Learning Linear Regression with Low-Rank Tasks In-Context

Kaito Takanami
The University of Tokyo

Takashi Takahashi
The University of Tokyo
RIKEN AIP

Yoshiyuki Kabashima
The University of Tokyo

Abstract

In-context learning (ICL) is a key building block of modern large language models, yet its theoretical mechanisms remain poorly understood. It is particularly mysterious how ICL operates in real-world applications where tasks have a common structure. In this work, we address this problem by analyzing a linear attention model trained on low-rank regression tasks. Within this setting, we precisely characterize the distribution of predictions and the generalization error in the high-dimensional limit. Moreover, we find that statistical fluctuations in finite pre-training data induce an implicit regularization. Finally, we identify a sharp phase transition of the generalization error governed by task structure. These results provide a framework for understanding how transformers learn to learn the task structure.

1 INTRODUCTION

A defining characteristic of modern large-scale language models (LLMs) is their capacity for in-context learning (ICL). This allows them to perform new tasks by conditioning on a few examples provided in context, without requiring any parameter updates. This capability has led to impressive performance on a wide array of tasks, including few-shot classification [Brown et al., 2020], complex reasoning [Ahn et al., 2024] and code generation [Wang and Chen, 2023, Zhong and Wang, 2024]. The surprising effectiveness of ICL has naturally attracted significant theoretical interest in uncovering underlying principles.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

However, relatively few studies of ICL go beyond settings with independent tasks to consider structured families of tasks [Oko et al., 2024]. In practice, tasks are often not independent but share underlying structures, as widely recognized in multi-task learning (MTL) [Caruana, 1997]. In this setting, assuming a shared low-rank structure among tasks [Han and Zhang, 2016] has led to effective algorithms across domains such as computer vision [Zhang et al., 2022], natural language processing [Zhang et al., 2005, Ando and Zhang, 2005], and bioinformatics [Pong et al., 2010]. This suggests that it is equally important to examine how ICL behaves when tasks share such common structure.

Another limitation of current ICL theory is its lack of precise analysis and interpretability. On the one hand, existing studies can predict overall performance and reproduce key phenomena such as double descent [Lu et al., 2025] and emergence [Wei et al., 2022, Raventós et al., 2023] in toy models, but these studies offer limited interpretability. On the other hand, prior works have shown that transformers *can* implement useful algorithms for regression [Garg et al., 2022, Akyürek et al., 2023, Von Oswald et al., 2023], yet the precise mathematical form of the predictors they *actually* acquire through training has remained elusive.

In this work, we aim to gain insight into these issues by analyzing an attention-only transformer trained on linear regression tasks [Garg et al., 2022] with a low-rank structure. Although the model is simple, it exhibits rich and nontrivial phenomena. It has therefore been extensively studied and shown to be instrumental in explaining various aspects of ICL [Von Oswald et al., 2023, Akyürek et al., 2023, Ahn et al., 2023, Wu et al., 2024, Zhang et al., 2023, Cui et al., 2025, Lu et al., 2025, Zhang et al., 2025, Samet and Zafer, 2025] both theoretically and empirically. To interpret the mechanisms of ICL effectively, we situate our analysis in the high-dimensional asymptotic regime, where the system's complex behavior can be precisely characterized by a small number of macro-

scopic order parameters [Zavatone-Veth et al., 2025].

At the core of our analysis is a precise solution of a linear attention model trained on regression tasks in the high-dimensional limit. This framework reveals a decomposition of the ICL prediction into two parts: an algorithmic component that performs linear regression, and a noise component whose effect depends on the information acquired during pre-training (Section 4). When the tasks share a low-rank structure, additional phenomena emerge. First, the model learns an efficient algorithm that exploits the low-rank structure of the tasks. Moreover, when the prompt contains memorized tasks or the learned task structure, the noise component can be suppressed (Section 4). Second, the learning of low-rank tasks is stabilized by an implicit regularization induced by finite-data statistics during pre-training (Section 5). Finally, we show that the rank of the pre-training tasks induces a sharp phase transition in the model's capabilities, from a regime where the model cannot fully exploit the low-rank structure to one where the low-rank structure is fully exploited. In the latter regime, the model faces a fundamental trade-off between specialization to pre-trained (in-distribution) tasks and robustness to unseen out-of-distribution tasks (Section 6)[1].

2 RELATED WORK

Theory of ICL. A growing body of theoretical studies has investigated ICL through simplified regression settings. Early analyses [Garg et al., 2022, Akyürek et al., 2023, Von Oswald et al., 2023, Ahn et al., 2023] showed that Transformers can implement regression procedures in-context and characterized the learned predictors in tractable toy models. Subsequent works [Lu et al., 2025, Letey et al., 2026] employing asymptotic analyses of linear attention have established a high-dimensional theory that predicts macroscopic behavior. This perspective makes it possible to analyze the typical outcome of pretraining in tractable Transformer models, going beyond representability to characterize its generalization. Building on these foundations, our work focuses on structured low-rank task families and provides a mechanism-level characterization of how pretraining statistics shape the learned in-context predictor through a decomposition into an algorithmic signal and noise terms.

Distribution Shift Across Tasks in ICL. Recent work has begun to characterize ICL under distribution shift. [Kwon et al., 2026] show that when task vectors lie in a union of low-dimensional subspaces, ICL can generalize to any subspace in their span. [Letey et al., 2026] study pretrain-test mismatch and show that performance degrades as the two task distributions separate, while perfect matching is not always optimal. Complementing these results, our theory links out-of-distribution degradation to an explicit structure-dependent error term that grows under task-structure mismatch.

3 MODEL

We model the mechanism of ICL using a transformer-based architecture. The overall goal is to train a model that can infer and execute the algorithm for linear regression purely from a sequence of examples provided in its context. To this end, we formalize a problem setup consisting of a pre-training phase, where the model learns the general algorithm, and an inference phase, where its ability to generalize to new instances is evaluated[2].

3.1 Pretraining Phase

We begin by constructing a base collection of tasks $\mathcal{W}_0 = \{\boldsymbol{w}^\mu \in \mathbb{R}^D \mid \mu = 1, \dots, M_0\}$, which represents a diverse pool of possible task vectors. Each $\boldsymbol{w}^\mu$ is generated from a low-dimensional latent structure: we fix a shared feature matrix $A \in \mathbb{R}^{D\times r}$ with orthonormal columns ($r \leq D$), draw a latent task feature vector $\boldsymbol{v}^\mu \sim \mathcal{N}(\boldsymbol{0}, I_r)$, and map it to the D-dimensional space via $\boldsymbol{w}^\mu = \sqrt{D/r} A \boldsymbol{v}^\mu$.

From this base collection, we then form the actual training set $\mathcal{W}$ by sampling $M (\gg M_0)$ tasks uniformly with replacement. This step reflects the fact that, in the thermodynamic limit, obtaining nontrivial learning behavior requires that individual tasks may recur many times. In other words, $\mathcal{W}_0$ captures the diversity of available tasks, while $\mathcal{W}$ specifies the effective workload that the learner repeatedly encounters.

For each task $\boldsymbol{w}^\mu \in \mathcal{W}$, we create a training instance by first drawing $L+1$ input vectors, $\{\boldsymbol{x}_i^\mu\}_{i=1}^{L+1}$, independently from a normal distribution $\mathcal{N}(\boldsymbol{0}, \mathbf{I}_D/D)$. The corresponding outputs are then generated as $y_i^\mu = \boldsymbol{w}^\mu \cdot \boldsymbol{x}_i^\mu + \epsilon_i^\mu$, where each ϵ_i^μ is an independent noise term drawn from $\mathcal{N}(0, \sigma^2)$. The first L pairs, $\{(\boldsymbol{x}_l^\mu, y_l^\mu)\}_{l=1}^L$, serve as **demonstrations**, while the final pair, $(\boldsymbol{x}_{L+1}^\mu, y_{L+1}^\mu)$, constitutes the **query** and its **label**. The entire generative process for these instances is denoted by $\mathcal{D}_{\text{pretrain}}$. These elements are concatenated to form the input sequence for the model,

[1] The code is available in https://github.com/taka255/icl-replica-analysis.

[2] A comprehensive list of notations is provided in Appendix A for reference.

which we refer to as the **context** $C^\mu \in \mathbb{R}^{(D+1)\times(L+1)}$:

$$C^\mu = \begin{pmatrix} \boldsymbol{x}_1^\mu & \boldsymbol{x}_2^\mu & \cdots & \boldsymbol{x}_L^\mu & \boldsymbol{x}_{L+1}^\mu \\ y_1^\mu & y_2^\mu & \cdots & y_L^\mu & 0 \end{pmatrix}, \tag{1}$$

where the label corresponding to the query $\boldsymbol{x}_{L+1}^\mu$ is replaced with a placeholder value of zero, indicating that it is the unknown quantity the model is required to predict.

Our model is an attention-based architecture that maps an input context to a prediction. This is expressed as the composition of an attention function $\mathsf{Atten}_\Theta : \mathbb{R}^{(D+1)\times(L+1)} \to \mathbb{R}^{(D+1)\times(L+1)}$ with parameters Θ, and a readout function $\mathsf{Read} : \mathbb{R}^{(D+1)\times(L+1)} \to \mathbb{R}$. The prediction for the query is given by:

$$\hat{y}_{L+1}^\mu = \mathsf{Read}(\mathsf{Atten}_\Theta(C^\mu)). \tag{2}$$

The model is trained by minimizing the mean squared error (MSE) loss between the predictions and the true labels over all M training instances. The loss function is:

$$\mathcal{L}(\Theta) = \frac{1}{M}\sum_{\mu=1}^{M}(\hat{y}_{L+1}^\mu - y_{L+1}^\mu)^2. \tag{3}$$

Through optimization, we obtain the learned parameters $\Theta^* = \arg\min_\Theta \mathcal{L}(\Theta)$.

3.2 Inference Phase

During the inference phase, we evaluate the generalization ability of the pre-trained model with fixed optimal parameters Θ^*. Performance is assessed on new test instances generated from a fixed evaluation task $\boldsymbol{w}^* \in \mathbb{R}^D$. Each test instance is created by first drawing $\tilde{L}+1$ input vectors, $\{\boldsymbol{x}_i\}_{i=1}^{\tilde{L}+1}$, independently from $\mathcal{N}(\mathbf{0}, \mathbf{I}_D/D)$, where $\tilde{L} \le L$ reflects a few-shot learning scenario. The corresponding outputs are generated without noise as $y_i = \boldsymbol{w}^* \cdot \boldsymbol{x}_i$ to assess the model's pure algorithmic reasoning. The first $\tilde{L}$ pairs serve as the **demonstrations**, while the final pair serves as the **query** $\boldsymbol{x}^{\tilde{L}+1} = \boldsymbol{x}$ and its **label** $y^{\tilde{L}+1} = y$. We refer to this test data distribution for a given task $\boldsymbol{w}^*$ as $\mathcal{D}_{\text{test}}$. Since the length of demonstrations $\tilde{L}$ may be less than L, the input sequence, referred to as the **prompt** P, is padded with zero vectors to match the context length of $L+1$ expected by the model. The prompt $P \in \mathbb{R}^{(D+1)\times(L+1)}$ is constructed as:

$$P = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{\tilde{L}} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{x} \\ y_1 & \cdots & y_{\tilde{L}} & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{4}$$

where the final column is the query input $\boldsymbol{x}$ with its label hidden. The prediction $\hat{y}$ is compared with the true label y. The model's prediction for the query is obtained by applying the frozen network; $\hat{y} = \mathsf{Read}(\mathsf{Atten}_{\Theta^*}(P))$.

3.3 Evaluation Protocol

We evaluate the model's ICL capabilities across three distinct protocols.

- **Task Memorization (TM):** The task is sampled from the pre-training set, $\boldsymbol{w}^* \sim \mathsf{Unif}(\mathcal{W}_0)$. This protocol tests the model's ability to recall and execute a task seen during training, given a new set of demonstrations.
- **In-Distribution Generalization (IDG):** The task is novel but generated from the same underlying structure, $\boldsymbol{w}^* = \sqrt{D/r}A\boldsymbol{v}^*$, for a new latent vector $\boldsymbol{v}^* \sim \mathcal{N}(\mathbf{0}, I_r)$. This protocol measures the ability to generalize to unseen tasks that share the same low-dimensional structure as the training tasks.
- **Out-of-Distribution Generalization (ODG):** The task is drawn from a standard isotropic Gaussian distribution, $\boldsymbol{w}^* \sim \mathcal{N}(\mathbf{0}, I_D)$. These tasks, by design, lack the low-rank structure inherent in the training distribution, thereby testing the model's robustness to a fundamental structural mismatch.

The model's performance on each protocol is quantified by its specific MSE, which we denote generically as $\mathcal{E}_{\text{protocol}}$ for protocol $\in$ {TM, IDG, ODG}. This error is formally defined as:

$$\mathcal{E}_{\text{protocol}} = \mathbb{E}_{\boldsymbol{w}^*\sim\mathcal{D}_{\text{protocol}}}\mathbb{E}_{\mathcal{D}_{\text{test}},\mathcal{D}_{\text{pretrain}}}\left[(y-\hat{y})^2\right], \tag{5}$$

where $\mathcal{D}_{\text{protocol}}$ represents the task distribution corresponding to each protocol as defined above.

3.4 Attention Architecture

For analytical tractability, we now specify the general architecture to a single-layer linear attention model. In this simplified setting, the attention mechanism is defined as:

$$\mathsf{Atten}_\Theta(C) = C + \frac{1}{L}VC(KC)^\top(QC), \tag{6}$$

where the learnable parameters are $\Theta = \{V, K, Q\}$, which are matrices in $\mathbb{R}^{(D+1)\times(D+1)}$. The readout function simply extracts the final scalar output at the query position: $\mathsf{Read}(C') = C'_{D+1,L+1}$.

Under this setting and certain simplifying assumptions, it is known that the pre-training process is equivalent to training a one-layer neural network (see [Zhang et al., 2023, Zhang et al., 2025] and Appendix B). Specifically, the input context C^μ can be mapped to an effective feature matrix $H^\mu \in \mathbb{R}^{D\times D}$ for each training instance μ:

$$H^\mu = \frac{D}{L}\left[\boldsymbol{x}_{L+1}^\mu \sum_{l\le L} y_l^\mu (\boldsymbol{x}_l^\mu)^\top\right]. \tag{7}$$

The attention model's prediction is then equivalent to the output of a neural network, which is a linear function of this effective feature:

$$\hat{y}^{\mu}_{L+1} = \mathrm{Tr}\Big(W H^{\mu\top}\Big) = \sum_{i,j=1}^{D} W_{ij} H^{\mu}_{ij}, \tag{8}$$

where the trainable weights of the attention mechanism $\{V, K, Q\}$ are implicitly mapped to a single effective weight matrix $W \in \mathbb{R}^{D\times D}$.

This equivalence extends directly to the inference phase. For a given prompt P with $\tilde{L}$ demonstrations, the effective feature matrix H and the corresponding prediction $\hat{y}$ are given by:

$$H = \frac{D}{\tilde{L}}\left[\boldsymbol{x}\sum_{l\le\tilde{L}} y_l(\boldsymbol{x}_l)^{\top}\right] \quad \hat{y} = \mathrm{Tr}\big(W^* H^{\top}\big), \tag{9}$$

where W^* is the learned weight matrix from the pre-training phase.

3.5 High-Dimensional Analysis

To analyze the model's typical performance, we consider the high-dimensional limit where the problem dimensions $(D, L, \tilde{L}, M_0, M, r)$ grow to infinity. The system's behavior in this regime is characterized by several dimensionless parameters, which are assumed to be fixed constants in $(0, \infty)$. We define the **Pre-training Sample Ratio** as $\alpha = L/D$ and the **Inference Sample Ratio** as $\tilde{\alpha} = \tilde{L}/D$, which represent the ratio of examples to features during pre-training and at inference, respectively, with the constraint that $\tilde{\alpha} \le \alpha$. Furthermore, we introduce the **Task Difficulty** $\rho = r/D$, which is the relative dimension of the latent task space where $\rho \in (0, 1]$. Finally, we define the **Task Diversity** as $\kappa = M_0/D$ to capture the richness of the base task set, and the **Training Data Density** as $\gamma = M/(DM_0)$ to represent the number of training instances per task.

To characterize the model's behavior, we define two regimes based on data density and task diversity. For data density, we distinguish the data-rich regime ($\kappa\gamma > 1$) and the data-deficient regime ($\kappa\gamma < 1$). Independently, for task diversity, we distinguish the *task-rich regime* ($\rho < \kappa$), where the variety of tasks spans the latent space, and the task-deficient regime ($\rho > \kappa$), where it does not.

4 DECOMPOSITION OF ICL PREDICTION

Our high-dimensional analysis reveals how structured tasks are learned in context. The model's prediction decomposes into a core algorithmic signal and two noise terms that can be suppressed by context. This view reinterprets ICL as a context-dependent noise-reduction mechanism, providing a framework for how general-purpose models adapt their computation. Formally, the prediction is given by the following result:

Result 4.1 (Decomposition of ICL prediction)**.** *The prediction $\hat{y}$ for a given prompt is:*

$$\hat{y} = \mathrm{tr}(W^* H^{\top}) \overset{\mathrm{d}}{=} \hat{y}_{\mathrm{algo}} + \hat{y}_{\mathrm{mem}} + \hat{y}_{\mathrm{struct}}, \tag{10}$$

*where $\overset{\mathrm{d}}{=}$ represents equality in distribution. This decomposition yields a deterministic **algorithmic signal** ($\hat{y}_{\mathrm{algo}}$) and two independent noise terms: a **memorization noise** ($\hat{y}_{\mathrm{mem}}$) correlated with the pre-training tasks, and a **structure noise** ($\hat{y}_{\mathrm{struct}}$) uncorrelated with them. This decomposition holds for any pre-training task distribution $\mathcal{W}_0$*[3]*.*

The next result characterizes each component analytically for $\alpha \gg 1$.

Result 4.2 (Asymptotic Characterization of the Components)**.** *Assume data-rich and task-rich regimes. Each term in Eq. (10) is given by:*

$$\hat{y}_{\mathrm{algo}} \overset{\mathrm{d}}{=} (P_A \boldsymbol{w}_{\mathrm{prompt}})^{\top}\boldsymbol{x} + \mathcal{O}\Big(\frac{1}{\alpha}\Big) \tag{11}$$

$$\hat{y}_{\mathrm{mem}} \overset{\mathrm{d}}{=} \frac{\sigma\sqrt{\rho}}{\gamma - 1/\kappa} Z_{\mathrm{mem}} + \mathcal{O}\Big(\frac{1}{\alpha}\Big) \tag{12}$$

$$\hat{y}_{\mathrm{struct}} \overset{\mathrm{d}}{=} \frac{\sigma\rho}{\sqrt{(1+\sigma^2)(\gamma - 1/\kappa)}} Z_{\mathrm{struct}}\sqrt{\alpha} + \mathcal{O}\Big(\frac{1}{\sqrt{\alpha}}\Big), \tag{13}$$

where random variables Z_{mem} and Z_{struct} are given by:

$$Z_{\mathrm{mem}} \sim \mathcal{N}\Big(0, (1/M_0)\big(A^{\top}\boldsymbol{w}_{\mathrm{prompt}}\big)^{\top} S_v^{-1}\big(A^{\top}\boldsymbol{w}_{\mathrm{prompt}}\big)\Big) \tag{14}$$

$$Z_{\mathrm{struct}} \sim \mathcal{N}\Big(0, (1/M_0)\big|P_A^{\perp}\boldsymbol{w}_{\mathrm{prompt}}\big|^2\Big) \tag{15}$$

respectively. Here, $S_v = (1/M_0)\sum_{\mu}\boldsymbol{v}^{\mu}\boldsymbol{v}^{\mu\top}$, $P_A = AA^{\top}$, $P_A^{\perp} = I_D - P_A$ and $\boldsymbol{w}_{\mathrm{prompt}} = 1/\tilde{\alpha}\sum_l y_l\boldsymbol{x}_l$.

