Optimal Attention Temperature Improves the Robustness of In-Context Learning under Distribution Shift in High Dimensions

Samet Demir [1] **Zafer Doğan** [1 2]

Abstract

Pretrained Transformers can perform in-context learning (ICL) from a few demonstrations, but this ability can fail sharply when the test distribution differs from pretraining—a common deployment setting. We study attention temperature as a simple inference-time control for improving ICL robustness under such shifts. In a high-dimensional linear-regression framework, we analyze a Transformer with "approximate softmax" attention, which preserves softmax's normalization and temperature-dependent selectivity while remaining tractable. We derive a closed-form expression for the ICL generalization error under distribution shift, and show that it is minimized by an explicit optimal attention temperature. This characterization yields interpretable guidance by linking the best temperature to moments of the pre-softmax attention scores, and predicts when temperature adjustment can recover near Bayes-optimal performance. We validate the theory with extensive simulations, and further demonstrate gains for pretrained LLMs on question-answering benchmarks under distribution shift induced by noisy in-context demonstrations. Overall, attention temperature emerges as a principled, lightweight knob for improving the robustness of ICL in pretrained Transformers.

1. Introduction

Transformers (Vaswani et al., 2017) have emerged as the foundational architecture of contemporary AI systems, underpinning state-of-the-art models such as ChatGPT, Gemini, and DeepSeek. Central to their remarkable success is *in-context learning* (ICL)—the capability to adapt to novel tasks directly from prompts without any gradient-based weight updates (Brown et al., 2020). This property, often described as emergent, has catalyzed a surge of research aimed at uncovering the underlying mechanisms of ICL (Akyürek et al., 2023; Von Oswald et al., 2023) and at characterizing how factors such as task diversity and model scale shape its performance (Wei et al., 2022; Wu et al., 2024).

Yet, despite its transformative potential, ICL exhibits pronounced sensitivity to distributional shifts between pretraining and downstream tasks. Both empirical and theoretical investigations demonstrate that even mild shifts can substantially degrade performance (Zhang et al., 2024), underscoring unresolved questions about the robustness, generalization, and adaptability of pretrained Transformer models. Addressing these limitations is crucial for realizing the full promise of ICL in reliable, deployable AI systems.

At the core of the Transformer architecture is the self-attention mechanism, defined as

$$\text{Attention}(\boldsymbol{Z}) := \boldsymbol{V}\boldsymbol{Z} \cdot \text{softmax}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \quad (1)$$

where $\boldsymbol{Z}$ denotes the input representation and $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value weight matrices, respectively. The parameter $\tau > 0$, referred to as the *attention temperature*, controls the variance of the softmax outputs and hence the selectivity of attention weights. This quantity is distinct from the "sampling temperature" commonly used to adjust the output distribution of generative models such as large language models (LLMs) (Renze & Guven, 2024). Throughout this work, we exclusively focus on the attention temperature as an intrinsic component of the attention mechanism.

While the original Transformer fixes $\tau = \sqrt{d_k}$ (Vaswani et al., 2017), where d_k is the key dimension, subsequent empirical studies have demonstrated that adjusting the attention temperature can enhance performance across diverse NLP and computer vision benchmarks (Lin et al., 2018; Zhang et al., 2022; Peng et al., 2024; Lee et al., 2021; Chen et al., 2023; Zou et al., 2024). Yet, to the best of our knowledge, its role within *in-context learning* (ICL) remains unexplored. Because the attention temperature directly governs

[1]MLIP Research Group, KUIS AI Center, Koç University [2]Department of EEE, Koç University, İstanbul, Turkey. Correspondence to: Zafer Doğan <zdogan@ku.edu.tr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

how sharply the model concentrates on specific inputs, it is poised to critically influence ICL behavior under distributional shift—a setting of central practical relevance, where mismatches between training and inference distributions are ubiquitous.

This work — In this paper, we present a unified theoretical and empirical study of the *attention temperature* in the context of in-context learning (ICL). We focus on how adjusting this parameter can systematically improve the ICL performance of pretrained Transformers under distributional shift. We address this question in the setting of linear regression tasks, which offer a well-controlled yet expressive framework for dissecting the mechanisms of ICL (Garg et al., 2022; Zhang et al., 2024). Departing from prior work restricted to linear attention, we analyze a Transformer with *approximate softmax* attention—an architecture that preserves the essential temperature-dependent behavior of standard attention while remaining mathematically tractable.

Our analysis yields a closed-form characterization of the *optimal temperature*—the value of τ that minimizes generalization error during inference. We show that this optimal temperature depends explicitly on the nature of the distribution shift and that setting it appropriately can recover or even surpass baseline ICL performance. We validate our theoretical predictions through extensive experiments on both synthetic (linear regression) and real-world (question answering with LLMs) tasks, demonstrating that temperature selection constitutes a simple yet powerful mechanism for improving robustness.

Contributions — Our work makes the following key contributions:

1. We provide, to our knowledge, the first theoretical characterization of the optimal attention temperature for pretrained Transformers with *approximate softmax attention* in ICL tasks.
2. We analyze the generalization behavior of such models under a broad spectrum of distributional shifts, employing weaker assumptions than prior studies.
3. We establish a clear theoretical and empirical link between distribution shift and attention temperature, showing that principled temperature selection can substantially enhance ICL performance across diverse tasks.

Taken together, these results offer new insights into the interplay between temperature, distribution shift, and generalization in in-context learning, and highlight a practical avenue for improving the robustness of pretrained Transformers.

Notation — We follow standard notation from Goodfellow et al. (2016). The spectral norm of matrix $\boldsymbol{M}$ is denoted by $\|\boldsymbol{M}\|$, and the trace by $\mathrm{Tr}(\boldsymbol{M})$. Matrix entries and slices are denoted as $M_{i,j}$, $\boldsymbol{M}_{:,j}$, and $\boldsymbol{M}_{i,:}$.

2. Related work

ICL by Transformers — The ICL capability of Transformers was first brought to prominence by (Brown et al., 2020), leading to a surge of empirical and theoretical investigations. Several works have demonstrated that ICL performance improves with model scale (Wei et al., 2022; Olsson et al., 2022; Schaeffer et al., 2023), underscoring its importance in modern AI systems. To better understand this phenomenon, synthetic tasks such as linear regression have served as controlled testbeds for analyzing ICL in Transformers (Garg et al., 2022; Zhang et al., 2024; Raventós et al., 2023). A prevailing hypothesis in recent theoretical work is that Transformers implicitly learn algorithms during pretraining, which they subsequently execute during inference (Bai et al., 2023; Li et al., 2023; Akyürek et al., 2023; Ahn et al., 2023; Von Oswald et al., 2023; Mahankali et al., 2024; Fu et al., 2024; Zhang et al., 2024; Li et al., 2024; Park et al., 2024). There remains ongoing debate over the precise nature of these learned procedures. However, our work focuses on a fundamentally different question, which is how attention temperature affects the ICL performance of pretrained Transformers under distribution shifts.

Theory of in-context learning — Simplified Transformer variants—particularly those using linear attention—have proven useful for gaining analytical insights about ICL (Garg et al., 2022; Zhang et al., 2024; Raventós et al., 2023). Notably, Zhang et al. (2024) showed that linear Transformers approximate Bayes-optimal inference in linear regression tasks, even under distribution shift. We build on this line of research but focus explicitly on the role of the *attention temperature*. In contrast to Zhang et al. (2024), we (i) employ *approximate softmax attention* to isolate the effect of temperature, (ii) study how temperature adjustments can mitigate the impact of distribution shifts, and (iii) derive and empirically evaluate the *optimal temperature* for improving ICL performance. These advances extend prior analyses and yield a deeper theoretical and empirical understanding of how principled temperature selection enhances the robustness of Transformers under distributional shift.

Linear vs. softmax attention — Although linear attention has gained traction for its computational efficiency, it typically lags behind softmax-based counterparts in predictive performance, spurring efforts to narrow this gap (Choromanski et al., 2021; Qin et al., 2022). A pivotal advance in this direction is due to Han et al. (2024), who showed that a *approximate variant of softmax attention* can closely approximate the performance of standard softmax attention. Building on this insight, we adopt the *approximate softmax* formulation, which preserves the essential temperature-dependent behavior of standard attention while enabling tractable theoretical analysis. This choice provides a princi-

pled framework for investigating how attention-temperature selection shapes ICL performance in pretrained Transformers.

Attention temperature — Research on attention temperature remains limited. Veličković et al. (2025) recently proposed an adaptive temperature scheme to sharpen softmax outputs, and several empirical studies in natural language processing and computer vision (Lin et al., 2018; Zhang et al., 2022; Peng et al., 2024; Lee et al., 2021; Chen et al., 2023; Zou et al., 2024) suggest that adjusting the attention temperature can enhance Transformer performance. However, these works do not examine ICL under distributional shift. To our knowledge, no prior study has systematically analyzed how attention temperature influences ICL in such settings—a gap our work directly addresses.

3. Setting

We describe the setup for analyzing ICL in linear regression using pretrained Transformers, covering the data model, approximate softmax attention with reparameterization, evaluation metrics, and the Bayes-optimal benchmark.

3.1. In-context learning for linear regression

We study the ICL abilities of pretrained Transformers on linear regression tasks. Given a sequence of tokens, i.e., input-label pairs, $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{l-1}, y_{l-1}), (\boldsymbol{x}_l, ?)\}$ where each input vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and corresponding label $y_i \in \mathbb{R}$ are independently sampled from an unknown joint distribution, the model must predict y_l using only the context $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l-1}$ and the query $\boldsymbol{x}_l$, where $l-1$ is referred as the "context length". Each $(\boldsymbol{x}_i, y_i)$ pair is sampled i.i.d. from a joint distribution defined by:

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x),\ y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (2)$$

where the task vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ is fixed within a context but varies across tasks.

Assumption 3.1 (Well-behaved data distributions)**.** There exist constants $c_1, c_2, c_3 > 0$ such that:

$$\|\boldsymbol{\mu}_x\|, \|\boldsymbol{\mu}_w\| \le c_1, \quad \lambda_{\min}(\boldsymbol{\Sigma}_x), \lambda_{\min}(\boldsymbol{\Sigma}_w) \ge c_2, \quad (3)$$

$$\lambda_{\max}(\boldsymbol{\Sigma}_x), \lambda_{\max}(\boldsymbol{\Sigma}_w) \le c_3. \quad (4)$$

This assumption ensures that the input and task distributions have bounded means and covariances, offering greater flexibility than the more restrictive setup of Zhang et al. (2024).

Assumption 3.2 (High-dimensional regime)**.** The context length l and input dimension d diverge jointly: $l, d \to \infty$.

This assumption reflects realistic settings where both context length and input dimension grow simultaneously, aligning with modern ML trends and enabling analysis of generalization in high-dimensional regimes.

Under this set of assumptions, we define ICL for linear regression tasks as follows:

Definition 3.3 (In-Context Learning (ICL))**.** A model succeeds at ICL for linear regression if its generalization error nearly matches that of the Bayes-optimal linear model (defined in Section 3.6).

3.2. Modeling attention with transformers

Following the convention established by Zhang et al. (2024), we represent the input sequence by an embedding matrix:

$$\boldsymbol{Z} := \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{l-1} & \boldsymbol{x}_l \\ y_1 & \cdots & y_{l-1} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times l}, \quad (5)$$

where the last column corresponds to the query input with no label.

Given this embedding, the softmax self-attention output is

$$\boldsymbol{S} := \boldsymbol{Z} + \boldsymbol{V}\boldsymbol{Z} \cdot \text{softmax}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \quad (6)$$

where $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ are the key, query, and value matrices, respectively, and τ is the temperature.

Here, we denote the model's prediction as $S_{d+1,l}$ — the last element in the final row.

3.3. Approximate softmax attention

To analytically characterize the effect of temperature on ICL, we employ an approximation of softmax attention :

$$\boldsymbol{E} := \boldsymbol{Z} + \boldsymbol{V}\boldsymbol{Z} \cdot \widehat{\text{softmax}}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \quad (7)$$

the explicit form of which is provided together with its derivation in Appendix B. In contrast to linear attention (Zhang et al., 2024),

$$\boldsymbol{Z} + \frac{1}{l}\boldsymbol{V}\boldsymbol{Z}(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z}), \quad (8)$$

our formulation in (7) explicitly preserves normalization, which is essential for both interpretability and robustness. This difference is described in the following remark.

Remark 3.4 (Linear vs. approximate softmax attention). Approximate softmax attention maintains row-wise normalization, making it inherently more robust to shifts in input means — a critical failure mode of linear attention in ICL.

Another key distinction between the linear case and the approximate softmax case is that linear (with temperature scaling) fails to capture the temperature behavior of softmax.

