IN-CONTEXT LEARNING OF SINGLE-INDEX TARGETS: COMPARING KERNEL AND FEATURE LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) enables a pretrained model to infer a task from demonstrations without updating its parameters. While much of the existing theory focuses on linear target functions, in this paper we study nonlinear cases by comparing two one-layer attention architectures on the same family of single-index tasks. A kernel learner first maps inputs through a fixed nonlinear feature map and then applies linear attention, whereas a feature learner applies attention to the original input, followed by a learned nonlinear readout. We derive predictions for their memorization and generalization errors using the replica method, retaining the effects of pretraining size, task-pool diversity, and training and inference context lengths. The resulting predictions closely match numerical experiments across a broad range of regimes. Our analysis yields phase diagrams that characterize when each architecture is advantageous as the amount of pretraining data, task diversity, and context lengths vary. We further identify qualitatively different context-length scalings for the two learners. Together, these results clarify how architectural choices interact with the dataset and govern nonlinear in-context learning.

1 INTRODUCTION

In-context learning (ICL) is the ability of a pretrained model to infer a task from examples provided in its input context and make predictions without updating its parameters. This behavior is one of the most distinctive capabilities of modern Transformers (Brown et al., 2020; Dong et al., 2024), and has motivated a growing theoretical literature that studies ICL in controlled problems. Early work showed that Transformers trained on families of synthetic tasks can learn new functions from demonstrations (Garg et al., 2022), and connected it to familiar learning algorithms such as gradient descent, least squares, and ridge regression (Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Mahankali et al., 2024). These results established that attention-based models can implement nontrivial learning procedures within their forward pass.

A more recent line of work asks how such in-context algorithms are acquired during pretraining. For linear attention, whether a model memorizes the pretraining tasks or generalizes to unseen tasks depends jointly on the number of pretraining prompts, task diversity, and the number of demonstrations available at inference time (Raventós et al., 2023; Nguyen & Nallamala, 2025; Goddard et al., 2025). Recent high-dimensional analyses make these dependencies quantitative and characterize how the shared task structure affects in-context generalization (Lu et al., 2025; Wu et al., 2024; Takanami et al., 2026; Kwon et al., 2026; Letey et al., 2026). Other works also study the learning dynamics of ICL (Singh et al., 2023; Zhang et al., 2025) and the influence of various factors including data structures, distribution shift, hyperparameters and so on (Dragutinović et al., 2025; Takanami et al., 2026; Kwon et al., 2026; Letey et al., 2026; Bordelon et al., 2026).

Most of this theory, however, focuses on linear targets. Nonlinear tasks introduce an additional question: beyond aggregating demonstrations in context, the learner must determine how the nonlinearity on the input should be learned. Single-index models provide a natural setting for this question, and create a classical separation between fixed-kernel methods and feature-learning methods (Ghorbani et al., 2020; Bietti et al., 2022; Ba et al., 2022; Dandi et al., 2024). ICL introduces a further distinction: the relevant direction changes between tasks and must be extracted from the demonstrations without updating the parameters. Recent work has shown that pretrained Transformers can exploit low-dimensional structure to learn single-index targets efficiently in context (Oko et al., 2024),

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.

and that nonlinear Transformers can perform inference-time feature learning on single-index tasks (Nishikawa et al., 2025).

Our goal is thus to compare fixed-kernel methods and feature-learning methods across various amounts of pretraining data, task diversities, and context lengths. We consider two learners that use the same one-layer attention but place the nonlinear transformation on opposite sides of attention. The *kernel learner* first maps every input through a fixed nonlinear feature map and then performs linear attention in the resulting feature space. In contrast, the *feature learner* performs attention directly on the original inputs and applies a learned nonlinear readout after attention.

We study the two architectures on the same learning problem. Pretraining prompts are generated from a finite pool of tasks, allowing us to separately measure performance on tasks that appeared during pretraining and on fresh tasks drawn from the underlying task distribution. We refer to these quantities as the *memorization error* (on seen tasks) and the *generalization error* (on unseen tasks), respectively. Varying the task-pool size therefore lets us study how architectural choices interact with the memorization–generalization tradeoff emphasized in previous work (Raventós et al., 2023; Lu et al., 2025; Goddard et al., 2025; Takanami et al., 2026; Farné et al., 2026).

For analytical tractability, we instantiate the kernels using spherical harmonics and analyze both learners using the non-rigorous yet well-established replica method (Mézard et al., 1987; Zdeborová & Krzakala, 2016). The comparison is defined for common values of the input dimension d, feature dimension D, pretraining size N, task-pool size N_0, and context lengths $L, \widetilde{L}$. We validate these predictions numerically over a broad range of parameters.

Our main contributions are threefold.

- **Comparison of kernel and feature learners on the same dataset.** We formulate the two learners on a common learning problem, with the same task pool, pretraining prompts, and inference prompts, and characterize their memorization and generalization performance as functions of $d, D, N_0, N, L, \widetilde{L}$. For each learner, we derive replica-based predictions and the resulting formulas accurately predict the performance of both learners over a broad range of regimes.
- **Phase diagrams for when each architecture is advantageous.** We map out how the relative performance of the two learners depends jointly on problem parameters. The feature learner benefits from its lower-dimensional representation and typically reaches low error with fewer pretraining prompts, while the kernel learner can achieve lower memorization error when the task pool is small and pretraining is sufficient. We further identify regimes in which context-length mismatch changes the relative advantage of the two architectures.
- **A theoretical characterization of context-length scaling.** In the large-pretraining and large-task-diversity limit, we prove explicit context-length scaling. For a kernel learner containing degree-k spherical harmonics, recovering that component requires a context length of order d^k, which produces a stage-wise learning curve. By contrast, the feature learner exhibits a single transition at $L = \Theta(d)$ and can exploit all Hermite components of the teacher at the same time.

2 Settings

2.1 Task

We consider an ICL problem in which the training dataset consists of N prompts $\mathcal{D} = \{(x_0^\mu, y_0^\mu, x_1^\mu, y_1^\mu, \dots, x_L^\mu, y_L^\mu)\}_{\mu=1}^N$. Each prompt contains L demonstrations $\{(x_\ell^\mu, y_\ell^\mu)\}_{\ell=1}^L$ together with a query-label pair (x_0^μ, y_0^μ). At inference time, the learner observes the demonstrations and the query x_0 and aims to predict the corresponding label y_0.

The inputs are sampled independently and uniformly on the sphere of radius $\sqrt{d}$: $x_\ell^\mu \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$ for $\ell = 0, 1, \dots, L$ and $\mu = 1, 2, \dots, N$. Each prompt is associated with a latent task vector drawn from a finite task pool. Specifically, we first generate $\mathcal{A}_0 = \{a^k\}_{k=1}^{N_0}$ with $a^k \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$ and then independently sample a task vector $a^\mu := a^{\tau^\mu}$ with $\tau^\mu \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{1, \dots, N_0\})$ for each training prompt μ. Conditional on a^μ, the labels are gener-

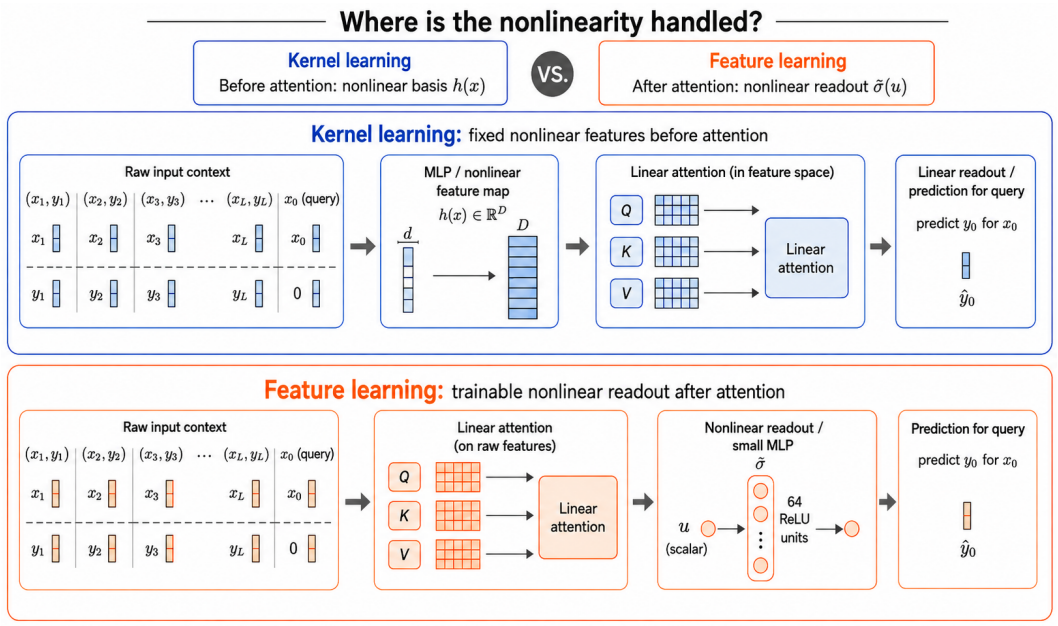

Figure 1: The kernel and feature learners. The kernel learner applies a fixed map $h : \mathbb{R}^d \to \mathbb{R}^D$ before self-attention, whereas the feature learner applies a trainable readout $\widehat{\sigma}$ after self-attention.

ated by the noisy single-index model

$$y_\ell^\mu = \sigma_\star\left(\frac{(a^\mu)^\top x_\ell^\mu}{\sqrt{d}}\right) + \epsilon_\ell^\mu, \qquad \epsilon_\ell^\mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Delta), \tag{1}$$

for $\ell = 0, 1, \ldots, L$. Throughout the paper, $\sigma_\star : \mathbb{R} \to \mathbb{R}$ is a fixed nonlinear teacher function satisfying $\mathbb{E}[\sigma_\star(G)] = 0$, $\mathbb{E}[\sigma_\star(G)^2] = 1$ and $\mu_1 := \mathbb{E}[\sigma_\star(G)G] \neq 0$, where the expectation is over $G \sim \mathcal{N}(0, 1)$. While the first two assumptions are imposed for simplicity, $\mu_1 \neq 0$ is essential. It guarantees a nonvanishing linear component of the teacher and places the problem in the $\mathcal{O}(d^2)$ learning regime (Nishikawa et al., 2025).

We evaluate the trained learner on an inference prompt containing $\widetilde{L}$ demonstrations: $\{(x_\ell, y_\ell)\}_{\ell=1}^{\widetilde{L}}$ and a query x_0 associated with a task vector $a^\star$. To separate the learner performance from irreducible label noise, we use noiseless labels at inference time: $y_\ell = \sigma_\star\left(\frac{(a^\star)^\top x_\ell}{\sqrt{d}}\right)$ for $\ell = 0, 1, \ldots, \widetilde{L}$, where $x_\ell \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right)$. Given the demonstrations and the query, let $\widehat{y}$ denote the learner prediction for y_0.

We distinguish two notions of test performance according to whether the inference task has appeared in the training task pool. We define the memorization and the generalization error by

$$\mathcal{E}_M := \mathbb{E}_{a^\star \sim \mathrm{Unif}(\mathcal{A}_0)} \mathbb{E}_{x_{0:\widetilde{L}}} \left[(y_0 - \widehat{y})^2\right], \; \mathcal{E}_G := \mathbb{E}_{a^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))} \mathbb{E}_{x_{0:\widetilde{L}}} \left[(y_0 - \widehat{y})^2\right]. \tag{2}$$

For memorization, the test task is sampled uniformly from the same pool $\mathcal{A}_0$, whereas for generalization, a fresh task is sampled uniformly from the sphere. Here $x_{0:\widetilde{L}} := (x_0, x_1, \ldots, x_{\widetilde{L}})$, and the expectation is taken over a fresh inference prompt, conditional on the trained learner and, for $\mathcal{E}_M$, on the task pool $\mathcal{A}_0$. Related works have also studied in-distribution and out-of-distribution task generalization by restricting the task vectors to lower-dimensional subspaces (Takanami et al., 2026; Lu et al., 2025). We do not consider the distribution shift here because this is not our main focus.

The above task model is closely related to existing models of ICL (Takanami et al., 2026; Lu et al., 2025; Oko et al., 2024; Nishikawa et al., 2025). In particular, Oko et al. (2024); Nishikawa et al. (2025) characterize the sample complexity required to learn such task families. Our goal is instead to characterize and compare the memorization and generalization errors of the two classes of learners introduced below. A schematic comparison is shown in Figure 1.

2.2 KERNEL LEARNER

We first consider the *kernel learner*, which applies a fixed nonlinear feature map to each input before a single self-attention layer. Let $h : \mathbb{R}^d \to \mathbb{R}^D$ be a fixed feature map. For the μ-th training prompt, we form the embedded prompt

$$E^\mu = \begin{bmatrix} h(x_1^\mu) & \cdots & h(x_L^\mu) & h(x_0^\mu) \\ y_1^\mu & \cdots & y_L^\mu & 0 \end{bmatrix} \in \mathbb{R}^{(D+1)\times(L+1)}, \tag{3}$$

where the final zero corresponds to the masked query label. We apply a single linear self-attention layer,

$$\mathrm{Attn}(E^\mu) = E^\mu + W^V E^\mu \frac{(W^K E^\mu)^\top W^Q E^\mu}{L}, \tag{4}$$

and take the bottom-right entry as the prediction: $\widehat{y}^\mu = \mathrm{Attn}(E^\mu)_{D+1,L+1}$. Following the standard parameterization used in theoretical analyses of self-attention (Takanami et al., 2026; Lu et al., 2025), we restrict the value and key-query matrices to satisfy

$$W^V = \begin{bmatrix} * & * \\ 0 & 1 \end{bmatrix}, \qquad (W^K)^\top W^Q = \begin{bmatrix} W^\top & * \\ 0 & * \end{bmatrix}, \tag{5}$$

where $W \in \mathbb{R}^{D\times D}$ is the trainable attention matrix and the blocks denoted by $*$ do not affect the predicted query label. Under this parameterization, the learner output reduces to

$$\widehat{y}^\mu = \mathrm{Tr}\big(W(H^{\mathrm{K}})^\top\big), \text{ where } H^{\mathrm{K}} := \frac{1}{L} h(x_0^\mu) \left(\sum_{\ell=1}^{L} y_\ell^\mu h(x_\ell^\mu)\right)^\top \in \mathbb{R}^{D\times D} \tag{6}$$

Thus, after the fixed embedding h, training the attention layer is equivalent to a linear regression problem in the D^2 entries of H_μ^{K}. We train W by empirical risk minimization,

$$\widehat{W}_{\mathrm{K}} = \arg\min_{W\in\mathbb{R}^{D\times D}} \left\{ \sum_{\mu=1}^{N} \left[\mathrm{Tr}\big(W(H_\mu^{\mathrm{K}})^\top\big) - y_0^\mu\right]^2 + \lambda N_0 \|W\|_F^2 \right\}, \tag{7}$$

where $\lambda \geq 0$ denotes the weight decay strength. We denote by $\mathcal{E}_M^{\mathrm{K}}$ and $\mathcal{E}_G^{\mathrm{K}}$ the memorization and generalization errors defined in (2), respectively, when the prediction is produced by $\widehat{W}_{\mathrm{K}}$.

For analytical tractability, we specialize the fixed feature map h to a finite collection of spherical harmonics. For each degree $k \geq 1$, let $h_k(x) \in \mathbb{R}^{D_{k,d}}$ denote an orthonormal basis of the degree-k spherical harmonics on $\mathbb{S}^{d-1}(\sqrt{d})$, where $D_{k,d} = \frac{2k+d-2}{k}\binom{k+d-3}{k-1}$. Let $\mathcal{K} \subset \mathbb{Z}_{\geq 1}$ be a fixed finite set of harmonic degrees. We choose $h(x) = \bigoplus_{k\in\mathcal{K}} h_k(x) \in \mathbb{R}^D$, where $D = D(d) := \sum_{k\in\mathcal{K}} D_{k,d}$ and $\bigoplus$ denotes vertical concatenation. This spherical-harmonic construction should be viewed as an analytically convenient realization of a fixed nonlinear feature map. More practical embeddings, such as random features (Hu et al., 2024; Aguirre-López et al., 2025; Defilippis et al., 2024), lead to closely related architectures. We leave the analysis of such random-feature learners to future work.

2.3 FEATURE LEARNER

We next consider the *feature learner*, which reverses the order of the nonlinear transformation and the attention module. For the μ-th training prompt, we form

$$E^\mu = \begin{bmatrix} x_1^\mu & \cdots & x_L^\mu & x_0^\mu \\ y_1^\mu & \cdots & y_L^\mu & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(L+1)}. \tag{8}$$

We apply the same single linear self-attention layer as in (4) and impose the parameterization (5), now with $W \in \mathbb{R}^{d\times d}$. The bottom-right entry of the attention output is then

$$u_W^\mu := \mathrm{Attn}(E^\mu)_{d+1,L+1} = \frac{1}{L}\sum_{\ell=1}^{L} y_\ell^\mu \, (x_0^\mu)^\top W x_\ell^\mu. \tag{9}$$

We equivalently have

$$u_W^\mu = \mathrm{Tr}\big(W(H_\mu^{\mathrm{F}})^\top\big), \text{ where } H^{\mathrm{F}} := \frac{1}{L} x_0^\mu \left(\sum_{\ell=1}^{L} y_\ell^\mu x_\ell^\mu\right)^\top \in \mathbb{R}^{d\times d}. \tag{10}$$

The final prediction is obtained by applying a scalar nonlinear readout $\widehat{y}^\mu = \sigma(u_W^\mu)$.

We train the attention module and the nonlinear readout in two stages. In the first stage, we fix a prescribed scalar function $\sigma_0 : \mathbb{R} \to \mathbb{R}$ and optimize only the attention matrix:

$$\widehat{W}_{\mathrm{F}} = \arg\min_{W \in \mathbb{R}^{d\times d}} \left\{ \sum_{\mu=1}^{N} \left[\sigma_0\big(\mathrm{Tr}\big(W(H_\mu^{\mathrm{F}})^\top\big)\big) - y_0^\mu\right]^2 + \lambda N_0 \|W\|_F^2 \right\}. \tag{11}$$

Here σ_0 is fixed throughout the first stage and is not itself trained.

In the second stage, we freeze $\widehat{W}_{\mathrm{F}}$ and optimize only the scalar readout. Let $U_{\widehat{W}_{\mathrm{F}}} = \mathrm{Tr}\Big(\widehat{W}_{\mathrm{F}}(H^{\mathrm{F}})^\top\Big)$ be the scalar representation produced by $\widehat{W}_{\mathrm{F}}$ on an independent fresh prompt of length L drawn from the same pretraining distribution, with its task vector sampled uniformly from the same task pool $\mathcal{A}_0$. We define

$$\widehat{\sigma} \in \arg\min_{\sigma} \mathbb{E}\left[\Big(\sigma(U_{\widehat{W}_{\mathrm{F}}}) - y_0\Big)^2 \,\middle|\, \widehat{W}_{\mathrm{F}}, \mathcal{A}_0\right], \tag{12}$$

where the expectation is over the fresh prompt, including its task index, inputs, and label noise. The solution of (12) is $\widehat{\sigma}(u) = \mathbb{E}\left[y_0 \,\middle|\, U_{\widehat{W}_{\mathrm{F}}} = u,\, \widehat{W}_{\mathrm{F}}, \mathcal{A}_0\right]$.

The population optimization in (12) is an idealization of training a one-dimensional multi-layer perceptron (MLP) on an independent set of fresh prompts after the attention matrix has been learned. More explicitly, in experiments we use N' additional prompts and fit

$$\widehat{\sigma}_{N'} = \arg\min_{\sigma\in\mathcal{S}} \frac{1}{N'} \sum_{\mu=1}^{N'} \left[\sigma\Big(\mathrm{Tr}\Big(\widehat{W}_{\mathrm{F}}(H_\mu^{\mathrm{F}})^\top\Big)\Big) - y_0^\mu\right]^2, \tag{13}$$

where $\mathcal{S}$ is chosen to be a one-dimensional MLP. Since this stage involves only scalar regression, we idealize its estimation error away by taking $N' \to \infty$ and $\frac{N'}{N} \to 0$. Consequently, $N + N' = N(1 + o(1))$, so the additional samples used to learn the scalar readout form a vanishing fraction of the total number of training prompts and do not change the leading-order sample complexity.

We denote by $\mathcal{E}_M^{\mathrm{F}}$ and $\mathcal{E}_G^{\mathrm{F}}$ the memorization and generalization errors defined in (2), respectively, for the learner specified by $(\widehat{W}_{\mathrm{F}}, \widehat{\sigma})$.

3 Main results

In this section we present and compare the memorization and generalization errors of the kernel learner and the feature learner. The results are derived using the replica method together with a Gaussian-equivalence ansatz, in which the local fields are replaced by Gaussian variables with matching leading-order covariances. Such Gaussian equivalence has been established rigorously for several related models (Goldt et al., 2022; Hu & Lu, 2023; Montanari & Saeed, 2022). In our ICL setting, however, we use it as an ansatz and validate the resulting predictions numerically.

Let c_k denote the Hermite coefficients of $\sigma_\star$, $\theta(a) := \bigoplus_{k\in\mathcal{K}} c_k \sqrt{\frac{D}{D_{k,d}}}\, h_k(a) \in \mathbb{R}^D$ and $\nu_{\mathrm{in}} := \sum_{k\in\mathcal{K}} c_k^2$, $\nu_{\mathrm{out}} := \sum_{k\notin\mathcal{K}} c_k^2$. Result 1 gives the memorization error and the generalization error in the limit $D \to \infty$ with $\kappa_K := N_0/D$, $\alpha_K := L/D$, $\widetilde{\alpha}_K := \widetilde{L}/D$, and $\gamma_K := N/(DN_0)$ fixed.

Result 1 (kernel learner)**.** *For $\nu_{\mathrm{in}} > 0$, let $(m_K, q_K, \overline{q}_K, \chi_K, \overline{\chi}_K, \widehat{m}_K, \widehat{q}_K, \widehat{\chi}_K)$ solve Eqs.* (39)*. The distribution of the learned attention matrix is represented by*

$$\widehat{W}_K \overset{\mathrm{d}}{=} \frac{1}{\widehat{\chi}_K}\left[\widehat{m}_K S_K + \frac{\sqrt{\nu_{\mathrm{in}}\widehat{q}_K}}{\sqrt{N_0}} Z_K \left(S_K + \frac{1+\Delta}{\alpha_K} I_D\right)^{1/2}\right]\left[S_K + \left(\frac{1+\Delta}{\alpha_K} + \frac{\lambda}{\widehat{\chi}_K}\right) I_D\right]^{-1}, \tag{14}$$

where $S_K := \frac{1}{N_0}\sum_{k=1}^{N_0} \theta(a^k)\theta(a^k)^\top$ and $(Z_K)_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. For $\mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}$, the memorization error and the generalization error can be written as

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}} \approx \nu_{\mathrm{out}} + \nu_{\mathrm{in}}(1 - 2m_{K,\mathcal{P}} + q_{K,\mathcal{P}}) + \frac{\overline{q}_K}{\widetilde{\alpha}_K}, \tag{15}$$

where $m_{K,\mathrm{M}} := m_K, q_{K,\mathrm{M}} := q_K$ *and* $m_{K,\mathrm{G}} := \frac{1}{D\nu_{\mathrm{in}}}\mathbb{E}\,\mathrm{Tr}(\widehat{W}_K S_{K,\mathrm{pop}}), q_{K,\mathrm{G}} := \frac{1}{D\nu_{\mathrm{in}}}\mathbb{E}\,\mathrm{Tr}(\widehat{W}_K S_{K,\mathrm{pop}}\widehat{W}_K^\top)$ *with* $S_{K,\mathrm{pop}} := \mathbb{E}_a[\theta(a)\theta(a)^\top]$.

Moreover, for large γ_K *we have* $\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}} = f_{\mathrm{M}}(\alpha_K, \widetilde{\alpha}_K, S_K) + O\Big(\frac{D^2}{N}\Big)$ *and* $\mathcal{E}_{\mathrm{G}}^{\mathrm{kernel}} = f_{\mathrm{G}}(\alpha_K, \widetilde{\alpha}_K, S_K, S_{K,\mathrm{pop}}) + O\Big(\frac{D^2}{N}\Big)$, *where the functions* $f_{\mathcal{P}}$ *are given in Eqs.* (110)–(111).

Result 1 is derived in Appendix A. It is based on the approximation

$$\sigma_\star\left(\frac{a^\top x}{\sqrt{d}}\right) \approx \sum_{k\in\mathcal{K}} \frac{c_k}{\sqrt{D_{k,d}}} h_k(a)^\top h_k(x) + \xi, \qquad \xi \sim \mathcal{N}(0, \nu_{\mathrm{out}}), \tag{16}$$

where the residual component is modeled as independent Gaussian noise under a Gaussian-equivalence ansatz, as detailed in Appendix A.1.

