DUAL OPERATING MODES OF IN-CONTEXT LEARNING

Ziqian Lin
Department of Computer Science
University of Wisconsin-Madison
zlin284@wisc.edu

Kangwook Lee
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
kangwook.lee@wisc.edu

ABSTRACT

In-context learning (ICL) exhibits dual operating modes: *task learning*, *i.e.*, acquiring a new skill from in-context samples, and *task retrieval*, *i.e.*, locating and activating a relevant pretrained skill. Recent theoretical work proposes various mathematical models to analyze ICL, but they cannot fully explain the duality. In this work, we analyze the dual operating modes leveraging assumptions on the pretraining data. Based on our analysis, we obtain a quantitative understanding of the two operating modes of ICL. We first explain an unexplained phenomenon observed with real-world large language models (LLMs), where the ICL risk initially increases and then decreases with more in-context examples. We also analyze ICL with biased labels, *e.g.*, zero-shot ICL, where in-context examples are assigned random labels, and predict the bounded efficacy of such approaches. We corroborate our analysis and predictions with extensive experiments with real-world LLMs.

1 INTRODUCTION

In-context learning (ICL), an emergent ability of large language models (LLMs), **operates in two distinct modes: task learning and task retrieval** (Pan et al., 2023). They can learn unseen functions from in-context examples, demonstrating the learning mode (Brown et al., 2020; Razeghi et al., 2022; Garg et al., 2022). Concurrently, LLMs can also retrieve a *pretrained* skill. Motivated by this, our work seeks to address the following questions: *How do we rigorously explain the dual operating modes of ICL? Can we define the conditions when retrieval mode is dominant and vice versa?*

A New Model for Pretraining Data To find the answers, we first propose a new probabilistic model for pretraining data. We consider in-context learning of linear functions (Garg et al., 2022; Akyürek et al., 2023; Li et al., 2023; von Oswald et al., 2023; Raventos et al., 2023; Wu et al., 2024). We extend the existing model for pretraining data (Raventos et al., 2023) by introducing multiple task groups and task-dependent input distributions. Shown on the left-most panel in Fig. 1 is a simple visualization of our model. As illustrated, the red task group is modeled as the cluster of linear functions with negative coefficients ($\boldsymbol{w} \approx -1$), with input distribution centered at $\mathbb{E}[\boldsymbol{x}] = -1$.

Analysis With our new model for pretraning data, we analyze the Bayes-optimal pretrained model under the squared loss. Here, the pretraining task distribution (of multiple task groups) is the prior, and in-context examples are the observations. By fully quantifying the posterior distribution, we characterize how in-context examples are used to update those task groups. We will call updates of mixture probabilities as *task group (component) re-weighting* and updates of task group center as *task group (component) shifting*. See the central panel in Fig. 1 for visualization. By analyzing these two effects, we obtain a quantitative understanding of how two different operating modes emerge.

Explanation of Two Real-World Phenomena We demonstrate the practical value of our new insights by explaining and predicting two phenomena observed with LLMs in practice.

- **The *early ascent* phenomenon** refers to the observation that, the ICL risk initially increases and then decreases when more in-context examples are introduced (Brown et al., 2020; Xie et al., 2022). (Fig. 1 left) We offer a plausible explanation for this early ascent phenomenon–a limited number of in-context samples may lead to the retrieval of an incorrect skill, thereby increasing the risk.
- **Bounded efficacy of biased-label ICL** is predicted by our model. ICL performs well even with biased labels (Lyu et al., 2023; Min et al., 2022). Our model provides a rigorous justification that if in-context examples with biased labels carry sufficient information for retrieving a correct pretrained

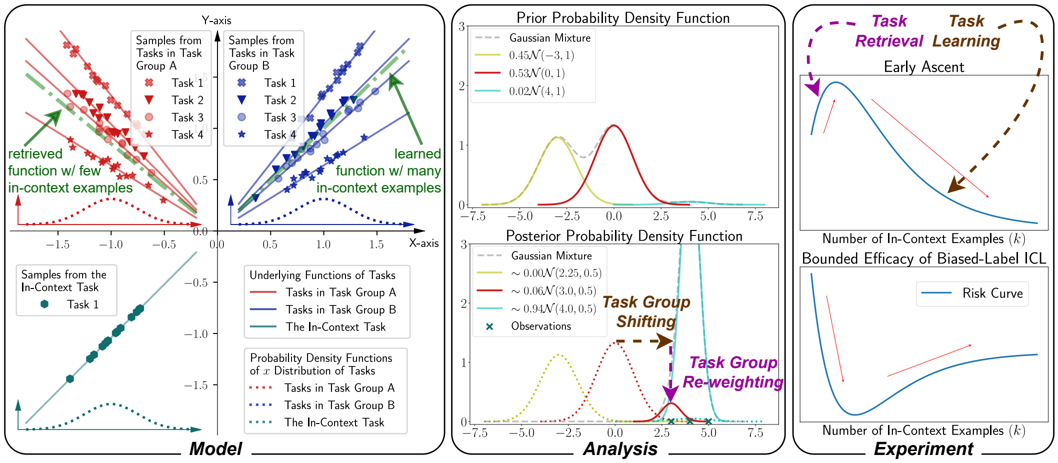


Figure 1: **A summary of our contributions.** We propose a model for pretraining data and in-context examples. Via analysis, we obtain a quantitative understanding of the duality of ICL, and explain two real-world phenomena observed with LLMs.

task, then this approach would work. Meanwhile, our analysis suggests that when the learning mode starts taking place, the test risks of such methods will start increasing. (Fig. 1 left) We observe the predicted phenomenon with real-world LLMs such as Mistral, Mixtral, Llama 2, and GPT-4.

2 Related Work

Dual Operating Modes of ICL. Pan et al. (2023) empirically disentangle the two operating modes of ICL: task recognition, which we refer to as task retrieval, and task learning.

Explaining ICL via Bayesian Inference. Xie et al. (2022) model the pretraining data with a Hidden Markov Model (HMM) (Ghahramani & Jordan, 1995; Rabiner, 1989). On the other hand, Garg et al. (2022); Raventos et al. (2023) consider the setting where a next-token prediction model is pretrained on token sequences in the form of $(\boldsymbol{x}_1, y_1, \boldsymbol{x}_2, y_2, \ldots)$. While this linear regression model facilitates a tractable analysis and elucidates certain aspects of the dual operating modes, it falls short in modeling the clustered characteristic of nature language. Han et al. (2023) show that ICL asymptotically approaches kernel regression as the number of in-context samples increases. On the other hand, our proposed model allows for tractable analysis and captures the clustered characteristic.

Explaining ICL via Gradient Descent. Garg et al. (2022) hint that under ICL, the pretrained Transformer might implicitly execute gradient descent. Akyürek et al. (2023); von Oswald et al. (2023); Dai et al. (2023) show that one attention layer can be exactly constructed to perform gradient descent. Further, Ahn et al. (2023); Mahankali et al. (2024); Zhang et al. (2023) show that the pretrained transformer will implement gradient descent algorithm.

3 Pretraining and Data Generative Model

A next-token predictor is a sequential prediction model that predicts the next token given an initial token sequence. During pretraining, this model receives sequences $\mathcal{S}_K = (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_K, y_K)$ with $2K$ tokens to predict only the y values in the sequences. During inference, the model receives a sequence of $2k+1$ tokens with k labeled samples $(\boldsymbol{x}_i, y_i), i \in \{1, \ldots, k\}$, and an unlabeled $\boldsymbol{x}_{k+1}$.

3.1 Pretraining Data Generative Model

We assume pretraining data consists of sequences generated based on Assumption 1.

Assumption 1 (Pretraining Data Generative Model)**.** *Given $\mathcal{D}^{prior}$, $\mathcal{D}_{\boldsymbol{x}}$, and $\mathcal{D}_{y|\boldsymbol{x}}$, we generate $\mathcal{S}_K$:*
(a) $(\boldsymbol{\mu}, \boldsymbol{w}) \sim \mathcal{D}^{prior} : P(\boldsymbol{\mu}, \boldsymbol{w}) = \sum_{m=1}^{M} \pi_m P(\boldsymbol{\mu}, \boldsymbol{w}|T_m)$, where T_m is the mth mixture component[1] of the Gaussian mixture, i.e., $P(\boldsymbol{\mu}, \boldsymbol{w}|T_m) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_m, \sigma_\mu^2 \boldsymbol{I}) \cdot \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_m, \sigma_w^2 \boldsymbol{I})$, and π_m is the

[1]The concept "mixture component" is analogous to the term "Task Group" depicted in the left panel of Fig. 1.

mixture weight. $\sum_{m=1}^{M} \pi_m = 1$, $0 < \pi_m < 1$, $(\boldsymbol{\mu}_m, \boldsymbol{w}_m)$ *is the center of the mixture component* T_m*, and all components share the same covariance matrix controlled by* σ_μ *and* σ_w*;*
(b) $\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu}), P(\boldsymbol{x}|\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sigma_x^2 \boldsymbol{I})$*; (c)* $y|\boldsymbol{x} \sim \mathcal{D}_{y|\boldsymbol{x}}(\boldsymbol{w}) : P(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma_y^2)$*;*
$\forall m$*, (d)* $\|\boldsymbol{\mu}_m\| = \|\boldsymbol{w}_m\| = 1$*; (e)* $\exists r > 1$ *that* $\frac{1}{r} \leq \frac{\pi_\alpha}{\pi_m} \leq r$*; (f)* $\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\mu}_m, \boldsymbol{w}, \boldsymbol{w}_m \in \mathbb{R}^d$.

In the sequence $\mathcal{S}_K$, the first $2k$ elements of $\mathcal{S}_K$ is denoted as $\mathcal{S}_k$, and the first $2k+1$ elements will be indicated by $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$, *e.g.*, $\mathcal{S}_0 = [\,]$, and $\mathcal{S}_1 \oplus \boldsymbol{x}_2 = [\boldsymbol{x}_1, y_1, \boldsymbol{x}_2]$.

3.2 Bayes-Optimal Next-Token Predictor

We consider the pretraining objective: $\mathcal{L}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_K}\left[\frac{1}{K}\sum_{k=0}^{K-1}(\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1})^2\right]$, where $\mathcal{F}$ is a next-token predictor and $\mathcal{S}_K$ is generated from $\mathcal{D}^{\text{prior}}$ following Assumption 1. We show that the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ by a Bayes-optimal pretrained $\mathcal{F}^*$ satisfies:

$$\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \mathop{\mathbb{E}}_{\boldsymbol{\mu}, \boldsymbol{w}}\left[\mathop{\mathbb{E}}_{y_{k+1}}[y_{k+1}|\boldsymbol{w}, \boldsymbol{x}_{k+1}]\middle|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}\right]. \tag{1}$$

(See Appendix A.1 for derivation.) Thus, $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ is the expectation (over task posterior) of $\mathop{\mathbb{E}}_{y_{k+1}}[y_{k+1}|\boldsymbol{w}, \boldsymbol{x}_{k+1}]$ regarding $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ as observation.

4 Inference and Dual Operating Modes

4.1 In-Context Task and In-Context Function

We introduce Assumption 2 for the in-context task and the in-context function of in-context examples:

Assumption 2 (Gaussian/Linear Assumptions for Inference)**.** *(a) The input sequence* $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ *of ICL satisfies,* $\forall i$, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I})$, $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle$*; (b)* $\|\boldsymbol{\mu}^*\| = \|\boldsymbol{w}^*\| = 1$.

Assumption 2(a) states that $(\boldsymbol{x}_i, y_i)$ follows in-context task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$ and in-context function $\boldsymbol{w}^*$.