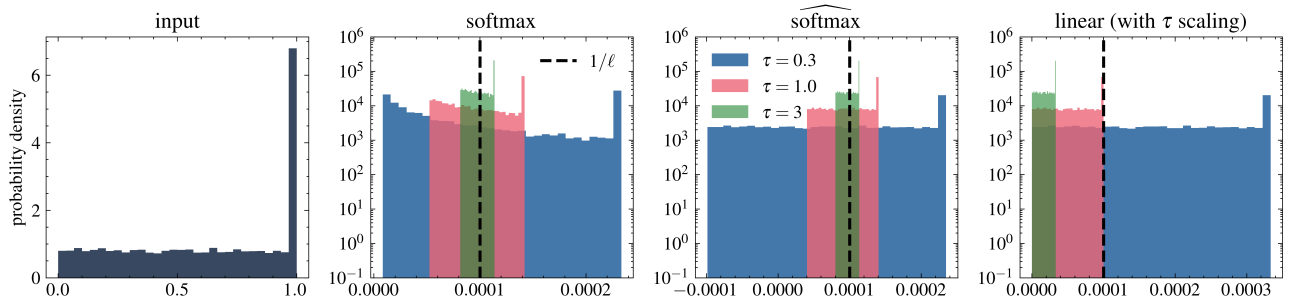


Figure 1. Comparison of temperature effects in softmax, approximate softmax, and linear (with temperature scaling) cases. We consider an input vector $\boldsymbol{x} \in \mathbb{R}^l$ whose histogram is illustrated on the left-most plot. Rest of the plots illustrates histograms of the elements of $\text{softmax}(\boldsymbol{x}/\tau)$, approximation $\widehat{\text{softmax}}(\boldsymbol{x}/\tau)$ derived in Appendix B and $\boldsymbol{x}/(l\tau)$ from left to right, respectively.

However, while this may not be immediately apparent, approximate softmax closely mirrors the behavior of softmax with respect to temperature variation. Figure 1 visualizes the effects of temperature across softmax, approximate softmax, and linear (with temperature scaling) settings under an example input. The figure demonstrates our claim that approximate softmax can be a useful tool to study the temperature effect exhibited by softmax. We expand on this point in Appendix C.

3.4. Reparametrization of approximate softmax attention

To streamline analysis, we reparametrize the matrices $\boldsymbol{V}$ and $\boldsymbol{M} := \boldsymbol{K}^T\boldsymbol{Q}$ as:

$$\boldsymbol{V} = \begin{bmatrix} * & * \\ \boldsymbol{v}_{21}^T & v_{22} \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & * \\ \boldsymbol{m}_{21}^T & * \end{bmatrix}, \tag{9}$$

where only $\boldsymbol{v}_{21}, v_{22}, \boldsymbol{m}_{21}$, and $\boldsymbol{M}_{11}$ influence the prediction $\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M})$. The remaining terms are denoted by $*$ as they are not relevant for predicting y_l in this context. The prediction from the Transformer model (7) can thus be expressed as a function of $\boldsymbol{M}$ and $\boldsymbol{V}$, i.e., $\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) := E_{d+1,l}$. This form parallels the approach by Zhang et al. (2024), allowing for tractable theoretical analysis.

By analyzing this reparameterization, we gain a deeper understanding of how the model parameters interact with the data to address the ICL problem effectively.

3.5. Evaluating generalization performance

We focus on evaluating the performance of our attention model by assessing its generalization error, measuring the ICL performance. For a given set of parameters $(\boldsymbol{V}, \boldsymbol{M})$, the model's generalization (ICL) error is defined as:

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) := \mathop{\mathbb{E}}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(y_l - \hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) \right)^2 \right], \tag{10}$$

where $\mathcal{D}^{test}$ denotes the distribution of the test set, which includes input-output pairs generated with tasks that the model has not encountered during training. Since the task vectors in the test set differ from those encountered during training, the model is required to infer these new vectors based solely on the provided context. Therefore, the ICL/generalization error (10) assesses the genuine ICL capabilities of the model.

3.6. Bayes-optimal ridge estimator

The Bayes-optimal ridge estimator provides a principled framework for estimating the task vector $\boldsymbol{w}$ given a prior distribution and a set of $l-1$ samples. It is defined as:

$$\hat{\boldsymbol{w}}_{Bayes} = \left(\frac{\bar{\boldsymbol{X}}^T \bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1} \right)^{-1} \left(\frac{\bar{\boldsymbol{X}}^T \bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_w \right), \tag{11}$$

where $\bar{\boldsymbol{X}}$ is the centered input matrix and $\bar{\boldsymbol{y}}$ is the centered label vector. This estimator combines information from observed data with prior knowledge of the distribution of $\boldsymbol{w}$, thereby balancing bias and variance. It serves as the gold standard against which we benchmark model predictions. The terms involving $\boldsymbol{\Sigma}_w^{-1}$ introduce a regularization effect, which is particularly advantageous in high-dimensional regimes.

The derivation, provided in Appendix A, illustrates how Bayesian principles inform regression by integrating data evidence with prior distributions to yield more reliable predictions. In our setting, the inputs and labels are derived from the prompt matrix $\boldsymbol{Z}$, and the Bayes-optimal linear model predicts any input $\boldsymbol{x}$ as $\hat{\boldsymbol{w}}_{Bayes}^T \boldsymbol{x}$.

4. Theoretical results

In this section, we present our main theoretical results on the ICL under distribution shifts for the Transformer with an approximate softmax attention without MLP layers, denoted by (7). We provide an asymptotic characterization of its generalization error, offering a principled framework for analyzing performance. Next, we investigate the role

of the temperature parameter and demonstrate that adjusting it appropriately can substantially improve generalization—especially in cases where the model initially fails to perform effective in-context learning. We then identify specific conditions for which the model fails to generalize under distribution shifts at test time, revealing key limitations of the model in ICL. We describe the interpretation of the found analytical form of the optimal attention temperature in mitigating the adverse effects of distribution shift. Finally, we explain a possible extension of our optimal temperature to other ICL settings involving softmax attention.

4.1. In-context learning performance

We analyze the in-context learning (ICL) performance of the Transformer model (7) by evaluating the generalization error defined in (10). To establish a general setting for the subsequent results, we impose the following assumption on the pretrained parameters.

Assumption 4.1. There exists a constant $c > 0$ such that

$$\|\boldsymbol{M}_{11}\| \leq cd, \quad \|\boldsymbol{m}_{21}\| = 0, \quad \|\boldsymbol{v}_{21}\| \leq \frac{c}{dl}, \quad |v_{22}| \leq \frac{c}{d}.$$

The generalization error result stated below holds for any parameters $\boldsymbol{M}, \boldsymbol{V}$ that satisfy Assumption 4.1.

Theorem 4.2 (Generalization error for ICL)**.** *Suppose Assumptions 3.1–3.2 and 4.1 hold. At test time, assume the input, task, and noise distributions are given by $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, and $\mathcal{N}(0, \sigma^2)$, respectively. Define*

$$\boldsymbol{A} := \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x^T, \quad \boldsymbol{B} := \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T.$$

Then, the generalization error is

$$\begin{aligned}\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = &\frac{1}{\tau^2} \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T \boldsymbol{F}_1 \boldsymbol{M}_{11}\right) \\ &- \frac{1}{\tau} \operatorname{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2 \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T \boldsymbol{F}_2^T\right)\right) \\ &+ \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) + \sigma^2, \end{aligned} \tag{12}$$

where

$$\boldsymbol{F}_1 := \left(\boldsymbol{\Sigma}_x \hat{\boldsymbol{B}} + \frac{1}{l}\left(v_{22}^2 \sigma^2 + \operatorname{Tr}(\hat{\boldsymbol{B}} \boldsymbol{\Sigma}_x)\right) \boldsymbol{I}\right) \boldsymbol{\Sigma}_x, \tag{13}$$

$$\boldsymbol{F}_2 := (\boldsymbol{\mu}_w \boldsymbol{v}_{21}^T + v_{22} \boldsymbol{B}) \boldsymbol{\Sigma}_x, \tag{14}$$

$$\hat{\boldsymbol{B}} := v_{22} \boldsymbol{\mu}_w \boldsymbol{v}_{21}^T + v_{22} \boldsymbol{v}_{21} \boldsymbol{\mu}_w^T + v_{22}^2 \boldsymbol{B}. \tag{15}$$

Proof. The generalization error is derived using Isserlis' theorem (Isserlis, 1918) to compute higher-order moments. See Appendix F for the full derivation. □

Theorem 4.2 illustrates how the parameters $\boldsymbol{M}$, $\boldsymbol{V}$, and the test-time data distribution affect the generalization error. Notably, the temperature parameter τ plays a critical role. Although temperature can be implicitly encoded in $\boldsymbol{M}$ during pretraining, it becomes especially important under distribution shifts that the model is not equipped to handle. In such cases, one can optimize generalization performance by choosing the temperature τ_{opt} that minimizes the generalization error, as discussed next.

4.2. Optimal attention temperature

To address distribution shifts, we define the optimal attention temperature in the following theorem.

Theorem 4.3 (Optimal attention temperature)**.** *Suppose Assumptions 3.1, 3.2, and 4.1. To minimize the generalization error, the optimal attention temperature for inference is given by*

$$\tau_{opt} = \frac{2 Tr\left(\boldsymbol{A}\boldsymbol{M}_{11}^T \boldsymbol{F}_1 \boldsymbol{M}_{11}\right)}{Tr\left(\boldsymbol{A}\left(\boldsymbol{F}_2 \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T \boldsymbol{F}_2^T\right)\right)}, \tag{16}$$

provided that $Tr\left(\boldsymbol{A}\left(\boldsymbol{F}_2 \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T \boldsymbol{F}_2^T\right)\right) > 0$ and $Tr\left(\boldsymbol{A}\boldsymbol{M}_{11}^T \boldsymbol{F}_1 \boldsymbol{M}_{11}\right) > 0$.

Proof. We minimize the generalization error from Theorem 4.2 with respect to τ (Appendix H). □

Consider the optimal temperature τ_{opt} from Theorem 4.3. When $\tau_{\text{opt}} \neq 1$, using an unadjusted temperature leads to suboptimal generalization error. Thus, incorporating the optimal temperature improves generalization in in-context learning under distribution shift.

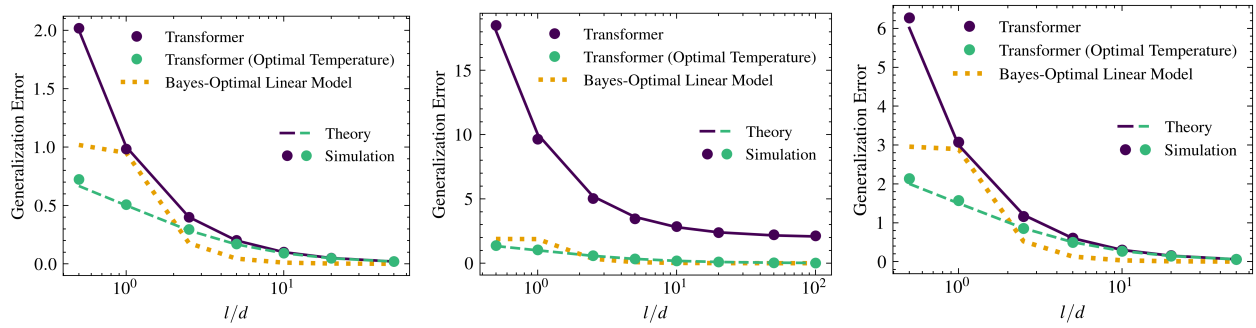
4.3. Effect of distribution shift

In this section, we explore scenarios where $\mathcal{D}^{test} \neq \mathcal{D}^{train}$, indicating a shift in the input, task, or noise distribution after pretraining the model. We consider three cases: (1) a shift in the input distribution (altered mean or covariance), (2) a shift in the task distribution, and (3) a change in the noise levels.

Pretraining for explaining distribution shift effects — To streamline our explanations below, for pretraining, we optimize the parameters $\boldsymbol{V}$ and $\boldsymbol{M}$ using m samples of $(\boldsymbol{Z}, y_l)$ drawn from the distribution $\mathcal{D}^{train}$, where each $\boldsymbol{Z}$ contains $l-1$ $(\boldsymbol{x}, y)$ pairs intended for ICL.