Result 2 gives the memorization and the generalization error of the feature learner with two-stage training in the limit $d \to \infty$ with $\kappa_F := N_0/d$, $\alpha_F := L/d$, $\widetilde{\alpha}_F := \widetilde{L}/d$, and $\gamma_F := N/(dN_0)$ fixed, which generalizes previous work on linear settings (Takanami et al., 2026; Lu et al., 2025).

Result 2 (feature learner)**.** *Let* $m, q, \bar{q}, \chi_Q, \chi_{\bar{Q}}$ *and their conjugates* $\widehat{m}, \widehat{\chi}_Q, \widehat{\chi}_{\bar{Q}}, \widehat{q}_Q, \widehat{q}_{\bar{Q}}$ *be the order parameters that solve the state equations in Eq.* (122)*. The distribution of the learned attention matrix is*

$$\widehat{W}_F \overset{\mathrm{d}}{=} \left(\widehat{m} S_F + \frac{1}{\sqrt{N_0}} Z(\widehat{q}_Q S_F + \widehat{q}_{\bar{Q}} I_d)^{1/2}\right)\left((\lambda + \widehat{\chi}_{\bar{Q}}) I_d + \widehat{\chi}_Q S_F\right)^{-1}, \tag{17}$$

where $S_F := \frac{1}{N_0}\sum_{k=1}^{N_0} a^k (a^k)^\top$ *and* $Z_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$*. The learned stage-II readout is*

$$\hat{\sigma}(v) = \mathbb{E}_\psi\left[\sigma_\star\left(\frac{\mu_1 m}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha_F} v + \sqrt{1 - \frac{\mu_1^2 m^2}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha_F}}\,\psi\right)\right], \tag{18}$$

where the expectation is over $\psi \sim \mathcal{N}(0,1)$*. For* $\mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}$*, the memorization error and the generalization error can be written as*

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} \approx \mathbb{E}_{g,z}\left[\left(\sigma_\star(g) - \widehat{\sigma}\left(C_{\mathcal{P}} g + \sqrt{V_{\mathcal{P}} - C_{\mathcal{P}}^2}\, z\right)\right)^2\right], \qquad g, z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1), \tag{19}$$

where we define $C_{\mathrm{M}} = \mu_1 m, V_{\mathrm{M}} = \mu_1^2 q + \frac{\bar{q}}{\widetilde{\alpha}_F}$ *and* $C_{\mathrm{G}} = \mu_1 d^{-1}\mathbb{E}_Z\,\mathrm{Tr}\,\widehat{W}_F, V_{\mathrm{G}} = \left(\mu_1^2 + \frac{1}{\widetilde{\alpha}_F}\right)\bar{q}$*.*

Moreover, for large γ_F, $\sigma_0(v) = v$ *and* $\mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}$ *we have* $\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} = g_{\mathcal{P}}(\alpha_F, \widetilde{\alpha}_F, S_F) + O\Big(\frac{d^2}{N}\Big)$, *where the function* $g_{\mathcal{P}}$ *is given in Eq.* (198)*.*

The following theorem studies the context scalings. For simplicity we choose $\sigma_0(v) = v$. While it can be derived from Results 1 and 2, we also prove it rigorously. See Appendix D.

Theorem 1 (Informal)**.** *Take* $\kappa_F, \kappa_K \to \infty$ *after* $\gamma_F, \gamma_K \to \infty$*, and set* $\Delta = \lambda = 0$ *and* $\widetilde{L} = L$*. We obtain*

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} = 1 - \sum_{k\geq 1} c_k^2 r_{\alpha_F}^k + o_{d,L}(1), \quad \mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}} = \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K}} c_k^2 \frac{D_{k,d}}{c_k^2 L + D_{k,d}} + o_{d,L}(1), \tag{20}$$

for the feature learner and the kernel learner respectively, where $r_{\alpha_F} := \frac{\alpha_F c_1^2}{1+\alpha_F c_1^2}$.

One immediate corollary is that if we choose $\widetilde{L} = L = \alpha d^s$ with fixed $s > 0$ and $d \to \infty$, we have

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} \longrightarrow \begin{cases} 1, & 0 < s < 1, \\ 1 - \sum_{k=1}^{\infty} c_k^2 r_\alpha^k, & s = 1, \\ 0, & s > 1, \end{cases} \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}, \tag{21}$$

while for the kernel learner:

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}} \longrightarrow \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K},\, k>s} c_k^2 + \mathbf{1}_{\{s\in\mathcal{K}\}} \frac{c_s^2}{1 + \alpha s! c_s^2}, \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}. \tag{22}$$

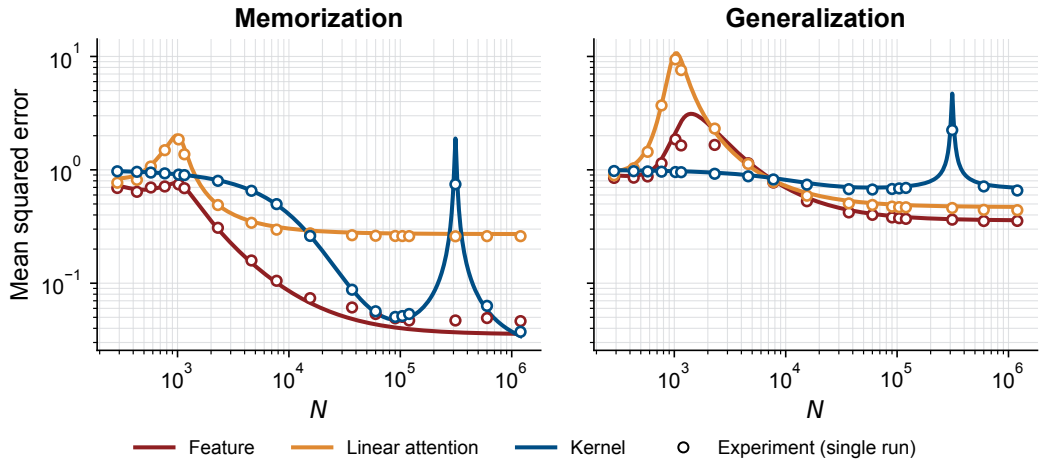

Figure 2: Memorization (left) and generalization (right) errors as a function of N. We fix $N_0 = 24$ and $L = \widetilde{L} = 1024$. Red, orange, and blue denote the feature learner, linear attention, and the kernel learner, respectively. Solid lines show replica predictions, and circles show numerical results from a single training run. The three learners use the same training and inference prompts.

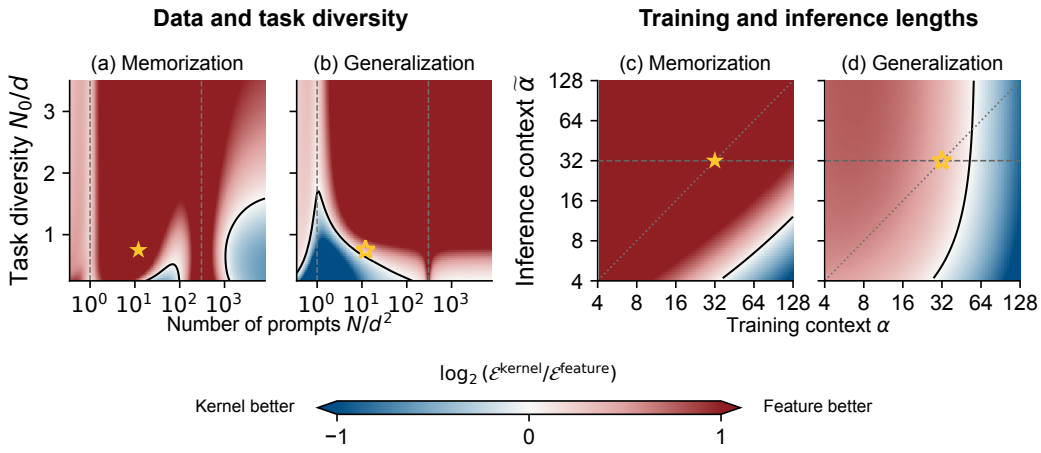

Figure 3: Replica predictions for the relative errors of the two learners. In all panels, the color shows $\log_2(\mathcal{E}^{\text{kernel}}/\mathcal{E}^{\text{feature}})$. (a,b) Memorization and generalization as N/d^2 and N_0/d vary, with $\alpha = \widetilde{\alpha} = 32$; vertical dashed lines mark $N = d^2$ and $N = D^2$. (c,d) Memorization and generalization as the training length $\alpha = L/d$ and inference length $\widetilde{\alpha} = \widetilde{L}/d$ vary, with $N_0 = 24$ and $N = 12288$; the diagonal dotted line marks equal lengths, and the horizontal dashed line marks $\widetilde{\alpha} = 32$. Gold stars mark the shared parameter setting ($N = 12288, N_0 = 24, \alpha = \widetilde{\alpha} = 32$). Blue favors the kernel learner, red favors the feature learner, and black curves indicate equal errors.

This suggests that for the kernel learner, a degree-k harmonic sector is asymptotically fully recovered when $L \gg d^k$, is only partially recovered when $L = \Theta(d^k)$, and is asymptotically unrecovered when $L \ll d^k$. Moreover, assuming $s = 1$ and $1 \in \mathcal{K}$, we obtain $\mathcal{E}_{\mathcal{P}}^{\text{kernel}} - \mathcal{E}_{\mathcal{P}}^{\text{feature}} \approx \sum_{k \geq 2} c_k^2 r_\alpha^k$. For a nonlinear teacher this gap is strictly positive for every $\alpha > 0$.

Thus there are two advantages of the feature learner. First, Results 1 and 2 suggest that the two learners operate at different pretraining sample scales. Denote $p := \max \mathcal{K}$. The feature and kernel attention matrices contain d^2 and $D^2 = \Theta(d^{2p})$ parameters, respectively, and hence converge on the scales $N \gg d^2$ and $N \gg D^2$. Thus, when $p > 1$, there is a broad regime $d^2 \ll N \ll D^2$ in which the feature learner is already close to its infinite-pretraining limit while the kernel learner is not.

Second, the two architectures use the demonstrations differently. For the kernel learner, a degree-k spherical-harmonic sector requires a context length of order d^k to be recovered. Under $L = \Theta(d)$, only the degree-one sector contributes. By contrast, the feature learner only requires $L = \Theta(d)$ demonstrations and exploits all Hermite components of the teacher with factors r_α^k simultaneously.

4 EXPERIMENTS

We compare the two learners at finite $d, D, N_0, N, L, \widetilde{L}$, using the same task pool and prompts, and evaluate the state equations at the corresponding values. Experiments use the common teacher:

$$\sigma_\star\left(\frac{a^\top x}{\sqrt{d}}\right) = \frac{\sqrt{3}}{2}\frac{a^\top x}{\sqrt{d}} + \frac{\sqrt{2}}{4}\left(\frac{(a^\top x)^2}{d} - 1\right). \tag{23}$$

For the kernel learner, we use the complete degree-one and degree-two spherical harmonic basis. Note that because the degree-one and degree-two harmonic basis contains the teacher, the kernel learner has no representation error ($\nu_{\rm out} = 0$).

Unless stated otherwise, we fix $d = 32$, $D = 559$, $\Delta = 0.01$ and $\lambda = 0.001$. We first increase N with a small fixed task pool, then vary task diversity and the training and inference lengths to compare the two learners across these settings. In this section we denote $\alpha := L/d$ and $\widetilde{\alpha} := \widetilde{L}/d$.

- **Numerics.** We use the attention block in Eq. (5). In terms of the feature learner, we use the following two-stage protocol.
  - *Stage I.* Initialize the attention block, fix the scalar readout to $\sigma_0(u) = u$ by initializing a width-64 ReLU readout (which remains frozen during Stage I), and do full-batch L-BFGS.
  - *Stage II.* Freeze the learned attention block, independently draw 1024 prompts[1] from the same task pool, and train the readout by AdamW followed by full-batch L-BFGS.

  For the kernel learner, we map each input to the fixed kernel basis $h(x)$ and optimize the attention block corresponding to Eq. (5) by full-batch L-BFGS. We also compare them with the linear attention baseline, which is equivalent to the feature learner after Stage I. All three learners use the same Stage-I and inference prompts. Each inference error averages 32768 prompts.
- **Replica curves.**
  - *Feature learner.* We solve Eq. (122), substitute the order parameters into Eq. (19) with the readout in Eq. (18).
  - *Kernel learner.* We solve Eqs. (39) and substitute their order parameters into Eq. (15).

4.1 VARYING THE NUMBER OF TRAINING PROMPTS

We first compare the numerics and the replica prediction with different numbers of training prompts. We fix $N_0 = 24$ and $\alpha = \widetilde{\alpha} = 32$, and vary N from 288 to 1200000. Increasing N allows us to compare the two learners both before and after their interpolation peaks. In Figure 2, we can observe a good agreement between the numerics and the replica prediction. In Appendix C.1 we further check the consistency of corresponding order parameters between experiments and replica prediction. The feature learner, linear attention and kernel learner all show the familiar double-descent behavior (Lu et al., 2025), with interpolation peaks near d^2 and D^2, respectively. Stage II substantially reduces the interpolation peak for the feature learner, although a smaller peak remains. Our main interest, however, is the relative performance of the two learners across the different data regimes. We highlight the main observations below.

- **The feature learner reaches low error earlier.** After its first interpolation peak, the feature learner (red line) reaches low error well before the kernel learner (blue line), since $D \gg d$. At $N = 36864$, its memorization/generalization errors are $0.061/0.421$, compared with $0.088/0.675$ for the kernel learner, which later passes through its interpolation peak near D^2. The feature learner is therefore better over $d^2 \ll N \ll D^2$.
- **The kernel learner wins on memorization at large N, but fails on generalization.** At $N = 1200000$, the kernel learner achieves memorization error 0.037, below the feature learner's 0.046. Generalization still favors the feature learner, with mean error 0.354 compared with 0.657 for the kernel learner. In fact, as $N \to \infty$, Results 1 and 2 give the limiting errors:

  $$\mathcal{E}_{\rm M}^{\rm feature} \simeq 0.03536, \quad \mathcal{E}_{\rm G}^{\rm feature} \simeq 0.35682, \quad \mathcal{E}_{\rm M}^{\rm kernel} \simeq 0.02264, \quad \mathcal{E}_{\rm G}^{\rm kernel} \simeq 0.67073.$$

  The kernel learner can thus learn the training tasks more accurately with more trainable parameters once enough prompts are available, but generalizes worse at low task diversity.

[1] This choice is for a generic MLP readout. In general, Stage II requires far fewer samples than Stage I; for example, in our case fitting a quadratic readout is enough.

4.2 Varying task diversity and context lengths

The previous comparison uses a fixed value of $N_0, \alpha, \widetilde{\alpha}$. In this section, we vary these hyperparameters to obtain a more comprehensive phase diagram (Figure 3).

Figure 3(a,b) shows that the kernel learner has a memorization advantage when the training set size N is large and the task pool N_0 is small. Increasing N_0 removes this kernel memorization advantage. At smaller N, the relative performance is further affected by the two interpolation peaks near $N \sim d^2$ and $N \sim D^2$. Hence, the kernel learner has a generalization advantage only in the low-N regime, where the feature learner is affected by its $N \sim d^2$ interpolation peak. This is consistent with the observation in Figure 2. We next vary the training and inference lengths in Figure 3(c,d).

Longer training prompts do not always improve inference performance. Figure 3(c,d) suggests that the relative advantage depends strongly on the training and inference context lengths. Two regimes can favor the kernel learner. First, under a large context-length mismatch, the feature learner can deteriorate because its readout is learned on training prompts and then kept fixed at inference, and therefore does not adapt to changes in $\widetilde{\alpha}$. Second, when both α and $\widetilde{\alpha}$ are large, the kernel learner can reconstruct the teacher with smaller fluctuations, since the kernel basis is complete.

Overall, these replica predictions show that the relative advantage of feature and kernel learners depends jointly on the amount of pretraining data N, task diversity N_0, and the training and inference context lengths $\alpha, \widetilde{\alpha}$. Additional numerical results are provided in Appendix C.2, where we compare the replica prediction with the experiments with different N_0.

4.3 Learning different harmonic degrees

Theorem 1 predicts that the two learners have different context scalings. Figure 4 illustrates this difference for the common teacher (23). The solid curves are obtained by substituting $d = 16$ and $L = 16^s$ into Eq. (20), which agree closely with the numerical experiments. The feature learner has lower error throughout the displayed range. The dashed curves evaluate the same formulas at $d = 10^6$ and show the separation of sample complexities emerge as d becomes larger. The kernel learner acquires the degree-one and degree-two components at the separate scales $L \asymp d$ and $L \asymp d^2$, whereas the feature learner recovers both components at the same time.

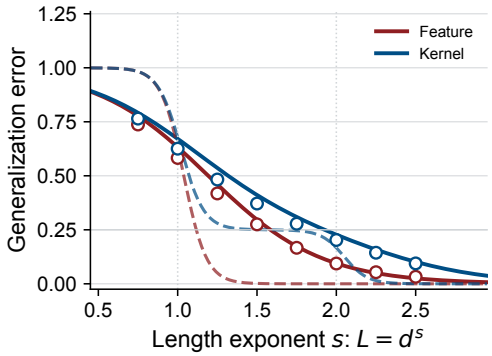

Figure 4: Learning different harmonic degrees. The horizontal coordinate is $s = \log L / \log d$, with equal training and inference lengths $L = \widetilde{L} = d^s$. The teacher is (23). Circles show experiments at $d = 16$, $N_0 = 8192$, $N = 1200000$, $\Delta = 0$, and $\lambda = 0$. Solid curves are Theorem 1 (Eq. (20)) evaluated at $d = 16$. Dashed curves are the same formulas evaluated at $d = 10^6$ and display the separation between the scales $L \asymp d$ and $L \asymp d^2$.

5 Conclusion and Future Work

In this work, we studied nonlinear ICL by comparing the kernel learner and the feature learner on the same single-index task. Using replica calculations, we characterized how their memorization and generalization errors depend jointly on pretraining size, task diversity, and the training and inference context lengths. In the population limit, we rigorously prove two context-scaling mechanisms. More broadly, our results suggest that the relative advantage of kernel and feature learning in ICL is determined by how the architectures interact with the amount of pretraining data, task diversity, and the information available in the context.

Several directions remain open. Our analysis focuses on one-layer linear attention, spherical-harmonic features, and an idealized two-stage training procedure in which the scalar readout is learned from sufficient additional data. It would be useful to understand whether the same separation persists for random features or learned embeddings, jointly trained readouts, softmax attention, and deeper Transformer architectures. On the theoretical side, important directions include proving the finite-dimensional results and analyzing ICL beyond single-index targets with $\mu_1 \neq 0$.

REFERENCES

Fabián Aguirre-López, Silvio Franz, and Mauro Pastore. Random features and polynomial rules. *SciPost Physics*, 18(1):039, 2025.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations*, volume 2026, pp. 21749–21779, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349): 1–65, 2024.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104630–104693. Curran Associates, Inc., 2024. doi: 10.52202/079017-3323. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bd18189308a4c45c7d71ca83acf3deaa-Abstract-Conference.html.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Sara Dragutinović, Andrew M Saxe, and Aaditya K Singh. Softmax $\geq$ linear: Transformers may learn to classify in-context by kernel gradient descent. *arXiv preprint arXiv:2510.10425*, 2025.

S. F. Edwards and P. W. Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5 (5):965–974, 1975. doi: 10.1088/0305-4608/5/5/017.

Gabriele Farné, Fabrizio Boncoraglio, and Lenka Zdeborová. The rules-and-facts model for simultaneous generalization and memorization in neural networks. *arXiv preprint arXiv:2603.25579*, 2026.

Christopher Frye and Costas J Efthimiou. Spherical harmonics in p dimensions. *arXiv preprint arXiv:1205.3548*, 2012.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

Chase Goddard, Lindsay M Smith, Vudtiwat Ngampruetikorn, and David J Schwab. When can in-context learning generalize out of task distribution? *arXiv preprint arXiv:2506.05574*, 2025.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022.

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.

Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

Soo Min Kwon, Alec S Xu, Can Yaras, Dogyoon Song, Laura Balzano, and Qing Qu. The effect of training task diversity on in-context learning through the lens of low-dimensional subspaces. *arXiv preprint arXiv:2606.06814*, 2026.

Mary Letey, Jacob Zavatone-Veth, Yue Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 128723–128761, 2026.

Bruno Loureiro. Four lectures on statistical physics of learning, 2024. URL `https://brloureiro.github.io/assets/pdf/NotesPrinceton_BL.pdf`. Lecture notes, Princeton Machine Learning Theory Summer School.

Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Arvind Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, volume 2024, pp. 32077–32096, 2024.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pp. 3351–3418. PMLR, 2021.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 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.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4310–4312. PMLR, 2022. URL `https://proceedings.mlr.press/v178/montanari22a.html`.

Alex Nguyen and Gautam Reddy Nallamala. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 8677–8697, 2025.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Forty-second International Conference on Machine Learning*, 2025.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365, 2024.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246, 2023.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in neural information processing systems*, 36:27801–27819, 2023.

Kaito Takanami, Takashi Takahashi, and Yoshiyuki Kabashima. Learning linear regression with low-rank tasks in-context. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4672–4680. PMLR, 2026.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, volume 2024, pp. 56587–56641, 2024.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

CONTENTS

1 Introduction **1**

2 Settings **2**

2.1 Task 2

2.2 Kernel learner 4

2.3 Feature learner 4

3 Main results **5**

4 Experiments **8**

4.1 Varying the number of training prompts 8

4.2 Varying task diversity and context lengths 9

4.3 Learning different harmonic degrees 9

5 Conclusion and Future Work **9**

A Derivation of the kernel learner **13**

A.1 Teacher expansion 13

A.2 Fixed point equations 15

A.3 Limiting spectrum 15

A.4 Reduction to Takanami et al. 16

A.5 Derivation of the main results 19

B Derivation of the feature learner **24**
B.1 Fixed point equations . . . 24
B.2 Derivation of the state equations . . . 24
B.2.1 The replica trick . . . 24
B.2.2 Teacher score and attention outputs . . . 25
B.2.3 Order parameters and replica symmetry . . . 27
B.2.4 Calculation of f_1 . . . 29
B.2.5 Calculation of f_2 . . . 30
B.2.6 Calculation of f_3 . . . 30
B.2.7 Stationarity . . . 31
B.3 Derivation of the main results . . . 32

C Additional experiments **36**
C.1 Consistency of the Replica method with numerical experiments . . . 36
C.2 Varying task diversity . . . 36
C.3 Derivation of Eq. (217) . . . 37

D A formal statement of Theorem 1 **39**

E Proof of Theorem 2 **40**
E.1 Gegenbauer polynomials . . . 40
E.2 Kernel learner . . . 42
E.3 Feature learner Stage-I . . . 43
E.4 Concentration of the context . . . 44
E.5 Feature learner Stage-II . . . 45
E.6 Completion of the proof . . . 46

F Declaration of LLM usage **47**

A Derivation of the kernel learner

A.1 Teacher expansion

We first express the nonlinear teacher in the spherical-harmonic basis used by the kernel learner. Let

$$g_d := \frac{a^\top x}{\sqrt{d}}, \qquad a, x \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})). \tag{24}$$

The random variable g_d has mean zero and variance one and converges in distribution to $G \sim \mathcal{N}(0,1)$ as $d \to \infty$.