For the formal statement and derivation of Results. 4.1 and 4.2, see Appendix E.

To interpret Result 4.1 at the level of generalization, we convert the prediction decomposition into an error decomposition. Because the two noise terms are

[3] We use a replica method [Mézard et al., 1987, Charbonneau et al., 2023] to analyze the asymptotic limit. This method involves a non-rigorous mathematical step. Consequently, we conservatively state our findings as "Results" rather than "Theorems". However, the validity of our theoretical predictions is confirmed by numerical experiments, which show excellent agreement (see Appendix G).

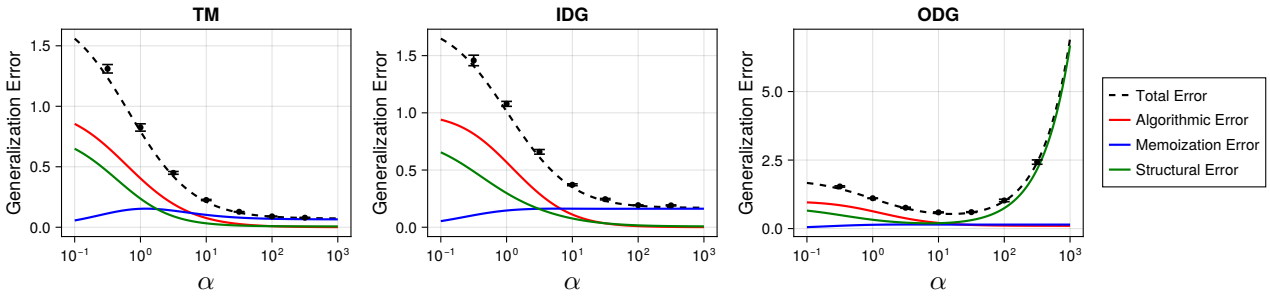

Figure 1: **Decomposition of generalization error reveals ICL's noise-reduction mechanism.** The total generalization error (in theory with dashed lines, and in numerical experiments with error bars) and its decomposition into algorithmic (signal), memorization (noise), and structural (noise) components (solid lines) for each protocol. Parameters: $\gamma = 1.5, \kappa = 1.5, \rho = 0.9, \sigma = 0.3, \tilde{\alpha} = \alpha, D = 60$. The error bars represent the standard errors of the mean over 5 independent numerical experiments per point.

independent and centered, their cross terms vanish after averaging, and we obtain $\mathcal{E}_{\text{protocol}} = \mathbb{E}[(y - \hat{y})^2] = \mathbb{E}[(y - \hat{y}_{\text{algo}})^2] + \mathbb{E}[\hat{y}^2_{\text{mem}}] + \mathbb{E}[\hat{y}^2_{\text{struct}}]$. Figure 1 plots these three contributions, showing how the decomposition of the predictor translates directly into the generalization error. These decompositions reveal that the prediction is governed by one algorithmic signal and two distinct noise sources, which we explain in detail below.

The Algorithmic Signal ($\hat{y}_{\text{algo}}$): The leading term, $\hat{y}_{\text{algo}}$, represents the model's core algorithmic competency. Eq. (11) reveals that ICL can acquire a sophisticated two-step procedure learned implicitly from pre-training without any explicit regularization. First, the model computes a naive (Matched Filter) estimator $\boldsymbol{w}_{\text{prompt}} = 1/\tilde{\alpha} \sum_l y_l \boldsymbol{x}_l$ from the prompt examples. Second, it projects this estimate onto the low-rank subspace defined by $P_A = AA^\top$, which represents the structural prior learned from the pre-training task distribution. This term is the primary signal of the prediction. As shown in Figure 1, the error contribution from this algorithmic component consistently decreases as α increases, indicating that a long context improves the model's algorithmic accuracy.

The Suppressible Noise Terms ($\hat{y}_{\text{mem}}, \hat{y}_{\text{struct}}$): Our analysis shows that the model's versatility across tasks inherently introduces noise, and that ICL operates by selectively suppressing these noise terms through context. This dual effect explains how ICL supports general-purpose adaptability, and the following two terms illustrate how it manifests in practice.

The term $\hat{y}_{\text{mem}}$ represents a memory-dependent noise. Eq. (12) shows that its magnitude is determined by the squared Mahalanobis distance between the projected task estimate $A^\top \boldsymbol{w}_{\text{prompt}}$ and the memorized task matrix S_v in the latent space. This distance becomes small when the current task estimate is statistically typical of the memorized tasks, i.e., it aligns with directions in which the tasks $\{\boldsymbol{v}^\mu\}$ have high variance. As seen in Figure 1, this memorization noise is lower for TM tasks compared to unfamiliar IDG tasks, as the model can successfully recall the familiar task. However, this term reveals a trade-off with respect to α. While a longer context reduces algorithmic and structural errors, it paradoxically worsens this noise, as shown in Figure 1 in the IDG protocol. This is a consequence of task-overfitting: with large α, the model's over-specialization to the specific patterns of pre-training instances leads to poorer generalization on new tasks, thereby increasing the memorization error.

The term $\hat{y}_{\text{struct}}$ represents a structure-dependent noise. The estimator $P_A^\perp \boldsymbol{w}_{\text{prompt}}$ in Eq. (13) explicitly measures the component of the in-context estimate $\boldsymbol{w}_{\text{prompt}}$ that is orthogonal to the learned structure A. This term can be interpreted as an error signal generated by the portion of the prompt that violates the learned structural prior. Figure 1 clearly shows this: for ODG tasks, the fundamental structural mismatch causes this noise component to become pronounced and dominate the error. Conversely, for in-distribution tasks (TM and IDG), a longer context allows the model to better identify the underlying structure, leading to a decrease in this structural noise.

The decomposition connects to practical ICL settings. The memorization and structural noise terms explain empirical patterns [Mueller et al., 2024, Wang et al., 2025, Letey et al., 2026], including strong performance on familiar or aligned tasks and degradation on unfamiliar or mismatched ones. Furthermore, the growth of memorization noise with longer context in in-distribution settings explains the task-overfitting-driven performance drop, a phenomenon that has received limited attention in

prior work. In this way, the decomposition matches existing empirical findings and generates new testable predictions.

5 FINITE PRE-TRAINING DATA INDUCES IMPLICIT REGULARIZATION

Section 4 showed that the model learns a low-rank regression algorithm; here, we investigate how this learning remains stable. This is particularly puzzling in low-rank settings ($\rho < 1$) where, despite the absence of explicit regularization in Eq. (3), our model achieves robust generalization. To understand this mechanism, we first examine the **idealized limit** of infinite pre-training context length ($\alpha \to \infty$ before learning). Contrary to the intuition that perfect data should improve performance, we demonstrate that this limit, lacking finite-sample fluctuations, renders the learning problem ill-posed. This leads to an initialization-sensitive solution and poor generalization, a stark contrast to the stable learning in the **asymptotic limit** ($\alpha \to \infty$ after learning). Our theoretical explanation attributes this stability to an implicit regularization induced by finite-data statistics.

5.1 Instability in the Idealized Limit

We conduct a numerical experiment to compare the model's IDG error under two conditions: a standard setting with a finite number of demonstrations α, and an idealized limit. The latter is implemented by training on the expected feature matrix $H^{\mu}_{\text{ideal}} = \lim_{\alpha\to\infty}[H^{\mu}] = \boldsymbol{x}^{\mu}_{L+1}\boldsymbol{w}^{\mu\top}$. In both cases, the model is trained via gradient descent from a random initialization $\boldsymbol{w}_0$ satisfying $\|\boldsymbol{w}_0\| = D$. The resulting IDG error is plotted as a function of ρ in Figure 2. The plot shows that for any $\rho < 1$, the error in the idealized limit fails to decrease, whereas for finite α, the error systematically improves as α increases. This result is counter-intuitive; the idealized model, which is given perfect information about the ground-truth task $\boldsymbol{w}^{\mu}$ through H^{μ}_{ideal}, was expected to perform better, yet it is outperformed by the model that must infer the task from finite demonstrations.

5.2 Theoretical Explanation

To explain this phenomenon, our theoretical analysis reveals that the learned weight matrix W^* behaves as if it were the solution to an effective optimization problem, which differs critically depending on whether the idealized limit or the asymptotic limit is considered.

Result 5.1 (Effective Objective Function)**.** *There exist scalars $\hat{q}$ and $\hat{\hat{q}}$, determined solely by the system parameters, such that the statistics of the learned weight matrix W^* are equivalent to those of the solution to the following optimization problem, i.e., $W^* \overset{\text{d}}{=} \operatorname{argmin} f(W)$, where the effective objective function $f(W)$ is given by:*

$$f(W) = \operatorname{Tr}\left[\frac{\hat{q}}{2}WSW^\top - \mathcal{M}_S W^\top + \frac{\hat{\hat{q}}}{2}WW^\top\right]. \quad (16)$$

Here, the matrix $S = (1/M_0)\sum_{\mu=1}^{M_0}\boldsymbol{w}^{\mu}\boldsymbol{w}^{\mu\top}$ is the task covariance matrix, and $\mathcal{M}_S$ is a random noise matrix whose structure depends on S.

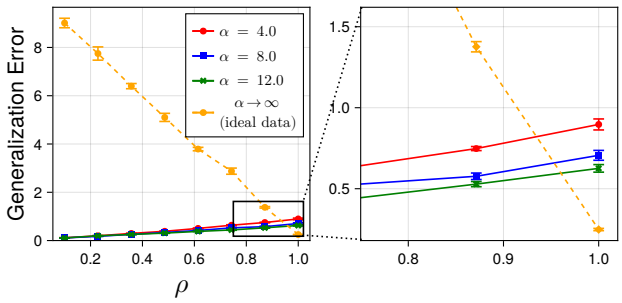

Figure 2: **Implicit regularization from finite data prevents learning instability.** IDG error as a function of the task subspace dimensionality ρ. The figure compares the performance of the idealized limit (dashed line), with standard setting using finite sample ratio α (solid lines). Parameters: $D = 40, M_0 = 60, \tilde{L} = 160, M = 2400, \sigma = 0.01$. The error bars represent the standard errors of the mean over 5 trials per point.

Within Eq. (16), the two scalars $\hat{q}$ and $\hat{\hat{q}}$ arise from marginalizing over the statistical fluctuations in the multi-body problem Eq. (3). The parameter $\hat{q}$ represents the strength of the quadratic interaction between the model parameters and the task structure encoded by S. The parameter $\hat{\hat{q}}$ acts as an effective regularization strength. This regularization is essential for learning low-rank tasks ($\rho < 1$), where the task covariance matrix S is rank-deficient. Without it, the objective function possesses flat directions corresponding to the null space of S. We therefore next characterize the magnitude of this effective regularization.

Result 5.2 (Asymptotics of Regularization)**.** *When $\kappa\gamma > 1$, the scalars in the effective objective function (Result 5.1) scale as $\hat{q} = \mathcal{O}(1)$ and $\|\mathcal{M}_S\| = \mathcal{O}(1)$. The effective regularization coefficient $\hat{\hat{q}}$ is given by:*

$$\hat{\hat{q}} = \begin{cases} \mathcal{O}(1/\alpha) > 0 & (\textit{in the asymptotic limit}) \\ 0 & (\textit{in the idealized limit}). \end{cases} \quad (17)$$

For the formal statement and derivation, see Appendix E.

These results provide a clear explanation for the observed phenomenon. Result 5.2 reveals that the regularization emerges only when the pre-training dataset is finite ($\hat{\tilde{q}} > 0$). By contrast, in the idealized limit, $\hat{\tilde{q}} = 0$, so the minimization problem becomes ill-posed and may fail to have a unique minimizer, leading to the learning instability seen in Figure 2.

Furthermore, this framework explains why performance improves as α increases. The effective regularization strength, given by the ratio $\hat{\tilde{q}}/\hat{q} \sim \mathcal{O}(1/\alpha)$, is naturally annealed as more data becomes available. This allows the model to rely more on the data and converge to a more accurate solution while avoiding instability.

5.3 Implications

This finding has several significant implications for theory and practice.

ICL as a Two-step Automatic Learning Process. We hypothesize that the origin of this implicit regularization lies in the shared structure and the two-stage process inherent to ICL. We can think of the first stage as task identification, where the model identifies the task from the pre-training dataset. The statistical fluctuations in this data ensure the resulting internal task representation is never perfectly low-rank and always contains a small, full-rank noise component. This unavoidable noise then becomes a beneficial regularizer during the second stage, which we can call inference execution. It breaks the flat directions of the optimization landscape that would otherwise cause learning instability in the idealized limit. In this way, the statistical imperfection in identifying the task is precisely what provides the stability for its execution in low-rank tasks.

Data as a Regularizer. In deep learning, implicit regularization is often attributed to optimization algorithms like SGD [Zhang et al., 2017]. While Transformers are also known to exhibit implicit regularization [Vasudeva et al., 2025, Frei and Vardi, 2025], existing theories often treat it as an extension of standard supervised learning. In contrast, our findings reveal a distinct form of regularization originating from the statistical fluctuations of a finite dataset. We attribute this phenomenon to the unique ICL objective of automatically acquiring an algorithm. This implies that for certain structures, learning is ironically hindered by idealized, noise-free data, suggesting that the natural statistical noise in a dataset can be a crucial feature, not a bug to be removed.

6 TRADE-OFFS FROM TASK DIFFICULTY

Building on our finding that the model learns a stable low-rank regression algorithm, we now explore how the model's capability is governed by the properties of this low-rank structure. We develop a theoretical framework based on eigenvalue spectral analysis, which predicts a transition in the model's capabilities. This prediction is later validated by our results, revealing a sharp phase transition and fundamental trade-off between specialization and out-of-distribution robustness.

6.1 Spectral Analysis of Task Structure

To understand the model's behavior, we draw an analogy to classical linear regression, where the Gram matrix is $G = X^\top X$. In the eigenbasis of G, the variance of the estimated coefficient along the direction associated with eigenvalue λ_i is $\mathsf{Var}(\hat{w}_i) \propto \sigma^2/\lambda_i$, where σ is the noise variance. A large eigenvalue of G corresponds to a feature direction with a high signal-to-noise ratio, leading to statistical stability. A small eigenvalue indicates an ill-conditioned direction that is highly sensitive to noise. Thus, the number and magnitude of nonzero eigenvalues jointly determine the model's learnable patterns and their stability. A zero eigenvalue means that no learnable pattern exists in the corresponding direction, so any apparent estimate comes only from regularization or noise.

Applying this analogy to our ICL framework, we can explain the model's capabilities by analyzing the eigenvalue distribution of the task covariance matrix $S = (1/M_0)\sum_{\mu=1}^{M_0} \boldsymbol{w}^\mu \boldsymbol{w}^{\mu\top}$ from Eq. (16), which encapsulates the task information learned during pre-training. This follows from the fact that, in the eigenbasis of S with eigenvalues s_j, the variance of the corresponding estimator scales as $\mathsf{Var}(W_{ij}) \propto 1/(\hat{q}s_j + \hat{\tilde{q}})^2$ from Eq. (16). The following theorem precisely characterizes this distribution of eigenvalues and predicts that its structure changes at the transition point $\rho = \kappa$.

Theorem 6.1 (Spectrum of the Task Matrix)**.** *Recall that $\rho = r/D$ and $\kappa = M_0/D$. The empirical spectral distribution of S converges almost surely to*

$$\nu_S(s) = (1 - \min(\rho, \kappa))\,\delta(s) + \mathbf{1}_{[s_-, s_+]}(s)\,\frac{\rho\kappa}{2\pi s}\sqrt{(s_+ - s)(s - s_-)}. \tag{18}$$

Here, $\mathbf{1}_{[s_-,s_+]}(s)$ is the indicator function. The support edges are given by $s_\pm = \left(\sqrt{\rho} \pm \sqrt{\kappa}\right)^2/(\rho\kappa)$.

The proof is given in Appendix F.

This theorem provides a quantitative basis for predicting the model's performance by revealing how the

learning bottleneck shifts. In the task-deficient regime ($\rho > \kappa$), the number of learnable patterns is limited by κ (the fraction of nonzero eigenvalues). Moreover, the typical eigenvalue magnitude scales as $\mathcal{O}(1/\kappa)$, implying that the statistical stability of these patterns is also controlled by κ. As a result, the model's capabilities do not change even as tasks get simpler, indicating the model's learning is bottlenecked by task diversity. In contrast, in the task-rich regime ($\rho < \kappa$), the number of nonzero eigenvalues is ρ, and their magnitudes scale as $\mathcal{O}(1/\rho)$. As tasks become simpler, the larger eigenvalues concentrate useful information, allowing the model to fully exploit the simplified low-rank structure for more stable and efficient learning.

6.2 Phase Transition in ICL Capabilities

To connect our spectral analysis with practical ICL performance, we derive the theoretical generalization errors for three settings: TM, IDG, and ODG.

Result 6.2. *The generalization errors $\mathcal{E}_{\text{TM}}$, $\mathcal{E}_{\text{IDG}}$, and $\mathcal{E}_{\text{ODG}}$ can be analytically obtained by solving a finite system of self-consistent equations. In particular, when the asymptotic case $\alpha \to \infty$, $\kappa\gamma > 1$, and $\sigma = 0$, each metric is given by:*

$$\mathcal{E}_{\text{TM}} = \frac{\min(\rho,\kappa)}{\tilde{\alpha}}\left(1 + \frac{1-\min(\rho,\kappa)}{\kappa\gamma - 1}\right) \tag{19}$$

$$\mathcal{E}_{\text{IDG}} = \begin{cases} \mathcal{E}_{\text{TM}} & \text{for } \rho < \kappa \\ 1 - \frac{\kappa}{\rho} - \frac{\kappa(\kappa-\rho)}{\rho(\kappa\gamma-1)} + \mathcal{E}_{\text{TM}} & \text{for } \rho \geq \kappa \end{cases} \tag{20}$$

$$\mathcal{E}_{\text{ODG}} = 1 - 2\min(\rho,\kappa) + (1+\tilde{\alpha})\mathcal{E}_{\text{TM}}. \tag{21}$$

For the formal statement and derivation, see Appendix E.

As shown in Result 6.2, these errors exhibit a sharp phase transition at the predicted critical point $\rho = \kappa$ when $\alpha \to \infty$. This extends prior work that identified phase transitions with respect to task diversity when no shared task structure is assumed [Lu et al., 2025]. Figure 3 illustrates this behavior by showing how the generalization errors change as ρ varies with fixed κ: for finite α we observe a crossover, which sharpens into a distinct phase transition as $\alpha \to \infty$, dividing the generalization behavior into two regimes. In the task-deficient regime ($\rho > \kappa$), both TM and ODG errors remain on a plateau, consistent with our spectral analysis. In this regime, the model's performance is unaffected by task difficulty, since learning is bottlenecked by task diversity. In the task-rich regime ($\rho < \kappa$), as task difficulty decreases, the ODG error increases, while the IDG and TM errors decrease in tandem and remain identical. This outcome demonstrates that the model can perfectly capture the in-distribution task structure when $\alpha \to \infty$, as reflected in the alignment of IDG and TM errors. It also shows that strong specialization to the low-rank task structure enhances memorization and in-distribution performance, but at the expense of out-of-distribution robustness.

Next, we analyze the role of $\tilde{\alpha}$. Result 6.2 shows that in the task-rich regime the $\mathcal{E}_{\text{TM}}$ and $\mathcal{E}_{\text{IDG}}$ scale as $\mathcal{O}(1/\tilde{\alpha})$. This means the model has learned the in-distribution task structure, but finite $\tilde{\alpha}$ limits the retrieval of this structure from the prompt. In the limit $\tilde{\alpha} \to \infty$, the errors vanish as $\mathcal{E}_{\text{TM}} = \mathcal{E}_{\text{IDG}} = 0$, showing that the model acquires the ability to achieve complete in-distribution generalization once given enough prompt examples, with the transition occurring at $\rho = \kappa$[4].