We start by observing that the prediction of the approximate softmax attention model in (7) can be written in the form

$$\begin{aligned}\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) = &\tfrac{1}{\tau} \hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V})^T \boldsymbol{x}_l \\ &+ b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}), \end{aligned} \tag{17}$$



(a) Without Distribution Shift ($\mathcal{D}^{train} = \mathcal{D}^{test}$)

(b) With Shift in Input Covariance ($\mathbf{\Sigma}_{test} = 2\mathbf{\Sigma}_{train}$)

(c) With Shift in Task Distribution ($\boldsymbol{\mu}_w^{test} = \mathbf{1}/\sqrt{d}$, $\mathbf{\Sigma}_w^{test} = 3\mathbf{\Sigma}_w^{train}$)

Figure 2. Experiments with Transformer (7) on ICL under distribution shifts. Parameters are set using (20) while the optimal temperature is calculated by Theorem 4.3. Here, $d = 50$, $m = 5000$ (with a new task per sample), $\sigma = 0.1$, $\boldsymbol{\mu}_x^{train} = \boldsymbol{\mu}_w^{train} = \mathbf{0}$, and $\mathbf{\Sigma}_x^{train} = \mathbf{\Sigma}_w^{train} = \boldsymbol{I}$.

where the *context-summary statistics*

$$\begin{aligned} \boldsymbol{s}_x &:= \tfrac{1}{l}\textstyle\sum_{i\le l} \boldsymbol{x}_i, & s_y &:= \tfrac{1}{l}\textstyle\sum_{i\le l-1} y_i, \\ \boldsymbol{C}_{xx} &:= \tfrac{1}{l}\textstyle\sum_{i\le l} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{s}_x\boldsymbol{s}_x^T, & C_{yy} &:= \tfrac{1}{l}\textstyle\sum_{i\le l-1} y_i^2 - s_y^2 \\ \boldsymbol{C}_{xy} &:= \tfrac{1}{l}\textstyle\sum_{i\le l-1} y_i\boldsymbol{x}_i - s_y\boldsymbol{s}_x, \end{aligned} \tag{18}$$

summarise the in-context examples, and the *effective predictor* and *effective bias* read

$$\begin{aligned} \hat{\boldsymbol{w}}_{Att} &= \boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right) \\ &\quad + \left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}, \\ b_{Att} &= \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}\, s_y. \end{aligned} \tag{19}$$

Here, $\hat{\boldsymbol{w}}_{Att}$ plays the role of an in-context regression coefficient and b_{Att} the role of a bias; both are induced by the attention parameters. The full derivation is given in Appendix D.

Building upon prior work that connects ICL in linear regression to the Bayes-optimal ridge estimator (Zhang et al., 2024; Raventós et al., 2023), we configure $\boldsymbol{M}$ and $\boldsymbol{V}$ to emulate Bayes-optimal ridge regression. Specifically, we aim for $\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V}) \approx \hat{\boldsymbol{w}}_{Bayes}$ and $b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) \approx 0$.

Proposition 4.4 (Pretrained parameters)**.** *Suppose Assumption 3.1 and 3.2. When the temperature parameter is set to $\tau = 1$ during pretraining, the following parameter configuration approximates the Bayes-optimal estimator in (11):*

$$\begin{aligned} \boldsymbol{M}_{11} &= d\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml} + \frac{\sigma^2}{l}\mathbf{\Sigma}_w^{-1}\right)^{-1}, & \boldsymbol{m}_{21} &= \mathbf{0}, \\ \boldsymbol{v}_{21} &= \frac{\sigma^2}{dl}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\mathbf{\Sigma}_w^{-1}\boldsymbol{\mu}_w, & v_{22} &= \frac{1}{d}, \end{aligned} \tag{20}$$

where $\hat{\boldsymbol{X}} \in \mathbb{R}^{ml\times d}$ is the centered input matrix formed from ml samples of $\boldsymbol{x}$.

This configuration aligns our model with Bayes-optimal ridge regression. The quantities $\boldsymbol{\mu}_w$ and $\mathbf{\Sigma}_w$ can be estimated from the pretraining data. A detailed derivation is provided in Appendix E.

Remark 4.5. Aligning $\hat{\boldsymbol{w}}_{Att} \approx \hat{\boldsymbol{w}}_{Bayes}$ via (11) would, in principle, require the *per-task* sample covariance $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$, which varies from one in-context prompt to the next. Since the pretrained parameters $\boldsymbol{M}, \boldsymbol{V}$ must be fixed after training, we replace this per-task quantity by its expectation over the pretraining input distribution, estimated by $\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}/(ml)$ from all ml pretraining samples. By the law of large numbers, $\frac{1}{ml}\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} \to \mathbf{\Sigma}_x$ as $ml \to \infty$, so this substitution is consistent. Together with the high-dimensional regime (Assumption 3.2), this allows the pretrained model to emulate the Bayes-optimal predictor.

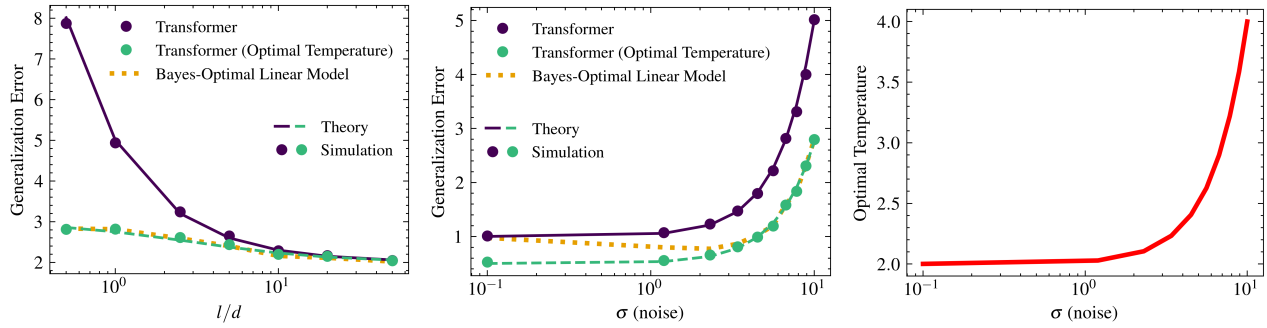
Remark 4.6. While the pretrained parameters specified in Proposition 4.4 are not guaranteed to be optimal in all settings, they are analytically useful for examining the effects of distribution shifts.

Based on Proposition 4.4, we arrive at the following:

Corollary 4.7. *Suppose there is no distribution shift between training and inference. Then, under the parameter configuration of Proposition 4.4, the Transformer model (7) emulates the Bayes-optimal linear model, implying that it is capable of in-context learning according to Definition 3.3.*

Since the pretrained model succeeds in ICL for $\mathcal{D}^{test} = \mathcal{D}^{train}$, we next investigate how distribution shifts affect its ICL performance.

ICL under distribution shift — To evaluate the impact of these distribution shifts on ICL performance, we assess whether adjustments to $\boldsymbol{M}$ and/or $\boldsymbol{V}$ are necessary to match the Bayes-optimal linear model under the new distribution.



(a) Effect of l/d when $\sigma_{test} = 10$ *(b)* Effect of σ_{test} when $l/d = 1$ *(c)* Optimal temperature

Figure 3. Effect of noise shift on Transformer (7). The pretraining noise is $\sigma_{train} = 0.1$, while σ_{test} varies across plots. The optimal temperature is set by Theorem 4.3. This setting matches Figure 2a, except for changes in test-time noise σ_{test}.

If so, the model is considered sensitive to the shift. Otherwise, it is deemed robust.

Case I: Shift in input distribution — Recall that inputs are drawn as $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, as defined in (2). Let $\boldsymbol{\mu}_x^{train}, \boldsymbol{\Sigma}_x^{train}$ and $\boldsymbol{\mu}_x^{test}, \boldsymbol{\Sigma}_x^{test}$ denote the input means and covariances for pretraining and testing, respectively. We consider two subcases:

(i) Mean shift ($\boldsymbol{\mu}_x^{train} \neq \boldsymbol{\mu}_x^{test}$): Centering renders the approximate softmax model invariant to mean shifts, but the uncentered linear attention model remains sensitive, as noted in Remark 3.4.

(ii) Covariance shift ($\boldsymbol{\Sigma}_x^{train} \neq \boldsymbol{\Sigma}_x^{test}$): Since $\boldsymbol{M}_{11}$ is fitted to the pretraining covariance, a mismatch drives the estimator away from Bayes-optimality, echoing prior results on linear attention (Zhang et al., 2024).

Case II: Shift in task distribution — The task vectors follow $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$. Let $\boldsymbol{\mu}_w^{train}, \boldsymbol{\Sigma}_w^{train}$ and $\boldsymbol{\mu}_w^{test}, \boldsymbol{\Sigma}_w^{test}$ be the mean and covariance of the task distribution during pretraining and testing, respectively. The Transformer model can incorporate $\boldsymbol{\mu}_w^{train}$ and $\boldsymbol{\Sigma}_w^{train}$ via the pretrained parameters $\boldsymbol{M}_{11}$ and $\boldsymbol{v}_{21}$ (see Proposition 4.4). However, as the context length l increases, the model's dependence on the task distribution diminishes. Thus, shifts in the task distribution primarily affect ICL performance for small l.

Case III: Shift in noise distribution — Finally, consider a change in the noise distribution: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with σ_{train}^2 and σ_{test}^2 denoting pretraining and testing noise variances. If $\sigma_{train}^2 \neq \sigma_{test}^2$, the parameters $\boldsymbol{M}_{11}$ and $\boldsymbol{v}_{21}$ become suboptimal relative to the Bayes-optimal linear model. However, as with the task distribution, the impact of noise shift diminishes as $l \to \infty$.

Summary — The Transformer model is robust to shifts in input mean but sensitive to input covariance changes. Shifts in task or noise distribution reduce ICL performance at small l, though increasing l mitigates these effects.

4.4. Interpretation of the optimal attention temperature (Theorem 4.3)

To obtain a simplified expression, we consider a simple but representative family of shifts as follows. The training distributions are

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \qquad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \qquad \epsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2).$$

We then introduce three independent shift parameters for the test distributions:

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, a\boldsymbol{I}), \qquad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, b\boldsymbol{I}), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $a > 0$ controls the input variance shift, $b > 0$ controls the task-parameter variance shift, and $\sigma > 0$ controls the noise-variance shift. This setting preserves isotropy, which makes it possible to derive a clean closed-form expression while still connecting directly to realistic distribution shifts.

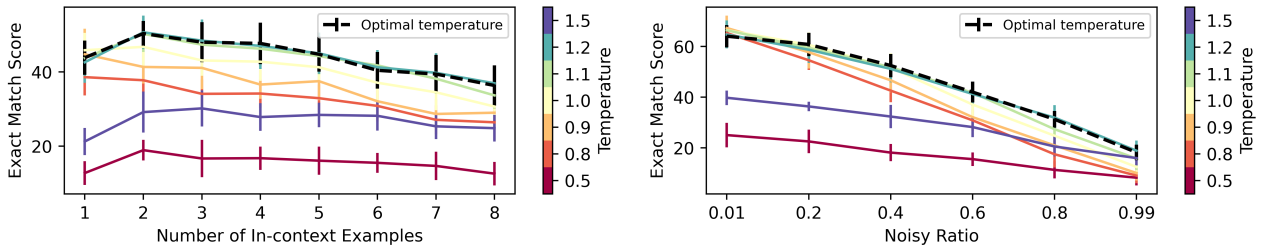
Substituting these shifted distributions into the optimal-temperature expression in (16) yields

$$\tau_{\text{opt}} = \frac{2\text{Tr}\left(a\boldsymbol{I}\boldsymbol{M}_{11}^T\left(ab\boldsymbol{I} + \frac{1}{l}(\sigma^2 + abd)\boldsymbol{I}\right)a\boldsymbol{I}\boldsymbol{M}_{11}\right)}{\text{Tr}\left(a\boldsymbol{I}\left(ab\boldsymbol{I}\left(\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\right)\right)\right)}, \tag{21}$$

$$= \left(a + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}^T\boldsymbol{M}_{11}\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}. \tag{22}$$

This concrete formula makes several effects fully explicit:

- *Input shift* — Increasing input variance a directly scales τ_{opt} upward. This aligns with our earlier results (Figure 2) and the heuristic derived in Appendix J, which suggests that a greater variance of pre-softmax scores requires a higher temperature to maintain robustness to input shifts.



(a) Effect of context length

(b) Effect of noisy ratio

Figure 4. Effect of attention temperature on the ICL performance of Llama2-7B (Touvron et al., 2023) on the SCIQ dataset (Welbl et al., 2017). Distribution shift is induced by injecting noisy yet "relevant" labels into in-context demonstrations following Gao et al. (2024). Panel (a) fixes the noisy ratio at 0.6; panel (b) fixes the number of in-context examples at 6. Results (averaged over 12 Monte Carlo runs) include error bars showing one standard deviation. Attention temperature of all the layers is set to $\tau\sqrt{d_k}$ for dimension independence, where d_k denotes the key dimension of the corresponding layer and τ refers to the temperature. Furthermore, the dashed black line marks the "optimal temperature" computed from the variance-to-mean ratio of pre-softmax scores, which is an insight derived from Theorem 4.3, as explained in Appendix I. Full experimental details appear in Appendix J.

- *Noise shift* — Increasing noise variance σ^2 also increases τ_{opt}, but only through the $\frac{1}{l}$ term, reflecting the diminishing effect of noise when more in-context examples are available.
- *Task shift* — Increasing task variance b reduces the effect of noise (via σ^2/b), slightly lowering the optimal temperature.
- *Context length* — As $l \to \infty$, the $\frac{1}{l}$ term vanishes, giving a simplified asymptotic rule: $\tau_{\text{opt}} \to a \cdot \frac{\text{Tr}(\boldsymbol{M}_{11}^T \boldsymbol{M}_{11})}{\text{Tr}(\boldsymbol{M}_{11})}$.