For each $k \geq 0$, let $P_{k,d}$ denote the degree-k Gegenbauer polynomial orthonormal with respect to the distribution of g_d, namely

$$\mathbb{E}[P_{j,d}(g_d)P_{k,d}(g_d)] = \delta_{jk}. \tag{25}$$

With the spherical harmonics normalized with respect to the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$, the spherical addition theorem (Frye & Efthimiou, 2012) gives

$$P_{k,d}\left(\frac{a^\top x}{\sqrt{d}}\right) = \frac{1}{\sqrt{D_{k,d}}} h_k(a)^\top h_k(x). \tag{26}$$

Define the Gegenbauer coefficients $c_{k,d} := \mathbb{E}[\sigma_\star(g_d) P_{k,d}(g_d)]$. Then, in L^2 with respect to the distribution of g_d,

$$\sigma_\star(g_d) = \sum_{k\geq 0} c_{k,d} P_{k,d}(g_d) = \sum_{k\geq 0} \frac{c_{k,d}}{\sqrt{D_{k,d}}} h_k(a)^\top h_k(x). \tag{27}$$

Let H_k denote the normalized Hermite polynomial and define the Hermite coefficients of the teacher by

$$c_k := \mathbb{E}[\sigma_\star(G) H_k(G)]. \tag{28}$$

For every fixed k, the normalized Gegenbauer polynomial $P_{k,d}$ converges to H_k as $d \to \infty$ (Mei et al., 2021). We assume throughout that $\sigma_\star$ is sufficiently regular so that

$$c_{k,d} \longrightarrow c_k \quad \text{for every fixed } k, \qquad \mathbb{E}[\sigma_\star(g_d)^2] \longrightarrow \mathbb{E}[\sigma_\star(G)^2] = 1. \tag{29}$$

We now separate the harmonic degrees represented by the learner from the remaining component. For a fixed finite set $\mathcal{K} \subset \mathbb{N}_{\geq 1}$, define

$$\begin{aligned} f_{\mathcal{K},d}(a,x) &:= \sum_{k\in\mathcal{K}} \frac{c_{k,d}}{\sqrt{D_{k,d}}} h_k(a)^\top h_k(x), \\ r_{\mathcal{K},d}(a,x) &:= \sigma_\star\left(\frac{a^\top x}{\sqrt{d}}\right) - f_{\mathcal{K},d}(a,x). \end{aligned} \tag{30}$$

By orthogonality of the Gegenbauer polynomials, for every fixed a,

$$\mathbb{E}_x[r_{\mathcal{K},d}(a,x) h_k(x)] = 0, \qquad k \in \mathcal{K}, \tag{31}$$

and

$$\mathbb{E}_x[r_{\mathcal{K},d}(a,x)^2] = \sum_{k\notin\mathcal{K}} c_{k,d}^2. \tag{32}$$

Using Parseval's identity together with (29),

$$\sum_{k\notin\mathcal{K}} c_{k,d}^2 = \mathbb{E}[\sigma_\star(g_d)^2] - \sum_{k\in\mathcal{K}} c_{k,d}^2 \longrightarrow 1 - \sum_{k\in\mathcal{K}} c_k^2 = \sum_{k\notin\mathcal{K}} c_k^2. \tag{33}$$

Define

$$\theta(a) := \bigoplus_{k\in\mathcal{K}} c_k \sqrt{\frac{D}{D_{k,d}}}\, h_k(a). \tag{34}$$

Since $\mathcal{K}$ is finite and $c_{k,d} \to c_k$ for every $k \in \mathcal{K}$,

$$\mathbb{E}_{a,x}\left[f_{\mathcal{K},d}(a,x) - \frac{1}{\sqrt{D}}\theta(a)^\top h(x)\right]^2 = \sum_{k\in\mathcal{K}} (c_{k,d} - c_k)^2 \longrightarrow 0. \tag{35}$$

Hence

$$f_{\mathcal{K},d}(a,x) = \frac{1}{\sqrt{D}}\theta(a)^\top h(x) + o_{L^2}(1). \tag{36}$$

The residual $r_{\mathcal{K},d}(a,x)$, however, is in general neither Gaussian nor independent of $h(x)$. To obtain a tractable effective model, we make an additional Gaussian-equivalence approximation: at the level of the high-dimensional learning problem, we replace the residual by an independent Gaussian random variable with the same variance. This approximation is motivated by Gaussian-equivalence results for high-dimensional random feature models (Hu et al., 2024; Aguirre-López et al., 2025; Defilippis et al., 2024).

Writing

$$\nu_{\rm in} := \sum_{k\in\mathcal{K}} c_k^2, \qquad \nu_{\rm out} := \sum_{k\notin\mathcal{K}} c_k^2 = 1 - \nu_{\rm in}, \tag{37}$$

we therefore approximate the original teacher by the effective linear model

$$y = \frac{1}{\sqrt{D}}\theta(a)^\top h(x) + \widetilde{\epsilon}, \qquad \widetilde{\epsilon} \sim \mathcal{N}(0, \Delta + \nu_{\rm out}), \tag{38}$$

where $\widetilde{\epsilon}$ is taken to be independent of (a, x) and of the feature vector $h(x)$.

A.2 Fixed point equations

Result A.1 (kernel learner: state equations)**.** *Given $(\alpha_K, \gamma_K, \kappa_K, \nu_{\rm in}, \Delta, \lambda, S_K)$, the state equations are*

$$\begin{aligned}
\widehat{m}_K = \widehat{\chi}_K &= \frac{\gamma_K \nu_{\rm in}}{\nu_{\rm in}(1+\chi_K) + \frac{1+\Delta}{\alpha_K}\overline{\chi}_K}, \\
\widehat{q}_K &= \frac{\gamma_K \nu_{\rm in}\left[1 + \Delta - 2\nu_{\rm in} m_K + \nu_{\rm in} q_K + \frac{1+\Delta}{\alpha_K}\overline{q}_K\right]}{\left[\nu_{\rm in}(1+\chi_K) + \frac{1+\Delta}{\alpha_K}\overline{\chi}_K\right]^2}, \\
\chi_K &= \frac{1}{\widehat{\chi}_K}\left[1 - \zeta_K r_K(\zeta_K)\right], \\
\overline{\chi}_K &= \frac{\nu_{\rm in}}{\widehat{\chi}_K}\left[\frac{1-\kappa_K}{\zeta_K} + \kappa_K r_K(\zeta_K)\right], \\
m_K &= \frac{\widehat{m}_K}{\nu_{\rm in}\widehat{\chi}_K}(\nu_{\rm in} - \zeta_K + \zeta_K^2 r_K(\zeta_K)), \\
q_K &= \frac{\widehat{m}_K^2}{\nu_{\rm in}\widehat{\chi}_K^2}(\nu_{\rm in} - 2\zeta_K + 3\zeta_K^2 r_K(\zeta_K) - \zeta_K^3 t_K(\zeta_K)) \\
&\quad + \frac{\widehat{q}_K}{\widehat{\chi}_K^2}\left[1 + (\beta_K - 2\zeta_K) r_K(\zeta_K) + (\zeta_K^2 - \beta_K\zeta_K) t_K(\zeta_K)\right], \\
\overline{q}_K &= \frac{\kappa_K \widehat{m}_K^2}{\widehat{\chi}_K^2}\left[1 - 2\zeta_K r_K(\zeta_K) + \zeta_K^2 t_K(\zeta_K)\right] \\
&\quad + \frac{\nu_{\rm in}\widehat{q}_K}{\widehat{\chi}_K^2}\left[(1-\kappa_K)\frac{\beta_K}{\zeta_K^2} + \kappa_K\left\{r_K(\zeta_K) + (\beta_K - \zeta_K) t_K(\zeta_K)\right\}\right],
\end{aligned} \tag{39}$$

where $r_K(\zeta) := \lim_{N_0\to\infty} \frac{1}{N_0}\operatorname{Tr}(G_K + \zeta I_{N_0})^{-1}$ and $t_K(\zeta) := \lim_{N_0\to\infty} \frac{1}{N_0}\operatorname{Tr}(G_K + \zeta I_{N_0})^{-2}$. Here we define $G_K := \frac{1}{D}\Theta_K^\top\Theta_K \in \mathbb{R}^{N_0\times N_0}$ with $\Theta_K := \begin{bmatrix}\theta(a^1) & \cdots & \theta(a^{N_0})\end{bmatrix} \in \mathbb{R}^{D\times N_0}$. We also define $\beta_K := (1+\Delta)\frac{\kappa_K}{\alpha_K}$ and $\zeta_K := \beta_K + \frac{\kappa_K\lambda}{\widehat{\chi}_K}$. These equations determine five order parameters $(m_K, q_K, \overline{q}_K, \chi_K, \overline{\chi}_K)$ and three conjugates $(\widehat{m}_K, \widehat{q}_K, \widehat{\chi}_K)$.

A.3 Limiting spectrum

Result A.1 does not specialize to a particular spectrum of G_K. Meanwhile, we can derive the analytical limiting spectrum of G_K to simplify the equations. In Section 4, we use a different scaling $\frac{N_0}{d} \longrightarrow \varrho \in (0, \infty)$ for our comparison with the feature learner. Next we derive the limiting resolvents $r_K(\zeta)$ and $t_K(\zeta)$ in this scaling. Let

$$\Theta_K := \begin{bmatrix}\theta(a^1) & \cdots & \theta(a^{N_0})\end{bmatrix} \in \mathbb{R}^{D\times N_0}, \qquad S_K = \frac{1}{N_0}\Theta_K\Theta_K^\top, \tag{40}$$

and introduce the $N_0 \times N_0$ task Gram matrix $G_K := \frac{1}{D}\Theta_K^\top\Theta_K$. We retain the notation $\kappa_K := \frac{N_0}{D}$. By the definition of $\theta(a)$,

$$G_K = \sum_{k\in\mathcal{K}} c_k^2 \frac{1}{D_{k,d}} H_k^\top H_k, \qquad H_k := \begin{bmatrix}h_k(a^1) & \cdots & h_k(a^{N_0})\end{bmatrix}. \tag{41}$$

In the regime $N_0/d \to \varrho$, the degree-one block has a nontrivial Marchenko–Pastur spectrum, whereas every degree $k \geq 2$ satisfies $N_0/D_{k,d} \to 0$ and contributes asymptotically as an identity matrix. More precisely, the polynomial-regime spectral decomposition of Misiakiewicz (2022) gives

$$G_K = c_{1,\mathcal{K}}^2 \frac{1}{d} H_1^\top H_1 + \nu_{>1,\mathcal{K}} I_{N_0} + o_{\mathrm{op}}(1), \tag{42}$$

where

$$c_{1,\mathcal{K}}^2 := c_1^2 \mathbf{1}_{\{1 \in \mathcal{K}\}}, \qquad \nu_{>1,\mathcal{K}} := \nu_{\mathrm{in}} - c_{1,\mathcal{K}}^2. \tag{43}$$

Here $H_1 = [a^1, \ldots, a^{N_0}]$. Let ν_κ denote the Marchenko–Pastur distribution

$$\nu_\kappa(\mathrm{d}s) = \big(1 - \min\{1,\kappa\}\big)\delta_0(\mathrm{d}s) + \mathbf{1}_{[s_-,s_+]}(s)\frac{\kappa\sqrt{(s_+ - s)(s - s_-)}}{2\pi s}\,\mathrm{d}s, \qquad s_\pm = (1 \pm \kappa^{-1/2})^2. \tag{44}$$

The empirical spectral distribution of $d^{-1}H_1^\top H_1$ converges to $\nu_{1/\varrho}$. Hence the empirical spectral distribution of G_K converges to the law of

$$g_\varrho = \nu_{>1,\mathcal{K}} + c_{1,\mathcal{K}}^2 s, \qquad s \sim \nu_{1/\varrho}. \tag{45}$$

In particular, if $1 \notin \mathcal{K}$, the limiting distribution reduces to the point mass at ν_{in}. The deterministic limits of the resolvents are therefore

$$\mathcal{R}_\varrho(\zeta) := \int \frac{\nu_{1/\varrho}(\mathrm{d}s)}{\nu_{>1,\mathcal{K}} + c_{1,\mathcal{K}}^2 s + \zeta}, \tag{46}$$

and

$$\mathcal{T}_\varrho(\zeta) := \int \frac{\nu_{1/\varrho}(\mathrm{d}s)}{\Big(\nu_{>1,\mathcal{K}} + c_{1,\mathcal{K}}^2 s + \zeta\Big)^2}. \tag{47}$$

For every fixed $\zeta > 0$,

$$r_K(\zeta) \longrightarrow \mathcal{R}_\varrho(\zeta), \qquad t_K(\zeta) \longrightarrow \mathcal{T}_\varrho(\zeta). \tag{48}$$

Substituting $r_K(\zeta_K) \simeq \mathcal{R}_\varrho(\zeta_K)$ and $t_K(\zeta_K) \simeq \mathcal{T}_\varrho(\zeta_K)$ into (39) gives the state equations used in our numerical experiments.

A.4 Reduction to Takanami et al.

We now relate the kernel learner to the linear in-context learning model analyzed by Takanami et al. (2026). The correspondence consists of an exact algebraic equivalence, followed by a Gaussian-equivalence approximation for the effective statistics entering the replica calculation.

We first remove the normalization $\nu_{\mathrm{in}} \neq 1$. Recall from Eq. (38) that the effective teacher obtained in the previous subsection is

$$y = \frac{1}{\sqrt{D}}\theta(a)^\top h(x) + \widetilde{\epsilon}, \qquad \widetilde{\epsilon} \sim \mathcal{N}(0, \Delta + \nu_{\mathrm{out}}). \tag{49}$$

For $\nu_{\mathrm{in}} > 0$, define

$$\theta'(a) := \frac{\theta(a)}{\sqrt{\nu_{\mathrm{in}}}}, \qquad y' := \frac{y}{\sqrt{\nu_{\mathrm{in}}}}, \qquad \Delta' := \frac{\Delta + \nu_{\mathrm{out}}}{\nu_{\mathrm{in}}}, \qquad S_K' := \frac{S_K}{\nu_{\mathrm{in}}}, \qquad \lambda' := \frac{\lambda}{\nu_{\mathrm{in}}}. \tag{50}$$

Since H^K is linear in the labels, the rescaled labels satisfy $(H^K)' = H^K/\sqrt{\nu_{\mathrm{in}}}$, and hence

$$\sum_{\mu=1}^{N} \big[\mathrm{Tr}\big(W(H_\mu^K)^\top\big) - y_0^\mu\big]^2 + \lambda N_0 \|W\|_F^2 = \nu_{\mathrm{in}} \left\{ \sum_{\mu=1}^{N} \big[\mathrm{Tr}\big(W((H_\mu^K)')^\top\big) - (y_0^\mu)'\big]^2 + \lambda' N_0 \|W\|_F^2 \right\}. \tag{51}$$

Therefore the minimizing matrix W is unchanged. It is thus sufficient to derive the result in the normalized setting

$$\nu_{\mathrm{in}} = 1, \qquad \nu_{\mathrm{out}} = 0, \qquad \|\theta(a)\|^2 = D, \tag{52}$$

with the understanding that the replacements in Eq. (50) are restored at the end. The last identity follows from $\|h_k(a)\|^2 = D_{k,d}$ and the definition of $\theta(a)$.

Correspondence of the linearized variables. In the notation of Takanami et al. (2026), define

$$x^{\mathrm{Tak}} := \frac{h(x)}{\sqrt{D}}, \qquad w^{\mathrm{Tak}} := \theta(a), \qquad \sigma^2_{\mathrm{Tak}} := \Delta. \tag{53}$$

Then

$$\mathbb{E}[x^{\mathrm{Tak}}] = 0, \qquad \mathbb{E}[x^{\mathrm{Tak}}(x^{\mathrm{Tak}})^\top] = \frac{1}{D} I_D, \tag{54}$$

and the effective teacher takes the same linear form

$$y = (w^{\mathrm{Tak}})^\top x^{\mathrm{Tak}} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \Delta). \tag{55}$$

Moreover, the effective feature matrix used by the attention model is exactly

$$H^{\mathrm{Tak}} = \frac{D}{L}\, x_0^{\mathrm{Tak}} \sum_{\ell=1}^{L} y_\ell (x_\ell^{\mathrm{Tak}})^\top = \frac{1}{L} h(x_0) \sum_{\ell=1}^{L} y_\ell h(x_\ell)^\top = H^K. \tag{56}$$

Thus the correspondence between the task matrices is also exact:

$$S^{\mathrm{Tak}} = \frac{1}{N_0} \sum_{k=1}^{N_0} w^k (w^k)^\top = \frac{1}{N_0} \sum_{k=1}^{N_0} \theta(a^k)\theta(a^k)^\top = S_K. \tag{57}$$

For replicas W^a, W^b, the corresponding macroscopic overlaps are

$$m_K^a := \frac{1}{D} \operatorname{Tr}(W^a S_K), \qquad Q_K^{ab} := \frac{1}{D} \operatorname{Tr}(W^a S_K (W^b)^\top), \qquad \overline{Q}_K^{ab} := \frac{1}{D} \operatorname{Tr}(W^a (W^b)^\top), \tag{58}$$

which coincide with the order parameters used in the linear model of Takanami et al. (2026).

Gaussian equivalence for the effective statistic. The correspondence above does not imply that $x^{\mathrm{Tak}} = h(x)/\sqrt{D}$ is Gaussian. We therefore require one additional approximation before transferring the replica calculation.

For a fixed task a, the first moments needed below are:

$$\mathbb{E}[y_0 \mid a] = 0, \qquad \mathbb{E}[y_0^2 \mid a] = 1 + \Delta, \tag{59}$$

and

$$\mathbb{E}[y_0 H_{ij}^K \mid a] = \mathbb{E}[y_0 h_i(x_0) \mid a]\, \mathbb{E}[y h_j(x) \mid a] = \frac{\theta_i(a)\theta_j(a)}{D}. \tag{60}$$

For the second moment of H^K, the calculation gives

$$\mathbb{E}[H_{ij}^K H_{i'j'}^K \mid a] = \delta_{ii'} \left[\frac{L-1}{LD} \theta_j(a)\theta_{j'}(a) + \frac{1}{L} B_{jj'}(a) \right], \tag{61}$$

where

$$B(a) := \mathbb{E}_x[y^2 h(x) h(x)^\top \mid a]. \tag{62}$$

By $\|h(x)\|^2 = D$ for every x, we have

$$\frac{1}{D} \operatorname{Tr} B(a) = \mathbb{E}_x[y^2 \mid a] = 1 + \Delta. \tag{63}$$

Moreover, rotational invariance gives

$$\mathbb{E}_a[B(a)] = (1 + \Delta) I_D. \tag{64}$$

Conditioned on a, however, $B(a)$ is not asymptotically isotropic. Nevertheless, for the following we replace $B(a)$ with its expectation (i.e., a self-averaging ansatz) and write

$$\mathbb{E}[H_{ij}^K H_{i'j'}^K \mid a] \simeq \frac{\delta_{ii'}}{D} \left(\theta_j(a)\theta_{j'}(a) + \frac{D(1+\Delta)}{L} \delta_{jj'} \right). \tag{65}$$

For the following, we invoke the Gaussian-equivalence step used in Takanami et al. (2026): the effective learning problem is characterized through the leading first- and second-order statistics,

while higher-order correlations are neglected in the high-dimensional limit. To make the connection explicit, define the task-dependent quadratic teacher matrix

$$W_{\text{teacher}}(a) := \frac{\theta(a)\theta(a)^\top}{D} \tag{66}$$

and the surrogate query label

$$\overline{y}_0 := \mathrm{Tr}\big(W_{\text{teacher}}(a)(H^K)^\top\big) + \epsilon_0, \qquad \epsilon_0 \sim \mathcal{N}(0, \Delta). \tag{67}$$

Under Eq. (65), $(\overline{y}_0, H^K)$ and (y_0, H^K) have the same leading first- and second-order statistics. Hence, following the Gaussian-equivalence argument of Takanami et al. (2026), we replace the original training problem by the corresponding many-teacher regression problem.

More specifically, for replicas $\{W^a\}_{a=1}^n$, introduce the scalar fields

$$u^a := \mathrm{Tr}\big[(W_{\text{teacher}}(a) - W^a)\,(H^K)^\top\big] + \epsilon_0. \tag{68}$$

Under the Gaussian-equivalence approximation, the vector $(u^1, \dots, u^n)$ is replaced by a centered Gaussian vector with covariance determined only by the overlaps in Eq. (58). After averaging over the task pool, its covariance is

$$\mathcal{M}_K^{ab} = Q_K^{ab} - m_K^a - m_K^b + 1 + \Delta + \frac{1+\Delta}{\alpha_K}\overline{Q}_K^{ab}, \qquad \alpha_K := \frac{L}{D}. \tag{69}$$

This is precisely the covariance entering the data-averaged replicated partition function of Takanami et al. (2026).

Consequently, under the same replica-symmetric ansatz, the remainder of the replica calculation has the same form as in the linear model, with the identifications

$$M_0 \leftrightarrow N_0, \qquad M \leftrightarrow N, \qquad \alpha \leftrightarrow \alpha_K, \qquad \kappa \leftrightarrow \kappa_K, \qquad \gamma \leftrightarrow \gamma_K, \qquad S \leftrightarrow S_K, \qquad \sigma^2 \leftrightarrow \Delta. \tag{70}$$

Restoring the original normalization with the replacements

$$S_K \mapsto \frac{S_K}{\nu_{\text{in}}}, \qquad \Delta \mapsto \frac{\Delta + \nu_{\text{out}}}{\nu_{\text{in}}}, \qquad \lambda \mapsto \frac{\lambda}{\nu_{\text{in}}}, \tag{71}$$

this yields the saddle-point equations

$$\begin{aligned}
\widehat{m}_K = \widehat{\chi}_K &= \frac{\gamma_K \nu_{\text{in}}}{\nu_{\text{in}}(1+\chi_K) + \frac{1+\Delta}{\alpha_K}\overline{\chi}_K}, \\
\widehat{q}_K &= \frac{\gamma_K \nu_{\text{in}}\left[1 + \Delta - 2\nu_{\text{in}} m_K + \nu_{\text{in}} q_K + \frac{1+\Delta}{\alpha_K}\overline{q}_K\right]}{\left[\nu_{\text{in}}(1+\chi_K) + \frac{1+\Delta}{\alpha_K}\overline{\chi}_K\right]^2}, \\
\chi_K &= \frac{1}{\kappa_K \widehat{\chi}_K D}\,\mathrm{Tr}(S_K A_K^{-1}), \\
\overline{\chi}_K &= \frac{\nu_{\text{in}}}{\kappa_K \widehat{\chi}_K D}\,\mathrm{Tr}(A_K^{-1}), \\
m_K &= \frac{\widehat{m}_K}{\nu_{\text{in}}\widehat{\chi}_K D}\,\mathrm{Tr}(S_K^2 A_K^{-1}), \\
q_K &= \frac{\widehat{m}_K^2}{\nu_{\text{in}}\widehat{\chi}_K^2 D}\,\mathrm{Tr}(S_K^3 A_K^{-2}) + \frac{\widehat{q}_K}{\kappa_K \widehat{\chi}_K^2 D}\,\mathrm{Tr}\left[S_K\left(S_K + \frac{1+\Delta}{\alpha_K} I_D\right) A_K^{-2}\right], \\
\overline{q}_K &= \frac{\widehat{m}_K^2}{\widehat{\chi}_K^2 D}\,\mathrm{Tr}(S_K^2 A_K^{-2}) + \frac{\nu_{\text{in}}\widehat{q}_K}{\kappa_K \widehat{\chi}_K^2 D}\,\mathrm{Tr}\left[\left(S_K + \frac{1+\Delta}{\alpha_K} I_D\right) A_K^{-2}\right],
\end{aligned} \tag{72}$$

where

$$A_K := S_K + \left(\frac{1+\Delta}{\alpha_K} + \frac{\lambda}{\widehat{\chi}_K}\right) I_D. \tag{73}$$

We finally obtain (39) using

$$\begin{aligned}
\frac{1}{D}\operatorname{Tr}(A_K^{-1}) &= \kappa_K\left[\frac{1-\kappa_K}{\zeta_K}+\kappa_K r_K(\zeta_K)\right],\\
\frac{1}{D}\operatorname{Tr}(S_K A_K^{-1}) &= \kappa_K\left[1-\zeta_K r_K(\zeta_K)\right],\\
\frac{1}{D}\operatorname{Tr}(S_K^2 A_K^{-1}) &= \nu_{\text{in}}-\zeta_K+\zeta_K^2 r_K(\zeta_K),\\
\frac{1}{D}\operatorname{Tr}(S_K^3 A_K^{-2}) &= \nu_{\text{in}}-2\zeta_K+3\zeta_K^2 r_K(\zeta_K)-\zeta_K^3 t_K(\zeta_K),\\
\frac{1}{D}\operatorname{Tr}(S_K^2 A_K^{-2}) &= \kappa_K\left[1-2\zeta_K r_K(\zeta_K)+\zeta_K^2 t_K(\zeta_K)\right],\\
\frac{1}{D}\operatorname{Tr}\big[S_K(S_K+b_K I_D)A_K^{-2}\big] &= \kappa_K\left[1+(\beta_K-2\zeta_K)r_K(\zeta_K)+(\zeta_K^2-\beta_K\zeta_K)t_K(\zeta_K)\right],\\
\frac{1}{D}\operatorname{Tr}\big[(S_K+b_K I_D)A_K^{-2}\big] &= \kappa_K\left[(1-\kappa_K)\frac{\beta_K}{\zeta_K^2}+\kappa_K\left\{r_K(\zeta_K)+(\beta_K-\zeta_K)t_K(\zeta_K)\right\}\right],
\end{aligned} \tag{74}$$

where we also used the identity

$$\frac{1}{N_0}\operatorname{Tr} G_K = \frac{1}{DN_0}\operatorname{Tr}(\Theta_K^\top\Theta_K) = \nu_{\text{in}}. \tag{75}$$

We emphasize that the reduction in this subsection concerns the pretraining problem and its saddle-point equations. The inference errors require a separate calculation because the original test labels are noiseless and there is an additional irreducible term; this is carried out in the following subsection.

