FIRST LEARN, THEN MEMORIZE: THE SPECTRAL BIAS OF DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models trained on a finite dataset first learn to generate novel, high-quality samples and only much later collapse onto their training set. We identify the mechanism behind this separation of timescales and the object that probes it. The training dynamics of the score function are governed—exactly, and at any width—by the Gram matrix of the Neural Tangent Kernel (NTK) evaluated on the noisy training data, so the timescales of generalization and of memorization must be encoded in its spectrum. We show that they are, and that the structure responsible has no analogue in standard kernel settings. The use of multiple noise realizations per sample (m noised copies at a fixed noise level) in the score-matching loss is what restructures the Gram matrix spectrum into two distinct parts. The first, of large eigenvalues, carries the global features of the target distribution and is present already for $m = 1$. The second, which the repeated noising creates, consists of the smallest eigenvalues and is supported on eigenvectors aligned with the sample-specific noise directions; it sets a memorization timescale parametrically larger in the training set size n. We establish this picture on two fronts. Analytically, we solve the spectrum in the lazy high-dimensional limit for both linear ($n \asymp d$) and polynomial ($n \asymp d^k$) sample complexities, and prove through a bias–variance decomposition that the first bulk minimizes the approximation error while the second drives the error associated with memorization. Empirically, we show the same two-bulk structure in Convolutional NTKs on CelebA and in finite-width U-Nets trained well beyond the lazy regime, and we make the link causal: truncating the Gram matrix at rank r tunes the generalization–memorization transition, and an L_2 penalty targeting the second bulk suppresses memorization in feature-learning U-Nets.

1 INTRODUCTION

Generative diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019) have established a new state of the art in high-dimensional data generation, achieving unprecedented success across modalities ranging from image (Rombach et al., 2021) and video generation (Liu et al., 2024) to applications in the natural sciences (Price et al., 2025; Li et al., 2024b). Rooted in the principles of out-of-equilibrium statistical mechanics, these models operate via two stochastic processes: a *forward process* that progressively degrades the data into white noise, and a reverse-time generative process, called the *backward process*, which generates new samples by simulating a diffusion process driven by a drift vector field, called the score function (Hyvärinen, 2005; Vincent, 2011), defined precisely as the gradient of the logarithm of the noisy data marginals. Recently, this paradigm has been unified under the broader framework of stochastic interpolants (Albergo et al., 2025; Lipman et al., 2023; Liu et al., 2022), where the noising process can be arbitrary and the generative dynamics can be recast as a deterministic flow.

Despite their empirical success, we still lack a rigorous theoretical understanding of how the training dynamics finds a score that captures the intrinsic structure of the target distribution. In particular, it remains unclear how highly overparametrized models of the score generate novel, high-quality samples without simply memorizing the finite training set. One prominent line of explanation is implicit dynamical regularization (Bonnaire et al., 2026; Favero et al., 2025). The learning dynamics is governed by a strict separation of timescales: a fast timescale, τ_{gen}, during which the model learns the underlying structural features required to generate high-quality samples, and a much longer timescale, τ_{mem}, at which the network memorizes the dataset. Here, we identify the mechanism

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

behind this separation of timescales, and the object that reveals it: the spectrum of the Gram matrix of the Neural Tangent Kernel (NTK).

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

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

where $\Delta_t = 1 - e^{-2t}$ and the expectation is over $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_d)$. Its global minimizer is the *empirical score*, the score of the noised empirical distribution (Pidstrigach, 2022; Biroli et al., 2024; Li et al., 2024a), which inevitably collapses the backward trajectories onto the training samples unless n grows exponentially with the intrinsic data dimension (Biroli et al., 2024; Achilli et al., 2024).

The expectation in (1) is never performed exactly. Training proceeds by stochastic gradient descent over batches of the n samples, drawing a fresh noise $\boldsymbol{\xi}$ each time a sample is visited. Each sample is thus seen with m different noise realizations, m being the number of epochs. For the sake of our analysis, we replace the expectation by a finite sum. This will be useful to make comparisons with standard supervised learning, corresponding to $m = 1$, and allows us to recover (1) taking $m \to \infty$. We attach to each sample $\mathbf{x}^\nu$ a set of m noise realizations $\{\boldsymbol{\xi}^{\nu\alpha}\}_{\alpha=1,\dots,m}$, drawn once and then *fixed* throughout training, and consider the empirical loss

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

While this resembles a regression problem, it is a non-standard one: the score is evaluated on $N = nm$ points $\mathbf{Y}^{\nu\alpha}$ that are *correlated*, forming clusters of m points around each clean sample $e^{-t}\mathbf{x}^\nu$. The number of repetitions m is the parameter that controls this cluster structure, and it interpolates between two problems that are usually treated separately: at $m = 1$ every training input carries its own independent noise, the clusters are single points, and (2) is an ordinary supervised kernel regression on n examples with i.i.d. label noise; as $m \to \infty$, instead, the within-cluster average reproduces the exact noise expectation, $\mathcal{L}_{\text{train}} \to \mathcal{L}_{\text{DSM}}$, and the minimizer is the empirical score. Generalization is measured by the standard test loss

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

where the expectation is taken over both fresh noises and fresh samples. We further decompose the test loss over the realizations of the learned score as $\mathcal{L}_{\text{test}} = \frac{1}{2}\left(C_t + \mathcal{B}^2 + \mathcal{V}\right)$, with a bias term $\mathcal{B}^2$ and a variance term $\mathcal{V}$ given by

$$\mathcal{B}^2 = \frac{1}{d}\, \mathbb{E}_{\mathbf{y}} \left[\|\mathbf{s}_{\text{exact}}(\mathbf{y}) - \langle \mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}(\mathbf{y}) \rangle \|^2 \right], \quad \mathcal{V} = \frac{1}{d}\, \mathbb{E}_{\mathbf{y}} \left[\left\langle \|\mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}(\mathbf{y}) - \langle \mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}(\mathbf{y}) \rangle \|^2 \right\rangle \right], \tag{4}$$

where $\mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}$ denotes the minimizer of the training loss $\mathcal{L}_{\text{train}}$ for a given dataset $\mathcal{D}$ and initialization $\boldsymbol{\theta}_0$ of the parameters. Here, $\mathbb{E}_{\mathbf{y}}$ is the average over test samples $\mathbf{y} \sim P_t$, whereas $\langle\cdot\rangle$ denotes the average over the realizations of the training set, and of other sources of randomness such as $\boldsymbol{\theta}_0$. The constant C_t depends only on the data distribution and is independent of the learned score (see Appendix D.8).

In practice the score is a neural network $\mathbf{s}_{\boldsymbol{\theta}}$ and (2) is minimized by variants of gradient descent (Robbins & Monro, 1951; Kingma & Ba, 2015). Here we focus on the gradient flow $\mathrm{d}\boldsymbol{\theta}/\mathrm{d}\tau =$

[1] Fixing t is a standard simplification in theoretical works (Cui et al., 2024; George et al., 2025a). Each noise level defines a separate regression problem, and the memorization transition we study is a small-t phenomenon; Sect. 3 reports a sweep over t. For simplicity, we drop the time argument t of the score.

$-d^2\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{train}}(\boldsymbol{\theta})$ of the training loss (2); the factor d^2 is chosen so that the high-dimensional limit is well defined. We shall occasionally add to the loss (2) an L_2 penalty $\frac{\gamma}{2nmd}\|\boldsymbol{\theta}\|^2$ of strength γ; as shown in Appendix A.4.1 this amounts to shifting the Gram matrix introduced below by $\gamma\mathbf{I}_N$. The relevant scale for that ridge is nm/d, and we will use the rescaled ridge $\tilde{\gamma} = \gamma\, d/(nm)$. The corresponding evolution of the function $\mathbf{s}_{\boldsymbol{\theta}}$ reads

$$\dot{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{y}) = -d^2\sum_{\nu,\alpha} K_\tau(\mathbf{y},\mathbf{Y}^{\nu\alpha})\nabla_{\mathbf{s}}\mathcal{L}_{\text{train}}(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha})) = \frac{-d}{nm}\sum_{\nu,\alpha} K_\tau(\mathbf{y},\mathbf{Y}^{\nu\alpha})\left(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha}) + \frac{\boldsymbol{\xi}^{\nu\alpha}}{\sqrt{\Delta_t}}\right),$$

for any test point $\mathbf{y}\in\mathbb{R}^d$, where $K_\tau(\mathbf{x},\mathbf{y}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})\cdot\nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{y})$ is the neural tangent kernel (NTK) of the model (Jacot et al., 2018; Chizat et al., 2020; Geiger et al., 2020). Taken at pairs of training points, the NTK defines the $N\times N$ **Gram matrix** $\mathbf{G}^{\nu\alpha,\mu\beta} = K_\tau(\mathbf{Y}^{\nu\alpha},\mathbf{Y}^{\mu\beta})$, which closes the dynamics on the training set and hence dictates the whole trajectory.[2] It is therefore natural to expect the distinct timescales of generalization and memorization to be encoded in its spectrum (Yao et al., 2007; Canatar et al., 2021; Bonnaire et al., 2026; Favero et al., 2025), which makes it possible to probe the mechanism underpinning memorization.

Main contributions. The picture supported by our results is that the Gram matrix of a diffusion model is a different object from the one encountered in standard supervised learning. In the latter case each training point carries its own independent input, whereas in denoising score matching each clean sample is presented with $m>1$ independent noise realizations, so the Gram matrix is built on clusters of correlated points. This repetition leads to new features. It restructures the spectrum of the Gram matrix into two well-separated parts: a *generalization* part of large eigenvalues, already present in standard supervised learning, whose modes encode the global features of the target distribution; and a new *memorization* part, made of the smallest eigenvalues and absent at $m=1$, whose modes encode the sample-specific noise directions. Since the training dynamics learns spectral modes in order of decreasing eigenvalue, this gap translates directly into the separation of timescales between τ_{gen} and τ_{mem}. We establish this picture analytically and numerically.

- **Theoretical characterization of the spectral hierarchy.** We provide a precise characterization of the spectral properties of the Gram matrix in the high-dimensional limit $d\gg 1$, in the lazy regime (Chizat et al., 2020), covering both the linear ($n\asymp d$) and polynomial ($n\asymp d^k$) sample complexity regimes (see Fig. 1). In both, the generalization part is associated with the timescales over which the model captures the global features of the target distribution by learning successive approximations of the score function, while the memorization part leads to a timescale that is parametrically larger in the training set size n, effectively delaying the onset of memorization.
- **Bias–variance decomposition.** In the linear regime, we derive closed-form equations for the bias and the variance of the score estimator in the large-d limit. We find that, contrary to the supervised learning case $m=1$, in diffusion models where each sample is noised $m>1$ times, the behavior of the bias changes: the contribution of the memorization part leads to a $\Theta(1)$ bias instead of vanishing at large sample complexity.
- **Validation on realistic settings.** We validate our predictions on CelebA images using Convolutional NTKs in the lazy regime and finite-width U-Net architectures. We show that the two-part structure persists beyond the lazy regime. To assess the role of the two parts of the spectrum, we show that the truncation rank r of the Gram matrix is a control knob for the generalization–memorization transition, and that a targeted L_2 regularization of the memorization part suppresses memorization in feature-learning U-Nets, establishing a direct link between this spectral component and memorization.

Related work. State-of-the-art image models are known to reproduce a non-negligible part of their training data (Carlini et al., 2023; Somepalli et al., 2023a;b). Further works examined how this phenomenon is influenced by factors such as data distribution, model architecture, and training procedure (Gu et al., 2025; Yoon et al., 2023; Bonnaire et al., 2026; Favero et al., 2025), as well as which samples are memorized first (Merger & Goldt, 2026b). The emergence of memorization in the reverse generative trajectories has been analyzed under the exact empirical score hypothesis by Biroli et al. (2024); Achilli et al. (2024); Ventura et al. (2025), who showed that avoiding collapse onto the

[2] In practice, diffusion models are trained with fresh noise at every step, i.e. online in the noise and offline in the clean data. The training dynamics are then governed by the NTK acting as an integral operator on the empirical noisy distribution $\hat{P}_t = \frac{1}{n}\sum_{\nu=1}^n \mathcal{N}(e^{-t}\mathbf{x}^\nu, \Delta_t\mathbf{I}_d)$ (Jacot et al., 2018), a setting recovered by the large-m limit of our results.

training data requires the sample complexity to scale exponentially with the intrinsic data dimension (Achilli et al., 2025; George et al., 2025b). To explain why diffusion models often avoid this collapse in practice, recent literature has uncovered several sources of regularization. It has been shown that architectural biases and limited network capacity constrain memorization (Kamb & Ganguli, 2025; Kadkhodaie et al., 2024; Baptista et al., 2025; Buchanan et al., 2026), while a finite learning rate prevents the network from collapsing to the exact empirical score (Wu et al., 2025). Closely related to our setting, several works have highlighted the beneficial role of early stopping in preventing memorization (Li et al., 2025; Baptista et al., 2025; Bonnaire et al., 2026; Favero et al., 2025), even though it may be insufficient (Garnier-Brun et al., 2026). The performance of neural networks in the lazy regime has been studied rigorously in the supervised learning setting (Ghorbani et al., 2021; Mei et al., 2022; Misiakiewicz, 2022). These analyses rely on the spectrum of the Gram matrix associated with the NTK operator. In the polynomial regime $n \asymp d^k$, the non-linear Gram matrix is equivalent to a deterministic polynomial matrix whose spectrum can be analyzed rigorously (El Karoui, 2010; Lu & Yau, 2022; Pandit et al., 2025). Recent works have begun to study how this picture changes in the diffusion setting. Closest to ours, Latourelle-Vigeant et al. (2026) analyze denoising score matching in the lazy regime and also find a transition from generalization to memorization. They work, however, with the exact expectation over the noise and with the infinite-dimensional kernel operator rather than the Gram matrix. Han et al. (2024) instead leverage these results to derive bounds on the training dynamics of two-layer neural networks on the score-matching task. An extended discussion can be found in Appendix C.

2 Analytical results

Notation and assumptions. We write $N = nm$ for the cardinality of the training set. $\mathbf{I}_d$ is the identity in dimension d and $\mathbf{1}_m$ the all-ones vector in dimension m (likewise $\mathbf{I}_N$ and $\mathbf{1}_N$), and $\mathbf{B}_m = \mathbf{I}_n \otimes \mathbf{1}_m \mathbf{1}_m^\top$. We set $\mathbf{\Sigma} = \mathbb{E}_{P_0}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d\times d}$ for the data covariance and $\mathbf{v}_\lambda$ for its eigenvector associated with the eigenvalue λ. We assume that the data are of the form $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{z}$, where $\mathbf{z}$ has independent zero-mean, unit-variance, sub-Gaussian entries, with a covariance $\mathbf{\Sigma}$ such that $\lambda_{\max}(\mathbf{\Sigma}) = O(1)$ and $\mathrm{Tr}(\mathbf{\Sigma})/d$ converges to a constant denoted by σ^2 as $d \to \infty$. We focus on inner-product kernels of the form $K(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}^\top \mathbf{y}/d)$, where f is a smooth function. A canonical example is provided by the NTK of a multilayer neural network in the infinite-width limit, given in Appendix A.

Lazy regime. In the following we consider the infinite-width limit, with the initialization scaling for which the network enters the *lazy* regime (Jacot et al., 2018; Chizat et al., 2020; Geiger et al., 2020): the weights barely move, the NTK freezes to a deterministic kernel $K_\tau = K$, and the dynamics above becomes linear. Assuming that $\mathbf{s}_{\boldsymbol{\theta}(0)} = 0$ at initialization, it integrates to (see Appendix A.4)

$$\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{x}) = K(\mathbf{x}, \mathbf{Y})^\top \mathbf{G}^{-1} \left(\mathbf{I}_N - e^{-\frac{d\tau}{nm}\mathbf{G}}\right) \left(-\frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right), \tag{5}$$

where $K(\mathbf{x}, \mathbf{Y}) \in \mathbb{R}^N$ collects the kernel evaluations between the test point $\mathbf{x}$ and the training set. The filter $\mathbf{I}_N - e^{-d\tau\mathbf{G}/nm}$ activates spectral modes in decreasing eigenvalue order, so fixing a training time τ is equivalent to imposing a spectral cutoff $\lambda_c \sim nm/(d\tau)$. A bulk of eigenvalues at λ is therefore learned on the timescale $\tau_\lambda \sim nm/(d\lambda)$, and a gap in the spectrum becomes a separation of timescales.

2.1 Linear regime: structure of the Gram matrix

Linear equivalent. Following El Karoui (2010), we extend the linear equivalence of non-linear Gram matrices to correlated data points $\{\mathbf{Y}^{\nu\alpha}\}_{1\le\nu\le n}^{1\le\alpha\le m}$. As $n, d \to \infty$ with $n/d \to \psi_n$ and $m = O(1)$, the Gram matrix $\mathbf{G}^{\nu\alpha,\mu\beta} = f(\mathbf{Y}^{\nu\alpha\top}\mathbf{Y}^{\mu\beta}/d)$ is spectrally equivalent to

$$\mathbf{G}_{\mathrm{lin}} = \mu_I \mathbf{I}_N + \mu_B \mathbf{B}_m + \mu_0 \mathbf{1}_N \mathbf{1}_N^\top + \mu_1 \frac{\mathbf{Y}^\top \mathbf{Y}}{d}, \tag{6}$$

with more details given in Appendix D.1. Crucially, in comparison to the standard case $m = 1$, the noisy repetition of clean data points induces a novel regularization term $\mu_B \mathbf{B}_m$. This term accounts for the intra-sample correlations across different noise realizations and is the primary driver of the spectral separation between generalization and memorization. Figure 1 (*middle*) shows, at $m = 4$, the convergence of the spectrum of the non-linear Gram matrix (colored continuous lines) to the linear equivalent (6) (black dashed line) as d grows.

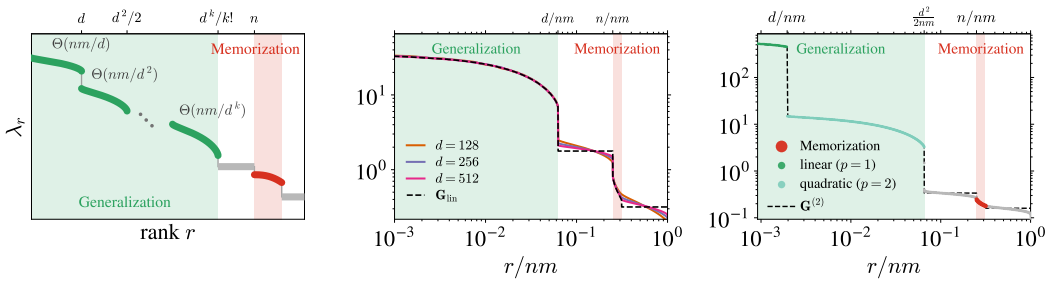

Figure 1: **Gram matrix spectrum.** Eigenvalues λ_r in decreasing order against their rank. (*Left*) Sketch of the spectrum predicted by Theorem 2.3 for the Gram matrix $\mathbf{G}^{\nu\alpha,\mu\beta} = f(\mathbf{Y}^{\nu\alpha\top}\mathbf{Y}^{\mu\beta}/d)$. (*Middle*) Spectrum of $\mathbf{G}$ with isotropic $\mathbf{\Sigma} = \sigma^2\mathbf{I}_d$, $t = 0.1$, $m = 4$, $\psi_n = 4$ and f the NTK of a two-layer ReLU network (Appendix A.8) in the linear regime at several values of d; the dashed black curve is the linear equivalent $\mathbf{G}_{\text{lin}}$ of (6) at $d = 512$. (*Right*) Quadratic regime $n = 2d^2$ at $d = 64$; the dashed black curve is the polynomial equivalent of Theorem 2.3. Details in Appendix D.7.

Spectral properties of the Gram matrix. We now characterize the spectral properties of the linearized Gram matrix. The term $\mu_0\mathbf{1}_N\mathbf{1}_N^\top$ induces a single outlier eigenvalue at $nm\mu_0$, which is not relevant for our results and will not be discussed further. The remainder of the spectrum is determined by the Stieltjes transform $q(z) = \frac{1}{N}\operatorname{Tr}(\mathbf{G} - z\mathbf{I}_N)^{-1}$. Using the replica method (Mézard et al., 1987), we derive a set of self-consistent equations for $q(z)$ in Appendix D.5. They allow us to characterize the spectrum of the Gram matrix in the high-dimensional regime.

Theorem 2.1 (Spectrum of the Gram matrix)**.** *Let the empirical spectral density of $\mathbf{\Sigma}$ converge to $\rho_{\mathbf{\Sigma}}$, and take $d, n \to \infty$ at fixed $\psi_n = n/d$, then $\psi_n \gg 1$. For each eigenvalue λ of $\mathbf{\Sigma}$, let $\lambda_t = e^{-2t}\lambda + \Delta_t$. The spectrum of $\mathbf{G} - \mu_0\mathbf{1}_N\mathbf{1}_N^\top$ consists of two bulks of d eigenvalues each:*

- ***generalization bulk:*** $\lambda_{\text{gen}} \simeq \mu_1\psi_n m\lambda_t = \Theta(nm/d)$,
- ***memorization bulk:*** $\lambda_{\text{mem}} \simeq \mu_I + \mu_B(m-1)\Delta_t/\lambda_t = \Theta(m)$,

and two atoms: μ_I with weight $(m - 1 - 1/\psi_n)/m$ and $\bar{\mu} = \mu_I + m\mu_B$ with weight $(1 - 1/\psi_n)/m$.

While the generalization bulk is present in standard kernels ($m = 1$), the multi-noise setting ($m > 1$) triggers the emergence of the memorization bulk, cf. Fig. 1. To understand why these components correspond to distinct learning regimes, we must first examine their associated eigenvectors.

Proposition 2.1 (Eigenvectors of the Gram matrix)**.** *For $\psi_n \gg 1$ and every eigenvalue λ of $\mathbf{\Sigma}$ with associated eigenvector $\mathbf{v}_\lambda$, the eigenvectors of $\mathbf{G}$ associated with the two bulks of Theorem 2.1 are asymptotically*

$$\mathbf{u}_1^{\nu\alpha} \propto \mathbf{v}_\lambda^\top\mathbf{Y}^{\nu\alpha}, \qquad \mathbf{u}_2^{\nu\alpha} \propto \mathbf{v}_\lambda^\top\left[(me^{-2t}\lambda + \Delta_t)\sqrt{\Delta_t}\,\boldsymbol{\xi}_\perp^{\nu\alpha} - (m-1)\Delta_t\,\bar{\mathbf{Y}}^\nu\right], \tag{7}$$

with $\bar{\mathbf{Y}}^\nu = \frac{1}{m}\sum_\beta \mathbf{Y}^{\nu\beta}$ and $\boldsymbol{\xi}_\perp^{\nu\alpha} = \boldsymbol{\xi}^{\nu\alpha} - \frac{1}{m}\sum_\beta \boldsymbol{\xi}^{\nu\beta}$. The eigenvectors of the two atoms are the $\boldsymbol{\varphi} \otimes \mathbf{1}_m/\sqrt{m}$ with $\boldsymbol{\varphi} \in \operatorname{Ker}(\bar{\mathbf{Y}})$ and the elements of $\operatorname{Ker}(\mathbf{Y}) \cap \operatorname{Ker}(\mathbf{B}_m)$; they do not contribute to the estimated score.

Theorem 2.1 and Proposition 2.1 reveal a spectral hierarchy. The eigenvectors of the two bulks have a transparent interpretation. The generalization modes $\mathbf{u}_1$ are the projections of the training points on the principal directions $\mathbf{v}_\lambda$ of the data covariance, so learning them fits the component of the score carried by the population covariance, exactly as a standard kernel does at $m = 1$. The memorization modes $\mathbf{u}_2$ contain $\boldsymbol{\xi}_\perp$, the noise components that distinguish the different realizations of the same data point, capturing the irregularity of the empirical score close to each data point. These two bulks are separated by a wide spectral gap: the *generalization* bulk scales as $\Theta(nm/d)$, while the *memorization* bulk remains $\Theta(m)$. Because of the filtering of Eq. (5), this gap enforces a strict separation of timescales: the generalization modes are learned first whereas the memorization ones are learned much later, on a timescale that is parametrically larger in n.

2.2 Linear regime: bias–variance decomposition

To get a sharper understanding of the phenomenon we now turn to the bias–variance decomposition of the estimator of the score. In the high-dimensional limit, we derive closed-form equations for the

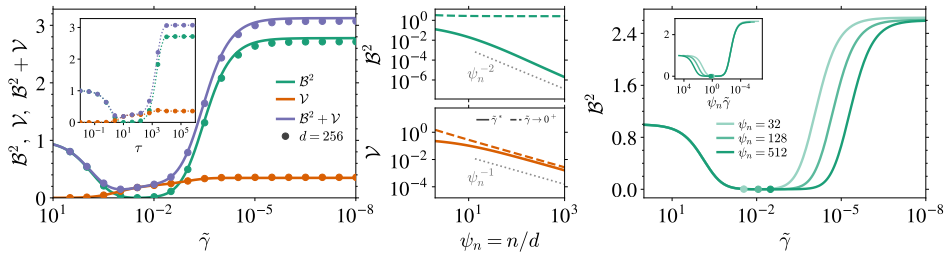

Figure 2: **Bias–variance decomposition.** $\mathcal{B}^2$, $\mathcal{V}$ and $\mathcal{B}^2+\mathcal{V}$ for the kernel $f(u)=\mathbb{E}[\tanh(z_1)\tanh(z_2)]$, with (z_1,z_2) unit-variance Gaussians of correlation u (Appendix A.8), at $\sigma^2=1$, $m=4$ and $t=0.1$. (*Left*) Vs. the rescaled ridge $\tilde{\gamma}$ at $\psi_n=8$, axis reversed to show the correspondence with training time; the *inset* shows the same vs. the training time τ of (5). Solid lines: Theorem 2.2; markers: experiments at $d=256$; in the inset, markers and dotted line are the finite-d gradient flow. (*Middle*) $\mathcal{B}^2$ (top) and $\mathcal{V}$ (bottom) vs. ψ_n, for the optimal ridge $\tilde{\gamma}^*=\arg\min_{\tilde{\gamma}}\mathcal{L}_{\text{test}}$ (solid) and the ridgeless limit $\tilde{\gamma}\to 0^+$ (dashed). (*Right*) $\mathcal{B}^2$ vs. $\tilde{\gamma}$ for several ψ_n, dots marking $\tilde{\gamma}^*$; the *inset* shows the same vs. $\psi_n\tilde{\gamma}$. Details in Appendix D.10.

bias and the variance of the kernel ridge predictor,

$$\mathbf{s}_\gamma(\mathbf{y}) = K(\mathbf{y},\mathbf{Y})^\top\left(\mathbf{G}+\gamma\,\mathbf{I}_N\right)^{-1}\left(-\frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right), \tag{8}$$

where $\mathbf{Y}$ and $\boldsymbol{\xi}$ are the noisy data and noises used during training.

Theorem 2.2 (Bias and variance of the kernel ridge predictor)**.** *In the limit $n,d\to\infty$ at fixed $\psi_n=n/d$, with $m=O(1)$ and $\boldsymbol{\Sigma}=\sigma^2\mathbf{I}_d$, the bias and the variance (4) of the kernel ridge predictor (8) concentrate, and are given by a closed system of algebraic equations stated in Appendix D.9.*

We use the ridge regularization as a proxy for the gradient flow dynamics in τ, studying the evolution of the equations of Theorem 2.2 as a function of $\tilde{\gamma}$ through the equivalence $\tilde{\gamma}\Leftrightarrow 1/\tau$ (Ali et al., 2019). As shown in the left panel of Fig. 2, both ridge and gradient flow show qualitatively the same evolution; we therefore study the closed-form equations on the kernel ridge estimator as a function of $\tilde{\gamma}$ and infer the training-time behavior from it. Solving these equations numerically gives the scalings of the bias and the variance as a function of ψ_n and t that can be read off Fig. 2 (*middle*) and Fig. 11.

Result 2.1 (Scalings of the bias and the variance)**.** *Let $m>1$ and $1/(m\psi_n)\ll t\ll 1$.*

- Ridgeless, *$\tilde{\gamma}\to 0^+$: $\mathcal{B}^2=\Theta(t^{-2})$, independently of ψ_n, and $\mathcal{V}=\Theta(\psi_n^{-1}t^{-2})$.*
- Optimal ridge, *$\tilde{\gamma}^*$: $\mathcal{B}^2=\Theta(\psi_n^{-2}t^{-2})$ and $\mathcal{V}=\Theta(\psi_n^{-1}t^{-1})$.*

At the end of training, i.e. $\tilde{\gamma}\to 0^+$, the test loss is dominated by a persistent bias[3] that does not vanish with the sample complexity and that diverges as the diffusion time $t\to 0$. On the other hand, when the model is optimally early-stopped (equivalently, at the optimal ridge $\tilde{\gamma}^*$), the bias vanishes with the sample complexity at fixed t, like the variance, as illustrated in Fig. 2 (*middle*). In the right panel of Fig. 2, the bias starts to rise at $\tilde{\gamma}\propto 1/\psi_n$ (*inset*), i.e. when the ridge γ falls to the $O(1)$ eigenvalues of the memorization bulk. Below this value the ridge is too small to damp these modes: the regularization no longer acts on the memorization bulk, which is the source of the persistent bias.

2.3 Polynomial scaling

These results can be extended beyond the linear regime under a few additional assumptions (see Appendix D.7). In the polynomial regime $d^k\ll n\ll d^{k+1}$, the analogue of the linear equivalent (6) is a polynomial equivalent, with the spectrum below, sketched in Fig. 1 (*left*) and shown at finite d in Fig. 1 (*right*).

Theorem 2.3 (Spectrum of the Gram matrix, polynomial regime)**.** *In the limit $d,n\to\infty$ with $d^k\ll n\ll d^{k+1}$ and $m=O(1)$, the spectrum of $\mathbf{G}$ consists, besides two atoms (see Appendix D.7), of:*

[3] There is a fundamental difference between the statistics of the score estimator on a test set and those of the generated distribution by integrating the backward process. At large m, the latter is the empirical one: it averages to the population distribution, hence has zero bias, but depends strongly on the training set, hence has large variance.

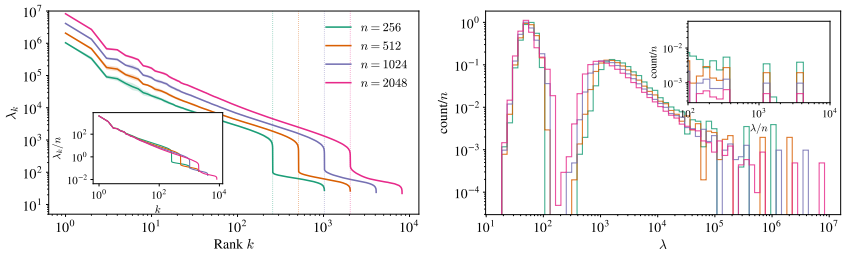

Figure 3: **Structure of the generalization and memorization bulks in the CNTK spectrum on CelebA.** *(Left)* Ordered eigenvalues for several n at $m = 4$. The dotted lines indicate $r = n$. Inset: Same, rescaled by n. *(Right)* Eigenvalue histogram with counts rescaled by n. Inset: collapse of the top eigenvalues under rescaling by n. $s = 0.05$, averaged over 10 training sets.

- *for each degree* $p \in \{1, \dots, k\}$*, a group of* $\Theta(d^p/p!)$ *eigenvalues* $\Theta(nm/d^p)$ (***generalization** sub-bulks);*
- *a bulk of* $\Theta(d^k/k!)$ *eigenvalues* $\Theta(m)$*, independent of* n *and absent at* $m = 1$ (***memorization** bulk).*

See Appendix D.7 for the precise statement. Since a mode of eigenvalue λ is learned on the timescale $\tau_\lambda \sim nm/(d\lambda)$, the generalization sub-bulks are learned degree by degree, on timescales $\tau_p \sim d^{p-1}$: at $m = 1$ this is the classical hierarchical picture of kernel regression (Ghorbani et al., 2021), in which low-order moments of the data are learned first and higher-order statistics later (Ricci et al., 2025; Bardone et al., 2026). Data repetition ($m > 1$) preserves this hierarchy but adds the memorization bulk at $\Theta(m)$, learned on $\tau_{\text{mem}} \sim n/d$. As in the linear regime, memorization is delayed relative to every degree the model can learn, by a factor that grows with the sample complexity. Note that, strictly speaking, full memorization—the ability of the learned score to reproduce the training samples—requires n to grow faster than any power of d, i.e. $n \gg d^k$ for every fixed k. At any finite k only an approximation of the empirical score can be learned, and memorization and overfitting are correspondingly partial (Misiakiewicz, 2022; Latourelle-Vigeant et al., 2026).