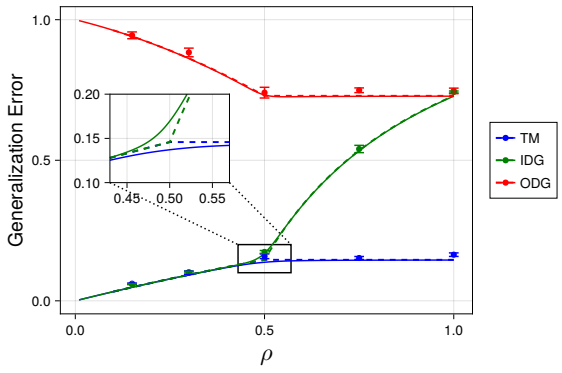

Figure 3: **Phase transition in the model's capabilities.** Generalization errors ($\mathcal{E}_{\text{TM}}, \mathcal{E}_{\text{IDG}}, \mathcal{E}_{\text{ODG}}$) as a function of the task difficulty ρ. The results confirm the predicted phase transition in generalization errors at $\rho = \kappa$. The error bars represent the standard errors of the mean over 5 trials per point. Parameters: $\alpha = 200$ (solid lines), $\alpha \to \infty$ (dashed lines), $\tilde{\alpha} = 4.0, \gamma = 8.0, \kappa = 0.5, \sigma = 0, D = 40$.

6.3 Implications for Pre-training Strategy

These theoretically grounded findings reveal that the balance between task difficulty and diversity is a critical lever for tailoring a model's capabilities. For instance, developing a specialized model for a specific domain mandates operating in the task-rich regime. This approach, which involves curating a dataset with high diversity of relatively simple tasks, enables the model to efficiently master domain-specific patterns and achieve high TM and IDG performance.

[4]The sudden acquisition of this IDG ability may represent a novel type of *emergence*. Unlike standard emergent phenomena in LLMs [Wei et al., 2022, Raventós et al., 2023], which are typically driven by increasing dataset size or model scale, this transition is induced by a change in the intrinsic structure of the task distribution itself, as predicted by our theory.

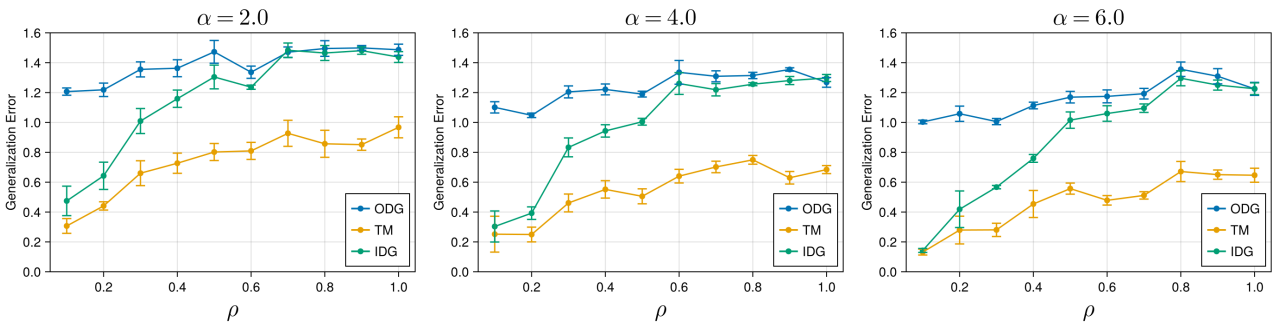

Figure 4: **Nonlinear softmax attention reproduces the qualitative ρ-dependence predicted by the linear theory.** Generalization errors (TM/IDG/ODG) of a one-layer softmax self-attention model as a function of task difficulty ρ. The error bars represent the standard errors of the mean over 5 trials per point. Parameters: $D = 20$, $\gamma = 3.0$, $\kappa = 0.5$, $\sigma = 0.01$.

For many practical applications, however, the most effective strategy is to target the "sweet spot" near the phase transition ($\rho \approx \kappa$). This regime offers a favorable trade-off while maintaining robust ODG performance, and it avoids the need to prepare unnecessarily difficult training tasks.

Moreover, our findings are consistent with the intuition of curriculum learning [Bengio et al., 2009], which introduces simple tasks first and gradually shifts to more complex ones. In our framework, the early stage of training corresponds to low task diversity ($\rho > \kappa$), where the model cannot yet learn complex tasks and thus only benefits from simpler ones. As training progresses and task diversity increases, the model enters a regime where it can exploit the richer structure to master more complex and abstract tasks. This explains why a curriculum that progresses from simple to difficult tasks is effective.

7 EXPERIMENTS ON NON-LINEAR ATTENTION

Our theoretical analysis has so far focused on linear attention, where the predictive behavior and phase transitions can be characterized precisely through macroscopic order parameters. To examine whether the same qualitative trends persist beyond this tractable setting, we also study a one-layer nonlinear model.

Specifically, we replace the linear mechanism in Eq. (6) with a one-layer softmax self-attention model, while keeping the pre-training and evaluation protocols identical to those in Sections 3.1, 3.2, and 3.3. We train the model on regression tasks with fixed input dimension D and task diversity κ, and vary the task difficulty ρ to evaluate the resulting generalization performance. The detailed architecture and optimization setup are described in Appendix H.

Figure 4 shows the TM, IDG, and ODG errors as functions of ρ for the one-layer softmax attention model. As ρ decreases, both TM and IDG errors improve systematically, indicating that nonlinear attention can also exploit shared task structures. Moreover, the improvement in TM begins to plateau as ρ approaches 1; this is consistent with our prediction in Section 6, which suggests that performance is eventually bottlenecked by task diversity (κ) rather than by the structure of individual tasks. We also note that we do not observe a sharp phase transition, and the ODG error remains higher than the random-guessing baseline in the present setting. These effects are likely to be most visible in the large-α regime, which we do not explore in the present nonlinear experiments.

8 CONCLUSION

In summary, our study sheds light on the mechanisms of ICL in linear transformers. We show that ICL functions as a context-dependent noise-reduction mechanism, enabling the model to perform a precise low-rank regression algorithm when the task structure is aligned with the context. We further demonstrate that statistical fluctuations in finite training data give rise to an implicit regularization, which stabilizes the learning process. Finally, we uncover a sharp phase transition governed by task structure, highlighting a fundamental trade-off between specialization and robustness under distributional shifts. These results together provide a coherent framework for understanding how transformers learn to learn.

Acknowledgements

KT was supported by JST BOOST NAIS (Grant No. JPMJBS2418). TT was supported by JSPS KAK-

ENHI (Grant No. 23K16960) and JST ACT-X (Grant No. JPMJAX24CG). TT and YK were supported by JSPS KAKENHI (Grant No. 22H05117).

References

[Ahn et al., 2024] Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges. In Falk, N., Papi, S., and Zhang, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

[Ahn et al., 2023] Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650.

[Akyürek et al., 2023] Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

[Ando and Zhang, 2005] Ando, R. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6(61):1817–1853.

[Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Caruana, 1997] Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

[Charbonneau et al., 2023] Charbonneau, P., Marinari, E., Parisi, G., Ricci-tersenghi, F., Sicuro, G., Zamponi, F., and Mezard, M. (2023). *Spin glass theory and far beyond: replica symmetry breaking after 40 years*. World Scientific.

[Cui et al., 2025] Cui, Y., Ren, J., He, P., Liu, H., Tang, J., and Xing, Y. (2025). Superiority of multi-head attention: A theoretical study in shallow transformers in in-context linear regression. In *The 28th International Conference on Artificial Intelligence and Statistics*.

[Frei and Vardi, 2025] Frei, S. and Vardi, G. (2025). Trained transformer classifiers generalize and exhibit benign overfitting in-context. In *The Thirteenth International Conference on Learning Representations*.

[Garg et al., 2022] Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598.

[Han and Zhang, 2016] Han, L. and Zhang, Y. (2016). Multi-stage multi-task learning with reduced rank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

[Kwon et al., 2026] Kwon, S. M., Xu, A. S., Yaras, C., Balzano, L., and Qu, Q. (2026). Out-of-distribution generalization of in-context learning: A low-dimensional subspace perspective. In *The 29th International Conference on Artificial Intelligence and Statistics*.

[Letey et al., 2026] Letey, M., Zavatone-Veth, J. A., Lu, Y. M., and Pehlevan, C. (2026). Pretrain–test task alignment governs generalization in in-context learning. In *The Fourteenth International Conference on Learning Representations*.

[Lu et al., 2025] Lu, Y. M., Letey, M., Zavatone-Veth, J. A., Maiti, A., and Pehlevan, C. (2025). Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122.

[Mei and Montanari, 2022] Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

[Mézard et al., 1987] Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company.

[Mueller et al., 2024] Mueller, A., Webson, A., Petty, J., and Linzen, T. (2024). In-context learning generalizes, but not always robustly: The case of syntax. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4761–4779, Mexico

City, Mexico. Association for Computational Linguistics.

[Oko et al., 2024] Oko, K., Song, Y., Suzuki, T., and Wu, D. (2024). Pretrained transformer efficiently learns low-dimensional target functions in-context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[Pong et al., 2010] Pong, T. K., Tseng, P., Ji, S., and Ye, J. (2010). Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM J. Optim.*, 20(6):3465–3489.

[Potters and Bouchaud, 2020] Potters, M. and Bouchaud, J.-P. (2020). *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press.

[Raventós et al., 2023] Raventós, A., Paul, M., Chen, F., and Ganguli, S. (2023). Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246.

[Samet and Zafer, 2025] Samet, D. and Zafer, D. (2025). Asymptotic study of in-context learning with random transformers through equivalent models. *arXiv [stat.ML]*.

[Vasudeva et al., 2025] Vasudeva, B., Deora, P., and Thrampoulidis, C. (2025). Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*.

[Von Oswald et al., 2023] Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

[Wang and Chen, 2023] Wang, J. and Chen, Y. (2023). A review on code generation with llms: Application and evaluation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pages 284–289. IEEE.

[Wang et al., 2025] Wang, Q., Wang, Y., Ying, X., and Wang, Y. (2025). Can in-context learning really generalize to out-of-distribution tasks? In *The Thirteenth International Conference on Learning Representations*.

[Wei et al., 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

[Wu et al., 2024] Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q., and Bartlett, P. (2024). How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*.

[Zavatone-Veth et al., 2025] Zavatone-Veth, J. A., Bordelon, B., and Pehlevan, C. (2025). Summary statistics of learning link changing neural representations to behavior. *Frontiers in Neural Circuits*, Volume 19 - 2025.

[Zhang et al., 2017] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

[Zhang et al., 2005] Zhang, J., Ghahramani, Z., and Yang, Y. (2005). Learning multiple related tasks using latent independent component analysis. *Neural Inf Process Syst*, 18:1585–1592.

[Zhang et al., 2023] Zhang, R., Frei, S., and Bartlett, P. (2023). Trained transformers learn linear models in-context. *J. Mach. Learn. Res.*, 25(49):49:1–49:55.

[Zhang et al., 2025] Zhang, Y., Singh, A. K., Latham, P. E., and Saxe, A. M. (2025). Training dynamics of in-context learning in linear attention. In *Forty-second International Conference on Machine Learning*.

[Zhang et al., 2022] Zhang, Y., Zhang, Y., and Wang, W. (2022). Learning linear and nonlinear low-rank structure in multi-task learning. *IEEE Trans. Knowl. Data Eng.*, 35(8):1–12.

[Zhong and Wang, 2024] Zhong, L. and Wang, Z. (2024). Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21841–21849.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes; provided in Section 3.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable; our work does not involve new algorithmic design requiring time or space complexity analysis.]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes; anonymized source code is provided at the end of Section 1.]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes; provided in Section 3 and Appendix E.]

(b) Complete proofs of all theoretical results. [Yes; provided in Appendix E and Appendix F]

(c) Clear explanations of any assumptions. [Yes; provided in Appendix E.]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes; provided at the end of Section 1.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes; provided in Section 3.]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes; provided in the caption of each figure.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes; the experiments were run on standard local compute resources.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator if your work uses existing assets. [Not Applicable; This work does not rely on any existing assets.]

(b) The license information of the assets, if applicable. [Not Applicable; No existing assets were used, so licensing is not a concern.]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable; No new assets are being released with this submission.]

(d) Information about consent from data providers/curators. [Not Applicable; Our work is based on theoretical analysis and simulations with synthetically generated data, not existing datasets requiring consent.]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable; Our work is theoretical and does not involve data with any personally identifiable or sensitive content.]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable; Our research is purely theoretical and does not involve crowdsourcing or human subjects.]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable; Our research is purely theoretical and does not involve crowdsourcing or human subjects.]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable; Our research is purely theoretical and does not involve crowdsourcing or human subjects.]

Learning Linear Regression with Low-Rank Tasks in-Context: Supplementary Materials

A Summary of Notations

In this section, we summarize the notation used in the main text for the reader's convenience.

Table 1: Summary of Notation

Symbol	Description
Dimensions and Counts	
D	Dimension of the input and task vector space.
r	Dimension of the latent space for tasks ($r \leq D$).
M	Total number of instances in the pre-training set.
M_0	Number of base tasks in the initial pool $\mathcal{W}_0$.
L	Number of demonstrations (example pairs) in a pre-training context.
$\tilde{L}$	Number of demonstrations in an inference prompt ($\tilde{L} \leq L$).
Sets, Distributions, and Protocols	
$\mathcal{W}_0$	Base set of M_0 task vectors, $\lvert\mathcal{W}_0\rvert = M_0$.
$\mathcal{W}$	Full pre-training set of M tasks, sampled from $\mathcal{W}_0$, $\lvert\mathcal{W}\rvert = M$.
$\mathcal{D}_{\text{pretrain}}$	Generative distribution for pre-training instances.
$\mathcal{D}_{\text{test}}$	Generative distribution for inference instances.
$\mathcal{D}_{\text{protocol}}$	Task distribution for an evaluation protocol (TM, IDG, or ODG).
TM	Task Memorization evaluation protocol.
IDG	In-Distribution Generalization evaluation protocol.
ODG	Out-of-Distribution Generalization evaluation protocol.
$\mathcal{E}_{\text{protocol}}$	Generalization error for an evaluation protocol.
Task and Data	
$\boldsymbol{w}^\mu \in \mathbb{R}^D$	Ground-truth weight vector defining task μ.
$A \in \mathbb{R}^{D\times r}$	Shared, low-dimensional feature matrix with orthonormal columns.
$\boldsymbol{v}^\mu \in \mathbb{R}^r$	Latent feature vector for base task μ.
$\boldsymbol{x}_l^\mu, \boldsymbol{x} \in \mathbb{R}^D$	Input vectors for demonstrations or queries.
$y_l^\mu, y \in \mathbb{R}$	Output labels for demonstrations or queries.
Model Components in the attention model	
$\Theta = \{V, K, Q \in \mathbb{R}^{D\times D}\}$	Set of learnable parameters in the linear attention model.
Atten_Θ	The attention function: $\mathbb{R}^{(D+1)\times(L+1)} \to \mathbb{R}^{(D+1)\times(L+1)}$.
Read	The readout function: $\mathbb{R}^{(D+1)\times(L+1)} \to \mathbb{R}$.
$C^\mu, P \in \mathbb{R}^{(D+1)\times(L+1)}$	Feature matrix derived from the context/prompt, the input of the attention model.
Model Components in the equivalent linear neural network	
$\Theta = \{W \in \mathbb{R}^{D\times D}\}$	Set of learnable parameters in the linear neural network.
$H^\mu, H \in \mathbb{R}^{D\times D}$	Effective feature matrix derived from the context/prompt, the input of the linear neural network.
High-Dimensional Analysis Parameters	

Table 1 – continued from previous page

Symbol	Description
$\alpha = L/D$	Pre-training sample ratio.
$\tilde{\alpha} = \tilde{L}/D$	Inference sample ratio.
$\rho = r/D$	Task subspace dimensionality.
$\kappa = M_0/D$	Task diversity.
$\gamma = M/(DM_0)$	Training data density.

B Justification of Equivalence of the Linear Attention Model to a Linear Neural Network

In this section, we demonstrate that the single-layer linear attention model employed in our study can be mapped to an equivalent single-layer linear neural network. This equivalence provides a tractable framework for our theoretical analysis.

First, from the definition of the attention function Atten_Θ, we have:

$$\mathsf{Atten}_\Theta(C) = C + \frac{1}{L} VC(KC)^\top (QC) \tag{22}$$

$$= C + \frac{1}{L} VCC^\top (K^\top Q) C \tag{23}$$

The dependency on the key and query matrices, K and Q, occurs only through the product $K^\top Q$. We can therefore simplify the parameterization by defining a single matrix $R = K^\top Q \in \mathbb{R}^{(D+1)\times(D+1)}$.

To analyze the interaction between features and labels, we partition the parameter matrices V and R into blocks corresponding to the feature dimensions (D) and the label dimension (1):

$$R = \begin{pmatrix} R_{11} & \boldsymbol{r}_{12} \\ \boldsymbol{r}_{21}^\top & r_{22} \end{pmatrix}, \quad V = \begin{pmatrix} V_{11} & \boldsymbol{v}_{12} \\ \boldsymbol{v}_{21}^\top & v_{22} \end{pmatrix}, \tag{24}$$

where $R_{11}, V_{11} \in \mathbb{R}^{D\times D}$, $\boldsymbol{r}_{12}, \boldsymbol{v}_{12}, \boldsymbol{r}_{21}, \boldsymbol{v}_{21} \in \mathbb{R}^D$ and $r_{22}, v_{22} \in \mathbb{R}$.

In this notation, a direct calculation of the model's prediction, which is the scalar value at the query position $\{\dots\}_{D+1,L+1}$, yields:

$$\mathsf{Read}(\mathsf{Atten}_\Theta(C^\mu)) = \left\{ C^\mu + \frac{1}{L} VC^\mu {C^\mu}^\top R C^\mu \right\}_{D+1,L+1} \tag{25}$$

$$= \frac{1}{L} \Big(\boldsymbol{v}_{21}^\top X^\mu {X^\mu}^\top R_{11} + v_{22} \boldsymbol{y}^\mu {X^\mu}^\top R_{11} + \boldsymbol{v}_{21}^\top X^\mu {\boldsymbol{y}^\mu}^\top \boldsymbol{r}_{21} + \boldsymbol{r}_{21} {\boldsymbol{y}^\mu}^\top \boldsymbol{y}^\mu \Big) \boldsymbol{x}_{L+1}^\mu, \tag{26}$$

where $X^\mu = \begin{pmatrix} \boldsymbol{x}_1^\mu & \boldsymbol{x}_2^\mu & \cdots & \boldsymbol{x}_L^\mu \end{pmatrix} \in \mathbb{R}^{D\times L}$ and $\boldsymbol{y}^\mu = \begin{pmatrix} y_1^\mu & y_2^\mu & \cdots & y_L^\mu & 0 \end{pmatrix}^\top \in \mathbb{R}^{L+1}$.

To make the model analytically tractable, we follow recent works [Wu et al., 2024, Frei and Vardi, 2025, Zhang et al., 2025] and introduce a key simplification by setting the off-diagonal block matrices to zero: $\boldsymbol{v}_{21} = \boldsymbol{0}$ and $\boldsymbol{r}_{21} = \boldsymbol{0}$. This simplification is well-founded, as it has been shown that for networks initialized with these parameters at zero, they remain zero throughout training under gradient flow dynamics [Zhang et al., 2025]. While this reduces complexity, the resulting architecture is still sufficiently rich to exhibit nontrivial in-context learning phenomena.

Under this assumption, the prediction formula simplifies dramatically. By defining an effective weight matrix $W = (v_{22} R_{11})^\top \in \mathbb{R}^{D\times D}$, the model's output becomes:

$$\mathsf{Read}(\mathsf{Atten}_\Theta(C^\mu)) = \frac{1}{L} v_{22} \boldsymbol{y}^{\mu\top} X^{\mu\top} R_{11} \boldsymbol{x}_{L+1}^\mu = \mathrm{Tr}\Big(W H^{\mu\top} \Big) \tag{27}$$

This final expression demonstrates that the attention model's output is equivalent to that of a single-layer linear neural network. This equivalent network maps an effective feature matrix H^μ to a scalar prediction via the trace operator, corresponding to a mapping from $\mathbb{R}^{D\times D} \to \mathbb{R}$.