4.2 Closed-Form Expression of Posterior

The following lemma gives the closed-form expression of posterior $\mathcal{D}^{\text{post}}$ given any $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$:

Lemma 1 (Conjugate Distributions with Noisy Linear Regression Likelihood)**.** *Under Assumption 1, the posterior probability of task* $(\boldsymbol{\mu}, \boldsymbol{w})$ *given observation* $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ *is:*

$$P(\boldsymbol{\mu}, \boldsymbol{w}|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \sum_{m=1}^{M} \tilde{\pi}_m P(\boldsymbol{\mu}, \boldsymbol{w}|\widetilde{T}_m) = \sum_{m=1}^{M} \tilde{\pi}_m \cdot \mathcal{N}(\boldsymbol{\mu}|\tilde{\boldsymbol{\mu}}_m, \tilde{\sigma}_\mu^2 \boldsymbol{I}) \cdot \mathcal{N}(\boldsymbol{w}|\tilde{\boldsymbol{w}}_m, \tilde{\sigma}_w^2 \boldsymbol{I}),$$

where, $\widetilde{T}_m$ *is the mixture component in the posterior with mixture weight* $\tilde{\pi}_m$ *and component center* $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$*. (See Appendix A.2 for closed-form expressions of* $\tilde{\pi}_m$*,* $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$*,* $\tilde{\sigma}_\mu$*, and* $\tilde{\sigma}_w$*.)*

Lemma 1 states that the task posterior remains a Gaussian mixture, with component centers shifted to $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$, namely *component shifting* (CS) and with mixture weights re-weighted to $\tilde{\pi}_m$, namely *component re-weighitng* (CR).

4.3 Closed-form Expression of ICL Prediction

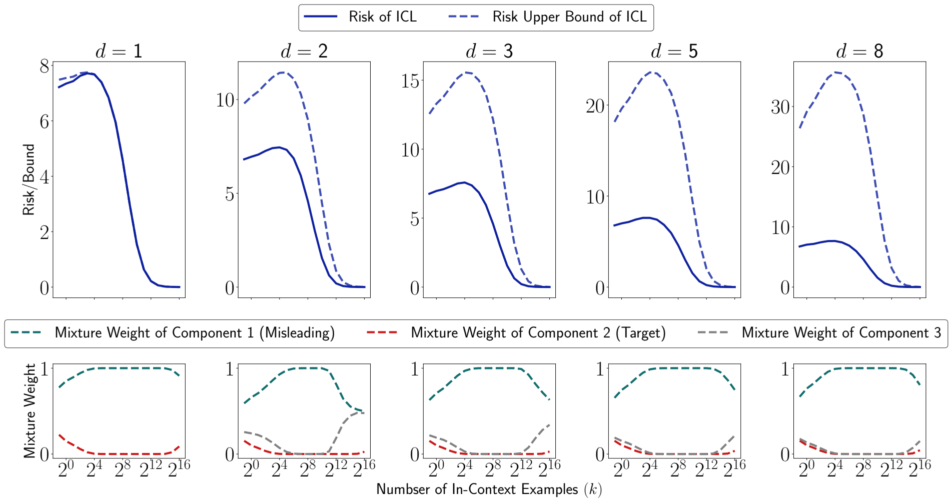
With Assumption 1 and Lemma 1, we have the following corollary for the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$:

Corollary 1. *Let* $\tilde{\boldsymbol{w}} = \sum_{m=1}^{M} \tilde{\pi}_m \tilde{\boldsymbol{w}}_m$*. With pretraining data generative model 1, if the pretrained model* $\mathcal{F}^*$ *minimizes the pretraining risk, then the prediction on any sequence* $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ *by* $\mathcal{F}^*$ *is:* $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \left\langle \boldsymbol{x}_{k+1}, \sum_{m=1}^{M} \tilde{\pi}_m \tilde{\boldsymbol{w}}_m \right\rangle = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle$*. See Appendix A.3 for proof details.*

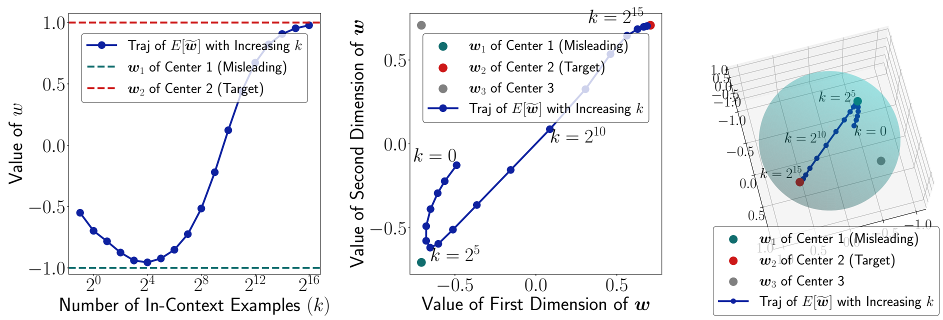
Under Assumption 2, we collected mathematical analyses and numerical computations of CS, CR, and Prediction in Appendix J, exploring the impacts of pretraining task noises and the number of in-context examples on $\tilde{\pi}_m$, $\tilde{\boldsymbol{w}}_m$, and $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$.

4.4 Dual Operating Modes

"Task retrieval" mode occurs when component re-weighting outweighs shifting, making predictions rely more on the interplay between pretraining prior and in-context examples. In contrast, ***"task learning"*** mode occurs when component shifting prevails, leading predictions to be based mostly on in-context examples, neglecting the pretraining prior.



(a) The mixture weights with increasing k under d equal to 1, 3 and 8.



(b) The trajectory of the expectation of $\tilde{\boldsymbol{w}}$ with increasing k under d equal to 1, 2 and 3.

Figure 2: **The early ascent phenomenon.** Fig. 2(a) shows expected losses, upper bounds, and mixture weights, while Fig. 2(b) shows the expectation of $\tilde{\boldsymbol{w}}$. Under these settings (see Appendix D.3 for setting details) As k increases, task retrieval happens first and retrieves a misleading task, causing increasing risks, and then risks decrease due to task learning.

5 Early Ascent

Brown et al. (2020) report that GPT-3 on LAMBADA shows a lower one-shot accuracy (72.5%) than zero-shot accuracy (76.2%), but the few-shot accuracy (86.4%) is higher than the zero-shot accuracy. Xie et al. (2022) replicate this phenomenon with their synthetic dataset and explain this by "the few-shot setting introduces the distracting prompt structure, which can initially lower accuracy." Based on our analysis, we take a further step to formalize this explanation. See Appendix H where we derive that the early ascent phenomenon provably occurs under a certain assumption. We also reproduce this performance tendency in Fig. 2(a) under our model, where the upper bound and the risk initially increase due to the misleading task (component 1) is retrieved first. Figure 2(b) further demonstrates the relative locations between the retrieved functions and functions of prior centers. Finally, we give the formal theorem on the early ascent phenomenon:

Theorem 2 (Early Ascent)**.** *Assume* $\mathbb{E}_{\boldsymbol{x}_1}\left[\frac{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2}{2\sigma_x^2})\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2}{2\sigma_x^2})}\right] <$ $\mathbb{E}_{\boldsymbol{x}_1}\left[\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2\right]$*, where* $\alpha = \arg\min_m \frac{\|\boldsymbol{\mu}_m-\boldsymbol{\mu}^*\|^2}{2\sigma_x^2} + \frac{\|(\boldsymbol{w}_m-\boldsymbol{w}^*)^\top\boldsymbol{\mu}^*\|^2+d\tau_x^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{2\sigma_y^2}$*. Then, when* δ_μ *and* δ_w *are small enough, we have the early ascent phenomenon on the risk upper bound:*

$$\exists k \geq 1 \text{ s.t. } \mathbb{E}_{\boldsymbol{x}_1}\left[\frac{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2}{2\sigma_x^2})\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2}{2\sigma_x^2})}\right]$$

$$< \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\frac{\sum_{m=1}^{M} \tilde{\pi}_m \|\boldsymbol{x}_1\|^2 \|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{\sum_{m=1}^{M} \tilde{\pi}_m} \right],$$

where $\mathbb{E}[\|\boldsymbol{x}_1\|^2 \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2]$ equals to the risk when the prediction is fully depends on the a misleading task function $\boldsymbol{w}_\alpha$ of prior center α. See Appendix H.3 for proof details.

6 Bounded Efficacy of Biased-Label ICL

The following theorem shows an upper bound for ICL risk with biased labels under mild assumption to describe the bias of the labels (see Appendix Assumption 3 for details):

Theorem 3 (Upper Bound for ICL with Biased Labels)**.** *Consider a next-token predictor attaining the optimal pretraining risk. When $\delta_\mu = \sigma_\mu^2/\sigma_x^2$ and $\delta_w = \sigma_w^2/\sigma_y^2$ are sufficiently small, there exists a particular interval for k such that ICL risk with biased labels is upper bounded by:*

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < C_1 \exp(-k(\frac{d_{\boldsymbol{\mu}}^2}{8\sigma_x^2} + \frac{u_{\boldsymbol{w}}^2 \tau_x^2}{8\sigma_y^2})) + C_2 \exp(-\frac{k^{\frac{1}{2}}}{8}) + C_3 \min\{1, 4k^2 {\delta_w}^2 (1+\tau_x^2)^2\},$$

where $\mathcal{L}_k^\alpha = (\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle)^2$, indicating the target function $\boldsymbol{w}_\alpha$ is associated with a prior center. C_1, C_2, and C_3 are constants depending on the prior setting. When k is small, the first and second terms dominate and exponential decay. When k is large, the second term dominates and increases. Thus, we predict a bounded efficacy phenomenon. See Appendix K.3 for proof details.

Table 1: **Bounded efficacy in GPT-4.** Error rate measured with respect to "addition (+)" and "biased +". The error rate of "+" goes down to $k = 2$, but it increases afterward.

Number of In-context Examples (k)	0	1	2	4	8	16
Error Rate of "Addition"	75.0%	36.2%	**33.9%**	49.3%	79.3%	85.1%
Error Rate of "Off-by-one Addition"	100.0%	98.3%	95.9%	60.5%	24.4%	**16.8%**

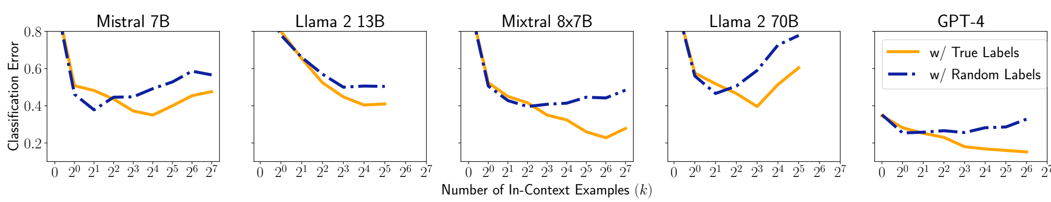


Figure 3: **Bounded efficacy.** The error rates of ICL with random labels increases at large k.

To verify that such a phenomenon exists in real-world LLMs, we first conducted an experiment with GPT-4 to check which function GPT-4 will predict using "biased +" as the in-context function. The results of Table 1 verify the existence of the bounded efficacy phenomenon, where GPT-4 will first retrieve "+" and then learn "biased +". See experiment details in Appendix F. We further extend the experiments of Min et al. (2022) to show this phenomenon exists in real-world ICL algorithms in Fig 3, such as zero-shot ICL (Lyu et al., 2023) with random labels. The results highlight the bounded efficacy phenomenon in the error curve associated with random labels compared to gold labels. See Appendix G for experimental results on all five LLMs and experiment setup.

7 Conclusion

In this paper, we introduced a probabilistic model for understanding the dual operating modes of in-context learning: task learning and task retrieval. Our analysis allowed us to explain the existing early ascent phenomenon observed in real-world ICL applications, and predict a new bounded efficacy phenomenon of biased-label ICL. We validated our findings and predictions via experiments involving real-world LLMs. Our work lays the groundwork for future research in further exploration and improvement of ICL.

8 Reproducibility Statement

The code for all experiments reported in this paper is publicly accessible. For the purpose of reproducibility, the code can be found at the following GitHub repository: `https://github.com/UW-Madison-Lee-Lab/Dual_Operating_Modes_of_ICL`.

References

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 (NeurIPS)*, 2023.

