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

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

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 on pretrained LLMs (GPT-2 and Llama2-7B) 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 [28] underpin contemporary AI systems, and a key driver of their success is *in-context learning* (ICL): adaptation to novel tasks directly from prompts, without gradient updates [4]. ICL has motivated extensive efforts to understand its mechanisms and scaling behavior [2, 30, 31, 34]. Yet ICL is fragile: even mild distributional shifts between pretraining and downstream tasks can sharply degrade performance [36], raising practical concerns about robustness and adaptability.

Within softmax self-attention, the *attention temperature*[1] $\tau > 0$ rescales the pre-softmax scores, controlling their variance and hence the selectivity of attention. Although the original Transformer fixes $\tau = \sqrt{d_k}$ (key dimension) [28], empirical work shows that adjusting τ can substantially improve performance across NLP and vision benchmarks [5, 13, 16, 21, 37, 38]. To the best of our knowledge, however, its role in ICL under distributional shift remains largely unexplored[2].

This work — We give a unified theoretical and empirical study of attention temperature for ICL under distribution shift, in the linear-regression framework that has been productive for analyzing ICL [9, 36]. Departing from prior work on linear attention, we analyze a Transformer with *ap-*

1. We distinguish this attention temperature from the LLM *sampling temperature* [25] that is used to adjust the output distribution of generative models.
2. Related work is deferred to Appendix A.

proximate softmax attention — which preserves softmax's temperature-dependent selectivity while remaining tractable — and derive a closed-form *optimal temperature* τ_{opt} that minimizes the ICL generalization error. The expression depends explicitly on the type of shift, and setting the temperature $\tau = \tau_{\text{opt}}$ recovers near-Bayes-optimal performance in our experiments. We also derive a moment-ratio heuristic that transfers the insight to standard softmax attention, which we evaluate on Llama2-7B.

Contributions —

1. We derive, to our knowledge, the first closed-form expression for the *optimal attention temperature* τ_{opt} that minimizes the ICL generalization error of a pretrained approximate-softmax Transformer (Theorems 2–3).
2. We characterize the effect of input-, task-, and noise-distribution shifts on ICL: the model is invariant to input-mean shifts but sensitive to input-covariance, task, and noise shifts, all of which τ_{opt} partially mitigates.
3. We distill the analysis into a *moment-ratio heuristic* (9) that extends the optimal-temperature insight to standard softmax attention, and empirically validate it with Llama2-7B on SCIQ.

Notation — We follow Goodfellow et al. [10] for the notation. The spectral norm of a matrix $\boldsymbol{M}$ is $\|\boldsymbol{M}\|$ and the trace is $\mathrm{Tr}(\boldsymbol{M})$. Matrix entries and slices are denoted by $M_{i,j}$, $\boldsymbol{M}_{:,j}$, $\boldsymbol{M}_{i,:}$.

2. Setting

We study ICL on linear-regression tasks: from context $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l-1}$ and query $\boldsymbol{x}_l \in \mathbb{R}^d$, the model predicts y_l, with each pair drawn i.i.d. as

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x),\ y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

with task vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ fixed within a context and varying across tasks. We impose two assumptions; the first is weaker than the corresponding setup of Zhang et al. [36].

Assumption 1 (Well-behaved data distributions) *There exist $c_1, c_2, c_3 > 0$ such that $\|\boldsymbol{\mu}_x\|, \|\boldsymbol{\mu}_w\| \le c_1$, $\lambda_{\min}(\boldsymbol{\Sigma}_x), \lambda_{\min}(\boldsymbol{\Sigma}_w) \ge c_2$, and $\lambda_{\max}(\boldsymbol{\Sigma}_x), \lambda_{\max}(\boldsymbol{\Sigma}_w) \le c_3$.*

Assumption 2 (High-dimensional regime) *The context length l and input dimension d diverge jointly: $l, d \to \infty$.*

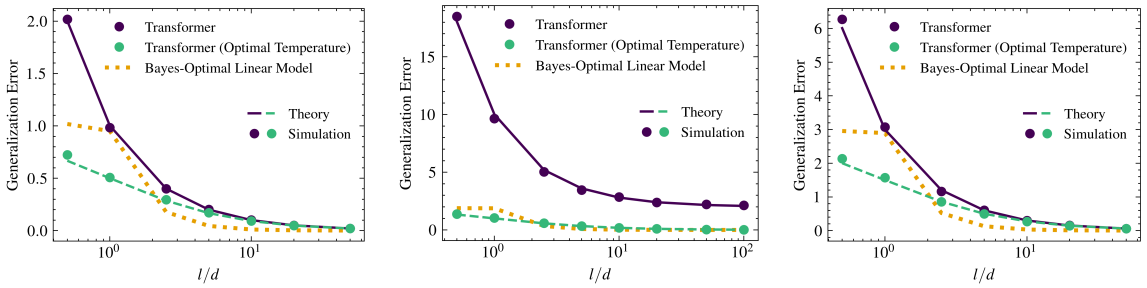
Definition 1 (In-context learning) *A model succeeds at ICL if its generalization error nearly matches that of the Bayes-optimal linear model (Appendix B).*

Approximate softmax attention — Following Zhang et al. [36], we encode the context as the embedding matrix

$$\boldsymbol{Z} := \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{l-1} & \boldsymbol{x}_l \\ y_1 & \cdots & y_{l-1} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times l}, \tag{2}$$

whose last column is the query. We then analyze the approximate-softmax attention

$$\boldsymbol{E} := \boldsymbol{Z} + \boldsymbol{V}\boldsymbol{Z} \cdot \widehat{\text{softmax}}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T (\boldsymbol{Q}\boldsymbol{Z})}{\tau} \right), \tag{3}$$



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

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

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

Figure 1: Experiments with Transformer (3) on ICL under distribution shifts. Parameters are set using (52) while the optimal temperature is by Theorem 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 $\boldsymbol{\Sigma}_x^{train} = \boldsymbol{\Sigma}_w^{train} = \boldsymbol{I}$.

where $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V}$ are key/query/value matrices, τ is the temperature, and the explicit $\widehat{\text{softmax}}$ form is derived in Appendix C. Unlike linear attention [36], $\widehat{\text{softmax}}$ preserves row-wise normalization (so the model is robust to mean shifts; Remark 4) and mirrors softmax's temperature behavior (Figure 3; Appendix D).

Reparametrization and generalization error — With $\boldsymbol{M} := \boldsymbol{K}^\top \boldsymbol{Q}$, the prediction $\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) := E_{d+1,l}$ depends on $(\boldsymbol{V}, \boldsymbol{M})$ only through their lower blocks:

$$\boldsymbol{V} = \begin{bmatrix} * & * \\ \boldsymbol{v}_{21}^T & v_{22} \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & * \\ \boldsymbol{m}_{21}^T & * \end{bmatrix}, \tag{4}$$

where the entries marked $*$ are irrelevant for predicting y_l. The ICL (generalization) error is

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) := \mathbb{E}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{\text{test}}} \left[(y_l - \hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}))^2 \right], \tag{5}$$

evaluated on a test distribution $\mathcal{D}^{\text{test}}$ whose task vectors were not seen during training.

3. Theoretical results

We characterize the ICL generalization error of the Transformer model (3) and identify the temperature that minimizes it. Throughout, Assumptions 1–2 are in force, together with a boundedness condition on the pretrained $(\boldsymbol{M}, \boldsymbol{V})$ (Assumption 3, Appendix G).