3 Experiments

Data, noising process, and model. We work with the CelebA face dataset (Liu et al., 2015), which we convert to grayscale, downsample to 32×32 pixels, and standardize. For computational convenience, our experiments adopt a flow-matching parametrization (Liu et al., 2022; Lipman et al., 2023; Albergo et al., 2025) in which the noising process is the linear interpolant[4] $\mathbf{Y}_s^{\nu\alpha} = (1 - s)\,\mathbf{x}^\nu + s\,\boldsymbol{\xi}^{\nu\alpha}$, with $s \in [0, 1]$, between a clean data point $\mathbf{x}^\nu \sim P_0$ and an independent Gaussian noise $\boldsymbol{\xi}^{\nu\alpha} \sim \mathcal{N}(0, \mathbf{I}_d)$. The regression targets are the conditional velocities $\boldsymbol{\xi}^{\nu\alpha} - \mathbf{x}^\nu$. Throughout the section, we use n clean images and m independent noise realizations per data point such that the empirical Gram matrix $\mathbf{G}^{\nu\alpha,\mu\beta} = K(\mathbf{Y}_s^{\nu\alpha}, \mathbf{Y}_s^{\mu\beta})$ has size $N \times N$ with $N = nm$. We take the velocity model to be a depth-D vanilla convolutional neural network (CNN) with 3×3 filters and ReLU activations, and a final readout layer. Our choice is motivated by the existence of a closed-form Convolutional Neural Tangent Kernel (CNTK) in the infinite-width (lazy) limit (Arora et al., 2019). This allows us to replace the network by its kernel predictor, whose Gram matrix is computed in closed form rather than from parameter gradients (see Appendix A.7). We then test whether these spectral properties persist in the feature-learning regime by training a finite-width U-Net in Sect. 3.2.

Gram-matrix eigendecomposition and truncated kernel regression. At every flow time s, we approximate the top-r eigenpairs of the Gram matrix $\mathbf{G}$ via the randomized method of Halko et al. (2011) in which, for memory efficiency, the full Gram matrix is never materialized. Keeping only the top r modes of $\mathbf{G}$ defines the spectrally truncated estimator $\hat{\mathbf{v}}_r$ of the velocity field $\mathbb{E}[\boldsymbol{\xi} - \mathbf{x}_0 \mid \mathbf{Y}_s = \mathbf{x}]$, and approximately corresponds to early stopping of training in the lazy regime at $\tau \sim nm/(d\lambda_r)$, cf. Eq. (5). More details can be found in Appendix E.2.

[4]This differs from the Ornstein–Uhlenbeck interpolant of Sect. 2 only by a time reparametrization and a scaling of the input and the target, so that the analytical results remain unchanged (see Appendix B).

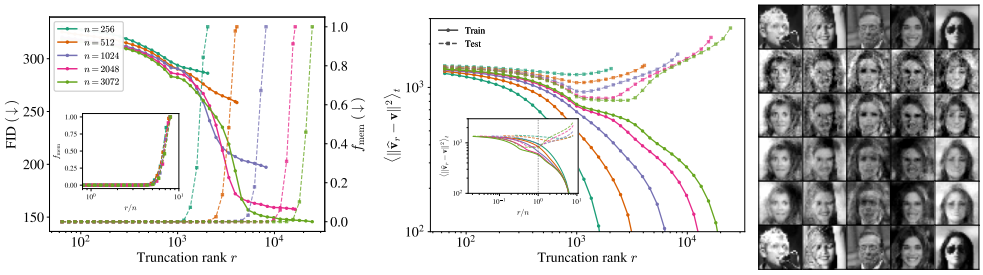

Figure 4: **Spectrally truncated kernel regression and generalization–memorization transition.** *(Left)* FID (solid) and f_{mem} (dashed) vs. truncation rank r for several n. Inset: f_{mem} vs. r/n. *(Middle)* Time-averaged L_2 training (solid) and test (dashed) losses; the inset shows the same curves against r/n. *(Right)* Top row: training images. Rows 2–6: generated samples ($n = 3072$) starting from a fixed training noise associated with the top-row clean image at $r \in \{2048,\ 4096,\ 8192,\ 16384,\ N = nm\}$.

Metrics. To quantify the ability of the approximated velocity field to memorize, we adopt the first-to-second nearest-neighbor ratio criterion used in several previous studies (Yoon et al., 2023; Gu et al., 2025; Bonnaire et al., 2026) and compute the fraction $f_{\mathrm{mem}} \in [0, 1]$ of samples $\tilde{\mathbf{x}}_0$ generated by the backward dynamics that are memorized. We initialize the trajectories at $\tilde{\mathbf{x}}_1 \in \{\boldsymbol{\xi}^{\mu\alpha}\}$; our experiment can therefore be viewed as a long-time version of the U-turn protocol of Sclocchi et al. (2024); Behjoo & Chertkov (2025). As a measure of the quality of the generated images, we also report the Fréchet Inception Distance (FID, Heusel et al., 2017) between 2,048 samples generated from fresh noise and 2,048 held-out test images. See Appendix E.1 for more details.

3.1 Lazy regime: Gram matrix for real data

Structure of the Gram matrix spectrum in CelebA. Figure 3 displays the spectrum of the CNTK Gram matrix computed from noised images of CelebA at fixed $s = 0.05$. It exhibits the two-bulk structure predicted by Theorem 2.1, including a component absent in the standard $m = 1$ case (see Appendix E.2 for the scaling with m). The first bulk of large eigenvalues behaves like that of a classical Gram matrix: at fixed m, the top eigenvalues scale linearly with n and collapse once rescaled (see insets). This generalization bulk contains n eigenvalues rather than the d predicted for sub-Gaussian data in the proportional regime, a known feature of image statistics (Canatar et al., 2021). This likely reflects the strong anisotropy and low intrinsic dimension of natural images, whose covariance spectrum decays as a power law. For such data, the bulks of Theorem 2.3 are not well separated and can merge (Wortsman & Loureiro, 2025). The key qualitative phenomenon nonetheless persists: the spectrum is partitioned into a first bulk scaling with n, and a second which is independent of it. We show below that they correspond to generalization and memorization, respectively.

Generalization–memorization transition in truncated kernel regression. Figure 4 varies the truncation rank r of the CNTK Gram matrix on CelebA for $n \in \{256, 512, 1024, 2048, 3072\}$ and $m = 8$. At small r, the FID (left panel) of samples generated with the truncated velocity field $\hat{\mathbf{v}}_r$ decreases steadily while f_{mem} (in inset) stays near zero. For small n, however, f_{mem} rises before the FID reaches its minimum: the estimator starts memorizing training points before it produces plausible images. As n grows, the FID drops at roughly the same r but reaches lower minima, saturating beyond $n = 2048$. In this regime, the rank-r approximated velocity generates CelebA-like samples without copying any training point over a growing range of r. The threshold marking the onset of generalization is therefore essentially independent of n, whereas the memorization threshold scales linearly with n, as shown by the inset, where rescaling r by n makes all curves collapse. The train and test losses (middle panel) exhibit the same pattern: the training error decreases monotonically with r, while the test error develops a clean minimum around $r \approx n$ for n sufficiently large, in agreement with Sect. 2. The right panel illustrates this transition at fixed $n = 3072$: integrating the trajectories backward yields blurry reconstructions at small r, which sharpen as r grows. Across an entire intermediate regime, the model produces *new* samples, that differ from their paired training images (top row) in gender, accessories and hairstyles. At larger r, these features are progressively morphed into those of the associated training image, until eventually, at $r = N = nm$ (last row), every column reproduces its training image. These three diagnostics agree on a transition from generalization to memorization as r increases, confirming the role of the successive bulks of eigenvalues in the CNTK Gram matrix in separating

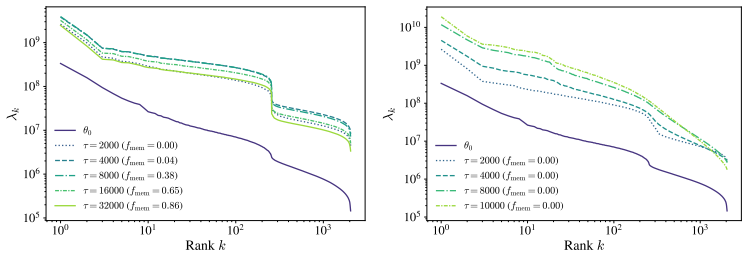

Figure 5: **Training-time-dependent empirical Gram eigenspectrum of a U-Net.** Eigenvalues of the empirical Gram matrix computed for several training times τ at $s = 0.23$ for *(Left)* the unregularized dynamics, and *(Right)* the L_2-regularized dynamics.

the two phases. This behavior is in agreement with the previous findings of Bonnaire et al. (2026): as the two bulks of eigenvalues are learned on different timescales, they create a dynamical separation between a generalization and a memorization phase, whose timescale ratio is proportional to n.

3.2 Feature-learning regime and training-time-dependent NTK

We now investigate whether the spectral properties persist beyond the lazy regime, when the kernel evolves throughout training. To this end, we train a finite-width U-Net in the feature-learning regime on $n = 256$ CelebA images with $m = 8$ and track the eigenvalues of its empirical NTK Gram matrix, shown in the left panel of Fig. 5. At initialization, the spectrum displays the same two-bulk structure as the CNTK, with a gap at $r \approx n$. While eigenvalue amplitudes shift non-monotonically during training, the bulk separation persists and the spectral gap even sharpens for $\tau > 0$. Strikingly, even though training spans the entire generalization–memorization transition (with f_{mem} going from 0 to 0.86), the shape of the spectrum (two bulks with a gap at n) is preserved. This suggests that our analysis and the crucial role of the spectral bias we exhibit extend beyond the lazy regime. To further connect the second bulk to memorization, we perform an L_2-regularized training with a regularization parameter set to the n-th eigenvalue of the Gram matrix, tracked dynamically during training at each flow time s (right panel of Fig. 5). This procedure therefore effectively damps the modes with eigenvalues below λ_n and has two striking consequences: f_{mem} remains exactly zero over 10,000 training steps (compared to ≈ 0.5 with no regularization) without sacrificing the test loss (see Appendix E.3), and the second bulk gets progressively smoothed out during training. This is consistent with memorization being indeed controlled by this second component and provides a spectral interpretation of the findings of Baptista et al. (2025), according to which regularization suppresses memorization in diffusion models.

4 Conclusion and limitations

We analyze the spectrum of the Gram matrix that governs the lazy training dynamics of diffusion models and identify two well-separated bulks: a *generalization* bulk, already present in standard kernel regression, and a *memorization* bulk created by the repetition of each training sample with several noise realizations—specific to diffusion models. The two bulks are separated by a spectral gap that grows with the training set size n, leading to a separation of training timescales. We establish this structure analytically for both linear and polynomial sample complexities and, in the linear regime, derive closed-form equations for the bias and the variance: the generalization bulk yields a bias and a variance that vanish with the sample complexity, whereas the memorization bulk contributes a $\Theta(1)$ bias that has no counterpart in the supervised setting. Experiments on CelebA with the convolutional NTK exhibit the same spectral hierarchy, and a spectral truncation controls the generalization–memorization transition. The two-bulk structure further persists in a finite-width U-Net trained in the feature-learning regime, suggesting that this spectral mechanism is a general property of diffusion models.

Limitations. Our analytical results rely on several assumptions: the lazy regime, a fixed diffusion time t, and a finite number m of noise realizations frozen throughout training. While the experiments strongly suggest that the mechanism extends well beyond these hypotheses, a rigorous treatment of the feature-learning regime remains open. Finally, the gap between the predicted and observed numbers of generalization eigenvalues calls for more realistic data models.

REFERENCES

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

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, jul 2025. doi: 10.1088/1742-5468/ade136. URL https://doi.org/10.1088/1742-5468/ade136.

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL http://jmlr.org/papers/v26/23-1605.html.

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

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

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Francis Bach. Polynomial magic iii: Hermite polynomials. https://francisbach.com/hermite-polynomials/, 2023. Accessed: 2025-10-09.

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

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

Lorenzo Bardone, Claudia Merger, and Sebastian Goldt. A theory of statistics in diffusion models, from easy to hard. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026.

Hamidreza Behjoo and Michael Chertkov. U-turn diffusion. *Entropy*, 27(4):343, March 2025. ISSN 1099-4300. doi: 10.3390/e27040343. URL http://dx.doi.org/10.3390/e27040343.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c4ef9c39b300931b69a36fb3dbb8d60e-Paper.pdf.

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

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*, 2021.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=BSZqpqgqM0.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. *Advances in Neural Information Processing Systems*, 38:96113–96157, 2026.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

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

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020. URL https://arxiv.org/abs/1812.07956.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

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

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

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38 (1):1–50, February 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.

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

Jerome Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mézard, and Luca Saglietti. Biased generalization in diffusion models, 2026. URL https://arxiv.org/abs/2603.03469.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, November 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc4de. URL http://dx.doi.org/10.1088/1742-5468/abc4de.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves, 2025a. URL https://arxiv.org/abs/2502.00336.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Analysis of diffusion models for manifold data. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2025b. doi: 10.1109/ISIT63088.2025.11195641.

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

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

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, Apr 2021. doi: 10.1214/20-AOS2001. URL https://doi.org/10.1214/20-AOS2001.

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, 2021. URL https://arxiv.org/abs/2006.14709.

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

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

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

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

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization, 2026. URL https://arxiv.org/abs/2401.15604v4. Extended version of the ICLR 2024 paper.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

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

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

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

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

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ilpL2qACla.

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

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

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

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models, 2025. URL https://arxiv.org/abs/2311.01797.

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

T. Li, L. Biferale, F. Bonaccorso, and et al. Synthetic lagrangian turbulence by generative diffusion models. *Nat Mach Intell*, 6:393–403, 2024b. doi: 10.1038/s42256-024-00810-0. URL https://doi.org/10.1038/s42256-024-00810-0.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 3730–3738, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.425. URL https://doi.org/10.1109/ICCV.2015.425.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. doi: 10.1214/17-AAP1328.

Yue M. Lu and Horng-Tzer Yau. An Equivalence Principle for the Spectrum of Random Inner-Product Kernel Matrices with Polynomial Scalings. *arXiv e-prints*, art. arXiv:2205.06308, May 2022. doi: 10.48550/arXiv.2205.06308.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, 2020. URL https://arxiv.org/abs/1908.05355.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

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

Claudia Merger and Sebastian Goldt. Local coverage governs memorization in diffusion models, 2026b. URL https://arxiv.org/abs/2606.14390.

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

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

Parthe Pandit, Zhichao Wang, and Yizhe Zhu. Universality of kernel random matrices and kernel regression in the quadratic regime, 2025. URL https://arxiv.org/abs/2408.01062.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AiNrnIrDfD9.

I. Price, A. Sanchez-Gonzalez, F. Alet, and et al. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

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

Fabiola Ricci, Lorenzo Bardone, and Sebastian Goldt. Feature learning from non-gaussian inputs: the case of independent component analysis in high dimensions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kmg7hweySi.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data, 2024.

J.W. Silverstein and Z.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. ISSN 0047-259X. doi: https://doi.org/10.1006/jmva.1995.1051. URL https://www.sciencedirect.com/science/article/pii/S0047259X85710512.

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

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

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

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

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

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KlN00vQEY2.

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

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

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. *Advances in Neural Information Processing Systems*, 38:95865–95963, 2026. URL https://arxiv.org/abs/2503.03206.

Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization, 2025. URL https://arxiv.org/abs/2510.04780.

Yu-Han Wu, Pierre Marion, Gérard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization, 2025.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315, 2007.

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

Appendix

This appendix provides detailed derivations and additional results supporting the main text. Sect. A reviews the Neural Tangent Kernel framework and gives explicit expressions of the kernel for common architectures. Sect. B shows the equivalence between the Ornstein–Uhlenbeck process used in Sect. 2 and the rectified-flow interpolant used in Sect. 3. Sect. C discusses related work in more depth and the differences with this work. Sect. D then contains the proofs of the analytical results of Sect. 2. Finally, Sect. E gives the numerical details of Sect. 3, for both the closed-form CNTK and the finite-width U-Net.

A Neural Tangent Kernel

In this section we present in more detail the lazy regime and the Neural Tangent Kernel.

A.1 Introduction

Consider a parametric family of functions $\mathbf{s}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^{d'}$ with $\boldsymbol{\theta} \in \mathbb{R}^P$ and a training set $\{\mathbf{x}^\nu, \mathbf{y}^\nu\}_\nu$ with $\mathbf{x}^\nu \in \mathbb{R}^d$ and $\mathbf{y}^\nu \in \mathbb{R}^{d'}$. Trained under gradient flow on the square loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_\nu \|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}^\nu) - \mathbf{y}^\nu\|^2$, the parameters evolve as $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$, which induces dynamics in function space

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{x}) \;=\; -\sum_\nu K\left(\boldsymbol{\theta}(\tau);\, \mathbf{x}, \mathbf{x}^\nu\right) \cdot \left(\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{x}^\nu) - \mathbf{y}^\nu\right), \tag{9}$$

where the matrix-valued *Neural Tangent Kernel* is

$$K\left(\boldsymbol{\theta};\, \mathbf{x}, \mathbf{x}'\right)_{k,k'} \;=\; \sum_{i=1}^{P} \partial_{\boldsymbol{\theta}_i}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})_k\, \partial_{\boldsymbol{\theta}_i}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}')_{k'}, \qquad k, k' \in \{1, \dots, d'\}. \tag{10}$$

A priori the kernel (10) depends on the current parameters $\boldsymbol{\theta}(\tau)$ and is therefore a stochastic, time-dependent operator. The crucial observation of Jacot et al. (2018) is that in the infinite-width limit, with the appropriate variance scaling at initialization, the kernel converges to a *deterministic* limit $K^\star(\mathbf{x}, \mathbf{x}')$, and the deviation of the parameters from their initial value vanishes: $\|\boldsymbol{\theta}(\tau) - \boldsymbol{\theta}(0)\| \to 0$. In this lazy regime, the dynamics on $\mathbf{s}_{\boldsymbol{\theta}}$ becomes a linear kernel flow with the constant kernel $K^\star$.

A.2 Lazy regime and linearization

The lazy regime can equivalently be stated as a first-order linearization around the initialization (Chizat et al., 2020):

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}) \;\simeq\; \mathbf{s}_{\boldsymbol{\theta}(0)}(\mathbf{x}) \;+\; \nabla_{\boldsymbol{\theta}}\,\mathbf{s}_{\boldsymbol{\theta}(0)}(\mathbf{x}) \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}(0)\right). \tag{11}$$

The features $\nabla_{\boldsymbol{\theta}}\,\mathbf{s}_{\boldsymbol{\theta}(0)}(\mathbf{x}) \in \mathbb{R}^{d' \times P}$ are random at initialization and *frozen* throughout training. Training the network is equivalent to ridgeless linear regression on these random features, and the associated kernel is exactly the NTK at initialization (10) evaluated at $\boldsymbol{\theta}(0)$.

A.3 Vector-valued outputs and decomposable kernels

For a score-matching task one needs a *vector*-valued output ($d' = d$). For the standard NTK parametrizations of fully-connected and convolutional networks, the matrix-valued NTK has the convenient decomposable form

$$K(\mathbf{x}, \mathbf{x}')_{k,k'} \;=\; K_{\mathrm{scal}}(\mathbf{x}, \mathbf{x}')\, \delta_{k,k'}, \tag{12}$$

i.e. a scalar kernel $K_{\mathrm{scal}}(\mathbf{x}, \mathbf{x}')$ tensored with the identity in output space. This factorization holds whenever the output layer is fully connected with i.i.d. initialization and no output-coupling structure, and it reduces the spectral analysis of the matrix-valued kernel to that of the scalar kernel: the operator K acts diagonally across output coordinates and its eigenvalues are d-fold copies of the eigenvalues of K_{scal}. In what follows, we therefore present only the scalar kernels.

A.4 KERNEL GRADIENT FLOW

We derive here the closed-form solution (5) of the main text. The argument is the standard lazy-regime computation of Jacot et al. (2018), adapted to the DSM loss (13); its only difference is that the regression target is the frozen noise $-\boldsymbol{\xi}/\sqrt{\Delta_t}$ rather than a label.

Recall the training loss (2) at fixed noise level,

$$\mathcal{L}_{\text{train}}(\boldsymbol{\theta}) = \frac{1}{2nmd}\sum_{\nu=1}^{n}\sum_{\alpha=1}^{m}\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha}) + \frac{\boldsymbol{\xi}^{\nu\alpha}}{\sqrt{\Delta_t}}\right\|^2, \qquad \mathbf{Y}^{\nu\alpha} = e^{-t}\mathbf{x}^{\nu} + \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\alpha}, \tag{13}$$

and the gradient flow $\mathrm{d}\boldsymbol{\theta}/\mathrm{d}\tau = -d^2\,\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{train}}(\boldsymbol{\theta})$. Differentiating,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{train}}(\boldsymbol{\theta}) = \frac{1}{nmd}\sum_{\nu,\alpha}\nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha})^{\top}\left(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha}) + \frac{\boldsymbol{\xi}^{\nu\alpha}}{\sqrt{\Delta_t}}\right),$$

so that, by the chain rule, the induced evolution of the function at an arbitrary point $\mathbf{y}\in\mathbb{R}^d$ is

$$\dot{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{y}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{y})\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\frac{d}{nm}\sum_{\nu,\alpha}K_\tau(\mathbf{y},\mathbf{Y}^{\nu\alpha})\left(\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{Y}^{\nu\alpha}) + \frac{\boldsymbol{\xi}^{\nu\alpha}}{\sqrt{\Delta_t}}\right), \tag{14}$$

with $K_\tau(\mathbf{x},\mathbf{y}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x})\cdot\nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{y})$ the NTK at training time τ. The factor d^2 of the learning rate combines with the normalization $1/(nmd)$ of the loss into the rate $d/(nm)$ appearing in (5); with this choice the generalization eigenvalues $\Theta(nm/d)$ are learned on times $\tau = O(1)$.

In the infinite-width (lazy) limit, $K_\tau = K$ is constant. By the decomposable structure (12) the d output components decouple and it suffices to treat each of them separately; we therefore drop the output index and regard all quantities below as vectors in $\mathbb{R}^N$, $N = nm$, indexed by the pair $(\nu\alpha)$.

Let $\mathbf{f}(\tau)\in\mathbb{R}^N$ collect the values of the score on the training points, $\mathbf{f}_{\nu\alpha}(\tau) = \mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{Y}^{\nu\alpha})$, let $\boldsymbol{\Xi}_{\nu\alpha} = \boldsymbol{\xi}^{\nu\alpha}/\sqrt{\Delta_t}$, and let $\mathbf{G}\in\mathbb{R}^{N\times N}$, $\mathbf{G}^{\mu\beta,\nu\alpha} = K(\mathbf{Y}^{\mu\beta},\mathbf{Y}^{\nu\alpha})$, be the Gram matrix. Taking $\mathbf{y} = \mathbf{Y}^{\mu\beta}$ in (14) gives a closed system,

$$\dot{\mathbf{f}} = -\frac{d}{nm}\,\mathbf{G}\,(\mathbf{f}+\boldsymbol{\Xi}). \tag{15}$$

With the initialization $\mathbf{s}_{\boldsymbol{\theta}(0)} = 0$, integrating gives

$$\mathbf{f}(\tau) = -\left(\mathbf{I}_N - e^{-\frac{d\tau}{nm}\mathbf{G}}\right)\boldsymbol{\Xi}. \tag{16}$$

The training loss along the flow is therefore $\mathcal{L}_{\text{train}}(\boldsymbol{\theta}(\tau)) = \frac{1}{2nmd}\,\mathrm{Tr}\big(e^{-\frac{2d\tau}{nm}\mathbf{G}}\,\boldsymbol{\Xi}\boldsymbol{\Xi}^{\top}\big)$, summed over output components.

For a general test point $\mathbf{y}$, integrating from 0 to τ with $\mathbf{s}_{\boldsymbol{\theta}(0)}(\mathbf{y}) = 0$ gives

$$\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{y}) = -\frac{d}{nm}\,K(\mathbf{y},\mathbf{Y})^{\top}\int_0^{\tau} e^{-\frac{d\sigma}{nm}\mathbf{G}}\,\mathrm{d}\sigma\;\boldsymbol{\Xi} \tag{17}$$

$$= -K(\mathbf{y},\mathbf{Y})^{\top}\mathbf{G}^{-1}\left(\mathbf{I}_N - e^{-\frac{d\tau}{nm}\mathbf{G}}\right)\boldsymbol{\Xi} \tag{18}$$

$$= -K(\mathbf{y},\mathbf{Y})^{\top}\varphi_\tau(\mathbf{G})\,\boldsymbol{\Xi}, \tag{19}$$

where $\varphi_\tau(\lambda) = \big(1-e^{-\frac{d\tau}{nm}\lambda}\big)/\lambda$ is the gradient-flow spectral filter and $K(\mathbf{y},\mathbf{Y})\in\mathbb{R}^N$ has entries $K(\mathbf{y},\mathbf{Y}^{\nu\alpha})$. Recalling $\boldsymbol{\Xi} = \boldsymbol{\xi}/\sqrt{\Delta_t}$, this is exactly (5).

Diagonalizing $\mathbf{G} = \sum_\lambda \lambda\,\mathbf{u}_\lambda\mathbf{u}_\lambda^{\top}$, with $\mathbf{u}_\lambda$ the unit eigenvectors of $\mathbf{G}$, turns (17) into a filter acting mode by mode. Since $\varphi_\tau(\lambda)\simeq 1/\lambda$ for $\lambda \gg nm/(d\tau)$ and $\varphi_\tau(\lambda)\simeq d\tau/(nm)$ for $\lambda \ll nm/(d\tau)$, at training time τ the modes above

$$\lambda_c \;\sim\; \frac{nm}{d\tau}$$

have been learned — their coefficient has saturated at the ridgeless value $1/\lambda$ — whereas those below are still essentially untouched. This is the spectral cutoff of Sect. 2: fixing a training time is equivalent, up to the width of the crossover, to a hard spectral cutoff at λ_c, and a mode of eigenvalue λ is learned on the timescale $\tau_\lambda \sim nm/(d\lambda)$. Letting $\tau\to\infty$ recovers the ridgeless kernel interpolator $\mathbf{s}_\infty(\mathbf{y}) = -K(\mathbf{y},\mathbf{Y})^{\top}\mathbf{G}^{+}\boldsymbol{\xi}/\sqrt{\Delta_t}$, with $\mathbf{G}^{+}$ the Moore–Penrose pseudo-inverse; adding an L_2 penalty instead replaces $\mathbf{G}$ by $\mathbf{G}+\gamma\mathbf{I}_N$, as discussed in Sect. A.4.1 below.

A.4.1 THE ROLE OF REGULARIZATION

Adding an L_2 penalty $\frac{\gamma}{2nmd}\|\boldsymbol{\theta}\|^2$ to the loss (13) amounts to shifting the Gram matrix by $\gamma\mathbf{I}_N$ in (5),

$$\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{y}) = K(\mathbf{y},\mathbf{Y})^\top \left(\mathbf{G}+\gamma\mathbf{I}_N\right)^{-1}\left(\mathbf{I}_N - e^{-\frac{d(\mathbf{G}+\gamma\mathbf{I}_N)\tau}{nm}}\right)\left(-\frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\right). \tag{20}$$

Derivation. In the lazy regime the model is linear in its parameters around the initialization $\boldsymbol{\theta}(0)$ that we assume to be equal to 0 for simplicity, at which $\mathbf{s}_{\boldsymbol{\theta}(0)} = 0$. As in Sect. A.4 we treat one output component at a time and write $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{\Phi}(\mathbf{y})\boldsymbol{\theta}$, with $\boldsymbol{\Phi}(\mathbf{y}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{y})$ the (constant) feature map, $\boldsymbol{\Phi}_{\mathbf{Y}} \in \mathbb{R}^{N\times|\boldsymbol{\theta}|}$ its rows at the training points, so that $K(\mathbf{y},\mathbf{Y}) = \boldsymbol{\Phi}_{\mathbf{Y}}\boldsymbol{\Phi}(\mathbf{y})^\top$ and $\mathbf{G} = \boldsymbol{\Phi}_{\mathbf{Y}}\boldsymbol{\Phi}_{\mathbf{Y}}^\top$. With the penalty, the gradient flow $\mathrm{d}\boldsymbol{\theta}/\mathrm{d}\tau = -d^2\nabla_{\boldsymbol{\theta}}\big(\mathcal{L}_{\text{train}} + \frac{\gamma}{2nmd}\|\boldsymbol{\theta}\|^2\big)$ reads

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\frac{d}{nm}\Big[\boldsymbol{\Phi}_{\mathbf{Y}}^\top\big(\boldsymbol{\Phi}_{\mathbf{Y}}\boldsymbol{\theta}+\boldsymbol{\Xi}\big)+\gamma\,\boldsymbol{\theta}\Big], \qquad \boldsymbol{\Xi} = \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}. \tag{21}$$

Since $\boldsymbol{\theta}(0) = 0$ and the right-hand side lies in the row space of $\boldsymbol{\Phi}_{\mathbf{Y}}$ whenever $\boldsymbol{\theta}$ does, $\boldsymbol{\theta}(\tau) = \boldsymbol{\Phi}_{\mathbf{Y}}^\top\mathbf{a}(\tau)$ for some $\mathbf{a}(\tau) \in \mathbb{R}^N$. Substituting and using $\boldsymbol{\Phi}_{\mathbf{Y}}\boldsymbol{\Phi}_{\mathbf{Y}}^\top = \mathbf{G}$ gives $\boldsymbol{\Phi}_{\mathbf{Y}}^\top\dot{\mathbf{a}} = -\frac{d}{nm}\boldsymbol{\Phi}_{\mathbf{Y}}^\top\big[(\mathbf{G}+\gamma\mathbf{I}_N)\mathbf{a}+\boldsymbol{\Xi}\big]$, which is solved by

$$\dot{\mathbf{a}} = -\frac{d}{nm}\big[(\mathbf{G}+\gamma\mathbf{I}_N)\mathbf{a}+\boldsymbol{\Xi}\big], \qquad \mathbf{a}(\tau) = -(\mathbf{G}+\gamma\mathbf{I}_N)^{-1}\Big(\mathbf{I}_N - e^{-\frac{d\tau}{nm}(\mathbf{G}+\gamma\mathbf{I}_N)}\Big)\boldsymbol{\Xi}. \tag{22}$$

Finally $\mathbf{s}_{\boldsymbol{\theta}(\tau)}(\mathbf{y}) = \boldsymbol{\Phi}(\mathbf{y})\boldsymbol{\Phi}_{\mathbf{Y}}^\top\mathbf{a}(\tau) = K(\mathbf{y},\mathbf{Y})^\top\mathbf{a}(\tau)$, which is (20). At $\tau\to\infty$ it reduces to kernel ridge regression with ridge γ, $K(\mathbf{y},\mathbf{Y})^\top(\mathbf{G}+\gamma\mathbf{I}_N)^{-1}(-\boldsymbol{\Xi})$, i.e. (8).

A regularization strength $m \ll \gamma \ll \psi_n m$ suppresses the low eigenvalues and thus prevents overfitting, as early stopping does: the window sits between the memorization bulk, at $\Theta(m)$, and the generalization bulk, at $\Theta(\psi_n m)$.

Throughout the bias–variance analysis it is convenient to measure the ridge in units of $m\psi_n = N/d$, the scale of the generalization eigenvalues, and we write

$$\tilde{\gamma} \;=\; \frac{\gamma}{m\psi_n} \tag{23}$$

for the resulting *rescaled* ridge.

A.5 TWO-LAYER FULLY-CONNECTED NETWORK

Consider the two-layer network

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \;=\; \frac{1}{\sqrt{p}}\,\mathbf{a}^\top\,\sigma\left(\frac{\mathbf{W}\mathbf{x}}{\sqrt{d}}\right), \qquad \mathbf{W}\in\mathbb{R}^{p\times d}, \quad \mathbf{a}\in\mathbb{R}^p, \tag{24}$$

with i.i.d. Gaussian initialization $\mathbf{W}_{ij}, \mathbf{a}_i \sim \mathcal{N}(0,1)$ and a pointwise activation σ. In the limit $p\to\infty$, the NTK splits into two contributions, coming from the gradients with respect to the second-layer and first-layer weights respectively, and converges to the deterministic limit

$$K^\star(\mathbf{x},\mathbf{x}') = \underbrace{\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\sigma\left(\tfrac{\mathbf{w}^\top\mathbf{x}}{\sqrt{d}}\right)\sigma\left(\tfrac{\mathbf{w}^\top\mathbf{x}'}{\sqrt{d}}\right)\right]}_{K_\sigma(\mathbf{x},\mathbf{x}')} + \underbrace{\frac{\mathbf{x}^\top\mathbf{x}'}{d}\,\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\sigma'\left(\tfrac{\mathbf{w}^\top\mathbf{x}}{\sqrt{d}}\right)\sigma'\left(\tfrac{\mathbf{w}^\top\mathbf{x}'}{\sqrt{d}}\right)\right]}_{K_{\sigma'}(\mathbf{x},\mathbf{x}')}. \tag{25}$$

The first term K_σ is the so-called *conjugate* (or NNGP) kernel; the second is the *tangent* kernel of the first layer.

For inputs of comparable norm $\|\mathbf{x}\|^2, \|\mathbf{x}'\|^2 \asymp d$, both expectations (25) depend only on the normalized inner products $\mathbf{x}^\top\mathbf{x}'/d$, $\|\mathbf{x}\|^2/d$, $\|\mathbf{x}'\|^2/d$. When $\|\mathbf{x}\|^2 = \|\mathbf{x}'\|^2 = d$ exactly, $K^\star$ is a pure dot-product kernel,

$$K^\star(\mathbf{x},\mathbf{x}') \;=\; f\left(\tfrac{\mathbf{x}^\top\mathbf{x}'}{d}\right), \tag{26}$$

for some scalar function f depending on σ.