C Justification of Equivalence of the Linear Attention Model to a Many-Teacher-Student Model

To enable a tractable high-dimensional analysis using the replica method, we map our ICL model to an equivalent teacher-student framework with multiple teachers. This reformulation is not only a crucial technical step, but it also provides a clear interpretation of the pre-training phase. Furthermore, we believe this analytical strategy can serve as a valuable starting point for studying a broader class of similar ICL models. The core of this equivalence is the following statement:

Statement C.1. *The macroscopic statistical quantities of the learning process based on Eq. (8) remain unchanged if the true training labels y^{μ}_{L+1} are replaced by the outputs of a task-specific teacher model. The teacher for task μ is defined by the weight matrix $W^{\mu}_{\text{teacher}} = \boldsymbol{w}^{\mu}(\boldsymbol{w}^{\mu})^{\top}/D$, leading to the substitution:*

$$y^{\mu}_{L+1} \to \mathrm{Tr}(W^{\mu}_{\text{teacher}} H^{\mu}) + \epsilon^{\mu}. \tag{28}$$

For the justification of Statement C.1 and a discussion about its relevance to the traditional teacher-student model, see the following subsections.

This equivalence is based on an analogous to the Gaussian Equivalence Theorem [Mei and Montanari, 2022], as the first two moments of the true labels and the teacher's outputs are identical, while higher-order correlations become negligible in the high-dimensional limit.

The student model in our framework is not merely memorizing the solutions to specific tasks. Because it learns from an ensemble of M different teachers, one for each training instance, it is forced to distill the underlying, universal problem-solving algorithm common to all of them.

C.1 Justification of Equivalence of the Linear Attention Model to a Many-Teacher-Student Model

In this subsection, we will justify the equivalence to a many-teacher-student model, by checking the consistency of the first and second-order statistics. Specifically, we will show the following lemma:

Lemma C.2. *Let $W^{\mu}_{\text{teacher}} = \boldsymbol{w}^{\mu}(\boldsymbol{w}^{\mu})^{\top}/D$ be the weight matrix of a task-specific teacher model, and $\bar{y}^{\mu}_{L+1} = \mathrm{Tr}(W^{\mu}_{\text{teacher}} H^{\mu}) + \epsilon^{\mu}$ be the output of a task-specific teacher model. Then, for the fixed $\mathcal{W}$, the first and second-order statistics of the teacher model are consistent with the ICL model, i.e.,*

$$\mathbb{E}\big[\bar{y}^{\mu}_{L+1}\big] = \mathbb{E}\big[y^{\mu}_{L+1}\big] = 0 \tag{29}$$

$$\mathbb{E}\big[\bar{y}^{\mu}_{L+1}\bar{y}^{\nu}_{L+1}\big] = \mathbb{E}\big[y^{\mu}_{L+1}y^{\nu}_{L+1}\big] = \delta_{\mu\nu}\big(1+\sigma^2\big) + \mathcal{O}\bigg(\frac{1}{D}\bigg) \tag{30}$$

$$\mathbb{E}\big[\bar{y}^{\mu}_{L+1}H^{\nu}_{ij}\big] = \mathbb{E}\big[y^{\mu}_{L+1}H^{\nu}_{ij}\big] = \frac{1}{D}\delta_{\mu\nu}w^{\mu}_{i}w^{\mu}_{j} + \mathcal{O}\bigg(\frac{1}{D^2}\bigg). \tag{31}$$

Proof. Due to $H^{\mu}_{ij} = 0$, we have $\mathbb{E}\big[\bar{y}^{\mu}_{L+1}\big] = \mathbb{E}\big[y^{\mu}_{L+1}\big] = 0$. Also, straightforwardly we have $\mathbb{E}\big[y^{\mu}_{L+1}y^{\nu}_{L+1}\big] = \delta_{\mu\nu}\big(1+\sigma^2\big)$ and

$$\mathbb{E}\big[H^{\mu}_{ij}y^{\nu}_{L+1}\big] = \frac{D}{L}\delta_{\mu\nu}x^{\mu}_{L+1,i}\bigg(\sum_{l=1}^{L} y^{\mu}_{l}x^{\mu}_{l,j}\bigg)\bigg(\sum_{k=1}^{D} w^{\mu}_{k}x^{\mu}_{l,k} + \epsilon^{\mu}_{l}\bigg) \tag{32}$$

$$= \frac{D}{L}\delta_{\mu\nu}\sum_{k,k'} w^{\mu}_{k}w^{\mu}_{k'}\mathbb{E}\bigg[x^{\mu}_{L+1,i}x_{L+1,k'}\sum_{l=1}^{L} x^{\mu}_{l,j}x^{\mu}_{l,k}\bigg] \tag{33}$$

$$= \frac{1}{D}\delta_{\mu\nu}w^{\mu}_{i}w^{\mu}_{j}. \tag{34}$$

For the second-order statistics of the teacher model, we have:

$$\mathbb{E}\big[H^{\mu}_{ij}H^{\nu}_{i'j'}\big] = \mathbb{E}\left[\frac{D^2}{L^2}\left[x^{\mu}_{L+1,i}\sum_{l=1}^{L}\left(\sum_{k=1}^{D}w^{\mu}_{k}x^{\mu}_{l,k}+\epsilon^{\mu}_{l,k}\right)x^{\mu}_{l,j}\right]\left[x^{\nu}_{L+1,i'}\sum_{l'=1}^{L}\left(\sum_{k'=1}^{D}w^{\nu}_{k'}x^{\nu}_{l',k'}+\epsilon^{\nu}_{l'}\right)x^{\nu}_{l',j'}\right]\right] \tag{35}$$

$$= \delta_{\mu\nu}\mathbb{E}\left[\frac{D^2}{L^2}\sum_{k,k'}w^{\mu}_{k}w^{\nu}_{k'}\left[x^{\mu}_{L+1,i}x^{\nu}_{L+1,i'}\left(\sum_{l=1}^{L}\sum_{l'=1}^{L}x^{l}_{l,k}x^{l}_{l,j}x^{l}_{l',k'}x^{l}_{l',j'}\right)\right]\right.$$

$$\left.+\frac{D^2}{L^2}x^{\mu}_{L+1,i}x^{\mu}_{L+1,i'}\sum_{l=1}^{L}\sum_{l'=1}^{L}x^{l}_{l,j}x^{l}_{l',j'}\epsilon^{l}_{l}\epsilon^{l}_{l'}\right] \tag{36}$$

$$= \delta_{\mu\nu}\frac{\delta_{ii'}}{D}\left(w_jw_{j'}+\frac{D}{L}(1+\sigma^2)\delta_{jj'}\right)+\mathcal{O}\left(\frac{1}{D^2}\right) \tag{37}$$

$$\mathbb{E}\big[\bar{y}^{\mu}_{L+1}H^{\nu}_{ij}\big] = \frac{1}{D}\sum_{i',j'}w^{\mu}_{i'}w^{\nu}_{j'}\mathbb{E}\big[H^{\mu}_{ij}H^{\nu}_{i'j'}\big] \tag{38}$$

$$= \frac{1}{D}\sum_{i',j'}w^{\mu}_{i'}w^{\nu}_{j'}\delta_{\mu\nu}\frac{\delta_{ii'}}{D}\left(w_jw_{j'}+\frac{D}{L}(1+\sigma^2)\delta_{jj'}\right)+\mathcal{O}\left(\frac{1}{D^2}\right) \tag{39}$$

$$= \frac{1}{D}\delta_{\mu\nu}w^{\mu}_{i}w^{\mu}_{j}+\mathcal{O}\left(\frac{1}{D^2}\right) \tag{40}$$

$$\mathbb{E}\big[\bar{y}^{\mu}_{L+1}\bar{y}^{\nu}_{L+1}\big] = \frac{1}{D^2}\sum_{i,j}\sum_{i',j'}w^{\mu}_{i}w^{\nu}_{j}w^{\mu}_{i'}w^{\nu}_{j'}\mathbb{E}\big[H^{\mu}_{ij}H^{\nu}_{i'j'}\big] \tag{41}$$

$$= \frac{1}{D^2}\sum_{i,j}\sum_{i',j'}w^{\mu}_{i}w^{\nu}_{j}w^{\mu}_{i'}w^{\nu}_{j'}\delta_{\mu\nu}\frac{\delta_{ii'}}{D}\left(w_jw_{j'}+\frac{D}{L}(1+\sigma^2)\delta_{jj'}\right)+\mathcal{O}\left(\frac{1}{D^2}\right) \tag{42}$$

$$= \delta_{\mu\nu}(1+\sigma^2)+\mathcal{O}\left(\frac{1}{D}\right), \tag{43}$$

which completes the proof. □

C.2 Relation to Committee Machine

This process reveals a crucial distinction from conventional multi-teacher models like committee machines, which learn the average behavior of a teacher ensemble to capture their consensus. Our student, in contrast, is not trained to find this middle ground but must learn to imitate every individual teacher. Statement C.2 reveals that the essence of ICL is not knowledge aggregation, but the acquisition of a universal meta-algorithm capable of acting as any specialized teacher based on the context.

D Preliminary Lemmas in Random Matrix Theory

In this section, we provide the proofs for several technical lemmas required in the calculation of the replica method in the main text. These lemmas establish fundamental properties of random matrices that arise in our analysis, particularly concerning resolvent functions and their derivatives.

D.1 Resolvent Functions for Wishart-Type Matrices

We begin by establishing the resolvent functions for a class of random matrices that play a central role in our replica calculation.

Lemma D.1. *Consider the random matrix*

$$\tilde{S} = \frac{D}{r}\frac{1}{M_0}\sum_{\mu=1}^{M_0}\boldsymbol{v}^{\mu}\boldsymbol{v}^{\mu\top}\in\mathbb{R}^{r\times r} \tag{44}$$

$$S = \frac{D}{r}\frac{1}{M_0}\sum_{\mu=1}^{M_0}A\boldsymbol{v}^{\mu}\boldsymbol{v}^{\mu\top}A^{\top}\in\mathbb{R}^{D\times D}, \tag{45}$$

where $0 < r < D$, $\boldsymbol{v}^\mu \in \mathbb{R}^r$ are random standard normal vectors and $A \in \mathbb{R}^{D\times r}$ is a random orthogonal basis. Then, the resolvent of S and $\tilde{S}$ are given by

$$g_{\tilde{S}}(z) = \frac{1}{D}\operatorname{Tr}\left(\left(\tilde{S} - zI_D\right)^{-1}\right) = \frac{-(\kappa\rho z + \rho - \kappa) - \sqrt{(\kappa\rho z + \rho - \kappa)^2 - 4\rho^2\kappa z}}{2\rho z} \tag{46}$$

$$g_S(z) = \frac{1}{D}\operatorname{Tr}\left((S - zI_D)^{-1}\right) = \frac{\kappa + \rho - \kappa\rho z - 2 - \sqrt{(\rho\kappa z + \rho - \kappa)^2 - 4\rho^2\kappa z}}{2z} \tag{47}$$

for $z < 0$ respectively, where $\rho = r/D$, $\kappa = M_0/D$ and $D, r, M_0 \to \infty$.

Proof. The random matrix $\tilde{S}$ is classified as a Wishart matrix (or a scaled sample covariance matrix). The empirical spectral distribution of such a matrix is known to converge to the Marchenko-Pastur distribution in the given asymptotic limit. Therefore, Eq. (46), which represents its Stieltjes transform, follows directly from this standard result in random matrix theory (e.g., see [Potters and Bouchaud, 2020]).

For the random matrix S, let $P_S(\lambda)$ and $P_{\tilde{S}}(\lambda)$ denote the characteristic polynomials of S and $\tilde{S}$, respectively. Then we have

$$P_S(\lambda) = \det(\lambda I_D - S) = \lambda^{D-r}\det\left(\lambda I_r - \tilde{S}\right) = \lambda^{D-r}P_{\tilde{S}}(\lambda) \tag{48}$$

which implies that the non-zero eigen values of S and $\tilde{S}$ coincide, including multiplicities. Therefore,

$$g_S(z) = -\frac{D-r}{D}\frac{1}{z} + \frac{r}{D}g_{\tilde{S}}(z) = -(1-\rho)\frac{1}{z} + \rho g_{\tilde{S}}(z) \tag{49}$$

$$= \frac{1}{D}\operatorname{Tr}\left((S - zI_D)^{-1}\right) = \frac{\kappa + \rho - \kappa\rho z - 2 - \sqrt{(\rho\kappa z + \rho - \kappa)^2 - 4\rho^2\kappa z}}{2z} \tag{50}$$

as desired in Eq. (47). □

D.2 Recurrence Relations for Matrix Moments

Having established the resolvent functions, we now derive recurrence relations that allow us to compute higher-order moments involving powers of the random matrices and their resolvents. These relations are essential for the subsequent calculations in the saddle-point equations.

Lemma D.2. *Let $\mathcal{M} = aS + bI_D \in \mathbb{R}^{D\times D}$, where $a, b > 0$ and S is defined in Lemma D.1. Define $E_{n,m}$ by*

$$E_{n,m} = \frac{1}{D}\mathbb{E}\left[\operatorname{Tr} S^n\mathcal{M}^{-m}\right] = \frac{1}{D}\mathbb{E}\left[\operatorname{Tr} S^n(aS + bI_D)^{-m}\right] \tag{51}$$

for integer $n, m \geq 0$. Then, the following recurrence relation holds:

$$E_{0,m} = \frac{1}{b^m}\frac{1}{(m-1)!}g_S^{(m-1)}(z) \quad (\text{for } m \geq 1) \tag{52}$$

$$E_{n,0} = 1 \quad (n \geq 1) \tag{53}$$

$$E_{n,m} = \frac{1}{b}E_{n-1,m-1} + zE_{n-1,m} \quad (n \geq 1, m \geq 1), \tag{54}$$

where $z = -b/a$ and $g_S^{(m-1)}(z)$ denotes the $(m-1)$-th derivative of $g_S(z)$.

Proof. The inverse powers of $\mathcal{M}$ are given by:

$$\mathcal{M}^{-m} = \frac{1}{b^m}(S - zI_D)^{-m} \tag{55}$$

We use the standard definition of the Stieltjes transform, $g_S(z) = \frac{1}{D}\mathbb{E}[\operatorname{tr}((S - zI_D)^{-1})]$. The k-th derivative of the resolvent $(S - zI_D)^{-1}$ with respect to z is:

$$\frac{d^k}{dz^k}(S - zI_D)^{-1} = k!(S - zI_D)^{-(k+1)} \tag{56}$$

From this, we can express the m-th power of the resolvent as:

$$(S - zI_D)^{-m} = \frac{1}{(m-1)!}\frac{d^{m-1}}{dz^{m-1}}(S - zI_D)^{-1} \tag{57}$$

The k-th derivative of the Stieltjes transform is therefore:

$$g_S^{(k)}(z) = \frac{d^k g_S(z)}{dz^k} = \frac{1}{D}\mathbb{E}\left[\mathrm{Tr}\left(\frac{d^k}{dz^k}(S - zI_D)^{-1}\right)\right] = \frac{k!}{D}\mathbb{E}\left[\mathrm{tr}\left((S - zI_D)^{-(k+1)}\right)\right] \tag{58}$$

First, we prove the expression for $E_{0,m}$. By definition, for $m \geq 1$:

$$\begin{aligned}
E_{0,m} &= \frac{1}{D}\mathbb{E}\left[\mathrm{tr}(\mathcal{M}^{-m})\right] && (59)\\
&= \frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(\frac{1}{b^m}(S - zI_D)^{-m}\right)\right] && (60)\\
&= \frac{1}{b^m}\frac{1}{D}\mathbb{E}\left[\mathrm{Tr}\left(\frac{1}{(m-1)!}\frac{d^{m-1}}{dz^{m-1}}(S - zI_D)^{-1}\right)\right] && (61)\\
&= \frac{1}{b^m}\frac{g_S^{(m-1)}(z)}{(m-1)!} && (62)
\end{aligned}$$

Second, the expression for $E_{n,0}$ for $n \geq 1$ follows directly from the definition:

$$E_{n,0} = \frac{1}{D}\mathbb{E}\left[\mathrm{Tr}(S^n\mathcal{M}^0)\right] = \frac{1}{D}\mathbb{E}\left[\mathrm{tr}(S^n)\right] = 1$$

Third, we prove the recurrence relation for $n \geq 1$ and $m \geq 1$. We start from the definition of $E_{n,m}$ and use the identity $S^n = S^{n-1}((S - zI_D) + zI_D)$.

$$\begin{aligned}
E_{n,m} &= \frac{1}{D}\mathbb{E}\left[\mathrm{Tr}(S^n\mathcal{M}^{-m})\right] && (63)\\
&= \frac{1}{D}\mathbb{E}\left[\mathrm{Tr}\left(\left(S^{n-1}(S - zI_D) + zS^{n-1}\right)\mathcal{M}^{-m}\right)\right] && (64)\\
&= \frac{1}{D}\mathbb{E}\left[\mathrm{Tr}\left(S^{n-1}(S - zI_D)\mathcal{M}^{-m}\right)\right] + \frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(zS^{n-1}\mathcal{M}^{-m}\right)\right]. && (65)
\end{aligned}$$

For the first term of Eq. (65), we substitute $(S - zI_D) = \frac{1}{b}\mathcal{M}$:

$$\frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(S^{n-1}(S - zI_D)\mathcal{M}^{-m}\right)\right] = \frac{1}{b}\frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(S^{n-1}\mathcal{M}^{-(m-1)}\right)\right] = \frac{1}{b}E_{n-1,m-1}. \tag{66}$$

The second term is:

$$\frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(zS^{n-1}\mathcal{M}^{-m}\right)\right] = z\left(\frac{1}{D}\mathbb{E}\left[\mathrm{tr}\left(S^{n-1}\mathcal{M}^{-m}\right)\right]\right) = zE_{n-1,m}. \tag{67}$$

Combining the two terms gives the recurrence relation:

$$E_{n,m} = \frac{1}{b}E_{n-1,m-1} + zE_{n-1,m}, \tag{68}$$

which completes the proof. □

D.3 Explicit Expressions for Matrix Moments

The recurrence relations established in the previous lemma can be used to derive explicit expressions for specific combinations of matrix powers and resolvents that frequently appear in our calculations.

Proposition D.3. *Let S, $\mathcal{M}$ and z be as defined in Lemma D.1 and D.2. Then, the following relations hold:*

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}\mathcal{M}^{-1}\right] = \frac{g_S(z)}{b} \tag{69}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}\mathcal{M}^{-2}\right] = \frac{g_S'(z)}{b^2} \tag{70}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}S\mathcal{M}^{-1}\right] = \frac{1+zg_S(z)}{b} \tag{71}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}S\mathcal{M}^{-2}\right] = \frac{g_S(z)+zg_S'(z)}{b^2} \tag{72}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}S^2\mathcal{M}^{-1}\right] = \frac{1+z+z^2g_S(z)}{b} \tag{73}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}S^2\mathcal{M}^{-2}\right] = \frac{1+2zg_S(z)+z^2g_S'(z)}{b^2} \tag{74}$$

$$\frac{1}{D}\mathbb{E}\left[\operatorname{Tr}S^3\mathcal{M}^{-2}\right] = \frac{1+2z+3z^2g_S(z)+z^3g_S'(z)}{b^2}. \tag{75}$$

Proof. The proof follows directly from Lemma D.2. □

D.4 Trace Identity for Projected Matrices

Finally, we establish a trace identity that relates the trace of a projected matrix to the trace of its lower-dimensional counterpart.