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 (ICLR)*, 2023.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *PASCAL Machine Learning Challenges Workshop (MLCW)*, 2005.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *International Workshop on Paraphrasing (IWP@IJCNLP)*, 2005.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can Transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zoubin Ghahramani and Michael Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1995.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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 (ICML)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Arvind Mahankali, Tatsunori B. 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 (ICLR)*, 2024.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation (LREC)*, 2014.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

OpenAI. GPT-4 technical report, 2023.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. The effects of pretraining task diversity on in-context learning of ridge regression. In *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*, 2023.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

Emily Sheng and David Uthus. Investigating societal biases in a poetry composition system. In *Workshop on Gender Bias in Natural Language Processing*, 2020.

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.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research (JMLR)*, 2023.

Harry L Van Trees. *Detection, estimation, and modulation theory, Part I: Detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning (ICML)*, 2023.

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

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. In *Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo)*, 2023.

Appendix

We organize the appendix with the following structure:

A PREDICTION AND POSTERIOR

A.1 BAYES-OPTIMAL NEXT-TOKEN PREDICTOR

We consider the pretraining objective: $\mathcal{L}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_K}\left[\frac{1}{K}\sum_{k=0}^{K-1}(\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1})^2\right]$, where $\mathcal{F}$ is a next-token predictor and $\mathcal{S}_K$ is generated from $\mathcal{D}^{\text{prior}}$ following Assumption 1. In other words, for each sequence, we pretrain $\mathcal{F}$ to predict each label y based on preceding samples, measuring risk with the squared loss. Due to the linearity of expectation, we have: $\mathcal{L}(\mathcal{F}) = \frac{1}{K}\sum_{k=0}^{K-1}\mathop{\mathbb{E}}_{\mathcal{S}_K}\left[(\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1})^2\right]$. A variable-input-length next-token predictor $\mathcal{F}$ can be viewed as K fixed-input-length next-token predictors $\mathcal{F}_0, \ldots, \mathcal{F}_{K-1}$, where $\mathcal{F}_k$ takes a sequence of exactly $2k+1$ tokens as input. Thus, assuming the sufficient expressiveness of $\mathcal{F}$, the pretraining process of minimizing $\mathcal{L}(\mathcal{F})$ can be decomposed into K separate optimization problems:

$$\mathcal{F}_k^* = \mathop{\mathrm{argmin}}_{\mathcal{F}_k} \mathop{\mathbb{E}}_{\mathcal{S}_K}[(\mathcal{F}_k(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1})^2],\ \forall k \in [K].$$

The solution denoted $\mathcal{F}_k^*$ is an MMSE estimator (Van Trees, 2004, page 63) for each k. Thus, the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \mathcal{F}_k^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ satisfies:

$$\begin{aligned}
\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) &= \mathop{\mathbb{E}}_{\mathcal{S}_K}\left[y_{k+1}|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}\right] \qquad (2)\\
&= \mathop{\mathbb{E}}_{\mathcal{D}_{\boldsymbol{x},y}}\left[\mathop{\mathbb{E}}_{y_{k+1}}\left[y_{k+1}|\mathcal{D}_{\boldsymbol{x},y}, \mathcal{S}_k \oplus \boldsymbol{x}_{k+1}\right]\middle|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{D}_{\boldsymbol{x},y}}\left[\mathop{\mathbb{E}}_{y_{k+1}}\left[y_{k+1}|\mathcal{D}_{\boldsymbol{x},y}, \boldsymbol{x}_{k+1}\right]\middle|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}\right].
\end{aligned}$$

Thus, $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ is the expectation (over task posterior) of $\mathop{\mathbb{E}}_{y_{k+1}}\left[y_{k+1}|\mathcal{D}_{\boldsymbol{x},y}, \boldsymbol{x}_{k+1}\right]$ regarding $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ as observation. We show that the pretrained Transformer can approximate Bayesian inference in Appendix E.

A.2 Closed-Form Expression of Posterior

The following lemma gives the closed-form expression of posterior $\mathcal{D}^{\text{post}}$ given any $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$:

Lemma 2 (Conjugate Distributions with Noisy Linear Regression Likelihood)**.** *Under Assumption 1, the posterior probability of task* $(\boldsymbol{\mu}, \boldsymbol{w})$ *given observation* $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ *is:*

$$P(\boldsymbol{\mu}, \boldsymbol{w}|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \textstyle\sum_{m=1}^M \tilde{\pi}_m P(\boldsymbol{\mu}, \boldsymbol{w}|\widetilde{T}_m) = \textstyle\sum_{m=1}^M \tilde{\pi}_m \cdot \mathcal{N}(\boldsymbol{\mu}|\tilde{\boldsymbol{\mu}}_m, \tilde{\sigma}_\mu^2 \boldsymbol{I}) \cdot \mathcal{N}(\boldsymbol{w}|\tilde{\boldsymbol{w}}_m, \tilde{\sigma}_w^2 \boldsymbol{I}).$$

Here, the mixture component T_m *in the prior is mapped to the mixture component* $\widetilde{T}_m$ *in the posterior with mixture weight* $\tilde{\pi}_m$ *and component center* $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$*:*

$$\begin{aligned}
\tilde{\pi}_m &= \pi_m C_1 c_m^{\boldsymbol{\mu}} c_m^{\boldsymbol{w}},\\
c_m^{\boldsymbol{\mu}} &= \exp\left(-\|\boldsymbol{\mu}_m\|^2 - \|\boldsymbol{\mu}_m + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+(k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}} / 2\sigma_\mu^2\right),\\
c_m^{\boldsymbol{w}} &= \exp\left(-\|\boldsymbol{w}_m\|^2 - \|\boldsymbol{w}_m + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} / 2\sigma_w^2\right),\\
\tilde{\boldsymbol{\mu}}_m &= (\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}),\\
\tilde{\boldsymbol{w}}_m &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m + k\delta_w \bar{\boldsymbol{w}}),\\
\tilde{\sigma}_\mu^2 &= \sigma_\mu^2 (\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1},\\
\tilde{\sigma}_w^2 &= \sigma_w^2 (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1},
\end{aligned}$$

where C_1 *is a normalizing constant, i.e.,* $\sum_m \tilde{\pi}_m = 1$*,* $\delta_\mu = \frac{\sigma_\mu^2}{\sigma_x^2}$*,* $\delta_w = \frac{\sigma_w^2}{\sigma_y^2}$*,* $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \boldsymbol{I}$*,* $\bar{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{k+1} \boldsymbol{x}_i}{k+1}$*,* $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$*, and* $\bar{\boldsymbol{w}} = \frac{\sum_{i=1}^{k} \boldsymbol{x}_i y_i}{k}$*. See Appendix I for the proof.*

A.3 Closed-Form Expression of ICL Prediction

With Assumption 1 and Lemma 1, we have the following corollary for the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$:

Corollary 4. *Let* $\tilde{\boldsymbol{w}} = \sum_{m=1}^M \tilde{\pi}_m \tilde{\boldsymbol{w}}_m$*. With pretraining data generative model 1, if the pretrained model* $\mathcal{F}^*$ *minimizes the pretraining risk, then the prediction on any sequence* $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ *by* $\mathcal{F}^*$ *is as follows:* $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \left\langle \boldsymbol{x}_{k+1}, \sum_{m=1}^M \tilde{\pi}_m \tilde{\boldsymbol{w}}_m \right\rangle = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle$*.*

Proof. By applying Assumption 1 to Eq. 2, $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{w})\sim\mathcal{D}^{\text{prior}}}[\langle \boldsymbol{x}_{k+1}, \boldsymbol{w}\rangle|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}]$. Using Lemma 1, this reduces to $\sum_{m=1}^M \tilde{\pi}_m \mathop{\mathbb{E}}_{(\boldsymbol{\mu},\boldsymbol{w})\sim\widetilde{T}_m}[\langle \boldsymbol{x}_{k+1}, \boldsymbol{w}\rangle]$. Due to the linearity of expectation and inner product, the prediction can be simplified as $\langle \boldsymbol{x}_{k+1}, \sum_{m=1}^M \tilde{\pi}_m \tilde{\boldsymbol{w}}_m \rangle = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle$. $\square$

B Coarse Upper Bound of ICL Risk

The following theorem shows a coarse upper bound of ICL risk to learn a task:

Theorem 5 (Coarse Upper Bound for ICL)**.** *Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, ICL risk is upper bounded by:*

$$\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^*] < \frac{4(1 + d\tau_x^2)}{\tau_x^4 {\delta_w}^2 k^2} + O(k^{\delta - \frac{5}{2}}),$$

where $\mathcal{L}_k^* = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^*)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle)^2$ and δ is an arbitrarily small positive constant. See Appendix K.2 for proof details. The upper bound decreases as the square of the inverse of k. Notice there is no noise for y labels of in-context examples under our setting, which leads to a faster decay rate than standard $1/k$ for ridge regression (Tsigler & Bartlett, 2023).

The notations δ_μ, δ_w and k are colored for easier observation.

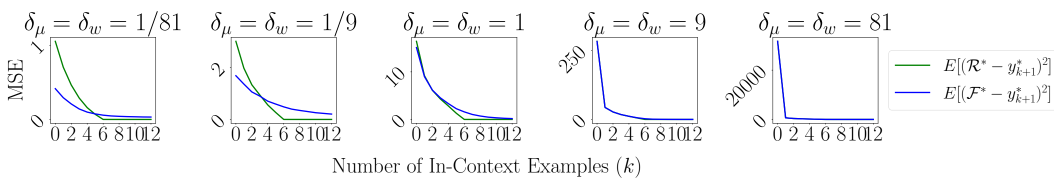


Figure 4: $\mathcal{R}^*$ indicates the prediction by Ridge regression, $\mathcal{F}^*$ indicates the prediction by ICL with a Bayes-optimal next-token predictor, and $y_{k+1}^* = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle$. Let the k samples draw from a task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$, which is drawn from the pretraining prior distribution. The dimension d of x equals 6. We observe that ICL performs better than Ridge regression of small k, and Ridge regression performs better than ICL when $k \geq d$. Especially, when the task prior distribution has high task variance (big δ_μ and δ_w values), ICL and Ridge regression have very similar performance.

We further compare the risk $\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^*]$ and the risk under Ridge regression with L2 regularization parameter equal to 10^{-6}, where the same k samples without label noises are used as in-context examples for ICL and training samples for Ridge regression. Fig. 4 shows the experiment results. In practice, we can observe more than the simple monotone loss decreasing phenomenon for ICL. The later sections will introduce and explain them.

C Notations

This section collects all notations used in the main paper.

Notations initially introduced in Sec. 3:

- $\mathcal{F}$: a next-token predictor.
- $\hat{\mathcal{F}}$: a pretrained next-token predictor.
- $\mathcal{F}^*$: a Bayes-optimal next-token predictor that attains Bayes risk minimization.
- $\mathcal{F}_k$: a Bayes-optimal next-token predictor for k in-context examples.
- $\mathcal{F}_k^*$: a Bayes-optimal next-token predictor that attains Bayes risk minimization for k in-context examples.
- $\boldsymbol{x}$ and y: input and label for a task, *e.g.*, $\boldsymbol{x}$ and y of a linear regression task $y = \boldsymbol{x}^\top \boldsymbol{w}$.
- k: the number of in-context examples.
- K: the max number of in-context examples in a sequence.
- $\mathcal{S}_k$: a sequence of k in-context examples, $[\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k]$.
- $\mathcal{S}_K$: a sequence of K in-context examples, $[\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_K, y_K]$.
- $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$: $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} = [\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1}]$, a sequence of k in-context examples and $\boldsymbol{x}_{k+1}$ pending to be predicted.
- $\boldsymbol{\mu}$ and $\boldsymbol{w}$: the parameters control a task. $\boldsymbol{\mu}$ controls the distribution of $\boldsymbol{x}$ and $\boldsymbol{w}$ controls the function mapping $\boldsymbol{x}$ to y.

