{c_t-e^{-2t}\sigma_x^2\mu_1^2/\Gamma_t^2}\,\Omega\Big)\Big(e^{-t}\mu_1\frac{\sigma_x}{\Gamma_t}\frac{WX}{\sqrt{d}}+\sqrt{c_t-e^{-2t}\sigma_x^2\mu_1^2/\Gamma_t^2}\,\Omega\Big)^\top. \quad (.31)$$

To finish the analysis we need to take care of the diagonal term. It must be equal to $\|\sigma\|^2$. The zero order term give $c_t-e^{-2t}\mu_1^2+e^{-t}\mu_1=\Delta_t\mu_1^2$ while the first order gives $\Delta_t\mu_1^2$, hence we need to add a term $\|\sigma\|^2-c_t-\Delta_t\mu_1^2=s_t^2$. Hence the GEP of U reads

$$U=\frac{G}{\sqrt{n}}\frac{G^\top}{\sqrt{n}}+\Delta_t\mu_1^2\frac{WW^\top}{d}+s_t^2I_p. \quad (.32)$$

□

.4 Appendix D

Proof. Our goal is to compute the Stieltjes transform of the matrix U.

$$\begin{aligned} q &= \lim_{p\to\infty}\frac{1}{p}\,\mathbb{E}_{W,X,\Omega}\big[\mathrm{Tr}(U-zI_p)^{-1}\big] && (.33)\\ &= -\partial_z\lim_{p\to\infty}\frac{1}{p}\,\mathbb{E}_{W,X,\Omega}[\log\det(U-zI_p)] && (.34)\\ &= 2\,\partial_z\lim_{p\to\infty}\frac{1}{p}\,\mathbb{E}_{W,X,\Omega}\Big[\log\det(U-zI_p)^{-1/2}\Big]. && (.35)\end{aligned}$$

The so-called replica trick consists of replacing the $\log x$ by $\lim_{m\to 0}\frac{x^m-1}{m}$. Applying this identity, we obtain

$$q=2\,\partial_z\lim_{m\to 0}\lim_{p\to\infty}\frac{1}{pm}\Big(\mathbb{E}_{W,X,\Omega}\Big[\det(U-zI_p)^{-m/2}\Big]-1\Big). \quad (.36)$$

We define the partition function Z as

$$Z=\det(U-zI_p)^{-1/2}\;=\;\int\frac{d\varphi}{(2\pi)^{p/2}}\exp\Big\{-\frac{1}{2}\,\varphi^\top(U-zI_p)\varphi\Big\}. \quad (.37)$$

We now use the fact that for any integer m

$$\mathbb{E}_{W,X,\Omega}[Z^m] = \int \prod_{a=1}^{m} \frac{d\varphi_a}{(2\pi)^{p/2}} \, \mathbb{E}_{W,X,\Omega}\left[\exp\left\{-\frac{1}{2}\sum_{a=1}^{m} \varphi_a^\top (U - zI_p)\varphi_a\right\}\right] \tag{.38}$$

$$= \int \prod_{a=1}^{m} \frac{d\varphi_a}{(2\pi)^{p/2}} \, \exp\left\{-\frac{1}{2}\sum_{a=1}^{m} \varphi_a^\top (s_t^2 - z)\varphi_a\right\} \times \mathbb{E}_{W,X,\Omega}\left[\exp\left\{-\frac{1}{2}\sum_{a=1}^{m} \varphi_a^\top \left(\frac{GG^\top}{n} + \Delta_t \mu_1^2 \frac{WW^\top}{d}\right)\varphi_a\right\}\right]. \tag{.39}$$

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

$$\int dx \, \exp\left\{-\frac{1}{2}xGx^\top + Jx^\top\right\} = \exp\left\{-\frac{1}{2}\log\det G + \frac{1}{2}JG^{-1}J^\top\right\}, \tag{.40}$$

where G is a square matrix and J a vector.