Lemma D.4. *Let A, S and $\tilde{S}$ be as defined in Lemma D.1. For scalars $a, b, l > 0$, the following identity holds:*

$$\operatorname{Tr}\left(AA^\top(aS+bI_D)^{-l}\right) = \operatorname{Tr}\left((a\tilde{S}+bI_r)^{-l}\right) \tag{76}$$

Proof. We prove this identity by performing a change of basis. Since the r columns of A are orthonormal, we can extend this set to form a complete orthonormal basis for $\mathbb{R}^D$. Let $B \in \mathbb{R}^{D\times(D-r)}$ be a matrix whose columns form an orthonormal basis for the orthogonal complement of the column space of A. The resulting matrix $U = [A\,B] \in \mathbb{R}^{D\times D}$ is an orthogonal matrix, satisfying $U^\top U = UU^\top = I_D$.

The condition $AA^\top S = S$ implies that the column space of S is contained within that of A. Consequently, S must annihilate any vector orthogonal to the column space of A. This means $B^\top S = 0$, which can be verified as follows:

$$B^\top S = B^\top(AA^\top S) = (B^\top A)A^\top S = 0. \tag{77}$$

In the new basis defined by U, the matrix S is represented by $U^\top SU$. This similarity transformation reveals a block upper-triangular structure:

$$U^\top SU = \begin{pmatrix}A^\top\\B^\top\end{pmatrix} S \begin{pmatrix}A & B\end{pmatrix} = \begin{pmatrix}A^\top SA & A^\top SB\\ B^\top SA & B^\top SB\end{pmatrix} = \begin{pmatrix}A^\top SA & A^\top SB\\ 0 & 0\end{pmatrix}. \tag{78}$$

Now, we express the left-hand side (LHS) of Eq. (78) in this basis. Using the cyclic property of the trace, $\operatorname{Tr}(X) = \operatorname{Tr}(U^\top XU)$, we have:

$$\text{LHS} = \operatorname{Tr}\left(U^\top(AA^\top)UU^\top(aS+bI_D)^{-l}U\right). \tag{79}$$

The projection matrix $AA^\top$ and the term $(aS+bI_D)$ transform as:

$$U^\top AA^\top U = \begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix} \tag{80}$$

$$U^\top(aS+bI_D)U = a(U^\top SU)+bI_D = \begin{pmatrix}aA^\top SA+bI_r & aA^\top SB\\ 0 & bI_{D-r}\end{pmatrix}. \tag{81}$$

The inverse of a block upper-triangular matrix is also block upper-triangular, and squaring it preserves this structure. We only need the diagonal blocks for the trace calculation:

$$\left(U^\top(aS+bI_D)U\right)^{-l} = \begin{pmatrix}(aA^\top SA+bI_r)^{-l} & *\\ 0 & (bI_{D-r})^{-l}\end{pmatrix}, \tag{82}$$

where $*$ denotes the off-diagonal block, which is irrelevant for the trace. Substituting these into the expression for the LHS:

$$\text{LHS} = \text{Tr}\left(\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} (aA^\top SA + bI_r)^{-l} & * \\ 0 & b^{-l}I_{D-r} \end{pmatrix}\right) \tag{83}$$

$$= \text{Tr}\begin{pmatrix} (aA^\top SA + bI_r)^{-l} & * \\ 0 & 0 \end{pmatrix} \tag{84}$$

$$= \text{Tr}\big((aA^\top SA + bI_r)^{-l}\big) \tag{85}$$

This is identical to the right-hand side of the proposition, which completes the proof. □

E Replica Calculation

E.1 Results with Complete Statements

In this subsection, we present the main results (Result 4.1, 4.2, 5.1, 5.2, 6.2) with complete statements, including the expression for the generalization error, which are derived using the replica method. We analyze a more general model that incorporates a regularization term into the loss function. The results for the unregularized model can be recovered by simply setting the regularization strength $\lambda = 0$.

Result E.1. *(Complete Statement of Optimal Parameter Matrix) Let the optimal parameter matrix $W^* \in \mathbb{R}^{D\times D}$ be the solution that minimizes the cost function $\mathcal{L}(W)$, defined as:*

$$\mathcal{L}(W) = \frac{1}{2}\sum_{\mu=1}^{M}\left(\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\frac{w_i^\mu w_j^\mu}{D} - W_{ij}\right)H_{ij}^\mu\right)^2 + \frac{M_0\lambda}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}W_{ij}^2 \tag{86}$$

For a given fixed set of tasks $\mathcal{W}_0 = \{\boldsymbol{w}^1, \dots, \boldsymbol{w}^{M_0}\}$, there exist non-negative scalar constants $\hat{q}, \hat{m}, \hat{\chi}, \hat{\bar{q}}, \hat{\bar{\chi}}$ such that the optimal matrix W^* is given by the expression:

$$W^* \overset{\mathrm{d}}{=} \underset{W}{\operatorname{argmin}}\left[\frac{\lambda+\hat{\bar{q}}}{2}\operatorname{Tr}(WW^\top) + \frac{\hat{q}}{2}\operatorname{Tr}(WSW^\top) - \operatorname{Tr}\left(\left(\sqrt{\hat{\chi}}T + \sqrt{\hat{\bar{\chi}}}R + \hat{m}S\right)W^\top\right)\right] \tag{87}$$

$$\overset{\mathrm{d}}{=} \left(\sqrt{\hat{\chi}}T + \sqrt{\hat{\bar{\chi}}}R + \hat{m}S\right)\left(\hat{q}S + (\lambda+\hat{\bar{q}})I_D\right)^{-1}, \tag{88}$$

where the matrices $T, R, S \in \mathbb{R}^{D\times D}$ are defined as follows:

$$T = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\boldsymbol{\xi}\boldsymbol{w}^{\mu\top}, \quad R = (R_{ij})_{i,j=1}^{D} \sim \mathcal{N}\left(0, \frac{1}{M_0}\right), \quad S = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top}, \tag{89}$$

with $\boldsymbol{\xi} \in \mathbb{R}^D \sim \mathcal{N}(0, I_D)$.

Result E.2. *(Parameter Derivation) The order parameters $q, m, q_0, m_0, \bar{q}, \bar{m}$ (with auxiliary parameters χ) in Result. E.1 and the conjugate parameters $\hat{q}, \hat{m}, \hat{\chi}, \hat{\bar{q}}, \hat{\bar{\chi}}$ in Result. E.3 are given by the solution of the following system of equations:*

$$q = \frac{1}{\kappa\hat{q}^2}\left[\kappa\hat{m}^2\left(1 + 2z + 3z^2 g_S(z) + z^3 g_S'(z)\right) + \hat{\bar{\chi}}(g_S(z) + z g_S'(z)) + \hat{\chi}\left(1 + 2z g_S(z) + z^2 g_S'(z)\right)\right] \tag{90}$$

$$\bar{q} = \frac{1}{\kappa\hat{q}^2}\left[\kappa\hat{m}^2\left(1 + 2z g_S(z) + z^2 g_S'(z)\right) + \hat{\bar{\chi}} g_S'(z) + \hat{\chi}(g_S(z) + z g_S'(z))\right] \tag{91}$$

$$q_0 = \frac{1}{\rho\kappa\hat{q}^2}\left[\kappa\hat{m}^2\left(1 + 2z g_S(z) + z^2 g_S'(z)\right) + \rho\hat{\bar{\chi}} g_{\tilde{S}}'(z) + \hat{\chi}(g_S(z) + z g_S'(z))\right] \tag{92}$$

$$m = \frac{\hat{m}}{\hat{q}}\left(1 + z + z^2 g_S(z)\right) \tag{93}$$

$$\bar{m} = \frac{\hat{m}}{\hat{q}}(1 + z g_S(z)) \tag{94}$$

$$m_0 = \frac{1}{\rho}\bar{m} \tag{95}$$

$$\chi = \frac{1}{\kappa\hat{q}}(1 + z g_S(z)) \tag{96}$$

$$\bar{\chi} = \frac{1}{\kappa\hat{q}} g_S(z), \tag{97}$$

where $z = -\frac{\hat{\bar{q}}+\lambda}{\hat{q}}$ *and*

$$\hat{q} = \gamma \frac{1}{1+\chi+\frac{1}{\alpha}(1+\sigma^2)\bar{\chi}} \tag{98}$$

$$\hat{\bar{q}} = \gamma \frac{\frac{1}{\alpha}(1+\sigma^2)}{1+\chi+\frac{1}{\alpha}(1+\sigma^2)\bar{\chi}} \tag{99}$$

$$\hat{m} = \gamma \frac{1}{1+\chi+\frac{1}{\alpha}(1+\sigma^2)\bar{\chi}} \tag{100}$$

$$\hat{\chi} = \gamma \frac{\frac{1}{\alpha}(1+\sigma^2)\bar{q}+q-2m+1+\sigma^2}{\left(1+\chi+\frac{1}{\alpha}(1+\sigma^2)\bar{\chi}\right)^2} \tag{101}$$

$$\hat{\bar{\chi}} = \gamma \frac{1}{\alpha}(1+\sigma^2) \frac{\frac{1}{\alpha}(1+\sigma^2)\bar{q}+q-2m+1+\sigma^2}{\left(1+\chi+\frac{1}{\alpha}(1+\sigma^2)\bar{\chi}\right)^2}. \tag{102}$$

Here, we define $z = -\frac{\hat{\bar{q}}+\lambda}{\hat{q}}$ *and the resolvents* $g_S(z)$ *and* $g_{\tilde{S}}(z)$ *are defined in Lemma D.1.*

Result E.3. *(Complete Statement of Generalization Error) There exist non-negative scalar constants* $q, m, q_0, m_0, \bar{q}, \bar{m}$ *such that the generalization error of the optimal parameter matrix* W^* *is given by:*

$$\mathcal{E}_{\text{TM}} = 1 - 2m + q + \frac{\bar{q}}{\tilde{\alpha}} \tag{103}$$

$$\mathcal{E}_{\text{IDG}} = 1 - 2m_0 + q_0 + \frac{\bar{q}}{\tilde{\alpha}} \tag{104}$$

$$\mathcal{E}_{\text{ODG}} = 1 - 2\bar{m} + \bar{q} + \frac{\bar{q}}{\tilde{\alpha}}. \tag{105}$$

Proposition E.4. *Under the condition that* $\kappa\gamma > 1$*, the order parameters in* $\alpha \gg 1$ *are given by:*

$$m = 1 + \mathcal{O}\left(\frac{1}{\alpha}\right), \qquad \hat{m} = \gamma - \frac{1}{\kappa} + \mathcal{O}\left(\frac{1}{\alpha}\right) \tag{106}$$

$$\bar{m} = \min(\kappa,\rho) + \mathcal{O}\left(\frac{1}{\alpha}\right) \tag{107}$$

$$m_0 = \frac{1}{\rho}\min(\kappa,\rho) + \mathcal{O}\left(\frac{1}{\alpha}\right) \tag{108}$$

$$q = 1 + \frac{\sigma^2}{\kappa\gamma-1}\min(\kappa,\rho) + \mathcal{O}\left(\frac{1}{\alpha}\right) \qquad \hat{q} = \gamma - \frac{1}{\kappa} + \mathcal{O}\left(\frac{1}{\alpha}\right) \tag{109}$$

$$\bar{q} = \frac{\sigma^2}{1+\sigma^2}\frac{1-\min(\kappa,\rho)}{\kappa\gamma-1}\alpha + \min(\kappa,\rho)\left[1+\frac{1-\min(\kappa,\rho)}{\kappa\gamma-1}\right] + \mathcal{O}\left(\frac{1}{\alpha}\right) \qquad \hat{\bar{q}} = \frac{1+\sigma^2}{\alpha}\left(\gamma-\frac{1}{\kappa}\right) + \mathcal{O}\left(\frac{1}{\alpha^2}\right) \tag{110}$$

$$q_0 = \begin{cases} \frac{\kappa}{\rho} + \frac{\kappa(\rho-\kappa)}{\rho}\frac{1}{\kappa\gamma-1} + \frac{\rho-\kappa}{\rho(\kappa\gamma-1)}\frac{\sigma^2}{1+\sigma^2}\alpha + \mathcal{O}\left(\frac{1}{\alpha}\right) & \text{for } \rho > \kappa \\ 1 + \frac{\sigma^2}{\kappa\gamma-1}\frac{\kappa\rho}{\kappa-\rho} + \mathcal{O}\left(\frac{1}{\alpha}\right) & \text{for } \rho < \kappa \end{cases} \tag{111}$$

$$\chi = \frac{\min(\kappa,\rho)}{\kappa\gamma-1} + \mathcal{O}\left(\frac{1}{\alpha}\right) \tag{112}$$

$$\hat{\chi} = \left(\gamma-\frac{1}{\kappa}\right)\left(\sigma^2 + \min(\kappa,\rho)\frac{1+\sigma^2}{\alpha}\right) + \mathcal{O}\left(\frac{1}{\alpha^2}\right) \tag{113}$$

$$\bar{\chi} = \frac{1-\min(\kappa,\rho)}{\kappa\gamma-1}\frac{1}{1+\sigma^2}\alpha + \mathcal{O}(1) \tag{114}$$

$$\hat{\bar{\chi}} = \frac{1+\sigma^2}{\alpha}\left(\gamma-\frac{1}{\kappa}\right)\left(\sigma^2 + \min(\kappa,\rho)\frac{1+\sigma^2}{\alpha}\right) + \mathcal{O}\left(\frac{1}{\alpha^3}\right) \tag{115}$$

The key findings in the main paper (Results 4.1, 4.2, 5.1, and 5.2) are recovered from Result E.1 and Proposition E.4. Similarly, Result 6.2 is recovered from Result E.3 and Proposition E.4. In the following, we provide the detailed replica calculations to sequentially derive these foundational results: Results E.1, E.2, E.3, and Proposition E.4.

E.2 Replica System and Replicated Partition Function

In this subsection, we outline the replica formalism for evaluating the moments of the solution W^*, which forms the basis for the statistical characterization of the estimator in the high-dimensional limit. To derive Result E.1, we first rely on two fundamental assumptions. The first assumption addresses the well-posedness of the statistical problem itself.

Assumption E.5 (Identifiability from Moments of a Scalar Statistic)**.** *Let X be a random matrix taking values in $\mathbb{R}^{D\times D}$, and let $g : \mathbb{R}^{D\times D} \to \mathbb{R}$ be an arbitrary measurable function. We assume that the probability law of X is uniquely determined by the sequence of moments of the scalar statistic $g(X)$. More formally, for any two random matrices X and X', the condition*

$$\mathbb{E}[(g(X))^n] = \mathbb{E}[(g(X'))^n] \quad \text{for all } n \in \mathbb{N} \tag{116}$$

implies that X and X' are identically distributed.

This is a technical assumption, positing that the random variable X does not exhibit pathological behavior (such as that of a log-normal distribution) where its moments fail to uniquely define its distribution. Assuming that W^* satisfies this property for a fixed task set $\mathcal{W}_0$, our goal is to compute its integer moments, $\mathbb{E}[g(W^*)^p]$ for $p \in \mathbb{N}$.

Our starting point is to re-interpret the solution of the optimization problem, W^*, from a statistical mechanics perspective. For a fixed set of training instances $\mathcal{D} = \{H^\mu \mid 1 \le \mu \le M\}$, we treat W^* as a random variable drawn from a Gibbs-Boltzmann distribution, where the loss function $\mathcal{L}(W)$ acts as the energy function. In the zero-temperature limit ($\beta \to \infty$), this distribution concentrates on the global minimum of the loss:

$$W^* = \operatorname*{argmin}_W \mathcal{L}(W) \sim p(W \mid \mathcal{D}) = \lim_{\beta\to\infty} \frac{\exp(-\beta\mathcal{L}(W \mid \mathcal{D}))}{\int \mathrm{d}W \exp(-\beta\mathcal{L}(W \mid \mathcal{D}))}. \tag{117}$$

Here, $\mathcal{L}(W \mid \mathcal{D})$ is the loss function for the model W on the fixed learning instances $\mathcal{D}$, and $\mathrm{d}W = \prod_{ij} \mathrm{d}W_{ij}$. The normalization factor of Eq. (117) is the **partition function** Z of the system:

$$Z = \int \mathrm{d}W \exp(-\beta\mathcal{L}(W \mid \mathcal{D})). \tag{118}$$

Using this notation, the p-th moment of W^*, averaged over the training data, is expressed as:

$$\mathbb{E}_{\mathcal{D}}[g(W^*)^p] = \mathbb{E}_{\mathcal{D}} \lim_{\beta\to\infty} \left(\frac{1}{Z} \int \mathrm{d}W\, g(W) \exp(-\beta\mathcal{L}(W \mid \mathcal{D})) \right)^p \tag{119}$$

$$= \mathbb{E}_{\mathcal{D}} \lim_{\beta\to\infty} \lim_{n\to 0} Z^{n-p} \left(\int \mathrm{d}W\, g(W) \exp(-\beta\mathcal{L}(W \mid \mathcal{D})) \right)^p. \tag{120}$$

However, the expression in Eq. (120) is difficult to average over the data distribution $\mathcal{D}$ because the data-dependent partition function Z appears in the denominator.

To circumvent this difficulty, we employ the replica trick. The key insight is to represent the problematic term $Z^{n'-p}$ as an integral over $n'-p$ independent copies (or 'replicas') of the system, which is valid for any integer $n' > p$:

$$Z^{n'-p} = \int \mathrm{d}W_{n'-p} \exp\left(-\beta \sum_{a=1}^{n'-p} \mathcal{L}(W^a \mid \mathcal{D}) \right). \tag{121}$$

Here, $\mathrm{d}W_s = \mathrm{d}W^1\, \mathrm{d}W^2 \cdots \mathrm{d}W^s$. This identity forms the basis for our second key assumption, which involves analytically continuing this expression from the domain of integers n' to the limit $n' \to 0$.

Assumption E.6 (Analytic Continuation of the Partition Function)**.** *The identity in Eq. (121), which is guaranteed to hold for integers $n' > p$, is assumed to remain valid in the limit $n' \to 0$. That is, for any integer p,*

$$\lim_{n'\to 0} Z^{n'-p} = \lim_{n'\to 0} \int \mathrm{d}W_{n'-p} \exp\left(-\beta \sum_{a=1}^{n'-p} \mathcal{L}(W^a \mid \mathcal{D}) \right). \tag{122}$$

While not mathematically rigorous for all systems, this assumption is standard practice in the replica method. With this assumption, Eq. (120) can be rewritten by combining all terms into a single expectation over n replicas:

$$\mathbb{E}_{\mathcal{D}}[g(W^*)^p] = \mathbb{E}_{\mathcal{D}} \lim_{\beta\to\infty} \lim_{n\to 0} Z^{n-p} \left(\int \mathrm{d}W\, g(W) \exp(-\beta \mathcal{L}(W \mid \mathcal{D})) \right)^p \tag{123}$$

$$= \mathbb{E}_{\mathcal{D}} \lim_{\beta\to\infty} \lim_{n\to 0} Z^{n-p} \int \mathrm{d}W_p \left(\prod_{a=1}^{p} g(W^a) \right) \exp\left(-\beta \sum_{a=1}^{p} \mathcal{L}(W^a \mid \mathcal{D}) \right) \tag{124}$$

$$= \lim_{\beta\to\infty} \lim_{n\to 0} \mathbb{E}_{\mathcal{D}} \int \mathrm{d}W_n \left(\prod_{a=1}^{p} g(W^a) \right) \exp\left(-\beta \sum_{a=1}^{n} \mathcal{L}(W^a \mid \mathcal{D}) \right) \tag{125}$$

$$= \lim_{n\to 0} \frac{\lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}} \int \mathrm{d}W_n \left(\prod_{a=1}^{p} g(W^a) \right) \exp(-\beta \sum_{a=1}^{n} \mathcal{L}(W^a \mid \mathcal{D}))}{\lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}}[Z^n]} \tag{126}$$

$$= \lim_{n\to 0} \frac{\lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}} \int \mathrm{d}W_n \left(\prod_{a=1}^{p} g(W^a) \right) \exp(-\beta \sum_{a=1}^{n} \mathcal{L}(W^a \mid \mathcal{D}))}{\lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}} \int \mathrm{d}W_n \, \exp(-\beta \sum_{a=1}^{n} \mathcal{L}(W^a \mid \mathcal{D}))}. \tag{127}$$

The final expression in Eq. (127) is the central result of this formalism. It shows that the desired moment, $\mathbb{E}_{\mathcal{D}}[g(W^*)^p]$, can be interpreted as a p-point correlation function of the replicated variables $\{W^1, \ldots, W^n\}$ drawn from an effective probability distribution:

$$p(W^1, \ldots, W^n) \propto \lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}} \exp\left(-\beta \sum_{a=1}^{n} \mathcal{L}(W^a \mid \mathcal{D}) \right), \tag{128}$$

in the $n \to 0$ limit. The main advantage of this approach is that the difficult average over the data distribution $\mathcal{D}$ has been absorbed into the structure of a new joint distribution over n replicated systems. We call this new system the **replica system**, and its partition function, $\mathbb{E}_{\mathcal{D}}[Z^n]$, the data-averaged **replicated partition function**.