- $\mathcal{D}^{\text{prior}}$ and $\mathcal{D}_{\boldsymbol{\mu},\boldsymbol{w}}$: $\mathcal{D}^{\text{prior}} = \mathcal{D}_{\boldsymbol{\mu},\boldsymbol{w}}$, and they represent the task prior distribution where each task is controlled by parameters $\boldsymbol{\mu}$ and $\boldsymbol{w}$. The task prior is also named pretraining prior, pretraining task prior, pretraining prior distribution, pretraining task prior distribution, or simply prior.
- $\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu})$: the conditional distribution of $\boldsymbol{x}$ conditioned on $\boldsymbol{\mu}$ of the task $(\boldsymbol{\mu}, \boldsymbol{w})$.
- $\mathcal{D}_{\boldsymbol{x},y}(\boldsymbol{\mu}, \boldsymbol{w})$: the joint distribution of $(\boldsymbol{x}, y)$ in the task $(\boldsymbol{\mu}, \boldsymbol{w})$.
- $\mathcal{D}_{y|\boldsymbol{x}}(\boldsymbol{w})$: y distribution conditioned on the input $\boldsymbol{x}$ and parameter $\boldsymbol{w}$ of the task $(\boldsymbol{\mu}, \boldsymbol{w})$.
- $P(\boldsymbol{\mu}, \boldsymbol{w})$: the task probability of $(\boldsymbol{\mu}, \boldsymbol{w})$ in the task prior $\mathcal{D}^{\text{prior}}$.
- $P(\boldsymbol{x} \mid \boldsymbol{\mu})$: the probability of $\boldsymbol{x}$ in $\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu})$.
- $P(y \mid \boldsymbol{x}, \boldsymbol{w})$: the probability of y in $\mathcal{D}_{y|\boldsymbol{x}}(\boldsymbol{w})$.
- $\mathcal{L}(\mathcal{F})$: the risk of $\mathcal{F}$ on samples generated from generative model 1.
- M: the number of mixture components in a Gaussian mixture prior.
- α, β: the indexes of a mixture component in a Gaussian mixture prior.
- T_β: the β^{the} mixture component in a Gaussian mixture prior.
- π_β: the mixture weight of the βth mixture component in a Gaussian mixture prior.
- $\boldsymbol{\mu}_\beta$ and $\boldsymbol{w}_\beta$: $(\boldsymbol{\mu}_\beta, \boldsymbol{w}_\beta)$ is the center of the βth mixture component.
- $\boldsymbol{\mu}^*$ and $\boldsymbol{w}^*$: $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$ is the in-context task, *i.e.*, in-context examples are drawn from this task without label noises.
- σ_μ and σ_w: the task noises, *i.e.*, the noise scales of $\boldsymbol{\mu}$ and $\boldsymbol{w}$.
- σ_x and σ_y: the sample noises, *i.e.*, the noise scales of $\boldsymbol{x}$ and y of pretraining samples.
- τ_x: the sample noise, *i.e.*, the noise scale of $\boldsymbol{x}$ of in-context examples.
- d: the dimension of $\boldsymbol{x}$.
- r: the max ratio of two mixture weights of two mixture components.

Notations initially introduced in Sec. 4:

- $\mathcal{D}^{\text{post}}$: The posterior distribution of the pretraining prior $\mathcal{D}^{\text{prior}}$ after observing $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$.
- $\|\boldsymbol{x}\|^2$: for any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x}$.
- $\|\boldsymbol{x}\|^2_{\boldsymbol{A}}$: for any vector $\boldsymbol{x}$ and matrix $\boldsymbol{A}$, $\|\boldsymbol{x}\|^2_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$.
- $P(\boldsymbol{\mu}, \boldsymbol{w} \mid \mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$: the probability of task $(\boldsymbol{\mu}, \boldsymbol{w})$ in the posterior after observing $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$.
- $\widetilde{T}_\beta$: the βth mixture component in the Gaussian mixture posterior.
- $\tilde{\pi}_\beta$: the mixture weight of the βth mixture component in the Gaussian mixture posterior.
- $\tilde{\boldsymbol{\mu}}_\beta$ and $\tilde{\boldsymbol{w}}_\beta$: $(\tilde{\boldsymbol{\mu}}_\beta, \tilde{\boldsymbol{w}}_\beta)$ is the center of the βth mixture component in the Gaussian mixture posterior.
- $P(\boldsymbol{\mu}, \boldsymbol{w} \mid \widetilde{T}_\beta)$: the probability of task $(\boldsymbol{\mu}, \boldsymbol{w})$ in the βth mixture component of posterior.
- δ_μ and δ_w: the ratio of squared task noise over squared sample noise. $\delta_\mu = \frac{\sigma_\mu^2}{\sigma_x^2}$, and $\delta_w = \frac{\sigma_w^2}{\sigma_y^2}$.
- $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}$: $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \boldsymbol{I}$.
- $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$: $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$.
- $\bar{\boldsymbol{\mu}}$: $\bar{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{k+1} \boldsymbol{x}_i}{k+1}$.
- $\bar{\boldsymbol{w}}$: $\bar{\boldsymbol{w}} = \frac{\sum_{i=1}^k \boldsymbol{x}_i y_i}{k}$.
- $\tilde{\boldsymbol{w}}$: the mean of $\boldsymbol{w}$ in the task posterior, *i.e.*, the predicted function by Bayes-optimal next-token predictor. $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle = \left\langle \boldsymbol{x}_{k+1}, \sum_{\beta=1}^M \tilde{\pi}_\beta \tilde{\boldsymbol{w}}_\beta \right\rangle$.
- $c_\beta^{\boldsymbol{\mu}}$ and $c_\beta^{\boldsymbol{w}}$: parts of the re-weighting coefficient of Component Re-weighting.
- $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ and $\Psi_{\boldsymbol{w}}(\alpha, \beta)$: functions to help analyze the phenomenon of Component Re-weighting.
- $r(\alpha, \beta)$: the ratio of the mixture weight $\tilde{\pi}_\alpha$ of $\widetilde{T}_\alpha$ over the mixture weight $\tilde{\pi}_\beta$ of $\widetilde{T}_\beta$.
- $\lambda_d(\boldsymbol{A})$: the dth largest eigenvalue of matrix $\boldsymbol{A}$. In this paper $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, thus $\lambda_d(\boldsymbol{A})$ represents the smallest eigenvalue of matrix $\boldsymbol{A}$.

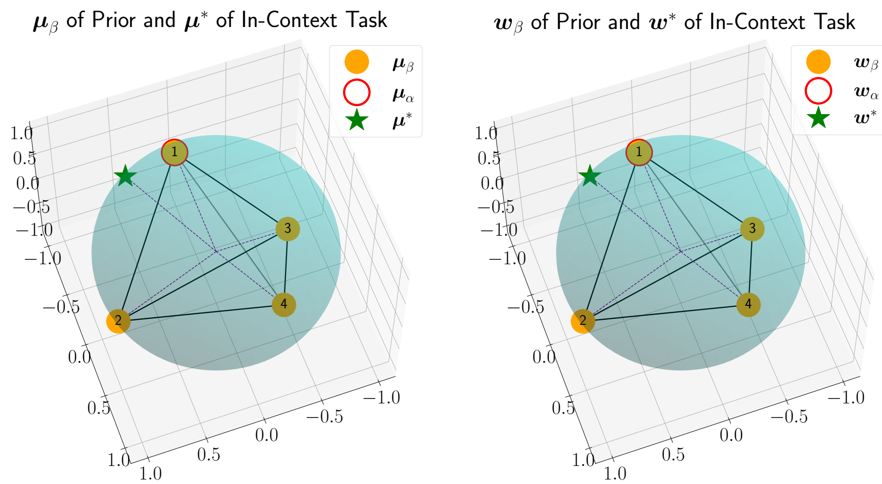


Figure 5: The figure shows the pretraining prior and in-context task. $(\boldsymbol{\mu}_\beta, \boldsymbol{w}_\beta)$ is mixture component center in the prior. $(\boldsymbol{\mu}_\alpha, \boldsymbol{w}_\alpha)$ for $\alpha = 1$ (numbers are noted in the center of circles) is the center of the target task for ICL with biased labels, while $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$ is the in-context task. The dotted purple lines highlight the distance of 1 from the origin $(0, 0, 0)$ to any point represented by $\boldsymbol{\mu}$ or $\boldsymbol{w}$.

- $\lambda_1(\boldsymbol{A})$: the 1st largest eigenvalue of matrix $\boldsymbol{A}$.
- y^*_{k+1}: the label of learning the function $\boldsymbol{w}^*$. $y^*_{k+1} = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle$.
- y^α_{k+1}: the label of retrieving the function $\boldsymbol{w}_\alpha$. $y^\alpha_{k+1} = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle$.

Notations initially introduced in Sec. 5

- $d^2_{\boldsymbol{\mu}}$: $\forall \beta \neq \alpha, \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 \geq d^2_{\boldsymbol{\mu}}$, the $\boldsymbol{\mu}$-margin of any other $\boldsymbol{\mu}_\beta$ over $\boldsymbol{\mu}_\alpha$.
- $d^2_{\boldsymbol{w}}$: $\forall \beta \neq \alpha, \|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq d^2_{\boldsymbol{w}}$, the $\boldsymbol{w}$-margin of any other $\boldsymbol{w}_\beta$ over $\boldsymbol{w}_\alpha$.
- $u^2_{\boldsymbol{w}}$: $\forall \beta \neq \alpha, \tau_x^2\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - (1+\tau_x^2)\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq \tau_x^2 u^2_{\boldsymbol{w}}$, the weighted $\boldsymbol{w}$-margin of any other $\boldsymbol{w}_\beta$ over $\boldsymbol{w}_\alpha$.
- The L2 loss of ICL learning to learn the function $\boldsymbol{w}^*$. $\mathcal{L}^*_k = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y^*_{k+1})^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle)^2$.
- The L2 loss of ICL learning to retrieve the function $\boldsymbol{w}_\alpha$ of pretraining prior. $\mathcal{L}^\alpha_k = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y^\alpha_{k+1})^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle)^2$.

D Prior Example

In this section, we introduce the prior settings we use in our numerical computations and small-scale Transformer experiments. We split the setting based on the shape of the centers in the priors. Those shapes include 3-dimensional regular polyhedrons in Sec. D.1, d-dimensional examples in Sec. D.2, and a special setting in Sec. D.3 for the early ascent phenomenon.

D.1 Regular Polyhedrons

For the abstract of the task prior (consider the mixture component centers), we consider 3-dimensional regular polyhedrons including Tetrahedron (4 vertices/centers), Octahedron (6 vertices/centers), Hexahedron (8 vertices/centers), Icosahedron (12 vertices/centers), and Dodecahedron (20 vertices/centers), listed with increasing density of the centers on a sphere.

A regular polyhedron setting with M centers is set as follows with all the parameters in Assumption 1:

- Dimension $d = 3$, number of mixture components $M = M$, and $\beta \in \{1, \ldots, M\}$;

- The centers of mixture components shape a regular polyhedron with M vertices;
- All component's mixture weights are the same, $\pi_\beta = 1/M$, and $\boldsymbol{\mu}_\beta = \boldsymbol{w}_\beta$, for all $\beta \in \{1, \dots, M\}$;
- For noises of $\boldsymbol{x}$ and y, we have $\sigma_x = \sigma_y = 1$, and $\tau_x = 1$;
- For noises of $\boldsymbol{\mu}$ and $\boldsymbol{w}$, we have $\sigma_\mu = \sigma_w = 0.25$ if not specified;
- For in-context task, $\boldsymbol{\mu}^* = \frac{2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{\|2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\|}$ and $\boldsymbol{w}^* = \frac{2\boldsymbol{w}_1+\boldsymbol{w}_2}{\|2\boldsymbol{w}_1+\boldsymbol{w}_2\|}$, where $\boldsymbol{\mu}_2$ is the closest center to $\boldsymbol{\mu}_1$.