A.5 Derivation of the main results

We now derive the distribution of the learned attention matrix and the memorization and generalization errors stated in Result 1.

Recall that

$$\nu_{\text{in}} := \sum_{k\in\mathcal{K}} c_k^2, \qquad \nu_{\text{out}} := \sum_{k\notin\mathcal{K}} c_k^2, \qquad \nu_{\text{in}}+\nu_{\text{out}} = 1, \tag{76}$$

and

$$S_K := \frac{1}{N_0}\sum_{k=1}^{N_0}\theta(a^k)\theta(a^k)^\top. \tag{77}$$

The reduction in the previous subsection maps the kernel learner to the normalized linear problem with

$$S_K' = \frac{S_K}{\nu_{\text{in}}}, \qquad \Delta' = \frac{\Delta+\nu_{\text{out}}}{\nu_{\text{in}}}, \qquad \lambda' = \frac{\lambda}{\nu_{\text{in}}}. \tag{78}$$

Applying (Takanami et al., 2026, Result E.1) to the normalized problem gives

$$\widehat{W}_K \overset{\text{d}}{=} \frac{1}{\widehat{\chi}_K}\left[\widehat{m}_K S_K' + \sqrt{\frac{\widehat{q}_K}{N_0}}\,Z_K\left(S_K'+\frac{1+\Delta'}{\alpha_K}I_D\right)^{1/2}\right]\left[S_K'+\left(\frac{1+\Delta'}{\alpha_K}+\frac{\lambda'}{\widehat{\chi}_K}\right)I_D\right]^{-1}, \tag{79}$$

where $(Z_K)_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Substituting Eq. (78) into Eq. (79) yields

$$\boxed{\widehat{W}_K \overset{\text{d}}{=} \frac{1}{\widehat{\chi}_K}\left[\widehat{m}_K S_K + \sqrt{\frac{\nu_{\text{in}}\widehat{q}_K}{N_0}}\,Z_K\left(S_K+\frac{1+\Delta}{\alpha_K}I_D\right)^{1/2}\right]\left[S_K+\left(\frac{1+\Delta}{\alpha_K}+\frac{\lambda}{\widehat{\chi}_K}\right)I_D\right]^{-1}}. \tag{80}$$

This gives Eq. (14).

We next derive the test errors. Define

$$S_{K,\mathrm{pop}} := \mathbb{E}_a[\theta(a)\theta(a)^\top] = \bigoplus_{k\in\mathcal{K}} c_k^2 \frac{D}{D_{k,d}} I_{D_{k,d}}, \tag{81}$$

where orthogonality of spherical harmonics across different degrees was used. In particular,

$$\frac{1}{D}\operatorname{Tr} S_{K,\mathrm{pop}} = \nu_{\mathrm{in}}. \tag{82}$$

For notational convenience, define

$$S_{K,\mathrm{M}} := S_K, \qquad S_{K,\mathrm{G}} := S_{K,\mathrm{pop}}, \tag{83}$$

We use the test-error parameters in Eq. (15), together with

$$\overline{q}_K := \frac{1}{D}\mathbb{E}\operatorname{Tr}(\widehat{W}_K \widehat{W}_K^\top). \tag{84}$$

By construction,

$$m_{K,\mathrm{M}} = m_K, \qquad q_{K,\mathrm{M}} = q_K, \tag{85}$$

where m_K and q_K are the order parameters of the normalized problem in Eq. (39).

At inference time the original labels are noiseless. Under the same Gaussian-equivalence approximation introduced in Section A.1, we write the test labels as

$$y = \frac{1}{\sqrt{D}}\theta(a)^\top h(x) + \xi, \qquad \xi \sim \mathcal{N}(0, \nu_{\mathrm{out}}), \tag{86}$$

where ξ is independent of $h(x)$. Thus

$$\mathbb{E}[y^2 \mid a] = 1, \qquad \mathbb{E}[yh(x) \mid a] = \frac{\theta(a)}{\sqrt{D}}. \tag{87}$$

Let

$$H_{\mathrm{te}}^K = \frac{1}{\widetilde{L}} h(x_0) \left(\sum_{\ell=1}^{\widetilde{L}} y_\ell h(x_\ell) \right)^\top, \qquad \widehat{y}_0 = \operatorname{Tr}\left(\widehat{W}_K (H_{\mathrm{te}}^K)^\top \right). \tag{88}$$

For a fixed test task a,

$$\mathbb{E}[y_0 (H_{\mathrm{te}}^K)_{ij} \mid a] = \frac{\theta_i(a)\theta_j(a)}{D}. \tag{89}$$

Therefore,

$$\mathbb{E}[y_0 \widehat{y}_0] = \nu_{\mathrm{in}} m_{K,\mathcal{P}}, \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}. \tag{90}$$

For the second moment, define

$$B_{\mathrm{te}}(a) := \mathbb{E}_x[y^2 h(x) h(x)^\top \mid a]. \tag{91}$$

A direct calculation gives

$$\mathbb{E}[(H_{\mathrm{te}}^K)_{ij}(H_{\mathrm{te}}^K)_{i'j'} \mid a] = \delta_{ii'} \left[\frac{\widetilde{L}-1}{\widetilde{L}D} \theta_j(a)\theta_{j'}(a) + \frac{1}{\widetilde{L}} (B_{\mathrm{te}}(a))_{jj'} \right]. \tag{92}$$

As in Appendix A.4, we have $\mathbb{E}_a[B_{\mathrm{te}}(a)] = I_D$. For memorization, we analogously replace the empirical task average by the population quantity. Hence, using Eq. (92), we obtain

$$\begin{aligned} \mathbb{E}[\widehat{y}_0^2] &= \frac{\widetilde{L}-1}{\widetilde{L}} \nu_{\mathrm{in}} q_{K,\mathcal{P}} + \frac{1}{\widetilde{L}} \mathbb{E}\operatorname{Tr}(\widehat{W}_K \widehat{W}_K^\top) + o(1) \\ &= \nu_{\mathrm{in}} q_{K,\mathcal{P}} + \frac{\overline{q}_K}{\widetilde{\alpha}_K} + o(1), \end{aligned} \tag{93}$$

where we used $\widetilde{L} = \widetilde{\alpha}_K D$ in the second equality.

Finally, since the test labels are noiseless and normalized so that $\mathbb{E}[y_0^2] = 1 = \nu_{\text{in}} + \nu_{\text{out}}$,

$$\begin{aligned}\mathcal{E}_{\mathcal{P}}^{\text{kernel}} &= \mathbb{E}[(y_0 - \widehat{y}_0)^2] \\ &= 1 - 2\nu_{\text{in}} m_{K,\mathcal{P}} + \nu_{\text{in}} q_{K,\mathcal{P}} + \frac{\overline{q}_K}{\widetilde{\alpha}_K} + o(1) \\ &= \boxed{\nu_{\text{out}} + \nu_{\text{in}}\left(1 - 2m_{K,\mathcal{P}} + q_{K,\mathcal{P}}\right) + \frac{\overline{q}_K}{\widetilde{\alpha}_K}} + o(1), \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}.\end{aligned} \tag{94}$$

Dropping the asymptotically vanishing $o(1)$ term gives Eq. (15).

Large-pretraining limit. Now we study the regime in which $\alpha_K = L/D$, $\widetilde{\alpha}_K = \widetilde{L}/D$, and $\kappa_K = N_0/D$ remain fixed, while $\gamma_K = N/(DN_0) \to \infty$. Let

$$b_K := \frac{1+\Delta}{\alpha_K}, \qquad \delta_K := \frac{\lambda}{\widehat{\chi}_K}, \qquad A_{K,\gamma} := S_K + (b_K + \delta_K) I_D. \tag{95}$$

Since $\widehat{m}_K = \widehat{\chi}_K$, the first two conjugate equations in Result A.1 give

$$\frac{\nu_{\text{in}}\widehat{q}_K}{N_0 \widehat{\chi}_K^2} = \frac{D}{N}\left(1 + \Delta - 2\nu_{\text{in}} m_K + \nu_{\text{in}} q_K + \frac{1+\Delta}{\alpha_K}\overline{q}_K\right). \tag{96}$$

Define

$$\Xi_K := 1 + \Delta - 2\nu_{\text{in}} m_K + \nu_{\text{in}} q_K + \frac{1+\Delta}{\alpha_K}\overline{q}_K. \tag{97}$$

Then Eq. (80) can be written as

$$\widehat{W}_K \overset{\text{d}}{=} S_K A_{K,\gamma}^{-1} + \sqrt{\frac{D}{N}\Xi_K}\, Z_K (S_K + b_K I_D)^{1/2} A_{K,\gamma}^{-1}. \tag{98}$$

We first control the ridge correction δ_K. Using the equations for χ_K and $\overline{\chi}_K$ in Eq. (39), the equation for $\widehat{\chi}_K$ becomes

$$\widehat{\chi}_K + \frac{1}{\kappa_K D}\operatorname{Tr}\left[(S_K + b_K I_D) A_{K,\gamma}^{-1}\right] = \gamma_K. \tag{99}$$

For $\lambda \geq 0$,

$$0 \leq (S_K + b_K I_D) A_{K,\gamma}^{-1} \preceq I_D, \tag{100}$$

and hence

$$\gamma_K - \frac{1}{\kappa_K} \leq \widehat{\chi}_K \leq \gamma_K. \tag{101}$$

Therefore

$$\widehat{\chi}_K = \gamma_K + O(1), \qquad \delta_K = O(\gamma_K^{-1}) = O\left(\frac{D^2}{N}\right), \tag{102}$$

where the last relation uses $N = \gamma_K \kappa_K D^2$ with fixed κ_K.

It follows that the deterministic part of Eq. (98) converges to

$$W_{K,\infty} := S_K (S_K + b_K I_D)^{-1}. \tag{103}$$

Indeed, since $b_K > 0$,

$$\left\| S_K A_{K,\gamma}^{-1} - W_{K,\infty} \right\|_{\text{op}} = O(\delta_K) = O\left(\frac{D^2}{N}\right). \tag{104}$$

The second term in Eq. (98) is a zero-mean fluctuation conditioned on S_K. Its normalized squared Frobenius norm is

$$\frac{1}{D}\mathbb{E}_{Z_K}\left\| \widehat{W}_K - S_K A_{K,\gamma}^{-1} \right\|_F^2 = \frac{D}{N}\Xi_K \operatorname{Tr}\left[(S_K + b_K I_D) A_{K,\gamma}^{-2}\right]. \tag{105}$$

For fixed $\alpha_K > 0$, $\frac{1}{D}\operatorname{Tr}\left[(S_K + b_K I_D) A_{K,\gamma}^{-2}\right] = O(1)$,, and $\Xi_K = O(1)$ in the large-pretraining regime. Consequently,

$$\frac{1}{D}\mathbb{E}_{Z_K}\left\| \widehat{W}_K - S_K A_{K,\gamma}^{-1} \right\|_F^2 = O\left(\frac{D^2}{N}\right). \tag{106}$$

We next evaluate the limiting test errors by substituting $W_{K,\infty}$ into Eq. (94). Define

$$A_{K,\infty} := S_K + b_K I_D. \tag{107}$$

For memorization,

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \nu_{\mathrm{out}} + \nu_{\mathrm{in}} - \frac{2}{D}\operatorname{Tr}(S_K^2 A_{K,\infty}^{-1}) + \frac{1}{D}\operatorname{Tr}(S_K^3 A_{K,\infty}^{-2}) + \frac{1}{\widetilde{\alpha}_K D}\operatorname{Tr}(S_K^2 A_{K,\infty}^{-2}). \tag{108}$$

Using $D^{-1}\operatorname{Tr} S_K = \nu_{\mathrm{in}}$ and $S_K = A_{K,\infty} - b_K I_D$, the first three terms reduce to

$$\nu_{\mathrm{in}} - \frac{2}{D}\operatorname{Tr}(S_K^2 A_{K,\infty}^{-1}) + \frac{1}{D}\operatorname{Tr}(S_K^3 A_{K,\infty}^{-2}) = \frac{b_K^2}{D}\operatorname{Tr}(S_K A_{K,\infty}^{-2}). \tag{109}$$

Therefore

$$\boxed{\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \nu_{\mathrm{out}} + \frac{(1+\Delta)^2}{\alpha_K^2 D}\operatorname{Tr}(S_K A_{K,\infty}^{-2}) + \frac{1}{\widetilde{\alpha}_K D}\operatorname{Tr}(S_K^2 A_{K,\infty}^{-2}).} \tag{110}$$

Similarly, using $D^{-1}\operatorname{Tr} S_{K,\mathrm{pop}} = \nu_{\mathrm{in}}$, the generalization error is

$$\boxed{\mathcal{E}_{\mathrm{G}}^{\mathrm{kernel}}(\infty) = \nu_{\mathrm{out}} + \frac{(1+\Delta)^2}{\alpha_K^2 D}\operatorname{Tr}(S_{K,\mathrm{pop}} A_{K,\infty}^{-2}) + \frac{1}{\widetilde{\alpha}_K D}\operatorname{Tr}(S_K^2 A_{K,\infty}^{-2}).} \tag{111}$$

The deterministic correction caused by $\delta_K = \lambda/\widehat{\chi}_K$ is $O(D^2/N)$, while terms linear in the Gaussian fluctuation vanish after averaging over Z_K and its quadratic contribution is controlled by Eq. (106). Hence

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}} = \mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}}(\infty) + O\!\left(\frac{D^2}{N}\right), \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}.} \tag{112}$$

Infinite task-pool limit and context-length scalings. We next set $\Delta = \lambda = 0$ and $\widetilde{L} = L$, and take first $\gamma_K \to \infty$ and then $\kappa_K \to \infty$. In this ordered limit,

$$S_K \longrightarrow S_{K,\mathrm{pop}} = \mathbb{E}_a[\theta(a)\theta(a)^\top] = \bigoplus_{k\in\mathcal{K}} c_k^2 \frac{D}{D_{k,d}} I_{D_{k,d}}. \tag{113}$$

Consequently, the memorization and generalization errors coincide. Denoting $S := S_{K,\mathrm{pop}}$ and

$$A := S + \frac{1}{\alpha_K} I_D, \tag{114}$$

Eqs. (110) and (111) give

$$\begin{aligned} \mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}}(\infty) &= \nu_{\mathrm{out}} + \frac{1}{\alpha_K^2 D}\operatorname{Tr}(SA^{-2}) + \frac{1}{\alpha_K D}\operatorname{Tr}(S^2 A^{-2}) \\ &= \nu_{\mathrm{out}} + \frac{1}{\alpha_K D}\operatorname{Tr}\left[S\left(S + \frac{1}{\alpha_K} I_D\right)^{-1}\right], \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}. \end{aligned} \tag{115}$$

In the second equality we used $\frac{1}{\alpha_K} S + S^2 = S\left(S + \frac{1}{\alpha_K} I_D\right)$.

Since the degree-k block of S has eigenvalue $c_k^2 \frac{D}{D_{k,d}}$ with multiplicity $D_{k,d}$, Eq. (115) becomes

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}}(\infty) = \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K}} \frac{1}{\alpha_K D} D_{k,d} \frac{s_{k,d}}{s_{k,d} + 1/\alpha_K} = \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K}} c_k^2 \frac{D_{k,d}}{c_k^2 \alpha_K D + D_{k,d}}. \tag{116}$$

Recalling that $L = \alpha_K D$, we obtain the equivalent expression

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\mathrm{kernel}}(\infty) = \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K}} c_k^2 \frac{D_{k,d}}{c_k^2 L + D_{k,d}} = 1 - \sum_{k\in\mathcal{K}} \frac{c_k^4 L}{c_k^2 L + D_{k,d}}.} \tag{117}$$

Thus the recovery of the degree-k component depends only on the ratio $L/D_{k,d}$.

Although the replica calculation above was formulated with a fixed positive $\alpha_K = L/D$, the explicit infinite-pretraining expression (117) naturally allows us to examine d-dependent context scalings by substituting $\alpha_K(d) = L(d)/D(d)$ before taking $d \to \infty$. For every fixed k,

$$D_{k,d} = \frac{d^k}{k!}\left(1 + O(d^{-1})\right). \tag{118}$$

Consider a polynomial context-length scaling

$$L = \widetilde{L} = \alpha d^s, \qquad \alpha > 0, \quad s > 0. \tag{119}$$

For each $k \in \mathcal{K}$,

$$c_k^2 \frac{D_{k,d}}{c_k^2 L + D_{k,d}} \longrightarrow \begin{cases} 0, & k < s, \\ \dfrac{c_k^2}{1 + \alpha k! c_k^2}, & k = s, \\ c_k^2, & k > s, \end{cases} \tag{120}$$

where the middle case occurs only when s is an integer contained in $\mathcal{K}$. Therefore,

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\text{kernel}}(\infty) \longrightarrow \nu_{\text{out}} + \sum_{\substack{k \in \mathcal{K} \\ k > s}} c_k^2 + \mathbf{1}_{\{s \in \mathbb{N},\, s \in \mathcal{K}\}} \frac{c_s^2}{1 + \alpha s! c_s^2},} \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}. \tag{121}$$

Hence a degree-k harmonic component is asymptotically fully recovered when $L \gg d^k$, partially recovered when $L = \Theta(d^k)$, and unrecovered when $L \ll d^k$.

B Derivation of the feature learner

B.1 Fixed point equations

Result B.1 (feature learner: state equations)**.** *Given* $(\alpha_F, \gamma_F, \kappa_F, \Delta, \lambda, \sigma_\star, \sigma_0, S_F)$*, the state equations are*

$$
\begin{aligned}
&V := \mu_1^2(q - m^2) + \frac{1+\Delta}{\alpha_F}\bar{q}, \qquad d_\chi := \mu_1^2\chi_Q + \frac{1+\Delta}{\alpha_F}\chi_{\bar{Q}}, \qquad u_0 := \mu_1 m G + \sqrt{V}\, Z,\\
&v^\star := \arg\min_{v\in\mathbb{R}} \left\{ \frac{1}{2}\left[\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon - \sigma_0(v)\right]^2 + \frac{(v-u_0)^2}{2d_\chi} \right\},\\
&\widehat{m} = \gamma_F \mathbb{E}\left[\frac{v^\star - u_0}{d_\chi} \left(\mu_1 G - \frac{\mu_1^2 m}{\sqrt{V}} Z \right) \right],\\
&\widehat{\chi}_Q = -\gamma_F \mu_1^2 \mathbb{E}\left[\frac{(v^\star - u_0)Z}{d_\chi \sqrt{V}} \right],\\
&\widehat{\chi}_{\bar{Q}} = -\gamma_F \frac{1+\Delta}{\alpha_F} \mathbb{E}\left[\frac{(v^\star - u_0)Z}{d_\chi \sqrt{V}} \right],\\
&\widehat{q}_Q = \gamma_F \mu_1^2 \mathbb{E}\left[\frac{(v^\star - u_0)^2}{d_\chi^2} \right],\\
&\widehat{q}_{\bar{Q}} = \gamma_F \frac{1+\Delta}{\alpha_F} \mathbb{E}\left[\frac{(v^\star - u_0)^2}{d_\chi^2} \right],\\
&m = \widehat{m} \int \frac{s^2}{\lambda + \widehat{\chi}_{\bar{Q}} + \widehat{\chi}_Q s}\, \nu_{\kappa_F}(\mathrm{d}s),\\
&q = \int \frac{\widehat{m}^2 s^3 + \kappa_F^{-1} s(\widehat{q}_Q s + \widehat{q}_{\bar{Q}})}{(\lambda + \widehat{\chi}_{\bar{Q}} + \widehat{\chi}_Q s)^2}\, \nu_{\kappa_F}(\mathrm{d}s),\\
&\bar{q} = \int \frac{\widehat{m}^2 s^2 + \kappa_F^{-1}(\widehat{q}_Q s + \widehat{q}_{\bar{Q}})}{(\lambda + \widehat{\chi}_{\bar{Q}} + \widehat{\chi}_Q s)^2}\, \nu_{\kappa_F}(\mathrm{d}s),\\
&\chi_Q = \frac{1}{\kappa_F} \int \frac{s}{\lambda + \widehat{\chi}_{\bar{Q}} + \widehat{\chi}_Q s}\, \nu_{\kappa_F}(\mathrm{d}s),\\
&\chi_{\bar{Q}} = \frac{1}{\kappa_F} \int \frac{1}{\lambda + \widehat{\chi}_{\bar{Q}} + \widehat{\chi}_Q s}\, \nu_{\kappa_F}(\mathrm{d}s).
\end{aligned}
\tag{122}
$$

where $G, Z, \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ *and* ν_κ *denotes the spectrum in* (44)*. The fixed point gives the five order parameters* $(m, q, \bar{q}, \chi_Q, \chi_{\bar{Q}})$ *and their conjugates* $(\widehat{m}, \widehat{q}_Q, \widehat{q}_{\bar{Q}}, \widehat{\chi}_Q, \widehat{\chi}_{\bar{Q}})$*.*

B.2 Derivation of the state equations

We condition on the task pool throughout this calculation, and evaluate the resulting normalized traces using its limiting spectrum at the end. For notational simplicity, we will drop the subscript F.

B.2.1 The replica trick

Write $\ell(y, v) := \frac{1}{2}[y - \sigma_0(v)]^2$. The Stage-I objective in Eq. (11) is,

$$
\mathcal{L}_F(W) = \frac{1}{2}\sum_{t=1}^{N} \left[\sigma_\star\left(\frac{(a^t)^\top x_0^t}{\sqrt{d}} \right) + \epsilon_0^t - \sigma_0\big(\mathrm{Tr}(W(H^t)^\top)\big) \right]^2 + \frac{N_0\lambda}{2}\|W\|_F^2. \tag{123}
$$

The dataset $\mathcal{D}$ and the prompt task indices τ^t are defined in Section 2. Below, $\mathbb{E}_{\mathcal{D}|\mathcal{A}_0}$ averages over the task indices, inputs, and label noise, conditional on the task pool. We represent loss minimization

by the integral (Edwards & Anderson, 1975; Loureiro, 2024)

$$Z_\beta(\mathcal{D}) := \int \mathrm{d}W\, e^{-\beta \mathcal{L}_F(W)}, \qquad -\lim_{\beta\to\infty} \frac{\log Z_\beta(\mathcal{D})}{\beta d N_0} = \frac{\min_W \mathcal{L}_F(W)}{dN_0}. \tag{124}$$

Here $\beta > 0$ is the inverse temperature. To evaluate the data average of the logarithm, we use its relation to moments near $n = 0$:

$$\begin{aligned} \log \mathbb{E}[Z_\beta^n \mid \mathcal{A}_0] &= \log \mathbb{E}_{\mathcal{D}|\mathcal{A}_0} e^{n \log Z_\beta} = n\, \mathbb{E}_{\mathcal{D}|\mathcal{A}_0} \log Z_\beta + O(n^2), \\ \mathbb{E}_{\mathcal{D}|\mathcal{A}_0} \log Z_\beta &= \lim_{n\to 0} \frac{1}{n} \log \mathbb{E}[Z_\beta^n \mid \mathcal{A}_0]. \end{aligned} \tag{125}$$

We compute the moments for positive integers n and continue the expression to n near zero. The calculation first takes $d \to \infty$ at fixed n, β and fixed (α, γ, κ), then extracts the term linear in n, and finally takes $\beta \to \infty$. This continuation and the exchanges of limits needed to relate it to Eq. (124) are formal steps in the replica method (Mézard et al., 1987; Loureiro, 2024).

For integer n, Z_β^n introduces n copies $W^1, \dots, W^n$ of the attention matrix. Conditional on $\mathcal{A}_0$, the prompts are independent, so

$$\begin{aligned} \mathbb{E}[Z_\beta^n \mid \mathcal{A}_0] &= \int \prod_{r=1}^n \mathrm{d}W^r \exp\left[-\frac{\beta N_0 \lambda}{2} \sum_r \mathrm{Tr}(W^r (W^r)^\top)\right] \left[\frac{1}{N_0} \sum_{k=1}^{N_0} \phi_{\beta,k}(W^1, \dots, W^n)\right]^N, \\ \phi_{\beta,k} &:= \mathbb{E}_{x_{0:L}, \epsilon_{0:L} | a^k} \exp\left[-\beta \sum_{r=1}^n \ell\Big(\sigma_\star((a^k)^\top x_0/\sqrt{d}) + \epsilon_0, \mathrm{Tr}(W^r H^\top)\Big)\right]. \end{aligned} \tag{126}$$

The factor $N_0^{-1} \sum_k \phi_{\beta,k}$ averages one prompt over its task index; the power N accounts for all training prompts.