For $\sigma(u) = \max(u, 0)$ (ReLU), an explicit closed-form computation of the Gaussian integrals in (25) yields

$$K^\star(\mathbf{x}, \mathbf{x}') \;=\; \frac{\|\mathbf{x}\|\,\|\mathbf{x}'\|}{d}\, f_{\mathrm{ReLU}}\left(\frac{\mathbf{x}^\top \mathbf{x}'}{\|\mathbf{x}\|\,\|\mathbf{x}'\|}\right), \qquad f_{\mathrm{ReLU}}(u) \;=\; \frac{\sqrt{1-u^2} + 2u\,(\pi - \arccos u)}{2\pi}, \tag{27}$$

where the two terms come from $\mathbb{E}[\sigma(z_1)\sigma(z_2)] = \frac{\sqrt{1-u^2}+u(\pi-\arccos u)}{2\pi}$ and $\mathbb{E}[\sigma'(z_1)\sigma'(z_2)] = \frac{\pi-\arccos u}{2\pi}$ for standard Gaussians (z_1, z_2) of correlation u (Cho & Saul, 2009). The function f_{ReLU} is smooth on $(-1, 1)$, with $f_{\mathrm{ReLU}}(0) = 1/(2\pi)$ and $f'_{\mathrm{ReLU}}(0) = 1/2$. With the He-normalized activation $\sqrt{2}\max(u, 0)$ used in our experiments (Sect. A.8) the kernel is multiplied by 2.

A.6 Multilayer fully-connected network

For a depth-L fully-connected network with widths $d = n_0, n_1, \ldots, n_L = d'$ and activations σ,

$$\boldsymbol{\alpha}^{(0)}(\mathbf{x}) \;=\; \mathbf{x}, \qquad \tilde{\boldsymbol{\alpha}}^{(\ell+1)}(\mathbf{x}) \;=\; \frac{1}{\sqrt{n_\ell}}\, \mathbf{W}^{(\ell)}\, \boldsymbol{\alpha}^{(\ell)}(\mathbf{x}) + \mathbf{b}^{(\ell)}, \qquad \boldsymbol{\alpha}^{(\ell)} \;=\; \sigma\big(\tilde{\boldsymbol{\alpha}}^{(\ell)}\big), \tag{28}$$

with i.i.d. weights $\mathcal{N}(0, 1)$, biases $\mathcal{N}(0, \beta^2)$, and output $\tilde{\boldsymbol{\alpha}}^{(L)}(\mathbf{x})$, the conjugate and tangent kernels are computed by a coupled recursion on $\ell = 1, \ldots, L$ (Jacot et al., 2018). Writing $\Sigma^{(\ell)}(\mathbf{x}, \mathbf{x}')$ for the covariance of the pre-activations at the two inputs, define the 2×2 covariance matrix at layer ℓ,

$$\boldsymbol{\Sigma}^{(\ell)}(\mathbf{x}, \mathbf{x}') \;=\; \begin{pmatrix} \Sigma^{(\ell)}(\mathbf{x}, \mathbf{x}) & \Sigma^{(\ell)}(\mathbf{x}, \mathbf{x}') \\ \Sigma^{(\ell)}(\mathbf{x}', \mathbf{x}) & \Sigma^{(\ell)}(\mathbf{x}', \mathbf{x}') \end{pmatrix}, \tag{29}$$

which collects the covariances of the pre-activations $\tilde{\boldsymbol{\alpha}}^{(\ell)}$. The recursion is

$$\Sigma^{(1)}(\mathbf{x}, \mathbf{x}') \;=\; \frac{\mathbf{x}^\top \mathbf{x}'}{d} + \beta^2, \tag{30}$$

$$\Sigma^{(\ell+1)}(\mathbf{x}, \mathbf{x}') \;=\; \mathbb{E}_{(u,v)\sim\mathcal{N}(0, \boldsymbol{\Sigma}^{(\ell)}(\mathbf{x},\mathbf{x}'))}\big[\sigma(u)\,\sigma(v)\big] + \beta^2, \tag{31}$$

$$\dot{\Sigma}^{(\ell+1)}(\mathbf{x}, \mathbf{x}') \;=\; \mathbb{E}_{(u,v)\sim\mathcal{N}(0, \boldsymbol{\Sigma}^{(\ell)}(\mathbf{x},\mathbf{x}'))}\big[\sigma'(u)\,\sigma'(v)\big], \tag{32}$$

for $\ell = 1, \ldots, L-1$, where β^2 is the bias variance. The conjugate kernel is $K^\star_{\mathrm{conj}} = \Sigma^{(L)}$, and the NTK follows from $\Theta^{(1)} = \Sigma^{(1)}$ and $\Theta^{(\ell+1)} = \Theta^{(\ell)}\dot{\Sigma}^{(\ell+1)} + \Sigma^{(\ell+1)}$, i.e. from the unrolled product

$$K^\star(\mathbf{x}, \mathbf{x}') \;=\; \sum_{\ell=1}^{L} \Sigma^{(\ell)}(\mathbf{x}, \mathbf{x}') \prod_{\ell'=\ell+1}^{L} \dot{\Sigma}^{(\ell')}(\mathbf{x}, \mathbf{x}'). \tag{33}$$

At $L = 2$ and $\beta = 0$ this is $\Sigma^{(1)}\dot{\Sigma}^{(2)} + \Sigma^{(2)}$, i.e. (25). For a homogeneous activation such as ReLU, both $\Sigma^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}$ admit closed forms in terms of arc-cosine kernels, leading to a fully explicit (yet recursive) formula for $K^\star$.

A.7 Convolutional neural networks

Let the inputs lie in $\mathbb{R}^{C_0 \times P}$ with P pixels and C_0 channels. A depth-L CNN is built from convolutional layers

$$\tilde{\alpha}^{(\ell+1)}_{c,p}(\mathbf{x}) \;=\; \frac{1}{\sqrt{C_\ell\,|\mathcal{Q}|}} \sum_{c'=1}^{C_\ell} \sum_{q\in\mathcal{Q}} W^{(\ell)}_{c,c',q}\, \alpha^{(\ell)}_{c',p+q}(\mathbf{x}), \qquad \boldsymbol{\alpha}^{(\ell)} = \sigma(\tilde{\boldsymbol{\alpha}}^{(\ell)}), \tag{34}$$

with i.i.d. weights $W^{(\ell)}_{c,c',q} \sim \mathcal{N}(0, 1)$, patch support $\mathcal{Q} \subset \mathbb{Z}^k$ ($k = 1, 2$ in 1-D/2-D), and periodic boundary conditions. The readout aggregates the $C_L \times P$ last-layer activations into the d'-dimensional output.

Define the layer-ℓ pre-activation covariance tensor $\mathbf{\Sigma}^{(\ell)}(\mathbf{x},\mathbf{x}') \in \mathbb{R}^{P\times P}$, $\ell \geq 1$, by $\mathbf{\Sigma}^{(\ell)}_{p,p'}(\mathbf{x},\mathbf{x}') = \frac{1}{C_\ell}\sum_c \mathbb{E}[\tilde{\alpha}^{(\ell)}_{c,p}(\mathbf{x})\,\tilde{\alpha}^{(\ell)}_{c,p'}(\mathbf{x}')]$, and let $\mathbf{\Sigma}^{(0)}$ be the input overlap below. In the infinite-channel limit ($C_\ell \to \infty$ jointly), the recursion is

$$\mathbf{\Sigma}^{(0)}_{p,p'}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{C_0}\sum_{c=1}^{C_0} x_{c,p}\,x'_{c,p'}, \qquad \mathbf{\Sigma}^{(1)}_{p,p'}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{\Sigma}^{(0)}_{p+q,\,p'+q}(\mathbf{x},\mathbf{x}'), \tag{35}$$

$$\mathbf{\Sigma}^{(\ell+1)}_{p,p'}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathcal{T}\big[\mathbf{\Sigma}^{(\ell)}\big]_{p+q,\,p'+q}(\mathbf{x},\mathbf{x}'), \qquad \ell \geq 1, \tag{36}$$

the input $\boldsymbol{\alpha}^{(0)} = \mathbf{x}$ entering the first layer without activation, and where $\mathcal{T}$ and $\dot{\mathcal{T}}$ are defined as $\mathcal{T}[\mathbf{\Sigma}]_{p,p'} = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\,\mathbf{\Sigma}_{\{p,p'\}})}[\sigma(u)\sigma(v)]$ with $\mathbf{\Sigma}_{\{p,p'\}} = \left(\begin{smallmatrix}\mathbf{\Sigma}_{p,p} & \mathbf{\Sigma}_{p,p'}\\ \mathbf{\Sigma}_{p',p} & \mathbf{\Sigma}_{p',p'}\end{smallmatrix}\right)$, and similarly for $\dot{\mathcal{T}}$ with σ' in place of σ.

The convolutional NTK (Arora et al., 2019) of the pre-activations is the tensor $\mathbf{\Theta}^{(\ell)} \in \mathbb{R}^{P\times P}$ defined by $\mathbf{\Theta}^{(1)} = \mathbf{\Sigma}^{(1)}$ and

$$\mathbf{\Theta}^{(\ell+1)}_{p,p'}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \Big(\mathbf{\Theta}^{(\ell)} \odot \dot{\mathcal{T}}\big[\mathbf{\Sigma}^{(\ell)}\big]\Big)_{p+q,\,p'+q}(\mathbf{x},\mathbf{x}') \;+\; \mathbf{\Sigma}^{(\ell+1)}_{p,p'}(\mathbf{x},\mathbf{x}'), \tag{37}$$

with $\odot$ the entrywise product. Because the patch average acts on the product $\mathbf{\Theta}^{(\ell)} \odot \dot{\mathcal{T}}[\mathbf{\Sigma}^{(\ell)}]$, the recursion does not unroll into a product of layer-wise factors as in (33). The readout acts on the last activations $\boldsymbol{\alpha}^{(L)} = \sigma(\tilde{\boldsymbol{\alpha}}^{(L)})$, whose NTK tensor is $\bar{\mathbf{\Theta}}^{(L)} = \mathbf{\Theta}^{(L)} \odot \dot{\mathcal{T}}[\mathbf{\Sigma}^{(L)}] + \mathcal{T}[\mathbf{\Sigma}^{(L)}]$. The scalar NTK is then obtained by applying the readout to $\bar{\mathbf{\Theta}}^{(L)}$. For the two main readout choices, the CNTKs are

$$K^\star_{\mathrm{flat}}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{P}\sum_{p=1}^{P} \bar{\mathbf{\Theta}}^{(L)}_{p,p}(\mathbf{x},\mathbf{x}') \quad \text{(flatten + linear)}, \tag{38}$$

$$K^\star_{\mathrm{GAP}}(\mathbf{x},\mathbf{x}') \;=\; \frac{1}{P^2}\sum_{p,p'=1}^{P} \bar{\mathbf{\Theta}}^{(L)}_{p,p'}(\mathbf{x},\mathbf{x}') \quad \text{(Global Average Pooling)}. \tag{39}$$

A.8 Kernels used in the numerical experiments

We collect here the kernels used in the numerical experiments of the analytical section. In all the synthetic experiments the data are isotropic Gaussian, $\mathbf{\Sigma} = \mathbf{I}_d$ ($\sigma^2 = 1$), so that $\Gamma_t = 1$, and the kernel constants $\mu_I, \mu_B, \mu_0, \mu_1$ of Theorem D.1 are computed from the same f at the same t.

Two-layer ReLU NTK. We use the NTK of a two-layer network with the He-normalized activation $\sigma(z) = \sqrt{2}\max(0,z)$, both layers trained. In terms of the arc-cosine kernels of Cho & Saul (2009),

$$f(u) = \kappa_1(u) + u\,\kappa_0(u), \qquad \kappa_1(u) = \frac{\sqrt{1-u^2} + u(\pi - \arccos u)}{\pi}, \qquad \kappa_0(u) = 1 - \frac{\arccos u}{\pi}, \tag{40}$$

where $\kappa_1(u) = \mathbb{E}[\sigma(z_1)\sigma(z_2)]$ and $\kappa_0(u) = \mathbb{E}[\sigma'(z_1)\sigma'(z_2)]$ for standard Gaussians (z_1,z_2) of correlation u. Its Taylor coefficients at the origin are $f(0) = 1/\pi$, $f'(0) = 1$ and $f''(0) = 3/\pi$; the activation is not odd, so $\mu_0 \neq 0$ and the rank-one outlier at $\simeq nm\mu_0$ is present and removed as in Theorem 2.1. Since the arc-cosine kernels are defined only for $|u| \leq 1$, the noised points are projected onto the sphere of radius $\sqrt{d}$ in the kernel argument, $f(\hat{\mathbf{Y}}^{\nu\alpha\top}\hat{\mathbf{Y}}^{\mu\beta}/d)$ with $\hat{\mathbf{y}} = \sqrt{d}\,\mathbf{y}/\|\mathbf{y}\|$. At $\sigma^2 = 1$ this changes the arguments by $O(d^{-1/2})$ and makes $\mathbf{G}$ an exact inner-product kernel matrix.

Dual kernel of $\tanh$. The bias–variance theory and the kernel ridge experiments use

$$f(u) = \mathbb{E}\big[\tanh(z_1)\tanh(z_2)\big], \qquad f'(u) = \mathbb{E}\big[\tanh'(z_1)\tanh'(z_2)\big]. \tag{41}$$

This is the infinite-width limit of the random-features kernel of Sect. D.12, rather than an NTK, and it is an inner-product kernel of the form assumed throughout. The activation is odd, hence $f(0) = f''(0) = 0$ and $\mu_0 = 0$ exactly. In the finite-d experiments the kernel arguments are projected onto the sphere as above (Sect. D.10); scores are not rescaled.

Convolutional NTK. The CelebA experiments use the closed-form infinite-width NTK of a depth-D CNN with 3×3 filters, ReLU activations and a 1×1 readout, computed with the recursion of Sect. A.7; implementation details are in Sect. E.2.

Empirical NTK of a U-Net. In the feature-learning experiments the kernel is the finite-width NTK $\nabla_{\boldsymbol{\theta}} \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \nabla_{\boldsymbol{\theta}} \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{y})$ of the U-Net at training time τ, estimated by random projections as described in Sect. E.3.

B Equivalence between the Ornstein–Uhlenbeck and rectified-flow interpolants

B.1 Time reparametrization

The analytical part (Sect. 2) is formulated for the variance-preserving Ornstein–Uhlenbeck (OU) process

$$\mathbf{Y}_t^{\nu\alpha} \;=\; e^{-t}\,\mathbf{x}^{\nu} \;+\; \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\alpha}, \tag{42}$$

with $\Delta_t \;=\; 1 - e^{-2t}$ and t going from 0 to ∞, while the experimental section (Sect. 3) uses the rectified-flow (RF) interpolant

$$\mathbf{Y}_s^{\nu\alpha} \;=\; (1-s)\,\mathbf{x}^{\nu} \;+\; s\,\boldsymbol{\xi}^{\nu\alpha}, \tag{43}$$

with the time s going from 0 to 1. Both processes interpolate between the data distribution P_0 ($t = 0$ or $s = 0$) and a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I}_d)$ ($t \to \infty$ or $s = 1$), through different interpolants. Both are of the form

$$\mathbf{Y}^{\nu\alpha} = a\mathbf{x}^{\nu} + b\boldsymbol{\xi}^{\nu\alpha} \tag{44}$$

with $a, b \in [0, 1]$: $(a, b) = (e^{-t}, \sqrt{\Delta_t})$ with $a^2 + b^2 = 1$ for OU, and $(a, b) = (1 - s, s)$ for RF, which is not variance-preserving. The two are related by a change of time and a global rescaling of the inputs,

$$\mathbf{Y}_s^{\nu\alpha} = c_s\,\mathbf{Y}_{t(s)}^{\nu\alpha}, \qquad c_s = \sqrt{(1-s)^2 + s^2}, \qquad e^{-t(s)} = \frac{1-s}{c_s}, \qquad \Delta_{t(s)} = \frac{s^2}{c_s^2}, \tag{45}$$

with the same data and noises. For an inner-product kernel the rescaling only changes the kernel, $f(\mathbf{Y}_s^{\top}\mathbf{Y}'_s/d) = f_s(\mathbf{Y}_{t(s)}^{\top}\mathbf{Y}'_{t(s)}/d)$ with $f_s(u) = f(c_s^2 u)$, so that the RF Gram matrix at time s is the OU Gram matrix at time $t(s)$ built with the kernel f_s.

B.2 Substitution rule

All quantities appearing in the linear-equivalent theorem (Theorem D.1) and hence in the spectrum analysis (Theorem 2.1) depend on the OU process only through the scalar quantities $(e^{-t}, \sqrt{\Delta_t})$ and the kernel f. By (45), the results for RF at time s are those for OU at time $t(s)$ with the kernel f_s; equivalently, they are obtained from the OU formulas with the substitution

$$e^{-t} \;\longleftrightarrow\; 1 - s, \qquad \sqrt{\Delta_t} \;\longleftrightarrow\; s, \tag{46}$$

provided the identity $e^{-2t} + \Delta_t = 1$ is not used, i.e. keeping $\Gamma = (1-s)^2\sigma^2 + s^2$ wherever Γ_t appears.

B.3 Relation between score and velocity field

The OU and RF parametrizations differ in which target the network learns: the OU formulation trains a score model on $-\boldsymbol{\xi}/\sqrt{\Delta_t}$, while the RF formulation trains a velocity model on $\boldsymbol{\xi} - \mathbf{x}^{\nu}$. The two targets are related by the algebraic identity

$$\mathbf{v}^*(\mathbf{x}_s, s) \;=\; -\,\frac{\mathbf{x}_s + s\,\mathbf{s}^*(\mathbf{x}_s, s)}{1-s}, \qquad \mathbf{s}^*(\mathbf{x}_s, s) \;=\; -\,\frac{(1-s)\,\mathbf{v}^*(\mathbf{x}_s, s) + \mathbf{x}_s}{s}, \tag{47}$$

which follows from Tweedie's formula applied to either parametrization. This is a one-to-one affine map between the trained predictors, so the spectral analysis of $\mathbf{G}$ controls the training dynamics of both equivalently.

C Extended discussion of related work

This section discusses related work in more depth.

C.1 Effect of the architecture and training dynamics

To move beyond the empirical score, a recent line of work studies the effect of parametrizing the score with a neural network. The simplest such parametrization is a linear one, and it has been analyzed at both ends of the sample-complexity spectrum. Merger & Goldt (2026a) work in the proportional regime $n \asymp d$ and characterize the overfitting of the empirical risk minimizer for data with a power-law covariance. Wang & Pehlevan (2026), at the other end, work in the population limit and follow the training dynamics, again for power-law data; they exhibit a spectral bias, the modes associated with the largest eigenvalues being learned first. The same sequential picture survives beyond linear models: Bardone et al. (2026), for a two-layer network, show that the statistics of the data distribution are learned successively during training. Closer to the present setting, George et al. (2025a) and Bonnaire et al. (2026) both take the score to be a random-features network (Rahimi & Recht, 2007), an architecture whose infinite-width limit is a kernel method — see Appendix D.12 for the correspondence between random-features networks and the kernels used here. George et al. (2025a) compute the learning curves of the empirical risk minimizer, while Bonnaire et al. (2026) characterize the training dynamics and its timescales. We discuss the latter work at greater length in the next subsection.

C.2 Discussion of Bonnaire et al. (2026)

The phenomenon this paper analyzes was identified independently by Bonnaire et al. (2026) and Favero et al. (2025): a diffusion model trained on a finite dataset passes through a long window during which it generalizes, and only memorizes its training set much later, so that early stopping suffices to avoid memorization. Both papers establish that the width of that window grows with the number of samples n.

Bonnaire et al. (2026) characterize the two timescales $\tau_{\rm gen}$ and $\tau_{\rm mem}$ within a specific analytically tractable architecture, a random-features network (Rahimi & Recht, 2007). In that model the two timescales are read off the curvature of the loss at initialization. Since the loss is quadratic in the trainable weights, the Hessian at the origin governs the whole gradient flow: a direction of curvature λ relaxes on a timescale $1/\lambda$. The authors find that its spectrum splits into fast directions, learned first, and slow directions, learned much later, and the gap between the two groups of eigenvalues is what opens the window between $\tau_{\rm gen}$ and $\tau_{\rm mem}$. The connection with the present analysis is direct: for a quadratic loss the Hessian in parameter space and the Gram matrix in sample space are $\mathbf{\Phi}^\top\mathbf{\Phi}$ and $\mathbf{\Phi}\mathbf{\Phi}^\top$ for the same feature matrix $\mathbf{\Phi}$, and therefore have the same non-zero spectrum up to normalization. Favero et al. (2025) reach the same phenomenology from a kernel argument, stated for an arbitrary isotropic kernel.

The question addressed here is whether the same phenomenology can be established directly on the spectrum of the Gram matrix $\mathbf{G}^{\nu\alpha,\mu\beta} = K(\mathbf{Y}^{\nu\alpha}, \mathbf{Y}^{\mu\beta})$. This object is defined for any architecture, since the time-dependent neural tangent kernel exists whether or not training is lazy; what the lazy regime adds is that it is constant along training, which is what makes a closed-form analysis possible, while outside it the same matrix can still be measured, as we do in Sect. 3. This is indeed the case: the separation of timescales is read off that spectrum (Theorem 2.1), through the two bulks located at $\Theta(nm/d)$ and $\Theta(m)$, and hence the ratio $\tau_{\rm mem}/\tau_{\rm gen} \sim \psi_n$.

C.3 Concurrent work: Latourelle-Vigeant et al. (2026)

The closest work to this paper is the concurrent study of Latourelle-Vigeant et al. (2026). They also study the training dynamics of neural networks trained in the lazy regime on the score-matching task, in the linear regime $n \asymp d$. The main difference is that we fix the number of times each data point is noised to $1 \le m < \infty$ and study the associated Gram matrix, while they take $m = \infty$ and focus on the infinite-dimensional kernel operator $\mathcal{K}$ on a vector-valued RKHS

$$(\mathcal{K}f)(\mathbf{x}) = \int \mathrm{d}\mathbf{y}\; p_t^{\rm emp}(\mathbf{y})\, K(\mathbf{x},\mathbf{y})\, f(\mathbf{y}), \tag{48}$$

where $p_t^{\text{emp}}(\mathbf{y})$ is the noisy empirical distribution. They derive closed-form equations for the whole gradient-flow trajectory. Moreover, they characterize the generated distribution, which allows them to distinguish overfitting of the empirical score from memorization. Another difference is that they do not compute the bias–variance decomposition, which is an important part of our work.

Three-timescale picture. Instead of our two-timescale picture they describe three timescales: on $\tau = O(d)$ the model generalizes, as seen from a decrease in the test loss, then on timescales $\tau = O(n)$ the test loss starts to increase again. On $\tau = d^{1+\Theta(1)}$ the model starts to overfit without yet memorizing, and finally it memorizes on timescales of order $\tau = d^{\omega(1)}$.

Linear versus polynomial regime. Their analysis is confined to the linear scaling $n \asymp d$. Our Sect. 2.3 and Fig. 1 (*right*) show that the picture survives at $n \asymp d^k$: generalization splits into k sub-bulks indexed by polynomial degree, while memorization remains a single $\Theta(m)$, n-independent component.

Empirical evidence. Their numerical experiments use synthetic Gaussian data throughout, whereas our experiments of Sect. 3 use a CNTK on CelebA and a finite-width U-Net outside the lazy regime.

C.4 Discussion of Han et al. (2024)

Han et al. (2024; 2026) study the training of a two-layer fully-connected ReLU network by gradient descent to learn the score, and establish a minimax-optimal generalization bound for the resulting estimator. Their starting point is that the statistical literature on diffusion models is algorithm-agnostic: it assumes the empirical risk is minimized exactly, and says nothing about whether gradient descent on a non-convex network actually reaches such a minimizer. They close that gap by showing that the evolution of the trained network can be approximated by a sequence of localized kernel regression problems, and by deriving an early-stopping rule at which the minimax rate is attained. The extended version (Han et al., 2026) adds that explicit stopping rule together with the resulting estimation bound, neither of which appears in the ICLR 2024 paper (Han et al., 2024), and an experiment generating financial tabular data on a credit-default dataset.

Two differences in the setting are worth noting. Their network takes the diffusion time as an input, so the score is learned jointly across noise levels, whereas we work at fixed t and study the spectrum of the Gram matrix at that noise level. Their sampling procedure draws N i.i.d. triples $(t_j, \mathbf{x}_{0,j}, \mathbf{x}_{t_j})$, so that each clean sample is drawn afresh and used once, at one diffusion time, with one noise realization, which cannot lead to memorization. In our notation this is $m = 1$. We warn the reader that m denotes the network width in Han et al. (2024), whereas here it denotes the number of noise realizations per clean sample.

D Proofs of the analytical results

D.1 Assumptions

We assume that the data are of the form $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{z}$, where $\mathbf{z}$ has independent zero-mean, unit-variance, sub-Gaussian entries, with a covariance $\mathbf{\Sigma}$ such that $\lambda_{\max}(\mathbf{\Sigma}) = O(1)$ and $\operatorname{Tr}(\mathbf{\Sigma})/d$ converges to a constant denoted by σ^2 as $d \to \infty$. We focus on inner-product kernels of the form $K(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}^\top \mathbf{y}/d)$, where f is a smooth function. A canonical example is provided by the NTK of a multilayer neural network in the infinite-width limit.

To derive our results for the polynomial scaling regime ($n \asymp d^k$), we further assume that the kernel admits a diagonal representation in the basis of orthogonal polynomials:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{p \geq 0} \sum_{|\boldsymbol{j}|=p} \frac{\mu_p}{d^p \boldsymbol{j}!} \psi_p^{\boldsymbol{j}}(\mathbf{x}) \psi_p^{\boldsymbol{j}}(\mathbf{y}), \tag{49}$$

where $\{\psi_p^{\boldsymbol{j}}\}$ are the (monic, unnormalized) polynomials obtained by orthogonalizing the monomials in $L^2(P_t)$ (see Lemma D.2 for the formal definition). This assumption implies that these polynomials are the eigenfunctions of the kernel. For the uniform distribution on the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, every

dot-product kernel is diagonal in the spherical harmonics, with eigenvalues depending only on the degree (Bietti & Mairal, 2019; Ghorbani et al., 2021); for the isotropic Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ the degree-only form holds asymptotically in d, the mixing between degrees (e.g. $(\mathbf{x}^\top \mathbf{y})^2$ contains degree-0 terms) being subleading.

D.2 Notation

Indices $\nu = 1, \dots, n$ label the clean samples, $\alpha = 1, \dots, m$ the noise realizations and $i = 1, \dots, d$ the coordinates. We write $N = nm$ for the cardinality of the training set. $\mathbf{I}_d$ is the identity matrix in dimension d, $\mathbf{1}_m$ the all-ones vector in dimension m (likewise for N) and $\mathbf{B}_m = \mathbf{I}_n \otimes \mathbf{1}_m \mathbf{1}_m^\top$ is m times the projector on $U = \{\mathbf{v}^{\nu\alpha} \in \mathbb{R}^{nm} \ : \ \forall \nu,\ \alpha,\ \beta,\ \mathbf{v}^{\nu\alpha} = \mathbf{v}^{\nu\beta}\}$. We write $d_1 = O(d_2)$ when d_1/d_2 remains bounded as $d_1, d_2 \to \infty$, $d_1 = \Theta(d_2)$ when both d_1/d_2 and d_2/d_1 remain bounded, and $d_1 \asymp d_2$ when d_1/d_2 converges to a positive constant. The data distribution at $t = 0$ is denoted P_0 while the noisy distribution at time $t > 0$ is denoted P_t. We introduce $\mathbf{\Sigma} = \mathbb{E}_{P_0}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{d \times d}$ the data covariance and denote $\mathbf{v}_\lambda \in \mathbb{R}^d$ its eigenvector associated with the eigenvalue λ. We write $\Delta_t = 1 - e^{-2t}$ for the noise variance, $\mathbf{\Sigma}_t = e^{-2t}\mathbf{\Sigma} + \Delta_t \mathbf{I}_d$ for the covariance at noise level t and $\lambda_t = e^{-2t}\lambda + \Delta_t$ for the eigenvalue of $\mathbf{\Sigma}_t$ associated with $\mathbf{v}_\lambda$; $\sigma^2 = \mathrm{Tr}(\mathbf{\Sigma})/d$ and $\Gamma_t = \mathrm{Tr}(\mathbf{\Sigma}_t)/d = e^{-2t}\sigma^2 + \Delta_t$. Finally $\psi_n = n/d$ is the sample complexity and $\tilde{\gamma} = \gamma/(m\psi_n)$ the ridge rescaled by N/d, the scale of the generalization eigenvalues.

D.3 Proof of the linear equivalent of the Gram matrix

We first state the linear-equivalence theorem referred to in the main text.

Theorem D.1 (Linear Equivalent of the Gram Matrix with Repeated Data)**.** *Consider the Gram matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$ ($N = nm$) defined by $\mathbf{G}^{\nu\alpha,\mu\beta} = f\left(\frac{\sum_{i=1}^d \mathbf{Y}_i^{\nu\alpha}\mathbf{Y}_i^{\mu\beta}}{d}\right)$, with training samples $\mathbf{Y}^{\nu\alpha} = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\alpha}$. In the limit $n, d \to \infty$ with $n/d \to \psi_n$ and $m = O(1)$, the matrix $\mathbf{G}$ is asymptotically equivalent in spectrum to:*

$$\mathbf{G}_{\mathrm{lin}} = \mu_I \mathbf{I}_N + \mu_B \mathbf{B}_m + \mu_0 \mathbf{1}_N \mathbf{1}_N^\top + \mu_1 \frac{\mathbf{Y}^\top \mathbf{Y}}{d}, \tag{50}$$

where $\mathbf{B}_m = \mathbf{I}_n \otimes \mathbf{1}_m \mathbf{1}_m^\top$ and the coefficients are given by:

$$\mu_I = f(\sigma^2 e^{-2t} + \Delta_t) - f(\sigma^2 e^{-2t}) - \Delta_t f'(0), \qquad \mu_B = f(\sigma^2 e^{-2t}) - f(0) - \sigma^2 e^{-2t} f'(0),$$
$$\mu_1 = f'(0), \qquad \mu_0 = f(0) + \tfrac{1}{2d^2} f''(0)\,\mathrm{Tr}(\mathbf{\Sigma}_t^2).$$

Proof. Write $a = (\nu, \alpha) \in [n] \times [m]$ for the $N = nm$ training indices, and

$$q^{ab} \ = \ \frac{1}{d}\sum_{i=1}^d \mathbf{Y}_i^a \mathbf{Y}_i^b, \qquad \mathbf{Y}^{\nu\alpha} \ = \ e^{-t}\,\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\alpha},$$

with $\mathbf{x}^\nu \overset{\text{i.i.d.}}{\sim} P_0$ and $\boldsymbol{\xi}^{\nu\alpha} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$, so that $\mathbf{G}^{ab} = f(q^{ab})$. Three index regimes have to be distinguished, according to how much randomness the pair (a, b) shares: $a = b$; $a \neq b$ within one cluster ($\nu = \mu$, $\alpha \neq \beta$); and different clusters ($\nu \neq \mu$).

Step 1: uniform control of the overlaps.

Lemma D.1. *Under the assumptions of Sect. D there is a constant C, depending only on the sub-Gaussian norm of P_0 and on $\lambda_{\max}(\mathbf{\Sigma})$, such that with probability $1 - o(1)$,*

$$\max_a \left|q^{aa} - \Gamma_t\right| \ \le \ C\sqrt{\tfrac{\log d}{d}}, \qquad \max_{\nu,\, \alpha \neq \beta} \left|q^{\nu\alpha,\nu\beta} - \sigma^2 e^{-2t}\right| \ \le \ C\sqrt{\tfrac{\log d}{d}}, \tag{51}$$

$$\max_{\nu \neq \mu} \left|q^{\nu\alpha,\mu\beta}\right| \ \le \ C\sqrt{\tfrac{\log d}{d}}. \tag{52}$$

Each of the three families consists of quadratic forms in independent sub-Gaussian vectors, so each entry obeys a Bernstein-type bound with sub-exponential tails at scale $d^{-1/2}$; a union bound over the

$O(d^2)$ pairs costs the factor $\sqrt{\log d}$. The centering constants are the means: $\mathbb{E}[q^{aa}] = \mathrm{Tr}(\mathbf{\Sigma}_t)/d = \Gamma_t$, $\mathbb{E}[q^{\nu\alpha,\nu\beta}] = e^{-2t}\,\mathrm{Tr}(\mathbf{\Sigma})/d = \sigma^2 e^{-2t}$ for $\alpha \neq \beta$ (the two noises are independent, only $\mathbf{x}^\nu$ is shared), and $\mathbb{E}[q^{\nu\alpha,\mu\beta}] = 0$ for $\nu \neq \mu$.

We assume throughout that f is three times continuously differentiable on a neighborhood of $[-\delta, \Gamma_t + \delta]$ for some $\delta > 0$, which is where Lemma D.1 confines every overlap. All the expansions below are taken on the event of Lemma D.1.