Note that under the considered training distribution, we have $\text{Tr}(\boldsymbol{M}_{11}^T \boldsymbol{M}_{11})/\text{Tr}(\boldsymbol{M}_{11}) \approx 1$, which implies that $\tau_{\text{optimal}} \to a$ as $l \to \infty$.

4.5. Possible extensions to the general softmax setting: the moment-ratio heuristic

We propose a practical heuristic (derived in Appendix I): the optimal attention temperature is roughly proportional to the ratio of the second and first moments of pre-softmax attention scores. The expression in (22) provides further theoretical justification for that heuristic.

Indeed, for some $i \neq j$,

$$\tau_{\text{opt}} = a\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)} + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}, \tag{23}$$

$$= \frac{\mathbb{E}\left[(\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j)^2\right]}{\mathbb{E}\left[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_i\right]} + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}, \tag{24}$$

where we used the moments calculated in Appendix I to reach the final line. Here, since $\text{Tr}(\boldsymbol{M}_{11}^T \boldsymbol{M}_{11})/\text{Tr}(\boldsymbol{M}_{11}) \approx 1$ for the considered training distribution, this gives the approximation:

$$\tau_{\text{opt}} \approx \underbrace{\frac{\mathbb{E}\left[(\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j)^2\right]}{\mathbb{E}\left[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_i\right]}}_{\text{moment-ratio}} + \underbrace{\frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)}_{\text{correction for small } l}. \tag{25}$$

This demonstrates that the moment-ratio heuristic is not an ad-hoc rule, but a theoretically grounded approximation of the exact closed-form optimal temperature.

5. Experimental results

In this section, we empirically validate our theory and show that optimal attention temperature consistently enhances generalization[1]. We begin with controlled linear regression experiments using (i) the simplified Transformer model with approximate softmax attention (7) and (ii) GPT-2 (Radford et al., 2019), which combines multi-head softmax attention with MLP layers[2]. These experiments confirm that our theoretical insights transfer from simplified to expressive architectures. Finally, we evaluate Llama2-7B (Touvron et al., 2023) on SCIQ in-context learning tasks (Welbl et al., 2017), demonstrating that temperature selection is a principled and effective lever for improving robustness in large language models.

5.1. Experiments on linear regression tasks

We consider a Transformer architecture with approximate softmax attention and no MLP layers, as analyzed in our theoretical development. Figures 2 and 3 illustrate its behavior on linear regression tasks (2). Theoretical predictions closely match empirical performance across a range of conditions, confirming the robustness of our analysis. In

[1]Our code is available at https://github.com/KU-MLIP/ICL-Optimal-Attention-Temperature.

[2]GPT-2 results are in Appendix J.

Figure 2, we compare the ICL performance of the model with and without applying the optimal temperature. As context length l increases (Figure 2a), the model's predictions converge to those of the Bayes-optimal linear model, validating its ICL capability. Figure 2b shows that under an input covariance shift, model performance degrades—but applying the optimal temperature restores alignment with the Bayes-optimal solution. Additionally, Figure 2c shows that the influence of task distribution shift decreases as l increases.

We further evaluate robustness to label noise in Figure 3. In Figure 3a, we observe that noise effects diminish as the context length increases, consistent with our theoretical predictions. However, at small l, temperature adjustment becomes critical. In Figure 3b (for $l = d$), the Transformer increasingly diverges from the Bayes-optimal model as noise grows, yet optimal temperature correction closes this gap. Figure 3c shows that the optimal temperature increases with noise level, indicating a principled relationship between noise and temperature under limited context.

5.2. Experiments with LLMs for in-context Q&A

To assess the practical relevance of our theoretical framework, we investigate how attention temperature impacts the ICL behavior of LLMs. Since the optimal temperature in Theorem 4.3 is not directly applicable here due to setting differences, we derive insights regarding temperature choice in other settings based on the optimal temperature in Theorem 4.3. Specifically, the insight is that the temperature choice should be proportional to the ratio of the variance of pre-softmax scores to the mean of those, which is described in Appendix I in detail.

Following Gao et al. (2024), we generate SCIQ-based (Welbl et al., 2017) ICL tasks that introduce distribution shift via noisy labels, with prompt examples and label construction detailed in Appendix J. We evaluate Llama2-7B (Touvron et al., 2023) using exact-match score.

Figure 4 shows the results. In (a) and (b), optimal attention temperature enhances the ICL performance for various context lengths and noisy ratios (analogous to Figures 3a and 3b)[3]. Furthermore, in (b), higher noise ratios push the optimal temperature upward, matching our theoretical prediction (cf. Figure 3c). To make this trend quantitative, Table 1 reports the best-performing temperature from our grid at each noise ratio in Figure 4b. The optimal temperature increases monotonically with noise — from below the default ($\tau < 1$) when demonstrations are nearly clean, to above it ($\tau > 1$) as noise grows — supporting our theoretical prediction.

Noisy ratio	0.01	0.2	0.4	0.6	0.8	0.99
Best temperature	0.9	1.0	1.1	1.1	1.2	1.2

Table 1. Best-performing attention temperature in Figure 4b as a function of the noisy ratio.

Together, these experiments demonstrate that the optimal temperature is not only theoretically motivated but also an effective tool for improving ICL robustness in the real world.

6. Conclusion

This work provides a unified theoretical and empirical account of how attention temperature governs the in-context learning (ICL) performance of pretrained Transformers under distribution shift. Using a simplified yet expressive framework based on *approximate softmax attention*, we analytically show how shifts in input covariance and label noise degrade ICL and derive an *optimal temperature* that provably minimizes generalization error. Extensive experiments on synthetic regression tasks, GPT-2, and Llama-2 validate our predictions, demonstrating that temperature selection is not a mere heuristic but a principled mechanism for improving robustness. Taken together, our results advance the theoretical understanding of Transformer behavior under distribution shift and establish attention temperature as a powerful, practical lever for building more adaptive and generalizable foundation models.

Limitations and future work — Our analysis makes three deliberate simplifications. First, the closed-form characterization in Theorems 4.2 and 4.3 is derived for *approximate* softmax attention; although Appendix C and Figure 1 show that the approximate and exact softmax behave similarly across the relevant temperature range, the two can deviate at very low temperature, so the exact-softmax optimal temperature should be viewed as approximated, not recovered, by (16). Second, we analyse a single-layer attention model with no MLP and fixed parameters; the role of training dynamics and depth on the optimal temperature is left open, although our experiments suggest the qualitative picture transfers. Third, our real-world experiments use a single QA-style ICL benchmark (SCIQ) with a specific label-noise model from Gao et al. (2024); a broader empirical taxonomy of temperature sensitivity across task types and shift mechanisms is a natural direction for follow-up work. Finally, we used our heuristic (25) to set a single temperature for all layers; per-layer or per-head variants are a promising but non-trivial extension, as reliable per-head moment estimation introduces additional challenges.

[3]Although Figure 3a shows error decreasing monotonically with context length while Figure 4a is non-monotonic, the two are not in conflict: they use qualitatively different noise models—i.i.d. additive label noise, which averages out as context grows, versus a fixed fraction of demonstrations carrying incorrect (but semantically related) labels, which does not. Figure 6 in Appendix J verifies this via an ablation across noisy ratios $\{0.01, 0.3, 0.6\}$.

Acknowledgements

This work is supported in part by TÜBİTAK under project 124E063 within the ARDEB 1001 program, and we gratefully acknowledge their support. S.D. is further supported by an AI Fellowship from the KUIS AI Center and a PhD Scholarship (BİDEB 2211) from TÜBİTAK.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, 2023.

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

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

Chen, X., Hu, Q., Li, K., Zhong, C., and Wang, G. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Fu, D., Chen, T.-Q., Jia, R., and Sharan, V. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, 2024.

Gao, H., Zhang, F., Jiang, W., Shu, J., Zheng, F., and Wei, H. On the noise robustness of in-context learning for text generation. In *Advances in Neural Information Processing Systems*, 2024.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 2022.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Han, D., Pu, Y., Xia, Z., Han, Y., Pan, X., Li, X., Lu, J., Song, S., and Huang, G. Bridging the divide: Reconsidering softmax and linear attention. In *Advances in Neural Information Processing Systems*, 2024.

Isserlis, L. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Lee, S. H., Lee, S., and Song, B. C. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, 2023.

Li, Y., Rawat, A. S., and Oymak, S. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *Advances in Neural Information Processing Systems*, 2024.

Lin, J., Sun, X., Ren, X., Li, M., and Su, Q. Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2985–2990, 2018.

Liu, J., Shen, D., Zhang, Y., Dolan, W. B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022.

Mahankali, A. V., Hashimoto, T., and Ma, T. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Park, C. F., Lubana, E. S., Pres, I., and Tanaka, H. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., and Zhong, Y. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI*, 2019.

Raventós, A., Paul, M., Chen, F., and Ganguli, S. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 2023.

Renze, M. and Guven, E. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP*, 2024.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Veličković, P., Perivolaropoulos, C., Barbero, F., and Pascanu, R. Softmax is not enough (for sharp size generalisation). *International Conference on Machine Learning*, 2025.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

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

Wu, Z., Wang, Y., Ye, J., Feng, J., Xu, J., Qiao, Y., and Wu, Z. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*, 2023.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Zhang, S., Zhang, X., Bao, H., and Wei, F. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 127–141, 2022.

Zou, Y., Ma, R., Li, Y., and Li, R. Attention temperature matters in vit-based cross-domain few-shot learning. In *Advances in Neural Information Processing Systems*, 2024.

A. Derivation of Bayes-optimal ridge estimator for $\boldsymbol{w}$

We derive the Bayes-optimal ridge estimator for $\boldsymbol{w}$ given a set of context samples. We place a Gaussian prior on $\boldsymbol{w}$, assumed to be a random vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with prior mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. Let the observed (centered) inputs and labels be

$$\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_{l-1}]^T, \quad \bar{\boldsymbol{y}} = [\bar{y}_1, \ldots, \bar{y}_{l-1}]^T,$$

and assume i.i.d. Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The likelihood of $\bar{\boldsymbol{y}}$ given $\boldsymbol{w}$ is

$$p(\bar{\boldsymbol{y}} \mid \bar{\boldsymbol{X}}, \boldsymbol{w}) = \prod_{i=1}^{l-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\bar{y}_i - \boldsymbol{w}^T \bar{\boldsymbol{x}}_i)^2}{2\sigma^2}\right] \tag{26}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})^T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})\right], \tag{27}$$

where $\propto$ denotes proportionality.

By Bayes' rule, the posterior of $\boldsymbol{w}$ is proportional to the product of likelihood and prior:

$$p(\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}) \propto p(\bar{\boldsymbol{y}} \mid \bar{\boldsymbol{X}}, \boldsymbol{w})\, p(\boldsymbol{w}). \tag{28}$$

Substituting the Gaussian prior yields

$$p(\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}) \propto \exp\Big[-\frac{1}{2\sigma^2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})^T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})\Big] \exp\Big[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\Big]. \tag{29}$$

To determine the form of the posterior distribution, we complete the square in the exponent by collecting all terms involving $\boldsymbol{w}$. Expanding the exponent in the joint expression from above, we obtain:

$$-\frac{1}{2\sigma^2}\left(\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}} - 2\bar{\boldsymbol{y}}^T\bar{\boldsymbol{X}}\boldsymbol{w} + \boldsymbol{w}^T\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\boldsymbol{w}\right) - \frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{w} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right). \tag{30}$$

Grouping the quadratic and linear terms in $\boldsymbol{w}$, we arrive at:

$$-\frac{1}{2}\boldsymbol{w}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{w} + \boldsymbol{w}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \text{terms independent of } \boldsymbol{w}. \tag{31}$$

Defining the posterior precision and linear coefficient terms as $\boldsymbol{\Sigma}_l^{-1} = \frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}$ and $b_l = \frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, the exponent can be rewritten as

$$-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}_l^{-1}\boldsymbol{w} + \boldsymbol{w}^T b_l = -\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_l) + \text{const}, \tag{32}$$

where $\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_l b_l$ denotes the posterior mean. Expanding this expression gives:

$$\boldsymbol{\mu}_l = \left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right). \tag{33}$$

Hence, the posterior distribution of $\boldsymbol{w}$ given the observed data is Gaussian:

$$\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{34}$$

where $\boldsymbol{\mu}_l$ is the posterior mean and $\boldsymbol{\Sigma}_l$ is the posterior covariance matrix.

Under squared-error loss, the Bayes-optimal estimator coincides with the posterior mean, yielding the Bayes-optimal ridge estimator:

$$\hat{\boldsymbol{w}}_{\text{Ridge}} = \mathbb{E}[\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}] = \boldsymbol{\mu}_l = \Big(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\Big)^{-1}\Big(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\Big). \tag{35}$$

This expression provides the Bayes-optimal ridge estimate of $\boldsymbol{w}$ under a Gaussian prior and additive Gaussian noise—minimizing expected squared error with respect to the posterior.