Theorem 2 (Generalization error for ICL) *At test time, let the input, task, and noise distributions be $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, and $\mathcal{N}(0, \sigma^2)$; let $\boldsymbol{A} := \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x^\top$ and $\boldsymbol{B} := \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w \boldsymbol{\mu}_w^\top$ denote the corresponding second-moment matrices. Then*

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \frac{1}{\tau^2} \operatorname{Tr}\big(\boldsymbol{A} \boldsymbol{M}_{11}^T \boldsymbol{F}_1 \boldsymbol{M}_{11}\big) - \frac{1}{\tau} \operatorname{Tr}\big(\boldsymbol{A}\big(\boldsymbol{F}_2 \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T \boldsymbol{F}_2^T\big)\big) + \operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}) + \sigma^2, \tag{6}$$

where $\boldsymbol{F}_1, \boldsymbol{F}_2$ depend on $(\boldsymbol{V}, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_w, \sigma^2)$; their explicit forms and the proof (via Isserlis' theorem [12]) are in Appendix G.

The temperature enters $\mathcal{G}$ as $1/\tau^2$ and $1/\tau$, so a single scalar adjustment can rebalance both terms. Minimizing $\mathcal{G}$ over τ yields:

Theorem 3 (Optimal attention temperature) *The generalization error is minimized at*

$$\tau_{opt} = \frac{2\,\mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\big)}{\mathrm{Tr}\big(\boldsymbol{A}\big(\boldsymbol{F}_2\boldsymbol{M}_{11}+\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\big)\big)}, \tag{7}$$

provided both traces are positive (Appendix I).

When $\tau_{\text{opt}} \neq 1$, an unadjusted temperature is suboptimal. Without distribution shift, the Transformer with the pretrained parameters (Proposition 5 in Appendix F) emulates the Bayes-optimal linear model at $\tau_{\text{opt}} = 1$.

Effect of distribution shift — Under distribution shift ($\mathcal{D}^{\text{test}} \neq \mathcal{D}^{\text{train}}$), the pretrained $(\boldsymbol{M}, \boldsymbol{V})$ no longer match the test-time data, and three cases arise (Appendix I). *(i) Input shift*: centering renders the approximate model invariant to mean shifts, but a covariance mismatch ($\boldsymbol{\Sigma}_x^{\text{train}} \neq \boldsymbol{\Sigma}_x^{\text{test}}$) drives the resulting estimator away from Bayes-optimality at every l, echoing prior results on linear attention [36]. *(ii) Task shift*: $\boldsymbol{\mu}_w^{\text{train}}$ and $\boldsymbol{\Sigma}_w^{\text{train}}$ are encoded in $\boldsymbol{M}_{11}, \boldsymbol{v}_{21}$ via Proposition 5, but the model's dependence on the task distribution decays with l, so task-shift error is most pronounced at small context lengths. *(iii) Noise shift*: a mismatch $\sigma^2_{\text{train}} \neq \sigma^2_{\text{test}}$ similarly moves the resulting estimator away from Bayes-optimality at small l, with the effect vanishing as $l \to \infty$. In the three sensitive cases (input-covariance, task, and noise shifts), applying τ_{opt} from (7) minimizes the resulting generalization error and empirically recovers near-Bayes-optimal performance (Figure 1).

Closed form under isotropic shift — Consider isotropic training $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\epsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2)$ and test $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, a\boldsymbol{I})$, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, b\boldsymbol{I})$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with shift parameters $a, b, \sigma > 0$. Then (7) reduces to

$$\tau_{\text{opt}} = \left(a + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\right)\frac{\mathrm{Tr}(\boldsymbol{M}_{11}^T\boldsymbol{M}_{11})}{\mathrm{Tr}(\boldsymbol{M}_{11})}. \tag{8}$$

In particular, $\tau_{\text{opt}} \to a$ as $l \to \infty$, since $\mathrm{Tr}(\boldsymbol{M}_{11}^T\boldsymbol{M}_{11})/\,\mathrm{Tr}(\boldsymbol{M}_{11}) \approx 1$ under the training distribution.

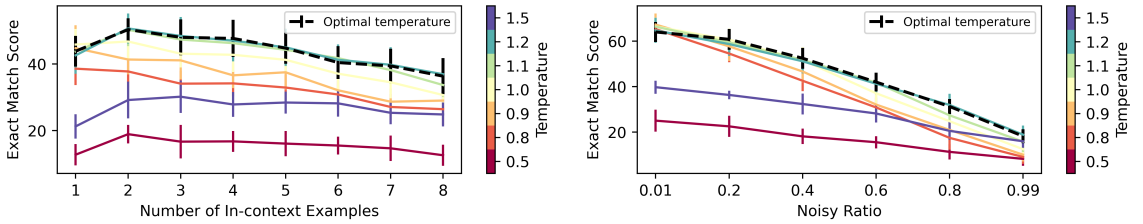
Moment-ratio heuristic — Rewriting (8) in terms of the moments of pre-softmax scores $\boldsymbol{z}_i^\top \boldsymbol{M}\boldsymbol{z}_j$ (Appendix J) gives, for $i \neq j$,

$$\tau_{\text{opt}} \approx \underbrace{\frac{\mathbb{E}[(\boldsymbol{z}_i^\top\boldsymbol{M}\boldsymbol{z}_j)^2]}{\mathbb{E}[\boldsymbol{z}_i^\top\boldsymbol{M}\boldsymbol{z}_i]}}_{\text{moment-ratio}} + \underbrace{\frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)}_{\text{small-}l\text{ correction}}. \tag{9}$$

The moment-ratio is computable for any softmax attention layer, providing a recipe to transfer the optimal-temperature insight beyond approximate softmax — in particular, to large language models (cf. Figure 2).

4. Experimental results

We empirically validate our theory on (i) linear-regression ICL with the approximate-attention Transformer (3) and GPT-2 [23] (the latter in Appendix K), and (ii) Llama2-7B [27] on SCIQ [32].



(*a*) Effect of context length (*b*) Effect of noisy ratio

Figure 2: Attention-temperature effect on Llama2-7B [27] ICL on SCIQ [32] under noisy-label distribution shift [8]. Panel (a) fixes the noisy ratio at 0.6; panel (b) fixes the number of in-context examples at 6. The dashed black line marks the moment-ratio temperature (Eq. (9) and Appendix J) derived from Theorem 3. Error bars show one standard deviation over 12 Monte Carlo runs. The experimental details are deferred to Appendix K.

4.1. Experiments on linear regression tasks

Figure 1 compares the approximate-attention Transformer to the Bayes-optimal linear model on the linear-regression setup (1). The pretrained model converges to Bayes-optimal as l grows (panel a) and recovers Bayes-optimal performance under input-covariance shift once τ_{opt} is applied (panel b); task-shift error decays with l (panel c). Robustness to label-noise shift is reported in Appendix I (Figure 4), where temperature adjustment is critical at small l and τ_{opt} increases with noise level. GPT-2 results (Appendix K) corroborate that the optimal-temperature insight transfers from approximate softmax to a full multi-head softmax Transformer with MLP layers.

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

Since Theorem 3's τ_{opt} is specific to approximate softmax attention, we use the moment-ratio heuristic (9) as a transferable proxy for standard softmax. Following Gao et al. [8], we generate SCIQ ICL tasks with distribution shift induced by noisy labels and evaluate Llama2-7B via exact-match score (Appendix K). Figure 2 shows that the moment-ratio temperature consistently improves ICL across context lengths and noise ratios; higher noise ratios push the optimal temperature upward, matching our theoretical prediction.

5. 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.

References

[1] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, 2023.

[2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, 2023.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

[5] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[6] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[7] Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, 2024.

[8] Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the noise robustness of in-context learning for text generation. In *Advances in Neural Information Processing Systems*, 2024.

[9] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 2022.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[11] Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In *Advances in Neural Information Processing Systems*, 2024.

[12] L. Isserlis. 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.

[13] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[14] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, 2023.

[15] Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *Advances in Neural Information Processing Systems*, 2024.