Step 2: the three regimes. On the diagonal, $f(q^{aa}) = f(\Gamma_t) + f'(\Gamma_t)(q^{aa} - \Gamma_t) + O(d^{-1}\log d)$, so the diagonal of $\mathbf{G}$ equals $f(\Gamma_t)\mathbf{I}_N$ up to a *diagonal* matrix $\mathbf{D}$. Within a cluster, likewise, $f(q^{\nu\alpha,\nu\beta}) = f(\sigma^2 e^{-2t}) + f'(\sigma^2 e^{-2t})(q^{\nu\alpha,\nu\beta} - \sigma^2 e^{-2t}) + O(d^{-1}\log d)$, so the same-cluster off-diagonal part equals $f(\sigma^2 e^{-2t})(\mathbf{B}_m - \mathbf{I}_N)$ up to a *block-diagonal* matrix $\mathbf{E}$ with n blocks of size $m \times m$. Across clusters the overlaps are small and we expand at the origin,

$$f(q^{ab}) \;=\; f(0) + f'(0)\,q^{ab} + \tfrac{1}{2}f''(0)\,(q^{ab})^2 + R^{ab}, \qquad |R^{ab}| \;\le\; \tfrac{1}{6}\sup_{|u|\le\delta}|f'''(u)| \cdot |q^{ab}|^3.$$

Collecting the exact (non-remainder) pieces in the basis $\{\mathbf{I}_N,\ \mathbf{B}_m - \mathbf{I}_N,\ \mathbf{1}_N\mathbf{1}_N^\top - \mathbf{B}_m,\ \mathbf{Y}^\top\mathbf{Y}/d\}$ and using $\mathbb{E}[(q^{ab})^2] = \mathrm{Tr}(\mathbf{\Sigma}_t^2)/d^2$ for $\nu \neq \mu$ gives exactly $\mathbf{G}_{\mathrm{lin}}$ with the four constants $\mu_I, \mu_B, \mu_0, \mu_1$ of the statement.

Step 3: operator-norm bookkeeping. It remains to bound the residuals. $\mathbf{D}$ is diagonal and $\mathbf{E}$ is block diagonal with blocks of fixed size $m = O(1)$, so their operator norms are controlled entrywise,

$$\|\mathbf{D}\|_{\mathrm{op}} \;=\; \max_a |\mathbf{D}^{aa}| \;=\; O\big(\sqrt{\log d/d}\big), \qquad \|\mathbf{E}\|_{\mathrm{op}} \;\le\; m \max_{a\neq b}|\mathbf{E}^{ab}| \;=\; O\big(\sqrt{\log d/d}\big).$$

For the third-order remainder $\mathbf{R}$ the Frobenius norm suffices, because $N \asymp d$:

$$\|\mathbf{R}\|_{\mathrm{op}} \;\le\; \|\mathbf{R}\|_F \;\le\; N \max_{\nu\neq\mu}|R^{ab}| \;=\; O\big(d\cdot(\log d/d)^{3/2}\big) \;=\; O\big(d^{-1/2}\log^{3/2} d\big).$$

The quadratic term is the one place where entrywise bounds are not enough. The matrix $\mathbf{Q}^{(2)}$ with entries $(q^{ab})^2$ for $\nu \neq \mu$ has entries of order $1/d$ and dimension $N \asymp d$, so $\|\mathbf{Q}^{(2)} - \mathbb{E}\mathbf{Q}^{(2)}\|_F = O(1)$: the Frobenius bound does not vanish, and $\mathbf{Q}^{(2)}$ concentrates around $\frac{\mathrm{Tr}(\mathbf{\Sigma}_t^2)}{d^2}(\mathbf{1}_N\mathbf{1}_N^\top - \mathbf{B}_m)$. Nonetheless, as in El Karoui (2010), one can prove that $\|\mathbf{Q}^{(2)} - \mathbb{E}\mathbf{Q}^{(2)}\|_{\mathrm{op}} = o(1)$.

Step 4: conclusion. Summing the four contributions,

$$\|\mathbf{G} - \mathbf{G}_{\mathrm{lin}}\|_{\mathrm{op}} \;=\; O\big(d^{-1/2}\log^{3/2} d\big) + o(1) \;=\; o(1) \qquad \text{with probability } 1 - o(1).$$

By Weyl's inequality every eigenvalue of $\mathbf{G}$ is within $o(1)$ of the corresponding eigenvalue of $\mathbf{G}_{\mathrm{lin}}$; in particular the two empirical spectral distributions have the same weak limit, and any eigenvalue of $\mathbf{G}_{\mathrm{lin}}$ separated from the rest of its spectrum by a gap bounded away from zero is tracked individually. This is the sense in which $\mathbf{G}$ and $\mathbf{G}_{\mathrm{lin}}$ are used interchangeably in the rest of the appendix. □

D.4 Spectrum of the Gram matrix

We now state Theorem 2.1 of the main text in full.

Theorem D.2 (Spectral Distribution of the Gram Matrix)**.** *Let the empirical spectral density of the data covariance $\mathbf{\Sigma}$ converge to $\rho_{\mathbf{\Sigma}}(\lambda)$. In the high-dimensional limit $d, n \to \infty$ with $n/d \to \psi_n$, the Stieltjes transform $q(z) = \frac{1}{N}\mathrm{Tr}(\mathbf{G} - z\mathbf{I}_N)^{-1}$ and the auxiliary order parameter $r(z)$ of (59) solve the system*

$$z - \mu_I - m\mu_B + \frac{1}{mr} = \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\frac{\mu_1\Delta_t + m\mu_1 e^{-2t}\lambda}{1 + \mu_1 e^{-2t}\psi_n m^2\lambda r + \mu_1\psi_n m\Delta_t q}, \tag{53}$$

$$z - \mu_I + \frac{m-1}{m(q-r)} = \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\frac{\mu_1\Delta_t}{1 + \mu_1 e^{-2t}\psi_n m^2\lambda r + \mu_1\psi_n m\Delta_t q}. \tag{54}$$

In the large-data regime ($\psi_n \gg 1$), the spectral density $\rho(\omega)$ of $\mathbf{G} - \mu_0\mathbf{1}_N\mathbf{1}_N^\top$, in the spectral variable ω, admits a decomposition into four distinct components:

$$\rho(\omega) \approx \underbrace{\frac{m-1-1/\psi_n}{m}\delta(\omega-\mu_I) + \frac{1-1/\psi_n}{m}\delta(\omega-\bar{\mu})}_{\textit{atoms}} + \underbrace{\frac{1}{\psi_n m}\rho_1(\omega)}_{\textit{generalization bulk}} + \underbrace{\frac{1}{\psi_n m}\rho_2(\omega)}_{\textit{memorization bulk}}, \tag{55}$$

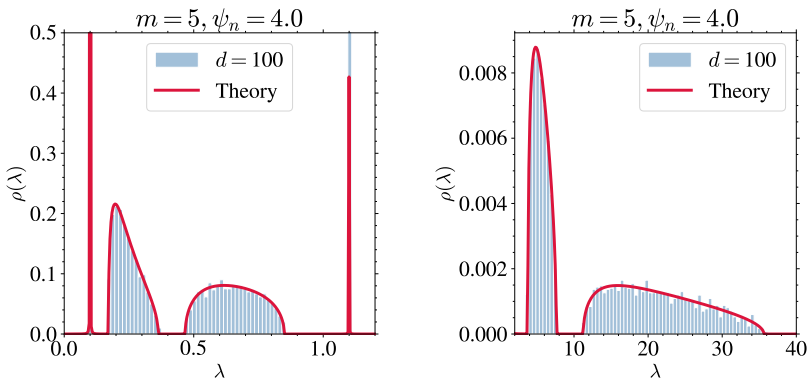

Figure 6: Empirical histogram of the eigenvalues of $\mathbf{G}_{\text{lin}}$ for $d = 100$ (blue) averaged over 10 runs and the analytical prediction obtained by solving the equations on the Stieltjes transform for $m = 5, \psi_n = 4.0$ and $t = 0.1$ for $\rho_{\mathbf{\Sigma}}(\lambda) = \frac{1}{2}\delta(\lambda - 1) + \frac{1}{2}\delta(\lambda - 0.1)$ and for the kernel parameters $\mu_1 = 1, \mu_I = 0.1, \mu_B = 0.2$. The left and right panels corresponds to different ranges of λ.

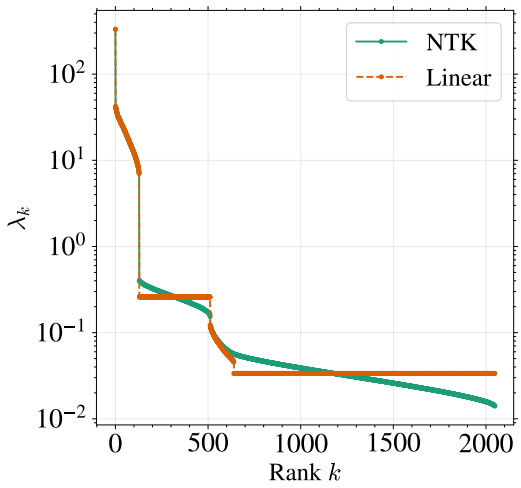

Figure 7: Comparison between the linear equivalent and the non-linear NTK spectrum in finite dimension for $d = 128,\ \psi_n = 4,\ m = 4$, and $t = 0.1$.

where $\bar{\mu} = m\mu_B + \mu_I$. To every eigenvalue λ of $\mathbf{\Sigma}$ there correspond two eigenvalues of $\mathbf{G}$, with the following asymptotic scalings, where $\lambda_t = e^{-2t}\lambda + \Delta_t$ is the associated eigenvalue of $\mathbf{\Sigma}_t = e^{-2t}\mathbf{\Sigma} + \Delta_t \mathbf{I}_d$:

- ***Generalization bulk:*** $\lambda_{\rm gen}(\lambda) \simeq \mu_1 \psi_n m \lambda_t = \Theta(nm/d)$.
- ***Memorization bulk:*** $\lambda_{\rm mem}(\lambda) \simeq \mu_I + \frac{\mu_B(m-1)\Delta_t}{\lambda_t} = \Theta(m)$.

The proof of this theorem is presented in the next two subsections. In Sect. D.5, we derive the self-consistent equations on the Stieltjes transform of $\mathbf{G}$, while in Sect. D.6 we establish the result on the asymptotic spectrum of $\mathbf{G}$. An example of the spectrum for $\mathbf{\Sigma} \neq \mathbf{I}_d$ is presented in Fig. 6, while the convergence of the two bulks $\rho_1,\ \rho_2$ is presented in Fig.7 and Fig. 8.

D.5 Replica computation of the Stieltjes transform of the Gram matrix

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

$$\log \mathcal{Z} = \lim_{s\to 0} \frac{\mathcal{Z}^s - 1}{s}. \tag{56}$$

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

We first derive the self-consistent equations on the Stieltjes transform of $\mathbf{G}$.

Proof. Since $\mathbf{G}$ and $\mathbf{G}_{\rm lin}$ have asymptotically the same spectrum, we replace the non-linear Gram matrix by its linear equivalent. We compute the Stieltjes transform $q(z) = \frac{1}{N}\operatorname{Tr}(\mathbf{G}_{\rm lin} - z\mathbf{I}_N)^{-1}$ of the linear-equivalent Gram matrix

$$\mathbf{G}_{\rm lin} = \mu_I\, \mathbf{I}_N + \mu_B\, \mathbf{B}_m + \mu_0\, \mathbf{1}_N \mathbf{1}_N^\top + \mu_1\, \frac{\mathbf{Y}^\top \mathbf{Y}}{d}.$$

The rank-one spike $\mu_0 \mathbf{1}_N \mathbf{1}_N^\top$ contributes a single outlier eigenvalue outside the bulk; we omit it from the replica calculation.

For $\Im z > 0$, write

$$q(z) \;=\; \frac{2}{N}\,\partial_z \log \mathcal{Z}(z), \qquad \mathcal{Z}(z) \;=\; \int \mathrm{d}\boldsymbol{\phi}\; e^{-\frac{1}{2}\boldsymbol{\phi}^\top (\mathbf{G}_{\rm lin} - z\mathbf{I}_N)\boldsymbol{\phi}}.$$

Introducing s replicas $\{\boldsymbol{\phi}^a\}_{a=1}^s$ and absorbing the identity term into a shifted spectral parameter $z - \mu_I$,

$$\mathbb{E}_{X,\Xi}[\mathcal{Z}^s] \;=\; \int \prod_a \mathrm{d}\boldsymbol{\phi}^a\, e^{\frac{1}{2}(z-\mu_I)\,\boldsymbol{\phi}^a \cdot \boldsymbol{\phi}^a}\; e^{-\frac{\mu_B}{2}\,\boldsymbol{\phi}^{a\top}\mathbf{B}_m\boldsymbol{\phi}^a}\; \mathbb{E}_Y\left[e^{-\frac{\mu_1}{2d}\,\boldsymbol{\phi}^{a\top}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\phi}^a}\right]. \tag{57}$$

Gaussian average over $\mathbf{Y}$. With sample-index pair $(\nu, \alpha) \in [n] \times [m]$ and $\mathbf{v} = \mathbf{1}_m$, the second moment of $\mathbf{Y}$ is

$$\mathbf{K}_{ij}^{\nu\nu',\alpha\alpha'} = \mathbb{E}\left[\mathbf{Y}_i^{\nu\alpha}\mathbf{Y}_j^{\nu'\alpha'}\right] = e^{-2t}\,\mathbf{\Sigma}_{ij}\mathbf{v}^\alpha\mathbf{v}^{\alpha'}\,\delta^{\nu\nu'} + \Delta_t\,\delta_{ij}\,\delta^{\nu\nu'}\delta^{\alpha\alpha'}.$$

Setting $\mathbf{A}_{ij}^{\nu\nu',\alpha\alpha'} = \delta_{ij}\sum_a \boldsymbol{\phi}^a_{\nu\alpha}\boldsymbol{\phi}^a_{\nu'\alpha'}$ and using the Gaussian identity $\mathbb{E}_Y\left[e^{-\frac{\mu_1}{2d}\operatorname{Tr}(\mathbf{Y}^\top\mathbf{A}\mathbf{Y})}\right] = e^{-\frac{1}{2}\log\det\left(\mathbf{I}_d+\frac{\mu_1}{d}\mathbf{K}\mathbf{A}\right)}$, we obtain

$$\mathbb{E}_{X,\Xi}[\mathcal{Z}^s] \;=\; \int\prod_a \mathrm{d}\boldsymbol{\phi}^a\, e^{\frac{z-\mu_I}{2}\,\boldsymbol{\phi}^a\cdot\boldsymbol{\phi}^a}\, e^{-\frac{\mu_B}{2}\,\boldsymbol{\phi}^{a\top}\mathbf{B}_m\boldsymbol{\phi}^a}\, e^{-\frac{1}{2}\log\det\left(\mathbf{I}_d+\frac{\mu_1}{d}\mathbf{K}\mathbf{A}\right)}. \tag{58}$$

Order parameters. Introduce the overlaps and their Lagrange-multiplier conjugates

$$\mathbf{Q}^{ab} \;=\; \tfrac{1}{nm}\,\boldsymbol{\phi}^a\cdot\boldsymbol{\phi}^b, \qquad \mathbf{R}^{ab} \;=\; \tfrac{1}{nm^2}\,\boldsymbol{\phi}^a_{\nu\alpha}\,\mathbf{v}^\alpha\mathbf{v}^{\alpha'}\,\boldsymbol{\phi}^b_{\nu\alpha'}, \tag{59}$$

via the identity

$$1 \;=\; \int \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\,\mathrm{d}\mathbf{R}\,\mathrm{d}\hat{\mathbf{R}}e^{\frac{1}{2}\hat{\mathbf{Q}}^{ab}(nm\,\mathbf{Q}^{ab}-\boldsymbol{\phi}^a\cdot\boldsymbol{\phi}^b)+\frac{1}{2}\hat{\mathbf{R}}^{ab}(nm^2\,\mathbf{R}^{ab}-\boldsymbol{\phi}^a_{\nu\alpha}\mathbf{v}^\alpha\mathbf{v}^{\alpha'}\boldsymbol{\phi}^b_{\nu\alpha'})}. \tag{60}$$

The $\mathbf{B}_m$ and $\mathbf{K}$-determinant terms then become functions of $\mathbf{Q},\mathbf{R}$ only:

$$e^{-\frac{\mu_B}{2}\,\boldsymbol{\phi}^{a\top}\mathbf{B}_m\boldsymbol{\phi}^a}\, e^{-\frac{1}{2}\log\det(\mathbf{I}+\frac{\mu_1}{d}\mathbf{K}\mathbf{A})} = e^{-\frac{\mu_B nm^2}{2}\operatorname{Tr}\mathbf{R}}e^{-\frac{1}{2}\log\det\left(\mathbf{I}+\frac{\mu_1}{d}(e^{-2t}nm^2\boldsymbol{\Sigma}\mathbf{R}+nm\Delta_t\mathbf{I}_d\,\mathbf{Q})\right)}. \tag{61}$$

The remaining $\boldsymbol{\phi}$-integral is Gaussian:

$$\int \mathrm{d}\boldsymbol{\phi}^a\; e^{-\frac{1}{2}\hat{\mathbf{Q}}^{ab}\,\boldsymbol{\phi}^a\cdot\boldsymbol{\phi}^b-\frac{1}{2}\hat{\mathbf{R}}^{ab}\,\boldsymbol{\phi}^a_{\nu\alpha}v^\alpha v^{\alpha'}\boldsymbol{\phi}^b_{\nu\alpha'}} \;=\; e^{-\frac{n}{2}\log\det(\hat{\mathbf{Q}}^{ab}\delta_{\alpha\alpha'}+\hat{\mathbf{R}}^{ab}\mathbf{v}^\alpha\mathbf{v}^{\alpha'})}. \tag{62}$$

Replica-symmetric ansatz. Set $\mathbf{Q}^{ab} = q\delta^{ab}$, $\hat{\mathbf{Q}}^{ab} = \hat{q}\delta^{ab}$, $\mathbf{R}^{ab} = r\delta^{ab}$, $\hat{\mathbf{R}}^{ab} = \hat{r}\delta^{ab}$ for all a,b. The matrix $\hat{q}\,\delta_{\alpha\alpha'} + \hat{r}\mathbf{v}^\alpha\mathbf{v}^{\alpha'}$ has eigenvalues $\hat{q}$ (multiplicity $m-1$) and $\hat{q}+m\hat{r}$ (multiplicity 1), so

$$\log\det(\hat{q}\,\delta_{\alpha\alpha'} + \hat{r}\,v^\alpha v^{\alpha'}) \;=\; (m-1)\log\hat{q} + \log(\hat{q}+m\hat{r}).$$

Replacing the determinant over $\boldsymbol{\Sigma}$ by an integral against its spectral measure $\rho_{\boldsymbol{\Sigma}}$ and dividing by sd, the per-replica action reads

$$\begin{aligned} S(q,\hat{q},r,\hat{r}) \;=\; &-\; m\psi_n\,(z-\mu_I)\,q \;+\; \psi_n(m-1)\log\hat{q} \;+\; \psi_n\log(\hat{q}+m\hat{r}) \\ &-\; \psi_n m\,q\hat{q} \;-\; \psi_n m^2\,\hat{r}\,r \;+\; \mu_B\,\psi_n m^2\,r \\ &+\; \int \mathrm{d}\rho_{\boldsymbol{\Sigma}}(\lambda)\;\log\left(1+\mu_1\left(e^{-2t}\,\psi_n m^2\,\lambda\,r + \psi_n m\,\Delta_t\,q\right)\right). \end{aligned} \tag{63}$$

Saddle-point equations. Stationarity $\partial_{\hat{q}}S = \partial_{\hat{r}}S = \partial_q S = \partial_r S = 0$ gives

$$m\,q \;=\; \frac{m-1}{\hat{q}} \;+\; \frac{1}{\hat{q}+m\hat{r}}, \tag{64}$$

$$m\,r \;=\; \frac{1}{\hat{q}+m\hat{r}}, \tag{65}$$

$$(z-\mu_I)+\hat{q} \;=\; \int \mathrm{d}\rho_{\boldsymbol{\Sigma}}(\lambda)\;\frac{\mu_1\,\Delta_t}{1+\mu_1\,e^{-2t}\psi_n m^2\,\lambda\,r + \mu_1\,\psi_n m\,\Delta_t\,q}, \tag{66}$$

$$\hat{r} \;=\; \mu_B \;+\; \int \mathrm{d}\rho_{\boldsymbol{\Sigma}}(\lambda)\;\frac{\mu_1\,e^{-2t}\,\lambda}{1+\mu_1\,e^{-2t}\psi_n m^2\,\lambda\,r + \mu_1\,\psi_n m\,\Delta_t\,q}. \tag{67}$$

These four equations on $(q,r,\hat{q},\hat{r})$ can be reduced to the two equations on (q,r) by eliminating $\hat{q},\hat{r}$. □

D.6 Asymptotic spectrum of the Gram matrix

There are two ways to obtain the scaling and the structure of the spectrum of $\mathbf{G}$: from the replica equations, and directly from linear-algebra arguments, the latter also giving the eigenvectors. We assume without loss of generality that $\mu_I = 0$ since it only amounts to a constant shift in the spectrum.

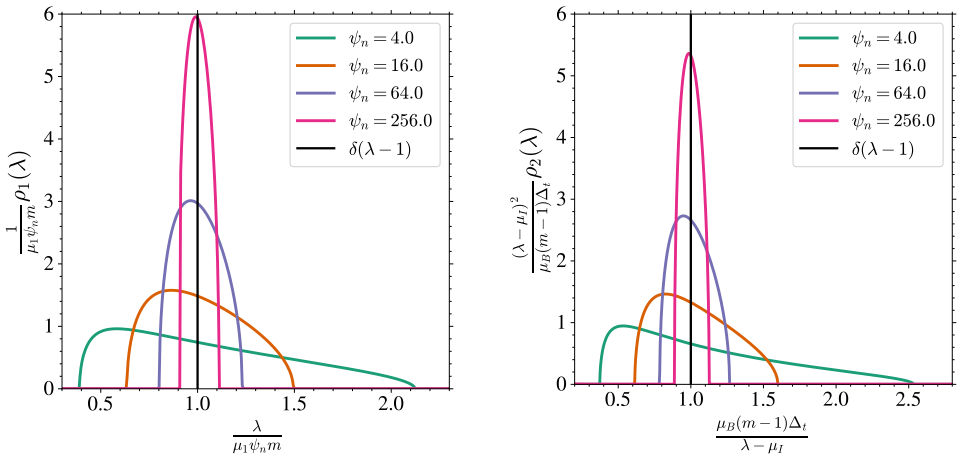

Figure 8: Analytical solutions of the rescaled bulks ρ_1, ρ_2 for different values of ψ_n for $m = 5,\ t = 0.1, \mathbf{\Sigma} = \mathbf{I}_d,\ \mu_1 = 1.0,\ \mu_B = 0.2,\ \mu_I = 0.1,\ \mu_0 = 0.0$ and the asymptotic limit $\delta(\lambda - 1)$.

D.6.1 Derivation from the replica equations

We solve the replica equations perturbatively to obtain the structure and the scalings of the spectrum of $\mathbf{G}_{\text{lin}}$.

Peak at $z = m\mu_B$. Assume $z = m\mu_B + \epsilon$ with $\epsilon \ll 1$ and plug in the ansatz $q = -\frac{q_0}{\epsilon}$ and $r = -\frac{r_0}{\epsilon}$. Then

$$\frac{1}{mr} = m\hat{r} + \hat{q} = m\mu_B - z + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{1 + \mu_1 e^{-2t} \psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} \tag{68}$$

$$= -\epsilon (1 + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{\mu_1 e^{-2t} \psi_n m^2 \lambda r_0 + \mu_1 \psi_n m \Delta_t q_0}) \tag{69}$$

and thus

$$mq = \frac{m-1}{\hat{q}} + mr = \frac{m-1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 \Delta_t}{1 + \mu_1 e^{-2t} \psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} - z} + mr \tag{70}$$

$$= -\frac{1}{\epsilon} (1 + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{\mu_1 e^{-2t} \psi_n m^2 \lambda r_0 + \mu_1 \psi_n m \Delta_t q_0})^{-1} \tag{71}$$

at first order

$$q = -\frac{1}{m} \frac{1}{\epsilon} \tag{72}$$

$$r = -\frac{1}{m} \frac{1}{\epsilon} \tag{73}$$

i.e. there are n eigenvalues at this peak. Inserting $q = -1/(m\epsilon)$ back into the equation for q, the integral reduces to $1/\psi_n$ and expanding $(1 + 1/\psi_n)^{-1} \simeq 1 - 1/\psi_n$ gives the correction

$$q = -\frac{1}{\epsilon} (\frac{1}{m} - \frac{1}{m\psi_n}) \tag{74}$$

The correction yields $n - d$ eigenvalues.

Peak at $z = 0$. Assume $q = \frac{-q_0}{z}$ and $r = \frac{-r_0}{z}$.

$$\frac{1}{mr} = m\hat{r} + \hat{q} = m\mu_B - z + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{1 + \mu_1 e^{-2t} \psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} \sim m\mu_B + O(z) \tag{75}$$

hence

$$mq = \frac{m-1}{\hat{q}} \tag{76}$$

$$= \frac{m-1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 \Delta_t}{1+\mu_1 e^{-2t}\psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} - z} \tag{77}$$

$$= \frac{m-1}{\frac{z}{-\psi_n m q_0} - z} \tag{78}$$

This yields at first order

$$q_0 = 1 - \frac{1}{\psi_n m} - \frac{1}{m}. \tag{79}$$

i.e. there are $nm - n - d$ eigenvalues at the peak.

Bulk at $z = O(\psi_n m)$. We solve the equations perturbatively, using the fact that at leading order $q \sim \frac{-1}{z}$ and $mr = \frac{-1}{z}$. Denote $z = \psi_n m \tilde{z}$. Then one has $mq = \frac{m-1}{\hat{q}} + mr$.

$$mq = \frac{m-1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 \Delta_t}{1+\mu_1 e^{-2t}\psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} - z} \tag{80}$$

$$+ \frac{1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{1+\mu_1 e^{-2t}\psi_n m^2 \lambda r + \mu_1 \psi_n m \Delta_t q} + m\mu_B - z} \tag{81}$$

$$= \frac{m-1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 \Delta_t}{1-\mu_1 e^{-2t}\lambda/\tilde{z} - \mu_1 \Delta_t/\tilde{z}} - z} + \frac{1}{\int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda) \frac{\mu_1 (e^{-2t}\lambda m + \Delta_t)}{1-\mu_1 e^{-2t}\lambda/\tilde{z} - \mu_1 \Delta_t/\tilde{z}} + m\mu_B - z} \tag{82}$$

$$\sim -\frac{m-1}{z}\left(1 + \frac{1}{z}\int \mathrm{d}\rho_{\mathbf{\Sigma}} \frac{\mu_1 \Delta_t \tilde{z}}{\tilde{z} - \mu_1(e^{-2t}\lambda + \Delta_t)}\right) \tag{83}$$

$$- \frac{1}{z}\left(1 + \frac{m}{z}\left(\mu_B + \int \mathrm{d}\rho_{\mathbf{\Sigma}} \frac{\mu_1 (e^{-2t}\lambda + \Delta_t/m)\tilde{z}}{\tilde{z} - \mu_1(e^{-2t}\lambda + \Delta_t)}\right)\right) \tag{84}$$

$$= -\frac{m}{z} - \frac{m}{z^2}\left(\mu_B + \int \mathrm{d}\rho_{\mathbf{\Sigma}} \frac{\mu_1 (e^{-2t}\lambda + \Delta_t)\tilde{z}}{\tilde{z} - \mu_1(e^{-2t}\lambda + \Delta_t)}\right) \tag{85}$$

Hence at first order

$$q = -\frac{1}{z} - \frac{1}{z^2}\left(\mu_B + \int \mathrm{d}\rho_{\mathbf{\Sigma}} \frac{\mu_1 (e^{-2t}\lambda + \Delta_t)\tilde{z}}{\tilde{z} - \mu_1(e^{-2t}\lambda + \Delta_t)}\right) \tag{86}$$

$$= -\frac{1}{z} - \frac{1}{z^2}\left(\mu_B + \int \mathrm{d}\bar{\rho} \frac{\bar{\lambda}\tilde{z}}{z - \bar{\lambda}}\right) \tag{87}$$

$$= -\frac{1}{z} - \frac{1}{\psi_n m z^2}\left(\psi_n m \mu_B + \int \mathrm{d}\bar{\rho} \frac{\bar{\lambda} z}{z - \bar{\lambda}}\right) \tag{88}$$

by denoting $\bar{\lambda} = \psi_n m \mu_1 (e^{-2t}\lambda + \Delta_t)$. Then

$$\frac{1}{q} = -z + \mu_B + \frac{1}{\psi_n m}\int \mathrm{d}\bar{\rho} \frac{\bar{\lambda} z}{z - \bar{\lambda}} \tag{89}$$

We match this with a Bai–Silverstein equation (Silverstein & Bai, 1995). For $\mathbf{Z} \in \mathbb{R}^{N \times d}$ with i.i.d. standard Gaussian entries and aspect ratio $c = d/N$, the Stieltjes transform $v(z)$ of the spectral density of $\mathbf{Z}\mathbf{\Sigma}'\mathbf{Z}^\top/d$, with $\mathbf{\Sigma}' = \mu_1(e^{-2t}\mathbf{\Sigma} + \Delta_t \mathbf{I}_d)$ of spectral density ρ', satisfies

$$\frac{1}{v(z)} = -z + \int \frac{\mathrm{d}\rho'(\lambda')\,\lambda'}{1 + \lambda' v(z)/c} \sim -z + \int \frac{\mathrm{d}\rho'(\lambda')\,\lambda' z}{z - \lambda'/c} = -z + c\int \frac{\mathrm{d}\bar{\rho}(\bar{\lambda})\,\bar{\lambda} z}{z - \bar{\lambda}}, \tag{90}$$

at first order, using $v \sim -1/z$. With $c = 1/(\psi_n m)$, this is the equation above: the density has total mass $1/(\psi_n m)$, i.e. d eigenvalues as expected, and at leading order the generalization bulk is the non-zero spectrum of

$$\mu_1 \frac{\mathbf{Z}(e^{-2t}\mathbf{\Sigma} + \Delta_t \mathbf{I}_d)\mathbf{Z}^\top}{d}, \qquad \mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_{N \times d}), \tag{91}$$

the constant μ_B being an $O(1)$ shift, below the resolution of this expansion since the bulk has width $\Theta(\sqrt{\psi_n m})$.

Bulk at $z = O(m)$. To study the second bulk, we assume $z = m\tilde{z}$ where $\tilde{z} = O(1)$. In this regime the leading-order behavior of the Stieltjes transforms is fixed by the two atoms found above: the atom at 0 carries a fraction $1 - \frac{1}{m} - \frac{1}{\psi_n m}$ of the spectrum and the atom at $m\mu_B$ a fraction $\frac{1}{m}(1 - \frac{1}{\psi_n})$, so that

$$q \sim -\frac{1}{z}\Big(1 - \frac{1}{m}\Big) + \frac{1}{m(m\mu_B - z)}, \qquad mr \sim \frac{1}{m\mu_B - z}. \tag{92}$$

(Note that $q \sim -1/z$, which would correspond to putting *all* the mass at the origin, is not consistent with the weight $q_0 = 1 - \frac{1}{m} - \frac{1}{\psi_n m}$ obtained in the previous paragraph; keeping the correct weight is what produces the factor $m - 1$ below.) We expand the saddle-point equation $mq = \frac{m-1}{\hat{q}} + \frac{1}{\hat{q}+m\hat{r}}$ by substituting the expressions for $\hat{q}$ and $\hat{r}$ and keeping terms up to $O(\frac{1}{\psi_n})$:

$$\begin{aligned} q &= \frac{m-1}{m}\,\frac{1}{-z + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\frac{\mu_1\Delta_t}{1+\mu_1\psi_n m(e^{-2t}\lambda\, mr+\Delta_t q)}} \\ &\quad + \frac{1}{m}\,\frac{1}{m\mu_B - z + \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\frac{\mu_1(e^{-2t}\lambda m+\Delta_t)}{1+\mu_1\psi_n m(e^{-2t}\lambda\, mr+\Delta_t q)}} \\ &\approx \frac{-1}{z}\Big(1 - \frac{1}{m}\Big) + \frac{1}{m(m\mu_B - z)} \qquad (93) \\ &\quad - \frac{1}{m\psi_n z(m\mu_B - z)} \int \mathrm{d}\rho_{\mathbf{\Sigma}}(\lambda)\frac{me^{-2t}\lambda z^2 + \Delta_t z^2 + (m-1)\Delta_t(m\mu_B - z)^2}{me^{-2t}\lambda z + \Delta_t z - (m-1)\Delta_t(m\mu_B - z)} \qquad (94) \end{aligned}$$

To find the density $\rho(\omega)$ in the spectral variable ω, we use the Stieltjes inversion formula $\rho(\omega) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}[q(\omega + i\epsilon)]$. The imaginary part is generated by the pole in the integrand where the denominator $D(\lambda, \omega) = me^{-2t}\lambda\omega + \Delta_t\omega - (m-1)\Delta_t(m\mu_B - \omega) = m\omega\lambda_t - m(m-1)\Delta_t\mu_B$ vanishes, with $\lambda_t = e^{-2t}\lambda + \Delta_t$. This occurs at:

$$\lambda^*(\omega) = e^{2t}\Delta_t\left(\frac{(m-1)\mu_B}{\omega} - 1\right) \tag{95}$$

equivalently $\omega = (m-1)\mu_B\Delta_t/\lambda_t$, in agreement with the memorization eigenvalue $\lambda_{\mathrm{mem}}(\lambda)$ (at $\mu_I = 0$) obtained by the linear-algebra route in Sect. D.6.2 below.