We will mainly use the **Tetrahedron** setting in the paper. Therefore, we further visualize the setting and note down the parameters. The 3D visualization of mixture component centers (those clean tasks) in the prior and the in-context task are shown in Fig. 5. The parameters are noted as follows:

- Dimension $d = 3$, number of mixture components $M = 4$, and $\beta \in \{1, 2, 3, 4\}$;
- The centers of topics shape a tetrahedron as shown in Fig. 5. $\boldsymbol{\mu}_1 = \boldsymbol{w}_1 = [0, 0, -1]^\top$, $\boldsymbol{\mu}_2 = \boldsymbol{w}_2 = [\sqrt{\frac{8}{9}}, 0, \frac{1}{3}]^\top$, $\boldsymbol{\mu}_3 = \boldsymbol{w}_3 = [-\sqrt{\frac{2}{9}}, +\sqrt{\frac{2}{3}}, \frac{1}{3}]^\top$, and $\boldsymbol{\mu}_4 = \boldsymbol{w}_4 = [-\sqrt{\frac{2}{9}}, -\sqrt{\frac{2}{3}}, \frac{1}{3}]^\top$;
- All component's mixture weights are the same, $\pi_\beta = 1/4$, and $\boldsymbol{\mu}_\beta = \boldsymbol{w}_\beta$, for all $\beta \in \{1, 2, 3, 4\}$;
- For noise of $\boldsymbol{x}$ and y, we have $\sigma_x = \sigma_y = 1$, and $\tau_x = 1$;
- For noises of $\boldsymbol{\mu}$ and $\boldsymbol{w}$, we have $\sigma_\mu = \sigma_w = 0.25$ if not specified;
- For in-context task, we have $\boldsymbol{\mu}^* = \frac{2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2+0.2\boldsymbol{\mu}_3}{\|2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2+0.2\boldsymbol{\mu}_3\|}$ and $\boldsymbol{w}^* = \frac{2\boldsymbol{w}_1+\boldsymbol{w}_2+0.2\boldsymbol{w}_3}{\|2\boldsymbol{w}_1+\boldsymbol{w}_2+0.2\boldsymbol{w}_3\|}$. We slightly shift the in-context task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$ towards $(\boldsymbol{\mu}_3, \boldsymbol{w}_3)$ for visualization purposes, to make $\beta = 3$ and $\beta = 4$ produce slightly different curves.

D.2 d-DIMENSIONAL EXAMPLES

We consider d-dimensional examples with d centers for $d \in \{2, 4, 8, 16, 32\}$. A d-dimensional example with d vertices is parametered as follows:

- Dimension $d = d$, number of mixture component $M = d$, and $\beta \in \{1, \dots, d\}$;
- For all β, $\boldsymbol{\mu}_{\beta,i} = \begin{cases} 1 & \text{if } i = \beta \\ 0 & \text{if } i \neq \beta \end{cases}$, *i.e.*, $\boldsymbol{\mu}_\beta$ is a vector with all elements 0 except the βth element is 1.
- All component's mixture weights are the same, $\pi_\beta = 1/d$, and $\boldsymbol{\mu}_\beta = \boldsymbol{w}_\beta$, for all $\beta \in \{1, \dots, d\}$;
- For noise of $\boldsymbol{x}$ and y, we have $\sigma_x = \sigma_y = 1$, and $\tau_x = 1$;
- For noises of $\boldsymbol{\mu}$ and $\boldsymbol{w}$, we have $\sigma_\mu = \sigma_w = 0.25$ if not specified;
- For in-context task, we have $\boldsymbol{\mu}^* = \frac{2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{\|2\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\|}$ and $\boldsymbol{w}^* = \frac{2\boldsymbol{w}_1+\boldsymbol{w}_2}{\|2\boldsymbol{w}_1+\boldsymbol{w}_2\|}$.

D.3 EARLY ASCENT EXAMPLES

Table 2 shows the setting for reproducing the early ascent phenomenon. The in-context task adopts a distribution of x close to a misleading task.

E TRANSFORMER PERFORMANCE IN APPROXIMATING BAYESIAN INFERENCE

We examine if a Transformer pretrained on samples generated from our pretraining data generative model matches the performance of Bayesian inference. We consider three factors of the task prior in our experiment: *prior task noise*, *number of components*, and *feature dimension*. For scalar y, we transform it to a d-dimensional vector $[y, 0, \dots, 0]$. Thus, $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ forms a $(2k+1) \times d$ matrix, comprising $\boldsymbol{x}_{k+1}$ and k pairs of $(\boldsymbol{x}_i, y_i)$.

Experiment Setting. We conduct experiments based on the module GPT2Model from the package Transformers supported by Hugging Face[2]. We use a 10-layer, 8-head Transformer decoder with 1024-dimensional feedforward layers, and the input dimension is set to d, equal to the dimension of $\boldsymbol{x}$. We train the model over three epochs, each consisting of 10,000 batches, with every batch containing 256 samples. We use AdamW (Loshchilov & Hutter, 2019) as the optimizer with weight decay as 0.00001 and set the learning rate to 0.00001.

[2]https://huggingface.co/

Case	Component /Task	Mixture Weight	$\boldsymbol{\mu}$	$\boldsymbol{w}$
d=1	Component 1	1/2	$\boldsymbol{\mu}_1 = [+1]$	$\boldsymbol{w}_1 = [-1]$
	Component 2	1/2	$\boldsymbol{\mu}_2 = [-1]$	$\boldsymbol{w}_2 = [+1]$
	Component 3	/	/	/
	In-context Task	/	$\boldsymbol{\mu}^* = [+1]$	$\boldsymbol{w}^* = [+1]$
d=2	Component 1	1/3	$\boldsymbol{\mu}_1 = [+1, +1]$	$\boldsymbol{w}_1 = [-1, -1]$
	Component 2	1/3	$\boldsymbol{\mu}_2 = [-1, -1]$	$\boldsymbol{w}_2 = [+1, +1]$
	Component 3	1/3	$\boldsymbol{\mu}_3 = [+1, -1]$	$\boldsymbol{w}_3 = [-1, +1]$
	In-context Task	/	$\boldsymbol{\mu}^* = [+1, +1]$	$\boldsymbol{w}^* = [+1, +1]$
d>=2	Component 1	1/3	$\boldsymbol{\mu}_1 = [+1] + [+1] \times (d-1)$	$\boldsymbol{w}_1 = [-1] + [-1] \times (d-1)$
	Component 2	1/3	$\boldsymbol{\mu}_2 = [-1] + [-1] \times (d-1)$	$\boldsymbol{w}_2 = [+1] + [+1] \times (d-1)$
	Component 3	1/3	$\boldsymbol{\mu}_3 = [+1] + [-1] \times (d-1)$	$\boldsymbol{w}_3 = [-1] + [+1] \times (d-1)$
	In-context Task	/	$\boldsymbol{\mu}^* = [+1] \times d$	$\boldsymbol{w}^* = [+1] \times d$

Table 2: In all cases with various dimensions, the pretraining task prior comprises two components for single dimension and three for two or more dimensions. The aim is to predict following the in-context function $\boldsymbol{w}^*$, equivalent to prior center 2's function $\boldsymbol{w}_2$ ($\boldsymbol{w}^* = \boldsymbol{w}_2$). The in-context task is characterized by having a closer x distribution to the task of prior center 1 but having a closer $x \to y$ mapping to the task of prior center 2. The parameters for all cases are set to $\sigma_\mu = \sigma_w = 0.05$, $\sigma_x = 1$, and $\sigma_y = 2$. Refer to Fig. 2(b) for visualization of the prior centers under dimension $d \in \{1, 2, 3\}$.

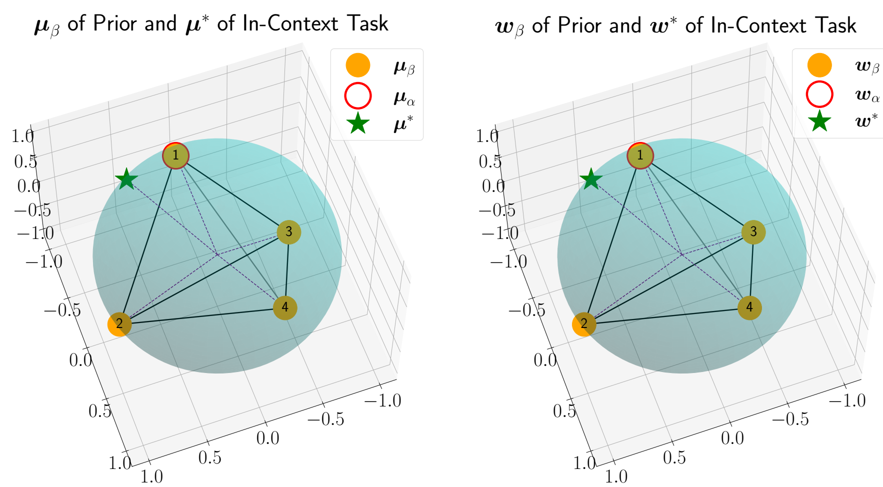


Figure 6: The figure shows the experiment results under varied noise levels. δ_μ and δ_w indicate the noise levels of the pretraining task prior. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer. One can observe that the lower the values of δ_μ and δ_w are, *i.e.*, the noise levels, the stronger the bounded efficacy phenomenon and the harder for the Transformer to approach the Bayesian inference.

Experiment Results. Fig. 6, 7, and 8 show the experimental results, where the prediction of the Transformer is denoted as $\hat{\mathcal{F}}$, and the prediction of Bayesian inference is denoted as $\mathcal{F}^*$. We can observe that the pretrained Transformer model can approximate the Bayes-optimal predictor under varied settings. In Fig. 6, we consider the Tetrahedron setting (see Apendix D.1 for setting details) under varied task noises ($\delta_\mu = \delta_w \in \{1/256, 1/64, 1/16, 1/4, 1\}$). The results show that the lower the task noises, the stronger the bounded efficacy phenomenon in both Bayesian and Transformer inference, and it is also harder, taking more training epochs, for the Transformer to capture the Bayesian prediction. In Fig. 7, we consider settings of regular shapes (see Appendix D.1 for setting details) with different numbers of vertices/components ($M \in \{4, 6, 8, 12, 20\}$). In Fig. 8, we consider settings with varied dimensions (see Appendix D.2 for setting details, $d \in \{2, 4, 8, 16, 32\}$). In Fig. 7 and 8, we observe that the higher the number of dimensions and the number of mixture components, the harder it is for the Transformer to approximate Bayesian prediction.

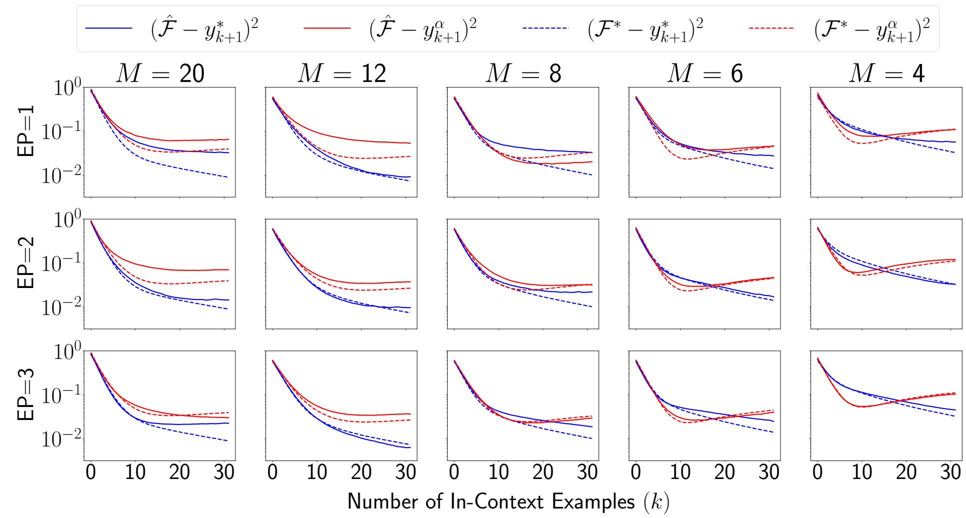


Figure 7: The figure shows the experiment results under varied component densities. M indicates the number of mixture components, and $\delta_\mu = \delta_w = \frac{1}{16}$. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer. It is observed that the higher the component density is, the harder it is for the Transformer to approach the Bayesian inference.