[16] Junyang Lin, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su. 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*, pages 2985–2990, 2018.

[17] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 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*, pages 100–114, 2022.

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

[19] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[20] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

[21] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

[22] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[24] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 2023.

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

[26] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[29] Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). *International Conference on Machine Learning*, 2025.

[30] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

[31] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[32] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, 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*, pages 38–45, 2020.

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

[35] Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*, 2023.

[36] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[37] Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141, 2022.

[38] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. In *Advances in Neural Information Processing Systems*, 2024.

Appendix A. Related work

Theory of in-context learning — Simplified Transformer variants—particularly those using linear attention—have proven useful for gaining analytical insights about ICL [9, 24, 36]. Notably, Zhang et al. [36] 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. [36], 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 [6, 22]. A pivotal advance in this direction is due to Han et al. [11], who showed that an *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 principled framework for investigating how attention-temperature selection shapes ICL performance in pretrained Transformers.

Attention temperature — Research on attention temperature remains limited. Veličković et al. [29] recently proposed an adaptive temperature scheme to sharpen softmax outputs, and several empirical studies in natural language processing and computer vision [5, 13, 16, 21, 37, 38] 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.

ICL by Transformers — The ICL capability of Transformers was first brought to prominence by [4], leading to a surge of empirical and theoretical investigations. Several works have demonstrated that ICL performance improves with model scale [19, 26, 31], 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 [9, 24, 36]. A prevailing hypothesis in recent theoretical work is that Transformers implicitly learn algorithms during pretraining, which they subsequently execute during inference [1–3, 7, 14, 15, 18, 20, 30, 36]. 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.

Appendix B. 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

$$p(\bar{\boldsymbol{y}} \mid \bar{\boldsymbol{X}}, \boldsymbol{w}) = \prod_{i=1}^{l-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\bar{y}_i - \boldsymbol{w}^T \bar{\boldsymbol{x}}_i)^2}{2\sigma^2}\right] \tag{10}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})^T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})\right], \tag{11}$$

where $\propto$ denotes proportionality.

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

$$p(\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}) \propto p(\bar{\boldsymbol{y}} \mid \bar{\boldsymbol{X}}, \boldsymbol{w})\, p(\boldsymbol{w}). \tag{12}$$

Substituting the Gaussian prior yields

$$p(\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}) \propto \exp\Big[-\frac{1}{2\sigma^2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})^T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\boldsymbol{w})\Big] \exp\Big[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\Big]. \tag{13}$$

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:

$$-\frac{1}{2\sigma^2}\left(\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}} - 2\bar{\boldsymbol{y}}^T\bar{\boldsymbol{X}}\boldsymbol{w} + \boldsymbol{w}^T\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\boldsymbol{w}\right) - \frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{w} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right). \tag{14}$$

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

$$-\frac{1}{2}\boldsymbol{w}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{w} + \boldsymbol{w}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \text{terms independent of } \boldsymbol{w}. \tag{15}$$

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

$$-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}_l^{-1}\boldsymbol{w} + \boldsymbol{w}^T b_l = -\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_l) + \text{const}, \tag{16}$$

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

$$\boldsymbol{\mu}_l = \left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right). \tag{17}$$

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

$$\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{18}$$

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:

$$\hat{\boldsymbol{w}}_{\text{Ridge}} = \mathbb{E}[\boldsymbol{w} \mid \bar{\boldsymbol{y}}, \bar{\boldsymbol{X}}] = \boldsymbol{\mu}_l = \Big(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\Big)^{-1}\Big(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\Big). \tag{19}$$

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.

Appendix C. Derivation of approximate softmax

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

$$\text{softmax}(\boldsymbol{z})_i := \frac{e^{z_i}}{\sum_{j=1}^{l} e^{z_j}} \quad \forall i \in \{1, \dots, l\}. \tag{20}$$

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

$$\text{softmax}(\boldsymbol{z}) \approx \text{softmax}(\mathbf{0}) + J_{\text{softmax}}(\mathbf{0})\boldsymbol{z}, \tag{21}$$

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:

$$\text{softmax}(\mathbf{0}) = \frac{e^0}{\sum_{j=1}^{l} e^0}\mathbf{1} = \frac{1}{l}\mathbf{1}. \tag{22}$$

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

$$J_{\text{softmax}}(\mathbf{0})_{ii} = \text{softmax}(\mathbf{0})_i\,(1 - \text{softmax}(\mathbf{0})_i) = \frac{l-1}{l^2}, \quad \forall i, \tag{23}$$

$$J_{\text{softmax}}(\mathbf{0})_{ij} = -\text{softmax}(\mathbf{0})_i \cdot \text{softmax}(\mathbf{0})_j = -\frac{1}{l^2}, \quad \forall i \neq j. \tag{24}$$

This yields the compact matrix form:

$$J_{\text{softmax}}(\mathbf{0}) = \frac{1}{l}\boldsymbol{I} - \frac{1}{l^2}\mathbf{1}\mathbf{1}^T. \tag{25}$$

Substituting back, we obtain the approximate softmax:

$$\text{softmax}(\boldsymbol{z}) \approx \frac{1}{l}\mathbf{1} + \left(\frac{1}{l}\boldsymbol{I} - \frac{1}{l^2}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{z}, \tag{26}$$

$$= \left(\frac{1}{l} - \frac{1}{l^2}\sum_{j=1}^{l} z_j\right)\mathbf{1} + \frac{1}{l}\boldsymbol{z}, \tag{27}$$

$$=: \widehat{\text{softmax}}(\boldsymbol{z}). \tag{28}$$

This derivation yields the approximate softmax attention formulation in (3). From a practical standpoint, approximate softmax attention mechanisms have been empirically evaluated and shown to achieve performance comparable to standard softmax attention [11].

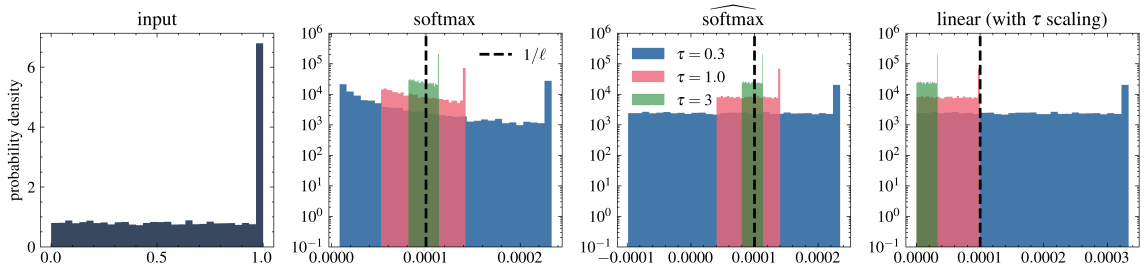


Figure 3: Comparison of temperature effects in softmax, approximate softmax, and linear (with temperature scaling) cases. We consider an input vector $\boldsymbol{x} \in \mathbb{R}^l$ whose histogram is illustrated on the left-most plot. Rest of the plots illustrates histograms of the elements of $\text{softmax}(\boldsymbol{x}/\tau)$, approximation $\widehat{\text{softmax}}(\boldsymbol{x}/\tau)$ derived in Appendix C and $\boldsymbol{x}/(l\tau)$ from left to right, respectively.

Appendix D. Temperature effects for softmax and approximate softmax

In contrast to linear attention [36], $\boldsymbol{Z} + \frac{1}{l}\boldsymbol{V}\boldsymbol{Z}(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})$, the approximate softmax formulation in (3) explicitly preserves normalization, which is essential for both interpretability and robustness. This difference is described in the following remark.

Remark 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.*

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 3 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.