Applying the Sokhotski–Plemelj theorem $\mathrm{Im}\frac{1}{x-i\epsilon} = \pi\delta(x)$ and using $|\partial_\lambda D| = m\omega e^{-2t}$ at the pole, we obtain, since at $\lambda = \lambda^*(\omega)$ the numerator reduces to $(m-1)\Delta_t\, m\mu_B(m\mu_B - \omega)$,

$$\rho(\omega) = \frac{e^{2t}\Delta_t(m-1)\mu_B}{\psi_n m\,\omega^2}\,\rho_{\mathbf{\Sigma}}\left(e^{2t}\Delta_t\left(\frac{(m-1)\mu_B}{\omega} - 1\right)\right) \tag{96}$$

which is exactly the pushforward of $\rho_{\mathbf{\Sigma}}$ by the map $\lambda \mapsto (m-1)\mu_B\Delta_t/\lambda_t$, carrying total mass

$$\int \mathrm{d}\omega\,\rho(\omega) = \frac{1}{\psi_n m}, \tag{97}$$

i.e. exactly d eigenvalues. Restoring the shift μ_I amounts to replacing ω by $\omega - \mu_I$ in (96).

Conclusion. The density of eigenvalues is composed of

- a delta peak at $\lambda = 0$ with $nm - n - d$ eigenvalues, the null space of $\mathbf{G}$ due to its finite rank;
- a bulk at $\lambda = \Theta(m)$ with d eigenvalues, whose density is given by (96);
- a delta peak at $\lambda = \mu_B m$ with $n - d$ eigenvalues;
- a bulk at $\lambda = \Theta(\psi_n m)$ with d eigenvalues, which is that of the population kernel operator of the noisy distribution P_t.

D.6.2 Derivation with linear algebra arguments

In this subsection we derive the eigenvalues and eigenvectors of the linearized Gram matrix. Besides re-deriving the eigenvalues of Theorem D.2, this route gives the eigenvectors, and proves the following proposition, which restates Proposition 2.1 of the main text.

Proposition D.1 (Eigenvectors of the Gram matrix)**.** *For $\psi_n \gg 1$ and every eigenvalue λ of $\mathbf{\Sigma}$ with associated eigenvector $\mathbf{v}_\lambda$, the eigenvectors of $\mathbf{G}$ associated with the generalization and the memorization bulks of Theorem D.2 are, at leading order,*

$$\mathbf{u}_1^{\nu\alpha} \propto \mathbf{v}_\lambda^\top \mathbf{Y}^{\nu\alpha}, \qquad \mathbf{u}_2^{\nu\alpha} \propto \mathbf{v}_\lambda^\top \left[(me^{-2t}\lambda + \Delta_t)\sqrt{\Delta_t}\,\boldsymbol{\xi}_\perp^{\nu\alpha} - (m-1)\Delta_t\,\bar{\mathbf{Y}}^\nu\right], \tag{98}$$

with $\bar{\mathbf{Y}}^\nu = \frac{1}{m}\sum_\beta \mathbf{Y}^{\nu\beta}$ and $\boldsymbol{\xi}_\perp^{\nu\alpha} = \boldsymbol{\xi}^{\nu\alpha} - \frac{1}{m}\sum_\beta \boldsymbol{\xi}^{\nu\beta}$. The eigenvectors of the two atoms are the $\boldsymbol{\varphi} \otimes \mathbf{1}_m/\sqrt{m}$ with $\boldsymbol{\varphi} \in \mathrm{Ker}(\bar{\mathbf{Y}})$ and the elements of $\mathrm{Ker}(\mathbf{Y}) \cap \mathrm{Ker}(\mathbf{B}_m)$; they do not contribute to the estimated score.

Proof. As above we set $\mu_I = 0$, which only shifts the whole spectrum by μ_I, and we drop the rank-one term $\mu_0 \mathbf{1}_N \mathbf{1}_N^\top$, which contributes a single outlier outside the bulk. We are left with the spectrum of the $nm \times nm$ matrix

$$\mathbf{G} = \mu_B \mathbf{B}_m + \mu_1 \frac{\mathbf{Y}^\top \mathbf{Y}}{d} \tag{99}$$

Let $\mathbf{e}_{\nu\alpha}$ denote the canonical orthonormal basis, where $\nu \in \{1, \dots, n\}$ and $\alpha \in \{1, \dots, m\}$. We define the n block-averaged vectors

$$\mathbf{u}^\nu = \frac{1}{\sqrt{m}} \sum_{\alpha=1}^m \mathbf{e}_{\nu\alpha} \tag{100}$$

which are orthonormal.

We introduce the d feature vectors $\boldsymbol{\phi}_i$

$$\boldsymbol{\phi}_i = \sqrt{m}e^{-t} \sum_\nu \mathbf{x}_i^\nu \mathbf{u}^\nu + \sqrt{\Delta_t} \sum_{\nu,\alpha} \boldsymbol{\xi}_i^{\nu\alpha} \mathbf{e}_{\nu\alpha} \tag{101}$$

where $\mathbf{x}^\nu \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $\boldsymbol{\xi}^{\nu\alpha} \sim \mathcal{N}(0, \mathbf{I}_d)$ are mutually independent. We can write our Gram matrix as

$$\mathbf{G} = \mu_B m \sum_\nu \mathbf{u}^\nu (\mathbf{u}^\nu)^\top + \frac{\mu_1}{d} \sum_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \tag{102}$$

We study the asymptotic spectrum in the limit $d \gg 1$ and $\psi_n = n/d \gg 1$, keeping m finite. Let λ and $\mathbf{v}_i^\lambda$ denote the eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}$.

Because m is finite, the block-averaged noise is not negligible. We split the noise into its block-average and a transverse component:

$$\boldsymbol{\eta}_i^\nu = \frac{1}{\sqrt{m}} \sum_{\alpha=1}^m \boldsymbol{\xi}_i^{\nu\alpha}, \qquad \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} = \boldsymbol{\xi}_i^{\nu\alpha} - \frac{1}{\sqrt{m}} \boldsymbol{\eta}_i^\nu \tag{103}$$

By definition, $\sum_\alpha \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} = 0$. Projecting onto the eigenbasis of $\mathbf{\Sigma}$, we define the coordinates:

$$z^{\lambda\nu} = \frac{1}{\sqrt{\lambda}} \sum_i \mathbf{v}_i^\lambda \mathbf{x}_i^\nu, \qquad \bar{t}^{\lambda\nu} = \sum_i \mathbf{v}_i^\lambda \boldsymbol{\eta}_i^\nu, \qquad s^{\lambda\nu\alpha} = \sum_i \mathbf{v}_i^\lambda \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} \tag{104}$$

By Gaussianity and the orthogonality of the $\mathbf{v}_i^\lambda$, the variables $z^{\lambda\nu}$ and $\bar{t}^{\lambda\nu}$ are independent $\mathcal{N}(0,1)$. The transverse variables $s^{\lambda\nu\alpha}$ are centered, orthogonal to the all-ones vector in α, and satisfy

$$\mathbb{E}[s^{\lambda\nu\alpha} s^{\lambda\nu\beta}] = \delta_{\alpha\beta} - \frac{1}{m} \tag{105}$$

Thus, for large n, $\sum_{\nu,\alpha} (s^{\lambda\nu\alpha})^2 \simeq n(m-1)$.

For each eigenvalue λ, we define three normalized vectors:

$$\mathbf{U}^\lambda = \frac{1}{\sqrt{n}} \sum_\nu z^{\lambda\nu} \mathbf{u}^\nu \tag{106}$$

$$\mathbf{X}^\lambda = \frac{1}{\sqrt{n}} \sum_\nu \bar{t}^{\lambda\nu} \mathbf{u}^\nu \tag{107}$$

$$\mathbf{P}^\lambda = \frac{1}{\sqrt{n(m-1)}} \sum_{\nu,\alpha} s^{\lambda\nu\alpha} \mathbf{e}_{\nu\alpha} \tag{108}$$

Because z, $\bar{t}$, and s are independent and we are in the high-dimensional limit $n \gg 1$, the family $\{\mathbf{U}^\lambda, \mathbf{X}^\lambda, \mathbf{P}^\lambda\}_{\lambda \in \mathrm{Spec}(\mathbf{\Sigma})}$ is asymptotically orthonormal. Furthermore, different λ sectors do not mix at leading order.

We compute the overlaps of the feature vectors $\boldsymbol{\phi}_i$ with our basis. By standard concentration of measure, we find:

$$\boldsymbol{\phi}_i^\top \mathbf{U}^\lambda \simeq \sqrt{mn\lambda} e^{-t} \mathbf{v}_i^\lambda \tag{109}$$

$$\boldsymbol{\phi}_i^\top \mathbf{X}^\lambda \simeq \sqrt{n\Delta_t} \mathbf{v}_i^\lambda \tag{110}$$

$$\boldsymbol{\phi}_i^\top \mathbf{P}^\lambda \simeq \sqrt{n(m-1)\Delta_t} \mathbf{v}_i^\lambda \tag{111}$$

Let $\kappa = \mu_1 \frac{n}{d}$. The operator $\mu_B \mathbf{B}_m$ acts as $\mu_B m \, \mathrm{diag}(1,1,0)$ on the subspace $(\mathbf{U}^\lambda, \mathbf{X}^\lambda, \mathbf{P}^\lambda)$, because $\mathbf{P}^\lambda$ is entirely transverse to the block structure. The sample covariance part yields a rank-one update. Defining the vector $\mathbf{w}_\lambda = (\sqrt{m\lambda} e^{-t}, \sqrt{\Delta_t}, \sqrt{(m-1)\Delta_t})^\top$, the operator $\mathbf{G}$ projected onto this 3×3 subspace is asymptotically:

$$\mathbf{M}_\lambda^{(3)} \simeq \mu_B m \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \kappa \mathbf{w}_\lambda \mathbf{w}_\lambda^\top \tag{112}$$

To block-diagonalize this system, we introduce $D_\lambda = m\lambda e^{-2t} + \Delta_t$ and rotate the parallel sector into a mode coupled to the perturbation and an orthogonal mode:

$$\mathbf{\Phi}^\lambda = \frac{\sqrt{m\lambda} e^{-t} \mathbf{U}^\lambda + \sqrt{\Delta_t} \mathbf{X}^\lambda}{\sqrt{D_\lambda}} \tag{113}$$

$$\mathbf{\Phi}_\perp^\lambda = \frac{\sqrt{\Delta_t} \mathbf{U}^\lambda - \sqrt{m\lambda} e^{-t} \mathbf{X}^\lambda}{\sqrt{D_\lambda}} \tag{114}$$

By construction, $\mathbf{\Phi}_\perp^\lambda$ is orthogonal to $\mathbf{w}_\lambda$ and remains an exact eigenvector of $\mathbf{G}$ with eigenvalue $\mu_B m$. The non-trivial spectrum is governed by the 2×2 matrix acting on the remaining coupled subspace $(\mathbf{\Phi}^\lambda, \mathbf{P}^\lambda)$:

$$\boldsymbol{\mathcal{M}}^\lambda \simeq \begin{pmatrix} \mu_B m + \kappa D_\lambda & \kappa\sqrt{(m-1)\Delta_t D_\lambda} \\ \kappa\sqrt{(m-1)\Delta_t D_\lambda} & \kappa(m-1)\Delta_t \end{pmatrix} \tag{115}$$

The characteristic equation for $\boldsymbol{\mathcal{M}}^\lambda$ yields two eigenvalues for each λ, $\lambda_{\mathrm{gen}}(\lambda)$ (upper sign) and $\lambda_{\mathrm{mem}}(\lambda)$ (lower sign):

$$\begin{aligned} \lambda_{\mathrm{gen/mem}}(\lambda) = &\frac{\mu_B m + \kappa(D_\lambda + (m-1)\Delta_t)}{2} \\ &\pm \frac{1}{2}\sqrt{\left[\mu_B m + \kappa(D_\lambda - (m-1)\Delta_t)\right]^2 + 4\kappa^2 (m-1)\Delta_t D_\lambda} \end{aligned} \tag{116}$$

In the strict limit $\kappa = \mu_1 \frac{n}{d} \gg 1$, we can expand these roots to isolate the dominant scaling:

$$\lambda_{\mathrm{gen}}(\lambda) = \mu_1 \frac{nm}{d} \left(e^{-2t}\lambda + \Delta_t\right) + \frac{\mu_B D_\lambda}{e^{-2t}\lambda + \Delta_t} + O\left(\frac{d}{n}\right) \tag{117}$$

$$\lambda_{\mathrm{mem}}(\lambda) = \frac{\mu_B (m-1)\Delta_t}{e^{-2t}\lambda + \Delta_t} + O\left(\frac{d}{n}\right) \tag{118}$$

This confirms that the spectrum separates into a large bulk scaling with $\psi_n = n/d$, which depends on the signal covariance $\mathbf{\Sigma}$, and a smaller memorization bulk whose eigenvalues $\lambda_{\mathrm{mem}}(\lambda)$ are related to the spectrum of $\mathbf{\Sigma}$ via the exact mapping:

$$\lambda_{\mathrm{mem}}(\lambda) = \frac{\mu_B (m-1)\Delta_t}{\lambda e^{-2t} + \Delta_t} \tag{119}$$

Eigenvectors. We have established that for each covariance mode λ, the operator $\mathbf{G} = \mu_B \mathbf{B}_m + \mu_1 \frac{\mathbf{Y}^\top \mathbf{Y}}{d}$ acts non-trivially on the three-dimensional subspace spanned by the asymptotically orthonormal vectors $(\mathbf{U}^\lambda, \mathbf{X}^\lambda, \mathbf{P}^\lambda)$ introduced above, the sample covariance acting as a rank-one perturbation along $\mathbf{w}_\lambda$. We now construct the eigenvectors of $\mathbf{G}$ explicitly and relate them to the raw signal components $\mathbf{x}_i^\nu$ and noise components $\boldsymbol{\xi}_i^{\nu\alpha}$.

The direction orthogonal to the perturbation $\mathbf{w}_\lambda$ within the parallel subspace $(\mathbf{U}^\lambda, \mathbf{X}^\lambda)$ remains an exact eigenvector of $\mathbf{G}$ with the unshifted eigenvalue $\mu_B m$. With $D_\lambda = m\lambda e^{-2t} + \Delta_t$ as above, this vector is

$$\mathbf{\Phi}_\perp^\lambda = \frac{\sqrt{\Delta_t}\mathbf{U}^\lambda - \sqrt{m\lambda}e^{-t}\mathbf{X}^\lambda}{\sqrt{D_\lambda}} \tag{120}$$

We now substitute the definitions of $\mathbf{U}^\lambda$ and $\mathbf{X}^\lambda$ in terms of the raw coordinates $z^{\lambda\nu} = \frac{1}{\sqrt{\lambda}} \sum_i \mathbf{v}_i^\lambda \mathbf{x}_i^\nu$ and $\bar{t}^{\lambda\nu} = \sum_i \mathbf{v}_i^\lambda \boldsymbol{\eta}_i^\nu$, where $\boldsymbol{\eta}_i^\nu = \frac{1}{\sqrt{m}} \sum_\alpha \boldsymbol{\xi}_i^{\nu\alpha}$ is the block-averaged noise.

$$\mathbf{\Phi}_\perp^\lambda = \frac{1}{\sqrt{nD_\lambda}} \sum_{i=1}^d \mathbf{v}_i^\lambda \sum_{\nu=1}^n \left(\sqrt{\frac{\Delta_t}{\lambda}} \mathbf{x}_i^\nu - \sqrt{m\lambda} e^{-t} \boldsymbol{\eta}_i^\nu \right) \mathbf{u}^\nu \tag{121}$$

The remaining two eigenvectors are orthogonal linear combinations of the coupled parallel state $\mathbf{\Phi}^\lambda$ and the transverse noise state $\mathbf{P}^\lambda$, parametrized by a mixing angle θ_λ (we call them $\mathbf{E}_{\lambda,\pm}$, rather than $\mathbf{u}$, to avoid a clash with the block-averaged basis $\mathbf{u}^\nu$; $\mathbf{E}_{\lambda,+} \propto \mathbf{u}_1$ and $\mathbf{E}_{\lambda,-} \propto \mathbf{u}_2$ are the eigenvectors of Proposition D.1, with eigenvalues $\lambda_{\text{gen}}(\lambda)$ and $\lambda_{\text{mem}}(\lambda)$ respectively):

$$\mathbf{E}_{\lambda,+} = \cos\theta_\lambda \mathbf{\Phi}^\lambda + \sin\theta_\lambda \mathbf{P}^\lambda \tag{122}$$

$$\mathbf{E}_{\lambda,-} = -\sin\theta_\lambda \mathbf{\Phi}^\lambda + \cos\theta_\lambda \mathbf{P}^\lambda \tag{123}$$

where $\mathbf{\Phi}^\lambda = \frac{1}{\sqrt{D_\lambda}} \left(\sqrt{m\lambda} e^{-t} \mathbf{U}^\lambda + \sqrt{\Delta_t} \mathbf{X}^\lambda \right)$.

The angle is determined by the effective 2×2 block $\boldsymbol{\mathcal{M}}^\lambda$. In the asymptotic limit $n/d \gg 1$, the coupling parameter $\kappa = \mu_1 \frac{n}{d} \to \infty$, and the mixing angle simplifies to:

$$\cos\theta_\lambda \approx \sqrt{\frac{D_\lambda}{D_\lambda + (m-1)\Delta_t}}, \qquad \sin\theta_\lambda \approx \sqrt{\frac{(m-1)\Delta_t}{D_\lambda + (m-1)\Delta_t}} \tag{124}$$

Substituting the asymptotic angle into $\mathbf{E}_{\lambda,+}$, we obtain:

$$\mathbf{E}_{\lambda,+} \approx \frac{1}{\sqrt{D_\lambda + (m-1)\Delta_t}} \left(\sqrt{D_\lambda} \mathbf{\Phi}^\lambda + \sqrt{(m-1)\Delta_t} \mathbf{P}^\lambda \right) \tag{125}$$

Expanding $\mathbf{\Phi}^\lambda$, the $\sqrt{D_\lambda}$ factor cancels:

$$\mathbf{E}_{\lambda,+} \propto \sqrt{m\lambda} e^{-t} \mathbf{U}^\lambda + \sqrt{\Delta_t} \mathbf{X}^\lambda + \sqrt{(m-1)\Delta_t} \mathbf{P}^\lambda \tag{126}$$

We now expand this entirely into the raw variables. Recall that $s^{\lambda\nu\alpha} = \sum_i \mathbf{v}_i^\lambda \boldsymbol{\xi}_{i,\perp}^{\nu\alpha}$, where $\boldsymbol{\xi}_{i,\perp}^{\nu\alpha}$ is the transverse noise.

$$\mathbf{E}_{\lambda,+} \propto \frac{1}{\sqrt{n}} \sum_{i=1}^d \mathbf{v}_i^\lambda \sum_{\nu=1}^n \left(\sqrt{m} e^{-t} \mathbf{x}_i^\nu \mathbf{u}^\nu + \sqrt{\Delta_t} \boldsymbol{\eta}_i^\nu \mathbf{u}^\nu + \sqrt{\Delta_t} \sum_{\alpha=1}^m \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} \mathbf{e}_{\nu\alpha} \right) \tag{127}$$

Recognizing that the block-averaged noise and transverse noise reconstruct the full noise vector exactly ($\boldsymbol{\eta}_i^\nu \mathbf{u}^\nu + \sum_\alpha \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} \mathbf{e}_{\nu\alpha} = \sum_\alpha \boldsymbol{\xi}_i^{\nu\alpha} \mathbf{e}_{\nu\alpha}$), the term in parentheses is precisely the empirical data vector $\boldsymbol{\phi}_i$. Thus,

$$\mathbf{E}_{\lambda,+} \propto \frac{1}{\sqrt{n}} \sum_{i=1}^d \mathbf{v}_i^\lambda \boldsymbol{\phi}_i \tag{128}$$

Using the orthogonal rotation, the eigenvector corresponding to the memorization bulk is:

$$\mathbf{E}_{\lambda,-} \approx \frac{1}{\sqrt{D_\lambda + (m-1)\Delta_t}} \left(\sqrt{(m-1)\Delta_t} \mathbf{\Phi}^\lambda - \sqrt{D_\lambda} \mathbf{P}^\lambda \right) \tag{129}$$

Dropping the overall normalization and substituting the basis vectors, we express this mode in terms of the raw variables:

$$\mathbf{E}_{\lambda,-} \propto \sum_{i=1}^{d} \mathbf{v}_i^\lambda \sum_{\nu=1}^{n} \left[\sqrt{(m-1)\Delta_t} \left(\sqrt{m} e^{-t} \mathbf{x}_i^\nu + \sqrt{\Delta_t} \boldsymbol{\eta}_i^\nu \right) \mathbf{u}^\nu - \frac{D_\lambda}{\sqrt{m-1}} \sum_{\alpha=1}^{m} \boldsymbol{\xi}_{i,\perp}^{\nu\alpha} \mathbf{e}_{\nu\alpha} \right] \quad (130)$$

This proves Proposition D.1. Indeed $\mathbf{E}_{\lambda,+} \propto \frac{1}{\sqrt{n}} \sum_i \mathbf{v}_i^\lambda \boldsymbol{\phi}_i$ has components $\mathbf{v}_\lambda^\top \mathbf{Y}^{\nu\alpha}$, which is $\mathbf{u}_1$. For $\mathbf{E}_{\lambda,-}$, using $\mathbf{u}^\nu = \frac{1}{\sqrt{m}} \sum_\alpha \mathbf{e}_{\nu\alpha}$ and $\boldsymbol{\eta}_i^\nu = \sqrt{m}\, \bar{\boldsymbol{\xi}}_i^\nu$, the bracket has $(\nu\alpha)$ component $\sqrt{(m-1)\Delta_t}\, \bar{\mathbf{Y}}^\nu - \frac{D_\lambda}{\sqrt{m-1}} \boldsymbol{\xi}_\perp^{\nu\alpha}$ once projected on $\mathbf{v}_\lambda$; multiplying by $-\sqrt{(m-1)\Delta_t}$ gives $D_\lambda \sqrt{\Delta_t}\, \boldsymbol{\xi}_\perp^{\nu\alpha} - (m-1)\Delta_t\, \bar{\mathbf{Y}}^\nu$, which is $\mathbf{u}_2$ with $D_\lambda = m e^{-2t}\lambda + \Delta_t$. Finally $\boldsymbol{\Phi}_\perp^\lambda$ and the kernel of $\mathbf{G}$ give the eigenvectors of the two atoms.

The atoms do not contribute to the estimated score. By (5), the k-th coordinate of the estimator is $\mathbf{s}^k(\mathbf{y}) = K(\mathbf{y}, \mathbf{Y})^\top h_\tau(\mathbf{G})\, (-\boldsymbol{\xi}_k/\sqrt{\Delta_t})$, where $\boldsymbol{\xi}_k \in \mathbb{R}^N$ collects the k-th coordinates of the training noises and $h_\tau(\lambda) = (1 - e^{-d\tau\lambda/(nm)})/\lambda$ (or $h(\lambda) = 1/(\lambda+\gamma)$ with a ridge). Let $\mathbf{P}$ be the orthogonal projector on an eigenspace E of $\mathbf{G}_{\text{lin}}$ on which $\mathbf{G}_{\text{lin}}$ acts as $c\,\mathbf{I}$. The contribution of E to the estimator is

$$\mathbf{s}_E^k(\mathbf{y}) = -\frac{h_\tau(c)}{\sqrt{\Delta_t}}\, K(\mathbf{y}, \mathbf{Y})^\top \mathbf{P}\, \boldsymbol{\xi}_k, \quad (131)$$

which vanishes at every τ and every ridge as soon as $\mathbf{P} K(\mathbf{y}, \mathbf{Y}) = 0$. We now restore the rank-one term and take the full linear equivalent $\mathbf{G}_{\text{lin}} = \mu_I \mathbf{I}_N + \mu_B \mathbf{B}_m + \mu_0 \mathbf{1}_N \mathbf{1}_N^\top + \mu_1 \mathbf{Y}^\top \mathbf{Y}/d$, and the linearized test kernel $K_{\text{lin}}(\mathbf{y}, \mathbf{Y}) = \mu_0 \mathbf{1}_N + \mu_1 \mathbf{Y}^\top \mathbf{y}/d$, valid for a test point $\mathbf{y}$ independent of the training set since $\mathbf{y}^\top \mathbf{Y}^{\nu\alpha}/d = O(d^{-1/2})$.

- *Atom at $\bar{\mu} = \mu_I + m\mu_B$.* For $\boldsymbol{\varphi} \in \mathbb{R}^n$, $\mathbf{G}_{\text{lin}}(\boldsymbol{\varphi} \otimes \mathbf{1}_m) = \bar{\mu}\, \boldsymbol{\varphi} \otimes \mathbf{1}_m + \mu_0 m (\mathbf{1}_n^\top \boldsymbol{\varphi}) \mathbf{1}_N + \mu_1 m\, \mathbf{Y}^\top \bar{\mathbf{Y}} \boldsymbol{\varphi}/d$. The eigenspace is therefore $E_{\text{par}} = \{\boldsymbol{\varphi} \otimes \mathbf{1}_m : \boldsymbol{\varphi} \in \text{Ker}(\bar{\mathbf{Y}}),\ \mathbf{1}_n^\top \boldsymbol{\varphi} = 0\}$. The second condition, which comes from the rank-one term, removes a single dimension out of $n - d$ and leaves the weight of the atom unchanged. For $\boldsymbol{\varphi} \otimes \mathbf{1}_m \in E_{\text{par}}$,

$$(\boldsymbol{\varphi} \otimes \mathbf{1}_m)^\top K_{\text{lin}}(\mathbf{y}, \mathbf{Y}) = \mu_0 m\, \mathbf{1}_n^\top \boldsymbol{\varphi} + \frac{\mu_1 m}{d}\, \mathbf{y}^\top \bar{\mathbf{Y}} \boldsymbol{\varphi} = 0. \quad (132)$$

- *Atom at μ_I.* For $\boldsymbol{\psi} \in E_{\text{trans}} = \text{Ker}(\mathbf{Y}) \cap \text{Ker}(\mathbf{B}_m)$, $\mathbf{B}_m \boldsymbol{\psi} = 0$ means $\sum_\alpha \boldsymbol{\psi}^{\nu\alpha} = 0$ for every ν, hence $\mathbf{1}_N^\top \boldsymbol{\psi} = 0$, and $\mathbf{Y} \boldsymbol{\psi} = 0$; thus $\mathbf{G}_{\text{lin}} \boldsymbol{\psi} = \mu_I \boldsymbol{\psi}$ and

$$\boldsymbol{\psi}^\top K_{\text{lin}}(\mathbf{y}, \mathbf{Y}) = \mu_0\, \mathbf{1}_N^\top \boldsymbol{\psi} + \frac{\mu_1}{d}\, \mathbf{y}^\top \mathbf{Y} \boldsymbol{\psi} = 0. \quad (133)$$

For the linearized kernel both atoms therefore contribute exactly zero to the estimated score, for every training time and every ridge. The non-linear remainder $K_{\text{res}} = K - K_{\text{lin}}$ has entries $f(\mathbf{y}^\top \mathbf{Y}^{\nu\alpha}/d) - f(0) - f'(0) \mathbf{y}^\top \mathbf{Y}^{\nu\alpha}/d = O(d^{-1})$. Its mean over the entries is $\simeq f''(0) \mathbf{y}^\top \boldsymbol{\Sigma}_t \mathbf{y}/(2d^2)$, i.e. proportional to $\mathbf{1}_N$, and is annihilated by both projectors, since $\mathbf{1}_N \perp E_{\text{par}}$ and $\mathbf{1}_N \perp E_{\text{trans}}$. Its fluctuating part has norm $O(\sqrt{N}/d) = O(\sqrt{m\psi_n/d})$; if it is asymptotically uncorrelated with $\mathbf{P}\boldsymbol{\xi}_k$, its contribution is of the same order and vanishes as $d \to \infty$ at fixed ψ_n and m. □

D.7 Derivation of the spectrum in the polynomial scaling

For the polynomial regime we assume that the eigenfunctions of the kernel are given by a family of orthogonal polynomials.

Lemma D.2. *Let P be a probability measure on $\mathbb{R}^d$ with finite moments of all orders. Assume further that the space of multivariate polynomials is dense in $L^2(P)$. Then, there exists a countable family of polynomials $\{\psi_p^{\boldsymbol{j}}\}_{\boldsymbol{j} \in \mathbb{N}^d}$, where $p = |\boldsymbol{j}|$ denotes the total degree, that forms an orthogonal basis of $L^2(P)$.*

Proof. We equip $L^2(P)$ with the inner product $\langle f, g \rangle = \int_{\mathbb{R}^d} f(x) g(x)\, \mathrm{d}P(x)$. Let $\boldsymbol{j} = (j_1, \ldots, j_d) \in \mathbb{N}^d$ be a multi-index, with total degree $p = |\boldsymbol{j}| = \sum_{i=1}^d j_i$. Because P has finite moments of all orders, every monomial $x^{\boldsymbol{j}} = x_1^{j_1} \cdots x_d^{j_d}$ belongs to $L^2(P)$.

To construct an orthogonal basis, we establish a strict well-ordering on the set of multi-indices $\mathbb{N}^d$. We choose the graded lexicographic order, denoted by $\prec$. For two multi-indices $\boldsymbol{j}, \boldsymbol{l}$, we say $\boldsymbol{j} \prec \boldsymbol{l}$ if $|\boldsymbol{j}| < |\boldsymbol{l}|$, or if $|\boldsymbol{j}| = |\boldsymbol{l}|$ and the first non-zero entry in the difference $\boldsymbol{j} - \boldsymbol{l}$ is negative.

We now apply the Gram–Schmidt orthogonalization process to the sequence of monomials ordered by $\prec$, by induction along $\prec$ (every multi-index has finitely many predecessors).

Let $\mathbf{0} = (0, \dots, 0)$. For the base case (where $p = 0$), we define:

$$\psi_0^{\mathbf{0}}(x) = 1. \tag{134}$$

For the inductive step, let $\boldsymbol{j} \in \mathbb{N}^d$ with total degree $p = |\boldsymbol{j}|$. Assume we have constructed a mutually orthogonal set of polynomials $\left\{\psi_q^{\boldsymbol{l}}(x)\right\}_{\boldsymbol{l} \prec \boldsymbol{j}}$, where $q = |\boldsymbol{l}|$. We define the polynomial corresponding to $\boldsymbol{j}$ by projecting the monomial $x^{\boldsymbol{j}}$ onto the orthogonal complement of the span of all preceding polynomials:

$$\psi_p^{\boldsymbol{j}}(x) = x^{\boldsymbol{j}} - \sum_{\boldsymbol{l} \prec \boldsymbol{j}} \frac{\langle x^{\boldsymbol{j}}, \psi_q^{\boldsymbol{l}} \rangle}{\langle \psi_q^{\boldsymbol{l}}, \psi_q^{\boldsymbol{l}} \rangle} \psi_q^{\boldsymbol{l}}(x). \tag{135}$$

By construction, $\psi_p^{\boldsymbol{j}}(x)$ is a polynomial of total degree p with leading term $x^{\boldsymbol{j}}$, and $\langle \psi_p^{\boldsymbol{j}}, \psi_q^{\boldsymbol{l}} \rangle = 0$ for all $\boldsymbol{l} \prec \boldsymbol{j}$.

This process yields a countable family of mutually orthogonal polynomials. Because the span of this family equals the span of all multivariate monomials, and polynomials are dense in $L^2(P)$, this family $\left\{\psi_p^{\boldsymbol{j}}\right\}_{\boldsymbol{j} \in \mathbb{N}^d}$ is an orthogonal basis of $L^2(P)$. □

Remark. *The orthogonal basis $\{\psi_p^{\boldsymbol{j}}\}$ constructed above reduces to classical families for the standard data distributions: products of one-dimensional Hermite polynomials for the standard Gaussian (Bach, 2023), and spherical harmonics for data uniform on the sphere.*

This kernel form is natural because, for data uniform on the sphere, every dot-product kernel is diagonalized by spherical harmonics (Ghorbani et al., 2021; Bietti & Mairal, 2019).

We first derive a polynomial equivalent of the Gram matrix in the regime $d^{k+1} \gg n \gg d^k$ for a kernel satisfying the diagonal hypothesis (49), with $\psi_p^{\boldsymbol{j}}$ the polynomials of Lemma D.2.