B.2.2 Teacher score and attention outputs

For a prompt generated by task a^μ, we first compute the moments of the teacher score $g := (a^\mu)^\top x_0^\mu/\sqrt{d}$ and the attention output $u^r := \mathrm{Tr}(W^r (H^\mu)^\top), r = 1, \dots, n$. We use $\alpha = L/d$, $\widetilde{\alpha} = \widetilde{L}/d$, $\gamma = N/(dN_0)$, and $\kappa = N_0/d$, as in Result B.1.

Eq. (10) gives

$$\begin{aligned} H^\mu &:= \frac{1}{L} x_0^\mu \left(\sum_{\ell=1}^L y_\ell^\mu x_\ell^\mu\right)^\top, \\ u^r &= \mathrm{Tr}\big(W^r (H^\mu)^\top\big) = \frac{1}{L} \sum_{\ell=1}^L y_\ell^\mu (x_0^\mu)^\top W^r x_\ell^\mu. \end{aligned} \tag{127}$$

We first compute the first- and second-order moments of these local fields.

Proposition 1. *For fixed matrices $W^r, W^s \in \mathbb{R}^{d\times d}$ with bounded operator norm and $d^{-1}\mathrm{Tr}(W^r(W^r)^\top) = O(1)$,*

$$\begin{aligned} \mathbb{E}[g \mid a^\mu] &= 0, \\ \mathbb{E}[u^r \mid a^\mu] &= 0, \\ \mathbb{E}[g^2 \mid a^\mu] &= 1, \\ \mathbb{E}[g\, u^r \mid a^\mu] &= \frac{\mu_1}{d} (a^\mu)^\top W^r a^\mu + o(1), \\ \mathbb{E}[u^r u^s \mid a^\mu] &= \frac{\mu_1^2}{d} (W^r a^\mu)^\top (W^s a^\mu) + \frac{1+\Delta}{\alpha d} \mathrm{Tr}\big(W^r (W^s)^\top\big) + o(1). \end{aligned} \tag{128}$$

Proof. Using $\mu_1 = \mathbb{E}[\sigma_\star(G)G]$, we decompose $x = g_d \frac{a^\mu}{\sqrt{d}} + x_\perp$. Then

$$\mathbb{E}[yx \mid a^\mu] = \mathbb{E}\left[y\left(g_d\frac{a^\mu}{\sqrt{d}} + x_\perp\right)\middle| a^\mu\right] = \mathbb{E}[yg_d \mid a^\mu]\frac{a^\mu}{\sqrt{d}} + \mathbb{E}[y\,\mathbb{E}[x_\perp \mid g_d, a^\mu]|a^\mu] = \mu_1\frac{a^\mu}{\sqrt{d}} + o(1),$$
$$\mathbb{E}[y^2xx^\top \mid a^\mu] = (1+\Delta)I_d + O(1)\frac{a^\mu(a^\mu)^\top}{d} + o(1). \tag{129}$$

The mixed moment of (g, u^r) is therefore

$$\begin{aligned}\mathbb{E}[g\,u^r \mid a^\mu] &= \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}\big[g\,y_\ell^\mu (x_0^\mu)^\top W^r x_\ell^\mu | a^\mu\big] \\ &= \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{\sqrt{d}}(a^\mu)^\top W^r \mathbb{E}[y_\ell^\mu x_\ell^\mu \mid a^\mu] \\ &= \frac{\mu_1}{d}(a^\mu)^\top W^r a^\mu + o(1). \end{aligned} \tag{130}$$

For the second-order student statistic u^r,

$$\mathbb{E}[u^\mu(W^r)u^\mu(W^s)|a^\mu] \tag{131}$$
$$= \frac{1}{L^2}\sum_{\ell\neq\ell'}\mathbb{E}\big[y_\ell^\mu y_{\ell'}^\mu (x_0^\mu)^\top W^r x_\ell^\mu (x_0^\mu)^\top W^s x_{\ell'}^\mu | a^\mu\big] + \frac{1}{L^2}\sum_{\ell=1}^{L}\mathbb{E}\big[(y_\ell^\mu)^2 (x_0^\mu)^\top W^r x_\ell^\mu (x_0^\mu)^\top W^s x_\ell^\mu | a^\mu\big]$$
$$= \frac{L-1}{L}\frac{\mu_1^2}{d}(W^r a^\mu)^\top (W^s a^\mu) + \frac{1}{L}\operatorname{Tr}\big[W^r\mathbb{E}[y^2xx^\top \mid a^\mu](W^s)^\top\big]$$
$$= \frac{\mu_1^2}{d}(W^r a^\mu)^\top (W^s a^\mu) + \frac{1+\Delta}{\alpha d}\operatorname{Tr}\big(W^r(W^s)^\top\big) + o(1). \tag{132}$$

These calculations prove Eq. (128). □

These definitions are imposed inside $\phi_{\beta,k}$ by the Fourier representations of Dirac delta functions

$$1 = \int \mathrm{d}g\,\delta(g - (a^k)^\top x_0/\sqrt{d}) = \int \frac{\mathrm{d}g\,\mathrm{d}\widehat{g}}{2\pi} e^{i\widehat{g}(g-(a^k)^\top x_0/\sqrt{d})}, \tag{133}$$

$$1 = \int \mathrm{d}u^r\,\delta(u^r - \operatorname{Tr}(W^r H^\top)) = \int \frac{\mathrm{d}u^r\,\mathrm{d}\widehat{u}^r}{2\pi} e^{i\widehat{u}^r(u^r - \operatorname{Tr}(W^r H^\top))}, \qquad r = 1,\dots,n. \tag{134}$$

The variables $\widehat{g}, \widehat{u}^1, \dots, \widehat{u}^n$ are temporary Fourier variables and are unrelated to the conjugate order parameters introduced later. If p_Δ denotes the density of $\mathcal{N}(0,\Delta)$, substituting Eqs. (133) and (134) gives

$$\begin{aligned}\phi_{\beta,k} = \int \mathrm{d}\epsilon_0 p_\Delta(\epsilon_0) \int \frac{\mathrm{d}g\,\mathrm{d}\widehat{g}}{2\pi}\prod_{r=1}^{n}\frac{\mathrm{d}u^r\,\mathrm{d}\widehat{u}^r}{2\pi} e^{-\beta\sum_r \ell(\sigma_\star(g)+\epsilon_0, u^r)} e^{i\widehat{g}g + i\sum_r \widehat{u}^r u^r} \\ \times\, \mathbb{E}_{x_{0:L},\epsilon_{1:L}|a^k}\exp\left[-i\widehat{g}(a^k)^\top x_0/\sqrt{d} - i\sum_r \widehat{u}^r \operatorname{Tr}(W^r H^\top)\right].\end{aligned} \tag{135}$$

Assumption 1 (Gaussian equivalence of the local field)**.** *For fixed n and $L/d \to \alpha \in (0,\infty)$, condition on $a^k, W^1, \dots, W^n$, independently of the current prompt. We assume that the teacher score and the attention outputs can be replaced, in an averaged prompt, by jointly Gaussian variables with the same leading first two moments:*

$$(g, u^1, \dots, u^n) \Longrightarrow \mathcal{N}(0, C_k), \tag{136}$$

where

$$\begin{aligned}(C_k)_{00} &= 1, \\ (C_k)_{0r} &= \frac{\mu_1}{d}(a^k)^\top W^r a^k + o(1), \\ (C_k)_{rs} &= \frac{\mu_1^2}{d}(W^r a^k)^\top (W^s a^k) + \frac{1+\Delta}{\alpha d}\operatorname{Tr}(W^r(W^s)^\top) + o(1).\end{aligned} \tag{137}$$

Here we assume the first two moments are matched, and higher-order correlations are negligible in the high-dimensional limit.

Let $\boldsymbol{z} := (g, u^1, \dots, u^n)^\top$ and $\widehat{\boldsymbol{z}} := (\widehat{g}, \widehat{u}^1, \dots, \widehat{u}^n)^\top$. Assumption 1 gives

$$\mathbb{E}_{x_{0:L},\epsilon_{1:L}|a^k} e^{-i\widehat{\boldsymbol{z}}^\top \boldsymbol{z}} = \exp\left(-\frac{1}{2}\widehat{\boldsymbol{z}}^\top C_k \widehat{\boldsymbol{z}}\right) + o(1). \tag{138}$$

Substituting this function into Eq. (135) and integrating the Fourier variables gives

$$\begin{aligned}\phi_{\beta,k} &= \int \mathrm{d}\epsilon_0 p_\Delta(\epsilon_0) \int \mathrm{d}g \prod_r \mathrm{d}u^r \int \frac{\mathrm{d}^{n+1}\widehat{\boldsymbol{z}}}{(2\pi)^{n+1}} e^{-\beta \sum_r \ell(\sigma_\star(g)+\epsilon_0, u^r)} e^{i\widehat{\boldsymbol{z}}^\top \boldsymbol{z} - \widehat{\boldsymbol{z}}^\top C_k \widehat{\boldsymbol{z}}/2} + o(1) \\ &= \int \mathrm{d}\epsilon_0 p_\Delta(\epsilon_0) \int \mathrm{d}g \prod_r \mathrm{d}u^r \, \mathcal{N}((g, u^1, \dots, u^n) \mid 0, C_k) e^{-\beta \sum_r \ell(\sigma_\star(g)+\epsilon_0, u^r)} + o(1).\end{aligned} \tag{139}$$

The Fourier integrals thus replace the average over the prompt by a Gaussian integral over $(g, u^1, \dots, u^n)$.

B.2.3 Order parameters and replica symmetry

As in Takanami et al. (2026), we assume self-averaging with respect to the tasks in the pool. The task-dependent quadratic forms can then be replaced by their averages over the task pool. We therefore introduce

$$\begin{aligned} m^r &:= \frac{1}{dN_0} \sum_{k=1}^{N_0} (a^k)^\top W^r a^k = \frac{1}{d} \operatorname{Tr}(W^r S), \\ Q^{rs} &:= \frac{1}{dN_0} \sum_{k=1}^{N_0} (W^r a^k)^\top (W^s a^k) = \frac{1}{d} \operatorname{Tr}(W^r S (W^s)^\top), \\ \bar{Q}^{rs} &:= \frac{1}{d} \operatorname{Tr}(W^r (W^s)^\top). \end{aligned} \tag{140}$$

Equations (137) give

$$\begin{aligned} C_{00} &= 1, \\ C_{0r} &= \mu_1 m^r, \\ C_{rs} &= \mu_1^2 Q^{rs} + \frac{1+\Delta}{\alpha} \bar{Q}^{rs}. \end{aligned} \tag{141}$$

Hence

$$\phi_\beta(m, Q, \bar{Q}) = \int \mathrm{d}\epsilon_0 p_\Delta(\epsilon_0) \int \mathrm{d}g \prod_r \mathrm{d}u^r \, \mathcal{N}((g, u^1, \dots, u^n) \mid 0, C) e^{-\beta \sum_r \ell(\sigma_\star(g)+\epsilon_0, u^r)}, \tag{142}$$

and

$$\mathbb{E}[Z_\beta^n \mid \mathcal{A}_0] \simeq \int \prod_r \mathrm{d}W^r \exp\left[-\frac{\beta N_0 \lambda}{2} \sum_r \operatorname{Tr}(W^r (W^r)^\top)\right] [\phi_\beta(m, Q, \bar{Q})]^N. \tag{143}$$

We insert Eq. (140) into the integral over the matrices using Dirac delta functions. Their Fourier representations introduce the conjugate parameters $\widehat{m}^r, \widehat{Q}^{rs}, \widehat{\bar{Q}}^{rs}$:

$$\begin{aligned} 1 &= \int \mathrm{d}m^r \, \delta\left(m^r - \frac{1}{d}\operatorname{Tr}(W^r S)\right) \propto \int \mathrm{d}m^r \, \mathrm{d}\widehat{m}^r \exp[-dN_0 \widehat{m}^r m^r + N_0 \widehat{m}^r \operatorname{Tr}(W^r S)], \\ 1 &= \int \mathrm{d}Q^{rs} \, \delta\left(Q^{rs} - \frac{1}{d}\operatorname{Tr}(W^r S (W^s)^\top)\right) \propto \int \mathrm{d}Q^{rs} \, \mathrm{d}\widehat{Q}^{rs} \exp\left[-\frac{dN_0}{2}\widehat{Q}^{rs} Q^{rs} + \frac{N_0}{2}\widehat{Q}^{rs} \operatorname{Tr}(W^r S (W^s)^\top)\right], \\ 1 &= \int \mathrm{d}\bar{Q}^{rs} \, \delta\left(\bar{Q}^{rs} - \frac{1}{d}\operatorname{Tr}(W^r (W^s)^\top)\right) \propto \int \mathrm{d}\bar{Q}^{rs} \, \mathrm{d}\widehat{\bar{Q}}^{rs} \exp\left[-\frac{dN_0}{2}\widehat{\bar{Q}}^{rs} \bar{Q}^{rs} + \frac{N_0}{2}\widehat{\bar{Q}}^{rs} \operatorname{Tr}(W^r (W^s)^\top)\right]. \end{aligned} \tag{144}$$

The conjugate integrations are along the Fourier contours, with the factors of i absorbed into the conjugate variables. Substituting these identities separates the dependence on the matrices from that on the order parameters:

$$\mathbb{E}[Z_\beta^n \mid \mathcal{A}_0] \simeq \int \mathrm{d}\Omega\, e^{-n\beta d N_0 f_{n,\beta}(\Omega)}, \tag{145}$$

where $\Omega = \{m^r, Q^{rs}, \bar{Q}^{rs}, \widehat{m}^r, \widehat{Q}^{rs}, \widehat{\bar{Q}}^{rs}\}$ and

$$\begin{aligned} f_{n,\beta}(\Omega) &= -\frac{\gamma}{n\beta}\log\phi_\beta + \frac{1}{n\beta}\sum_r \widehat{m}^r m^r + \frac{1}{2n\beta}\sum_{r,s}(\widehat{Q}^{rs}Q^{rs} + \widehat{\bar{Q}}^{rs}\bar{Q}^{rs}) - \frac{1}{n\beta d N_0}\log\int\prod_r \mathrm{d}W^r\, e^{G_n}, \\ G_n &= \frac{N_0}{2}\sum_{r,s}\widehat{Q}^{rs}\,\mathrm{Tr}(W^r S (W^s)^\top) + \frac{N_0}{2}\sum_{r,s}\widehat{\bar{Q}}^{rs}\,\mathrm{Tr}(W^r (W^s)^\top) \\ &\quad + N_0\sum_r \widehat{m}^r\,\mathrm{Tr}(W^r S) - \frac{\beta N_0 \lambda}{2}\sum_r \mathrm{Tr}(W^r (W^r)^\top). \end{aligned} \tag{146}$$

In the high-dimensional limit, the integral in Eq. (145) is evaluated at a stationary point of $f_{n,\beta}$. We next substitute the replica-symmetric form and take the limits described above.

Assumption 2 (Replica symmetry)**.** *We assume that the stationary solution is invariant under permutations of the replica indices. In the zero-temperature limit, write*

$$\begin{aligned} m^r &= m, \\ Q^{rs} &= q + \frac{\chi_Q}{\beta}\delta_{rs}, \\ \bar{Q}^{rs} &= \bar{q} + \frac{\chi_{\bar{Q}}}{\beta}\delta_{rs}. \end{aligned} \tag{147}$$

Here $q, \bar{q}$ are the off-diagonal values, and $\chi_Q, \chi_{\bar{Q}}$ give the leading difference between diagonal and off-diagonal values. We use the same symmetry for the conjugate parameters Takanami et al. (2026); Loureiro (2024).

The $1/\beta$ differences retain the variance around a stable minimum as the loss-weighted distribution concentrates. This scaling holds for a strictly convex quadratic objective, as in the identity-readout case with $\lambda > 0$; for a general σ_0, finite limits of these scaled differences are part of the ansatz Loureiro (2024). The corresponding scaling of the conjugate parameters is

$$\begin{aligned} \widehat{Q}^{rs} &= \beta^2 \widehat{q}_Q - \beta\widehat{\chi}_Q\delta_{rs}, \\ \widehat{\bar{Q}}^{rs} &= \beta^2 \widehat{q}_{\bar{Q}} - \beta\widehat{\chi}_{\bar{Q}}\delta_{rs}, \\ \widehat{m}^r &= \beta\widehat{m}. \end{aligned} \tag{148}$$

After substituting Eqs. (147) and (148), the remaining variables are

$$\boldsymbol{x} = (m, q, \bar{q}, \chi_Q, \chi_{\bar{Q}}, \widehat{m}, \widehat{q}_Q, \widehat{q}_{\bar{Q}}, \widehat{\chi}_Q, \widehat{\chi}_{\bar{Q}}). \tag{149}$$

Grouping the terms in Eq. (146) gives

$$f(\boldsymbol{x}) = \lim_{\beta\to\infty}\lim_{n\to 0} f_{n,\beta}(\Omega)|_{\mathrm{RS}} = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + f_3(\boldsymbol{x}), \tag{150}$$

up to additive constants independent of $\boldsymbol{x}$. With u_0, V, d_χ as defined in Eqs. (122) and $A_F := (\lambda + \widehat{\chi}_{\bar{Q}})I_d + \widehat{\chi}_Q S_F$, $T_F := \widehat{q}_Q S_F + \widehat{q}_{\bar{Q}} I_d$., we would obtain

$$\begin{aligned} f_1(\boldsymbol{x}) &= \gamma\mathbb{E}_{G,Z,\varepsilon}\min_{v\in\mathbb{R}}\left\{\ell\Big(\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon, v\Big) + \frac{(v-u_0)^2}{2d_\chi}\right\}, \\ f_2(\boldsymbol{x}) &= \widehat{m}m - \frac{1}{2}\widehat{\chi}_Q q - \frac{1}{2}\widehat{\chi}_{\bar{Q}}\bar{q} + \frac{1}{2}\widehat{q}_Q\chi_Q + \frac{1}{2}\widehat{q}_{\bar{Q}}\chi_{\bar{Q}}, \\ f_3(\boldsymbol{x}) &= -\frac{1}{2d}\left[\widehat{m}^2\,\mathrm{Tr}(S^2 A_F^{-1}) + \frac{1}{\kappa}\,\mathrm{Tr}(T_F A_F^{-1})\right]. \end{aligned} \tag{151}$$

Here G, Z, ε are independent standard Gaussian variables. The three functions correspond, respectively, to the term containing $\log\phi_\beta$, the sums over the order parameters and their conjugates, and the

logarithm of the matrix integral in Eq. (146). The conjugate parameters act as auxiliary multipliers enforcing the definitions of the order parameters.

The state equations follow from

$$\frac{\partial f}{\partial x_i} = 0, \qquad i = 1, \dots, 10. \tag{152}$$

We seek a stationary point of f. The function determines the training statistics. Below we calculate f_1, f_2, f_3 in turn and then take these derivatives.

B.2.4 Calculation of f_1

To calculate f_1, we evaluate the factor ϕ_β obtained above. The teacher score g and the replicated attention outputs $(u^1, \dots, u^n)$ can be replaced by jointly Gaussian variables with the covariance in Eq. (141). After imposing the replica-symmetric form in Eq. (147), this Gaussian vector can be represented as

$$g = G, \qquad u^r = \mu_1 m G + \sqrt{V}\, Z + \sqrt{\frac{d_\chi}{\beta}}\, \eta_r, \qquad G, Z, \eta_1, \dots, \eta_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{153}$$

where $V = \mu_1^2(q - m^2) + (1+\Delta)\bar{q}/\alpha$ and $d_\chi = \mu_1^2 \chi_Q + (1+\Delta)\chi_{\bar{Q}}/\alpha$, as defined in Result B.1.

The first two terms in u^r are shared by all replicas in Eq. (141). The last term is replica-dependent. Indeed, $\mathbb{E}[G u^r] = \mu_1 m$ and $\mathbb{E}[u^r u^s] = \mu_1^2 q + (1+\Delta)\bar{q}/\alpha + (d_\chi/\beta)\delta_{rs}$, which agrees with the covariance obtained from the order parameters.

Using this Gaussian representation and the definition of u_0 in Eq. (122), we have

$$\phi_\beta = \mathbb{E}_{G,Z,\varepsilon}\left\{ \left[\int \frac{e^{-\eta^2/2}\, \mathrm{d}\eta}{\sqrt{2\pi}}\, e^{-\beta \ell(\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon,\, u_0 + \sqrt{d_\chi/\beta}\,\eta)} \right]^n \right\}. \tag{154}$$

Extracting the coefficient linear in n gives

$$\lim_{n\to 0} \frac{1}{n} \log \phi_\beta = \mathbb{E}_{G,Z,\varepsilon} \log \int \frac{e^{-\eta^2/2}\, \mathrm{d}\eta}{\sqrt{2\pi}}\, \exp\left[-\frac{\beta}{2} \left(\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon - \sigma_0\left(u_0 + \sqrt{\frac{d_\chi}{\beta}}\,\eta \right) \right)^2 \right]. \tag{155}$$

Thus the original average over all inputs and demonstrations in a prompt has been reduced to an average over a few scalar Gaussian variables.

To take $\beta \to \infty$, set $v = u_0 + \sqrt{d_\chi/\beta}\,\eta$. Writing $y = \sigma_\star(G) + \sqrt{\Delta}\,\varepsilon$ within this calculation,

$$\int \frac{\mathrm{d}\eta}{\sqrt{2\pi}}\, e^{-\eta^2/2 - \beta \ell(y, u_0 + \sqrt{d_\chi/\beta}\,\eta)} = \sqrt{\frac{\beta}{2\pi d_\chi}} \int \mathrm{d}v\, e^{-\beta\{\ell(y,v) + (v-u_0)^2/(2d_\chi)\}}. \tag{156}$$

The quadratic term comes from the Gaussian density. Taking the negative logarithm and dividing by β makes the logarithm of the prefactor vanish, while the integral gives the minimum of the scalar objective. Hence

$$f_1(\boldsymbol{x}) = -\lim_{\beta\to\infty} \lim_{n\to 0} \frac{\gamma}{n\beta} \log \phi_\beta = \gamma \mathbb{E}_{G,Z,\varepsilon} \min_{v\in\mathbb{R}} \left\{ \frac{1}{2}[\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon - \sigma_0(v)]^2 + \frac{(v-u_0)^2}{2d_\chi} \right\}. \tag{157}$$

We denote the minimizer of the scalar problem in Eq. (157) by $v^\star$. It is this scalar minimization that enters the fixed-point equations. For the identity Stage-I readout $\sigma_0(v) = v$, the objective is strictly convex and $v^\star$ is unique.

B.2.5 CALCULATION OF f_2

For f_2, we evaluate the sums in Eq. (146). For the Q terms, substituting the two ansatzes gives

$$\frac{1}{2n\beta}\sum_{r,s}\widehat{Q}^{rs}Q^{rs} = \frac{n\beta}{2}\widehat{q}_Q q + \frac{1}{2}\widehat{q}_Q\chi_Q - \frac{1}{2}\widehat{\chi}_Q q - \frac{1}{2\beta}\widehat{\chi}_Q\chi_Q \xrightarrow{n\to 0 \text{ then } \beta\to\infty} \frac{1}{2}\widehat{q}_Q\chi_Q - \frac{1}{2}\widehat{\chi}_Q q$$
$$\frac{1}{n\beta}\sum_r \widehat{m}^r m^r = \widehat{m}m. \tag{158}$$

The $\bar{Q}$ terms give the same expression with bars. Combining these terms gives f_2 in Eq. (151).

B.2.6 CALCULATION OF f_3

To calculate f_3, we evaluate the matrix integral in Eq. (146).