Averaging over the dataset The dataset dependence enters through

$$\mathbb{E}_{\mathbf{X}}\left[e^{-\frac{1}{2}\phi^{aT}\left(\frac{\mathbf{G}\mathbf{G}^T}{n}\right)\phi^a}\right] = \mathbb{E}_{\mathbf{X}}\left[e^{-\frac{1}{2n}\phi^{aT}\left(e^{-t}\mu_1\frac{\mathbf{W}\mathbf{X}}{\sqrt{d}}+\nu_t\Omega\right)\left(e^{-t}\mu_1\frac{\mathbf{W}\mathbf{X}}{\sqrt{d}}+\nu_t\Omega\right)^T\phi^a}\right]$$

$$= \mathbb{E}_{\mathbf{X}}\left[e^{-\frac{e^{-2t}\mu_1^2}{2nd}\phi^{aT}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T\phi^a} e^{-\frac{e^{-t}\mu_1\nu_t}{2\sqrt{d}n}\phi^{aT}(\mathbf{W}\mathbf{X}\Omega^T+\Omega\mathbf{X}^T\mathbf{W}^T)\phi^a} e^{-\frac{\nu_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}\right].$$

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

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

Hence,

$$\mathbb{E}_{\mathbf{X}}\left[e^{-\frac{1}{2}\phi^{aT}\left(\frac{\mathbf{G}\mathbf{G}^T}{n}\right)\phi^a}\right] = \mathbb{E}_{\mathbf{X}}\left[e^{-\frac{e^{-2t}\mu_1^2 p}{2nd}\omega^{aT}\mathbf{X}\mathbf{X}^T\omega^a} e^{-\frac{e^{-t}\mu_1\sqrt{p}\nu_t}{\sqrt{dn}}\sum_{a,\nu}\Omega^\nu\phi^{a\nu}\omega^{aT}\mathbf{x}^\nu} e^{-\frac{\nu_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}\right].$$

Denote $\mathbf{G}_{\mathbf{X}} = \frac{e^{-2t}\mu_1^2\sigma_x^2 p}{dn}\sum_a \omega^a\omega^{aT}$ and $(\mathbf{J}_{\mathbf{X}})_k^\nu = \frac{e^{-t}\mu_1\sqrt{p}\sigma_x\nu_t}{\sqrt{dn}}\sum_a(\Omega^\nu\cdot\phi^a)\omega_k^a$, then

$$\mathbb{E}_{\mathbf{X}}\left[e^{-\frac{1}{2}\phi^{aT}\left(\frac{\mathbf{G}\mathbf{G}^T}{n}\right)\phi^a}\right] = e^{-\frac{n}{2}\log\det(I_d+\mathbf{G}_{\mathbf{X}})} e^{\frac{e^{-2t}\mu_1^2\sigma_x^2 p\nu_t^2}{2dn^2}\sum_{\nu,b}(\Omega^\nu\cdot\phi^a)(\Omega^\nu\cdot\phi^b)\omega_k^a(1+\mathbf{G}_{\mathbf{X}})^{-1}_{k,l}\omega_l^b e^{-\frac{\nu_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}}.$$

Averaging over Ω The terms that depend on Ω are

$$\mathbb{E}_\Omega\left[e^{\frac{e^{-2t}\mu_1^2 p\nu_t^2\sigma_x^2}{2dn^2}\sum_\nu(\Omega^\nu\cdot\phi^a)(\Omega^\nu\cdot\phi^b)\omega_k^a(1+\mathbf{G}_{\mathbf{X}})^{-1}_{k,l}\omega_l^b e^{-\frac{\nu_t^2}{2n}\phi^{aT}\Omega\Omega^T\phi^a}}\right] = e^{-\frac{n}{2}\log\det(1+\mathbf{G}_\Omega)},$$

with

$$(\mathbf{G}_\Omega)_{k,l} = \phi^a\left(\frac{\nu_t^2}{n}\delta_{ab} - \frac{e^{-2t}\mu_1^2 p\nu_t^2\sigma_x^2}{dn^2}\omega_k^a(1+\mathbf{G}_{\mathbf{X}})^{-1}_{k,l}\omega_l^b\right)\phi^b.$$