Proposition D.2 (Polynomial Equivalent of the Gram matrix)**.** *In the regime $d^{k+1} \gg n \gg d^k \gg 1$, the Gram matrix is equivalent to*

$$\mathbf{G} = \mu_I^{>k} \mathbf{I}_N + \mu_B^{>k} \mathbf{B}_m + \sum_{p=0}^{k} \sum_{|\boldsymbol{j}|=p} \frac{\mu_p}{d^p \boldsymbol{j}!} \psi_p^{\boldsymbol{j}}(\mathbf{Y}) \psi_p^{\boldsymbol{j}}(\mathbf{Y})^\top \tag{136}$$

with the constants defined as

$$\begin{aligned} \mu_B^{>k} &= \sum_{p>k} \frac{\mu_p}{d^p} \mathbb{E}_{\mathbf{Y}, \mathbf{Y}'} \left[\sum_{|\boldsymbol{j}|=p} \frac{\psi_p^{\boldsymbol{j}}(\mathbf{Y})\, \psi_p^{\boldsymbol{j}}(\mathbf{Y}')}{\boldsymbol{j}!} \right], \\ \mu_I^{>k} &= \sum_{p>k} \frac{\mu_p}{d^p} \mathbb{E}_{\mathbf{Y}} \left[\sum_{|\boldsymbol{j}|=p} \frac{\psi_p^{\boldsymbol{j}}(\mathbf{Y})^2}{\boldsymbol{j}!} \right] - \mu_B^{>k}, \end{aligned} \tag{137}$$

with $\mathbf{Y}, \mathbf{Y}'$ two conditionally independent same-cluster noisy copies (i.e., sharing the underlying $\mathbf{x}^\nu$ but with independent noise $\boldsymbol{\xi}$).

Proof. We split the kernel by polynomial degree at the cutoff k:

$$K = K_{\le k} + K_{>k}, \qquad K_{\le k}(\mathbf{x}, \mathbf{y}) := \sum_{p=0}^{k} \sum_{|\boldsymbol{j}|=p} \frac{\mu_p}{d^p \boldsymbol{j}!} \psi_p^{\boldsymbol{j}}(\mathbf{x})\, \psi_p^{\boldsymbol{j}}(\mathbf{y}), \tag{138}$$

which induces $\mathbf{G} = \mathbf{G}_{\le k} + \mathbf{G}_{>k}$. The low-degree part is, by construction, exactly the explicit double sum in the statement. We must therefore show that, in operator norm,

$$\mathbf{G}_{>k} \to \mu_I^{>k}\,\mathbf{I}_N \;+\; \mu_B^{>k}\,\mathbf{B}_m \tag{139}$$

We treat the three index regimes (diagonal, same cluster, distinct clusters) separately.

Recall $\mathbf{B}_m = \mathbf{I}_n \otimes \mathbf{1}_m\mathbf{1}_m^\top$ has entries $(\mathbf{B}_m)_{(\nu\alpha),(\mu\beta)} = \mathbf{1}\{\nu = \mu\}$. The matrix $\mu_I^{>k}\,\mathbf{I}_N + \mu_B^{>k}\,\mathbf{B}_m$ takes the value $\mu_I^{>k} + \mu_B^{>k}$ on the diagonal, $\mu_B^{>k}$ on the same-cluster off-diagonal, and 0 on cross-cluster pairs. We verify each.

Diagonal entries. For $(\nu,\alpha) = (\mu,\beta)$:

$$\big(\mathbf{G}_{>k}\big)_{(\nu\alpha),(\nu\alpha)} \;=\; K_{>k}(\mathbf{Y}^{\nu\alpha},\mathbf{Y}^{\nu\alpha}) \;=\; \sum_{p>k}\frac{\mu_p}{d^p}\sum_{|\boldsymbol{j}|=p}\frac{\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})^2}{\boldsymbol{j}!}.$$

Each summand is a function of the single sample $\mathbf{Y}^{\nu\alpha}$ alone. By the law of large numbers in $L^2(P_t)$, the inner sum $\sum_{|\boldsymbol{j}|=p}\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})^2/\boldsymbol{j}!$ concentrates around its expectation under $\mathbf{Y}^{\nu\alpha}\sim P_t$, with fluctuations a factor $\sqrt{1/d^p}$ smaller than the mean (since the rank of the Mercer block at degree p is $\sim d^p/p!$). Summing over $p>k$ and dividing by d^p leaves the deterministic value

$$\big(\mathbf{G}_{>k}\big)_{(\nu\alpha),(\nu\alpha)} \;=\; \mu_I^{>k} + \mu_B^{>k} + o_P(1)$$

uniformly in (ν,α).

Same-cluster off-diagonal entries. For $\nu=\mu, \alpha\neq\beta$, $\mathbf{Y}^{\nu\alpha}$ and $\mathbf{Y}^{\nu\beta}$ share the underlying $\mathbf{x}^\nu$ but are conditionally independent given $\mathbf{x}^\nu$ (through the noise $\boldsymbol{\xi}$):

$$\big(\mathbf{G}_{>k}\big)_{(\nu\alpha),(\nu\beta)} \;=\; K_{>k}(\mathbf{Y}^{\nu\alpha},\mathbf{Y}^{\nu\beta}) \;=\; \sum_{p>k}\frac{\mu_p}{d^p}\sum_{|\boldsymbol{j}|=p}\frac{\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\beta})}{\boldsymbol{j}!}.$$

Conditional on $\mathbf{x}^\nu$, each summand has expectation $\mathbb{E}[\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})\,|\,\mathbf{x}^\nu]\cdot\mathbb{E}[\psi_p^{\boldsymbol{j}}(\mathbf{Y}^{\nu\beta})\,|\,\mathbf{x}^\nu]$. Marginalizing over $\mathbf{x}^\nu$, this is precisely the joint expectation appearing in the definition of $\mu_B^{>k}$. The same concentration argument as for the diagonal entries gives, uniformly in $\nu,\alpha\neq\beta$,

$$\big(\mathbf{G}_{>k}\big)_{(\nu\alpha),(\nu\beta)} \;=\; \mu_B^{>k} + o_P(1).$$

Cross-cluster entries. For $\nu\neq\mu$, $\mathbf{Y}^{\nu\alpha}$ and $\mathbf{Y}^{\mu\beta}$ are fully independent. By orthogonality of $\{\psi_p^{\boldsymbol{j}}\}_{p\ge1}$ to constants in $L^2(P_t)$, $\mathbb{E}[\psi_p^{\boldsymbol{j}}(\mathbf{Y})]=0$ for all $p\ge1$, and therefore for $p>k\ge1$,

$$\mathbb{E}\Big[\big(\mathbf{G}_{>k}\big)_{(\nu\alpha),(\mu\beta)}\Big] \;=\; \sum_{p>k}\frac{\mu_p}{d^p}\sum_{|\boldsymbol{j}|=p}\frac{\mathbb{E}[\psi_p^{\boldsymbol{j}}(\mathbf{Y})]^2}{\boldsymbol{j}!} \;=\; 0.$$

Hence the cross-cluster entries are mean-zero. Since $\mathbf{Y}^{\nu\alpha}$ and $\mathbf{Y}^{\mu\beta}$ are independent and the $\psi_p^{\boldsymbol{j}}$ are orthogonal in $L^2(P_t)$, all the cross terms $(p,\boldsymbol{j})\neq(q,\boldsymbol{l})$ vanish in the second moment, which is therefore a sum of squares:

$$\mathrm{Var}\big[(\mathbf{G}_{>k})_{(\nu\alpha),(\mu\beta)}\big] \;=\; \sum_{p>k}\frac{\mu_p^2}{d^{2p}}\sum_{|\boldsymbol{j}|=p}\frac{\big(c_p^{\boldsymbol{j}}\big)^2}{(\boldsymbol{j}!)^2} \;\sim\; \sum_{p>k}\frac{C_p}{d^p},$$

where $c_p^{\boldsymbol{j}} = \mathbb{E}_{\mathbf{y}\sim P_t}[\psi_p^{\boldsymbol{j}}(\mathbf{y})^2]$ and the last estimate uses that there are $\Theta(d^p/p!)$ multi-indices of degree p. Let $\mathbf{G}_{>k}^{\mathrm{cross}} \in \mathbb{R}^{N\times N}$ denote the matrix obtained from $\mathbf{G}_{>k}$ by zeroing out the diagonal and same-cluster blocks; its entries are mean-zero and weakly correlated. By the standard operator-norm bound for random matrices with i.i.d.-like entries of variance σ^2, we have $\|\mathbf{G}_{>k}^{\mathrm{cross}}\|_{\mathrm{op}} \le C\sqrt{N\sigma^2}$. Substituting $N=nm$ and $\sigma^2\sim 1/d^p$ for the dominant degree $p=k+1$:

$$\|\mathbf{G}_{>k}^{\mathrm{cross}}\|_{\mathrm{op}} \;\lesssim\; \sqrt{\frac{nm}{d^{k+1}}} \;=\; \sqrt{m\cdot\frac{n}{d^{k+1}}}.$$

The hypothesis $n \ll d^{k+1}$ then gives $\|\mathbf{G}_{>k}^{\text{cross}}\|_{\text{op}} = o_d(1)$.

The entries of $\mathbf{G}_{>k}$ on the diagonal, same-cluster off-diagonal and cross-cluster blocks match those of $\mu_I^{>k}\mathbf{I}_N + \mu_B^{>k}\mathbf{B}_m$ up to fluctuations of order $o_P(1)$ entrywise on the structured part, and up to a residual matrix with operator norm $o_d(1)$ on the cross-cluster part. The structured residual fluctuations are themselves of operator norm $o_d(1)$ (by the same $\sqrt{N\sigma^2}$ bound applied to each cluster block). Therefore

$$\left\|\mathbf{G}_{>k} - \mu_I^{>k}\mathbf{I}_N - \mu_B^{>k}\mathbf{B}_m\right\|_{\text{op}} \;=\; o_d(1),$$

which combined with $\mathbf{G} = \mathbf{G}_{\le k} + \mathbf{G}_{>k}$ proves the claimed operator-norm equivalence. □

Theorem D.3 (Spectrum of the Gram matrix, polynomial regime)**.** *Let $c_p^{\boldsymbol{j}} = \mathbb{E}_{\mathbf{y}\sim P_t}[\psi_p^{\boldsymbol{j}}(\mathbf{y})^2]$ and $c_{X,p}^{\boldsymbol{j}} = \mathbb{E}_{\mathbf{x}\sim P_0,\, \boldsymbol{\xi},\, \boldsymbol{\xi}'\sim\mathcal{N}(0,\mathbf{I}_d)}[\psi_p^{\boldsymbol{j}}(e^{-t}\mathbf{x}+\sqrt{\Delta_t}\boldsymbol{\xi})\psi_p^{\boldsymbol{j}}(e^{-t}\mathbf{x}+\sqrt{\Delta_t}\boldsymbol{\xi}')]$. In the scaling $d^{k+1} \gg n \gg d^k$ with $m = O(1)$, the spectrum of the Gram matrix has the following structure:*

- *for every $p \in \{1,\dots,k\}$, a group of $\Theta(d^p/p!)$ eigenvalues of order nm/d^p, with eigenvectors asymptotically $\psi_p^{\boldsymbol{j}}(\mathbf{Y})$;*
- *a bulk of $\Theta(d^k/k!)$ eigenvalues $\mu_I^{>k} + \mu_B^{>k}(m-1)\big(1 - c_{X,k}^{\boldsymbol{j}}/c_k^{\boldsymbol{j}}\big)$, with eigenvectors $\Lambda_\perp^{\boldsymbol{j}}\frac{1}{m}\mathbf{B}_m\psi_k^{\boldsymbol{j}}(\mathbf{Y}) - \Lambda_\parallel^{\boldsymbol{j}}(\mathbf{I}_N - \frac{1}{m}\mathbf{B}_m)\psi_k^{\boldsymbol{j}}(\mathbf{Y})$, where $\Lambda_\parallel^{\boldsymbol{j}} = \frac{\mu_k}{\boldsymbol{j}!}\left(\frac{c_k^{\boldsymbol{j}}}{m} + \frac{m-1}{m}c_{X,k}^{\boldsymbol{j}}\right)$ and $\Lambda_\perp^{\boldsymbol{j}} = \frac{\mu_k}{\boldsymbol{j}!}(1-\frac{1}{m})\left(c_k^{\boldsymbol{j}} - c_{X,k}^{\boldsymbol{j}}\right)$ (at $k=1$ and for Gaussian data, $c_{X,1}/c_1 = e^{-2t}$ and this is the memorization bulk of Theorem D.2)*
- *a spike at $m\mu_B^{>k}$ with, at leading order, $n - d^k/k!$ eigenvalues;*
- *an atom at $\lambda = \mu_I^{>k}$ with $nm - n - d^k/k!$ eigenvalues.*

Proof. **Case $p < k$.** The rank of the term $\sum_{p=0}^{k-1}\sum_{\boldsymbol{j}}\frac{\mu_p}{d^p\boldsymbol{j}!}\psi_p^{\boldsymbol{j}}(\mathbf{Y})\psi_p^{\boldsymbol{j}}(\mathbf{Y})^\top$ is not extensive in the dimension of $\mathbf{G}$, so this part can only generate spikes. Moreover the vectors $\psi_p^{\boldsymbol{j}}(\mathbf{Y}), \psi_q^{\boldsymbol{l}}(\mathbf{Y})$ for $(p,\boldsymbol{j}) \neq (q,\boldsymbol{l})$ are asymptotically orthogonal in the large-n limit $\psi_p^{\boldsymbol{j}}(\mathbf{Y})^\top\psi_q^{\boldsymbol{l}}(\mathbf{Y}) \sim nm\mathbb{E}_{\mathbf{y}\sim P_t}[\psi_p^{\boldsymbol{j}}(\mathbf{y})\psi_q^{\boldsymbol{l}}(\mathbf{y})] = nmc_p^{\boldsymbol{j}}\delta_{p,q}\delta^{\boldsymbol{j},\boldsymbol{l}}$. Hence, this term will generate spikes in the spectrum with eigenvectors $\psi_p^{\boldsymbol{j}}(\mathbf{Y})$ and eigenvalues $\mu_I^{>k} + \frac{c_p^{\boldsymbol{j}}\mu_p nm}{d^p\boldsymbol{j}!} + O(1)$. For each $p \in \{1,\dots,k-1\}$, there are $\Theta(d^p/p!)$ such eigenvectors. Because of the term $\mu_B^{>k}\mathbf{B}_m$, the 2×2 reduction below applies to every degree $p \le k$: at each $p < k$ it also produces $\Theta(d^p/p!)$ eigenvalues $\mu_I^{>k} + \mu_B^{>k}(m-1)(1 - c_{X,p}^{\boldsymbol{j}}/c_p^{\boldsymbol{j}})$, which are negligible in number compared with the $\Theta(d^k/k!)$ ones of degree k and are removed from the atom counts at subleading order. The case $p = k$ is treated below.

Case $p = k$. We now focus on

$$\mathbf{G} = \mu_I^{>k}\mathbf{I}_N + \mu_B^{>k}\mathbf{B}_m + \sum_{\boldsymbol{j}}\frac{\mu_k}{d^k\boldsymbol{j}!}\psi_k^{\boldsymbol{j}}(\mathbf{Y})\psi_k^{\boldsymbol{j}}(\mathbf{Y})^\top \tag{140}$$

As in the linear case, we observe that in the asymptotic regime

$$\left[\sum_{\boldsymbol{j}}\frac{\mu_k}{d^k\boldsymbol{j}!}\psi_k^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})\psi_k^{\boldsymbol{j}}(\mathbf{Y}^{\nu'\alpha'})\right]\psi_k^{\boldsymbol{l}}(\mathbf{Y}^{\nu'\alpha'}) \sim \frac{nm\mu_k}{d^k\boldsymbol{l}!}c_k^{\boldsymbol{l}}\psi_k^{\boldsymbol{l}}(\mathbf{Y}^{\nu\alpha}) \tag{141}$$

$$\left[\sum_{\boldsymbol{j}}\frac{\mu_k}{d^k\boldsymbol{j}!}\psi_k^{\boldsymbol{j}}(\mathbf{Y}^{\nu\alpha})\psi_k^{\boldsymbol{j}}(\mathbf{Y}^{\nu'\alpha'})\right]\frac{1}{m}\mathbf{B}_m\psi_k^{\boldsymbol{l}}(\mathbf{Y}^{\nu'\alpha'}) \sim \frac{nm\mu_k}{d^k\boldsymbol{l}!}\left(\frac{c_k^{\boldsymbol{l}}}{m} + \frac{m-1}{m}c_{X,k}^{\boldsymbol{l}}\right)\psi_k^{\boldsymbol{l}}(\mathbf{Y}^{\nu\alpha}) \tag{142}$$

where the sums over (ν',α') are implicit, and we used that $c_{X,k}^{\boldsymbol{l}}$ is by definition the covariance of two noisy copies of the same clean sample. Up to the shift $\mu_I^{>k}$, the Gram matrix on the space spanned by

$\{\frac{1}{m}\mathbf{B}_m\psi_k^{\boldsymbol{j}}(\mathbf{Y}), (\mathbf{I}_N - \frac{1}{m}\mathbf{B}_m)\psi_k^{\boldsymbol{j}}(\mathbf{Y})\}$ reads

$$\begin{pmatrix} m\mu_B^{>k} + \frac{nm\mu_k}{d^k \boldsymbol{j}!}\left(\frac{c_k^{\boldsymbol{j}}}{m} + \frac{m-1}{m}c_{X,k}^{\boldsymbol{j}}\right) & \frac{nm\mu_k}{d^k \boldsymbol{j}!}\frac{m-1}{m}(c_k^{\boldsymbol{j}} - c_{X,k}^{\boldsymbol{j}}) \\ \frac{nm\mu_k}{d^k \boldsymbol{j}!}\left(\frac{c_k^{\boldsymbol{j}}}{m} + \frac{m-1}{m}c_{X,k}^{\boldsymbol{j}}\right) & \frac{nm\mu_k}{d^k \boldsymbol{j}!}\frac{m-1}{m}(c_k^{\boldsymbol{j}} - c_{X,k}^{\boldsymbol{j}}) \end{pmatrix} \tag{143}$$

$$= \begin{pmatrix} m\mu_B^{>k} + \frac{nm}{d^k}\Lambda_{\parallel}^{\boldsymbol{j}} & \frac{nm}{d^k}\Lambda_{\perp}^{\boldsymbol{j}} \\ \frac{nm}{d^k}\Lambda_{\parallel}^{\boldsymbol{j}} & \frac{nm}{d^k}\Lambda_{\perp}^{\boldsymbol{j}} \end{pmatrix} \tag{144}$$

Its trace is $m\mu_B^{>k} + \frac{nm}{d^k}(\Lambda_{\parallel}^{\boldsymbol{j}} + \Lambda_{\perp}^{\boldsymbol{j}})$ and its determinant $m\mu_B^{>k}\frac{nm}{d^k}\Lambda_{\perp}^{\boldsymbol{j}}$, with $\Lambda_{\parallel}^{\boldsymbol{j}} + \Lambda_{\perp}^{\boldsymbol{j}} = \mu_k c_k^{\boldsymbol{j}}/\boldsymbol{j}!$. For $nm/d^k \gg 1$, diagonalizing it yields, at leading order, the eigenvector $\psi_k^{\boldsymbol{j}}(\mathbf{Y})/\sqrt{nmc_k^{\boldsymbol{j}}}$ with eigenvalue

$$\mu_I^{>k} + \frac{nm\,\mu_k c_k^{\boldsymbol{j}}}{d^k \boldsymbol{j}!} + m\mu_B^{>k}\frac{\Lambda_{\parallel}^{\boldsymbol{j}}}{\Lambda_{\parallel}^{\boldsymbol{j}} + \Lambda_{\perp}^{\boldsymbol{j}}}, \tag{145}$$

and the eigenvector $\Lambda_{\perp}^{\boldsymbol{j}}\frac{1}{m}\mathbf{B}_m\psi_k^{\boldsymbol{j}} - \Lambda_{\parallel}^{\boldsymbol{j}}(\mathbf{I}_N - \frac{1}{m}\mathbf{B}_m)\psi_k^{\boldsymbol{j}}$ with eigenvalue

$$\mu_I^{>k} + m\mu_B^{>k}\frac{\Lambda_{\perp}^{\boldsymbol{j}}}{\Lambda_{\perp}^{\boldsymbol{j}} + \Lambda_{\parallel}^{\boldsymbol{j}}} = \mu_I^{>k} + \mu_B^{>k}(m-1)\left(1 - \frac{c_{X,k}^{\boldsymbol{j}}}{c_k^{\boldsymbol{j}}}\right). \tag{146}$$

Each of these two bulks has $\Theta(d^k/k!)$ eigenvalues. As a check, at $k = 1$ and for isotropic Gaussian data with $\sigma^2 = 1$, $\psi_1^{\boldsymbol{e}_i}(\mathbf{y}) = y_i$, $c_1 = \Gamma_t = 1$ and $c_{X,1} = e^{-2t}$, so that the second eigenvalue is $\mu_I + \mu_B(m-1)\Delta_t$, the memorization bulk of Theorem D.2.

□

D.8 Bias–variance decomposition

We record here the decomposition used in Sect. 2.2 of the main text and in the two subsections below. Write $\mathbf{y} = e^{-t}\mathbf{x} + \sqrt{\Delta_t}\,\boldsymbol{\xi}$ for a fresh test point, whose law is P_t, and let $\mathbf{s}_{\text{exact}}(\mathbf{y}) = \nabla_{\mathbf{y}} \log P_t(\mathbf{y})$. Differentiating $P_t(\mathbf{y}) = \int \mathrm{d}P_0(\mathbf{x})\,\mathcal{N}(\mathbf{y}; e^{-t}\mathbf{x}, \Delta_t \mathbf{I}_d)$ under the integral sign gives Tweedie's identity

$$\mathbb{E}\big[\boldsymbol{\xi}\,\big|\,\mathbf{y}\big] = -\sqrt{\Delta_t}\,\mathbf{s}_{\text{exact}}(\mathbf{y}). \tag{147}$$

Let $\mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}$ be the learned score and $\langle\cdot\rangle$ the average over the training set and the initialization, both independent of the test pair $(\mathbf{x}, \boldsymbol{\xi})$. Adding and subtracting $\mathbf{s}_{\text{exact}}(\mathbf{y})$ in the test loss (3),

$$2\,\mathcal{L}_{\text{test}} = \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\Big\langle \Big\| \big[\mathbf{s}_{\mathcal{D},\boldsymbol{\theta}_0}(\mathbf{y}) - \mathbf{s}_{\text{exact}}(\mathbf{y})\big] + \Big[\mathbf{s}_{\text{exact}}(\mathbf{y}) + \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\Big] \Big\|^2 \Big\rangle. \tag{148}$$

The cross term vanishes: conditionally on $\mathbf{y}$ the first bracket is a function of $\mathbf{y}$ and of $(\mathcal{D}, \boldsymbol{\theta}_0)$ alone, while the second has conditional mean zero by (147). Splitting what remains of the first bracket into its mean and its fluctuation produces the two terms (4) of the main text, and

$$\mathcal{L}_{\text{test}} = \frac{1}{2}\big(C_t + \mathcal{B}^2 + \mathcal{V}\big), \qquad C_t = \frac{1}{d}\,\mathbb{E}_{\mathbf{x},\boldsymbol{\xi}}\Big\|\mathbf{s}_{\text{exact}}(\mathbf{y}) + \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}}\Big\|^2. \tag{149}$$

The constant C_t is the Bayes error of the denoising problem at noise level t and it is the value taken by $\mathcal{L}_{\text{test}}$ at the exact score. It does not depend on the estimator. Using (147) once more, $C_t = \frac{1}{d\Delta_t}\mathbb{E}\,\mathrm{Tr}\,\mathrm{Cov}(\boldsymbol{\xi}\mid\mathbf{y})$, that is

$$C_t = \frac{1}{\Delta_t} - \frac{1}{d}\,\mathbb{E}_{\mathbf{y}\sim P_t}\big\|\mathbf{s}_{\text{exact}}(\mathbf{y})\big\|^2. \tag{150}$$

Let us specialize to the case $P_0 = \mathcal{N}(0, \boldsymbol{\Sigma})$. One has that $P_t = \mathcal{N}(0, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\Sigma}_t = e^{-2t}\boldsymbol{\Sigma} + \Delta_t \mathbf{I}_d$, hence $\mathbf{s}_{\text{exact}}(\mathbf{y}) = -\boldsymbol{\Sigma}_t^{-1}\mathbf{y}$ and $\mathbb{E}_{\mathbf{y}\sim P_t}\|\mathbf{s}_{\text{exact}}(\mathbf{y})\|^2 = \mathrm{Tr}(\boldsymbol{\Sigma}_t^{-1})$. Equation (150) becomes

$$C_t = \frac{1}{\Delta_t} - \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_t^{-1})}{d} = \frac{e^{-2t}}{d}\sum_{\lambda\in\mathrm{Spec}(\boldsymbol{\Sigma})}\frac{\lambda}{\Delta_t\,\lambda_t}, \qquad \lambda_t = e^{-2t}\lambda + \Delta_t. \tag{151}$$

For isotropic data, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_d$, every λ_t equals Γ_t and

$$C_t = \frac{\sigma^2 e^{-2t}}{\Delta_t\,\Gamma_t}. \tag{152}$$

D.9 Derivation of Theorem 2.2: fixed-point equations in the kernel limit

We derive here the closed-form equations of Theorem 2.2, used in Sect. 2.2. We set $\psi_n = n/d$, and the data are isotropic, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_d$. We first state the equations at $\sigma^2 = 1$, where $\Gamma_t = e^{-2t} + \Delta_t = 1$; the general case reduces exactly to this one (see the remark after Theorem D.4). We write γ for the ridge and $\mu_1 = f'(0)$, $\mu_I = f(1) - f(e^{-2t}) - \Delta_t f'(0)$, $\mu_B = f(e^{-2t}) - f(0) - e^{-2t} f'(0)$, $\bar{\mu} = \mu_I + m\mu_B$ for the kernel constants of Theorem 2.1.

Theorem D.4 (Bias and variance of the kernel ridge predictor)**.** *Under the assumptions above, let* $w_1 = me^{-2t} + \Delta_t$, $c_1 = \bar{\mu}$, $w_2 = \Delta_t$ *and* $c_2 = \mu_I$. *The bias and the variance* (4) *are*

$$\mathcal{B}^2 = (b-1)^2, \qquad \mathcal{V} = \frac{\mu_1 T_4}{\Delta_t} - b^2, \tag{153}$$

where b *and* T_4 *are given as follows.*

1. Ridge $\tilde{\gamma} > 0$. *Let* $\bar{q} = 1/\tilde{\gamma}$ *and*

$$L_a(\bar{r}) = 1 + \frac{\mu_1 w_a \bar{r} + c_a \bar{q}}{\psi_n m}, \quad g(\bar{r}) = \frac{1}{L_1} + \frac{m-1}{L_2}, \quad R(\bar{r}) = \frac{\mu_1}{m}\left(\frac{w_1}{L_1} + \frac{(m-1)w_2}{L_2}\right). \tag{154}$$

   Then

$$b = \frac{\mu_1 \, \bar{r} \, g(\bar{r})}{m}, \qquad T_4 = \frac{-\Phi'(\bar{r})}{\bar{r}^{-2} + R'(\bar{r})}, \qquad \Phi(\bar{r}) = \frac{g}{m} - \frac{\mu_1 \Delta_t}{m^2} \, \bar{r} \, g^2, \tag{155}$$

   where $\bar{r}$ *solves* $1/\bar{r} = \tilde{\gamma} + R(\bar{r})$.

2. Ridgeless limit $\tilde{\gamma} \to 0^+$. *For any* $\varrho > 0$, *let* $q = 1/\varrho$ *and*

$$e_a(r) = \mu_1 r \, w_a + q \, c_a, \quad g(r) = \frac{1}{e_1} + \frac{m-1}{e_2}, \quad R(r) = \mu_1 \psi_n \left(\frac{w_1}{e_1} + \frac{(m-1)w_2}{e_2}\right). \tag{156}$$

   Then (155) *holds with* $\bar{r}$ *replaced by the solution* r *of* $1/r = \varrho + R(r)$, $b = \mu_1 \psi_n \, r \, g(r)$ *and* $\Phi(r) = \psi_n g - \mu_1 \Delta_t \psi_n^2 \, r \, g^2$. *The result does not depend on* ϱ.

Case 1 is Theorem 2.2 of the main text. The index $a = 1$ refers to the average of the m noisy copies of a clean sample and $a = 2$ to the $m - 1$ directions orthogonal to it; c_1 and c_2 are the two atoms of Theorem D.2. For $m = 1$ only the terms with $a = 1$ remain.

Reduction of $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_d$ to $\sigma^2 = 1$. Let $\Gamma_t = \sigma^2 e^{-2t} + \Delta_t$ and define t' by $e^{-2t'} = \sigma^2 e^{-2t}/\Gamma_t$, so that $\Delta_{t'} = 1 - e^{-2t'} = \Delta_t/\Gamma_t$. With $\mathbf{x}' = \mathbf{x}/\sigma$, which has covariance $\mathbf{I}_d$,

$$\mathbf{Y}^{\nu\alpha} = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\,\boldsymbol{\xi}^{\nu\alpha} = \sqrt{\Gamma_t}\left(e^{-t'}\mathbf{x}'^\nu + \sqrt{\Delta_{t'}}\,\boldsymbol{\xi}^{\nu\alpha}\right) = \sqrt{\Gamma_t}\,\mathbf{Y}'^{\nu\alpha}, \tag{157}$$

where $\mathbf{Y}'$ is the noised data of the $\sigma^2 = 1$ problem at time t', with the *same* noises $\boldsymbol{\xi}$. Hence:

- the Gram matrix is unchanged, $f(\mathbf{Y}^\top \mathbf{Y}/d) = \tilde{f}(\mathbf{Y}'^\top \mathbf{Y}'/d)$ with $\tilde{f}(u) = f(\Gamma_t u)$, and so are the ridge γ and the training time τ;
- the target $-\boldsymbol{\xi}/\sqrt{\Delta_t} = \Gamma_t^{-1/2}\big(-\boldsymbol{\xi}/\sqrt{\Delta_{t'}}\big)$ and the exact score $\mathbf{s}_{\text{exact}}(\mathbf{y}) = -\mathbf{y}/\Gamma_t = \Gamma_t^{-1/2}\,\mathbf{s}'_{\text{exact}}(\mathbf{y}')$ are both rescaled by $\Gamma_t^{-1/2}$, so that the predictor, its bias and its variance are those of the $\sigma^2 = 1$ problem, with $\mathcal{B}^2$ and $\mathcal{V}$ multiplied by $1/\Gamma_t$.

Therefore, independently of the Gaussian-equivalence argument (which also extends to general σ^2 with modified coefficients, see the end of the proof of Proposition D.3), Theorem D.4 holds for $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_d$ with t replaced by t', f by $\tilde{f}$, and $\mathcal{B}^2, \mathcal{V}$ divided by Γ_t. In the original variables the kernel constants of $\tilde{f}$ are exactly those of Theorem D.1 at general σ^2,

$$\begin{aligned} \tilde{\mu}_I &= f(\Gamma_t) - f(\sigma^2 e^{-2t}) - \Delta_t f'(0) = \mu_I, \\ \tilde{\mu}_B &= f(\sigma^2 e^{-2t}) - f(0) - \sigma^2 e^{-2t} f'(0) = \mu_B, \qquad \tilde{\mu}_1 = \Gamma_t f'(0), \end{aligned} \tag{158}$$

and the channel weights are $\tilde{w}_1 = (m\sigma^2 e^{-2t} + \Delta_t)/\Gamma_t$ and $\tilde{w}_2 = \Delta_t/\Gamma_t$. As a check, the same map sends the irreducible term $C'_{t'} = e^{-2t'}/\Delta_{t'}$ to $C'_{t'}/\Gamma_t = \sigma^2 e^{-2t}/(\Delta_t \Gamma_t)$, which is (152).

Proof. We take the data isotropic and Gaussian, $\mathbf{x}^\nu \sim \mathcal{N}(0, \mathbf{I}_d)$, and treat the case $\sigma^2 = 1$; general σ^2 follows from the reduction below. The proof has four steps: we reduce the predictor to a linear map, use isotropy to fix its mean, and compute its first two moments with deterministic equivalents for the resolvent.

Step 1: linear structure. By Theorem D.1 (or, equivalently, the Gaussian equivalence of Proposition D.3), the Gram matrix can be replaced by its linear equivalent, and the test kernel by $K(\mathbf{y}, \mathbf{Y}) \simeq \mu_1 \mathbf{Y}^\top \mathbf{y}/d$ for a test point $\mathbf{y} \sim P_t$ independent of the training set. The rank-one terms $\mu_0 \mathbf{1}_N \mathbf{1}_N^\top$ and $\mu_0 \mathbf{1}_N$ change the normalized traces below by $O(1/d)$ and are dropped. With $\mathbf{E} = (\mu_I + \gamma)\mathbf{I}_N + \mu_B \mathbf{B}_m$ and $\mathbf{R} = (\mathbf{G} + \gamma \mathbf{I}_N)^{-1}$,

$$\mathbf{G} + \gamma \mathbf{I}_N = \mathbf{E} + \frac{\mu_1}{d} \mathbf{Y}^\top \mathbf{Y},$$
$$\mathbf{s}(\mathbf{y}) = K(\mathbf{y}, \mathbf{Y})^\top \mathbf{R} \Big(- \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}} \Big)^\top = -\mathbf{A}\mathbf{y}, \qquad \mathbf{A} = \frac{\mu_1}{d\sqrt{\Delta_t}} \boldsymbol{\xi}\, \mathbf{R}\, \mathbf{Y}^\top \in \mathbb{R}^{d \times d}, \tag{159}$$

where $\mathbf{Y}, \boldsymbol{\xi} \in \mathbb{R}^{d \times N}$ stack the noised training points and their noises as columns. The predictor is thus linear in the test point. The d rows $\mathbf{y}_i \in \mathbb{R}^N$ of $\mathbf{Y}$ are i.i.d. $\mathcal{N}(0, \mathbf{C})$ with

$$\mathbf{C} = e^{-2t} \mathbf{B}_m + \Delta_t \mathbf{I}_N, \tag{160}$$

since $\mathbb{E}[\mathbf{Y}_i^{\nu\alpha} \mathbf{Y}_i^{\mu\beta}] = \delta_{\nu\mu}(e^{-2t} + \Delta_t \delta_{\alpha\beta})$. The matrices $\mathbf{C}$ and $\mathbf{E}$ are diagonal in the same orthonormal basis: on the n block-average directions $\mathbf{1}_m/\sqrt{m}$ they take the values $w_1 = m e^{-2t} + \Delta_t$ and $c_1 + \gamma = \bar{\mu} + \gamma$, and on the $n(m-1)$ directions transverse to them the values $w_2 = \Delta_t$ and $c_2 + \gamma = \mu_I + \gamma$. This is the decoupling into the two channels of the statement.