Substitution into G_n gives

$$G_n = \frac{N_0\beta^2}{2}\operatorname{Tr}\left[\left(\sum_r W^r\right)T_F\left(\sum_s W^s\right)^\top\right] - \beta N_0\sum_r\left[\frac{1}{2}\operatorname{Tr}(W^r A_F (W^r)^\top) - \widehat{m}\operatorname{Tr}(W^r S)\right]. \tag{159}$$

The remaining difficulty in the matrix integral is the quadratic coupling between different replicas through $\sum_r W^r$. We remove this coupling by introducing an auxiliary Gaussian matrix. For $(Z_W)_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, the Hubbard–Stratonovich identity gives

$$\exp\left\{\frac{N_0\beta^2}{2}\operatorname{Tr}\left[\left(\sum_r W^r\right)T_F\left(\sum_s W^s\right)^\top\right]\right\} = \mathbb{E}_{Z_W}\exp\left[\beta\sqrt{N_0}\sum_r \operatorname{Tr}(W^r T_F^{1/2} Z_W^\top)\right]. \tag{160}$$

Conditioned on Z_W, the replicas are now independent. Therefore the matrix integral factorizes as

$$\begin{aligned}\int\prod_r \mathrm{d}W^r\, e^{G_n} &= \mathbb{E}_{Z_W}\prod_{r=1}^n\int \mathrm{d}W^r \exp\left\{-\beta N_0\begin{bmatrix}\frac{1}{2}\operatorname{Tr}(W^r A_F (W^r)^\top) - \widehat{m}\operatorname{Tr}(W^r S)\\ -\frac{1}{\sqrt{N_0}}\operatorname{Tr}(Z_W T_F^{1/2}(W^r)^\top)\end{bmatrix}\right\}\\ &= \mathbb{E}_{Z_W}\left\{\left[\int \mathrm{d}W\, e^{-\beta N_0\mathcal{L}_{\text{eff}}(W;Z_W)}\right]^n\right\},\end{aligned} \tag{161}$$

where

$$\mathcal{L}_{\text{eff}}(W;Z_W) := \frac{1}{2}\operatorname{Tr}(W A_F W^\top) - \widehat{m}\operatorname{Tr}(WS) - \frac{1}{\sqrt{N_0}}\operatorname{Tr}(Z_W T_F^{1/2} W^\top). \tag{162}$$

For positive definite A_F, the integral over W is Gaussian. The minimizer of its quadratic exponent satisfies

$$0 = WA_F - \widehat{m}S - \frac{1}{\sqrt{N_0}}Z_W T_F^{1/2} \qquad\Longrightarrow\qquad W_F(Z_W) = \left(\widehat{m}S + \frac{1}{\sqrt{N_0}}Z_W T_F^{1/2}\right)A_F^{-1}. \tag{163}$$

Completing the square gives the exact integral

$$\int \mathrm{d}W\, e^{-\beta N_0\mathcal{L}_{\text{eff}}(W;Z_W)} = e^{-\beta N_0\mathcal{L}_{\text{eff}}(W_F(Z_W);Z_W)}\left(\frac{2\pi}{\beta N_0}\right)^{d^2/2}(\det A_F)^{-d/2}. \tag{164}$$

Its contribution to Eq. (146), after taking $n\to 0$, is therefore

$$\begin{aligned}-\lim_{n\to 0}\frac{1}{n\beta d N_0}\log\int\prod_r \mathrm{d}W^r\, e^{G_n} &= -\frac{1}{\beta d N_0}\mathbb{E}_{Z_W}\log\int \mathrm{d}W\, e^{-\beta N_0\mathcal{L}_{\text{eff}}(W;Z_W)}\\ &= \frac{1}{d}\mathbb{E}_{Z_W}\min_W \mathcal{L}_{\text{eff}}(W;Z_W) + \frac{\log\det A_F}{2\beta N_0} + \frac{d}{2\beta N_0}\log\frac{\beta N_0}{2\pi}.\end{aligned} \tag{165}$$

The last term is independent of the order parameters. If $d^{-1}\log\det A_F$ remains bounded, the determinant term vanishes as $\beta\to\infty$. Substituting the minimizer into $\mathcal{L}_{\text{eff}}$ gives

$$\begin{aligned} f_3(\boldsymbol{x}) = \frac{1}{d}\mathbb{E}_{Z_W}\min_W \mathcal{L}_{\text{eff}}(W;Z_W) &= -\frac{1}{2d}\mathbb{E}_{Z_W}\operatorname{Tr}\left[\left(\widehat{m}S+\frac{Z_W T_F^{1/2}}{\sqrt{N_0}}\right)A_F^{-1}\left(\widehat{m}S+\frac{Z_W T_F^{1/2}}{\sqrt{N_0}}\right)^\top\right] \\ &= -\frac{1}{2d}\left[\widehat{m}^2\operatorname{Tr}(S^2A_F^{-1})+\frac{1}{N_0}\mathbb{E}_{Z_W}\operatorname{Tr}\Big(Z_W T_F^{1/2}A_F^{-1}T_F^{1/2}Z_W^\top\Big)\right] \\ &= -\frac{1}{2d}\left[\widehat{m}^2\operatorname{Tr}(S^2A_F^{-1})+\frac{d}{N_0}\operatorname{Tr}(T_FA_F^{-1})\right] \\ &= -\frac{1}{2d}\left[\widehat{m}^2\operatorname{Tr}(S^2A_F^{-1})+\frac{1}{\kappa}\operatorname{Tr}(T_FA_F^{-1})\right], \end{aligned} \tag{166}$$

where we used $\mathbb{E}[Z_W T_F^{1/2}A_F^{-1}T_F^{1/2}Z_W^\top]=\operatorname{Tr}(T_FA_F^{-1})I_d$ and $N_0=\kappa d$ in the last two lines.

Together, Eqs. (157), (158), and (166) give f_1, f_2, f_3 in Eq. (151), and hence $f=f_1+f_2+f_3$.

B.2.7 Stationarity

We first differentiate f with respect to the five order parameters $(m,q,\bar{q},\chi_Q,\chi_{\bar{Q}})$. At a differentiable scalar minimum, the derivative through $v^\star$ vanishes because the scalar objective is stationary there. Thus

$$\begin{aligned} 0=\partial_m f &= \widehat{m}-\gamma\mathbb{E}\left[\frac{v^\star-u_0}{d_\chi}\left(\mu_1 G-\frac{\mu_1^2 m}{\sqrt{V}}Z\right)\right], \\ 0=2\partial_q f &= -\widehat{\chi}_Q-\gamma\mu_1^2\mathbb{E}\left[\frac{(v^\star-u_0)Z}{d_\chi\sqrt{V}}\right], \\ 0=2\partial_{\bar{q}} f &= -\widehat{\chi}_{\bar{Q}}-\gamma\frac{1+\Delta}{\alpha}\mathbb{E}\left[\frac{(v^\star-u_0)Z}{d_\chi\sqrt{V}}\right], \\ 0=2\partial_{\chi_Q} f &= \widehat{q}_Q-\gamma\mu_1^2\mathbb{E}\left[\frac{(v^\star-u_0)^2}{d_\chi^2}\right], \\ 0=2\partial_{\chi_{\bar{Q}}} f &= \widehat{q}_{\bar{Q}}-\gamma\frac{1+\Delta}{\alpha}\mathbb{E}\left[\frac{(v^\star-u_0)^2}{d_\chi^2}\right]. \end{aligned} \tag{167}$$

These are the five conjugate equations in Eq. (122). Taking derivatives of f with respect to the conjugates gives the remaining equations:

$$\begin{aligned} 0=\partial_{\widehat{m}} f &= m-\frac{\widehat{m}}{d}\operatorname{Tr}(S^2A_F^{-1}), \\ 0=-2\partial_{\widehat{\chi}_Q} f &= q-\frac{\widehat{m}^2}{d}\operatorname{Tr}(S^3A_F^{-2})-\frac{1}{\kappa d}\operatorname{Tr}(T_FSA_F^{-2}), \\ 0=-2\partial_{\widehat{\chi}_{\bar{Q}}} f &= \bar{q}-\frac{\widehat{m}^2}{d}\operatorname{Tr}(S^2A_F^{-2})-\frac{1}{\kappa d}\operatorname{Tr}(T_FA_F^{-2}), \\ 0=2\partial_{\widehat{q}_Q} f &= \chi_Q-\frac{1}{\kappa d}\operatorname{Tr}(SA_F^{-1}), \\ 0=2\partial_{\widehat{q}_{\bar{Q}}} f &= \chi_{\bar{Q}}-\frac{1}{\kappa d}\operatorname{Tr}(A_F^{-1}). \end{aligned} \tag{168}$$

Together, the derivatives with respect to the five order parameters and their five conjugates recover Result B.1 under the fact that as $d, N_0\to\infty$ with $N_0/d\to\kappa_F\in(0,\infty)$, the empirical spectrum of S_F converges almost surely to $\nu_{\kappa_F}(\mathrm{d}s)$.

B.3 Derivation of the main results

Equation (163) gives the learned attention matrix stated in the main text:

$$\boxed{\widehat{W} \stackrel{\mathrm{d}}{=} \left(\widehat{m}S + \frac{1}{\sqrt{N_0}} Z(\widehat{q}_Q S + \widehat{q}_{\bar{Q}} I_d)^{1/2}\right)\left[(\lambda + \widehat{\chi}_{\bar{Q}})I_d + \widehat{\chi}_Q S\right]^{-1}}. \tag{169}$$

After the fixed point is solved, Stage II is trained on fresh prompts from the training distribution. Its attention output has moments

$$\begin{aligned} \mathbb{E}[G u_{\mathrm{tr}}] &= \mu_1 m, \\ \mathbb{E}[u_{\mathrm{tr}}^2] &= \mu_1^2 q + \frac{1+\Delta}{\alpha}\bar{q} \end{aligned} \tag{170}$$

Hence, the output of the attention can be represented as

$$u_{\mathrm{tr}} = \mu_1 m G + \sqrt{\mu_1^2(q - m^2) + \frac{1+\Delta}{\alpha}\bar{q}}\, Z, \qquad G, Z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1). \tag{171}$$

We next evaluate the same learned attention matrix on the noiseless inference prompts. The only changes are the task distribution and the prompt length $\widetilde{L} = \widetilde{\alpha} d$. For a task drawn from the training pool, the relevant overlaps are $m, q, \bar{q}$, whereas a fresh task averages over an isotropic direction and therefore involves $\mathbb{E}_Z \operatorname{Tr} \widehat{W}$. This gives

$$\begin{aligned} C_{\mathrm{M}} &= \mu_1 m, & V_{\mathrm{M}} &= \mu_1^2 q + \frac{\bar{q}}{\widetilde{\alpha}}, \\ C_{\mathrm{G}} &= \frac{\mu_1}{d}\mathbb{E}_Z \operatorname{Tr} \widehat{W}, & V_{\mathrm{G}} &= \left(\mu_1^2 + \frac{1}{\widetilde{\alpha}}\right)\bar{q}. \end{aligned} \tag{172}$$

Thus, for either test protocol $\mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}$, the attention output can be represented as

$$u_{\mathcal{P}} = C_{\mathcal{P}} G + \sqrt{V_{\mathcal{P}} - C_{\mathcal{P}}^2}\, Z, \qquad G, Z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1), \tag{173}$$

and the corresponding error is

$$\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} = \mathbb{E}_{G,Z}\left[\left(\sigma_\star(G) - \widehat{\sigma}(u_{\mathcal{P}})\right)^2\right]. \tag{174}$$

It remains to determine the Stage-II readout $\widehat{\sigma}$. The Stage-II training label is $\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon$, where ε is independent of (G, u_{tr}). Therefore, for any scalar function h, $\mathbb{E}\left[\left(\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon - h(u_{\mathrm{tr}})\right)^2\right] = \Delta + \mathbb{E}\left[\left(\sigma_\star(G) - h(u_{\mathrm{tr}})\right)^2\right]$. The query noise therefore adds only the constant Δ and does not affect the optimal readout. Conditioning on u_{tr} gives

$$\mathbb{E}\left[\left(\sigma_\star(G) - h(u_{\mathrm{tr}})\right)^2 \mid u_{\mathrm{tr}}\right] = \operatorname{Var}(\sigma_\star(G) \mid u_{\mathrm{tr}}) + \left(\mathbb{E}[\sigma_\star(G) \mid u_{\mathrm{tr}}] - h(u_{\mathrm{tr}})\right)^2. \tag{175}$$

The second term is minimized by $\widehat{\sigma}(v) = \mathbb{E}[\sigma_\star(G) \mid u_{\mathrm{tr}} = v]$.

Since (G, u_{tr}) is jointly Gaussian, its conditional distribution is

$$G \mid (u_{\mathrm{tr}} = v) \stackrel{\mathrm{d}}{=} \frac{\mu_1 m}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha}\, v + \sqrt{1 - \frac{\mu_1^2 m^2}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha}}\, \psi, \qquad \psi \sim \mathcal{N}(0,1). \tag{176}$$

Therefore,

$$\boxed{\widehat{\sigma}(v) = \mathbb{E}_\psi\left[\sigma_\star\left(\frac{\mu_1 m}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha}\, v + \sqrt{1 - \frac{\mu_1^2 m^2}{\mu_1^2 q + (1+\Delta)\bar{q}/\alpha}}\, \psi\right)\right]}, \tag{177}$$

where $\psi \sim \mathcal{N}(0,1)$. Equations (169)–(177) determine the learned attention matrix, the Stage-II readout, and the two inference errors, and therefore derive Result 2.

Now we set $\sigma_0(v) = v$. We first study the large-pretraining regime ($\gamma \to \infty$ with $\alpha, \widetilde{\alpha}$ fixed). Then we take the infinite-pretraining and infinite-task-pool limits with the more general context scaling $L = \widetilde{L} = \alpha d^s$.

Large-pretraining limit. For $\sigma_0(v) = v$, the scalar minimizer in Eq. (122) is

$$v^\star = \frac{u_0 + d_\chi y}{1 + d_\chi}, \qquad y := \sigma_\star(G) + \sqrt{\Delta}\,\varepsilon. \tag{178}$$

Hence the conjugate equations simplify to

$$\begin{aligned} \widehat{m} = \widehat{\chi}_Q &= \frac{\gamma\mu_1^2}{1 + d_\chi}, \\ \widehat{\chi}_{\bar{Q}} &= \frac{\gamma(1+\Delta)}{\alpha(1 + d_\chi)}, \\ \widehat{q}_Q &= \frac{\gamma\mu_1^2 R}{(1 + d_\chi)^2}, \\ \widehat{q}_{\bar{Q}} &= \frac{\gamma(1+\Delta)R}{\alpha(1 + d_\chi)^2}, \end{aligned} \tag{179}$$

where $R := 1 + \Delta - 2\mu_1^2 m + \mu_1^2 q + \frac{1+\Delta}{\alpha}\bar{q}$. Define $\tau := \frac{1+\Delta}{\alpha\mu_1^2}$. Equation (179) implies

$$\widehat{\chi}_{\bar{Q}} = \tau\widehat{m}, \qquad \widehat{q}_{\bar{Q}} = \tau\widehat{q}_Q. \tag{180}$$

Therefore

$$A = \widehat{m}\left[S + \left(\tau + \frac{\lambda}{\widehat{m}}\right) I_d\right], \qquad T = \widehat{q}_Q(S + \tau I_d). \tag{181}$$

Substituting these identities into Eq. (169) gives

$$\widehat{W} \overset{\mathrm{d}}{=} S\left[S + \left(\tau + \frac{\lambda}{\widehat{m}}\right) I_d\right]^{-1} + \sqrt{\frac{\widehat{q}_Q}{N_0\widehat{m}^2}}\, Z(S + \tau I_d)^{1/2}\left[S + \left(\tau + \frac{\lambda}{\widehat{m}}\right) I_d\right]^{-1}. \tag{182}$$

Moreover, Eq. (179) yields

$$\frac{\widehat{q}_Q}{\widehat{m}^2} = \frac{R}{\gamma\mu_1^2}. \tag{183}$$

For fixed $\alpha > 0$, the resolvents in Eq. (182) are uniformly bounded because $\tau > 0$. Using the equations for χ_Q and $\chi_{\bar{Q}}$,

$$d_\chi = \frac{\mu_1^2}{\kappa d\,\widehat{m}} \operatorname{Tr}\Big[(S + \tau I_d)\,(S + (\tau + \lambda/\widehat{m})\, I_d)^{-1}\Big] = O(\widehat{m}^{-1}), \tag{184}$$

and hence

$$\widehat{m} = \Theta(\gamma), \qquad \frac{\lambda}{\widehat{m}} = O(\gamma^{-1}). \tag{185}$$

Since $N = \gamma d N_0 = \gamma\kappa d^2$, we have,

$$\gamma^{-1} = O\left(\frac{d^2}{N}\right). \tag{186}$$

The large-γ limit of Eq. (182) is therefore

$$W_\infty := S(S + \tau I_d)^{-1}. \tag{187}$$

Write $\widehat{W} = W_{F,\infty} + \delta W_{\mathrm{det}} + \delta W_{\mathrm{fl}}$, where δW_{det} is the correction caused by $\lambda/\widehat{m}$, and δW_{fl} is the fluctuation (the last term of Eq. (182)). Equation (185) gives

$$\|\delta W_{\mathrm{det}}\|_{\mathrm{op}} = O\left(\frac{d^2}{N}\right). \tag{188}$$

Conditioned on S, the fluctuation δW_{fl} has zero mean. Moreover, using $\gamma N_0 = N/d$,

$$\begin{aligned} \frac{1}{d}\mathbb{E}_Z\|\delta W_{\mathrm{fl}}\|_F^2 &= \frac{R}{\gamma\mu_1^2 N_0} \operatorname{Tr}\Big[(S + \tau I_d)\,(S + (\tau + \lambda/\widehat{m})\, I_d)^{-2}\Big] \\ &= \frac{dR}{\mu_1^2 N} \operatorname{Tr}\Big[(S + \tau I_d)\,(S + (\tau + \lambda/\widehat{m})\, I_d)^{-2}\Big]. \end{aligned} \tag{189}$$

For fixed $\alpha > 0$, we have $\tau > 0$, so the normalized trace in Eq. (189) is $O(1)$. Furthermore,

$$R = \mathbb{E}\left[\left(\sigma_\star(G) + \sqrt{\Delta}\,\varepsilon - u_0\right)^2\right]. \tag{190}$$

As the large-γ solution converges to the finite population solution $(m_\infty, q_\infty, \bar{q}_\infty)$, we have

$$R \longrightarrow R_\infty := 1 + \Delta - 2\mu_1^2 m_\infty + \mu_1^2 q_\infty + \frac{1+\Delta}{\alpha}\bar{q}_\infty, \tag{191}$$

which is finite. Consequently,

$$\frac{1}{d}\mathbb{E}_Z\|\delta W_{\mathrm{fl}}\|_F^2 = O\left(\frac{d^2}{N}\right). \tag{192}$$

It follows that the order parameters differ from their infinite-pretraining values by $O(d^2/N)$:

$$\begin{aligned} m &= m_\infty + O\left(\frac{d^2}{N}\right), \\ q &= q_\infty + O\left(\frac{d^2}{N}\right), \\ \bar{q} &= \bar{q}_\infty + O\left(\frac{d^2}{N}\right), \\ \frac{1}{d}\mathbb{E}_Z \operatorname{Tr}\widehat{W}_F &= w_\infty + O\left(\frac{d^2}{N}\right), \end{aligned} \tag{193}$$

where

$$\begin{aligned} m_\infty &= \frac{1}{d}\operatorname{Tr}\left[S^2(S+\tau I_d)^{-1}\right], \\ q_\infty &= \frac{1}{d}\operatorname{Tr}\left[S^3(S+\tau I_d)^{-2}\right], \\ \bar{q}_\infty &= \frac{1}{d}\operatorname{Tr}\left[S^2(S+\tau I_d)^{-2}\right], \\ w_\infty &= \frac{1}{d}\operatorname{Tr}\left[S(S+\tau I_d)^{-1}\right]. \end{aligned} \tag{194}$$

The linear fluctuation terms vanish after averaging over Z, while their quadratic contribution is controlled by Eq. (189).

The Stage-II training at infinite pretraining has moments

$$C_{\mathrm{tr}}^{(\infty)} = \mu_1 m_\infty, \qquad V_{\mathrm{tr}}^{(\infty)} = \mu_1^2 q_\infty + \frac{1+\Delta}{\alpha}\bar{q}_\infty. \tag{195}$$

Accordingly, define the population-optimal readout

$$\widehat{\sigma}_\infty(v) := \mathbb{E}\left[\sigma_\star(G)\,\middle|\, C_{\mathrm{tr}}^{(\infty)} G + \sqrt{V_{\mathrm{tr}}^{(\infty)} - (C_{\mathrm{tr}}^{(\infty)})^2}\, Z = v\right]. \tag{196}$$

For memorization and generalization, respectively, the infinite-pretraining test moments are

$$\begin{aligned} C_{\mathrm{M}}^{(\infty)} &= \mu_1 m_\infty, \quad V_{\mathrm{M}}^{(\infty)} = \mu_1^2 q_\infty + \frac{\bar{q}_\infty}{\widetilde{\alpha}}, \\ C_{\mathrm{G}}^{(\infty)} &= \mu_1 w_\infty, \quad V_{\mathrm{G}}^{(\infty)} = \left(\mu_1^2 + \frac{1}{\widetilde{\alpha}}\right)\bar{q}_\infty. \end{aligned} \tag{197}$$

Hence we define

$$\boxed{g_{\mathcal{P}}(\alpha, \widetilde{\alpha}, S) := \mathbb{E}_{G,Z}\left[\left(\sigma_\star(G) - \widehat{\sigma}_\infty\left(C_{\mathcal{P}}^{(\infty)} G + \sqrt{V_{\mathcal{P}}^{(\infty)} - (C_{\mathcal{P}}^{(\infty)})^2}\, Z\right)\right)^2\right], \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}.} \tag{198}$$

For fixed positive $\alpha, \widetilde{\alpha}$, these Gaussian channels are nondegenerate. Their covariances, and hence the corresponding conditional expectation in Eq. (196), vary smoothly under the $O(d^2/N)$ perturbations in Eq. (193). Therefore

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\mathrm{feature}} = g_{\mathcal{P}}(\alpha_F, \widetilde{\alpha}_F, S_F) + O\left(\frac{d^2}{N}\right), \qquad \mathcal{P} \in \{\mathrm{M}, \mathrm{G}\}.} \tag{199}$$

This gives the feature-learner finite-pretraining correction.

Infinite task-pool limit and context-length scalings. We next set $\Delta = \lambda = 0$ and take first $\gamma \to \infty$ and then $\kappa \to \infty$. (187) gives

$$W_\infty = S\left(S + \frac{1}{\alpha\mu_1^2} I_d\right)^{-1}. \tag{200}$$

In the $\kappa \to \infty$ limit, isotropy of the task distribution implies $S \longrightarrow I_d$. Hence

$$W_\infty \longrightarrow r_\alpha I_d, \qquad r_\alpha := \frac{\alpha\mu_1^2}{1+\alpha\mu_1^2}. \tag{201}$$

The corresponding limiting order parameters therefore satisfy

$$m_\infty = r_\alpha, \qquad q_\infty = \bar{q}_\infty = r_\alpha^2, \qquad \frac{1}{d}\operatorname{Tr} W_\infty = r_\alpha. \tag{202}$$

Since the task pool is now infinite, the memorization and generalization have the same task distribution. Moreover, when $\widetilde{L} = L$, the Stage-II training channel and the noiseless inference channel coincide. Their covariance parameters are

$$C = \mu_1 r_\alpha, \qquad V = r_\alpha^2\left(\mu_1^2 + \frac{1}{\alpha}\right). \tag{203}$$

Consequently,

$$\frac{C^2}{V} = \frac{\alpha\mu_1^2}{1+\alpha\mu_1^2} = r_\alpha. \tag{204}$$

To evaluate the Stage-II error, standardize the attention output by writing $T := \frac{u}{\sqrt{V}}$, the pair (G, T) is jointly Gaussian with unit variances and correlation

$$\rho_\alpha := \mathbb{E}[GT] = \frac{C}{\sqrt{V}}. \tag{205}$$

By Eq. (204), $\rho_\alpha^2 = \frac{C^2}{V} = r_\alpha$. The optimal readout of Stage II is the conditional expectation

$$\widehat{\sigma}(u) = \mathbb{E}[\sigma_\star(G) \mid u] = \mathbb{E}[\sigma_\star(G) \mid T]. \tag{206}$$

Expand the teacher in the normalized Hermite basis,

$$\sigma_\star(G) = \sum_{k\geq 1} c_k H_k(G). \tag{207}$$

For two jointly standard Gaussian variables with correlation ρ_α, Mehler's identity implies

$$\mathbb{E}[H_k(G) \mid T] = \rho_\alpha^k H_k(T). \tag{208}$$

Therefore the readout is

$$\widehat{\sigma}(u) = \sum_{k\geq 1} c_k \rho_\alpha^k H_k\left(\frac{u}{\sqrt{V}}\right). \tag{209}$$

The test error is the conditional-mean squared error,

$$\begin{aligned} \mathcal{E}_{\mathcal{P}}^{\text{feature}} &= \mathbb{E}\Big[(\sigma_\star(G) - \mathbb{E}[\sigma_\star(G) \mid T])^2\Big] \\ &= \mathbb{E}[\sigma_\star(G)^2] - \mathbb{E}\big[\mathbb{E}[\sigma_\star(G) \mid T]^2\big]. \end{aligned} \tag{210}$$

Using the orthonormality of the Hermite polynomials,

$$\mathbb{E}[\sigma_\star(G)^2] = \sum_{k\geq 1} c_k^2 = 1, \tag{211}$$

and

$$\mathbb{E}\big[\mathbb{E}[\sigma_\star(G) \mid T]^2\big] = \sum_{k\geq 1} c_k^2 \rho_\alpha^{2k} = \sum_{k\geq 1} c_k^2 r_\alpha^k. \tag{212}$$

Hence

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\text{feature}} = 1 - \sum_{k\geq 1} c_k^2 r_\alpha^k, \qquad \mathcal{P} \in \{\text{M}, \text{G}\}.} \tag{213}$$

Although the replica calculation above was formulated with a fixed $\alpha = L/d$, (213) naturally allows us to examine d-dependent context scalings by substituting $\alpha_F(d) = \frac{L(d)}{d}$ before taking $d \to \infty$. Consider

$$L = \widetilde{L} = \alpha d^s, \qquad \alpha > 0, \quad s > 0. \tag{214}$$

Then $\alpha(d) = \alpha d^{s-1}$, and hence, using $\mu_1 = c_1$,

$$r_{\alpha(d)} = \frac{\alpha c_1^2 d^{s-1}}{1 + \alpha c_1^2 d^{s-1}} \longrightarrow \begin{cases} 0, & s < 1, \\ r_\alpha := \dfrac{\alpha c_1^2}{1 + \alpha c_1^2}, & s = 1, \\ 1, & s > 1. \end{cases} \tag{215}$$

Substituting this limit into Eq. (213) gives

$$\boxed{\mathcal{E}_{\mathcal{P}}^{\text{feature}} \longrightarrow \begin{cases} 1, & s < 1, \\ 1 - \displaystyle\sum_{k=1}^{\infty} c_k^2 r_\alpha^k, & s = 1, \\ 0, & s > 1, \end{cases} \qquad \mathcal{P} \in \{\text{M}, \text{G}\}.} \tag{216}$$

C Additional experiments

C.1 Consistency of the Replica method with numerical experiments

We first validate our analytical results from the replica method against numerical experiments. We use the same learned attention matrices and the same fixed-point solutions as in Section 4. All runs use the teacher in Eq. (23), $\Delta = 0.01$, and $\lambda = 10^{-3}$. In Figure 5 we compare the order parameters obtained from numerics against those obtained from replica under the following settings:

Panels	d	N	N_0	α	$\widetilde{\alpha}$
A–B	32	288 to 1200000	24	32	32
C–D	24	1200000	$24, 48, 96, 144, 192$	32	32

The plotted $m, q, \bar{q}$ are the corresponding learner's order parameters in Eqs. (15), (84), and (140), measured from the saved attention matrices. The additional quantity r is $d^{-1} \operatorname{Tr} \widehat{W}$ for the feature learner and $D^{-1} \operatorname{Tr} \widehat{W}_K$ for the kernel learner. The feature learner and linear attention share the same learned attention, and Stage II leaves these quantities unchanged. Figure 5 shows that the replica predictions capture the main trends of the measured order parameters, with visible deviations near the kernel interpolation peak.