We are left with

$$\mathbb{E}_{\mathbf{W},\mathbf{X},\Omega}[Z^m] = \int\prod_{a=1}^m\frac{d\phi^a}{(2\pi)^{p/2}}\int d\omega^a d\hat{\omega}^a$$
$$\mathbb{E}_{\mathbf{W}}\left[e^{-\frac{1}{2}\phi^{aT}(s_z^2-z)\phi^a} e^{i\hat{\omega}^{aT}(\sqrt{p}\omega^a-\mathbf{W}^T\phi^a)} e^{-\frac{\Delta_t\mu_1^2 p}{2d}\omega^{aT}\omega^a} e^{-\frac{n}{2}\log\det(I_d+\mathbf{G}_{\mathbf{X}})-\frac{n}{2}\log\det(I_d+\mathbf{G}_\Omega)}\right].$$

Averaging over $\mathbf{W}$, $\hat{\omega}$, and introducing order parameters Q^{ab} and R^{ab}, we obtain

$$1 = \int dQ^{ab}d\hat{Q}^{ab}e^{-\frac{p}{2}(\log\det\hat{Q}-2\mathrm{Tr}(Q\hat{Q}))},$$
$$1 = \int dR^{ab}d\hat{R}^{ab}e^{-\frac{d}{2}(\log\det\hat{R}-2\mathrm{Tr}(R\hat{R}))},$$

leading to

$$\mathbb{E}_{\mathbf{W},\mathbf{X},\Omega}[Z^m] = \int dQ\, d\hat{Q}\, dR\, d\hat{R}\, e^{-\frac{d}{2}mS(Q,\hat{Q},R,\hat{R})}.$$

In the high-dimensional limit, we evaluate the integral via the saddle point method. Taking derivatives with respect to the conjugate order parameters gives

$$\hat{Q} = -\frac{1}{2}Q^{-1}, \tag{.41}$$

$$\hat{R} = -\frac{1}{2}R^{-1}. \tag{.42}$$

We can now rewrite the integrand purely in terms of the order parameters Q and R. In particular, quadratic terms involving φ_a and ω_a simplify to

$$\exp\left\{-\frac{1}{2}\sum_a \varphi_a^\top (s_t^2 - z)\varphi_a\right\} \exp\left\{-\frac{\Delta_t \mu_1^2}{2}\frac{p}{2d}\sum_a \omega_a^\top \omega_a\right\} = \exp\left\{-\frac{p}{2}(s_t^2 - z)\operatorname{Tr} Q - \frac{\Delta_t \mu_1^2}{2}\frac{p}{2}\operatorname{Tr} R\right\}. \tag{.43}$$

The term $e^{-\frac{p}{2}\omega^\top G^{-1}\omega}$ is rewritten as

$$\exp\left\{-\frac{1}{2}\log\det G - \frac{p}{2}\omega^\top G^{-1}\omega\right\} = \exp\left\{-\frac{d}{2}\log\det Q - \frac{d}{2}\operatorname{Tr}(RQ^{-1})\right\}. \tag{.44}$$

Similarly,

$$\exp\left\{-\frac{n}{2}\log\det(I_d + GX)\right\} = \exp\left\{-\frac{n}{2}\log\det\left(I_m + \mu_1^2 e^{-2t}\,\frac{p\,\sigma_x^2}{n}\,R\right)\right\}. \tag{.45}$$

The remaining term involves $G\Omega$,

$$\exp\left\{-\frac{n}{2}\log\det(I_d + G\Omega)\right\}. \tag{.46}$$

Using for instance the Taylor expansion of $(1+X)^{-1}$, we obtain that

$$e^{-2t}\mu_1^2 p v_t^2 \sigma_x^2\,\frac{dn^2}{p^2}\,\omega_a^k\,(1+GX)^{-1}_{k\ell}\,\omega_b^\ell = e^{-2t}\mu_1^2 p v_t^2 \sigma_x^2\,\frac{n^2}{p}\,R\,(1 + e^{-2t}\mu_1^2 p v_t^2 \sigma_x^2\,\tfrac{n}{p}R)^{-1}. \tag{.47}$$