In the subsequent sections, we will analyze this replicated system. By calculating the replicated partition function, we will derive the correlations between replicas, which in turn will allow us to derive Result E.1.

E.3 Analysis of the Data-Averaged Replicated Partition Function

In this subsection, we analyze the data-averaged replicated partition function to derive the statistical properties of the optimal solution. Our immediate goal is to compute the partition function of the probability distribution Eq. (128), which is the data (disorder) average of the replicated partition function:

$$\mathbb{E}[Z^n] = \mathbb{E} \int \mathrm{d}W_n \, \exp\left[-\frac{\beta}{2} \sum_{\mu=1}^{M} \sum_{a=1}^{n} \left(\sum_{ij} \left(\frac{w_i^\mu w_j^\mu}{D} - W_{ij}^a \right) H_{ij}^\mu + \epsilon^\mu \right)^2 - \frac{D\lambda\beta}{2} \sum_{ij} \sum_{a} (W_{ij}^a)^2 \right]. \tag{129}$$

The overall disorder average $\mathbb{E}[\cdots]$ here is performed over three hierarchical stages of data generation. Each stage introduces a different source of randomness:

- $\mathbb{E}_{\mathcal{W}_0}$: This denotes the average over the generation of the base pool of tasks, $\mathcal{W}_0 = \{\boldsymbol{w}^1, \ldots, \boldsymbol{w}^{M_0}\}$. This expectation accounts for the randomness in the underlying low-dimensional structure from which all tasks are derived (i.e., the shared matrix A and latent vectors $\{\boldsymbol{v}^\mu\}$).

- $\mathbb{E}_{\mathcal{W}|\mathcal{W}_0}$: This denotes the average over the sampling of the final training set $\mathcal{W}$. Specifically, it is the average over the process of drawing M tasks uniformly and with replacement from the fixed base pool $\mathcal{W}_0$.

- $\mathbb{E}_{H^\mu|\boldsymbol{w}^\mu}$: This denotes the average over the generation of the data within a single training instance for a given task $\boldsymbol{w}^\mu$. This "in-context" randomness comes from sampling the $L+1$ input vectors $\{\boldsymbol{x}_l^\mu\}$ and the corresponding label noise $\{\epsilon_l^\mu\}$.

Using the above notation, we can isolate the average over the data distribution (H^μ and noise) for a fixed task $\boldsymbol{w}^\mu$ by defining the function $\phi(\boldsymbol{W}, \boldsymbol{w}^\mu)$:

$$\phi(\boldsymbol{W}, \boldsymbol{w}^\mu) = \mathbb{E}_{H^\mu|\boldsymbol{w}^\mu} \exp\left[-\frac{\beta}{2}\sum_a \left(\sum_{ij}\left(\frac{w_i^\mu w_j^\mu}{D} - W_{ij}^a\right)H_{ij}^\mu + \epsilon^\mu\right)^2\right]. \tag{130}$$

With this definition, the replicated partition function becomes:

$$\mathbb{E}[Z^n] = \int \mathrm{d}W_n \, \exp\left(-\frac{D\lambda\beta}{2}\sum_a\sum_{ij}\big(W_{ij}^a\big)^2\right)\mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\{\boldsymbol{w}^\mu\}|\mathcal{W}_0}\prod_{\mu=1}^{M}\phi(\boldsymbol{W}, \boldsymbol{w}^\mu) \tag{131}$$

$$= \int \mathrm{d}W_n \, \exp\left(-\frac{D\lambda\beta}{2}\sum_a\sum_{ij}\big(W_{ij}^a\big)^2\right) \prod_{\mu=1}^{M_0}\mathbb{E}_{\mathcal{W}_0}\left[\phi(\boldsymbol{W}, \boldsymbol{w}^\mu)^{\frac{M}{M_0}}\right] \tag{132}$$

$$= \mathbb{E}_{\mathcal{W}_0}\int \mathrm{d}W_n \, \exp\left(-\frac{D\lambda\beta}{2}\sum_a\sum_{ij}\big(W_{ij}^a\big)^2\right) \exp\left[\frac{M}{M_0}\sum_{\mu=1}^{M_0}\log\phi(\boldsymbol{W}, \boldsymbol{w}^\mu)\right]. \tag{133}$$

In the last step, we assume self-averaging with respect to the tasks $\boldsymbol{w}^\mu$, allowing us to replace the sum over logarithms with its average value, which simplifies the expression to depend on the average of $\log\phi$:

$$\mathbb{E}[Z^n] = \mathbb{E}_{\mathcal{W}_0}\int \mathrm{d}W_n \, \exp\left(-\frac{D\lambda\beta}{2}\sum_a\sum_{ij}\big(W_{ij}^a\big)^2\right) \exp\big[M \ \log\phi(\boldsymbol{W}, \bar{Q}, m)\big]. \tag{134}$$

To proceed, we introduce the following order parameters, which describe the macroscopic state of the system. We define the following order parameters:

$$Q^{ab} = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\frac{1}{D}(W^a\boldsymbol{w}^\mu)^\top(W^b\boldsymbol{w}^\mu) \tag{135}$$

$$\bar{Q}^{ab} = \frac{1}{D}\,\mathrm{Tr}\left((W^a)^\top W^b\right) \tag{136}$$

$$m^a = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\frac{1}{D}\boldsymbol{w}^{\mu\top}W^a\boldsymbol{w}^\mu, \tag{137}$$

where the index $a, b \in \{1, ..., n\}$ denotes the replica index. Q^{ab} measures the task-conditioned overlap between replicas, $\bar{Q}^{ab}$ measures the direct overlap of the weight matrices, and m^a measures the alignment of a replica's weights to the task structure.

We now re-evaluate the data-averaged term ϕ by performing the Gaussian integral. We define the vectors

$$\boldsymbol{v}^a = \begin{pmatrix}\mathsf{vec}\left(\frac{\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top}}{D} - W^a\right)\\ 1\end{pmatrix}, \qquad \boldsymbol{h} = \begin{pmatrix}\mathsf{vec}(H^\mu)\\ \epsilon^\mu\end{pmatrix}, \tag{138}$$

which allows us to write the exponent as a quadratic form. The average over the zero-mean Gaussian vector $\boldsymbol{h}$ yields:

$$\phi(\boldsymbol{W}, \boldsymbol{w}^\mu) = \mathbb{E}_{H^\mu|\boldsymbol{w}^\mu}\exp\left[-\frac{\beta}{2}\sum_a\left(\sum_{ij}\left(\frac{w_i^\mu w_j^\mu}{D} - W_{ij}^a\right)H_{ij}^\mu + \epsilon^\mu\right)^2\right] \tag{139}$$

$$= \mathbb{E}_{H^\mu|\boldsymbol{w}^\mu}\exp\left(-\frac{\beta}{2}\sum_a(\boldsymbol{v}^a\cdot\boldsymbol{h})^2\right) \tag{140}$$

$$= \left[\det\left(I_{D^2+1} + 2\beta C^H V V^\top\right)\right]^{-\frac{1}{2}} \tag{141}$$

$$= \left[\det\left(I_n + 2\beta V^\top C^H V\right)\right]^{-\frac{1}{2}}, \tag{142}$$

where $V \in \mathbb{R}^{(D^2+1)\times n}$ is the matrix whose columns are $\boldsymbol{v}^a$, and C^H is the covariance matrix of $\boldsymbol{h}$. The entries of the matrix $V^\top C^H V$ can be computed as:

$$\left(V^\top C^H V\right)_{ab} = \sum_{(ij),(kl)}^{D^2} V_{(ij),a} V_{(kl),b} C^H_{(ij)(kl)} + \sigma^2 \tag{143}$$

$$= \sum_{(ij),(kl)}^{D^2} \left(\frac{w_i^\mu w_j^\mu}{D} - W_{ij}^a\right)\left(\frac{w_k^\mu w_l^\mu}{D} - W_{kl}^b\right) \frac{\delta_{ik}}{D}\left(\frac{D}{L}\delta_{jl}(1+\sigma^2) + w_j w_l\right) + \sigma^2 \tag{144}$$

$$= \frac{D}{L}(1+\sigma^2)\bar{Q}^{ab} + \left(Q^{ab} - m^a - m^b + 1\right) + \sigma^2. \tag{145}$$

Defining $\mathcal{M} \in \mathbb{R}^{n\times n}$ as the matrix whose elements are given by the expression above, we finally obtain:

$$\phi(\boldsymbol{W}, \boldsymbol{w}^\mu) = \left[\det(I_n + \beta\mathcal{M})\right]^{-\frac{1}{2}} \tag{146}$$

$$= \mathbb{E}_{\boldsymbol{u}} \exp\left[-\frac{\beta}{2}\sum_a (u^a)^2\right] \tag{147}$$

$$= \exp\left[\frac{1}{2}\mathcal{M}^{ab}\frac{\partial}{\partial h^a h^b}\right] \exp\left(-\frac{\beta}{2}\sum_a (h^a)^2\right), \tag{148}$$

where $\boldsymbol{u} = (u^1, \cdots, u^n)^\top \in \mathbb{R}^n$ is a Gaussian random vector with zero mean and covariance given by

$$\mathcal{M}^{ab} = \frac{D}{L}(1+\sigma^2)\bar{Q}^{ab} + \left(Q^{ab} - m^a - m^b + 1\right) + \sigma^2. \tag{149}$$

We enforce the definitions of the order parameters by introducing their integral representations using the Dirac delta function:

$$1 = \int \mathrm{d}Q^{ab}\, \delta\left(DM_0 Q^{ab} - \sum_{\mu=1}^{M_0} \left(W^a \boldsymbol{w}^\mu\right)^\top \left(W^b \boldsymbol{w}^\mu\right)\right) \tag{150}$$

$$= \int \mathrm{d}Q^{ab}\, \mathrm{d}\hat{Q}^{ab}\, \exp\left[-\frac{\hat{Q}^{ab}}{2}\left(DM_0 Q^{ab} - \sum_{\mu=1}^{M_0} \left(W^a \boldsymbol{w}^\mu\right)^\top \left(W^b \boldsymbol{w}^\mu\right)\right)\right] \tag{151}$$

$$1 = \int \mathrm{d}\bar{Q}^{ab}\, \delta\Big(DM_0 \bar{Q}^{ab} - M_0 \operatorname{Tr}\Big(\left(W^a\right)^\top W^b\Big)\Big) \tag{152}$$

$$= \int \mathrm{d}\hat{\bar{Q}}^{ab}\, \mathrm{d}\hat{\bar{Q}}^{ab}\, \exp\left[-\frac{\hat{\bar{Q}}^{ab}}{2}\Big(DM_0 \bar{Q}^{ab} - M_0 \operatorname{Tr}\Big(\left(W^a\right)^\top W^b\Big)\Big)\right] \tag{153}$$

$$1 = \int \mathrm{d}m^a\, \delta\left(DM_0 m^a - \sum_{\mu=1}^{M_0} (\boldsymbol{w}^\mu)^\top W^a \boldsymbol{w}^\mu\right) \tag{154}$$

$$= \int \mathrm{d}\hat{m}^a\, \mathrm{d}m^a\, \exp\left[-\frac{\hat{m}^a}{2}\left(DM_0 m^a - \sum_{\mu=1}^{M_0} (\boldsymbol{w}^\mu)^\top W^a \boldsymbol{w}^\mu\right)\right], \tag{155}$$

where the conjugate parameters $\{\hat{Q}^{ab}, \hat{\bar{Q}}^{ab}\}_{ab}, \{\hat{m}^a\}_a$ are introduced[5]. This allows us to express the averaged replicated partition function as a product of three terms:

$$\mathbb{E}[Z^n] = \int \prod_{ab}\Big(\mathrm{d}Q^{ab}\, \mathrm{d}\hat{Q}^{ab}\, \mathrm{d}\bar{Q}^{ab}\, \mathrm{d}\hat{\bar{Q}}^{ab}\Big) \prod_a (\mathrm{d}m^a\, \mathrm{d}\hat{m}^a)\, G_I G_E G_S, \tag{156}$$

[5]The integration over the conjugate parameters is performed along the imaginary axis to ensure the positivity of the parameters.

where

$$G_I = \exp\left(-\frac{DM_0}{2}\sum_{ab} Q^{ab}\hat{Q}^{ab} - \frac{DM_0}{2}\sum_{ab}\bar{Q}^{ab}\hat{\bar{Q}}^{ab} - DM_0\sum_a m^a\hat{m}^a\right) \tag{157}$$

$$G_S = \mathbb{E}_{\mathcal{W}_0}\int \mathrm{d}W^n \exp\left[\frac{1}{2}\sum_\mu\sum_{ab}\hat{Q}^{ab}\left((W^a\boldsymbol{w}^\mu)^\top(W^b\boldsymbol{w}^\mu)\right) + \frac{M_0}{2}\sum_{ab}\hat{\bar{Q}}^{ab}\,\mathrm{Tr}\left(W^{a\top}W^b\right)\right.$$
$$\left. + \sum_\mu\sum_a \hat{m}^a\boldsymbol{w}^{\mu\top}W^a\boldsymbol{w}^\mu - \frac{\lambda\beta M_0}{2}\sum_a \mathrm{Tr}\left(W^{a\top}W^a\right)\right] \tag{158}$$

$$G_E = \exp\big[M\ \log\ \phi(Q,\bar{Q},m)\big]. \tag{159}$$

E.4 Replica-Symmetric (RS) Ansatz

In this subsection, we now apply the replica-symmetric (RS) ansatz, which is a foundational step in simplifying the replicated system. The core idea is to assume that all n replicas are statistically equivalent and interchangeable. It is important to note that the validity of the RS solution is not mathematically guaranteed for all complex systems; its correctness has only been rigorously proven for a limited class of models. However, particularly for convex optimization problems such as the one considered in this work, the replica method under the Assumption. E.6, E.5, and E.7 has an extensive and successful track record. To date, there are no known instances where the method, when applied to such problems under standard supporting assumptions, has led to physically inconsistent or contradictory results. Therefore, it serves as a crucial and often surprisingly accurate starting point in the analysis of disordered systems.

Assumption E.7 (Replica Symmetry)**.** *The RS ansatz parameterizes the order parameter matrices, which depend on pairs of replica indices (a, b), in terms of a small number of macroscopic variables. This symmetry is expressed as:*

$$Q^{ab} = \frac{\chi}{\beta}\delta_{ab} + q, \tag{160}$$

$$\bar{Q}^{ab} = \frac{\bar{\chi}}{\beta}\delta_{ab} + \bar{q}, \tag{161}$$

$$\hat{Q}^{ab} = -\beta\hat{q}\delta_{ab} + \beta^2\hat{\chi}, \tag{162}$$

$$\hat{\bar{Q}}^{ab} = -\beta\hat{\bar{q}}\delta_{ab} + \beta^2\hat{\bar{\chi}} \tag{163}$$

$$m^a = m \tag{164}$$

$$\hat{m}^a = \beta\hat{m}. \tag{165}$$

This assumption is essential to the replica method for two primary reasons. First, it drastically simplifies the problem by positing that the complex interactions between $n(n-1)/2$ pairs of replicas can be described by a small, fixed number of macroscopic order parameters. This reduces a problem with a large number of degrees of freedom to one of solving a few self-consistent equations. Second, the symmetric structure imposed by the ansatz is a necessary technical step to analytically continue the expressions from integer n to the $n \to 0$ limit, which is the core of the replica trick. Without this simplification, the combinatorial structure of the replica indices would prevent a well-defined continuation.

Upon substituting the RS ansatz into the expression for the replicated partition function, the average over the data distribution $\mathbb{E}[Z^n]$ (Eq. (156)) can be expressed as an integral over the relevant order parameters. Let the set of these parameters be denoted by $\boldsymbol{x} = \{q, \bar{q}, m, \chi, \bar{\chi}, \hat{q}, \hat{\bar{q}}, \hat{m}\}$. The resulting expression takes the form:

$$\mathbb{E}[Z^n] = \int\left(\prod_{i\in\boldsymbol{x}}\mathrm{d}x_i\right)\exp\left[-\beta DM_0 f(\boldsymbol{x}) + \mathcal{O}(1)\right] \tag{166}$$

where $f(\boldsymbol{x})$ is a function of the order parameters.

In the asymptotic limit where $D \to \infty$, this integral can be evaluated using the saddle-point method (also known as the method of steepest descent). Consequently, the values of the order parameters are determined by finding

the specific configuration $\boldsymbol{x}^*$ that extremizes the function $f(\boldsymbol{x})$. This procedure yields the saddle-point equations, which require the partial derivative of f with respect to each order parameter x_i to be zero:

$$\frac{\partial f}{\partial x_i} = 0, \quad \text{for all } x_i \in \boldsymbol{x}. \tag{167}$$

The solution to this system of equations determines the macroscopic state of the system, providing the values of the order parameters that characterize its typical behavior in the asymptotic limit.