C.2 Varying task diversity

The following experiment varies the task-diversity N_0 with $N = 1200000$ and $\alpha = \widetilde{\alpha} = 32$, and $N_0 \in \{24, 48, 96, 144, 192\}$. Increasing N_0 means a more diverse task pool and fewer training prompts per task.

Figure 6 shows that increasing task diversity narrows the gap between memorization and generalization for both learners. For the kernel learner, this improves performance on unseen tasks (generalization) at the cost of accuracy on the training tasks (memorization). For the feature learner, the main effect is an improvement in generalization.

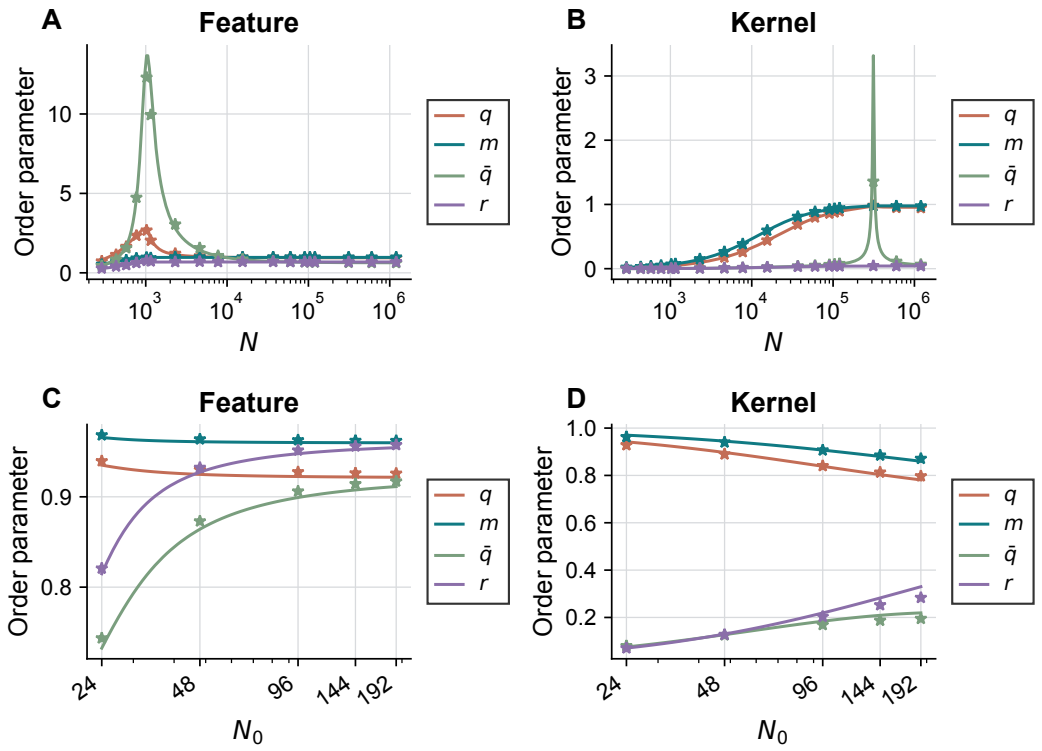

Figure 5: Consistency of the replica method with numerical experiments. Left: feature learner; right: kernel learner. The upper row varies N at $d = 32$, and the lower row varies N_0 at $d = 24$, $N = 1200000$. Solid lines are replica predictions, and stars show the order parameters measured from the learned attention matrices in one run. Red, teal, green, and purple denote $q, m, \bar{q}, r$, respectively, for the learner in that column.

The replica analysis also predicts this crossover. For the common teacher in Eq. (23), set $\Delta = \lambda = 0$, $\sigma_0(v) = v$, and $L = \widetilde{L} = \alpha d$. We first take $N \to \infty$, then $d, N_0 \to \infty$ at fixed $\kappa = N_0/d$, and finally $\alpha \to \infty$. In Appendix C.3, we show that the memorization errors satisfy

$$\mathcal{E}_M^F(\infty) = \frac{5\min(1,\kappa)}{3\alpha} + o(\alpha^{-1}), \quad \mathcal{E}_M^K(\infty) = \frac{\kappa}{\alpha} + o(\alpha^{-1}). \tag{217}$$

For $\kappa \geq 1$, which covers the scan in Fig. 6, the leading terms therefore give

$$\mathcal{E}_M^K(\infty) < \mathcal{E}_M^F(\infty) \quad \Longleftrightarrow \quad \kappa < \frac{5}{3}. \tag{218}$$

For $d = 24$, this gives $N_0 \simeq 40$, consistent with the crossover between $N_0 = 24$ and 48 in Figure 6.

C.3 Derivation of Eq. (217)

We set $\Delta = \lambda = 0$, $\sigma_0(v) = v$, and $L = \widetilde{L} = \alpha d$. As in Eq. (217), we first take $N \to \infty$, then $d, N_0 \to \infty$ with $N_0/d \to \kappa \in (0, \infty)$, and finally $\alpha \to \infty$. The common teacher in Eq. (23) is $\sigma_\star(G) = \frac{\sqrt{3}}{2} H_1(G) + \frac{1}{2} H_2(G)$, so $\mu_1^2 = 3/4$.

Feature learner. With $\tau_\alpha = 1/(\alpha\mu_1^2) = 4/(3\alpha)$, Eq. (194) gives

$$\begin{aligned} q_\infty + \tau_\alpha \bar{q}_\infty &= \int \frac{s^3 + \tau_\alpha s^2}{(s+\tau_\alpha)^2}\, \nu_\kappa(\mathrm{d}s) = \int \frac{s^2}{s+\tau_\alpha}\, \nu_\kappa(\mathrm{d}s) = m_\infty, \\ V_{\mathrm{M}}^{(\infty)} &= \mu_1^2 q_\infty + \frac{\bar{q}_\infty}{\alpha} = \mu_1^2 m_\infty, \qquad \frac{(C_{\mathrm{M}}^{(\infty)})^2}{V_{\mathrm{M}}^{(\infty)}} = \frac{\mu_1^2 m_\infty^2}{\mu_1^2 m_\infty} = m_\infty. \end{aligned} \tag{219}$$

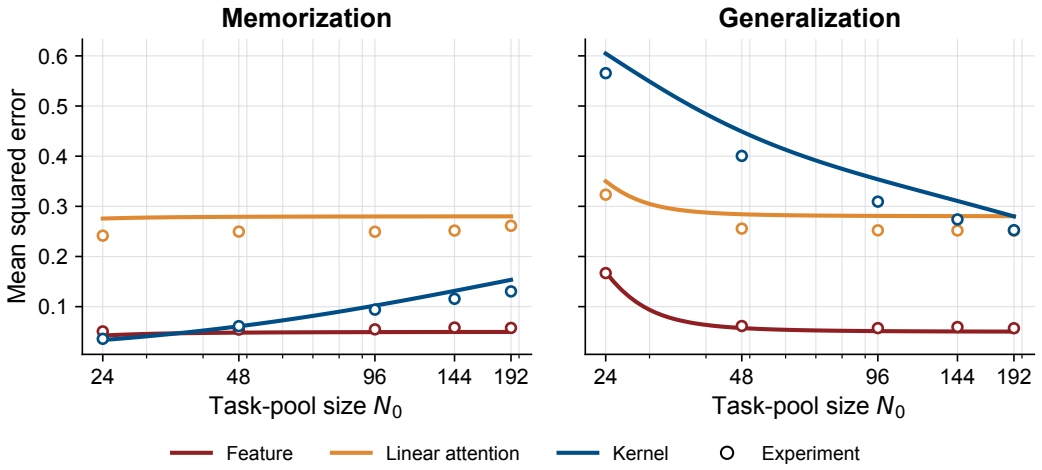

Figure 6: Memorization (left) and generalization (right) errors as a function of N_0. We fix $d = 24$, $D = 323$, $N = 1200000$, $L = \widetilde{L} = 768$, $\Delta = 0.01$, and $\lambda = 10^{-3}$. Red, orange, and blue denote the feature learner, linear attention, and the kernel learner, respectively. Solid lines show replica predictions, and circles show numerical results. The three learners use the same training and inference prompts at each N_0, with 32768 inference prompts per error estimate.

Since the Stage-II training and memorization distributions coincide, substituting this identity into Eq. (198) yields

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{feature}}(\infty) = \frac{3}{4}(1 - m_\infty) + \frac{1}{4}(1 - m_\infty^2) = \frac{5}{4}(1 - m_\infty) - \frac{1}{4}(1 - m_\infty)^2. \tag{220}$$

The Marchenko–Pastur spectrum satisfies $\int s\,\nu_\kappa(\mathrm{d}s) = 1$ and $\nu_\kappa((0,\infty)) = \min(1,\kappa)$. Therefore

$$\begin{aligned} 1 - m_\infty &= \int \left(s - \frac{s^2}{s+\tau_\alpha}\right) \nu_\kappa(\mathrm{d}s) = \tau_\alpha \int \frac{s}{s+\tau_\alpha}\,\nu_\kappa(\mathrm{d}s), \\ \lim_{\alpha\to\infty} \frac{1 - m_\infty}{\tau_\alpha} &= \int \mathbf{1}_{\{s>0\}}\,\nu_\kappa(\mathrm{d}s) = \min(1,\kappa). \end{aligned} \tag{221}$$

As $\alpha \to \infty$, $\tau_\alpha \to 0$, and the ratio $s/(s+\tau_\alpha)$ tends to 1 for every $s > 0$ and remains 0 at $s = 0$. Since this ratio always lies between 0 and 1, we can take its limit inside the integral, also when $\kappa = 1$. Combining Eqs. (220) and (221), we obtain

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{feature}}(\infty) = \frac{5}{4}\tau_\alpha \min(1,\kappa) + o(\tau_\alpha) = \frac{5\min(1,\kappa)}{3\alpha} + o(\alpha^{-1}). \tag{222}$$

Kernel learner. Since $\widetilde{\alpha}_K = \alpha_K$ and $\Delta = 0$, the two trace terms in Eq. (110) combine to give

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \frac{1}{\alpha_K D}\,\mathrm{Tr}\left[S_K \left(S_K + \frac{1}{\alpha_K} I_D\right)^{-1}\right]. \tag{223}$$

Let

$$G_K := \frac{1}{D}\Theta_K^\top \Theta_K, \qquad \Theta_K = \begin{bmatrix}\theta(a^1) & \cdots & \theta(a^{N_0})\end{bmatrix}. \tag{224}$$

The nonzero eigenvalues of G_K coincide with those of $(N_0/D)S_K$. Since $\frac{N_0}{L} \longrightarrow \frac{\kappa}{\alpha}$, Eq. (223) can therefore be written as

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \frac{N_0}{L}\,\frac{1}{N_0}\,\mathrm{Tr}\left[G_K \left(G_K + \frac{N_0}{L} I_{N_0}\right)^{-1}\right]. \tag{225}$$

For the common teacher, $c_1^2 = 3/4$ and $c_2^2 = 1/4$. By the spectral limit derived in Eq. (45), when $N_0/d \to \kappa$, the empirical spectral distribution of G_K converges to the law

$$g = \frac{1}{4} + \frac{3}{4}s, \qquad s \sim \nu_{1/\kappa}. \tag{226}$$

Hence, after taking $d, N_0 \to \infty$ at fixed κ,

$$\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \frac{\kappa}{\alpha} \int \frac{g}{g + \kappa/\alpha} \, \mu_\kappa(\mathrm{d}g) = \frac{\kappa}{\alpha} \left[1 - \frac{\kappa}{\alpha} \mathcal{R}_\kappa\Big(\frac{\kappa}{\alpha}\Big) \right], \tag{227}$$

where μ_κ denotes the law in Eq. (226) and

$$\mathcal{R}_\kappa(z) := \int \frac{\mu_\kappa(\mathrm{d}g)}{g + z}. \tag{228}$$

Since $g \geq 1/4$ almost surely,

$$0 \leq \mathcal{R}_\kappa(z) \leq 4, \qquad z \geq 0. \tag{229}$$

Therefore, for every fixed κ,

$$\boxed{\mathcal{E}_{\mathrm{M}}^{\mathrm{kernel}}(\infty) = \frac{\kappa}{\alpha} + O(\alpha^{-2}) = \frac{\kappa}{\alpha} + o(\alpha^{-1}).} \tag{230}$$

D A FORMAL STATEMENT OF THEOREM 1

In this section, we give a formal version of Theorem 1.

Population setting. For each fixed input dimension d and context length L, a prompt is generated by first sampling

$$a, x_0, x_1, \ldots, x_L \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\Big(\mathbb{S}^{d-1}(\sqrt{d})\Big), \tag{231}$$

and then setting

$$y_\ell = \sigma_\star\left(\frac{a^\top x_\ell}{\sqrt{d}}\right), \qquad \ell = 0, 1, \ldots, L. \tag{232}$$

Thus we study the problem with $\Delta = \lambda = 0$ and $\widetilde{L} = L$. The task vector is sampled independently for every prompt, and the training objectives are defined as population expectations. This is the population limit, corresponding to $\gamma \to \infty$ and $\kappa \to \infty$ in Theorem 1.

$\gamma \to \infty$ then implies that the learner minimizes the population loss, as we will detail below.

Let $G_d := \frac{a^\top x}{\sqrt{d}}$, where $a, x \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\Big(\mathbb{S}^{d-1}(\sqrt{d})\Big)$, and let $P_{k,d}$ denote the degree-k Gegenbauer polynomial ($k \geq 0$) normalized so that $\mathbb{E}[P_{j,d}(G_d) P_{k,d}(G_d)] = \delta_{jk}$. We write

$$v_d := \mathbb{E}[\sigma_\star(G_d)^2], \qquad c_{k,d} := \mathbb{E}[\sigma_\star(G_d) P_{k,d}(G_d)]. \tag{233}$$

Recall that the Gaussian Hermite coefficients are

$$c_k := \mathbb{E}[\sigma_\star(G) H_k(G)], \qquad G \sim \mathcal{N}(0,1). \tag{234}$$

Assumption 3 (Teacher regularity)**.** *The teacher $\sigma_\star : \mathbb{R} \to \mathbb{R}$ is a fixed measurable function, independent of d and L, satisfying*

$$\mathbb{E}[\sigma_\star(G)] = 0, \qquad \mathbb{E}[\sigma_\star(G)^2] = 1, \qquad c_1 := \mathbb{E}[G\sigma_\star(G)] \neq 0, \tag{235}$$

and

$$\mathbb{E}[|\sigma_\star(G)|^4] < \infty, \qquad G \sim \mathcal{N}(0,1). \tag{236}$$

Kernel learner. Fix a finite set of degrees $\mathcal{K} \subset \mathbb{N}_{\geq 1}$, independent of d, and let

$$h(x) = \bigoplus_{k \in \mathcal{K}} h_k(x) \in \mathbb{R}^D, \qquad D = \sum_{k \in \mathcal{K}} D_{k,d}, \tag{237}$$

where h_k is an orthonormal basis of the degree-k spherical harmonics and $D_{k,d} = \frac{2k+d-2}{k}\binom{k+d-3}{k-1}$. For a prompt define

$$H^{\mathrm{K}} := \frac{1}{L} h(x_0) \left(\sum_{\ell=1}^{L} y_\ell h(x_\ell) \right)^\top. \tag{238}$$

The population-trained attention matrix is

$$W_{d,L}^{\mathrm{K}} := \arg\min_{W \in \mathbb{R}^{D \times D}} \mathbb{E}\left[\Big(\mathrm{Tr}\big(W (H^{\mathrm{K}})^\top\big) - y_0 \Big)^2 \right], \tag{239}$$

where the expectation is over the prompt in Eq. (232). We define its population error by

$$\mathcal{E}_{d,L}^{\mathrm{kernel}} := \mathbb{E}\left[\Big(y_0 - \mathrm{Tr}\big(W_{d,L}^{\mathrm{K}} (H^{\mathrm{K}})^\top\big) \Big)^2 \right]. \tag{240}$$

Feature learner. For the feature learner, define

$$H^{\mathrm{F}} := \frac{1}{L} x_0 \left(\sum_{\ell=1}^{L} y_\ell x_\ell \right)^\top, \qquad U_W := \mathrm{Tr}\big(W (H^{\mathrm{F}})^\top\big). \tag{241}$$

We use the identity readout in Stage I, $\sigma_0(u) = u$, and define the populational attention matrix by

$$W_{d,L}^{\mathrm{F}} := \arg\min_{W \in \mathbb{R}^{d\times d}} \mathbb{E}\left[(U_W - y_0)^2\right]. \tag{242}$$

In Stage II, the attention matrix is frozen and the scalar readout is optimized over all measurable functions ϕ satisfying $\mathbb{E}[\phi(U_{W_{d,L}^{\mathrm{F}}})^2] < \infty$:

$$\widehat{\sigma}_{d,L} \in \arg\min_{\phi} \mathbb{E}\left[\left(\phi(U_{W_{d,L}^{\mathrm{F}}}) - y_0\right)^2\right]. \tag{243}$$

The resulting population test error is

$$\mathcal{E}_{d,L}^{\mathrm{feature}} := \mathbb{E}\left[\left(y_0 - \widehat{\sigma}_{d,L}\left(U_{W_{d,L}^{\mathrm{F}}}\right)\right)^2\right]. \tag{244}$$

We now state the rigorous version of the context scaling.

Theorem 2. *Suppose Assumption 3 holds. Define $\alpha_{d,L} := \frac{L}{d}$ and $r_\alpha := \frac{\alpha c_1^2}{1+\alpha c_1^2}$. Then the feature learner satisfies*

$$\mathcal{E}_{d,L}^{\mathrm{feature}} = 1 - \sum_{k\geq 1} c_k^2 r_{\alpha_{d,L}}^k + o_{d,L}(1). \tag{245}$$

For the kernel learner, let $\nu_{\mathrm{in}} := \sum_{k\in\mathcal{K}} c_k^2$ and $\nu_{\mathrm{out}} := 1 - \nu_{\mathrm{in}}$. Then

$$\mathcal{E}_{d,L}^{\mathrm{kernel}} = \nu_{\mathrm{out}} + \sum_{k\in\mathcal{K}} c_k^2 \frac{D_{k,d}}{c_k^2 L + D_{k,d}} + o_{d,L}(1). \tag{246}$$

Here $o_{d,L}(1)$ denotes a remainder that converges to zero along every sequence (d_n, L_n) satisfying $d_n \longrightarrow \infty$ and $L_n \longrightarrow \infty$.

E Proof of Theorem 2

E.1 Gegenbauer polynomials

We first recall some basic properties of the Gegenbauer polynomials. Define

$$Q_{k,d}(q) := \frac{P_{k,d}(\sqrt{d}\, q)}{\sqrt{D_{k,d}}}, \qquad -1 \leq q \leq 1 \tag{247}$$

for $k \geq 0$, where we choose $D_{0,d} := 1$, $P_{0,d}(t) \equiv 1$ and $P_{1,d}(t) = t$.

Lemma 1. *Let $e, b \in \mathbb{R}^d$ be deterministic unit vectors, let $q = e^\top b$, and let $x \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Set*

$$g = e^\top x, \qquad s = b^\top x. \tag{248}$$

Then the following statements hold.

(i) Spherical projection formula. *For every $k \geq 0$,*

$$\mathbb{E}[P_{k,d}(g) \mid s] = Q_{k,d}(q) P_{k,d}(s), \qquad |Q_{k,d}(q)| \leq 1. \tag{249}$$

(ii) Gegenbauer limits. *For every fixed k,*

$$\lim_{d\to\infty} \sup_{|q|\leq 1} |Q_{k,d}(q) - q^k| = 0. \tag{250}$$

Moreover, for every fixed M, there exists a constant $C_M < \infty$ such that, for all sufficiently large d,

$$\max_{0\leq k\leq M} |P_{k,d}(t)| \leq C_M(1 + |t|^M), \qquad t \in \mathbb{R}, \tag{251}$$

and for every fixed $A < \infty$,

$$\sup_{d \text{ sufficiently large}} \max_{0 \le k \le M} \sup_{|t| \le A} \left(|P_{k,d}(t)| + |P'_{k,d}(t)| \right) < \infty. \tag{252}$$

(iii) Convergence of the teacher coefficients. *Suppose Assumption 3 holds. Then*

$$v_d = \mathbb{E}[\sigma_\star(G_d)^2] \longrightarrow \mathbb{E}[\sigma_\star(G)^2] = 1, \tag{253}$$

and, for every fixed $k \ge 0$,

$$c_{k,d} = \mathbb{E}[\sigma_\star(G_d) P_{k,d}(G_d)] \longrightarrow \mathbb{E}[\sigma_\star(G) H_k(G)] = c_k. \tag{254}$$

Furthermore,

$$\lim_{d \to \infty} \sum_{k \ge 0} |c_{k,d}^2 - c_k^2| = 0, \ \lim_{M \to \infty} \limsup_{d \to \infty} \sum_{k > M} c_{k,d}^2 = 0. \tag{255}$$

Proof. The identity (249) is the standard formula for spherical harmonics, equivalently a consequence of the spherical addition theorem (Frye & Efthimiou, 2012). Since $P_{k,d}(g)$ is normalized to have unit second moment and $P_{k,d}(s)$ has the same distribution,

$$Q_{k,d}(q)^2 = \mathbb{E}\big[\mathbb{E}[P_{k,d}(g) \mid s]^2\big] \le \mathbb{E}[P_{k,d}(g)^2] = 1. \tag{256}$$

This proves (i).

For every fixed k, the normalized Gegenbauer polynomial converges, coefficientwise, to the normalized Hermite polynomial under $d \to \infty$, while $Q_{k,d}$ converges uniformly on $[-1, 1]$ to q^k; see, e.g., Mei et al. (2021). This proves (250). Since the degree is fixed, coefficientwise convergence immediately implies (251) and (252).