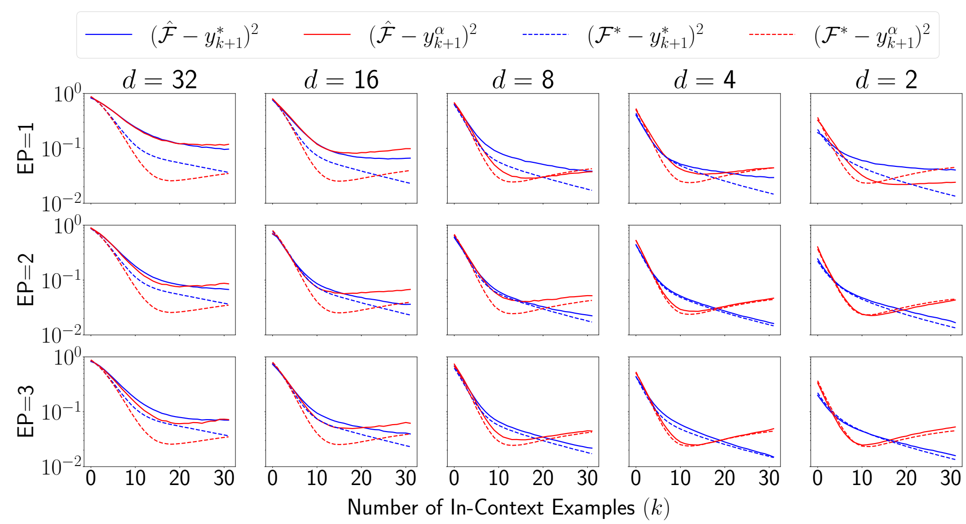


Figure 8: The figure shows the experiment results under varied dimensions. d indicates the dimension and the number of mixture components, and $\delta_\mu = \delta_w = \frac{1}{16}$. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer. It is observed that the higher the number of dimensions is, the harder it is for the Transformer to approach the Bayesian inference.

F Bounded Efficacy of Biased-Label ICL in GPT-4

This section first introduces the experiment design with GPT-4 in Table 3, and then reveals the bounded efficacy phenomenon of GPT-4 in Table. 1.

Table 3 introduces the experiment setting of GPT-4 including the system message, the prompt, the in-context task, the "add-1 addition" task, and the "addition" task. While the in-context task is the "add-1 addition" task, *i.e.*, $c_i = a_i + b_i + 1$, we measure the performances on two goals including learning the "add-1 addition" task and retrieving the "addition" task.

Setting	Description
LLM	GPT-4
System Message	You are a mathematician. Consider the following math problem and follow the exact instruction.
Prompt	You are given examples. Each example has two integers as input and one integer as output. Please provide an answer for the last problems in the math exercise: a_1(?)b_1=c_1 ... a_k(?)b_k=c_2 a_{k+1}(?)b_{k+1}= Provide your answer directly.
In-Context Task	a_i and b_i are uniformly sampled from $[10, 99]$, and $c_i = a_i + b_i + 1$.
Goal of Learning the "Add-1 Addition" Task with True Label	Aiming to learn the "add-1 addition" task, a(?)b=(a+b+1), with in-context examples following the same "add-1 addition" task, a(?)b=(a+b+1).
Goal of Retrieving the "Addition" Task with Biased Label	Aiming to retrieve the "addition" task, a(?)b=(a+b). However, the in-context examples are provided with a slightly different task "add-1 addition", a(?)b=(a+b+1).

Table 3: Experiment setting to reveal the bounded efficacy phenomenon of biased-label ICL in GPT-4.

Table 4: 0 in-context examples, $k = 0$. Prediction is colored red if it is correct for task retrieval ($a(?)b = (a + b)$), and colored blue if it is correct for task learning ($a(?)b = (a + b + 1)$).

prompt	... 51(?)36= ...	... 27(?)15= ...	... 76(?)82= ...	... 55(?)15= ...
results	Without knowing the operation or rule that connects the two input integers to the output integer in the examples, it's impossible to provide a correct answer. Please provide the examples or the rule.	Sorry, but your questionis not clear. Could you please provide more information about the operation between the two numbers?	Your question seems to be missing some information. Could you please provide the examples you mentioned? They are necessary to understand the relationship between the two input integers and the output integer.	70

Table 1 shows the experimental results. As the number of in-context examples increases, we observe the error rate of the "add-1 addition" task constantly decreases while the error rate of the "addition" task initially decreases and then increases, revealing a bounded efficacy phenomenon. We further randomly sample four pairs of prompts and predictions for $k = 0, 2, 8$ in Tables 4, 5, and 6 for references. The results show that ICL with biased labels will initially retrieve a commonsense pretraining task due to the task retrieval mode, and finally learn the in-context task since the task learning effect.

G Bounded Efficacy in Zero-shot ICL

We further introduce Lemma 3, a variation of the previous Theorem 3, to explain zero-shot ICL, an ICL algorithm capable of functioning with random label (Lyu et al., 2023).

Lemma 3 ((informal) Upper Bound for Zero-Shot ICL)**.** *Assume a next-token predictor attains the optimal pretraining risk, the risk of ICL with pure random label (provide no information) will reveal a bounded efficacy phenomenon. See Appendix L for proof details.*

Table 5: 2 in-context examples, $k = 2$. Prediction is colored red if it is correct for task retrieval ($a(?)b = (a + b)$), and colored blue if it is correct for task learning ($a(?)b = (a + b + 1)$).

prompt	... 73(?)80=154 59(?)22=82 54(?)97= ...	... 48(?)73=122 78(?)80=159 21(?)33= ...	... 21(?)28=50 69(?)29=99 47(?)10= ...	... 94(?)43=138 98(?)70=169 96(?)41= ...
results	151	54	57	187

Table 6: 8 in-context examples, $k = 8$. Prediction is colored red if it is correct for task retrieval ($a(?)b = (a + b)$), and colored blue if it is correct for task learning ($a(?)b = (a + b + 1)$).

prompt	... 37(?)70=108 41(?)18=60 19(?)12=32 82(?)67=150 42(?)13=56 26(?)41=68 80(?)39=120 58(?)23=82 40(?)90= ...	... 60(?)76=137 69(?)26=96 72(?)85=158 39(?)10=50 50(?)47=98 19(?)63=83 45(?)95=141 69(?)41=111 81(?)36= ...	... 66(?)40=107 46(?)81=128 63(?)31=95 41(?)24=66 70(?)43=114 89(?)84=174 76(?)82=159 46(?)28=75 49(?)46= ...	... 68(?)88=157 34(?)18=53 70(?)70=141 13(?)35=49 52(?)50=103 72(?)32=105 98(?)82=181 55(?)51=107 50(?)31= ...
results	130	118	96	82

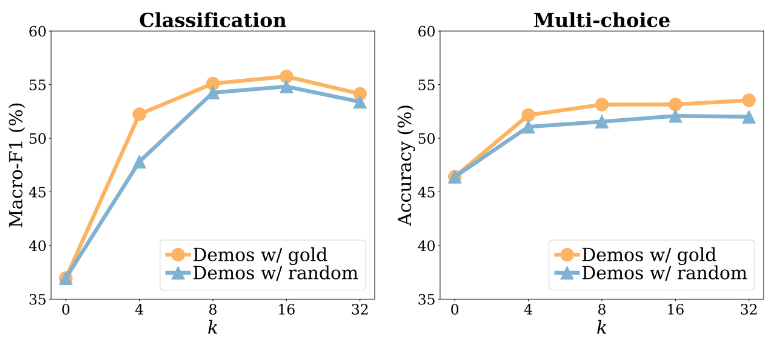


Figure 9: Ablations on varying numbers of examples in the demonstrations (k). Models that are the best under 13B in each task category (Channel MetaICL and Direct GPT-J, respectively) are used.

Lemma 3 says that as the number of in-context examples increases, the loss curve of zero-shot ICL with non-informative labels will have the bounded efficacy phenomenon, which conflicts with the observation from Min et al. (2022) that ICL with random labels has very similar performance as ICL with true labels using $0 \sim 32$ in-context examples. We believe this observation is due to the small number of in-context examples. Thus, we extend the experiment of Min et al. (2022) to explore the number of in-context examples beyond 32. Due to LLMs' context lengths constraining the maximum number of in-context examples, we choose different LLMs from Min et al. (2022) for a larger context length capacity.

Fig. 10 presents the experimental results, highlighting the bounded efficacy phenomenon in the error curve associated with random labels. First, we note that even with true labels, the error rates increase at a larger value of k. (We did not observe this with GPT-4 though.) This is possibly due to LLMs performance degrade when the input contexts become excessively large. However, the error rate of ICL with random labels are observed to increase at a much smaller k value, clearly exhibiting the bounded efficacy phenomenon we predicted.

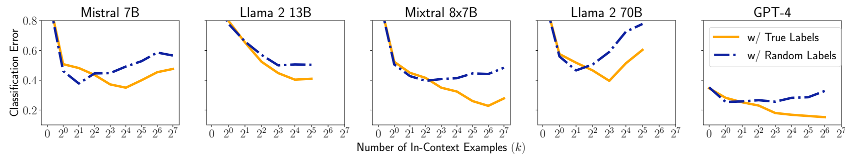


Figure 10: As k increases, the classification error curve of ICL with random labels exhibits the bounded efficacy phenomenon. The curve with true labels further confirms that this phenomenon is not due to models tending to perform worse on long sequences.

We than introduces the experiment setting of Fig. 10. We start by introducing the experiment results in Fig. 9 copied and pasted from the work of Min et al. (2022). While our theory shows the bounded efficacy phenomenon for ICL with non-informative labels (Lemma 3), Fig. 9 seems to imply a conflict phenomenon. Thus, we further extend the number of in-context examples in Fig. 9 left. The classification task adopts five datasets including (i) glue-mrpc (Dolan & Brockett, 2005), (ii) glue-rte (Dagan et al., 2005), (iii) tweet_eval-hate (Barbieri et al., 2020), (iv) sick (Marelli et al., 2014), and (v) poem-sentiment (Sheng & Uthus, 2020). We use the GitHub code[3] released by Min et al. (2022) to generate the same data and evaluate LLMs with a larger context length capacity aiming at a larger number of in-context examples. We selected Mistral 7B (32768) (Jiang et al., 2023), Mixtral 8×7B (32768) (Jiang et al., 2024), Llama2 13B (4096)Llama 2 (Touvron et al., 2023), Llama2 70B (4096) (Touvron et al., 2023), and GPT-4 (8192) (OpenAI, 2023) for our experiments, with the integers in parentheses indicating the maximum context length for each model. We perform inference on large models with 8 H100 with the package vllm[4].

H Mathematical Derivation for Early Ascent Phenomenon

H.1 Finegrained Upper Bound

We first introduce a finegrained upper bound for ICL as follows:

Theorem 6 (Finegrained Upper Bound for ICL)**.** *Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, ICL risk is upper bounded by:*

$$\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^*] < \textstyle\sum_{m=1}^{M} \|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_m \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2],$$

where $\|\boldsymbol{w}_m - \boldsymbol{w}^*\|$ is the distance between $\boldsymbol{w}^*$ and $\boldsymbol{w}_m$, $\tilde{\pi}_m$ is the posterior mixture weight, and $\boldsymbol{A} = (\boldsymbol{I} + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top)^{-1}$. See Appendix K.2 and Eq. 16 for proof details.

Notice that in-context examples affect the upper bound by affecting the two factors $\tilde{\pi}_\beta$ and $\lambda_1(\boldsymbol{A})$, corresponding to the component re-weighting and component shifting introduced in Sec. 4.2. When ignoring the component re-weighting effect and only considering component shifting, the finegrained upper bound with in Theorem 6, degrades to the coarse upper bound in Appendix B Theorem 5.