Appendix E. Expanded form of approximate softmax attention

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

$$\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) = A_{d+1,l}, \tag{29}$$

$$= \frac{1}{l}\boldsymbol{V}_{d+1,:}\boldsymbol{Z}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z}_{:,l})}{\tau} - \frac{\mathbf{1}}{l}\sum_{j=1}^{l}\frac{(\boldsymbol{K}\boldsymbol{Z}_{:,j})^T(\boldsymbol{Q}\boldsymbol{Z}_{:,l})}{\tau} + \mathbf{1}\right), \tag{30}$$

$$= \frac{1}{l}[\boldsymbol{v}_{21}^T \;\; v_{22}]\boldsymbol{Z}\left(\frac{\boldsymbol{Z}^T\boldsymbol{M}\boldsymbol{Z}_{:,l}}{\tau} - \frac{\mathbf{1}}{l}\sum_{j=1}^{l}\frac{(\boldsymbol{Z}_{:,j})^T\boldsymbol{M}\boldsymbol{Z}_{:,l}}{\tau} + \mathbf{1}\right), \tag{31}$$

$$= \frac{1}{l}[\boldsymbol{v}_{21}^T \;\; v_{22}][\boldsymbol{X} \;\; \boldsymbol{y}]^T\left(\frac{[\boldsymbol{X} \;\; \boldsymbol{y}]\boldsymbol{M}[\boldsymbol{x}_l^T \;\; 0]^T}{\tau} - \frac{\mathbf{1}}{l}\sum_{j=1}^{l}\frac{[\boldsymbol{x}_i^T \;\; y_i]\boldsymbol{M}[\boldsymbol{x}_l^T \;\; 0]^T}{\tau} + \mathbf{1}\right), \tag{32}$$

$$= \frac{1}{l}[\boldsymbol{v}_{21}^T \;\; v_{22}][\boldsymbol{X} \;\; \boldsymbol{y}]^T\left(\frac{1}{\tau}\left[\boldsymbol{X} - \mathbf{1}\boldsymbol{s}_x^T \;\; \boldsymbol{y} - s_y\mathbf{1}\right]\begin{bmatrix}\boldsymbol{M}_{11} & * \\ \boldsymbol{m}_{21}^T & *\end{bmatrix}[\boldsymbol{x}_l^T \;\; 0]^T + \mathbf{1}\right), \tag{33}$$

$$= \frac{1}{l}[\boldsymbol{v}_{21}^T \;\; v_{22}][\boldsymbol{X} \;\; \boldsymbol{y}]^T\left(\frac{1}{\tau}\left(\boldsymbol{X} - \mathbf{1}\boldsymbol{s}_x^T\right)\boldsymbol{M}_{11}\boldsymbol{x}_l + \frac{1}{\tau}\left(\boldsymbol{y} - s_y\mathbf{1}\right)\boldsymbol{m}_{21}^T\boldsymbol{x}_l + \mathbf{1}\right), \tag{34}$$

$$= \frac{1}{l}\left(\boldsymbol{v}_{21}^T\boldsymbol{X}^T + v_{22}\boldsymbol{y}^T\right)\left(\frac{1}{\tau}\left(\boldsymbol{X} - \mathbf{1}\boldsymbol{s}_x^T\right)\boldsymbol{M}_{11}\boldsymbol{x}_l + \frac{1}{\tau}\left(\boldsymbol{y} - s_y\mathbf{1}\right)\boldsymbol{m}_{21}^T\boldsymbol{x}_l + \mathbf{1}\right), \tag{35}$$

$$\begin{aligned}= \frac{1}{\tau}&\left(\boldsymbol{v}_{21}^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{l} - \boldsymbol{s}_x\boldsymbol{s}_x^T\right) + v_{22}\left(\frac{\boldsymbol{y}^T\boldsymbol{X}}{l} - s_y\boldsymbol{s}_x^T\right)\right)\boldsymbol{M}_{11}\boldsymbol{x}_l, \\ &+ \frac{1}{\tau}\left(\boldsymbol{v}_{21}^T\left(\frac{\boldsymbol{X}^T\boldsymbol{y}}{l} - s_y\boldsymbol{s}_x\right) + v_{22}\left(\frac{\boldsymbol{y}^T\boldsymbol{y}}{l} - s_y^2\right)\right)\boldsymbol{m}_{21}^T\boldsymbol{x}_l + \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y,\end{aligned} \tag{36}$$

$$= \frac{1}{\tau}\left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xx} + v_{22}\boldsymbol{C}_{xy}^T\right)\boldsymbol{M}_{11}\boldsymbol{x}_l + \frac{1}{\tau}\left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}^T\boldsymbol{x}_l + \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y, \tag{37}$$

$$= \frac{1}{\tau}\left(\left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xx} + v_{22}\boldsymbol{C}_{xy}^T\right)\boldsymbol{M}_{11} + \left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}^T\right)\boldsymbol{x}_l + \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y, \tag{38}$$

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

$$\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V}) = \boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right) + \left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}, \tag{39}$$

$$b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) = \boldsymbol{v}_{21}^T\boldsymbol{s}_x + v_{22}s_y, \tag{40}$$

which allows us to write

$$\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) = \frac{1}{\tau}\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V})^T\boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}). \tag{41}$$

Appendix F. 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 (Appendix B). The prediction of the approximate softmax attention model can be written as

$$\hat{y}(\boldsymbol{Z};\boldsymbol{V},\boldsymbol{M}) = \frac{1}{\tau}\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V})^T\boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}), \tag{42}$$

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

$$\hat{y}_{Bayes} = \hat{\boldsymbol{w}}_{Bayes}^T\boldsymbol{x}_l. \tag{43}$$

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

$$\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V}) \approx \hat{\boldsymbol{w}}_{Bayes}, \quad b_{Att}(\boldsymbol{s}_x,s_y;\boldsymbol{V}) \approx 0, \tag{44}$$

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

$$\begin{aligned}
&\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx},\boldsymbol{C}_{xy},C_{yy};\boldsymbol{M},\boldsymbol{V}) \\
&\quad= \left(\boldsymbol{M}_{11}^T\left(\boldsymbol{C}_{xx}\boldsymbol{v}_{21} + v_{22}\boldsymbol{C}_{xy}\right) + \left(\boldsymbol{v}_{21}^T\boldsymbol{C}_{xy} + v_{22}C_{yy}\right)\boldsymbol{m}_{21}\right), &(45)\\
&\quad= \left(\boldsymbol{M}_{11}^T\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{l}\boldsymbol{v}_{21} + v_{22}\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{l}\right) + \left(\boldsymbol{v}_{21}^T\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{l} + v_{22}\frac{\bar{\boldsymbol{y}}^T\bar{\boldsymbol{y}}}{l}\right)\boldsymbol{m}_{21}\right). &(46)
\end{aligned}$$

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

$$\hat{\boldsymbol{w}}_{Bayes} = \left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w\right), \tag{47}$$

as derived in Appendix B. Looking at equations (47) and (46) together, we can see that setting the parameters as follows would make $\hat{\boldsymbol{w}}_{Att} = \hat{\boldsymbol{w}}_{Bayes}$ hold

$$\boldsymbol{M}_{11} = \frac{l}{\sigma^2}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{l}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{l}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = 1. \tag{48}$$

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 (48) 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 (pre)training data consisting of ml samples.

$$\boldsymbol{M}_{11} = \frac{l}{\sigma^2}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{m\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{l}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = 1. \tag{49}$$

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

$$b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) = \boldsymbol{v}_{21}^T \boldsymbol{s}_x + v_{22} s_y, \tag{50}$$

where $\boldsymbol{s}_x$ and s_y are based on data so we have no control over them. Instead, by using Assumptions 1 and 2, we can choose $\boldsymbol{v}_{21}$ and v_{22} such that $b_{Att} \to 0$ as $l, d \to \infty$. Note that Assumption 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 (49). Therefore, multiplying $\boldsymbol{v}_{21}, v_{22}$ given in (49) 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

$$\boldsymbol{M}_{11} = \frac{dl}{\sigma^2}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{m\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \quad \boldsymbol{v}_{21} = \frac{\sigma^2}{dl}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \quad \boldsymbol{m}_{21} = \boldsymbol{0}, \quad v_{22} = \frac{1}{d}. \tag{51}$$

We summarize the resulting pretrained-parameters configuration as the following proposition.