We also need the fact that, for a matrix $A \in \mathbb{R}^{p\times p}$, the HubbardStratonovich transform yields

$$\log\det\Big(1+\frac{1}{n}\,\varphi_a A_{ab}\varphi_b^\top\Big) = \log\det\Big(1+\frac{p}{n}\,AQ\Big)\,, \tag{.48}$$

so that

$$\exp\Big\{-\frac{n}{2}\log\det(1+G\Omega)\Big\} = \exp\Big\{-\frac{n}{2}\log\Big(1+\tfrac{p}{n}\big(v_t^2-e^{-2t}\mu_1^2 p v_t^2\sigma_x^2\tfrac{n}{p}R(1+e^{-2t}\mu_1^2 p\sigma_x^2\tfrac{n}{p}R)^{-1}\big)Q\Big)\Big\}\,. \tag{.49}$$

Replica Symmetric Ansatz. We do the following Replica Symmetric (RS) Ansatz for Q and R, since in the replica approach to random matrices one does not expect any symmetry breaking (alternatively one could check the stability of the RS solution):

$$Q = \begin{pmatrix} q & \tilde q & \cdots & \tilde q \\ \tilde q & q & \cdots & \tilde q \\ \vdots & \vdots & \ddots & \vdots \\ \tilde q & \tilde q & \cdots & q \end{pmatrix}$$

$$\begin{pmatrix} r & \tilde r & \cdots & \tilde r \\ \tilde r & r & \cdots & \tilde r \\ \vdots & \vdots & \ddots & \vdots \\ \tilde r & \tilde r & \cdots & r \end{pmatrix}. \tag{.50}$$

Under this ansatz, the action reads

$$\begin{aligned} S = {}& \psi_n \log\Big(1+\frac{p}{n}\big(v_t^2(q-\tilde q)+\mu_1^2 e^{-2t}\sigma_x^2(r-\tilde r)\big)\Big) + \psi_n\frac{p}{n}\,\frac{v_t^2\tilde q+\mu_1^2e^{-2t}\sigma_x^2\tilde r}{1+\frac{p}{n}\big(v_t^2(q-\tilde q)+\mu_1^2e^{-2t}\sigma_x^2(r-\tilde r)\big)} \\ & + \psi_p(s_t^2-z)\,q + \Delta_t\mu_1^2\psi_p\,r + \log q + \frac{rq-2\tilde q\,r+\tilde r\,\tilde q}{(q-\tilde q)^2}. \end{aligned} \tag{.51}$$

Saddle point equations. The saddle point equations with respect to $\tilde q$ and $\tilde r$ yield $\tilde r = \tilde q = 0$. Evaluating the remaining saddle point conditions gives the coupled equations quoted in

Theorem 3.1:

$$\psi_p v_t^2 \frac{1}{1+e^{-2t}\mu_1^2\sigma_x^2\,\psi_p}\frac{\psi_p}{\psi_n}\,r \;+\; \frac{\psi_p}{\psi_n}v_t^2\,q \;+\; \psi_p(s_t^2-z) \;+\; 1-\psi_p q \;-\; rq^2 \;=\; 0, \qquad (.52)$$

$$\frac{e^{-2t}\mu_1^2\sigma_x^2\,\psi_p}{1+e^{-2t}\mu_1^2\sigma_x^2\,\psi_p}\frac{\psi_p}{\psi_n}\,r \;+\; \frac{\psi_p}{\psi_n}v_t^2\,q \;+\; \Delta_t\mu_1^2\psi_p \;+\; \frac{1}{q} \;-\; \frac{1}{r} \;=\; 0. \qquad (.53)$$

Finally, the eigenvalue distribution of U, $\rho(\lambda)$, can then be obtained using the SokhotskiPlemelj inversion formula

$$\rho(\lambda) \;=\; \lim_{\varepsilon\to 0^+}\frac{1}{\pi}\,\mathrm{Im}\;q(\lambda+i\varepsilon).$$