B. Derivation of approximate softmax

The function softmax : $\mathbb{R}^l \to \mathbb{R}^l$ is defined component-wise as

$$\text{softmax}(\boldsymbol{z})_i := \frac{e^{z_i}}{\sum_{j=1}^{l} e^{z_j}} \quad \forall i \in \{1, \ldots, l\}. \tag{36}$$

To obtain an approximation, we expand around the origin $\boldsymbol{z} = \mathbf{0}$ using a first-order Taylor series:

$$\text{softmax}(\boldsymbol{z}) \approx \text{softmax}(\mathbf{0}) + J_{\text{softmax}}(\mathbf{0})\boldsymbol{z}, \tag{37}$$

where $J_{\text{softmax}}(\mathbf{0})$ is the Jacobian matrix of the softmax function evaluated at $\boldsymbol{z} = \mathbf{0}$.

We first compute the zeroth-order term:

$$\text{softmax}(\mathbf{0}) = \frac{e^0}{\sum_{j=1}^{l} e^0}\mathbf{1} = \frac{1}{l}\mathbf{1}. \tag{38}$$

Next, we evaluate the Jacobian entries at $\boldsymbol{z} = \mathbf{0}$:

$$J_{\text{softmax}}(\mathbf{0})_{ii} = \text{softmax}(\mathbf{0})_i \left(1 - \text{softmax}(\mathbf{0})_i\right) = \frac{l-1}{l^2}, \quad \forall i, \tag{39}$$

$$J_{\text{softmax}}(\mathbf{0})_{ij} = -\text{softmax}(\mathbf{0})_i \cdot \text{softmax}(\mathbf{0})_j = -\frac{1}{l^2}, \quad \forall i \neq j. \tag{40}$$

This yields the compact matrix form:

$$J_{\text{softmax}}(\mathbf{0}) = \frac{1}{l}\boldsymbol{I} - \frac{1}{l^2}\mathbf{1}\mathbf{1}^T. \tag{41}$$

Substituting back, we obtain the approximate softmax:

$$\text{softmax}(\boldsymbol{z}) \approx \frac{1}{l}\mathbf{1} + \left(\frac{1}{l}\boldsymbol{I} - \frac{1}{l^2}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{z}, \tag{42}$$

$$= \left(\frac{1}{l} - \frac{1}{l^2}\sum_{j=1}^{l} z_j\right)\mathbf{1} + \frac{1}{l}\boldsymbol{z}, \tag{43}$$

$$=: \widehat{\text{softmax}}(\boldsymbol{z}). \tag{44}$$

This derivation yields the approximate softmax attention formulation in (7). From a practical standpoint, approximate softmax attention mechanisms have been empirically evaluated and shown to achieve performance comparable to standard softmax attention (Han et al., 2024).

C. Temperature effects for softmax and approximate softmax

The temperature parameter in softmax directly controls the variance of the output distribution. At higher temperatures, the variance across components decreases, and in the limit $\tau \to \infty$, all elements converge to $1/l$ with zero variance. Conversely, lower temperatures increase variance, and as $\tau \to 0^+$, the output approaches a one-hot vector, achieving maximal variance.

In the approximate case, temperature similarly acts as an inverse scaling of the variance of the output components, capturing the limit $\tau \to \infty$ (all elements equal to $1/l$). For $\tau \to 0^+$, approximate softmax also reflects the maximal variance, but it does not produce a true one-hot distribution. Thus, approximate softmax closely mirrors the temperature behavior of softmax, except in the degenerate limit $\tau \to 0^+$, which is not of practical relevance in this work.

To further illustrate these effects, Figure 1 compares softmax and approximate softmax across different temperatures. The figure demonstrates that approximate softmax faithfully captures the variance effect of temperature: the variance of the output components is inversely proportional to τ. Moreover, as $\tau \to \infty$, both softmax and approximate softmax concentrate around $1/l$, whereas linear attention with temperature scaling does not. Overall, the output distributions of softmax and approximate softmax are highly similar, except at very small values of τ, where approximate softmax may yield negative components while softmax tends toward sparsity with many zeros. By contrast, linear attention with temperature scaling produces qualitatively different distributions. This comparison highlights the advantage of approximate softmax as a faithful surrogate for analyzing temperature effects relevant to softmax.

D. Expanded form of approximate softmax attention

Using block matrix notation, the prediction from the approximate softmax attention model can be expanded as:

$$\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) = A_{d+1,l}, \tag{45}$$

$$= \frac{1}{l} \boldsymbol{V}_{d+1,:} \boldsymbol{Z} \left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T (\boldsymbol{Q}\boldsymbol{Z}_{:,l})}{\tau} - \frac{\boldsymbol{1}}{l} \sum_{j=1}^{l} \frac{(\boldsymbol{K}\boldsymbol{Z}_{:,j})^T (\boldsymbol{Q}\boldsymbol{Z}_{:,l})}{\tau} + \boldsymbol{1} \right), \tag{46}$$

$$= \frac{1}{l} [\boldsymbol{v}_{21}^T \;\; v_{22}] \boldsymbol{Z} \left(\frac{\boldsymbol{Z}^T \boldsymbol{M} \boldsymbol{Z}_{:,l}}{\tau} - \frac{\boldsymbol{1}}{l} \sum_{j=1}^{l} \frac{(\boldsymbol{Z}_{:,j})^T \boldsymbol{M} \boldsymbol{Z}_{:,l}}{\tau} + \boldsymbol{1} \right), \tag{47}$$

$$= \frac{1}{l} [\boldsymbol{v}_{21}^T \;\; v_{22}] [\boldsymbol{X} \;\; \boldsymbol{y}]^T \left(\frac{[\boldsymbol{X} \;\; \boldsymbol{y}] \boldsymbol{M} [\boldsymbol{x}_l^T \;\; 0]^T}{\tau} - \frac{\boldsymbol{1}}{l} \sum_{j=1}^{l} \frac{[\boldsymbol{x}_i^T \;\; y_i] \boldsymbol{M} [\boldsymbol{x}_l^T \;\; 0]^T}{\tau} + \boldsymbol{1} \right), \tag{48}$$

$$= \frac{1}{l} [\boldsymbol{v}_{21}^T \;\; v_{22}] [\boldsymbol{X} \;\; \boldsymbol{y}]^T \left(\frac{1}{\tau} \left[\boldsymbol{X} - \boldsymbol{1}\boldsymbol{s}_x^T \;\; \boldsymbol{y} - s_y \boldsymbol{1} \right] \begin{bmatrix} \boldsymbol{M}_{11} & * \\ \boldsymbol{m}_{21}^T & * \end{bmatrix} [\boldsymbol{x}_l^T \;\; 0]^T + \boldsymbol{1} \right), \tag{49}$$

$$= \frac{1}{l} [\boldsymbol{v}_{21}^T \;\; v_{22}] [\boldsymbol{X} \;\; \boldsymbol{y}]^T \left(\frac{1}{\tau} \left(\boldsymbol{X} - \boldsymbol{1}\boldsymbol{s}_x^T \right) \boldsymbol{M}_{11} \boldsymbol{x}_l + \frac{1}{\tau} \left(\boldsymbol{y} - s_y \boldsymbol{1} \right) \boldsymbol{m}_{21}^T \boldsymbol{x}_l + \boldsymbol{1} \right), \tag{50}$$

$$= \frac{1}{l} \left(\boldsymbol{v}_{21}^T \boldsymbol{X}^T + v_{22} \boldsymbol{y}^T \right) \left(\frac{1}{\tau} \left(\boldsymbol{X} - \boldsymbol{1}\boldsymbol{s}_x^T \right) \boldsymbol{M}_{11} \boldsymbol{x}_l + \frac{1}{\tau} \left(\boldsymbol{y} - s_y \boldsymbol{1} \right) \boldsymbol{m}_{21}^T \boldsymbol{x}_l + \boldsymbol{1} \right), \tag{51}$$

$$\begin{aligned} &= \frac{1}{\tau} \left(\boldsymbol{v}_{21}^T \left(\frac{\boldsymbol{X}^T \boldsymbol{X}}{l} - \boldsymbol{s}_x \boldsymbol{s}_x^T \right) + v_{22} \left(\frac{\boldsymbol{y}^T \boldsymbol{X}}{l} - s_y \boldsymbol{s}_x^T \right) \right) \boldsymbol{M}_{11} \boldsymbol{x}_l, \\ &\qquad + \frac{1}{\tau} \left(\boldsymbol{v}_{21}^T \left(\frac{\boldsymbol{X}^T \boldsymbol{y}}{l} - s_y \boldsymbol{s}_x \right) + v_{22} \left(\frac{\boldsymbol{y}^T \boldsymbol{y}}{l} - s_y^2 \right) \right) \boldsymbol{m}_{21}^T \boldsymbol{x}_l + \boldsymbol{v}_{21}^T \boldsymbol{s}_x + v_{22} s_y, \end{aligned} \tag{52}$$

$$= \frac{1}{\tau} \left(\boldsymbol{v}_{21}^T \boldsymbol{C}_{xx} + v_{22} \boldsymbol{C}_{xy}^T \right) \boldsymbol{M}_{11} \boldsymbol{x}_l + \frac{1}{\tau} \left(\boldsymbol{v}_{21}^T \boldsymbol{C}_{xy} + v_{22} C_{yy} \right) \boldsymbol{m}_{21}^T \boldsymbol{x}_l + \boldsymbol{v}_{21}^T \boldsymbol{s}_x + v_{22} s_y, \tag{53}$$

$$= \frac{1}{\tau} \left(\left(\boldsymbol{v}_{21}^T \boldsymbol{C}_{xx} + v_{22} \boldsymbol{C}_{xy}^T \right) \boldsymbol{M}_{11} + \left(\boldsymbol{v}_{21}^T \boldsymbol{C}_{xy} + v_{22} C_{yy} \right) \boldsymbol{m}_{21}^T \right) \boldsymbol{x}_l + \boldsymbol{v}_{21}^T \boldsymbol{s}_x + v_{22} s_y, \tag{54}$$

where the summary statistics are defined as:

$$\boldsymbol{s}_x := \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{x}_i, \qquad s_y := \frac{1}{l} \sum_{i=1}^{l-1} y_i,$$

$$\boldsymbol{C}_{xx} := \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{s}_x \boldsymbol{s}_x^T, \qquad \boldsymbol{C}_{xy} := \frac{1}{l} \sum_{i=1}^{l-1} y_i \boldsymbol{x}_i - s_y \boldsymbol{s}_x, \qquad C_{yy} := \frac{1}{l} \sum_{i=1}^{l-1} y_i^2 - s_y^2.$$

Then, we define

$$\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V}) = \boldsymbol{M}_{11}^T \left(\boldsymbol{C}_{xx} \boldsymbol{v}_{21} + v_{22} \boldsymbol{C}_{xy} \right) + \left(\boldsymbol{v}_{21}^T \boldsymbol{C}_{xy} + v_{22} C_{yy} \right) \boldsymbol{m}_{21}, \tag{55}$$

$$b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) = \boldsymbol{v}_{21}^T \boldsymbol{s}_x + v_{22} s_y, \tag{56}$$

which allows us to write

$$\hat{y}(\boldsymbol{Z};\boldsymbol{V},\boldsymbol{M}) = \frac{1}{\tau}\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V})^T\boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}). \tag{57}$$

E. Derivation of the pretraining for ICL by mimicking the Bayes-optimal estimator

Here, we derive the pretraining of the approximate softmax attention model by mimicking the Bayes-optimal ridge estimator (11). The prediction of the approximate softmax attention model can be written as

$$\hat{y}(\boldsymbol{Z};\boldsymbol{V},\boldsymbol{M}) = \frac{1}{\tau}\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V})^T\boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}), \tag{58}$$

which is derived in Appendix D. Furthermore, the Bayes-optimal ridge regression model's prediction is

$$\hat{y}_{Bayes} = \hat{\boldsymbol{w}}_{Bayes}^T\boldsymbol{x}_l. \tag{59}$$

Therefore, we select the parameters $\boldsymbol{M}$ and $\boldsymbol{V}$ such that

$$\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V}) \approx \hat{\boldsymbol{w}}_{Bayes}, \quad b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}) \approx 0, \tag{60}$$

which makes the prediction of the approximate softmax attention model approximately equal to that of the Bayes-optimal regression. Furthermore, we consider $\tau = 1$ for the pretraining. Let's first focus on $\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V})$ as follows

$$\begin{aligned}
&\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V}) \\
&\quad= \left(\boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right) + \left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}\right), && (61)\\
&\quad= \left(\boldsymbol{M}_{11}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{l}\boldsymbol{v}_{21} + v_{22}\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{l}\right) + \left(\boldsymbol{v}_{21}^T\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{l} + v_{22}\frac{\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}}}{l}\right)\boldsymbol{m}_{21}\right). && (62)
\end{aligned}$$

To reach the last line, we use the fact that $\boldsymbol{C}_{xx} := \boldsymbol{X}^T\boldsymbol{X}/l - \boldsymbol{s}_x\boldsymbol{s}_x^T = \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}/l$, $\boldsymbol{C}_{xy} := \boldsymbol{X}^T\boldsymbol{y}/l - s_y\boldsymbol{s}_x = \bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}/l$ and $C_{yy} = \bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}}/l$, where $\bar{\boldsymbol{X}} := \boldsymbol{X} - \boldsymbol{s}_x^T$ and $\bar{\boldsymbol{y}} := \boldsymbol{y} - s_y$ denote centered input matrix and centered label vector. Now, recall that the Bayes-optimal ridge estimator is

$$\hat{\boldsymbol{w}}_{Bayes} = \left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w\right), \tag{63}$$

as derived in Appendix A. Looking at equations (63) and (62) together, we can see that setting the parameters as follows would make $\hat{\boldsymbol{w}}_{Att} = \hat{\boldsymbol{w}}_{Bayes}$ hold

$$\boldsymbol{M}_{11} = \frac{l}{\sigma^2}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{l}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{l}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = 1. \tag{64}$$

However, while Bayes-optimal estimator $\hat{\boldsymbol{w}}_{Bayes}$ is different for each sample, the attention model should be pretrained and fixed. Thus, we replace $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ in (64) with $\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}/m$ as follows, where $\hat{\boldsymbol{X}} \in \mathbb{R}^{ml\times d}$ is the centred input matrix including all the pretraining data consisting of ml samples.

$$\boldsymbol{M}_{11} = \frac{l}{\sigma^2}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{m\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{l}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = 1. \tag{65}$$

In practice, the variance of noise σ^2, the mean $\boldsymbol{\mu}_w$, and covariance $\boldsymbol{\Sigma}_w$ of the task vectors are unknown. Yet, we can use their estimates based on the (pre)training data.