Step 2: isotropy. The joint law of $(\mathbf{Y}, \boldsymbol{\xi})$ is invariant under $(\mathbf{Y}, \boldsymbol{\xi}) \mapsto (\mathbf{O}\mathbf{Y}, \mathbf{O}\boldsymbol{\xi})$ for any orthogonal $\mathbf{O} \in O(d)$, under which $\mathbf{R}$ is unchanged and $\mathbf{A} \mapsto \mathbf{O}\mathbf{A}\mathbf{O}^\top$. Hence $\langle \mathbf{A} \rangle$ commutes with every rotation, and

$$\langle \mathbf{A} \rangle = b\, \mathbf{I}_d, \qquad b = \frac{1}{d} \big\langle \operatorname{Tr} \mathbf{A} \big\rangle. \tag{161}$$

This is the isotropic form of the mean predictor, obtained here without an ansatz. Since $\mathbf{s}_{\text{exact}}(\mathbf{y}) = -\mathbf{y}$ and $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \mathbf{I}_d$, the definitions (4) give

$$\mathcal{B}^2 = (b-1)^2, \qquad \mathcal{V} = \frac{1}{d} \big\langle \operatorname{Tr} \mathbf{A}\mathbf{A}^\top \big\rangle - b^2 = \frac{\mu_1 T_4}{\Delta_t} - b^2, \qquad T_4 = \frac{\mu_1}{d^3} \big\langle \operatorname{Tr} \big(\boldsymbol{\xi}\mathbf{R}\mathbf{Y}^\top \mathbf{Y}\mathbf{R}\boldsymbol{\xi}^\top \big) \big\rangle, \tag{162}$$

which is (153), with b and T_4 the first two moments of $\mathbf{A}$. It remains to compute them.

Step 3: the resolvent. Since $\mathbf{G} + \gamma \mathbf{I}_N$ is a deterministic matrix plus a sample covariance of d i.i.d. Gaussian vectors in $\mathbb{R}^N$, with $N/d = m\psi_n$ fixed, its resolvent has the deterministic equivalent (Silverstein & Bai, 1995; Louart et al., 2018)

$$\mathbf{R} \simeq \bar{\mathbf{R}} = \Big(\mathbf{E} + \frac{\mu_1 \mathbf{C}}{1 + \delta} \Big)^{-1}, \qquad \delta = \frac{\mu_1}{d} \operatorname{Tr} \big(\mathbf{C}\bar{\mathbf{R}} \big), \tag{163}$$

in the sense that $\frac{1}{d} \operatorname{Tr}(\mathbf{M}\mathbf{R}) - \frac{1}{d} \operatorname{Tr}(\mathbf{M}\bar{\mathbf{R}}) \to 0$ for deterministic $\mathbf{M}$ of bounded norm; we also use the leave-one-out identity $\mathbf{R}\mathbf{y}_i = \mathbf{R}_{-i}\mathbf{y}_i / (1 + \frac{\mu_1}{d} \mathbf{y}_i^\top \mathbf{R}_{-i} \mathbf{y}_i)$, with $\mathbf{R}_{-i}$ the resolvent without row i and $\frac{\mu_1}{d} \mathbf{y}_i^\top \mathbf{R}_{-i} \mathbf{y}_i \to \delta$. In the joint eigenbasis of $\mathbf{C}$ and $\mathbf{E}$, $\bar{\mathbf{R}}$ is diagonal with entries $(c_a + \gamma + \mu_1 w_a/(1+\delta))^{-1}$. Writing $\gamma = m\psi_n \tilde{\gamma} = m\psi_n/\bar{q}$ and setting

$$\bar{r} = \frac{\bar{q}}{1 + \delta}, \tag{164}$$

these entries are $\bar{q}/(m\psi_n L_a(\bar{r}))$ with L_a as in (154). Hence $\frac{1}{d} \operatorname{Tr} \bar{\mathbf{R}} = \bar{q}\, g(\bar{r})/m$, and the equation for δ becomes $\delta = \mu_1 \psi_n \big(w_1 \bar{R}_1 + (m-1) w_2 \bar{R}_2 \big) = \bar{q}\, R(\bar{r})$, i.e. $1/\bar{r} = (1+\delta)/\bar{q} = \tilde{\gamma} + R(\bar{r})$: this is the fixed point of the statement.

Step 4: the two moments. Row by row, $(\mathbf{y}_i, \boldsymbol{\xi}_i)$ is a centered Gaussian pair with $\mathbb{E}[\boldsymbol{\xi}_i \mathbf{y}_i^\top] = \sqrt{\Delta_t}\,\mathbf{I}_N$, so that

$$\boldsymbol{\xi} = \sqrt{\Delta_t}\,\mathbf{Y}\mathbf{C}^{-1} + \mathbf{H}, \tag{165}$$

with the rows of $\mathbf{H}$ i.i.d. $\mathcal{N}(0, \mathbf{I}_N - \Delta_t \mathbf{C}^{-1})$ and independent of $\mathbf{Y}$, hence of $\mathbf{R}$.

- *Bias.* The $\mathbf{H}$ part of $\mathrm{Tr}(\boldsymbol{\xi}\mathbf{R}\mathbf{Y}^\top) = \sum_i \boldsymbol{\xi}_i^\top \mathbf{R}\mathbf{y}_i$ has zero mean, and by the leave-one-out identity $\mathbf{y}_i^\top \mathbf{C}^{-1}\mathbf{R}\mathbf{y}_i \simeq \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{C}\bar{\mathbf{R}})/(1+\delta) = \mathrm{Tr}\,\bar{\mathbf{R}}/(1+\delta)$. Thus

$$b = \frac{\mu_1}{d^2\sqrt{\Delta_t}}\, d\sqrt{\Delta_t}\,\frac{\mathrm{Tr}\,\bar{\mathbf{R}}}{1+\delta} = \frac{\mu_1 \bar{q}\, g}{m(1+\delta)} = \frac{\mu_1 \bar{r}\, g(\bar{r})}{m}. \tag{166}$$

- *Second moment.* Let $\mathbf{R}_\varepsilon = \big(\mathbf{E} + (1-\varepsilon)\frac{\mu_1}{d}\mathbf{Y}^\top\mathbf{Y}\big)^{-1}$. Since $\partial_\varepsilon \mathbf{R}_\varepsilon|_0 = \mathbf{R}\,\frac{\mu_1}{d}\mathbf{Y}^\top\mathbf{Y}\,\mathbf{R}$,

$$T_4 = \partial_\varepsilon S(\varepsilon)\big|_{\varepsilon=0}, \qquad S(\varepsilon) = \frac{1}{d^2}\big\langle\, \mathrm{Tr}\,\big(\boldsymbol{\xi}\mathbf{R}_\varepsilon\boldsymbol{\xi}^\top\big)\big\rangle, \tag{167}$$

  which only involves first-order traces. Using the decomposition of $\boldsymbol{\xi}$ and, for the $\mathbf{Y}$ part, Sherman–Morrison, $\mathbf{y}_i^\top\mathbf{C}^{-1}\mathbf{R}_\varepsilon\mathbf{C}^{-1}\mathbf{y}_i \simeq \mathrm{Tr}(\mathbf{C}^{-1}\bar{\mathbf{R}}_\varepsilon) - \frac{\mu_\varepsilon}{d}(\mathrm{Tr}\,\bar{\mathbf{R}}_\varepsilon)^2/(1+\delta_\varepsilon)$ with $\mu_\varepsilon = (1-\varepsilon)\mu_1$, the terms in $\mathrm{Tr}(\mathbf{C}^{-1}\bar{\mathbf{R}}_\varepsilon)$ cancel between the two parts and

$$S(\varepsilon) = \frac{1}{d}\,\mathrm{Tr}\,\bar{\mathbf{R}}_\varepsilon - \frac{\Delta_t\mu_\varepsilon}{d^2}\,\frac{(\mathrm{Tr}\,\bar{\mathbf{R}}_\varepsilon)^2}{1+\delta_\varepsilon} = \bar{q}\left(\frac{g}{m} - \frac{\mu_\varepsilon\Delta_t}{m^2}\,\bar{r}_\varepsilon\, g^2\right), \tag{168}$$

  where $\bar{r}_\varepsilon$ and g are those of Step 3 with μ_1 replaced by μ_ε. All these quantities depend on μ_ε and $\bar{r}_\varepsilon$ only through $\rho = \mu_\varepsilon\bar{r}_\varepsilon$, except for the fixed point, which reads $1/\rho = 1/(\mu_\varepsilon\bar{q}) + h(\rho)/m$ with $h = w_1/L_1 + (m-1)w_2/L_2$. Differentiating it at $\varepsilon = 0$ gives $\partial_\varepsilon\rho = -\big[\mu_1\bar{q}\,(\rho^{-2} + h'(\rho)/m)\big]^{-1}$, and $S = \bar{q}\,\tilde{\Phi}(\rho)$ with $\tilde{\Phi}(\rho) = g/m - \Delta_t\rho g^2/m^2$, so that

$$T_4 = \bar{q}\,\tilde{\Phi}'(\rho)\,\partial_\varepsilon\rho = -\frac{\tilde{\Phi}'(\rho)}{\mu_1\big(\rho^{-2} + h'(\rho)/m\big)} = -\frac{\Phi'(\bar{r})}{\bar{r}^{-2} + R'(\bar{r})}, \tag{169}$$

  the last form following from $\Phi(\bar{r}) = \tilde{\Phi}(\mu_1\bar{r})$ and $R(\bar{r}) = \frac{\mu_1}{m}h(\mu_1\bar{r})$. This is (155).

Test loss. Combining with (149) and $C_t = e^{-2t}/\Delta_t$,

$$\mathcal{L}_{\text{test}} = \frac{1}{2}\big(C_t + \mathcal{B}^2 + \mathcal{V}\big) = \frac{1 - 2\Delta_t\, b + \mu_1 T_4}{2\Delta_t}. \tag{170}$$

Ridgeless limit. As $\tilde{\gamma} \to 0^+$, $\bar{q} \to \infty$ and $\bar{r} = \bar{q}\,r$ with r finite. Then $L_a \simeq \bar{q}\,e_a/(\psi_n m)$ with $e_a = \mu_1 r w_a + c_a$, so that $g = (\psi_n m/\bar{q})\,g_0$, $R = R_0/\bar{q}$ and $\Phi = \Phi_0/\bar{q}$, where g_0, R_0, Φ_0 are the functions of (156) at $q = 1$. The fixed point becomes $1/r = 1 + R_0(r)$, and $b = \mu_1\psi_n r g_0$ and $T_4 = -\Phi_0'/(r^{-2} + R_0')$, which is item 2 of the statement with $\varrho = q = 1$. The equations are invariant under $r \mapsto r/\varrho$, $q = 1/\varrho$, which is why the result does not depend on the gauge ϱ.

Derivatives. The derivatives entering T_4 follow from the chain rule. In the ridge parametrization (154), with $\partial L_a = \mu_1 w_a/(\psi_n m)$,

$$g' = -\frac{\partial L_1}{L_1^2} - \frac{(m-1)\partial L_2}{L_2^2}, \quad R' = -\frac{\mu_1}{m}\left[\frac{w_1\partial L_1}{L_1^2} + \frac{(m-1)w_2\partial L_2}{L_2^2}\right], \tag{171}$$

$$\Phi' = \frac{g'}{m} - \frac{\mu_1\Delta_t}{m^2}\big(g^2 + 2\bar{r}gg'\big), \tag{172}$$

and in the ridgeless parametrization (156),

$$g' = -\frac{\mu_1 w_1}{e_1^2} - \frac{(m-1)\mu_1 w_2}{e_2^2}, \quad R' = -\mu_1^2\psi_n\left[\frac{w_1^2}{e_1^2} + \frac{(m-1)w_2^2}{e_2^2}\right], \tag{173}$$

$$\Phi' = \psi_n g' - \mu_1\Delta_t\psi_n^2\big(g^2 + 2rgg'\big). \tag{174}$$

□

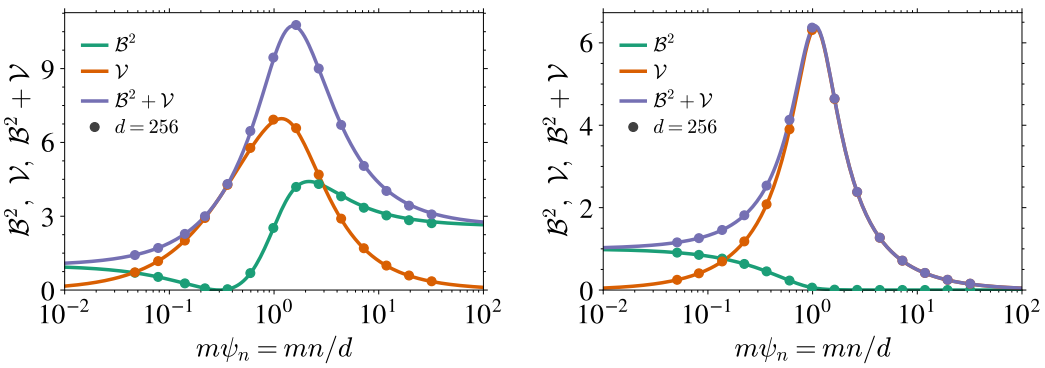

Figure 9: **Bias and variance against the sample complexity.** $\mathcal{B}^2$, $\mathcal{V}$ and $\mathcal{B}^2+\mathcal{V}$ against $m\psi_n = mn/d$ for the ridgeless estimator $\tilde{\gamma} \to 0^+$, at $t = 0.1$, $\sigma^2 = 1$, with the kernel of Figure 2, for $m = 4$ (*left*) and $m = 1$ (*right*). Solid lines are the equations of Theorem 2.2; markers are experiments at $d = 256$.

D.10 Numerical check of Theorem 2.2

This subsection documents the finite-dimensional experiments performed for kernel ridge regression.

Setup. Data are isotropic, $\mathbf{x}^\nu \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{Y}^{\nu\alpha} = e^{-t}\mathbf{x}^\nu + \sqrt{\Delta_t}\boldsymbol{\xi}^{\nu\alpha}$ with m noise realizations per data point, so that $\boldsymbol{\Sigma}_t = \mathbf{I}_d$ and the exact score is $\mathbf{s}_{\text{exact}}(\mathbf{y}) = -\mathbf{y}$. The kernel is the dual activation $f(u) = \mathbb{E}[\sigma(z_1)\sigma(z_2)]$ of $\sigma = \tanh$ at correlation u. Since f is a kernel only for $|u| \le 1$, and at finite d the norms $\|\mathbf{Y}^{\nu\alpha}\|^2/d$ fluctuate around their limit $\Gamma_t = 1$, the rows are rescaled to the sphere of radius $\sqrt{d}$ in the kernel argument only.

Bias and variance. With n_D independent training sets and n_z fresh test points, writing $\bar{\mathbf{s}}(\mathbf{y})$ for the average of the n_D predictors,

$$\widehat{\mathcal{V}}(\mathbf{y}) = \frac{1}{d\,(n_D-1)}\sum_{i=1}^{n_D}\left\|\mathbf{s}_i(\mathbf{y}) - \bar{\mathbf{s}}(\mathbf{y})\right\|^2, \quad \widehat{\mathcal{B}^2}(\mathbf{y}) = \frac{1}{d}\left\|\mathbf{s}_{\text{exact}}(\mathbf{y}) - \bar{\mathbf{s}}(\mathbf{y})\right\|^2 - \frac{\widehat{\mathcal{V}}(\mathbf{y})}{n_D}, \tag{175}$$

both averaged over the n_z test points. The $1/(n_D-1)$ and the $-\widehat{\mathcal{V}}/n_D$ subtraction make the two estimators unbiased at any n_D, so that neither the variance nor the bias is inflated by the finite number of realizations.

Parameters. The sweeps at fixed ψ_n of Figure 2 (left panel and inset) use $\psi_n = 8$, $n_z = 512$ test points and $d \in \{64, 128, 256\}$, with $n_D = 16$ training sets for $m = 1$ and $n_D = 16, 12, 8$ for $m = 4$ at $d = 64, 128, 256$ respectively, while the sweep over the budget $m\psi_n$ of Figure 9 uses $n_z = 1024$ and n_D between 32 at the smallest N and 8 at the largest.

Error bars. The error bars drawn on every finite-d point are a delete-one jackknife over the n_D realizations of the dataset.

Early stopping versus ridging. Fixing a training time τ in (5) imposes a spectral cutoff $\lambda_c \sim nm/(d\tau)$, whereas a ridge cuts at $\lambda \sim \gamma$. Equating the two cutoffs gives $\gamma_{\text{KRR}} = m\psi_n/\tau$, that is $\tilde{\gamma} = 1/\tau$ on the scale of Theorem 2.2 — the same calibration under which Ali et al. (2019) bound the risk of gradient flow by that of ridge regression along the whole path, for least squares with no assumption on the design. Figure 10 tests that identification by drawing the analytical $\mathcal{L}_{\text{test}}$, $\mathcal{B}^2$ and $\mathcal{V}$ at $\tilde{\gamma} = 1/\tau$ against the measured gradient flow at τ. The ridge looks therefore like a faithful proxy for early stopping for the values of the observables and for the scaling of the timescales.

D.11 Small-t behavior of the bias and the variance

Theorem 2.2 holds at a fixed noise level, and most figures are drawn at $t = 0.1$. Since memorization is a small-t phenomenon, we record here how $\mathcal{B}^2$ and $\mathcal{V}$ behave as $t \to 0$ (Fig. 11).

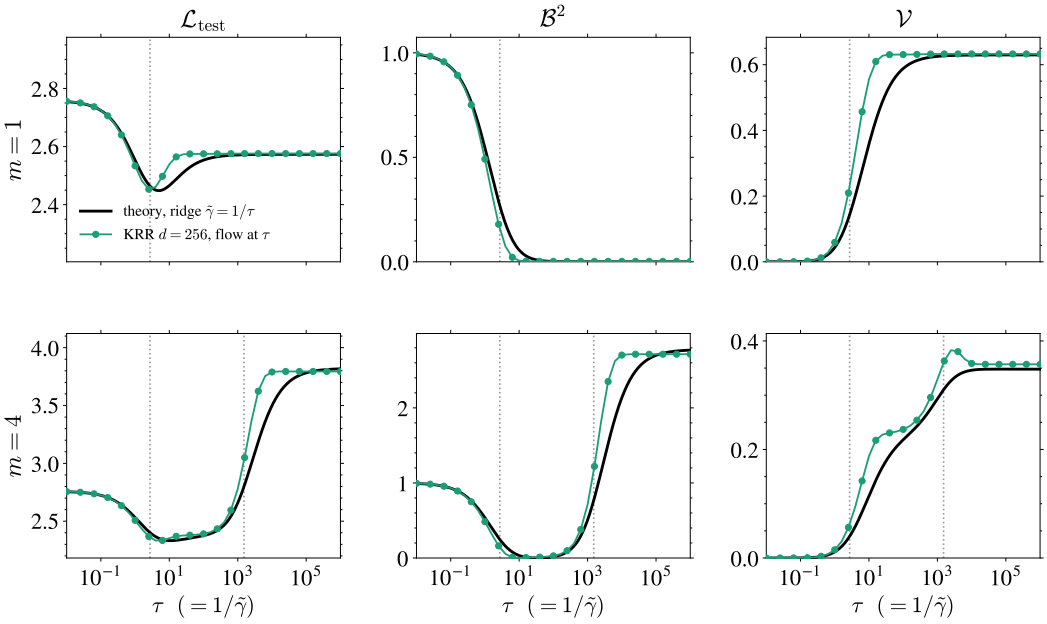

Figure 10: **The ridge as a proxy for early stopping.** $\mathcal{L}_{\text{test}}$, $\mathcal{B}^2$ and $\mathcal{V}$ against the training time τ, at $m = 1$ (*top*) and $m = 4$ (*bottom*). Black: the theory of Theorem 2.2 evaluated at $\tilde{\gamma} = 1/\tau$. Green: kernel gradient flow of (5) measured at finite $d = 256$, $\psi_n = 8$, averaged over 16 ($m = 1$) and 8 ($m = 4$) realizations of the dataset and 512 test points. Dotted verticals are τ_{gen} and τ_{mem}. Panels are scaled independently: at $m = 1$ the bias vanishes whereas at $m = 4$ it saturates at $\Theta(1)$.

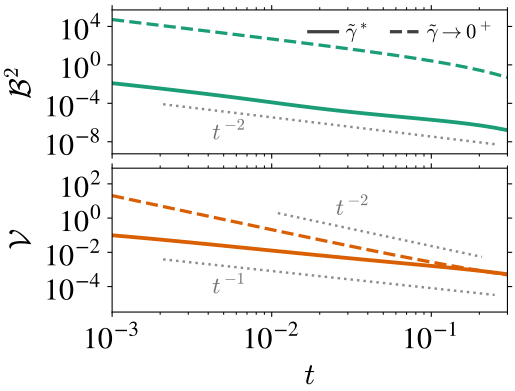

Figure 11: **Scalings with t.** $\mathcal{B}^2$ (*top*) and $\mathcal{V}$ (*bottom*) against t at $\psi_n = 10^3$ and $m = 4$, where $1/(m\psi_n) \ll t \ll 1$, for the optimally ridged estimator $\tilde{\gamma}^*$ (solid) and the ridgeless one $\tilde{\gamma} \to 0^+$ (dashed), with the kernel of Fig. 2. Lines are the equations of Theorem 2.2.

What collapses at small t. Expanding the constants of Theorem D.1 as $\Delta_t \to 0$,

$$\mu_I = \Delta_t\, \iota + O(\Delta_t^2), \quad \iota = f'(\sigma^2) - f'(0), \qquad \mu_B \to \mu_B^0 = f(\sigma^2) - f(0) - \sigma^2 f'(0), \tag{176}$$

so that, at $\sigma^2 = 1$ where $\lambda_t = 1$, the memorization eigenvalue of Theorem D.2 becomes

$$\lambda_{\text{mem}} = \mu_I + \frac{\mu_B (m-1)\Delta_t}{\lambda_t} \;\simeq\; \Delta_t\big(\iota + (m-1)\mu_B^0\big). \tag{177}$$

It vanishes linearly in Δ_t, while the generalization eigenvalue $\lambda_{\text{gen}} = \mu_1 \psi_n m \lambda_t$ stays $\Theta(\psi_n m)$: the two bulks separate further and further as the noise level drops. The memorization modes become both arbitrarily slow to learn and arbitrarily ill-conditioned for the ridgeless predictor, and the scalings below are the quantitative form of that second statement.

Result D.1 (Small-t scalings of the bias and the variance)**.** *Let $\sigma^2 = 1$ and $\Delta_t = 1 - e^{-2t} \simeq 2t$, and write $s = m\psi_n \Delta_t$.*

1. Ridgeless, $\Delta_t \to 0$ at fixed ψ_n and m.
   - $m = 1$*: $\mathcal{B}^2$ is independent of t, and $\psi_n \Delta_t \mathcal{V} \to 1$ as $\psi_n \to \infty$.*
   - $m > 1$*: $\mathcal{B}^2 \simeq R_{\psi_n}^2/\Delta_t^2$ and $\mathcal{V} \simeq c_{m,\psi_n}/(\psi_n \Delta_t^2)$, where the prefactors depend on ψ_n only through $O(1/\psi_n)$ corrections, with limits*

$$R = \lim_{\psi_n \to \infty} R_{\psi_n} = \frac{(m-1)\mu_B^0}{\iota + (m-1)\mu_B^0} \in (0,1),$$
$$c_m = \lim_{\psi_n \to \infty} c_{m,\psi_n} = \frac{m(m-1)\,\iota^2 (\mu_B^0)^2}{\left(\iota + (m-1)\mu_B^0\right)^4}. \tag{178}$$

2. Optimal ridge, $\Delta_t \to 0$ and $\psi_n \to \infty$ at fixed s, for every m: $\tilde{\gamma}^* = \mu_1/s$ *and*

$$\mathcal{B}^2 = \frac{1}{(1+s)^2}, \qquad \mathcal{V} = \frac{s}{(1+s)^2}, \qquad \mathcal{B}^2 + \mathcal{V} = \frac{1}{1+s}. \tag{179}$$

3. The window of Result 2.1. *For $m > 1$ and $1/(m\psi_n) \ll t \ll 1$, i.e. $s \gg 1$ with $\Delta_t \ll 1$, items 1 and 2 give $\mathcal{B}^2 \simeq R^2/\Delta_t^2 = \Theta(t^{-2})$ and $\mathcal{V} \simeq c_m/(\psi_n \Delta_t^2) = \Theta(\psi_n^{-1} t^{-2})$ without ridge, and $\mathcal{B}^2 \simeq s^{-2} = \Theta(\psi_n^{-2} t^{-2})$ and $\mathcal{V} \simeq s^{-1} = \Theta(\psi_n^{-1} t^{-1})$ at the optimal ridge. In the opposite regime $s \ll 1$ the optimally ridged estimator returns $\mathcal{B}^2 \to 1$ and $\mathcal{V} \simeq s$: it predicts almost nothing.*

The ridgeless prefactors. Take the ridgeless parametrization (156) with $\varrho = 1$ (the result does not depend on this choice) and let $\Delta_t \to 0$ with (176). Then $w_1 \to m$, $c_1 \to m\mu_B^0$, $w_2 = \Delta_t$ and $c_2 = \Delta_t \iota$, so that $e_1 \to m(\mu_1 r + \mu_B^0)$ and $e_2 = \Delta_t(\mu_1 r + \iota)$. The order parameter equation becomes independent of t,

$$\frac{1}{r} = 1 + \mu_1 \psi_n \left(\frac{1}{\mu_1 r + \mu_B^0} + \frac{m-1}{\mu_1 r + \iota} \right), \tag{180}$$

whereas $g = G/\Delta_t + O(1)$ with $G = (m-1)/(\mu_1 r + \iota)$ is dominated by the transverse directions. Hence $b = \mu_1 \psi_n r g$ diverges as R_{ψ_n}/Δ_t with $R_{\psi_n} = \mu_1 \psi_n r G$. As $\psi_n \to \infty$ the solution of (180) is $r \simeq \iota \mu_B^0 / [\mu_1 \psi_n (\iota + (m-1)\mu_B^0)]$, which gives the value of R in (178); $\mathcal{B}^2 = (b-1)^2 \simeq R_{\psi_n}^2/\Delta_t^2$. For the variance, $\Phi = \psi_n g - \mu_1 \Delta_t \psi_n^2 r g^2$ is also $\Theta(1/\Delta_t)$, so that both terms of $\mathcal{V} = \mu_1 T_4/\Delta_t - b^2$ are $\Theta(\Delta_t^{-2})$. Their leading parts in $1/\psi_n$ are both equal to $G^2/(\mu_B^{0\,-1} + (m-1)\iota^{-1})^2$ and cancel, and the next order, obtained by expanding (180) to $O(\psi_n^{-2})$, gives $\psi_n \Delta_t^2 \mathcal{V} \to c_m$. At $m = 1$ there are no transverse directions: $w_1 = 1$ and $c_1 = \mu_I + \mu_B = f(1) - f(0) - f'(0)$ are exactly independent of t, hence so is $\mathcal{B}^2$, and only the $1/\Delta_t$ in front of T_4 survives, which gives $\psi_n \Delta_t \mathcal{V} \to 1$.

Why the ridged branch is a function of s alone. The last item of Result D.1 is not only read off the numerics: it follows from Theorem D.4 in four steps, in the limit $\Delta_t \to 0$ at fixed $s = m\psi_n \Delta_t$, hence $\psi_n m = s/\Delta_t \to \infty$.

First, the two channel denominators saturate. With $w_1 = m e^{-2t} + \Delta_t \to m$, $c_1 = \bar{\mu} \to m\mu_*^2$, $w_2 = \Delta_t$ and $c_2 = \mu_I = \Delta_t \iota + O(\Delta_t^2)$, equation (154) gives

$$L_1 - 1 = \frac{m\Delta_t}{s}\left(\mu_1 \bar{r} + \frac{\mu_*^2}{\tilde{\gamma}}\right), \qquad L_2 - 1 = \frac{\Delta_t^2}{s}\left(\mu_1 \bar{r} + \frac{\iota}{\tilde{\gamma}}\right), \tag{181}$$

so $L_1, L_2 \to 1$ as soon as $\tilde{\gamma} \gg \mu_*^2/\psi_n$. Second, the fixed point becomes explicit: $g \to m$ and $R \to \mu_1$, so $1/\bar{r} = \tilde{\gamma} + R(\bar{r})$ collapses to $\bar{r} = (\tilde{\gamma} + \mu_1)^{-1}$ and

$$b = \frac{\mu_1 \bar{r} g}{m} = \frac{\mu_1}{\tilde{\gamma} + \mu_1}, \qquad \mathcal{B}^2 = (b-1)^2 = \frac{\tilde{\gamma}^2}{(\tilde{\gamma} + \mu_1)^2}. \tag{182}$$

Third, the trace. From $L_1' = \mu_1 m \Delta_t/s$ and $L_2' = \mu_1 \Delta_t^2/s$ one gets $g' \to -\mu_1 m \Delta_t/s$, hence $R' \to 0$, $\bar{r}^{-2} + R' \to (\tilde{\gamma} + \mu_1)^2$ and $\Phi' \to -\mu_1 \Delta_t (1 + 1/s)$, so that $\mu_1 T_4/\Delta_t = \mu_1^2 (1 + 1/s)/(\tilde{\gamma} + \mu_1)^2$ and

$$\mathcal{V} = \frac{\mu_1 T_4}{\Delta_t} - b^2 = \frac{\mu_1^2}{s\,(\tilde{\gamma} + \mu_1)^2}. \tag{183}$$

Equations (182) and (183) are stronger than Result D.1: the whole ridge curve, not only its minimum, depends on ψ_n, m and t through s alone, and $\mathcal{B}^2$ does not depend on s at all. Fourth, the optimum. The excess risk is

$$\mathcal{B}^2 + \mathcal{V} = \frac{\tilde{\gamma}^2 + \mu_1^2/s}{(\tilde{\gamma} + \mu_1)^2}, \tag{184}$$

whose derivative is proportional to $(\tilde{\gamma} + \mu_1)(\mu_1\tilde{\gamma} - \mu_1^2/s)$ and vanishes at $\tilde{\gamma}^* = \mu_1/s$; substituting gives (179).

D.12 Link between kernels and random features

Any dot-product kernel $K(\mathbf{x}, \mathbf{y}) = f(\frac{\mathbf{x}^\top \mathbf{y}}{d})$ can be written as the infinite-width limit of a random-features model (Rahimi & Recht, 2007) $K(\mathbf{x}, \mathbf{y}) = \lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \sigma(\frac{\mathbf{w}_i^\top \mathbf{x}}{\sqrt{d}})\sigma(\frac{\mathbf{w}_i^\top \mathbf{y}}{\sqrt{d}})$ with $\mathbf{W} \sim \mathcal{N}(0, \mathbf{I}_{p\times d})$ and an activation σ that depends on f. We show below that the large-p limit of the Gram matrix of the random-features network on our dataset yields the same linear equivalent as Theorem D.1. The difference with Bonnaire et al. (2026) is that they take the number of noises $m \to \infty$ first, whereas we keep $m = O(1)$. The results of this section are presented for $\mathbf{\Sigma} = \mathbf{I}_d$ but extend to arbitrary covariance. We first recall the choice of activation realizing a given dot-product kernel.

Lemma D.3. *For any dot-product kernel $K(\mathbf{x}, \mathbf{y}) = f(\frac{\mathbf{x}^\top \mathbf{y}}{d})$, there exists an activation function σ such that*

$$K(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}[\sigma(\frac{\mathbf{w}^\top \mathbf{x}}{\sqrt{d}})\sigma(\frac{\mathbf{w}^\top \mathbf{y}}{\sqrt{d}})] \tag{185}$$

Proof. Assume that f can be expanded as a Taylor series $f(u) = \sum_{k\geq 0} a_k u^k$ with $a_k \geq 0$[5] and that σ can be expanded on a basis of Hermite polynomials $\sigma(u) = \sum_{k\geq 0} b_k H_k(u)$. By Mehler's formula (Kibble, 1945),

$$\mathbb{E}_{\mathbf{w}}[\sigma(\frac{\mathbf{w}^\top \mathbf{x}}{\sqrt{d}})\sigma(\frac{\mathbf{w}^\top \mathbf{y}}{\sqrt{d}})] = \sum_{k\geq 0} b_k^2 k! \left(\frac{\mathbf{x}^\top \mathbf{y}}{d}\right)^k, \tag{186}$$

hence it suffices to choose the Hermite coefficients of σ such that $\forall k \geq 0,\ b_k^2 k! = a_k$. □

We then derive a Gaussian Equivalence Principle (GEP) (Gerace et al., 2020; Goldt et al., 2021) for the random-features model.