H.2 Actual Mathematical Derivation

To have a cleaner mathematical understanding of the early ascent phenomenon, this section uses the setting of $d = 1$ in Table. 3 to show the underlying mathematical logits leveraging Theorem 6. Under the setting of $d = 1$ in Table. 3, following Theorem 6, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^{\mathrm{L}}] &< \sum_{\beta=1}^{2} \|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \\
&= \|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_1 \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] + \|\boldsymbol{w}_2 - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_2 \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \\
&= 2^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{r(1,2)}{1 + r(1,2)} \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \\
&= 4 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{r(1,2)}{1 + r(1,2)} \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2].
\end{aligned}$$

[3]https://github.com/Alrope123/rethinking-demonstrations

[4]https://docs.vllm.ai/en/latest/

Noticing $\delta_\mu = \frac{0.05^2}{1^2}, \delta_w = \frac{0.05^2}{2^2}$ is very small, when k is small, we have $k\delta_w \approx 0$ and $\lambda_1(\boldsymbol{A}) = (\boldsymbol{I} + \delta_w \sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1} \approx \boldsymbol{I}$, thus $\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{r(1,2)}{1+r(1,2)}\|\boldsymbol{x}_{k+1}\|^2\lambda_1(\boldsymbol{A})^2] \approx \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{r(1,2)}{1+r(1,2)}\|\boldsymbol{x}_{k+1}\|^2]$ and a larger $r(1,2)$ means a larger upper bound.

Following Eq. 5:

$$\begin{aligned} r(1,2) &= \frac{1/2}{1/2}\exp(\Psi_{\boldsymbol{\mu}}(1,2) + \Psi_{\boldsymbol{w}}(1,2)) \\ &= \exp(\Psi_{\boldsymbol{\mu}}(1,2) + \Psi_{\boldsymbol{w}}(1,2)). \end{aligned}$$

Following Eq. 6:

$$\begin{aligned} \Psi_{\boldsymbol{\mu}}(1,2) &= (\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)/(2\sigma_x^2(1+(k+1)\delta_\mu^2)) \\ &= (\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)/(2\sigma_x^2) \\ &\approx \frac{4k}{2\times 1^2} \\ &= 2k. \end{aligned}$$

Following Eq. 8:

$$\begin{aligned} \Psi_{\boldsymbol{w}}(1,2) &= -\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}/(2\sigma_w^2) \\ &\approx -(\boldsymbol{w}_1 - \boldsymbol{w}^*)^\top k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}(\boldsymbol{w}_1 - \boldsymbol{w}^*)/(2\sigma_w^2) \\ &\text{(Notice } d=1, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^k\|\boldsymbol{x}_i\|^2}{k}) \\ &= -\frac{4\sum_{i=1}^k\|\boldsymbol{x}_i\|^2}{2\sigma_y^2} \\ &\approx -\frac{4k(1+1)}{2\times 2^2} = -k \end{aligned}$$

Therefore, when k is small, $r(1,2) \approx \exp(k)$, and the upper bound is approximately equal to:

$$4\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{\exp(k)}{1+\exp(k)}\|\boldsymbol{x}_{k+1}\|^2],$$

which increases as the number of in-context examples increases.

H.3 Theorem of Early Ascent

Theorem 2 (Early Ascent)**.** *Assume* $\mathbb{E}_{\boldsymbol{x}_1}\left[\frac{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m - \boldsymbol{x}_1\|^2}{2\sigma_x^2})\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m - \boldsymbol{x}_1\|^2}{2\sigma_x^2})}\right] < \mathbb{E}_{\boldsymbol{x}_1}\left[\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\right]$*, where* $\alpha = \arg\min_m \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}^*\|^2}{2\sigma_x^2} + \frac{\|(\boldsymbol{w}_m - \boldsymbol{w}^*)^\top\boldsymbol{\mu}^*\|^2 + d\tau_x^2\|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{2\sigma_y^2}$*. Then, when* δ_μ *and* δ_w *are small enough, we have the early ascent phenomenon on the risk upper bound:*

$$\begin{aligned} \exists k \geq 1 \ s.t.\ \mathbb{E}_{\boldsymbol{x}_1}&\left[\frac{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m - \boldsymbol{x}_1\|^2}{2\sigma_x^2})\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{\sum_{m=1}^M \pi_m \exp(-\frac{\|\boldsymbol{\mu}_m - \boldsymbol{x}_1\|^2}{2\sigma_x^2})}\right] \\ &< \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\frac{\sum_{m=1}^M \tilde{\pi}_m\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{\sum_{m=1}^M \tilde{\pi}_m}\right]. \end{aligned}$$

Proof. We examine the following case, when σ_μ and σ_w are small enough, and k is also big enough to retrieve a task, i.e., making a center dominate:

$$\lim_{k\to\infty}\lim_{(\sigma_\mu,\sigma_w)\to(0,0)}\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\frac{\sum_{m=1}^{M}\tilde{\pi}_m\|\boldsymbol{x}_{k+1}\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\tilde{\pi}_m}\right]$$

$$=\lim_{k\to\infty}\lim_{(\sigma_\mu,\sigma_w)\to(0,0)}\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\frac{\sum_{m=1}^{M}\pi_m\exp(\Psi_{\boldsymbol{\mu}}(m,1)+\Psi_{\boldsymbol{w}}(m,1))\|\boldsymbol{x}_{k+1}\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\pi_m\exp(\Psi_{\boldsymbol{\mu}}(m,1)+\Psi_{\boldsymbol{w}}(m,1))}\right]$$

(Following Eq. 10, we have $\lim_{(\sigma_\mu,\sigma_w)\to(0,0)}\Psi_{\boldsymbol{\mu}}(m,1)+\Psi_{\boldsymbol{w}}(m,1)=\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2-\|\boldsymbol{\mu}_1-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2}$

$$+\sum_{i=1}^{k}\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2-\|\boldsymbol{\mu}_1-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2-\|y_i^1-y_i^*\|^2}{2\sigma_y^2}\right))$$

$$=\lim_{k\to\infty}\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\frac{\sum_{m=1}^{M}\pi_m\exp\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2}+\sum_{i=1}^{k}(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2}{2\sigma_y^2})\right)\|\boldsymbol{x}_{k+1}\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\pi_m\exp\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2}+\sum_{i=1}^{k}(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2}{2\sigma_y^2})\right)}\right]$$

$$=\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2]$$

$$=\mathbb{E}_{\boldsymbol{x}_1}[\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2],$$

where $\alpha=\arg\min_m \frac{\|\boldsymbol{\mu}_m-\boldsymbol{\mu}^*\|^2}{2\sigma_x^2}+\frac{\|(\boldsymbol{w}_m-\boldsymbol{w}^*)^\top\boldsymbol{\mu}^*\|^2+d\tau_x^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{2\sigma_y^2}$. The limitation of limitation indicates that for any $\epsilon>0$, exists a large enough k such that exist small enough δ_μ and δ_w such that $\mathbb{E}[\frac{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)}]>\mathbb{E}[\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2]-\epsilon$. Therefore, we know when $\mathbb{E}[\frac{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)}]<\mathbb{E}[\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2]$ where $\alpha=\arg\min_m \frac{\|\boldsymbol{\mu}_m-\boldsymbol{\mu}^*\|^2}{2\sigma_x^2}+\frac{\|(\boldsymbol{w}_m-\boldsymbol{w}^*)^\top\boldsymbol{\mu}^*\|^2+d\tau_x^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{2\sigma_y^2}$, exists k, σ_μ and σ_w s.t. $\mathbb{E}[\frac{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\pi_m\exp(-\|\boldsymbol{\mu}_m-\boldsymbol{x}_1\|^2)}]<\mathbb{E}[\frac{\sum_{m=1}^{M}\tilde{\pi}_m\|\boldsymbol{x}_1\|^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{\sum_{m=1}^{M}\tilde{\pi}_m}]$. □

I THE DERIVATION OF POSTERIOR

This section provides detailed derivations for Lemma 1. We begin by showing the posterior is potentially still a Gaussian mixture in Sec. I.1. Then in Sec. I.2 we show how Eq. 3 is proportion to Eq. 4, which is exactly still a Gaussian mixture.

I.1 PRIOR TO POSTERIOR

We start by showing the posterior is potentially still a Gaussian mixture:

$$\begin{aligned}
&P(\boldsymbol{\mu},\boldsymbol{w}|\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})\\
\propto&P(\boldsymbol{\mu},\boldsymbol{w}|\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})\\
=&P(\boldsymbol{\mu},\boldsymbol{w},\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})\\
=&P(\boldsymbol{\mu},\boldsymbol{w})P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w})\\
=&(\sum_{\beta=1}^{M}\pi_\beta P(\boldsymbol{\mu},\boldsymbol{w}|T_\beta))P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w})\\
=&\sum_{\beta=1}^{M}\pi_\beta P(\boldsymbol{\mu},\boldsymbol{w}|T_\beta)P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w}) \qquad (3)\\
\propto&\sum_{\beta=1}^{M}\tilde{\pi}_\beta P(\boldsymbol{\mu},\boldsymbol{w}|\widetilde{T}_\beta). \qquad (4)
\end{aligned}$$

We give the derivation from Eq. 3 to Eq. 4 in the next Sec. I.2.

I.2 Closed-form Solution from Eq. 3 to Eq. 4

We analyze each component (indicated by a specific β) in Eq. 3. For all $\beta \in \{1, \dots, M\}$ and all $(\boldsymbol{\mu}, \boldsymbol{w})$, we have:

$$
\begin{aligned}
& P(\boldsymbol{\mu}, \boldsymbol{w}|\widetilde{T}_\beta) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}|\boldsymbol{\mu}, \boldsymbol{w}) \\
\propto & \exp(-\frac{\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}\|^2}{2\sigma_\mu^2}) \exp(-\frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu} - \boldsymbol{x}_i\|^2}{2\sigma_x^2}) \exp(-\frac{\|\boldsymbol{w}_\beta - \boldsymbol{w}\|^2}{2\sigma_w^2}) \exp(-\frac{\sum_{i=1}^{k} \|\boldsymbol{x}_i^\top \boldsymbol{w} - y_i\|^2}{2\sigma_y^2}) \\
& (\text{let } \delta_\mu = \frac{\sigma_\mu^2}{\sigma_x^2}, \delta_w = \frac{\sigma_w^2}{\sigma_y^2}) \\
= & \exp(-\frac{(\|\boldsymbol{\mu}_\beta\|^2 - 2\boldsymbol{\mu}_\beta^\top \boldsymbol{\mu} + \|\boldsymbol{\mu}\|^2) + \delta_\mu((k+1)\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}^\top \sum_{i=1}^{k+1} \boldsymbol{x}_i + \sum_{i=1}^{k+1} \|\boldsymbol{x}_i\|^2)}{2\sigma_\mu^2}) \\
& \exp(-\frac{(\|\boldsymbol{w}_\beta\|^2 - 2\boldsymbol{w}_\beta^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2) + \delta_w(\sum_{i=1}^{k} \boldsymbol{w}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w} - 2\boldsymbol{w}^\top \sum_{i=1}^{k} \boldsymbol{x}_i y_i + \sum_{i=1}^{k+1} y_i^2)}{2\sigma_\mu^2}) \\
\propto & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 + (1 + (k+1)\delta_\mu)\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}(\boldsymbol{\mu}_\beta + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i)}{2\sigma_\mu^2}) \\
& \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 + \boldsymbol{w}^\top(\boldsymbol{I} + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top)\boldsymbol{w} - 2\boldsymbol{w}(\boldsymbol{w}_\beta + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i)}{2\sigma_w^2}) \\
& (\text{let } \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \boldsymbol{I}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}) \\
= & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 + \|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+(k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}} - 2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_\beta + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i)}{2\sigma_\mu^2}) \\
& \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 + \|\boldsymbol{w}\|^2_{\boldsymbol{I}+k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - 2\boldsymbol{w}^\top(\boldsymbol{w}_\beta + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i)}{2\sigma_w^2}) \\
& (\text{let } \bar{\boldsymbol{\mu}} = \sum_{i=1}^{k+1} \boldsymbol{x}_i, \bar{\boldsymbol{w}} = \frac{\sum_{i=1}^{k} \boldsymbol{x}_i y_i}{k}) \\
= & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 + \|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+(k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}} - 2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_\beta + (k+1)\delta_\mu \bar{\boldsymbol{\mu}})}{2\sigma_\mu^2}) \\
& \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 + \|\boldsymbol{w}\|^2_{\boldsymbol{I}+k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - 2\boldsymbol{w}^\top(\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}})}{2\sigma_w^2}) \\
& (\text{Let } \Delta_\mu = (k+1)\delta_\mu, \Delta_w = k\delta_w) \\
= & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 + (\|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+\Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}} - 2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_\beta + \Delta_\mu \bar{\boldsymbol{\mu}}) + \|\boldsymbol{\mu}_\beta + \Delta_\mu \bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}) - \|\boldsymbol{\mu}_\beta + \Delta_\mu \bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}}{2\sigma_w^2}) \\
& \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 + (\|\boldsymbol{w}\|^2_{\boldsymbol{I}+\Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - 2\boldsymbol{w}^\top(\boldsymbol{w}_\beta + \Delta_w \bar{\boldsymbol{w}}) + \|\boldsymbol{w}_\beta + \Delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}) - \|\boldsymbol{w}_\beta + \Delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}) \\
= & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 - \|\boldsymbol{\mu}_\beta + \Delta_\mu \bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}}{2\sigma_\mu^2}) \cdot \exp(-\frac{\|\boldsymbol{\mu} - (\boldsymbol{I} + \Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_\beta + \Delta_\mu \bar{\boldsymbol{\mu}})\|^2_{\boldsymbol{I}+\Delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}}{2\sigma_\mu^2}). \\
& \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 - \|\boldsymbol{w}_\beta + \Delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}) \cdot \exp(-\frac{\|\boldsymbol{w} - (\boldsymbol{I} + \Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta + \Delta_w \bar{\boldsymbol{w}})\|^2_{\boldsymbol{I}+\Delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}}{2\sigma_w^2}) \\
\propto & \exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2 - \|\boldsymbol{\mu}_\beta + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+(k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}}{2\sigma_\mu^2}) \exp(-\frac{\|\boldsymbol{w}_\beta\|^2 - \|\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}).
\end{aligned}
$$