Now, we can focus on making $b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}) \approx 0$ hold as follows

$$b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}) = \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y, \tag{66}$$

where $\boldsymbol{s}_x$ and s_y are based on data so we have no control over them. Instead, by using Assumptions 3.1 and 3.2, we can choose $\boldsymbol{v}_{21}$ and v_{22} such that $b_{Att} \to 0$ as $l, d \to \infty$. Note that Assumption 3.1 makes $\boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y$ bounded with high

probability for $\boldsymbol{v}_{21}$ and v_{22} given in (65). Therefore, multiplying $\boldsymbol{v}_{21}, v_{22}$ given in (65) with $1/d$ would make $b_{Att} \to 0$ as $d \to \infty$. To fix the impact of the multiplication for $\hat{\boldsymbol{w}}_{Att}$, we can multiply $\boldsymbol{M}_{11}$ with d as well. So, by applying the mentioned multiplications, we reach the following pretrained parameters mimicking the Bayes-optimal regression model

$$\boldsymbol{M}_{11} = \frac{dl}{\sigma^2}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{m\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{dl}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = \frac{1}{d}. \tag{67}$$

F. Characterization of generalization error for ICL under distribution shift

Here, we characterize the generalization error for in-context learning under distribution shift, given that $\boldsymbol{M}$ and $\boldsymbol{V}$ are pretrained and fixed. So, the impact of pretraining distribution $\mathcal{D}^{train}$ is captured by $\boldsymbol{M}$ and $\boldsymbol{V}$. Suppose that $\mathcal{D}^{test}$ denotes the test distribution. To avoid additional notations, here, we again use $\boldsymbol{\mu}_x, \boldsymbol{\mu}_w, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_w, \sigma^2$ to denote means and covariances for input and task vectors and noise variance for the inference (test). However, note that these can be different from those used for pretraining. We begin studying the generalization error defined in (10) as follows

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) := \mathbb{E}_{(\boldsymbol{Z}, y_l)\sim\mathcal{D}^{test}}\left[\left(y_l - \hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M})\right)^2\right], \tag{68}$$

$$= \mathbb{E}_{(\boldsymbol{Z}, y_l)\sim\mathcal{D}^{test}}\left[\left(\frac{1}{\tau}\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V})^T\boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) - y_l\right)^2\right], \tag{69}$$

$$= \mathbb{E}_{(\boldsymbol{Z}, y_l)\sim\mathcal{D}^{test}}\left[\left(\frac{1}{\tau}\left(\boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right)\right)^T\boldsymbol{x}_l - y_l\right)^2\right], \tag{70}$$

where we use the parameters from pretraining (67) together with Assumptions 3.1 and 3.2 to reach the last line. Then,

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \mathbb{E}_{(\boldsymbol{Z}, y_l)\sim\mathcal{D}^{test}}\left[\left(\frac{1}{\tau}\left(\boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right)\right)^T\boldsymbol{x}_l - y_l\right)^2\right], \tag{71}$$

$$= \mathbb{E}\left[\left(\frac{1}{\tau}\left(\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\le l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i(\bar{\boldsymbol{x}}_i^T\boldsymbol{w} + \epsilon_i)\right)\right)^T\boldsymbol{x}_l - \boldsymbol{w}^T\boldsymbol{x}_l - \epsilon_l\right)^2\right], \tag{72}$$

$$= \mathbb{E}\left[\left(\frac{1}{\tau}\left(\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\le l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i(\bar{\boldsymbol{x}}_i^T\boldsymbol{w} + \epsilon_i)\right)\right)^T\boldsymbol{x}_l - \boldsymbol{w}^T\boldsymbol{x}_l\right)^2\right] + \sigma^2 \tag{73}$$

where $\bar{\boldsymbol{x}}_i := \boldsymbol{x}_i - \boldsymbol{s}_x = \boldsymbol{x}_i - \frac{1}{l}\sum_{i\le l}\boldsymbol{x}_i$ and we use $\epsilon_l \sim \mathcal{N}(0, \sigma^2)$ to reach the final line. We continue by defining

$$\boldsymbol{w}_{diff} := \frac{1}{\tau}\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i(\bar{\boldsymbol{x}}_i^T\boldsymbol{w} + \epsilon_i)\right) - \boldsymbol{w}, \tag{74}$$

which allows us to write

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \mathbb{E}\left[\left(\boldsymbol{w}_{diff}^T\boldsymbol{x}_l\right)^2\right] + \sigma^2, \tag{75}$$

$$= \mathbb{E}\left[\boldsymbol{w}_{diff}^T\mathbb{E}_{\boldsymbol{x}_l}[\boldsymbol{x}_l\boldsymbol{x}_l^T]\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{76}$$

$$= \mathbb{E}\left[\boldsymbol{w}_{diff}^T\left(\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_x\right)\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{77}$$

by the law of total expectation since $\boldsymbol{w}_{diff}$ is independent of $\boldsymbol{x}_l$. Note that when writing (75), we safely ignore terms with $(1/l)\bar{\boldsymbol{x}}_l\bar{\boldsymbol{x}}_l^T\boldsymbol{v}_{21}$ in (73) since they vanish as $l \to \infty$ by Assumptions 3.1-3.2 and 4.1. Letting $\boldsymbol{A} := \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_x$, we write

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \mathbb{E}\left[\boldsymbol{w}_{diff}^T\boldsymbol{A}\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{78}$$

$$= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{w}_{diff}^T\boldsymbol{A}\boldsymbol{w}_{diff}\right)\right] + \sigma^2, \tag{79}$$

$$= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T\right)\right] + \sigma^2, \tag{80}$$

$$= \mathrm{Tr}\left(\boldsymbol{A}\mathbb{E}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right) + \sigma^2, \tag{81}$$

where we first apply the cyclic property of trace and then use the linearity of expectation and trace to reach the last line. Now, we need to calculate $\mathbb{E}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]$, for which we first take the expectation over $\boldsymbol{w}$. To do so, we rewrite $\boldsymbol{w}_{diff}$ as

$$\boldsymbol{w}_{diff} = \underbrace{\frac{1}{\tau}\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right)}_{\boldsymbol{e}} + \underbrace{\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)}_{\boldsymbol{D}}\boldsymbol{w}, \tag{82}$$

$$= \boldsymbol{e} + \boldsymbol{D}\boldsymbol{w}, \tag{83}$$

where we define

$$\boldsymbol{e} := \frac{1}{\tau}\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right), \tag{84}$$

$$\boldsymbol{D} := \left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right). \tag{85}$$

Since $\boldsymbol{e}$ and $\boldsymbol{D}$ are independent of $\boldsymbol{w}$, we can easily calculate $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]$ as follows

$$\mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[(\boldsymbol{e}+\boldsymbol{D}\boldsymbol{w})(\boldsymbol{e}+\boldsymbol{D}\boldsymbol{w})^T]\right], \tag{86}$$

$$= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{e}\boldsymbol{\mu}_w^T\boldsymbol{D}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}(\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_w)\boldsymbol{D}^T\right], \tag{87}$$

$$= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]^T + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right], \tag{88}$$

where we first apply the law of total expectation, then take the expectation over $\boldsymbol{w}$ and finally, we define $\boldsymbol{B} := \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_w$ to reach the last line. Note that $\boldsymbol{\mu}_w$ and $\boldsymbol{B}$ are fixed while $\boldsymbol{e}$ and $\boldsymbol{D}$ are random in the last line. Therefore, we are required to calculate the three expectations that appeared in (88).

Before getting into the calculations of the aforementioned expectations, we provide the following lemma that is useful for the calculation of the expectations.

Lemma F.1. *Let $\bar{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\bar{\boldsymbol{x}} \in \mathbb{R}^d$. Let $\bar{\boldsymbol{x}}_i$ be $l-1$ independent samples of $\bar{\boldsymbol{x}}$ for $i = 1, \ldots, l-1$. Furthermore, let $\boldsymbol{A}$ be a fixed $d \times d$ matrix. Then, the following holds*

$$\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{A}\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right] = \frac{l-1}{l}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma} + \frac{1}{l}Tr(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}. \tag{89}$$

Proof. This is proven by using Isserlis' theorem (Isserlis, 1918) in Appendix G. □

Note that our inputs $\bar{\boldsymbol{x}}_i$ are centered, i.e., $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \frac{1}{l}\sum_{i\leq l}\boldsymbol{x}_i$, so their distribution is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_x)$ as $l \to \infty$. Therefore, Lemma F.1 is directly applicable in our setting.

Next, we start the calculations of the expectations in (88) with $\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right]$ as follows

$$\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] = \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right) \cdot \left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\boldsymbol{M}_{11}, \tag{90}$$

$$= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21}\right)\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right) + \left(v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right)\left(v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\right)\boldsymbol{M}_{11}, \tag{91}$$

$$= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right) + \left(v_{22}^2\frac{\sigma^2}{l^2}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\right)\boldsymbol{M}_{11}, \tag{92}$$

$$= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x + v_{22}^2\frac{\sigma^2(l-1)}{l^2}\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11}, \tag{93}$$

$$= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(v_{22}^2\frac{\sigma^2}{l}\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11}, \tag{94}$$

where we first use the independence of the random variables and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ to simplify the equation. Then, we apply Lemma F.1 and use the fact that $\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \boldsymbol{\Sigma}_x$ to get the penultimate line. Finally, we drop the vanishing terms and simplify the result using Assumptions 3.1-3.2 and 4.1 in order to reach the last line.

We continue with the calculation of $\mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]$ as

$$\mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]$$

$$= \frac{1}{\tau}\mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)\boldsymbol{\mu}_w\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\boldsymbol{M}_{11}, \tag{95}$$

$$= \frac{1}{\tau}\mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)\boldsymbol{\mu}_w\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11}, \tag{96}$$

$$= \frac{1}{\tau}\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11} - \boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\mathbb{E}\left[\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right]\boldsymbol{M}_{11}, \tag{97}$$

$$= \frac{1}{\tau}\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\boldsymbol{\Sigma}_x\boldsymbol{v}_{21}\boldsymbol{\mu}_w^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11} - \frac{1}{\tau}\frac{l-1}{l}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\boldsymbol{M}_{11}, \tag{98}$$

$$= \frac{v_{22}}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\boldsymbol{M}_{11}, \tag{99}$$

where we again first use the independence of the random variables and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$. Then, we apply basic algebraic manipulations. To reach the penultimate line, we utilize Lemma F.1 together with the fact that $\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \boldsymbol{\Sigma}_x$. Using Assumptions 3.1-3.2 and 4.1, we reach the last line.

Finally, we calculate $\mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right]$ as follows

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right] \\
&= \mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)\boldsymbol{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{M}_{11} - \boldsymbol{I}\right)\right], && (100)\\
&= \mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{M}_{11}\right)\right] \\
&\quad - \mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{B}\right] - \mathbb{E}\left[\boldsymbol{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{M}_{11}\right)\right] + \boldsymbol{B}, && (101)\\
&= \frac{v_{22}^2}{\tau^2}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{B}\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11} \\
&\quad - \frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{B} - \frac{v_{22}}{\tau}\boldsymbol{B}\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11} + \boldsymbol{B}, && (102)\\
&= \frac{v_{22}^2}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{B}\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11} - \frac{v_{22}}{\tau}\frac{l-1}{l}\boldsymbol{M}_{11}^T\boldsymbol{\Sigma}_x\boldsymbol{B} \\
&\quad - \frac{v_{22}}{\tau}\frac{l-1}{l}\boldsymbol{B}\boldsymbol{\Sigma}_x\boldsymbol{M}_{11} + \boldsymbol{B}, && (103)\\
&= \frac{v_{22}^2}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{B}\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11} - \frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\boldsymbol{\Sigma}_x\boldsymbol{B} - \frac{v_{22}}{\tau}\boldsymbol{B}\boldsymbol{\Sigma}_x\boldsymbol{M}_{11} + \boldsymbol{B}, && (104)
\end{aligned}$$

where we first do basic algebraic manipulations. Then, we use Lemma F.1 and $\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \boldsymbol{\Sigma}_x$ to get the penultimate line. For the final line, we utilize $l \to \infty$ by Assumption 3.2.