We next prove the statements involving the teacher. The random variable G_d has density

$$p_d(t) = \frac{\Gamma(d/2)}{\sqrt{\pi d}\, \Gamma((d-1)/2)} \left(1 - \frac{t^2}{d}\right)^{(d-3)/2} \mathbf{1}_{\{|t| \le \sqrt{d}\}}. \tag{257}$$

For every fixed t,

$$p_d(t) \longrightarrow \phi(t) := \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{258}$$

Moreover, there exists a universal constant $C < \infty$ such that, for all sufficiently large d,

$$p_d(t) \le C\phi(t), \qquad t \in \mathbb{R}. \tag{259}$$

Indeed, the prefactor in (257) remains uniformly bounded, and for $|t| \le \sqrt{d}$,

$$\left(1 - \frac{t^2}{d}\right)^{(d-3)/2} \le \exp\left(-\frac{d-3}{2d} t^2\right) \le e^{3/2} e^{-t^2/2}. \tag{260}$$

Since Assumption 3 gives $\mathbb{E}|\sigma_\star(G)|^4 < \infty$, (259) implies, in particular,

$$\sup_{d \text{ sufficiently large}} \mathbb{E}|\sigma_\star(G_d)|^4 < \infty. \tag{261}$$

By dominated convergence,

$$v_d = \int \sigma_\star(t)^2 p_d(t)\, \mathrm{d}t \longrightarrow \int \sigma_\star(t)^2 \phi(t)\, \mathrm{d}t = 1, \tag{262}$$

which proves (253).

Fix k. By Mei et al. (2021), $P_{k,d}(t) \to H_k(t)$ pointwise, and for all sufficiently large d,

$$|P_{k,d}(t)| \le C_k(1 + |t|^k). \tag{263}$$

Together with (259), this yields

$$|\sigma_\star(t) P_{k,d}(t) p_d(t)| \le C_k |\sigma_\star(t)| (1 + |t|^k) \phi(t). \tag{264}$$

The right-hand side is integrable by Hölder's inequality and Assumption 3. Dominated convergence therefore gives

$$c_{k,d} = \int \sigma_\star(t) P_{k,d}(t) p_d(t)\,\mathrm{d}t \longrightarrow \int \sigma_\star(t) H_k(t)\phi(t)\,\mathrm{d}t = c_k, \tag{265}$$

proving (254).

Finally, Parseval's identity gives

$$\sum_{k\geq 0} c_{k,d}^2 = v_d, \qquad \sum_{k\geq 0} c_k^2 = 1. \tag{266}$$

Thus, $\lim_{M\to\infty} \limsup_{d\to\infty} \sum_{k>M} c_{k,d}^2 = 0$. Moreover,

$$\sum_{k\geq 0} |c_{k,d}^2 - c_k^2| \leq \sum_{k=0}^{M} |c_{k,d}^2 - c_k^2| + \sum_{k>M} c_{k,d}^2 + \sum_{k>M} c_k^2. \tag{267}$$

First take $d \to \infty$ at fixed M, using (254), and then let $M \to \infty$. This proves (255). □

From now on Assumption 3 is in force. All limits are taken along an arbitrary deterministic sequence (d, L) such that $d \to \infty$ and $L \to \infty$. We work with sufficiently large d, so that $d \geq 3$, $v_d > 0$, and

$$\mu_d := \mathbb{E}[\sigma_\star(G_d) G_d] = c_{1,d} \neq 0. \tag{268}$$

These properties follow from $v_d \to 1$ and $c_{1,d} \to c_1 \neq 0$. Constants denoted by C may depend on the teacher and its uniform moment bound, but not on d or L. For a scalar random variable Z, write $\|Z\|_{L^2} := (\mathbb{E}[Z^2])^{1/2}$.

E.2 Kernel learner

Define

$$M_d := \bigoplus_{k\in\mathcal{K}} \frac{c_{k,d}^2}{D_{k,d}} I_{D_{k,d}}, \qquad A_{d,L} := \frac{L-1}{L} M_d + \frac{v_d}{L} I_D. \tag{269}$$

Let us evaluate all the terms in (239). Write

$$m_d(a) := \mathbb{E}[y h(x) \mid a] = \bigoplus_{k\in\mathcal{K}} \frac{c_{k,d}}{\sqrt{D_{k,d}}} h_k(a), \qquad V^{\mathrm{K}} := \frac{1}{L} \sum_{\ell=1}^{L} y_\ell h(x_\ell). \tag{270}$$

The first identity follows from the spherical harmonic expansion of the teacher. Averaging over the independent task gives $\mathbb{E}_a[m_d(a) m_d(a)^\top] = M_d$. For fixed x, $a^\top x/\sqrt{d}$ has the law of G_d. Hence,

$$\mathbb{E}_{a,x}[y^2 h(x) h(x)^\top] = v_d \mathbb{E}_x[h(x) h(x)^\top] = v_d I_D. \tag{271}$$

Conditional independence of the demonstrations now yields

$$\mathbb{E}[V^{\mathrm{K}} (V^{\mathrm{K}})^\top] = A_{d,L}. \tag{272}$$

Also, $\mathbb{E}[y_0 h(x_0) (V^{\mathrm{K}})^\top] = M_d$. Thus the kernel population risk in (239) can be written as

$$R_{\mathrm{K}}(W) = v_d - 2\,\mathrm{Tr}(W M_d) + \mathrm{Tr}(W A_{d,L} W^\top). \tag{273}$$

Since $A_{d,L} \succeq (v_d/L) I_D \succ 0$, completing the square gives the unique minimizer $W_{d,L}^{\mathrm{K}} = M_d A_{d,L}^{-1}$ and

$$\min_W R_{\mathrm{K}}(W) = v_d - \mathrm{Tr}(M_d A_{d,L}^{-1} M_d). \tag{274}$$

Evaluating the trace gives

$$\mathcal{E}_{d,L}^{\mathrm{kernel}} = v_d - \sum_{k\in\mathcal{K}} \frac{L c_{k,d}^4}{v_d D_{k,d} + (L-1) c_{k,d}^2}. \tag{275}$$

For a fixed $k \in \mathcal{K}$, write $s_{k,d} = c_{k,d}^2$, $s_k = c_k^2$, and $n_k = D_{k,d}$. Since $0 \le s_{k,d} \le v_d$,

$$\begin{aligned} 0 \le & \frac{L s_{k,d}^2}{v_d n_k + (L-1) s_{k,d}} - \frac{L s_{k,d}^2}{v_d n_k + L s_{k,d}} \\ & = \frac{L s_{k,d}^3}{(v_d n_k + (L-1) s_{k,d})(v_d n_k + L s_{k,d})} \le \frac{v_d}{n_k}. \end{aligned} \tag{276}$$

For $u \ge 0$, $v > 0$, and $t \ge 0$, define $F(u, v; t) := \frac{tu^2}{v+tu}$. Its derivatives satisfy

$$0 \le \partial_u F = 1 - \frac{v^2}{(v+tu)^2} \le 1, \qquad |\partial_v F| = \frac{tu^2}{(v+tu)^2} \le \frac{u}{4v}. \tag{277}$$

Because $v_d \to 1$ and $s_{k,d} \to s_k$ (Lemma 1), these bounds imply

$$\sup_{L \ge 1} \left| \frac{L s_{k,d}^2}{v_d n_k + L s_{k,d}} - \frac{L s_k^2}{n_k + L s_k} \right| \longrightarrow 0. \tag{278}$$

Every $k \in \mathcal{K}$ is positive, so $n_k \to \infty$. Since $\mathcal{K}$ is fixed and finite, Eqs. (275) and (278) give

$$\begin{aligned} \mathcal{E}_{d,L}^{\text{kernel}} &= 1 - \sum_{k \in \mathcal{K}} \frac{L c_k^4}{D_{k,d} + L c_k^2} + o_{d,L}(1) \\ &= \nu_{\text{out}} + \sum_{k \in \mathcal{K}} c_k^2 \frac{D_{k,d}}{D_{k,d} + L c_k^2} + o_{d,L}(1). \end{aligned} \tag{279}$$

This finishes the proof of the kernel part of Theorem 2.

E.3 Feature learner Stage-I

Now we evaluate the Stage-I population risk (242). Define $V_{d,L} := \frac{1}{L} \sum_{\ell=1}^{L} y_\ell x_\ell$ so that $U_W = x_0^\top W V_{d,L}$. Let $e := a/\sqrt{d}$. By rotational invariance,

$$\mathbb{E}[yx \mid a] = \mu_d e, \qquad \mathbb{E}_a[ee^\top] = \frac{1}{d} I_d, \tag{280}$$

where $\mu_d := \mathbb{E}[\sigma_\star(G_d) G_d]$. Since the query and the demonstrations are conditionally independent given a,

$$\mathbb{E}[y_0 x_0 V_{d,L}^\top] = \mu_d^2 \mathbb{E}_a[ee^\top] = \frac{\mu_d^2}{d} I_d. \tag{281}$$

Moreover, using the conditional independence of the demonstrations,

$$\begin{aligned} \mathbb{E}[V_{d,L} V_{d,L}^\top] &= \frac{L-1}{L} \mathbb{E}_a\left[\mathbb{E}[yx \mid a] \mathbb{E}[yx \mid a]^\top\right] + \frac{1}{L} \mathbb{E}[y^2 xx^\top] \\ &= \left(\frac{L-1}{L} \frac{\mu_d^2}{d} + \frac{v_d}{L} \right) I_d, \end{aligned} \tag{282}$$

where the last identity follows from rotational invariance and $\mathbb{E}[y^2 xx^\top] = v_d I_d$.

Therefore the Stage-I population risk (242) is

$$\mathbb{E}[(x_0^\top W V_{d,L} - y_0)^2] = v_d - \frac{2\mu_d^2}{d} \operatorname{Tr}(W) + \left(\frac{L-1}{L} \frac{\mu_d^2}{d} + \frac{v_d}{L} \right) \|W\|_F^2. \tag{283}$$

Thus the left side of (283) is strictly convex, and it has a unique minimizer

$$W_{d,L}^{\text{F}} = w_{d,L} I_d, \qquad w_{d,L} = \frac{L \mu_d^2}{d v_d + (L-1) \mu_d^2}. \tag{284}$$

For all sufficiently large d, $v_d > 0$ and $\mu_d \neq 0$, and hence $w_{d,L} > 0$.

Consequently, define $T_{d,L} := x_0^\top V_{d,L}$ and then the final risk is

$$\mathcal{E}_{d,L}^{\text{feature}} = \inf_{\phi} \mathbb{E}[(y_0 - \phi(T_{d,L}))^2], \tag{285}$$

where the infimum is over measurable functions with finite second moment under the law of $T_{d,L}$. Now our goal is to evaluate this infimum.

E.4 Concentration of the context

Set $e := a/\sqrt{d}$ and define quantities

$$B_{d,L} := \left(\mu_d^2 + \frac{dv_d}{L}\right)^{1/2}, \qquad \rho_{d,L} := \frac{\mu_d}{B_{d,L}} \in [-1, 1]. \tag{286}$$

For the context $V_{d,L}$, write

$$R := \|V_{d,L}\|, \qquad b := \begin{cases} V_{d,L}/R, & R > 0, \\ e, & R = 0, \end{cases} \qquad q := e^\top b, \qquad S := b^\top x_0. \tag{287}$$

In particular, $T_{d,L} = RS$, including on $\{R = 0\}$. Let $\mathcal{C} := \sigma(a, x_1, \ldots, x_L)$ be the sigma-algebra generated by the task and demonstrations.

Lemma 2. *There is a constant $C < \infty$ such that, for all sufficiently large d and every $L \geq 1$,*

$$\mathbb{E}\left[\left(\frac{R}{B_{d,L}} - 1\right)^2\right] + \mathbb{E}[(q - \rho_{d,L})^2] \leq C\left(\frac{1}{L} + \frac{1}{d}\right). \tag{288}$$

Consequently,

$$\mathbb{E}\left[\left(\frac{T_{d,L}}{B_{d,L}} - S\right)^2\right] = \mathbb{E}\left[\left(\frac{R}{B_{d,L}} - 1\right)^2\right] \longrightarrow 0. \tag{289}$$

Proof. Condition on a, and set

$$\eta_\ell := y_\ell x_\ell - \mu_d e, \qquad \bar{\eta} := \frac{1}{L}\sum_{\ell=1}^{L} \eta_\ell. \tag{290}$$

The vectors η_ℓ are conditionally independent and centered, and $V_{d,L} = \mu_d e + \bar{\eta}$. Let

$$t_d := \mathbb{E}[\sigma_\star(G_d)^2 G_d^2], \qquad \beta_d := \frac{dv_d - t_d}{d-1}. \tag{291}$$

The conditional covariance is given by

$$\Gamma_d(a) := \mathbb{E}[\eta_1 \eta_1^\top \mid a] = (t_d - \mu_d^2) ee^\top + \beta_d(I_d - ee^\top). \tag{292}$$

Assumption 3 implies

$$\|\Gamma_d(a)\|_{\mathrm{op}} \leq C, \qquad \mathrm{Tr}(\Gamma_d(a)^2) \leq Cd, \qquad \mathbb{E}[\|\eta_1\|^4 \mid a] \leq Cd^2, \tag{293}$$

uniformly in a. Indeed, t_d and v_d are bounded, $0 \leq \beta_d \leq dv_d/(d-1)$, and $\|y_1 x_1\|^4 = d^2 |y_1|^4$. Also,

$$\mathrm{Tr}\,\Gamma_d(a) = dv_d - \mu_d^2, \qquad \mathbb{E}[(e^\top V_{d,L} - \mu_d)^2 \mid a] = \frac{t_d - \mu_d^2}{L} \leq \frac{C}{L}. \tag{294}$$

Independence and centering give

$$\begin{aligned} \mathbb{E}[\|\bar{\eta}\|^2 \mid a] &= \frac{\mathrm{Tr}\,\Gamma_d(a)}{L}, \\ \mathrm{Var}(\|\bar{\eta}\|^2 \mid a) &= \frac{\mathrm{Var}(\|\eta_1\|^2 \mid a)}{L^3} + \frac{2(L-1)}{L^3}\,\mathrm{Tr}(\Gamma_d(a)^2) \leq C\left(\frac{d^2}{L^3} + \frac{d}{L^2}\right), \end{aligned} \tag{295}$$

where for the second line we use the fact that $\mathbb{E}[(\eta_\ell^\top \eta_j)^2 \mid a] = \mathrm{Tr}(\Gamma_d(a)^2)$ for $\ell \neq j$.

Since

$$R^2 = \mu_d^2 + 2\mu_d e^\top \bar{\eta} + \|\bar{\eta}\|^2, \qquad \mathbb{E}[R^2 \mid a] = B_{d,L}^2 - \frac{\mu_d^2}{L}, \tag{296}$$

(294)–(295) imply

$$\mathbb{E}[(R^2 - B_{d,L}^2)^2] \leq C\left(\frac{1}{L} + \frac{d}{L^2} + \frac{d^2}{L^3}\right). \tag{297}$$

For sufficiently large d, $B_{d,L}^2 \geq c(1 + d/L)$ for a fixed $c > 0$. Dividing by $B_{d,L}^4$ gives

$$\mathbb{E}\left[\left(\frac{R^2}{B_{d,L}^2} - 1\right)^2\right] \leq C\left(\frac{1}{L} + \frac{1}{d}\right). \tag{298}$$

The same bound for $R/B_{d,L} - 1$ follows from $|\sqrt{u} - 1| \leq |u - 1|$, $u \geq 0$.

On $\{R/B_{d,L} \geq 1/2\}$,

$$|q - \rho_{d,L}| \leq \frac{2|e^\top V_{d,L} - \mu_d|}{B_{d,L}} + 2\left|\frac{R}{B_{d,L}} - 1\right|. \tag{299}$$

On the complementary event, $|q - \rho_{d,L}| \leq 2$, and that event has probability at most $4\mathbb{E}[(R/B_{d,L} - 1)^2]$. Combining these facts with (294) proves (288).

Finally, b is a $\mathcal{C}$-measurable unit vector, whereas x_0 is independent of $\mathcal{C}$ and uniform on the sphere. Thus S is independent of $\mathcal{C}$ and $\mathbb{E}[S^2 \mid \mathcal{C}] = 1$. Using $T_{d,L} = RS$ proves (289). □

E.5 Feature learner Stage-II

The following two lemmas prove that the regression (285) on $T_{d,L}$ approximately give the oracle posterior $\mathbb{E}[y_0 \mid \mathcal{C}, S]$.

Lemma 3. *Define*

$$Z_{d,L} := \mathbb{E}[y_0 \mid \mathcal{C}, S], \qquad \mathcal{R}_{d,L}^{\mathrm{or}} := \mathbb{E}[(y_0 - Z_{d,L})^2]. \tag{300}$$

Then, with convergence of the series in L^2,

$$Z_{d,L} = \sum_{k \geq 0} c_{k,d} Q_{k,d}(q) P_{k,d}(S), \tag{301}$$

and

$$\mathcal{R}_{d,L}^{\mathrm{or}} = v_d - \sum_{k \geq 0} c_{k,d}^2 \mathbb{E}[Q_{k,d}(q)^2]. \tag{302}$$

Moreover,

$$\mathcal{E}_{d,L}^{\mathrm{feature}} - \mathcal{R}_{d,L}^{\mathrm{or}} = \inf_{\phi} \|Z_{d,L} - \phi(T_{d,L})\|_{L^2}^2 \geq 0. \tag{303}$$

Proof. Conditionally on $\mathcal{C}$, the unit vectors e, b are fixed and the query remains uniform on the sphere. Applying Lemma 1 to $y_0 = \sum_{k \geq 0} c_{k,d} P_{k,d}(e^\top x_0)$ proves (301). Orthonormality of $P_{k,d}(S)$ yields

$$\mathbb{E}[Z_{d,L}^2] = \sum_{k \geq 0} c_{k,d}^2 \mathbb{E}[Q_{k,d}(q)^2], \tag{304}$$

which gives (302). Since $T_{d,L} = RS$ is measurable with respect to $\mathcal{C} \vee \sigma(S)$, every ϕ satisfies

$$\mathbb{E}[(y_0 - \phi(T_{d,L}))^2] = \mathcal{R}_{d,L}^{\mathrm{or}} + \|Z_{d,L} - \phi(T_{d,L})\|_{L^2}^2. \tag{305}$$

Taking the infimum and using (285) proves (303). □

Lemma 4. *Along every sequence $d \to \infty$, $L \to \infty$,*

$$\inf_{\phi} \|Z_{d,L} - \phi(T_{d,L})\|_{L^2}^2 \longrightarrow 0. \tag{306}$$

Proof. For brevity, write $B = B_{d,L}$, $\rho = \rho_{d,L}$, $T = T_{d,L}$, and $Z = Z_{d,L}$. For a fixed integer $M \geq 1$, define

$$Z_M := \sum_{k=0}^{M} c_{k,d} Q_{k,d}(q) P_{k,d}(S), \qquad p_M(s) := \sum_{k=0}^{M} c_{k,d} \rho^k P_{k,d}(s). \tag{307}$$

Orthonormality and $|Q_{k,d}| \le 1$ give

$$\|Z - Z_M\|_{L^2}^2 \le \sum_{k>M} c_{k,d}^2. \tag{308}$$

For every fixed $k \ge 1$, Lemmas 1 and 2 give

$$\mathbb{E}[(Q_{k,d}(q) - \rho^k)^2] \le 2 \sup_{|u|\le 1} |Q_{k,d}(u) - u^k|^2 + 2k^2 \mathbb{E}[(q-\rho)^2] \longrightarrow 0. \tag{309}$$

For $k = 0$ the left side is zero. As q is $\mathcal{C}$-measurable and S is independent of $\mathcal{C}$, it follows that

$$\|Z_M - p_M(S)\|_{L^2}^2 = \sum_{k=0}^{M} c_{k,d}^2 \mathbb{E}[(Q_{k,d}(q) - \rho^k)^2] \longrightarrow 0. \tag{310}$$

It remains to replace S by the observable T/B. We use a truncation argument. For $A \ge 1$, define

$$\pi_A(t) := \max\{-A, \min\{t, A\}\}, \qquad \phi_{M,A,d,L}(t) := p_M(\pi_A(t/B)). \tag{311}$$

By Lemma 1, $|p_M(s)| \le C_M(1 + |s|^M)$, and using

$$\mathbb{E}[G_d^{2m}] = (2m-1)!! \prod_{j=0}^{m-1} \frac{d}{d+2j} \le (2m-1)!!, \qquad \mathbb{E}[G_d^{2m-1}] = 0, \tag{312}$$

we obtain

$$\|p_M(S) - p_M(\pi_A(S))\|_{L^2}^2 \le C_M \mathbb{E}[(1 + |S|^{2M}) \mathbf{1}_{\{|S|>A\}}] \le \frac{C_M}{A^2}. \tag{313}$$

For fixed M, A, the function $p_M \circ \pi_A$ is globally Lipschitz with a constant bounded uniformly in d, L by Lemma 1. Therefore Lemma 2 gives

$$\|p_M(\pi_A(S)) - \phi_{M,A,d,L}(T)\|_{L^2}^2 \le C_{M,A} \mathbb{E}[(S - T/B)^2] \longrightarrow 0. \tag{314}$$

Combining (308), (310), (313), and (314) by the triangle inequality shows that, for every fixed M, A,

$$\limsup_{d,L\to\infty} \inf_{\phi} \|Z - \phi(T)\|_{L^2} \le \left(\sum_{k>M} c_k^2 \right)^{1/2} + \frac{C_M}{A}. \tag{315}$$

First let $A \to \infty$ for fixed M, and then let $M \to \infty$. This proves (306). □

E.6 Completion of the proof

Now we prove Theorem 2 for the feature learner.

Lemmas 3 and 4 give

$$\mathcal{E}_{d,L}^{\text{feature}} = v_d - \sum_{k\ge 0} c_{k,d}^2 \mathbb{E}[Q_{k,d}(q)^2] + o_{d,L}(1). \tag{316}$$

For every fixed k, (309) and the bounds $|Q_{k,d}|, |\rho_{d,L}| \le 1$ imply

$$\mathbb{E}[Q_{k,d}(q)^2] - \rho_{d,L}^{2k} \longrightarrow 0. \tag{317}$$

Using Lemma 1, we obtain

$$\sum_{k\ge 0} c_{k,d}^2 \mathbb{E}[Q_{k,d}(q)^2] = \sum_{k\ge 1} c_k^2 \rho_{d,L}^{2k} + o_{d,L}(1), \tag{318}$$

where we use the fact that the tails are controlled by the summable coefficients (Lemma 1) and $c_{0,d} \to c_0 = 0$.

It remains to replace $\rho_{d,L}^2$ by $r_{L/d}$. Set

$$b_d := \frac{\mu_d^2}{v_d} \longrightarrow b_\star := c_1^2 > 0, \qquad \alpha := \frac{L}{d}. \tag{319}$$

Then

$$\rho_{d,L}^2 = \frac{\alpha b_d}{1+\alpha b_d}, \qquad r_\alpha = \frac{\alpha b_\star}{1+\alpha b_\star}. \tag{320}$$

The derivative bound

$$\sup_{\alpha\geq 0} \frac{\partial}{\partial b} \frac{\alpha b}{1+\alpha b} = \sup_{\alpha\geq 0} \frac{\alpha}{(1+\alpha b)^2} = \frac{1}{4b}, \qquad b > 0, \tag{321}$$

implies

$$\sup_{\alpha\geq 0} \left| \frac{\alpha b_d}{1+\alpha b_d} - \frac{\alpha b_\star}{1+\alpha b_\star} \right| \leq \frac{|b_d - b_\star|}{4\min\{b_d, b_\star\}} \longrightarrow 0. \tag{322}$$

The function

$$\Psi(t) := \sum_{k\geq 1} c_k^2 t^k, \qquad 0 \leq t \leq 1, \tag{323}$$

is continuous, and hence uniformly continuous, because its defining series converges uniformly on $[0,1]$. Thus (322) gives

$$\sum_{k\geq 1} c_k^2 \rho_{d,L}^{2k} = \sum_{k\geq 1} c_k^2 r_\alpha^k + o_{d,L}(1). \tag{324}$$

Combining this identity with $v_d \to 1$, (316), and (318), we obtain

$$\mathcal{E}_{d,L}^{\text{feature}} = 1 - \sum_{k\geq 1} c_k^2 r_\alpha^k + o_{d,L}(1), \tag{325}$$

which finishes the proof.

F Declaration of LLM usage

We used LLMs in the following stages of the preparation of this work: editing (e.g., grammar, spelling, word choice), drafting sections of the paper, assisting in the writing of proofs, facilitating or running experiments, visualizing results for submission. All LLM-generated material used in this manuscript was manually verified by the authors to ensure that it performs as intended, in particular with regard to the derivation, proof and the numerical experiments.