Proposition 5 (Pretrained parameters) *When the temperature parameter is set to $\tau = 1$ during pretraining, the following parameter configuration approximates the Bayes-optimal estimator (Appendix B):*

$$\boldsymbol{M}_{11} = d\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml} + \frac{\sigma^2}{l}\boldsymbol{\Sigma}_w^{-1}\right)^{-1}, \qquad \boldsymbol{m}_{21} = \boldsymbol{0}, \tag{52}$$

$$\boldsymbol{v}_{21} = \frac{\sigma^2}{dl}\left(\frac{\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}}{ml}\right)^{-1}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w, \qquad v_{22} = \frac{1}{d},$$

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 $\boldsymbol{\Sigma}_w$ can be estimated from the pretraining data.

Remark 6 *While the pretrained parameters specified in Proposition 5 are not guaranteed to be optimal in all settings, they are analytically useful for examining the effects of distribution shifts.*

Based on Proposition 5, we arrive at the following:

Corollary 7 *Suppose there is no distribution shift between training and inference. Then, under the parameter configuration of Proposition 5, the Transformer model (3) emulates the Bayes-optimal linear model, implying that it is capable of in-context learning according to Definition 1.*

Appendix G. Characterization of generalization error for ICL under distribution shift

We first state the boundedness assumption on the pretrained parameters used throughout Theorems 2 and 3, and give the explicit forms of the matrices appearing in those statements.

Assumption 3 *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 matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F}_1, \boldsymbol{F}_2, \hat{\boldsymbol{B}}$ appearing in Theorem 2 and Theorem 3 are defined as

$$\boldsymbol{A} := \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x^T, \qquad \boldsymbol{B} := \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T, \tag{53}$$

$$\boldsymbol{F}_1 := \left(\boldsymbol{\Sigma}_x \hat{\boldsymbol{B}} + \frac{1}{l} \left(v_{22}^2 \sigma^2 + \mathrm{Tr}(\hat{\boldsymbol{B}} \boldsymbol{\Sigma}_x) \right) \boldsymbol{I} \right) \boldsymbol{\Sigma}_x, \tag{54}$$

$$\boldsymbol{F}_2 := (\boldsymbol{\mu}_w \boldsymbol{v}_{21}^T + v_{22} \boldsymbol{B}) \boldsymbol{\Sigma}_x, \tag{55}$$

$$\hat{\boldsymbol{B}} := v_{22} \boldsymbol{\mu}_w \boldsymbol{v}_{21}^T + v_{22} \boldsymbol{v}_{21} \boldsymbol{\mu}_w^T + v_{22}^2 \boldsymbol{B}. \tag{56}$$

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 (5) as follows

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) := \mathbb{E}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{test}} \left[(y_l - \hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}))^2 \right], \tag{57}$$

$$= \mathbb{E}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} \hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V})^T \boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) - y_l \right)^2 \right], \tag{58}$$

$$= \mathbb{E}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} \left(\boldsymbol{M}_{11}^T \left(\boldsymbol{C}_{xx} \boldsymbol{v}_{21} + v_{22} \boldsymbol{C}_{xy} \right) \right)^T \boldsymbol{x}_l - y_l \right)^2 \right], \tag{59}$$

where we use the parameters from pretraining (51) together with Assumptions 1 and 2 to reach the last line. Then,

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \mathbb{E}_{(\boldsymbol{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} \left(\boldsymbol{M}_{11}^T \left(\boldsymbol{C}_{xx} \boldsymbol{v}_{21} + v_{22} \boldsymbol{C}_{xy} \right) \right)^T \boldsymbol{x}_l - y_l \right)^2 \right], \tag{60}$$

$$= \mathbb{E} \left[\left(\frac{1}{\tau} \left(\boldsymbol{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^T \boldsymbol{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\boldsymbol{x}}_i (\bar{\boldsymbol{x}}_i^T \boldsymbol{w} + \epsilon_i) \right) \right)^T \boldsymbol{x}_l - \boldsymbol{w}^T \boldsymbol{x}_l - \epsilon_l \right)^2 \right], \tag{61}$$

$$= \mathbb{E} \left[\left(\frac{1}{\tau} \left(\boldsymbol{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^T \boldsymbol{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\boldsymbol{x}}_i (\bar{\boldsymbol{x}}_i^T \boldsymbol{w} + \epsilon_i) \right) \right)^T \boldsymbol{x}_l - \boldsymbol{w}^T \boldsymbol{x}_l \right)^2 \right] + \sigma^2 \tag{62}$$

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

$$\boldsymbol{w}_{diff} := \frac{1}{\tau} \boldsymbol{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^T \boldsymbol{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\boldsymbol{x}}_i (\bar{\boldsymbol{x}}_i^T \boldsymbol{w} + \epsilon_i) \right) - \boldsymbol{w}, \tag{63}$$

which allows us to write

$$\mathcal{G}(\boldsymbol{V},\boldsymbol{M}) = \mathbb{E}\left[\left(\boldsymbol{w}_{diff}^T\boldsymbol{x}_l\right)^2\right] + \sigma^2, \tag{64}$$

$$= \mathbb{E}\left[\boldsymbol{w}_{diff}^T\mathbb{E}_{\boldsymbol{x}_l}[\boldsymbol{x}_l\boldsymbol{x}_l^T]\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{65}$$

$$= \mathbb{E}\left[\boldsymbol{w}_{diff}^T\left(\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_x\right)\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{66}$$

by the law of total expectation since $\boldsymbol{w}_{diff}$ is independent of $\boldsymbol{x}_l$. Note that when writing (64), we safely ignore terms with $(1/l)\bar{\boldsymbol{x}}_l\bar{\boldsymbol{x}}_l^T\boldsymbol{v}_{21}$ in (62) since they vanish as $l\to\infty$ by Assumptions 1-2 and 3. Letting $\boldsymbol{A} := \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_x$, we write

$$\mathcal{G}(\boldsymbol{V},\boldsymbol{M}) = \mathbb{E}\left[\boldsymbol{w}_{diff}^T\boldsymbol{A}\boldsymbol{w}_{diff}\right] + \sigma^2, \tag{67}$$

$$= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{w}_{diff}^T\boldsymbol{A}\boldsymbol{w}_{diff}\right)\right] + \sigma^2, \tag{68}$$

$$= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T\right)\right] + \sigma^2, \tag{69}$$

$$= \mathrm{Tr}\left(\boldsymbol{A}\mathbb{E}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right) + \sigma^2, \tag{70}$$

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

$$\boldsymbol{w}_{diff} = \underbrace{\frac{1}{\tau}\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right)}_{\boldsymbol{e}} + \underbrace{\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)}_{\boldsymbol{D}}\boldsymbol{w}, \tag{71}$$

$$= \boldsymbol{e} + \boldsymbol{D}\boldsymbol{w}, \tag{72}$$

where we define

$$\boldsymbol{e} := \frac{1}{\tau}\boldsymbol{M}_{11}^T\left(\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right), \tag{73}$$

$$\boldsymbol{D} := \left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\le l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right). \tag{74}$$

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

$$\mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[(\boldsymbol{e}+\boldsymbol{D}\boldsymbol{w})(\boldsymbol{e}+\boldsymbol{D}\boldsymbol{w})^T]\right], \tag{75}$$

$$= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{e}\boldsymbol{\mu}_w^T\boldsymbol{D}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}(\boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_w)\boldsymbol{D}^T\right], \tag{76}$$

$$= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]^T + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right], \tag{77}$$

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 (77).