Proposition D.3 (Gaussian equivalence for the random-features model)**.** *Assume $\mu_0' = \mathbb{E}_{\mathcal{N}(0,1)}[\sigma(u)] = 0$. In the proportional regime $p \asymp n \asymp d \gg 1$, the random-features model $\sigma(\frac{\mathbf{WY}}{\sqrt{d}})$ is equivalent to the Gaussian*

$$\mu_1' \frac{\mathbf{WY}}{\sqrt{d}} + \mu_*' \mathbf{\Omega} \tag{187}$$

with $\mathbf{\Omega} \in \mathbb{R}^{p\times mn}$ with Gaussian entries and covariance $\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_j^{\nu'\alpha'}] = \delta_{ij}\delta^{\nu\nu'}\left(\delta^{\alpha\alpha'} + \kappa_t(1-\delta^{\alpha\alpha'})\right)$ with $\mu_1' = \mathbb{E}[\sigma(z)z]$, $\mu_*'^2 = \mathbb{E}[\sigma^2(z)] - \mu_1'^2$ and $\kappa_t = \frac{1}{\mu_*'^2}\mathbb{E}_{u,v,w}[(\sigma(e^{-t}u + \sqrt{\Delta_t}v) - \mu_1' e^{-t}u)(\sigma(e^{-t}u + \sqrt{\Delta_t}w) - \mu_1' e^{-t}u)]$.

Proof. According to the Gaussian Equivalence Principle, $\sigma(\frac{\mathbf{WY}}{\sqrt{d}})$ and $\frac{\mathbf{WY}}{\sqrt{d}}$ are jointly Gaussian; hence, to characterize their statistics, it suffices to compute their covariance.

For each $i \in [p]$, $\nu \in [n]$, $\alpha \in [m]$, define the pre-activation

$$\mathbf{z}_i^{\nu\alpha} = \frac{1}{\sqrt{d}} \sum_{k=1}^{d} \mathbf{W}_{ik}\, \mathbf{Y}_k^{\nu\alpha}.$$

[5]The positivity of the coefficients holds since we assumed that $K(\mathbf{x}, \mathbf{y})$ is a positive-definite kernel.

Conditioned on $\mathbf{Y}$, the $\mathbf{z}_i^{\nu\alpha}$ are jointly centered Gaussian with

$$\mathbb{E}\left[\mathbf{z}_i^{\nu\alpha}\,\mathbf{z}_j^{\nu'\alpha'}\,\middle|\,\mathbf{Y}\right] \;=\; \delta_{ij}\,\frac{\mathbf{Y}^{\nu\alpha}\cdot\mathbf{Y}^{\nu'\alpha'}}{d}.$$

Define the centered nonlinearity

$$\tilde{\Omega}(z) \;:=\; \frac{\sigma(z)-\mu_1' z}{\mu_*'}, \qquad \mu_1' = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z\,\sigma(z)], \quad (\mu_*')^2 = \mathbb{E}[\sigma(z)^2] - (\mu_1')^2.$$

Direct calculation, using the assumption $\mu_0' = \mathbb{E}[\sigma] = 0$ and Stein's identity $\mathbb{E}[z\sigma(z)] = \mathbb{E}[\sigma'(z)]$ for $z\sim\mathcal{N}(0,1)$, gives the orthonormality conditions

$$\mathbb{E}_z[\tilde{\Omega}(z)] = 0, \qquad \mathbb{E}_z[z\,\tilde{\Omega}(z)] = 0, \qquad \mathbb{E}_z[\tilde{\Omega}(z)^2] = 1.$$

Setting $\mathbf{\Omega}_i^{\nu\alpha} = \tilde{\Omega}(\mathbf{z}_i^{\nu\alpha})$ yields

$$\sigma\left(\tfrac{(\mathbf{WY})_i^{\nu\alpha}}{\sqrt{d}}\right) \;=\; \mu_1'\,\tfrac{(\mathbf{WY})_i^{\nu\alpha}}{\sqrt{d}} \;+\; \mu_*'\,\mathbf{\Omega}_i^{\nu\alpha}. \tag{188}$$

This equality holds pointwise, not just asymptotically. We still need to show that $\mathbf{\Omega}$ behaves as a Gaussian matrix with a specific covariance in the proportional regime; we verify the covariance below and invoke the GEP for the equivalence as Gaussian objects. Let us compute $\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_j^{\nu'\alpha'}]$ case by case.

Different rows of $\mathbf{W}$ ($i\neq j$). The rows $\mathbf{w}_i, \mathbf{w}_j$ are independent, so $\mathbf{z}_i^{\nu\alpha}, \mathbf{z}_j^{\nu'\alpha'}$ are conditionally independent and $\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_j^{\nu'\alpha'}] = \mathbb{E}[\tilde{\Omega}(\mathbf{z}_i^{\nu\alpha})]\mathbb{E}[\tilde{\Omega}(\mathbf{z}_j^{\nu'\alpha'})] = 0$. This produces the prefactor δ_{ij}.

Different clusters ($i=j$, $\nu\neq\nu'$). The overlap $\mathbf{Y}^{\nu\alpha}\cdot\mathbf{Y}^{\nu'\alpha'}/d$ vanishes asymptotically (by independence of $\mathbf{x}^\nu, \mathbf{x}^{\nu'}$ and Gaussian concentration), so $\mathbf{z}_i^{\nu\alpha}$ and $\mathbf{z}_i^{\nu'\alpha'}$ are asymptotically independent. Then $\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_i^{\nu'\alpha'}] \to \mathbb{E}[\tilde{\Omega}(z)]^2 = 0$, giving the prefactor $\delta^{\nu\nu'}$.

Diagonal ($i=j$, $\nu=\nu'$, $\alpha=\alpha'$). Same Gaussian, $\mathbb{E}[\tilde{\Omega}(z)^2] = 1$.

Same cluster, different noise copies ($i=j$, $\nu=\nu'$, $\alpha\neq\alpha'$). The overlap concentrates around

$$\frac{\mathbf{Y}^{\nu\alpha}\cdot\mathbf{Y}^{\nu\alpha'}}{d} \;\xrightarrow{d\to\infty}\; e^{-2t}\sigma^2 \;=\; e^{-2t}$$

(the last equality assumes $\sigma^2=1$; the general isotropic case is discussed at the end of this proof). Conditional on $\mathbf{Y}$, $(\mathbf{z}_i^{\nu\alpha}, \mathbf{z}_i^{\nu\alpha'})$ is a centered Gaussian pair with unit marginal variance and correlation $\rho = e^{-2t}$. We realize this pair as

$$\mathbf{z}_i^{\nu\alpha} = e^{-t}u + \sqrt{\Delta_t}\,v, \qquad \mathbf{z}_i^{\nu\alpha'} = e^{-t}u + \sqrt{\Delta_t}\,w,$$

with $u, v, w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then

$$(\mu_*')^2\,\mathbb{E}\left[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_i^{\nu\alpha'}\right] \;=\; \mathbb{E}\left[(\sigma(z)-\mu_1' z)(\sigma(z')-\mu_1' z')\right]. \tag{189}$$

Expand the product and use Stein's lemma at correlation $\rho = e^{-2t}$, which gives $\mathbb{E}[z'\,\sigma(z)] = \rho\,\mathbb{E}[\sigma'(z)] = \rho\,\mu_1'$ and similarly $\mathbb{E}[z\,\sigma(z')] = \rho\,\mu_1'$:

$$\begin{aligned}\mathbb{E}[(\sigma(z)-\mu_1' z)(\sigma(z')-\mu_1' z')] &= \mathbb{E}[\sigma(z)\sigma(z')] \;-\; 2\,\mu_1'\,\rho\,\mu_1' \;+\; (\mu_1')^2\,\mathbb{E}[zz'] \\ &= \mathbb{E}[\sigma(z)\sigma(z')] - (\mu_1')^2\,\rho.\end{aligned}$$

The proposition defines κ_t via the integrand $(\sigma(e^{-t}u+\sqrt{\Delta_t}v) - \mu_1' e^{-t}u)(\sigma(e^{-t}u+\sqrt{\Delta_t}w) - \mu_1' e^{-t}u)$, which subtracts only $\mu_1' e^{-t}u$ from each factor instead of the full $\mu_1' z$. The two definitions coincide because the missing $\mu_1'\sqrt{\Delta_t}v$ and $\mu_1'\sqrt{\Delta_t}w$ pieces cancel: they are mean-zero, mutually independent, and each of them is independent of the remaining terms in the cross expectation. Concretely, expanding the product

$$(\sigma(z) - \mu_1' e^{-t}u - \mu_1'\sqrt{\Delta_t}v)(\sigma(z') - \mu_1' e^{-t}u - \mu_1'\sqrt{\Delta_t}w)$$

and taking expectation, the cross terms involving v or w alone vanish (because v is independent of u and w, and $\mathbb{E}[v] = 0$; similarly for w), and the term $(\mu_1')^2 \Delta_t \, vw$ vanishes by $\mathbb{E}[vw] = 0$. Hence (189) equals

$$\mathbb{E}\left[(\sigma(e^{-t}u + \sqrt{\Delta_t}v) - \mu_1' e^{-t}u)(\sigma(e^{-t}u + \sqrt{\Delta_t}w) - \mu_1' e^{-t}u)\right] \;=\; (\mu_*')^2 \, \kappa_t,$$

which is the definition of κ_t in the proposition. So $\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_i^{\nu\alpha'}] = \kappa_t$ for $\alpha \neq \alpha'$.

Combining the four cases:

$$\mathbb{E}[\mathbf{\Omega}_i^{\nu\alpha}\mathbf{\Omega}_j^{\nu'\alpha'}] \;=\; \delta_{ij}\,\delta^{\nu\nu'}\left(\delta^{\alpha\alpha'} + \kappa_t\,(1 - \delta^{\alpha\alpha'})\right).$$

This is the claimed covariance.

General isotropic data $\mathbf{\Sigma} = \sigma^2\mathbf{I}_d$. The same computation goes through with two changes: the pre-activations have marginal variance $\|\mathbf{Y}^{\nu\alpha}\|^2/d \to \Gamma_t = \sigma^2 e^{-2t} + \Delta_t$ instead of 1, and two copies of the same clean sample have correlation $\rho = \sigma^2 e^{-2t}/\Gamma_t$ instead of e^{-2t}. The Gaussian equivalence then holds with μ_1', μ_*' and κ_t computed for $z \sim \mathcal{N}(0, \Gamma_t)$ and pairs of correlation ρ, and the resulting constants of the Gram matrix, $\mu_I = \mu_*'^2(1 - \kappa_t)$, $\mu_B = \mu_*'^2\kappa_t$ and $\mu_1 = \mu'^2_1$, are exactly those of Theorem D.1 at general σ^2. The equations of Theorem D.4 therefore hold with these constants; equivalently, the exact reduction to $\sigma^2 = 1$ stated after Theorem D.4 applies.

The Gaussian Equivalence Principle of Gerace et al. (2020); Goldt et al. (2021); Hu & Lu (2023); Mei & Montanari (2020) asserts that, in the proportional regime $p \asymp n \asymp d$, all asymptotic statistics of $\sigma(\mathbf{WY}/\sqrt{d})$ coincide with those obtained by replacing $\mathbf{\Omega}$ by a centered Gaussian matrix with the same covariance. Combining with the exact decomposition above gives the claimed distributional equivalence $\sigma(\mathbf{WY}/\sqrt{d}) \overset{\mathrm{d}}{\sim} \mu_1'\mathbf{WY}/\sqrt{d} + \mu_*'\mathbf{\Omega}^{\mathrm{Gauss}}$, where $\mathbf{\Omega}^{\mathrm{Gauss}}$ has the prescribed covariance and is otherwise jointly Gaussian.

□

Proposition D.4. *In the overparametrized regime $p \gg n \asymp d \gg 1$, the Gram matrix of the random-features model $\mathbf{G}_{\mathrm{RF}} = \frac{1}{p}\sum_{i=1}^{p} \sigma(\frac{\mathbf{W}_i\mathbf{X}}{\sqrt{d}})^\top \sigma(\frac{\mathbf{W}_i\mathbf{Y}}{\sqrt{d}})$ converges to*

$$\mu_I \mathbf{I}_N + \mu_B \mathbf{B}_m + \mu_1 \frac{\mathbf{Y}^\top\mathbf{Y}}{d} \tag{190}$$

with $\mu_I = \mu'^2_*(1 - \kappa_t)$, $\mu_B = \mu'^2_*\kappa_t$, $\mu_1 = \mu'^2_1$.

Proof. Replace the random features by their Gaussian equivalent and the sum over hidden units by an expectation over $\mathbf{W}$ and $\mathbf{\Omega}$.

$$\mathbf{G}_{\mathrm{RF}}^{\nu\alpha,\nu'\alpha'} = \mathbb{E}_{\mathbf{W},\mathbf{\Omega}}\left[\left(\mu_1'\frac{\mathbf{W}^\top\mathbf{Y}^{\nu\alpha}}{\sqrt{d}} + \mu_*'\mathbf{\Omega}^{\nu\alpha}\right)\left(\mu_1'\frac{\mathbf{W}^\top\mathbf{Y}^{\nu'\alpha'}}{\sqrt{d}} + \mu_*'\mathbf{\Omega}^{\nu'\alpha'}\right)\right] \tag{191}$$

$$= \mu'^2_1\frac{\mathbf{Y}^{\nu\alpha}\cdot\mathbf{Y}^{\nu'\alpha'}}{d} + \mu'^2_*\mathbb{E}[\mathbf{\Omega}^{\nu\alpha}\mathbf{\Omega}^{\nu'\alpha'}] \tag{192}$$

$$= \mu'^2_1\frac{\mathbf{Y}^{\nu\alpha}\cdot\mathbf{Y}^{\nu'\alpha'}}{d} + \mu'^2_*\left(\delta^{\nu\nu'}\delta^{\alpha\alpha'} + \kappa_t\delta^{\nu\nu'}(1 - \delta^{\alpha\alpha'})\right). \tag{193}$$

Hence

$$\mathbf{G}_{\mathrm{RF}} = \mu'^2_1\frac{\mathbf{Y}^\top\mathbf{Y}}{d} + \mu'^2_*(1 - \kappa_t)\mathbf{I}_N + \mu'^2_*\kappa_t\mathbf{B}_m. \tag{194}$$

□

The infinite-width limit of the Gram matrix of a random-features network is thus consistent with the linear equivalent of the Gram matrix of the associated kernel.

The same Gaussian Equivalence Principle also accounts for the finite-width, infinite-m limit studied by Bonnaire et al. (2026). These authors work with the Hessian of the loss at initialization,

$$\mathbf{U}^{ij} = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}\left[\sigma\left(\frac{(\mathbf{W}(e^{-t}\mathbf{x}^{\nu}+\sqrt{\Delta_t}\boldsymbol{\xi}))_i}{\sqrt{d}}\right)\sigma\left(\frac{(\mathbf{W}(e^{-t}\mathbf{x}^{\nu}+\sqrt{\Delta_t}\boldsymbol{\xi}))_j}{\sqrt{d}}\right)\right], \tag{195}$$

a $p \times p$ matrix in feature space whose counterpart at finite m is

$$\mathbf{U}^{ij}(m) = \frac{1}{nm}\sum_{\nu=1}^{n}\sum_{\alpha=1}^{m} \sigma\left(\frac{(\mathbf{W}\mathbf{Y}^{\nu\alpha})_i}{\sqrt{d}}\right)\sigma\left(\frac{(\mathbf{W}\mathbf{Y}^{\nu\alpha})_j}{\sqrt{d}}\right), \tag{196}$$

so that $\mathbf{U} = \lim_{m\to\infty}\mathbf{U}(m)$. Replacing the features by their Gaussian equivalent,

$$\mathbf{U}^{ij}(m) = \frac{1}{nm}\sum_{\nu=1}^{n}\sum_{\alpha=1}^{m}\left(\mu_1'\tfrac{(\mathbf{W}\mathbf{Y})_i^{\nu\alpha}}{\sqrt{d}}+\mu_*'\,\boldsymbol{\Omega}_i^{\nu\alpha}\right)\left(\mu_1'\tfrac{(\mathbf{W}\mathbf{Y})_j^{\nu\alpha}}{\sqrt{d}}+\mu_*'\,\boldsymbol{\Omega}_j^{\nu\alpha}\right), \tag{197}$$

and the three groups of terms average separately. In the Gaussian equivalent $\boldsymbol{\Omega}$ is centered and uncorrelated with $\mathbf{W}\mathbf{Y}/\sqrt{d}$, so the two cross terms vanish,

$$\frac{\mu_1'\mu_*'}{nm}\sum_{\nu,\alpha}\frac{(\mathbf{W}\mathbf{Y})_i^{\nu\alpha}}{\sqrt{d}}\,\boldsymbol{\Omega}_j^{\nu\alpha} \;\longrightarrow\; 0. \tag{198}$$

The $\boldsymbol{\Omega}$–$\boldsymbol{\Omega}$ term pairs every copy with itself and never with another, so the cross-copy covariance κ_t is never probed and

$$\frac{\mu_*'^2}{nm}\sum_{\nu,\alpha}\boldsymbol{\Omega}_i^{\nu\alpha}\boldsymbol{\Omega}_j^{\nu\alpha} \;\longrightarrow\; \mu_*'^2\,\delta_{ij}. \tag{199}$$

The average of the linear part over the copies becomes an expectation over $\boldsymbol{\xi}\sim\mathcal{N}(0,\mathbf{I}_d)$ at fixed $\mathbf{W}$ and fixed data, which gives

$$\frac{\mu_1'^2}{nmd}\sum_{\nu,\alpha}(\mathbf{W}\mathbf{Y})_i^{\nu\alpha}(\mathbf{W}\mathbf{Y})_j^{\nu\alpha} \;\longrightarrow\; \mu_1'^2\left[\Delta_t\frac{(\mathbf{W}\mathbf{W}^\top)_{ij}}{d}+e^{-2t}\frac{(\mathbf{W}\hat{\boldsymbol{\Sigma}}\mathbf{W}^\top)_{ij}}{d}\right], \tag{200}$$

with $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_\nu \mathbf{x}^\nu(\mathbf{x}^\nu)^\top$ the empirical covariance of the clean samples. Collecting the three pieces, and using $\mu_1 = \mu_1'^2$ together with $\mu_*'^2 = \mu_I + \mu_B$ from the proposition above,

$$\mathbf{U} \;=\; \mu_1\frac{\mathbf{W}\hat{\boldsymbol{\Sigma}}_t\mathbf{W}^\top}{d} \;+\; (\mu_I+\mu_B)\,\mathbf{I}_p, \qquad \hat{\boldsymbol{\Sigma}}_t = e^{-2t}\hat{\boldsymbol{\Sigma}}+\Delta_t\mathbf{I}_d. \tag{201}$$

Equation (201) is the feature-space mirror of the linear equivalent of Theorem D.1, and it reproduces the two-group spectrum from which Bonnaire et al. (2026) read off the two timescales: $\mathbf{W}\hat{\boldsymbol{\Sigma}}_t\mathbf{W}^\top/d$ has rank d, so $\mathbf{U}$ carries d large eigenvalues, the fast directions, above a plateau of $p-d$ eigenvalues pinned exactly at $\mu_I+\mu_B$, the slow ones.

The comparison with the finite-m Gram matrix is then transparent. Since κ_t drops out, μ_I and μ_B enter (201) only through their sum: the $m=\infty$ Hessian has exactly the structure of our own Gram matrix at $m=1$, where $\mathbf{B}_m = \mathbf{I}_N$ and $\mathbf{G}_{\text{lin}} = (\mu_I+\mu_B)\mathbf{I}_N + \mu_1\mathbf{Y}^\top\mathbf{Y}/d$. What $m>1$ does is resolve that single degenerate plateau into the three objects of Theorem D.2: the atom at μ_I, the atom at $\bar{\mu} = \mu_I + m\mu_B$ and the memorization bulk at $\mu_I+\mu_B(m-1)\Delta_t/\lambda_t$, each of which depends on μ_I and μ_B separately and therefore on the cross-copy correlation κ_t. Averaging over the copies before forming the matrix, as the $m=\infty$ limit does, removes precisely the structure that the memorization bulk lives on.

E Experimental details

E.1 General comments

Code. A reference implementation of all the numerical experiments will be made available upon publication of the paper.

Generation. Once the velocity field $\hat{\mathbf{v}}(\mathbf{x}_s, s)$ is estimated (either by a U-Net or by the spectrally truncated estimator $\hat{\mathbf{v}}_r$) on a uniform grid of S time slices $\{s_k\}_{k=0}^{S-1} \subset [0,1]$, we generate samples by integrating the rectified-flow ODE

$$\frac{\mathrm{d}\tilde{\mathbf{x}}_s}{\mathrm{d}s} = \hat{\mathbf{v}}(\tilde{\mathbf{x}}_s, s), \qquad \tilde{\mathbf{x}}_1 \sim \mathcal{N}(0, \mathbf{I}_d), \tag{202}$$

backward with an Euler scheme on the same S slices used at training. For the truncation experiment, at each timestep, the cross-kernel $K(\tilde{\mathbf{x}}_{s_k}, \mathbf{Y}_{s_k})$ between the current iterate and the noised training anchors is evaluated through the same closed-form NTK recursion and the velocity is computed with a single matrix-vector product against the precomputed coefficients $\boldsymbol{\alpha}_r$ of the truncated estimator $\hat{\mathbf{v}}_r$ of (203), i.e. $K(\tilde{\mathbf{x}}_{s_k}, \mathbf{Y}_{s_k})\, \boldsymbol{\alpha}_r(s_k)$.

Details on f_{mem}. For each generated sample $\tilde{\mathbf{x}}_0$ with the backward dynamics, let d_{NN_1} and d_{NN_2} denote the cosine distances to its first and second nearest neighbors in the training set $\{\mathbf{x}^\nu\}_{\nu=1}^n$. We declare $\tilde{\mathbf{x}}_0$ memorized whenever $\frac{d_{\mathrm{NN}_1}(\tilde{\mathbf{x}}_0)}{d_{\mathrm{NN}_2}(\tilde{\mathbf{x}}_0)} < \frac{1}{3}$, and we define $f_{\mathrm{mem}} \in [0,1]$ as the fraction of generated samples that meet this condition among 2048 generated samples. We emphasize that f_{mem} is evaluated on backward trajectories starting from training noises $\tilde{\mathbf{x}}_1 \in \{\boldsymbol{\xi}^{\mu\alpha}\}$ and can therefore be viewed as a long-time version of the U-turn protocol from Behjoo & Chertkov (2025); Sclocchi et al. (2024). A model that has learned the correct population velocity field should produce an image that resembles those of the data manifold but is distinct from its associated clean image, therefore producing $f_{\mathrm{mem}} = 0$, while a model in the memorization regime should fall back onto the clean training image and lead to $f_{\mathrm{mem}} = 1$.

E.2 Closed-form CNTK computation and top-r eigendecomposition

Kernel recursion. We follow the depth-D recursion of Arora et al. (2019), described in Sect. A.7 and implemented in the authors' `CNTK` repository, except we use a 1×1 readout which is equivalent to the flatten+linear output up to a $1/P$ normalization with $P = L^2$ which is a global multiplicative constant that is absorbed in the kernel-regression predictors.

Top-r eigenvalues. For $N = nm < 10^4$, we materialize the empirical Gram matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$ and call a standard Lanczos solver (`ARPACK` package) to find its eigenvalues. For $N > 10^4$, we use the randomized SVD method from Halko et al. (2011) with a matrix-free linear operator that evaluates matrix-vector products in chunks without ever forming $\mathbf{G}$ explicitly. We use an oversampling of $p = 10$ columns and a second-order ($q = 2$) power iteration, i.e. $(q+2)(r+p) \approx 4r$ matrix-vector products in total to compute the top-r eigenvalues.

Spectrally truncated estimator. Writing the resulting decomposition as $\mathbf{G} \approx \mathbf{U}_r \boldsymbol{\Lambda}_r \mathbf{U}_r^\top$, with $\mathbf{U}_r$ collecting in column the top-r eigenvectors of $\mathbf{G}$ and $\boldsymbol{\Lambda}_r$ the corresponding descending-ordered eigenvalues, the spectrally truncated kernel regression estimator of the velocity field used in Sect. 3 is

$$\hat{\mathbf{v}}_r(\mathbf{x}, s) = K(\mathbf{x}, \mathbf{Y})\, \boldsymbol{\alpha}_r = K(\mathbf{x}, \mathbf{Y})\, \mathbf{U}_r \boldsymbol{\Lambda}_r^{-1} \mathbf{U}_r^\top \mathbf{V}, \tag{203}$$

where $\mathbf{V} \in \mathbb{R}^{N \times d}$ stacks the training target velocities.

Truncation rank and training time. Truncating at rank r is equivalent to fixing a training time. The gradient-flow filter of (5) activates the modes above $\lambda_c \sim nm/(d\tau)$ and leaves the others untouched, so discarding every eigenvalue below λ_r is the hard-threshold counterpart of stopping at $\tau \sim nm/(d\lambda_r)$. This is what makes the truncation rank of Sect. 3 a proxy for the training time, and the generalization to memorization transition it produces the same one that the theory of Sect. 2 describes as a function of τ. The three rules (truncation, gradient flow, ridge) differ in the shape of the filter, not in the order in which they activate the modes; the paragraph *Early stopping versus ridging* of Sect. D.10 quantifies how far the ridge and the flow depart from one another.

Scaling with m of the CNTK. Figure 12 displays the evolution of the CNTK spectrum for several values of $m \in \{1, 2, 4, 8, 16, 32, 64\}$ at fixed $n = 256$. The entire spectrum shifts with m: the first bulk linearly in m, in agreement with the theoretical prediction, and the second bulk sub-linearly, as emphasized by the inset. The figure also shows the absence of the memorization bulk at $m = 1$, where the standard single-bulk spectrum is recovered.

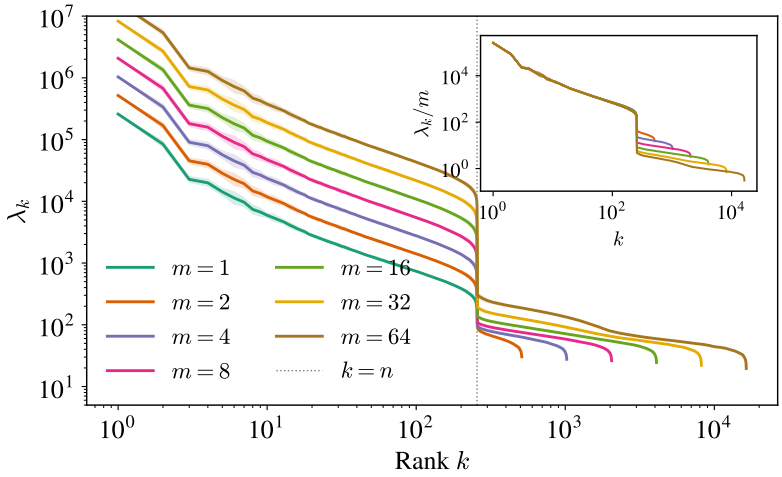

Figure 12: Ordered eigenvalues for several m at fixed $n = 256$ and their rescaling by m in the inset. Spectra obtained at $s = 0.05$ and averaged over 10 realizations of the training set.

E.3 U-Net training and empirical NTK on a finite-width network

At the end of Sect. 3, we compute the spectrum of the time-dependent empirical NTK (eNTK) $K(\boldsymbol{\theta}; \mathbf{x}, \mathbf{x}') = \sum_{i,k} \partial_{\theta_i} \mathbf{v}_{\boldsymbol{\theta}}^{(k)}(\mathbf{x})\, \partial_{\theta_i} \mathbf{v}_{\boldsymbol{\theta}}^{(k)}(\mathbf{x}')$ of a U-Net architecture.

U-Net architecture and training. The model follows the implementation of Bonnaire et al. (2026) with three residual blocks, base width $W = 32$ and channel multipliers $\{1, 2, 4\}$, trained on CelebA with $n = 256$ images and $m = 8$ noise realizations per data point. The model is trained on the rectified-flow objective with Adam at a learning rate of 10^{-4} for a total of $\tau = 60\,000$ steps.

Empirical NTK estimator. Materializing the full Jacobian $\partial_{\boldsymbol{\theta}} \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathbb{R}^{d \times |\boldsymbol{\theta}|}$ is intractable, so we use a Hutchinson-style projection estimator of the empirical NTK matrix

$$K_{\text{eNTK}} \approx \frac{1}{n_{\text{proj}}} \sum_{r=1}^{n_{\text{proj}}} \langle \nabla_{\boldsymbol{\theta}} (\mathbf{w}_r^\top \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x})), \nabla_{\boldsymbol{\theta}} (\mathbf{w}_r^\top \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}')) \rangle, \tag{204}$$

with $\mathbf{w}_r \sim \mathcal{N}(0, \mathbf{I}_d)$ and $n_{\text{proj}} = 16$. Each scalar $\mathbf{w}_r^\top \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x})$ is differentiated with a single reverse-mode pass so that the full Jacobian is never formed. The resulting gradients are batched over data points and contracted into the kernel in chunks.

Regularized rectified-flow training. We train the velocity network $\mathbf{v}_{\boldsymbol{\theta}}$ on the training set $\{(\mathbf{x}^\nu, \boldsymbol{\xi}^{\nu\alpha})\}$ with a spectrum-aware Tikhonov penalty

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{nmd} \sum_{\nu=1}^{n} \sum_{\alpha=1}^{m} \mathbb{E}_s \left\| \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{Y}_s^{\nu\alpha}, s) - (\boldsymbol{\xi}^{\nu\alpha} - \mathbf{x}^\nu) \right\|_2^2 + \frac{\gamma_{s,\tau}}{nmd^2} \|\boldsymbol{\theta}\|_2^2, \tag{205}$$

where $\mathbf{Y}_s^{\nu\alpha} = (1 - s)\mathbf{x}^\nu + s\boldsymbol{\xi}^{\nu\alpha}$, and $\gamma_{s,\tau}$ is the regularization parameter, chosen as the n-th eigenvalue of the empirical NTK Gram matrix $\mathbf{G}_\tau(s) \in \mathbb{R}^{nm \times nm}$ evaluated on the full training set at flow-matching time s. The additional $1/d$ factor in the prefactor of the regularization converts the traced-eNTK eigenvalues we measure into those of the full NTK under the output-isotropy assumption (12). In the linearization around $\boldsymbol{\theta}_0$, the stationarity of Eq. (205) is equivalent to a kernel-ridge problem $(\mathbf{G}(s) + \gamma \mathbf{I})\boldsymbol{\alpha} = \mathbf{V}$. A regularization of amplitude $\gamma_{s,\tau} = \lambda_n$ then damps the contributions of the eigenmodes of $\mathbf{G}$ below the n-th. As the eNTK drifts with training time in the feature-learning regime, the parameter $\gamma_{s,\tau}$ has to be tracked. At step 0 we eigendecompose $\mathbf{G}(s)$ on the current parameters $\boldsymbol{\theta}_0$ for each s in the discrete grid of size S, and refresh every $\Delta\tau = 1000$ training steps by repeating the same procedure on $\boldsymbol{\theta}(\tau)$. In Fig. 13, we display the evolution of the train and test errors corresponding to the training runs of Fig. 5. Without regularization (left panel), the test loss increases sharply at large training time, together with a steep rise in memorization, whereas with the regularization (right panel) the network stays close to the configuration of optimal test loss.

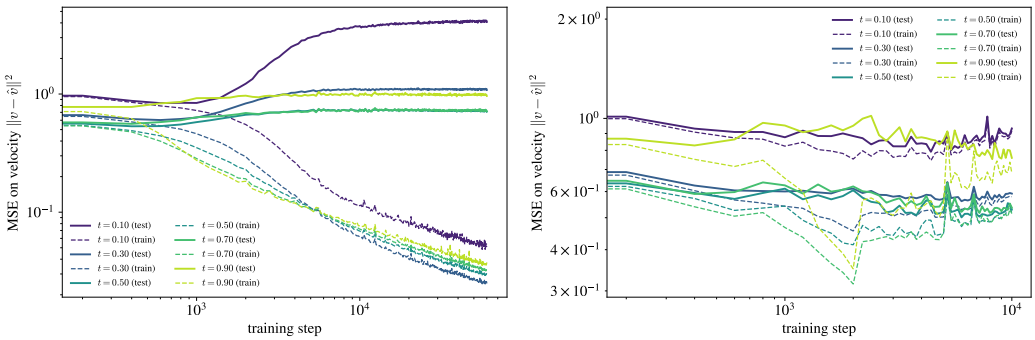

Figure 13: Training and test losses along training for the *(Left)* unregularized and *(Right)* regularized dynamics of Fig. 5.

E.4 Computational resources

All experiments were run on NVIDIA H100 GPUs: a single GPU for $n < 2048$ and four GPUs in parallel for $n \geq 2048$. As an example, the truncation experiment (kernel regression and generation) for $n = 2048$ takes about 4 hours on four GPUs.

F LLM usage

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

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

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

Proof assistant. LLMs were used as a proof assistant: to check derivations step by step, to look for errors in the proofs, and to write the code performing the numerical verifications of the analytical results.

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