Putting the found expectation results into (88), we get

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right] &= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]^T + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right], && (105)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{F}_2\boldsymbol{M}_{11} + \frac{1}{\tau}\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T + \boldsymbol{B}. && (106)
\end{aligned}$$

where matrices $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are defined as

$$\begin{aligned}
\boldsymbol{F}_1 &:= v_{22}^2\frac{\sigma^2}{l}\boldsymbol{\Sigma}_x + v_{22}\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right) && (107)\\
&\quad + v_{22}\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)^T + v_{22}^2\left(\boldsymbol{\Sigma}_x\boldsymbol{B}\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right), \\
&= \left(\boldsymbol{\Sigma}_x\hat{\boldsymbol{B}} + \left(v_{22}^2\frac{\sigma^2}{l} + \frac{1}{l}\mathrm{Tr}\left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_x\right)\right)\boldsymbol{I}\right)\boldsymbol{\Sigma}_x, && (108)\\
\boldsymbol{F}_2 &:= \boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + v_{22}\boldsymbol{B}\boldsymbol{\Sigma}_x = (\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{B})\boldsymbol{\Sigma}_x, && (109)
\end{aligned}$$

with $\hat{\boldsymbol{B}} := v_{22}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{v}_{21}\boldsymbol{\mu}_w^T + v_{22}^2\boldsymbol{B}$.

Going back to generalization error in (81), we have

$$\begin{aligned}
\mathcal{G}(\boldsymbol{V},\boldsymbol{M}) &= \mathrm{Tr}\left(\boldsymbol{A}\mathbb{E}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right) + \sigma^2, && (110)\\
&= \mathrm{Tr}\left(\boldsymbol{A}\left(\frac{1}{\tau^2}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{F}_2\boldsymbol{M}_{11} + \frac{1}{\tau}\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T + \boldsymbol{B}\right)\right) + \sigma^2, && (111)
\end{aligned}$$

where $\boldsymbol{F}_1 = \left(\boldsymbol{\Sigma}_x\hat{\boldsymbol{B}} + \frac{1}{l}\left(v_{22}^2\sigma^2 + \mathrm{Tr}\left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_x\right)\right)\boldsymbol{I}\right)\boldsymbol{\Sigma}_x$, and $\boldsymbol{F}_2 = (\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{B})\boldsymbol{\Sigma}_x$. Furthermore, $\hat{\boldsymbol{B}}$ is defined as $\hat{\boldsymbol{B}} := v_{22}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{v}_{21}\boldsymbol{\mu}_w^T + v_{22}^2\boldsymbol{B}$.

G. Proof of Lemma F.1

We first restate the lemma as follows.

Let $\bar{\boldsymbol{x}} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\bar{\boldsymbol{x}} \in \mathbb{R}^d$. Let $\bar{\boldsymbol{x}}_i$ be l independent samples of $\bar{\boldsymbol{x}}$ for $i = 1, \ldots, l$. Let $\boldsymbol{A}$ be a fixed $d \times d$ matrix. Then, the following holds

$$\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{A}\left(\frac{1}{l}\sum_{i\leq l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right] = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma} + \frac{1}{l}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}. \tag{112}$$

Proof. Let $\boldsymbol{S}_x = \frac{1}{l}\sum_{i=1}^{l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T$. First, note that $E[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \boldsymbol{\Sigma}$ since $\bar{\boldsymbol{x}}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$.

Thus, $\mathbb{E}[\boldsymbol{S}_x] = \frac{1}{l}\sum_{i=1}^{l}\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \frac{1}{l}\sum_{i=1}^{l}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$. We have

$$\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x = \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T \tag{113}$$

Taking the expectation, we get

$$\mathbb{E}[\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x] = \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] \tag{114}$$

When $i \neq j$, $\bar{\boldsymbol{x}}_i$ and $\bar{\boldsymbol{x}}_j$ are independent, so

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] = \mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T]\boldsymbol{A}\mathbb{E}[\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} \tag{115}$$

When $i = j$,

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \mathbb{E}[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\boldsymbol{A}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T] \tag{116}$$

Let $\bar{\boldsymbol{x}} = [x_1, x_2, \ldots, x_d]^T$. Then, from Isserlis' theorem (Isserlis, 1918), we have

$$\mathbb{E}[x_ix_jx_kx_l] = \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{kl} + \boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{jl} + \boldsymbol{\Sigma}_{il}\boldsymbol{\Sigma}_{jk} \tag{117}$$

Let $\boldsymbol{A} = [a_{ij}]$. Then, $\bar{\boldsymbol{x}}^T\boldsymbol{A}\bar{\boldsymbol{x}} = \sum_{i,j}a_{ij}x_ix_j$. Thus, we reach

$$\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\boldsymbol{A}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T = \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\sum_{i,j}a_{ij}x_ix_j, \tag{118}$$

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma}. \tag{119}$$

There are l^2 terms in the double sum. l terms are of the form $\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T]$ and $l^2 - l$ terms are of the form $\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}$. Therefore, we can write

$$\mathbb{E}[\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x] = \frac{1}{l^2}[l(\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma}) + l(l-1)\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}], \tag{120}$$

$$= \frac{1}{l}(\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma}) + \frac{l-1}{l}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}, \tag{121}$$

$$= \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma} + \frac{1}{l}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}, \tag{122}$$

which completes the proof. □

H. Analysis of optimal temperature for ICL under distribution shift

Here, we find the optimal temperature minimizing the generalization error. First, recall that we have the following generalization error.

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \frac{1}{\tau^2}\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right) - \frac{1}{\tau}\mathrm{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right) + \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) + \sigma^2, \tag{123}$$

as specified in Theorem 4.2. So, we can express the generalization error as,

$$\mathcal{G}(\tau; \boldsymbol{V}, \boldsymbol{M}) = \frac{a}{\tau^2} - \frac{b}{\tau} + c, \tag{124}$$

where $a := \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right)$, $b := \mathrm{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right)$, and $c = \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) + \sigma^2$. Therefore, we have the following optimization problem

$$\tau_{\text{opt}} := \arg\min_{\tau} \mathcal{G}(\tau; \boldsymbol{V}, \boldsymbol{M}), \tag{125}$$

$$= \arg\min_{\tau} \left\{\frac{a}{\tau^2} - \frac{b}{\tau} + c\right\}. \tag{126}$$

To find the optimal value of τ that minimizes the given function, we can take the derivative of the expression with respect to τ and set it to zero. From now on, we consider generalization error as a function of τ, written as $\mathcal{G}(\tau)$.

Next, find the derivative of $\mathcal{G}(\tau)$ with respect to τ as

$$\mathcal{G}'(\tau) = -2a\tau^{-3} + b\tau^{-2}. \tag{127}$$

To find the critical points, set $\mathcal{G}'(\tau) = 0$ as follows

$$\mathcal{G}'(\tau) = -2a\tau^{-3} + b\tau^{-2} = 0, \tag{128}$$

Solving this equation for τ, we reach the following critical point

$$\tau = \frac{2a}{b}. \tag{129}$$

Now, we need to check if this is a minimum by taking the second derivative, which is

$$\mathcal{G}''(\tau) = 6a\tau^{-4} - 2b\tau^{-3}. \tag{130}$$

Evaluate $\mathcal{G}''(\tau)$ at $\tau = \frac{2a}{b}$ as follows

$$\mathcal{G}''\left(\frac{2a}{b}\right) = 6a\left(\frac{2a}{b}\right)^{-4} - 2b\left(\frac{2a}{b}\right)^{-3} = 6a\left(\frac{b^4}{16a^4}\right) - 2b\left(\frac{b^3}{8a^3}\right) = \frac{b^4}{8a^3}. \tag{131}$$

Since $a, b > 0$, we reach $\mathcal{G}''\left(\frac{2a}{b}\right) = \frac{b^4}{8a^3} > 0$, which means the function has a minimum at $\tau = \frac{2a}{b}$. Therefore, $\tau_{\text{opt}} = \frac{2a}{b}$ is the solution minimizing the generalization error $\mathcal{G}(\tau)$. Writing a, b back into the optimal solution, we get

$$\tau_{\text{opt}} = \frac{2\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right)}{\mathrm{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right)}, \tag{132}$$

which concludes our derivation of the optimal temperature τ_{opt}.

I. An insight driven from optimal temperature for other settings

In this section, we extract a mathematical heuristic from the optimal temperature in Theorem 4.3 that can be applied to ICL settings beyond our existing setting involving approximate softmax attention and regression tasks. Specifically, we consider Transformers employing standard softmax attention. Recall that the attention temperature scales the pre-softmax scores (i.e., $(\boldsymbol{KZ})^T(\boldsymbol{QZ})$ in (1)), thereby controlling the variance of the final scores. Since the optimal temperature depends on the distribution of these scores, it can be naturally characterized by the moments of that distribution. Our central intuition is that the optimal temperature identified in Theorem 4.3 relates directly to the first two moments of the pre-softmax scores. Although this optimal temperature was derived for *approximate softmax* attention, the insight remains relevant for softmax attention because the two mechanisms behave similarly in the regime considered (see Appendix C).

We now illustrate how the optimal temperature in Theorem 4.3 can be related to the first two moments of the pre-softmax scores. For simplicity, we consider the case $\boldsymbol{\mu}_x = \boldsymbol{\mu}_w = \boldsymbol{m}_{21} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_w = \boldsymbol{I}$, under which the optimal temperature reduces to

$$\tau_{\text{opt}} = \frac{v_{22}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^T \boldsymbol{\Sigma}_x\right)}{\frac{1}{2}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_x\left(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T \boldsymbol{\Sigma}_x^T\right)\right)}. \tag{133}$$

We next show how this expression connects to the first two moments of $(\boldsymbol{KZ})^T(\boldsymbol{QZ})$. Let $\boldsymbol{z}_i$ denote the i-th column of $\boldsymbol{Z}$ from (5) and recall $\boldsymbol{K}^T\boldsymbol{Q} = \boldsymbol{M}$. We therefore compute $\mathbb{E}[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j]$ and $\mathbb{E}[(\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j)^2]$ for $i, j \in \{1, \ldots, l\}$. Starting with the first moment for $i = j$:

$$\mathbb{E}[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_i] = \mathrm{Tr}\left(\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^T]\boldsymbol{M}\right) = \mathrm{Tr}\left(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11}\right), \tag{134}$$

where the block structure (and zero entries) of $\boldsymbol{M}$ is used in the last step. For $i \neq j$,

$$\mathbb{E}[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j] = \mathrm{Tr}\left(\boldsymbol{M}\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_j^T]\right) = 0, \tag{135}$$

by independence of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. For the second moment with $i \neq j$:

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j)^2\right] &= \mathbb{E}\left[\boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j \boldsymbol{z}_j^T \boldsymbol{M}^T \boldsymbol{z}_i\right] && (136)\\
&= \mathbb{E}\left[\boldsymbol{x}_i^T \boldsymbol{M}_{11} \boldsymbol{x}_j \boldsymbol{x}_j^T \boldsymbol{M}_{11}^T \boldsymbol{x}_i\right] && (137)\\
&= \mathbb{E}_{\boldsymbol{x}_i}\left[\boldsymbol{x}_i^T \boldsymbol{M}_{11} \mathbb{E}_{\boldsymbol{x}_j}[\boldsymbol{x}_j \boldsymbol{x}_j^T] \boldsymbol{M}_{11}^T \boldsymbol{x}_i\right] && (138)\\
&= \mathbb{E}_{\boldsymbol{x}_i}\left[\boldsymbol{x}_i^T \boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^T \boldsymbol{x}_i\right] && (139)\\
&= \mathrm{Tr}\left(\boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^T \mathbb{E}_{\boldsymbol{x}_i}[\boldsymbol{x}_i \boldsymbol{x}_i^T]\right) && (140)\\
&= \mathrm{Tr}\left(\boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^T \boldsymbol{\Sigma}_x\right), && (141)
\end{aligned}$$

where we again exploit the block structure of $\boldsymbol{M}$ and apply straightforward manipulations.

We observe a parallel between the numerator of (133) and the computed second moment (for $i \neq j$), and between the denominator and the first moment (for $i = j$). This motivates the heuristic that the optimal temperature should be roughly proportional to the ratio of the second moment (for $i \neq j$) to the first moment (for $i = j$). Accordingly, in our LLM experiments (Figure 4), we select the temperature proportional to this ratio while taking care to avoid numerical issues.