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

Lemma 8 *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 $i = 1, \dots, l-1$. Furthermore, let $\boldsymbol{A}$ be a fixed $d \times d$ matrix. Then, the following holds*

$$\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{A}\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right] = \frac{l-1}{l}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma} + \frac{1}{l}Tr(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}. \tag{78}$$

Proof *This is proven by using Isserlis' theorem [12] in Appendix H.* ■

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 8 is directly applicable in our setting.

Next, we start the calculations of the expectations in (77) with $\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right]$ as follows

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] &= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21} + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right)\right. \\
&\qquad\left.\cdot\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\boldsymbol{M}_{11}, \qquad (79)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{v}_{21}\right)\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right.\right. \\
&\qquad\left.\left.+\left(v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\epsilon_i\right)\left(v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\right)\boldsymbol{M}_{11}, \qquad (80)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right) + \left(v_{22}^2\frac{\sigma^2}{l^2}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\right)\boldsymbol{M}_{11}, \qquad (81)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{v}_{21}\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x + v_{22}^2\frac{\sigma^2(l-1)}{l^2}\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11}, \qquad (82)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\left(v_{22}^2\frac{\sigma^2}{l}\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11}, \qquad (83)
\end{aligned}$$

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 8 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 1-2 and 3 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

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

$$= \frac{1}{\tau}\mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)\boldsymbol{\mu}_w\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T + v_{22}\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i^T\epsilon_i\right)\right]\boldsymbol{M}_{11}, \tag{84}$$

$$= \frac{1}{\tau}\mathbb{E}\left[\left(\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T - \boldsymbol{I}\right)\boldsymbol{\mu}_w\left(\frac{1}{l}\sum_{i\leq l-1}\boldsymbol{v}_{21}^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11}, \tag{85}$$

$$= \frac{1}{\tau}\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right]\boldsymbol{M}_{11}$$
$$- \boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\mathbb{E}\left[\frac{1}{l}\sum_{i\leq l-1}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right]\boldsymbol{M}_{11}, \tag{86}$$

$$= \frac{1}{\tau}\frac{v_{22}}{\tau}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\boldsymbol{\Sigma}_x\boldsymbol{v}_{21}\boldsymbol{\mu}_w^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11}$$
$$- \frac{1}{\tau}\frac{l-1}{l}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\boldsymbol{M}_{11}, \tag{87}$$

$$= \frac{v_{22}}{\tau^2}\boldsymbol{M}_{11}^T\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\boldsymbol{M}_{11}, \tag{88}$$

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 8 together with the fact that $\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \boldsymbol{\Sigma}_x$. Using Assumptions 1-2 and 3, 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], &(89)\\
&= \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}, &(90)\\
&= \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}, &(91)\\
&= \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}, &(92)\\
&= \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}, &(93)
\end{aligned}$$

where we first do basic algebraic manipulations. Then, we use Lemma 8 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 2.

Putting the found expectation results into (77), we get

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right] &= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right]^T + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{\mu}_w\boldsymbol{e}^T\right] + \mathbb{E}\left[\boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^T\right], &(94)\\
&= \frac{1}{\tau^2}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{F}_2\boldsymbol{M}_{11} + \frac{1}{\tau}\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T + \boldsymbol{B}. &(95)
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{F}_1 &:= v_{22}^2\frac{\sigma^2}{l}\boldsymbol{\Sigma}_x + v_{22}\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right) &(96)\\
&\quad + v_{22}\left(\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right)^T + v_{22}^2\left(\boldsymbol{\Sigma}_x\boldsymbol{B}\boldsymbol{\Sigma}_x + \frac{1}{l}\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}_x\right)\boldsymbol{\Sigma}_x\right), \\
&= \left(\boldsymbol{\Sigma}_x\hat{\boldsymbol{B}} + \left(v_{22}^2\frac{\sigma^2}{l} + \frac{1}{l}\mathrm{Tr}\left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_x\right)\right)\boldsymbol{I}\right)\boldsymbol{\Sigma}_x, &(97)\\
\boldsymbol{F}_2 &:= \boldsymbol{\mu}_w\boldsymbol{v}_{21}^T\boldsymbol{\Sigma}_x + v_{22}\boldsymbol{B}\boldsymbol{\Sigma}_x = (\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{B})\boldsymbol{\Sigma}_x, &(98)
\end{aligned}$$

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 (70), we have

$$\mathcal{G}(\boldsymbol{V},\boldsymbol{M}) = \text{Tr}\left(\boldsymbol{A}\mathbb{E}[\boldsymbol{w}_{diff}\boldsymbol{w}_{diff}^T]\right) + \sigma^2, \tag{99}$$

$$= \text{Tr}\left(\boldsymbol{A}\left(\frac{1}{\tau^2}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11} - \frac{1}{\tau}\boldsymbol{F}_2\boldsymbol{M}_{11} + \frac{1}{\tau}\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T + \boldsymbol{B}\right)\right) + \sigma^2, \tag{100}$$

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

Appendix H. Proof of Lemma 8

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

$$\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\le l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\boldsymbol{A}\left(\frac{1}{l}\sum_{i\le l}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right)\right] = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma} + \frac{1}{l}\text{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}. \tag{101}$$

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

$$\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x = \frac{1}{l^2}\sum_{i=1}^l\sum_{j=1}^l \bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T \tag{102}$$

Taking the expectation, we get

$$\mathbb{E}[\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x] = \frac{1}{l^2}\sum_{i=1}^l\sum_{j=1}^l \mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] \tag{103}$$

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

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] = \mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T]\boldsymbol{A}\mathbb{E}[\bar{\boldsymbol{x}}_j\bar{\boldsymbol{x}}_j^T] = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} \tag{104}$$

When $i = j$,

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \mathbb{E}[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\boldsymbol{A}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T] \tag{105}$$

Let $\bar{\boldsymbol{x}} = [x_1, x_2, \ldots, x_d]^T$. Then, from Isserlis' theorem [12], we have

$$\mathbb{E}[x_ix_jx_kx_l] = \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{kl} + \boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}_{jl} + \boldsymbol{\Sigma}_{il}\boldsymbol{\Sigma}_{jk} \tag{106}$$

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

$$\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\boldsymbol{A}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T = \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\sum_{i,j} a_{ij}x_ix_j, \tag{107}$$

$$\mathbb{E}[\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\boldsymbol{A}\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T] = \text{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma}. \tag{108}$$

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

$$\mathbb{E}[\boldsymbol{S}_x\boldsymbol{A}\boldsymbol{S}_x] = \frac{1}{l^2}[l(\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma})+l(l-1)\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}], \tag{109}$$

$$= \frac{1}{l}(\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma})+\frac{l-1}{l}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}, \tag{110}$$

$$= \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}+\frac{1}{l}\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{\Sigma}+\frac{1}{l}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{\Sigma}, \tag{111}$$

which completes the proof. ■

Appendix I. 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.

$$\mathcal{G}(\boldsymbol{V},\boldsymbol{M}) = \frac{1}{\tau^2}\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right)-\frac{1}{\tau}\mathrm{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11}+\boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right)+\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right)+\sigma^2, \tag{112}$$

as specified in Theorem 2. So, we can express the generalization error as,

$$\mathcal{G}(\tau;\boldsymbol{V},\boldsymbol{M}) = \frac{a}{\tau^2}-\frac{b}{\tau}+c, \tag{113}$$