E.5 Statistics of the solution W^* (derivation of the Result. E.1)

In this subsection, we derive an equivalent and interpretable expression that characterizes the statistics of W^*. Substituting the RS ansatz into the expression for G_S (Eq. (158)) gives:

$$\begin{aligned} G_S = \mathbb{E}_{\mathcal{W}_0} \int \mathrm{d}W_n \exp\Bigg[& \frac{1}{2}\sum_{\mu}\sum_{ab}\hat{Q}^{ab}\left((W^a\boldsymbol{w}^\mu)^\top(W^b\boldsymbol{w}^\mu)\right) + \frac{M_0}{2}\sum_{ab}\hat{\hat{Q}}^{ab}\,\mathrm{Tr}\left(W^{a\top}W^b\right) \\ & + \sum_{\mu}\sum_{a}\hat{m}^a\boldsymbol{w}^{\mu\top}W^a\boldsymbol{w}^\mu - \frac{\lambda\beta M_0}{2}\sum_a \mathrm{Tr}\left(W^{a\top}W^a\right)\Bigg] \end{aligned} \tag{168}$$

$$\begin{aligned} = \mathbb{E}_{\mathcal{W}_0} \int \mathrm{d}W_n \exp\Bigg[& -\frac{\beta\hat{q}}{2}\sum_{\mu}\sum_{a}(W^a\boldsymbol{w}^\mu)^\top(W^a\boldsymbol{w}^\mu) + \frac{\beta^2\hat{\chi}}{2}\sum_{\mu}\sum_{a}(W^a\boldsymbol{w}^\mu)^\top\sum_b(W^b\boldsymbol{w}^\mu) \\ & - \frac{M_0\beta\hat{\hat{q}}}{2}\sum_a\mathrm{Tr}\left(W^{a\top}W^a\right) + \frac{M_0\beta^2\hat{\hat{\chi}}}{2}\mathrm{Tr}\left(\sum_a W^{a\top}\sum_b W^b\right) \\ & + \beta\hat{m}\sum_{\mu}\sum_a \boldsymbol{w}^{\mu\top}W^a\boldsymbol{w}^\mu - \frac{\lambda\beta M_0}{2}\sum_a\mathrm{Tr}\left(W^{a\top}W^a\right)\Bigg]. \end{aligned} \tag{169}$$

This allows us to decouple the replicas through the following Hubbard-Stratonovich transformation:

$$\exp\left[\frac{\beta^2\hat{\chi}}{2}\sum_{\mu}\sum_a(W^a\boldsymbol{w}^\mu)^\top\sum_b(W^b\boldsymbol{w}^\mu)\right] = \int \mathrm{D}\boldsymbol{\xi}_{M_0}\exp\left[\beta\sqrt{\hat{\chi}}\sum_{\mu}\sum_a(W^a\boldsymbol{w}^\mu)^\top\boldsymbol{\xi}^\mu\right] \tag{170}$$

$$\exp\left[\frac{M_0\beta^2\hat{\hat{\chi}}}{2}\mathrm{Tr}\left(\sum_a W^{a\top}\sum_b W^b\right)\right] = \int \mathrm{D}\Lambda\exp\left[\beta\sqrt{M_0\hat{\hat{\chi}}}\,\mathrm{Tr}\left(W^{a\top}\Lambda\right)\right], \tag{171}$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}^\mu \in \mathbb{R}^D)_{1\le\mu\le M_0}$, $\Lambda \in \mathbb{R}^{D\times D}$, and $\mathrm{D}\boldsymbol{\xi}_{M_0}$ and $\mathrm{D}\Lambda$ are the standard Gaussian measures defined as $\mathrm{D}\boldsymbol{\xi}_{M_0} = \prod_{\mu=1}^{M_0}\exp\left(-\boldsymbol{\xi}^{\mu\top}\boldsymbol{\xi}^\mu/2\right)/\sqrt{2\pi}^D\,\mathrm{d}\boldsymbol{\xi}^\mu$, $\mathrm{D}\Lambda = \prod_{ij}\exp\left(-\Lambda_{ij}^2/2\right)/\sqrt{2\pi}\,\mathrm{d}\Lambda_{ij}$. Using these transformations, we can rewrite the expression for G_S as:

$$\begin{aligned} G_S = \mathbb{E}_{\mathcal{W}_0}\int \mathrm{d}W_n\mathrm{D}\boldsymbol{\xi}\mathrm{D}\Lambda\,\exp\Bigg[-\beta\sum_a\Bigg[& \frac{1}{2}\mathrm{Tr}\left(W^a\left(\hat{q}\sum_\mu\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top} + (\lambda+\hat{\hat{q}})M_0I_D\right)W^{a\top}\right) \\ & - \sum_\mu(W^a\boldsymbol{w}^\mu)^\top\left(\sqrt{\hat{\chi}}\boldsymbol{\xi}^\mu + \hat{m}\boldsymbol{w}^\mu\right) - \sqrt{M_0}\sqrt{\hat{\hat{\chi}}}\,\mathrm{Tr}(W^a\Lambda)\Bigg]\Bigg] \end{aligned} \tag{172}$$

$$= \mathbb{E}_{\mathcal{W}_0}\int \mathrm{d}W_n\mathrm{D}\boldsymbol{\xi}\mathrm{D}\Lambda\;\;\exp\left[-\frac{\beta}{2}\sum_a f(W^a)\right], \tag{173}$$

where we have defined

$$f(W) = \mathrm{Tr}\left[W\left(\hat{q}\sum_\mu\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top} + (\lambda+\hat{\hat{q}})M_0I_D\right)W^\top\right] - 2\sum_\mu\left(\sqrt{\hat{\chi}}\boldsymbol{\xi}^\mu + \hat{m}\boldsymbol{w}^\mu\right)^\top W\boldsymbol{w}^\mu - 2\sqrt{M_0}\sqrt{\hat{\hat{\chi}}}\,\mathrm{Tr}(W\Lambda^\top). \tag{174}$$

By extremizing the exponent of the integrand with respect to the order parameters, we obtain the saddle-point equations:

$$q = \frac{1}{M_0}\mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\boldsymbol{\xi},\Lambda}\left[\frac{1}{D}\sum_{\mu=1}^{M_0}\left(\hat{W}\boldsymbol{w}^\mu\right)^\top\left(\hat{W}\boldsymbol{w}^\mu\right)\right] \tag{175}$$

$$\bar{q} = \frac{1}{D}\mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\boldsymbol{\xi},\Lambda}\,\mathrm{Tr}\left(\hat{W}^\top\hat{W}\right) \tag{176}$$

$$m = \frac{1}{M_0}\mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\boldsymbol{\xi},\Lambda}\left[\frac{1}{D}\sum_{\mu=1}^{M_0}\boldsymbol{w}^{\mu\top}\hat{W}\boldsymbol{w}^\mu\right] \tag{177}$$

$$\chi = \frac{1}{M_0}\mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\boldsymbol{\xi},\Lambda}\left[\sum_{\mu=1}^{M_0}\left(\boldsymbol{w}^{\mu\top}\left(\hat{q}\sum_{\nu=1}^{M_0}\boldsymbol{w}^\nu\boldsymbol{w}^{\nu\top} + (\lambda+\hat{\bar{q}})M_0 I_D\right)^{-1}\boldsymbol{w}^\mu\right)\right] \tag{178}$$

$$\bar{\chi} = \mathbb{E}_{\mathcal{W}_0}\mathbb{E}_{\boldsymbol{\xi},\Lambda}\,\mathrm{Tr}\left[\left(\hat{q}\sum_{\mu=1}^{M_0}\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top} + (\lambda+\hat{\bar{q}})M_0 I_D\right)^{-1}\right]. \tag{179}$$

Here, the optimal weight matrix $\hat{W}$ that minimizes the effective problem $f(W)$ is given by:

$$\hat{W} = \underset{W}{\mathrm{argmin}}\, f(W) \tag{180}$$

$$= \underset{W}{\mathrm{argmin}}\left[\frac{\lambda+\hat{\bar{q}}}{2}\,\mathrm{Tr}\left(WW^\top\right) + \frac{\hat{q}}{2}\,\mathrm{Tr}\left(WSW^\top\right) - \mathrm{Tr}\left(\left(\sqrt{\hat{\chi}}T + \sqrt{\hat{\bar{\chi}}}R + \hat{m}S\right)W^\top\right)\right] \tag{181}$$

$$= \left(\sqrt{\hat{\chi}}T + \sqrt{\hat{\bar{\chi}}}R + \hat{m}S\right)\left(\hat{q}S + (\lambda+\hat{\bar{q}})I_D\right)^{-1}, \tag{182}$$

where we have defined the following matrices:

$$T = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\boldsymbol{\xi}^\mu\boldsymbol{w}^{\mu\top}, \quad R = (R_{ij})_{i,j=1}^{D} \sim \mathcal{N}\left(0, \frac{1}{M_0}\right), \quad S = \frac{1}{M_0}\sum_{\mu=1}^{M_0}\boldsymbol{w}^\mu\boldsymbol{w}^{\mu\top}, \tag{183}$$

The crucial insight from Eq. (173) is that the expression for G_S completely decouples over the replica indices a. This implies that the joint probability distribution for the n-replica system factorizes into a product of identical and independent distributions for each replica W^a.

Our original goal is to compute the moments of a statistic $g(W^*)$, where W^* is the solution to the original optimization problem. Within the replica framework, as seen in Section E.2, the p-th moment $\mathbb{E}_{\mathcal{D}}[g(W^*)^p]$ corresponds to the expectation of a p-body correlation, $\mathbb{E}[g(W^1)g(W^2)\cdots g(W^p)]$, in the $n \to 0$ limit. Due to the factorization, this simplifies significantly:

$$\mathbb{E}_{\mathcal{D}}\left[g(W^*)^p\right] = \mathbb{E}_{\{W^a\}}\left[g(W^1)g(W^2)\cdots g(W^p)\right] \tag{184}$$

$$= \mathbb{E}_{\mathcal{W}_0}\int \mathrm{D}\boldsymbol{\xi}\mathrm{D}\Lambda\left[\lim_{\beta\to\infty}\frac{\int \mathrm{d}W\, g(W)\exp\left[-\frac{\beta}{2}f(W)\right]}{\int \mathrm{d}W\,\exp\left[-\frac{\beta}{2}f(W)\right]}\right]^p \tag{185}$$

$$= \mathbb{E}_{\mathcal{W}_0,\boldsymbol{\xi},\Lambda}\left[g(\hat{W})^p\right]. \tag{186}$$

We now invoke our technical assumption, Assumption E.5, which posits that if the moments of a scalar statistic of two random matrices are identical, then the matrices themselves are identically distributed. This directly leads to our main result:

$$W^* \overset{\mathrm{d}}{=} \hat{W}. \tag{187}$$

This establishes that the solution to the original, complex optimization problem, W^*, is distributionally equivalent to $\hat{W}$, the solution of the much simpler effective quadratic problem defined by $f(W)$. This completes the derivation of Result E.1.

E.5.1 Interpretation of the Auxiliary Fields

It is instructive to provide a physical interpretation of the auxiliary fields, $\boldsymbol{\xi}$ and Λ in Eq. (173). The Hubbard-Stratonovich transformation replaced the difficult average over the data distribution with an average over simpler, data-independent Gaussian fields $\boldsymbol{\xi}$ and Λ. Eq. (186) shows that the auxiliary field $\boldsymbol{\xi}$ and Λ can be interpreted as an effective random field that embodies the statistical fluctuations of the training data. The solution $\hat{W}$ is then understood as the optimal response to a particular realization of these effective data fluctuations.

E.6 Simplified Expressions for the Saddle-Point Equations (derivation of the Result. E.2)

In this subsection, we derive a set of equations that determine the order parameters and their conjugate variables. By simplifying the saddle-point equations under the RS assumption, we obtain expressions suitable for numerical computation.

E.6.1 Order Parameters

Next, we derive the simplified expressions for the saddle-point equations in Eqs. (175)-(179). First, we define $\mathcal{M} = \hat{q}S + (\lambda + \hat{\bar{q}})I_D$. Noting that $\mathcal{M}^{-1}$ is commuting with S, $T^\top T = D/M_0 S$, $AA^\top S = S$, and Lemma D.3, we have the following relations:

$$
\begin{aligned}
q &= \frac{1}{D}\mathbb{E}\Big[\mathrm{Tr}\Big(SW^{*\top}W^*\Big)\Big] && (188)\\
&= \frac{1}{D}\mathbb{E}\big[\mathrm{Tr}\big(\hat{\chi}S\mathcal{M}^{-1}T^\top T\mathcal{M}^{-1} + \hat{\bar{\chi}}S\mathcal{M}^{-1}R^\top R\mathcal{M}^{-1} + \hat{m}^2 S\mathcal{M}^{-1}S^2\mathcal{M}^{-1}\big)\big] && (189)\\
&= \frac{1}{M_0}\big[\hat{\chi}\mathbb{E}\big[\mathrm{Tr}\big(S^2\mathcal{M}^{-2}\big)\big] + \hat{\bar{\chi}}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-2}\big)\big] + \kappa\hat{m}^2\mathbb{E}\big[\mathrm{Tr}\big(S^3\mathcal{M}^{-2}\big)\big]\big] && (190)\\
&= \frac{1}{\kappa\hat{q}^2}\big[\hat{\chi}\big(1 + 2zg_S(z) + z^2 g_S'(z)\big) + \hat{\bar{\chi}}(g_S(z) + zg_S'(z)) + \kappa\hat{m}^2\big(1 + 2z + 3z^2 g_S(z) + z^3 g_S'(z)\big)\big] && (191)
\end{aligned}
$$

$$
\begin{aligned}
\bar{q} &= \frac{1}{D}\mathbb{E}\Big[\mathrm{Tr}\Big(W^{*\top}W^*\Big)\Big] && (192)\\
&= \frac{1}{D}\mathbb{E}\big[\mathrm{Tr}\big(\hat{\chi}\mathcal{M}^{-1}T^\top T\mathcal{M}^{-1} + \hat{\bar{\chi}}\mathcal{M}^{-1}R^\top R\mathcal{M}^{-1} + \hat{m}^2\mathcal{M}^{-1}S^2\mathcal{M}^{-1}\big)\big] && (193)\\
&= \frac{1}{M_0}\big[\hat{\chi}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-2}\big)\big] + \hat{\bar{\chi}}\mathbb{E}\big[\mathrm{Tr}\big(\mathcal{M}^{-2}\big)\big] + \kappa\hat{m}^2\mathbb{E}\big[\mathrm{Tr}\big(S^2\mathcal{M}^{-2}\big)\big]\big] && (194)\\
&= \frac{1}{\kappa\hat{q}^2}\big[\hat{\chi}(g_S(z) + zg_S'(z)) + \hat{\bar{\chi}}g_S'(z) + \kappa\hat{m}^2\big(1 + 2zg_S(z) + z^2 g_S'(z)\big)\big] && (195)
\end{aligned}
$$

$$
m = \frac{1}{D}\mathbb{E}[\mathrm{Tr}(SW^*)] = \frac{\hat{m}}{D}\mathbb{E}\big[\mathrm{Tr}\big(S^2\mathcal{M}^{-1}\big)\big] = \frac{\hat{m}}{\hat{q}}\big(1 + z + z^2 g_S(z)\big) \tag{196}
$$

$$
\bar{m} = \frac{1}{D}\mathbb{E}[\mathrm{Tr}(W^*)] = \frac{\hat{m}}{D}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-1}\big)\big] = \frac{\hat{m}}{\hat{q}}(1 + zg_S(z)) \tag{197}
$$

$$
\chi = \frac{1}{M_0}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-1}\big)\big] = \frac{1}{\kappa\hat{q}}(1 + zg_S(z)) \tag{198}
$$

$$
\bar{\chi} = \frac{1}{M_0}\mathbb{E}\big[\mathrm{Tr}\big(\mathcal{M}^{-1}\big)\big] = \frac{1}{\kappa\hat{q}}g_S(z). \tag{199}
$$

Here, we used Lemma D.4 to simplify Eq. (202) to Eq. (203). We further introduce the order parameters q_0 and m_0, which characterize the correlation between the weight matrix and the structure A in the context of the IDG error, as follows:

$$q_0 = \frac{1}{r}\mathbb{E}\Big[\mathrm{Tr}\Big(AA^\top W^{*\top} W^*\Big)\Big] \tag{200}$$

$$= \frac{1}{r}\mathbb{E}\big[\mathrm{Tr}\big(\hat{\chi}AA^\top \mathcal{M}^{-1}T^\top T\mathcal{M}^{-1} + \hat{\tilde{\chi}}AA^\top \mathcal{M}^{-1}R^\top R\mathcal{M}^{-1} + \hat{m}^2 AA^\top \mathcal{M}^{-1}S^2\mathcal{M}^{-1}\big)\big] \tag{201}$$

$$= \frac{1}{r\kappa}\big[\hat{\chi}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-2}\big)\big] + \hat{\tilde{\chi}}\mathbb{E}\big[\mathrm{Tr}\big(A^\top A\mathcal{M}^{-2}\big)\big] + \hat{m}^2\kappa\mathbb{E}\big[\mathrm{Tr}\big(S^2\mathcal{M}^{-2}\big)\big]\big] \tag{202}$$

$$= \frac{1}{\rho\kappa\hat{q}^2}\big[\hat{\chi}(g_S(z) + zg_S'(z)) + \rho\hat{\tilde{\chi}}g_{\tilde{S}}'(z) + \kappa\hat{m}^2\big(1 + 2zg_S(z) + z^2 g_S'(z)\big)\big] \tag{203}$$

$$m_0 = \frac{1}{r}\mathbb{E}\big[\mathrm{Tr}\big(A^\top W^* A\big)\big] = \frac{\hat{m}}{r}\mathbb{E}\big[\mathrm{Tr}\big(S\mathcal{M}^{-1}\big)\big] = \frac{\hat{m}}{\rho\hat{q}}(1 + zg_S(z)) \tag{204}$$

E.6.2 Conjugate Parameters

The saddle point equations which define the conjugate parameters are followed by G_I (Eq. (157)) and G_S (Eq. (158)) in Eq. (156). In the RS ansatz, there is permutation symmetry among the replica indices. As a result, the saddle point equations can be written for any $a \neq b$ as follows:

$$\hat{\chi} = \gamma\int \mathrm{D}\boldsymbol{z}\mathrm{D}z\, u^a u^b \exp\left[-\frac{\beta}{2}\sum_c (u^c)^2\right] \tag{205}$$

$$\hat{\tilde{\chi}} = \gamma\frac{1}{\alpha}\big(1+\sigma^2\big)\int \mathrm{D}\boldsymbol{z}\mathrm{D}z\, u^a u^b \exp\left[-\frac{\beta}{2}\sum_c (u^c)^2\right] \tag{206}$$

$$\hat{q} = \gamma\int \mathrm{D}\boldsymbol{z}\mathrm{D}z\left(1 + \beta u^a u^b - \beta(u^a)^2\right)\exp\left[-\frac{\beta}{2}\sum_c (u^c)^2\right] \tag{207}$$

$$\hat{\tilde{q}} = \gamma\frac{1}{\alpha}\big(1+\sigma^2\big)\int \mathrm{D}\boldsymbol{z}\mathrm{D}z\left(1 + \beta u^a u^b - \beta(u^a)^2\right)\exp\left[-\frac{\beta}{2}\sum_c (u^c)^2\right] \tag{208}$$

$$\hat{m} = \gamma\int \mathrm{D}\boldsymbol{z}\mathrm{D}z\left(1 - u^a\sum_{b=1}^n u^b\right)\exp\left[-\frac{\beta}{2}\sum_c (u^c)^2\right], \tag{209}$$

where u^a is a Gaussian variable with covariance matrix $\mathcal{M}^{ab}$ (see Eq. (149)). Under the RS ansatz, u^a can be represented using $n+1$ independent standard normal random variables z and $\boldsymbol{z} = \{z^a\}_{a=1,\cdots,n}$ as:

$$u^a = z^a\sqrt{\frac{\frac{1}{\alpha}(1+\sigma^2)\bar{\chi} + \chi}{\beta}} + z\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{q} + q - 2m + 1 + \sigma^2} + h_{\mathrm{source}}, \tag{210}$$

where we introduced h_{source} as a source term used to generate moments of u^a, which is evaluated at $h_{\mathrm{source}} = 0$. Here, by changing variables as $\tilde{z}^a = \sqrt{z^a/\beta}$ and taking the limit $\beta \to \infty$, the integral of $\tilde{z}^a$ with Gaussian weight concentrates on $\tilde{z}^a = (\tilde{z}^a)^*$ as:

$$\int \mathrm{D}\boldsymbol{z}\left(\prod_a f(u^a)\right)\exp\left[-\frac{\beta}{2}\sum_a (u^a)^2\right] = \frac{\int \mathrm{D}\boldsymbol{z}\left(\prod_a f(u^a)\right)\exp\left[-\frac{\beta}{2}\sum_a (u^a)^2\right]}{\int \mathrm{D}\boldsymbol{z}\exp\left[-\frac{\beta}{2}\sum_a (u^a)^2\right]} \quad (n \to 0) \tag{211}$$

$$= \frac{\prod_a \int \mathrm{d}\tilde{z}^a\, f(u^a)\exp\left[-\frac{\beta}{2}\Big((u^a)^2 + (\tilde{z}^a)^2\Big)\right]}{\prod_a \int \mathrm{d}\tilde{z}^a\exp\left[-\frac{\beta}{2}\Big((u^a)^2 + (\tilde{z}^a)^2\Big)\right]} \tag{212}$$

$$= \prod_a f(u^a)\Big|_{\tilde{z}^a = (\tilde{z}^a)^*} \quad (\beta \to \infty), \tag{213}$$

where $f(u^a)$ is a function of u^a, and the value of $(\tilde{z}^a)^*$ is the solution of the following optimization problem:

$$(\tilde{z}^a)^* = \underset{\tilde{z}^a}{\operatorname{argmin}} \left[\frac{1}{2}(\tilde{z}^a)^2 + \frac{1}{2}(u^a)^2 \right] \tag{214}$$

$$= \underset{\tilde{z}^a}{\operatorname{argmin}} \left[\frac{1}{2}(\tilde{z}^a)^2 + \frac{1}{2}\left(\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{\chi} + \chi}\, \tilde{z}^a + \sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{q} + q - 2m + 1 + \sigma^2}\, z + h_{\text{source}} \right)^2 \right] \tag{215}$$

$$= -\frac{\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{\chi} + \chi}\left(\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{q} + q - 2m + 1 + \sigma^2}\, z + h_{\text{source}} \right)}{1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}}. \tag{216}$$

We denote u^a evaluated at $\tilde{z}^a = (\tilde{z}^a)^*$ as $\tilde{u}^*$, i.e.,

$$\tilde{u}^* = (\tilde{z}^a)^* \sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{\chi} + \chi} + z\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{q} + q - 2m + 1 + \sigma^2} + h_{\text{source}} \tag{217}$$

$$= \frac{\sqrt{\frac{1}{\alpha}(1+\sigma^2)\bar{q} + q - 2m + 1 + \sigma^2}\, z + h_{\text{source}}}{1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\hat{\bar{\chi}}}, \tag{218}$$

which does not depend on the replica index. By substituting this property and taking the limit $n \to 0$, the saddle point equations Eqs. (205)–(209) can be rewritten as follows:

$$\hat{\chi} = \gamma \int \mathrm{D}z\, (\tilde{u}^*)^2 \Big|_{h_{\text{source}}=0} = \gamma \frac{\frac{1}{\alpha}\left(1+\sigma^2\right)\bar{q} + q - 2m + 1 + \sigma^2}{\left(1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}\right)^2} \tag{219}$$

$$\hat{\bar{\chi}} = \frac{1}{\alpha}\left(1+\sigma^2\right)\gamma \int \mathrm{D}z\, (\tilde{u}^*)^2 \Big|_{h_{\text{source}}=0} = \gamma \frac{1}{\alpha}\left(1+\sigma^2\right) \frac{\frac{1}{\alpha}\left(1+\sigma^2\right)\bar{q} + q - 2m + 1 + \sigma^2}{\left(1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}\right)^2} \tag{220}$$

$$\hat{q} = \gamma \int \mathrm{D}z \frac{\mathrm{d}\tilde{u}^*}{\mathrm{d}h_{\text{source}}} \Big|_{h_{\text{source}}=0} = \gamma \frac{1}{1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}} \tag{221}$$

$$\hat{\bar{q}} = \frac{1}{\alpha}\left(1+\sigma^2\right)\gamma \int \mathrm{D}z \frac{\mathrm{d}\tilde{u}^*}{\mathrm{d}h_{\text{source}}} \Big|_{h_{\text{source}}=0} = \frac{1}{\alpha}\left(1+\sigma^2\right)\gamma \frac{1}{1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}} \tag{222}$$

$$\hat{m} = \gamma \int \mathrm{D}z \frac{\mathrm{d}\tilde{u}^*}{\mathrm{d}h_{\text{source}}} \Big|_{h_{\text{source}}=0} = \gamma \frac{1}{1 + \chi + \frac{1}{\alpha}(1+\sigma^2)\bar{\chi}}. \tag{223}$$

These results in Section E.6.1 and Section E.6.2 are summarized in Result E.2.