Finally, we note an important caveat: in order to obtain an insight of practical relevance, we intentionally relaxed the rigor applied in our main theoretical results. Consequently, the heuristic derived here—and the accompanying empirical findings—should be viewed as preliminary, intended to inspire future work on principled selection of attention temperature in practice.

J. Experimental details and GPT-2 experiments

This section describes our experimental setups for GPT-2 and large language models (LLMs), including the motivation for our distribution-shift scenarios.

J.1. GPT-2: Transformer with MLP layers

Building on the approximate-attention experiments, we investigate whether the optimal temperature also benefits more complex Transformer models on linear regression tasks. We evaluate GPT-2 (Radford et al., 2019) under a shift in input

covariance (Figure 5). Consistent with prior work (Garg et al., 2022; Zhang et al., 2024), such shifts substantially degrade performance and can even induce nonmonotonic generalization error with respect to context length l. Remarkably, applying the optimal temperature mitigates this nonmonotonicity and improves in-context generalization.

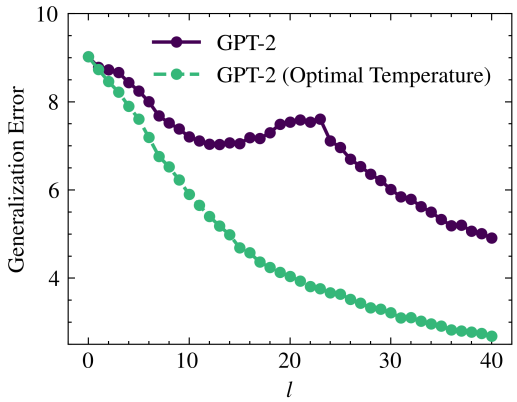


Figure 5. GPT-2 (Radford et al., 2019) under an input-covariance shift. GPT-2 exemplifies the Transformer architecture (Vaswani et al., 2017), combining multi-layer perceptrons with multi-head softmax self-attention. The model here is pretrained by Garg et al. (2022) on the linear regression tasks defined in (2). We consider a shift from $\boldsymbol{\Sigma}_x^{\text{train}} = \boldsymbol{I}$ to $\boldsymbol{\Sigma}_x^{\text{test}} = 3\boldsymbol{I}$. The attention temperature at each layer is scaled as $\tau\sqrt{d_k}$ (where d_k is the key dimension) to ensure dimension-independent τ values.

J.2. Details of the GPT-2 experiments in Figure 5

We use the standard GPT-2 architecture (Radford et al., 2019) as implemented in HuggingFace (Wolf et al., 2020), leveraging the pretrained model of Garg et al. (2022). Training data match ours, while their training procedure differs slightly: the loss is auto-regressive, i.e., the average over the entire context sequence of length $l = 40$. We adopt the same embedding method as in Garg et al. (2022). The input dimension is $d = 20$, with 12 layers and 8 heads. All GPT-2 experiments run on an NVIDIA Tesla V100 GPU and complete in approximately 10 minutes.

J.3. Details of the LLM experiments in Figure 4

For our large language model experiments, we employ Llama2-7B (Touvron et al., 2023) and the SCIQ dataset (Welbl et al., 2017), which contains science questions with supporting information. We generate ICL problems following Gao et al. (2024), selecting in-context demonstrations using the TopK retrieval technique (Liu et al., 2022) to ensure relevance. An example ICL sample from SCIQ appears in Table 2. To simulate distribution shift, we follow Gao et al. (2024) and introduce noisy labels—incorrect but semantically related—to the in-context demonstrations (Appendix J.4). Table 3 gives an example. The noisy ratio denotes the fraction of demonstrations with noisy labels (e.g., 0.6 means 60% noisy). We modify and use the codebase of Gao et al. (2024), built on HuggingFace (Wolf et al., 2020) and OpenICL (Wu et al., 2023). All LLM experiments run on an NVIDIA A40 GPU; a single Monte Carlo run per plot in Figure 4 takes a few hours.

J.4. Why in-context demonstrations with noisy labels as an example of distribution shift?

The link between noisy labels in demonstrations and distribution shift may not be immediately obvious. Quantifying pretraining–test shifts for pretrained LLMs is inherently difficult because their pretraining data are complex mixtures of sources (Touvron et al., 2023). However, we hypothesize—following Gao et al. (2024)—that high perplexity can serve as an empirical indicator of distribution shift. Inputs aligned with the training distribution tend to yield low perplexity (high-confidence generation), whereas contradictory or out-of-distribution inputs induce high perplexity. Since noisy demonstrations are expected to contradict training-set patterns, they yield high perplexity and thereby act as a proxy for distribution shift. Consequently, introducing noisy labels into in-context demonstrations constitutes a principled way to test the robustness of in-context learning under distribution shift.

In-context demonstration 1	
Support:	Cells are organized into tissues, tissues are organized into organs.
Question:	What is considered the smallest unit of the organ?
Answer:	Cells
In-context demonstration 2	
Support:	...four basic types of tissue: connective, muscle, nervous, and epithelial.
Question:	The four basic types of tissue are epithelial, muscle, connective, and what?
Answer:	nervous
⋮	
Test example	
Support:	All forms of life are built of at least one cell. A cell is the basic unit of life.
Question:	What are the smallest structural and functional units of all living organisms?
Output:	???

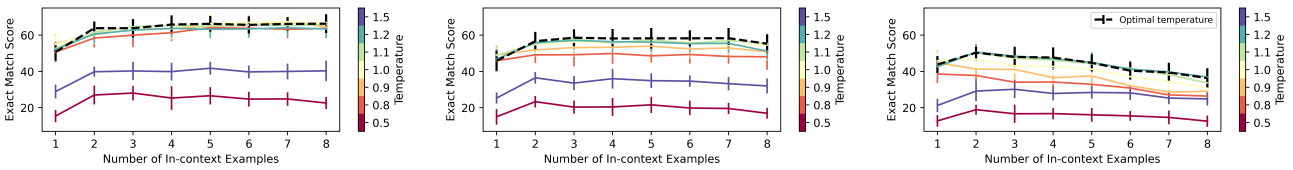
Table 2. A sample illustration of in-context learning on the SCIQ dataset.

Setting	**In-context demonstration**
True Label	**Support:** Cells are organized into tissues, tissues are organized into organs. **Question:** What is considered the smallest unit of the organ? **Label:** Cells
Noisy Label	**Support:** Cells are organized into tissues, tissues are organized into organs. **Question:** What is considered the smallest unit of the organ? **Label:** tissues

Table 3. An example of a true label vs. a relevant but noisy label. A relevant label is related to the question but is not necessarily true. Therefore, relevant labels can be considered noisy labels.

J.5. Reconciling the behaviors of Figures 3a and 4a

To verify that the non-monotonic behavior in Figure 4a is a direct consequence of the noise model (the fixed-fraction label corruption following Gao et al. (2024)), we re-run that experiment at three noisy ratios $\{0.01, 0.3, 0.6\}$. At noisy ratio 0.01, exact-match performance *increases* monotonically with context length, mirroring the monotonic behaviour of Figure 3a; as the noisy ratio grows, adding more demonstrations introduces proportionally more corrupted signal, eventually reversing the trend and recovering Figure 4a at noisy ratio 0.6. Across all three regimes, the optimal attention temperature improves over the default, confirming that the moment-ratio heuristic transfers across the full corruption spectrum.



(a) Low corruption (noisy ratio 0.01) *(b)* Moderate corruption (noisy ratio 0.3) *(c)* High corruption (noisy ratio 0.6)

Figure 6. Effect of context length on Llama2-7B (SCIQ) across three corruption levels; the setup matches Figure 4a except for the noisy ratio. The dashed black line marks the moment-ratio heuristic from Section 4.2.

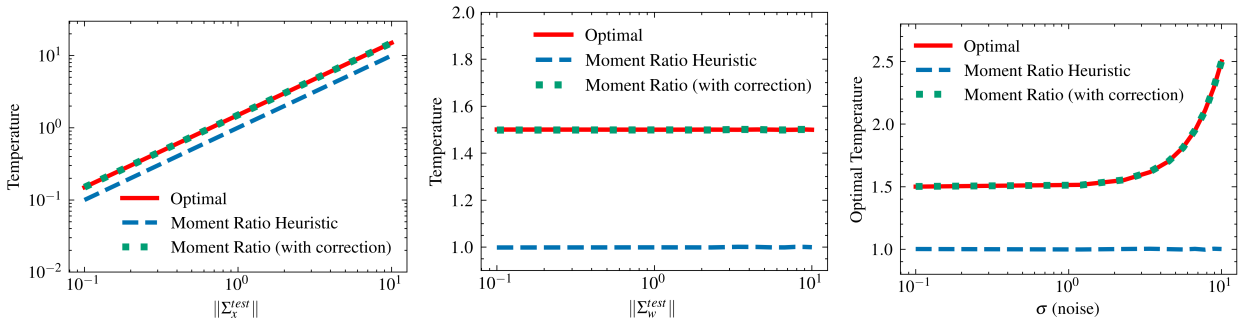
K. Numerical illustration of the optimal temperature and its generalization behavior in the setting of Section 4.4

To complement the analytical reductions in the setting of Section 4.4, we now present numerical experiments illustrating how the optimal attention temperature varies under different types of distribution shift. These simulations closely follow the structure predicted by the closed-form expression in Theorem 4.3 and its simplified forms in (22) and (25).

Figure 7 shows the optimal temperature as a function of (i) input covariance shift, (ii) task covariance shift, and (iii) noise-variance shift. In each subplot, a single shift parameter (a, b, or σ) is varied while the others remain fixed. The closed-form optimal temperatures (16) align closely with the moment-ratio heuristic with correction (25). As anticipated:

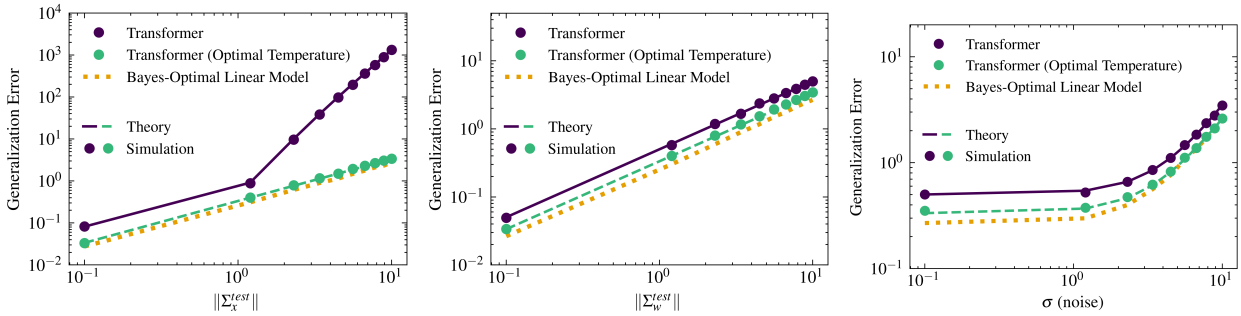
- higher input variance a or noise level σ increases the optimal temperature, while
- task variance b does not significantly change the optimal temperature.

Figure 8 presents the corresponding generalization errors. These results demonstrate that the closed-form characterization accurately captures the key dependencies of the optimal temperature under a range of distribution shifts.



(a) Input Covariance Shifted ($\mathbf{\Sigma}_x^{test} = a\mathbf{\Sigma}_x^{train}$) *(b)* Task Covariance Shifted ($\mathbf{\Sigma}_w^{test} = b\mathbf{\Sigma}_w^{train}$) *(c)* Noise Shifted ($\sigma^{test} = \sigma$)

Figure 7. Optimal temperature under different types of distribution shift. The moment-ratio heuristic (Appendix I) is derived from the closed-form optimal temperature, with its corrected form given in (25). During training, we use $m = 5000$ tasks, noise level $\sigma^{train} = 0.1$, means $\boldsymbol{\mu}_x^{train} = \boldsymbol{\mu}_w^{train} = \mathbf{0}$, and covariances $\mathbf{\Sigma}_x^{train} = \mathbf{\Sigma}_w^{train} = \boldsymbol{I}$. At test time, we set $\boldsymbol{\mu}_x^{test} = \boldsymbol{\mu}_w^{test} = \mathbf{0}$, $\mathbf{\Sigma}_x^{test} = aI$, $\mathbf{\Sigma}_w^{test} = bI$, and $\sigma^{test} = \sigma$. In each subplot, exactly one of a, b, or σ is varied, while the other two remain fixed at their training-distribution values to isolate the effect of a single shift dimension. The dimension and context length are set to $d = 50$ and $l = 2d$.



(a) Input Covariance Shifted ($\mathbf{\Sigma}_x^{test} = a\mathbf{\Sigma}_x^{train}$) *(b)* Task Covariance Shifted ($\mathbf{\Sigma}_w^{test} = b\mathbf{\Sigma}_w^{train}$) *(c)* Noise Shifted ($\sigma^{test} = \sigma$)

Figure 8. Generalization errors corresponding to Figure 7.