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

$$\tau_{\text{opt}} := \arg\min_{\tau}\mathcal{G}(\tau;\boldsymbol{V},\boldsymbol{M}), \tag{114}$$

$$= \arg\min_{\tau}\left\{\frac{a}{\tau^2}-\frac{b}{\tau}+c\right\}. \tag{115}$$

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

$$\mathcal{G}'(\tau) = -2a\tau^{-3}+b\tau^{-2}. \tag{116}$$

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

$$\mathcal{G}'(\tau) = -2a\tau^{-3}+b\tau^{-2} = 0, \tag{117}$$

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

$$\tau = \frac{2a}{b}. \tag{118}$$

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

$$\mathcal{G}''(\tau) = 6a\tau^{-4}-2b\tau^{-3}. \tag{119}$$

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

$$\mathcal{G}''\left(\frac{2a}{b}\right) = 6a\left(\frac{2a}{b}\right)^{-4} - 2b\left(\frac{2a}{b}\right)^{-3} = 6a\left(\frac{b^4}{16a^4}\right) - 2b\left(\frac{b^3}{8a^3}\right) = \frac{b^4}{8a^3}. \tag{120}$$

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

$$\tau_{\text{opt}} = \frac{2\text{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right)}{\text{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right)}, \tag{121}$$

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

I.1. 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. Building upon prior work that connects ICL in linear regression to the Bayes-optimal ridge estimator [24, 36], we configure $\boldsymbol{M}$ and $\boldsymbol{V}$ to emulate Bayes-optimal ridge regression (see Proposition 5).

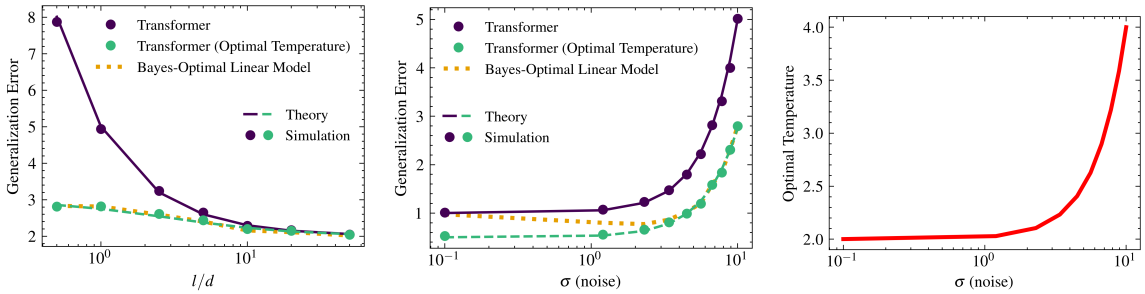
Since the pretrained model succeeds in ICL for $\mathcal{D}^{test} = \mathcal{D}^{train}$, we next investigate how distribution shifts affect its ICL performance.

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. 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 (1). 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 model invariant to mean shifts, but the uncentered linear attention model remains sensitive, as noted in Remark 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 [36].

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 5). 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.



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

Figure 4: Effect of noise shift on Transformer (3). The pretraining noise is $\sigma_{train} = 0.1$, while σ_{test} varies across plots. The optimal temperature is set by Theorem 3. This setting matches Figure 1a, except for changes in test-time noise σ_{test}.

Case III: Shift in noise distribution — Finally, consider a change in the noise distribution: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with σ^2_{train} and σ^2_{test} denoting pretraining and testing noise variances. If $\sigma^2_{train} \neq \sigma^2_{test}$, 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.

We further evaluate robustness to label noise in Figure 4. In Figure 4a, 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 4b (for $l = d$), the Transformer increasingly diverges from the Bayes-optimal model as noise grows, yet optimal temperature correction closes this gap. Figure 4c shows that the optimal temperature increases with noise level, indicating a principled relationship between noise and temperature under limited context.

Appendix J. An insight driven from optimal temperature for other settings

In this section, we extract a mathematical heuristic from the optimal temperature in Theorem 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 $(\boldsymbol{KZ})^\top(\boldsymbol{QZ})$, 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 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 D).

We now illustrate how the optimal temperature in Theorem 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

$$\tau_{\text{opt}} = \frac{v_{22}\,\mathrm{Tr}\big(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^\top \boldsymbol{\Sigma}_x\big)}{\frac{1}{2}\,\mathrm{Tr}\big(\boldsymbol{\Sigma}_x \left(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11} + \boldsymbol{M}_{11}^\top \boldsymbol{\Sigma}_x^\top\right)\big)}. \tag{122}$$

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

$$\mathbb{E}[\boldsymbol{z}_i^\top \boldsymbol{M} \boldsymbol{z}_i] = \mathrm{Tr}\big(\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^\top]\boldsymbol{M}\big) = \mathrm{Tr}\big(\boldsymbol{\Sigma}_x \boldsymbol{M}_{11}\big), \tag{123}$$

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

$$\mathbb{E}[\boldsymbol{z}_i^\top \boldsymbol{M} \boldsymbol{z}_j] = \mathrm{Tr}\big(\boldsymbol{M}\mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_j^\top]\big) = 0, \tag{124}$$

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

$$\begin{aligned}
\mathbb{E}\Big[(\boldsymbol{z}_i^\top \boldsymbol{M} \boldsymbol{z}_j)^2\Big] &= \mathbb{E}\Big[\boldsymbol{z}_i^\top \boldsymbol{M} \boldsymbol{z}_j \boldsymbol{z}_j^\top \boldsymbol{M}^\top \boldsymbol{z}_i\Big] && (125)\\
&= \mathbb{E}\Big[\boldsymbol{x}_i^\top \boldsymbol{M}_{11} \boldsymbol{x}_j \boldsymbol{x}_j^\top \boldsymbol{M}_{11}^\top \boldsymbol{x}_i\Big] && (126)\\
&= \mathbb{E}_{\boldsymbol{x}_i}\Big[\boldsymbol{x}_i^\top \boldsymbol{M}_{11} \mathbb{E}_{\boldsymbol{x}_j}[\boldsymbol{x}_j \boldsymbol{x}_j^\top] \boldsymbol{M}_{11}^\top \boldsymbol{x}_i\Big] && (127)\\
&= \mathbb{E}_{\boldsymbol{x}_i}\Big[\boldsymbol{x}_i^\top \boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^\top \boldsymbol{x}_i\Big] && (128)\\
&= \mathrm{Tr}\big(\boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^\top \mathbb{E}_{\boldsymbol{x}_i}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]\big) && (129)\\
&= \mathrm{Tr}\big(\boldsymbol{M}_{11} \boldsymbol{\Sigma}_x \boldsymbol{M}_{11}^\top \boldsymbol{\Sigma}_x\big), && (130)
\end{aligned}$$

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

We observe a parallel between the numerator of (122) 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 2), 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.

Appendix K. 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.

K.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 [23]

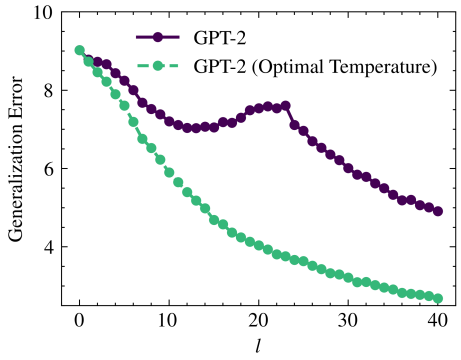


Figure 5: GPT-2 [23] under an input-covariance shift. GPT-2 exemplifies the Transformer architecture [28], combining multi-layer perceptrons with multi-head softmax self-attention. The model here is pretrained by Garg et al. [9] on the linear regression tasks defined in (1). We consider a shift from $\mathbf{\Sigma}_x^{\text{train}} = \boldsymbol{I}$ to $\mathbf{\Sigma}_x^{\text{test}} = 3\boldsymbol{I}$. 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.

under a shift in input covariance (Figure 5). Consistent with prior work [9, 36], 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.