E.7 Generalization Error (derivation of the Result E.3)

In this subsection, we compute the MSE in each protocol from Result. E.1 using the order parameters defined by Result. E.2. For the generalization error, we need to compute the correlation between the prediction using the optimal weight matrix W^* and the ground-truth label y.

E.7.1 Task Memorization

The effective feature matrix of the prompt in the task memorization setting is given by:

$$H = \frac{1}{M_0}\frac{D}{\tilde{L}} \sum_{\mu=1}^{M_0} \left[\boldsymbol{x}_{L+1} \boldsymbol{w}^{\mu\top} \sum_{l=1}^{\tilde{L}} \boldsymbol{x}_l \boldsymbol{x}_l^\top \right], \tag{224}$$

where $\boldsymbol{w}^* = 1/M_0 \sum_{\mu=1}^{M_0} \boldsymbol{w}^\mu$ is the averaged learned task in the pre-training. We have the following relation:

$$\mathcal{E}_{\mathrm{TM}} = \mathbb{E}\big[y - \mathrm{Tr}(W^* H^\top)\big]^2 \tag{225}$$
$$= 1 - 2\sum_{ij} \mathbb{E}\big[y H_{ij} W^*_{ij}\big] + \sum_{ijkl} \mathbb{E}\big[H_{ij} H_{kl} W^*_{ij} W^*_{kl}\big] \tag{226}$$
$$= 1 - 2\frac{1}{D}\frac{1}{M_0}\sum_{\mu} \mathrm{Tr}\Big(\boldsymbol{w}^{\mu\top} W^* \boldsymbol{w}^\mu\Big) + \frac{1}{\tilde{\alpha}}\frac{1}{D}\,\mathrm{Tr}\Big(W^{*\top} W^*\Big) + \frac{1}{D}\frac{1}{M_0}\sum_{\mu} (W^*\boldsymbol{w}^\mu)^\top (W^*\boldsymbol{w}^\mu) \tag{227}$$
$$= 1 - 2m + q + \frac{1}{\tilde{\alpha}}\bar{q}. \tag{228}$$

E.7.2 In-Distribution Generalization

The effective feature matrix of the prompt in the in-distribution setting is given by:

$$H = \frac{D}{\tilde{L}}\left[\boldsymbol{x}_{L+1}(A\boldsymbol{v}^*)^\top \sum_{l=1}^{\tilde{L}} \boldsymbol{x}_l \boldsymbol{x}_l^\top\right], \tag{229}$$

where $\boldsymbol{v}^* \sim \mathcal{N}(0, I_r) \in \mathbb{R}^r$. We have the following relation:

$$\mathcal{E}_{\mathrm{IDG}} = \mathbb{E}\big[y - \mathrm{Tr}(W^* H^\top)\big]^2 \tag{230}$$
$$= 1 - 2\sum_{ij} \mathbb{E}[y H_{ij}]\mathbb{E}\big[W^*_{ij}\big] + \sum_{ijkl} \mathbb{E}\big[H_{ij} H_{kl} W^*_{ij} W^*_{kl}\big] \tag{231}$$
$$= 1 - \frac{2}{r}\sum_{ijkl} \mathbb{E}\big[A_{ik} A_{jl} W^*_{ij}\big]\mathbb{E}[v_k v_l] + \frac{1}{D}\sum_{ijl} \mathbb{E}\left[\left(\frac{D}{\tilde{L}}\delta_{ij} + \sum_{st} A_{js} A_{lt} v_s v_t\right) W^*_{ij} W^*_{il}\right] \tag{232}$$
$$= 1 - 2\frac{1}{r}\mathbb{E}\big[\mathrm{Tr}\big(A^\top W^* A\big)\big] + \frac{1}{r}\mathbb{E}\Big[\mathrm{Tr}\Big((W^*A)^\top W^* A\Big)\Big] + \frac{1}{\tilde{\alpha}}\frac{1}{D}\,\mathrm{Tr}\Big(W^{*\top} W^*\Big) \tag{233}$$
$$= 1 - 2m_0 + q_0 + \frac{1}{\tilde{\alpha}}\bar{q}. \tag{234}$$

E.7.3 Out-of-Distribution Generalization

The effective feature matrix of the prompt in the out-of-distribution setting is given by:

$$H = \frac{D}{\tilde{L}}\left[\boldsymbol{x}_{L+1}\boldsymbol{w}^{*\top} \sum_{l=1}^{\tilde{L}} \boldsymbol{x}_l \boldsymbol{x}_l^\top\right], \tag{235}$$

where $\boldsymbol{w}^* \sim \mathcal{N}(0, I_D) \in \mathbb{R}^D$. We have the following relation:

$$\mathcal{E}_{\mathrm{ODG}} = \mathbb{E}\big[y - \mathrm{Tr}(W^* H^\top)\big]^2 \tag{236}$$
$$= 1 - 2\sum_{ij} \mathbb{E}[y H_{ij}]\mathbb{E}\big[W^*_{ij}\big] + \sum_{ijkl} \mathbb{E}[H_{ij} H_{kl}]\big[W^*_{ij} W^*_{kl}\big] \tag{237}$$
$$= 1 - \frac{2}{D}\sum_{ij} \mathbb{E}[w_i w_j]\mathbb{E}\big[W^*_{ij}\big] + \frac{1}{D}\sum_{ijl} \mathbb{E}\left[\frac{D}{\tilde{L}}\delta_{ij} + w_j w_l\right]\big[W^*_{ij} W^*_{il}\big] \tag{238}$$
$$= 1 - 2\frac{1}{D}\,\mathrm{Tr}\, W^* + \left(1 + \frac{1}{\tilde{\alpha}}\right)\frac{1}{D}\,\mathrm{Tr}\Big(W^{*\top} W^*\Big) \tag{239}$$
$$= 1 - 2\bar{m} + \left(1 + \frac{1}{\tilde{\alpha}}\right)\bar{q}. \tag{240}$$

These results are summarized in Result E.3.

E.8 Asymptotics under $\alpha \gg 1$ (derivation of Proposition E.4)

The proof proceeds by performing an asymptotic expansion of the saddle-point equations, presented in Result E.2, in the limit where $\alpha \to \infty$.

The key insight is that in this limit, the system of equations simplifies considerably. First, we observe that the conjugate parameters $\hat{\bar{q}}$ and $\hat{\bar{\chi}}$ vanish as $\mathcal{O}(1/\alpha)$ or faster. This causes the parameter $z = -(\lambda + \hat{\bar{q}})/\hat{q}$ to converge to a finite constant that is dependent of α.

By substituting these simplified forms back into the equations for the primary order parameters ($q, m, \bar{q}$, etc.) and retaining only the leading-order terms in α, the coupled system of equations becomes analytically solvable. This systematic expansion directly yields the explicit expressions for each parameter as stated in the proposition.

F Proof of Theorem 6.1

In this section, we proof Theorem 6.1.

From Lemma D.1, the resolvent of S is given by:

$$g_S(z) \;=\; -(1-\rho)\frac{1}{z} + \rho\frac{-(\kappa\rho z + \rho - \kappa) - \sqrt{(\kappa\rho z + \rho - \kappa)^2 - 4\rho^2\kappa z}}{2\rho z}. \tag{241}$$

By definition, the limiting spectral density $\nu_S(\lambda)$ is obtained from the imaginary part of the resolvent:

$$\nu_S(\lambda) = -\frac{1}{\pi}\lim_{\epsilon\to 0^+} \operatorname{Im} g_S(\lambda + i\epsilon). \tag{242}$$

For $\lambda \in [\lambda_-, \lambda_+]$, the square root term in $g_{\tilde{S}}(z)$ contributes an imaginary part. A direct computation yields:

$$\lim_{\epsilon\to 0^+} \operatorname{Im} g_{\tilde{S}}(\lambda + i\epsilon) \;=\; \frac{1}{2\rho\lambda}\sqrt{4\rho^2\kappa\lambda - (\kappa\rho\lambda + \rho - \kappa)^2}. \tag{243}$$

Substituting this into $g_S(z)$ gives:

$$\nu_S(\lambda) \;=\; \frac{\rho\kappa}{2\pi\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \quad \lambda \in [\lambda_-, \lambda_+], \tag{244}$$

with the edges

$$\lambda_\pm = \frac{1}{\rho\kappa}\left(\sqrt{\rho} \pm \sqrt{\kappa}\right)^2.$$

Finally, since S has rank at most $\min(r, M_0) = D\min(\rho, \kappa)$, there is a mass $1 - \min(\rho, \kappa)$ at zero. Hence, we have:

$$\nu_S(\lambda) \;=\; (1 - \min(\rho,\kappa))\,\delta(\lambda) \;+\; \mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda)\,\frac{\rho\kappa}{2\pi\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}.$$

G Consistency of the Replica Method with Numerical Experiments

In this section, we validate our analytical results from the replica method against numerical experiments. As depicted in Figure 5 and Figure 6, the theoretical predictions for both the order parameters and the generalization error show excellent agreement with the simulation results, confirming the accuracy of our results.

H Experimental Details for the Softmax-Attention Experiments

This section reports the reproducibility details of numerical experiments using a one-layer softmax self-attention model in Section 7. The task generation, prompt construction (including masking the query label), and the three evaluation protocols (TM/IDG/ODG) follow Section 3. We only describe the parts that differ from the linear-attention analysis.

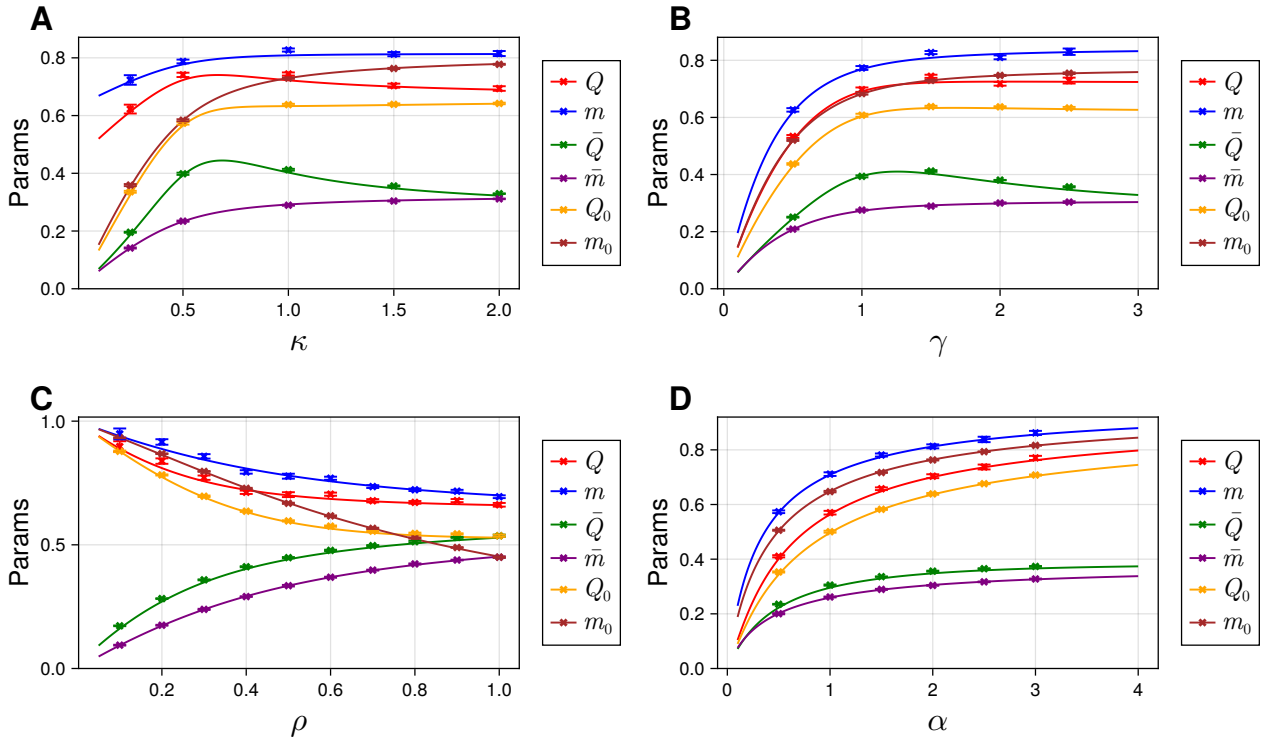

Figure 5: Comparison of the order parameters $Q, m, \bar{Q}, \bar{m}, Q_0, m_0$ obtained from the replica method (lines) with those obtained from the numerical experiments (error bars). Parameters for (A-D): $\lambda = 0.1, \sigma = 0.1, D = 70$; (A)$\gamma = 1.5, \rho = 0.4, \alpha = 2.0$; (B)$\kappa = 1.0, \rho = 0.4, \alpha = 2.0$; (C)$\kappa = 1.0, \gamma = 1.5, \alpha = 2.0$; (D)$\kappa = 1.0, \gamma = 1.5, \rho = 0.4$. Error bars represent the standard error of the mean over 10 trials per point.

H.1 Architecture

We replace the linear attention map in Eq. (10) with a standard scaled dot-product softmax attention layer with learned linear projections. Each token has dimension $D+1$, obtained by concatenating the input vector $\boldsymbol{x} \in \mathbb{R}^D$ and the scalar label channel. For a context matrix $C \in \mathbb{R}^{(D+1)\times(L+1)}$, we define $Q = W_Q C, \quad K = W_K C, \quad V = W_V C$, where $W_Q \in \mathbb{R}^{d_{\text{attn}}\times(D+1)}, W_K \in \mathbb{R}^{d_{\text{attn}}\times(D+1)}$, and $W_V \in \mathbb{R}^{(D+1)\times(D+1)}$. We use no biases in all projections. Unlike standard transformer blocks, we do not include an output projection W_O; instead, we set the value dimension to $D+1$ so that the value projection already returns the token dimension.

We use only the query vector of the last token. Writing

$$q_{\text{last}} \in \mathbb{R}^{d_{\text{attn}}} \tag{245}$$

for the last-column query and $K \in \mathbb{R}^{d_{\text{attn}}\times(L+1)}$ for all keys, the attention weights are

$$\boldsymbol{a} = \text{softmax}\left(\frac{q_{\text{last}}^\top K}{\sqrt{d_{\text{attn}}}}\right) \in \mathbb{R}^{L+1}. \tag{246}$$

The scalar prediction is read out only from the label channel (the $(D+1)$-st coordinate) of the values: $\hat{y} = \sum_{l=1}^{L+1} a_l V_{D+1,l}$. This matches the evaluation convention in Section 3, where the query label is masked in the input and the model predicts the missing scalar. Unless stated otherwise, we fix the attention width to $d_{\text{attn}} = 4$ in all reported experiments.

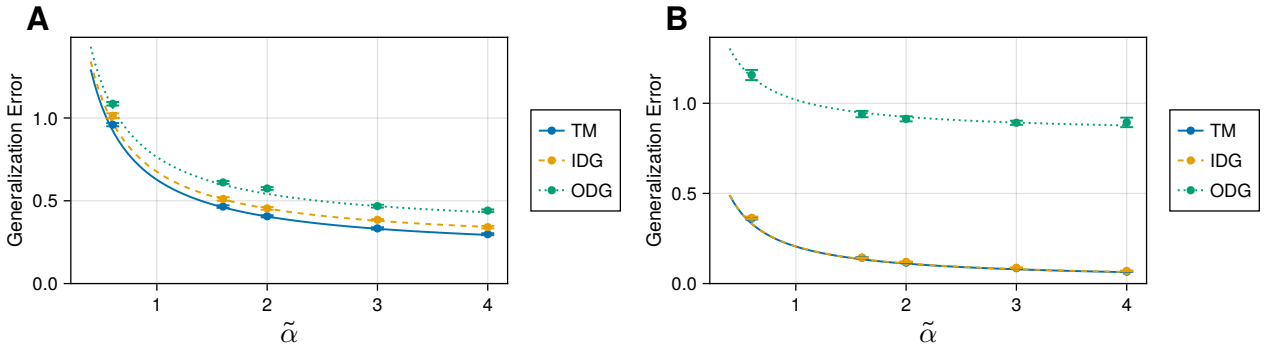

Figure 6: Comparison of the generalization errors $(\mathcal{E}_{\rm TM}, \mathcal{E}_{\rm IDG}, \mathcal{E}_{\rm ODG})$ obtained from the replica method (lines) with those obtained from the numerical experiments (error bars). Parameters for (A-D): $\kappa = 4.0, \gamma = 0.5, \alpha = 4.0, \lambda = 0.1, \sigma = 0.1, D = 60$; (A)$\rho = 0.9$; (B)$\rho = 0.2$. Error bars represent the standard errors of the mean over 5 trials per point.

H.2 Training objective and optimization

We train the model parameters $\Theta = \{W_Q, W_K, W_V\}$ by minimizing the empirical mean-squared error on the query label:

$$\mathcal{L}(\Theta) = \frac{1}{M} \sum_{\mu=1}^{M} \left(\hat{y}^\mu - y^\mu_{L+1}\right)^2, \tag{247}$$

using Adam with learning rate $\eta = 10^{-3}$. We use full-batch training: at each epoch, the loss $\mathcal{L}(\Theta)$ is evaluated over all M training instances and a single Adam update is performed. All projection matrices are initialized using Glorot (Xavier) uniform initialization.