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

We use the standard GPT-2 architecture [23] as implemented in HuggingFace [33], leveraging the pretrained model of Garg et al. [9]. 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. [9]. 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.

K.3. Details of the LLM experiments in Figure 2

For our large language model experiments, we employ Llama2-7B [27] and the SCIQ dataset [32], which contains science questions with supporting information. We generate ICL problems following Gao et al. [8], selecting in-context demonstrations using the TopK retrieval technique [17] to ensure relevance. An example ICL sample from SCIQ appears in Table 1. To simulate distribution shift, we follow Gao et al. [8] and introduce noisy labels—incorrect but semantically related—to the in-context demonstrations (Appendix K.4). Table 2 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. [8], built on HuggingFace [33] and OpenICL [35]. The attention temperature at each layer is scaled as $\tau\sqrt{d_k}$ (where d_k is the key dimension) so that the reported τ values are dimension-independent. All LLM experiments run on an NVIDIA A40 GPU; a single Monte Carlo run per plot in Figure 2 takes a few hours.

K.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 [27]. However, we hypothesize—following Gao et al. [8]—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 1: 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 2: 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.

Appendix L. Interpretation of the optimal attention temperature: full derivation

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

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

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

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

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 (7) yields

$$\tau_{\text{opt}} = \frac{2\text{Tr}\left(a\boldsymbol{I}\boldsymbol{M}_{11}^T\left(ab\boldsymbol{I} + \frac{1}{l}(\sigma^2 + abd)\boldsymbol{I}\right)a\boldsymbol{I}\boldsymbol{M}_{11}\right)}{\text{Tr}\left(a\boldsymbol{I}\left(ab\boldsymbol{I}\left(\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\right)\right)\right)}, \tag{131}$$

$$= \left(a + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}^T\boldsymbol{M}_{11}\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}, \tag{132}$$

recovering (8).

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 1) 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.
- *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: $\tau_{\text{opt}} \to a \cdot \frac{\text{Tr}(\boldsymbol{M}_{11}^\top \boldsymbol{M}_{11})}{\text{Tr}(\boldsymbol{M}_{11})}$.

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

Moment-ratio heuristic — We propose a practical heuristic (derived in Appendix J): the optimal attention temperature is roughly proportional to the ratio of the second and first moments of pre-softmax attention scores. The expression in (8) provides further theoretical justification for that heuristic. Indeed, for some $i \neq j$,

$$\tau_{\text{opt}} = a\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)} + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}, \tag{133}$$

$$= \frac{\mathbb{E}\left[(\boldsymbol{z}_i^\top \boldsymbol{M}\boldsymbol{z}_j)^2\right]}{\mathbb{E}\left[\boldsymbol{z}_i^\top \boldsymbol{M}\boldsymbol{z}_i\right]} + \frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)\frac{\text{Tr}\left(\boldsymbol{M}_{11}\boldsymbol{M}_{11}^T\right)}{\text{Tr}\left(\boldsymbol{M}_{11}\right)}, \tag{134}$$

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

$$\tau_{\text{opt}} \approx \underbrace{\frac{\mathbb{E}\left[(\boldsymbol{z}_i^\top \boldsymbol{M}\boldsymbol{z}_j)^2\right]}{\mathbb{E}\left[\boldsymbol{z}_i^\top \boldsymbol{M}\boldsymbol{z}_i\right]}}_{\text{moment-ratio}} + \underbrace{\frac{1}{l}\left(\frac{\sigma^2}{b} + ad\right)}_{\text{correction for small } l}, \tag{135}$$

recovering (9). 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.

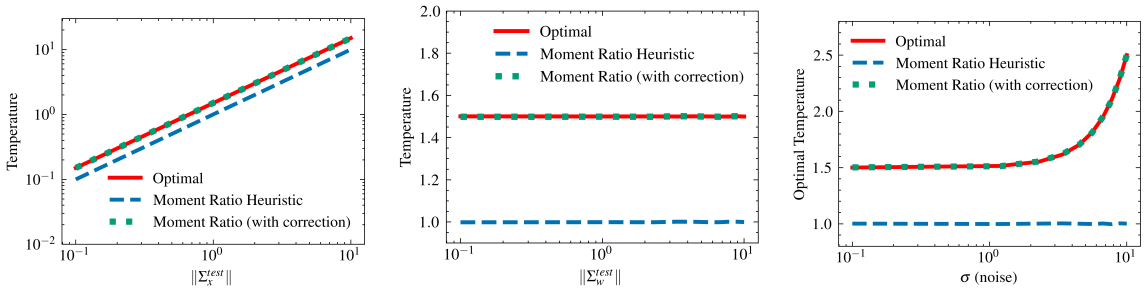
Appendix M. Numerical illustration of the optimal temperature and its generalization behavior

To complement the analytical reductions in the setting of Section 3, 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 3 and its simplified forms in (8) and (9).

Figure 6 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 (7) align closely with the moment-ratio heuristic with correction (9). As anticipated:

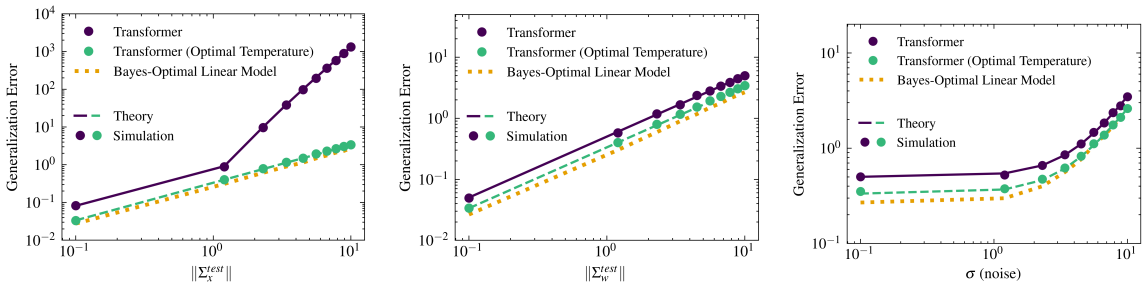
- higher input variance a or noise level σ increases the optimal temperature, while
- task variance b does not significantly change the optimal temperature.

Figure 7 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 ($\boldsymbol{\Sigma}_x^{test} = a\boldsymbol{\Sigma}_x^{train}$) (*b*) Task Covariance Shifted ($\boldsymbol{\Sigma}_w^{test} = b\boldsymbol{\Sigma}_w^{train}$) (*c*) Noise Shifted ($\sigma^{test} = \sigma$)

Figure 6: Optimal temperature under different types of distribution shift. The moment-ratio heuristic (Appendix J) is derived from the closed-form optimal temperature, with its corrected form given in (9). 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 $\boldsymbol{\Sigma}_x^{train} = \boldsymbol{\Sigma}_w^{train} = \boldsymbol{I}$. At test time, we set $\boldsymbol{\mu}_x^{test} = \boldsymbol{\mu}_w^{test} = \mathbf{0}$, $\boldsymbol{\Sigma}_x^{test} = aI$, $\boldsymbol{\Sigma}_w^{test} = bI$, 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 ($\boldsymbol{\Sigma}_x^{test} = a\boldsymbol{\Sigma}_x^{train}$) (*b*) Task Covariance Shifted ($\boldsymbol{\Sigma}_w^{test} = b\boldsymbol{\Sigma}_w^{train}$) (*c*) Noise Shifted ($\sigma^{test} = \sigma$)

Figure 7: Generalization errors corresponding to Figure 6.