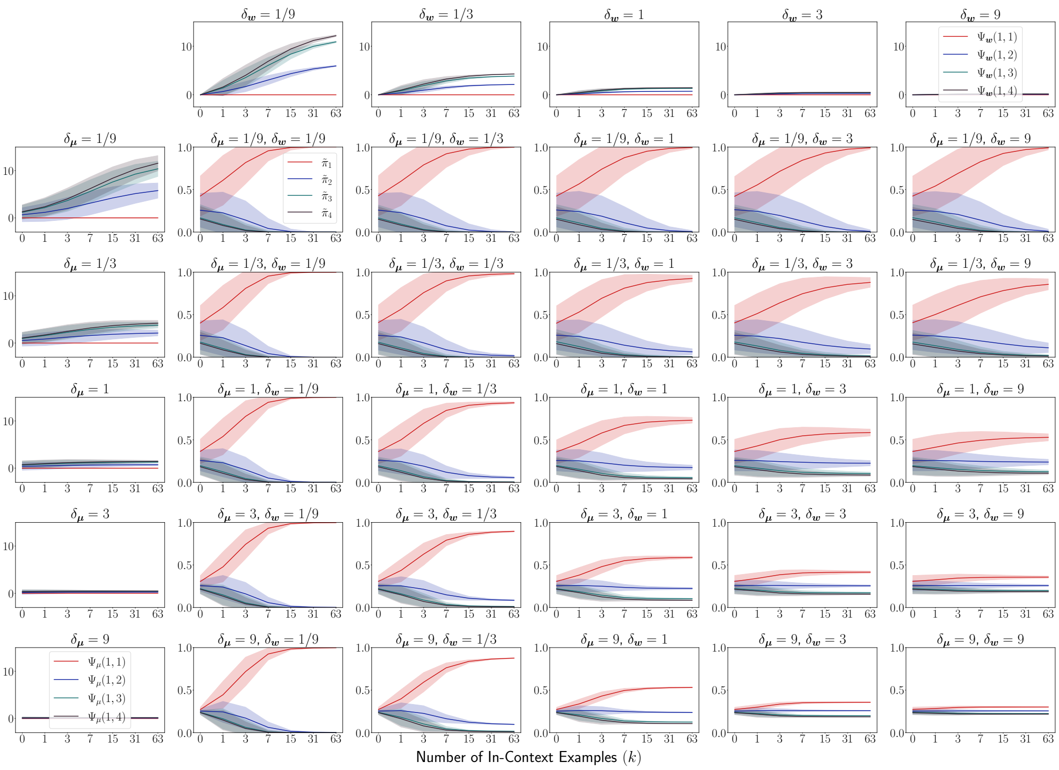


Figure 11: Numerical computations of $\Psi_{\boldsymbol{\mu}}$, $\Psi_{\boldsymbol{w}}$, and π for CR with varying task noise parameters.

$$\mathcal{N}(\boldsymbol{\mu}|(\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_\beta + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}), \sigma_\mu^2(\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}) \cdot$$
$$\mathcal{N}(\boldsymbol{w}|(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}), \sigma_w^2(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})$$

J Detailed Analysis of Component Shifting and Re-weighting

J.1 Analysis of Component Re-weighting

This section analyzes the CR effect on $\tilde{\pi}_\beta$ as k increases. We focus on whether $\tilde{\pi}_\alpha$ of $\widetilde{T}_\alpha$ surpasses $\tilde{\pi}_\beta$ of any other $\widetilde{T}_\beta$ with $\beta \neq \alpha$, where α is the index of the closest clean task to the task which in-context samples follow as Assumption 2. We assess this via the ratio $r(\alpha, \beta)$ of $\tilde{\pi}_\alpha$ to $\tilde{\pi}_\beta$:

$$r(\alpha, \beta) = \frac{\tilde{\pi}_\alpha}{\tilde{\pi}_\beta} = \frac{\pi_\alpha C_1 c_\alpha^{\boldsymbol{\mu}} c_\alpha^{\boldsymbol{w}}}{\pi_\beta C_1 c_\beta^{\boldsymbol{\mu}} c_\beta^{\boldsymbol{w}}} = \frac{\pi_\alpha}{\pi_\beta} \exp(\Psi_{\boldsymbol{\mu}}(\alpha, \beta) + \Psi_{\boldsymbol{w}}(\alpha, \beta)), \tag{5}$$

where we define two functions $\Psi_{\boldsymbol{\mu}}(\alpha, \beta) = \log(c_\alpha^{\boldsymbol{\mu}}/c_\beta^{\boldsymbol{\mu}})$ and $\Psi_{\boldsymbol{w}}(\alpha, \beta) = \log(c_\alpha^{\boldsymbol{w}}/c_\beta^{\boldsymbol{w}})$ to facilitate the analyses of how $r(\alpha, \beta)$ changes with increasing k.

Analysis of $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$. We further simplify the function $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ as follows:

$$\Psi_{\boldsymbol{\mu}}(\alpha, \beta) = \left(\sum_{i=1}^{k+1} \|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1} \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2\right)/(2\sigma_x^2(1 + (k+1)\delta_\mu)). \tag{6}$$

(See Appendix J.4.1 for derivation.) Since $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I})$, a value of $\boldsymbol{\mu}^*$ closer to $\boldsymbol{\mu}_\alpha$, tends to make $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ positive and grow large faster with increasing k. However, as k approaches infinity, $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ stabilizes rather than increasing infinitely, *i.e.*, $\lim_{k\to\infty} \Psi_{\boldsymbol{\mu}}(\alpha, \beta) = (\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2)/(2\sigma_\mu^2)$. The leftmost column of Fig. 11 shows the numerical computation of $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ with varied task noises under the setting "Tetrahedron" (see Appendix D.1 for setting details). The smaller the value of δ_μ ($= \frac{\sigma_\mu^2}{\sigma_x^2}$) is, the easier for $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ to grow large as k increases.

Meanwhile, we also have:

$$\lim_{\sigma_\mu \to 0} \Psi_{\boldsymbol{\mu}}(\alpha, \beta) = (\sum_{i=1}^{k+1} \|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1} \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)/(2\sigma_x^2) \tag{7}$$

Analysis of $\Psi_{\boldsymbol{w}}(\alpha, \beta)$. We further simplify the function $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ as follows:

$$\Psi_{\boldsymbol{w}}(\alpha, \beta) = (\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}})/(2\sigma_w^2). \tag{8}$$

(See Appendix J.4.2 for derivation.) Since $k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ ($= \delta_w \sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top$, see definition of $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ in Lemma 1) is at least semi-positive definite, thus choosing $\boldsymbol{w}^*$ closer to $\boldsymbol{w}_\alpha$ tends to make $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ positive and grow large faster with increasing k. However, as k approaches infinity, $\lim_{k\to\infty} k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \lim_{k\to\infty} k\delta_w \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k} = k\delta_w(\boldsymbol{\mu}^* \boldsymbol{\mu}^{*\top} + \tau_x^2 \boldsymbol{I})$. Thus, $\lim_{k\to\infty} \boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1} = \boldsymbol{I}$ and $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ stabilizes rather than increasing infinitely, *i.e.*, $\lim_{k\to\infty} \Psi_{\boldsymbol{w}}(\alpha, \beta) = (\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2)/(2\sigma_w^2)$. The topmost row of Fig. 11 shows the numerical computation of $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ with varied task noises under setting "Tetrahedron" (see Appendix D.1 for setting details). The smaller the value of δ_w ($= \frac{\sigma_w^2}{\sigma_y^2}$) is, the easier for $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ to grow large as k increases. However, one should be caution that $\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 \geq \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2$ does not necessarily imply $\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} \geq \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}$.

Meanwhile, we also have:

$$\begin{aligned}
\lim_{\sigma_w \to 0} \Psi_{\boldsymbol{w}}(\alpha, \beta) &= (\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}})/(2\sigma_w^2) \\
&= (\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2_{k\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2_{k\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}})/(2\sigma_y^2) \\
&= (\sum_{i=1}^k \|y_i^\beta - y_i^*\|^2 - \sum_{i=1}^k \|y_i^\alpha - y_i^*\|^2)/(2\sigma_y^2),
\end{aligned} \tag{9}$$

where $y_i^\beta = \langle \boldsymbol{x}_i, \boldsymbol{w}_\beta \rangle$, $y_i^\alpha = \langle \boldsymbol{x}_i, \boldsymbol{w}_\alpha \rangle$, and $y_i^* = \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle$.

Therefore, combine Eqs. 7 and 9 and we have:

$$\begin{aligned}
&\lim_{\sigma_\mu, \sigma_w \to 0} \Psi_{\boldsymbol{\mu}}(\alpha, \beta) + \Psi_{\boldsymbol{w}}(\alpha, \beta) \\
&= \frac{\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2} + \sum_{i=1}^k \left(\frac{\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2} + \frac{\|y_i^\beta - y_i^*\|^2 - \|y_i^\alpha - y_i^*\|^2}{2\sigma_y^2} \right)
\end{aligned} \tag{10}$$

Numerical Computations of Component Re-weighting. We have seen how noises σ_μ and σ_w of the task prior affect the values of $\Psi_{\boldsymbol{\mu}}$ and $\Psi_{\boldsymbol{w}}$ with increasing k. We further show the numerical computation of $\tilde{\pi}_\beta$ in the center of Fig. 11. The figure shows that the smaller δ_μ and δ_w are, the larger $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ and $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ will be with increasing k, and the easier for the mixture component $\widetilde{T}_\alpha$ to dominates in the posterior with an increasing number of in-context examples.

J.2 Analysis of Component Shifting

The Component Shifting described in Lemma 1 involves shifting the variables $\tilde{\boldsymbol{\mu}}_\beta$ and $\tilde{\boldsymbol{w}}_\beta$:

$$\tilde{\boldsymbol{\mu}}_\beta = (\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_\beta + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}), \tag{11}$$

$$\tilde{\boldsymbol{w}}_\beta = (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}). \tag{12}$$

The following analyses examine these two variables with increasing k.

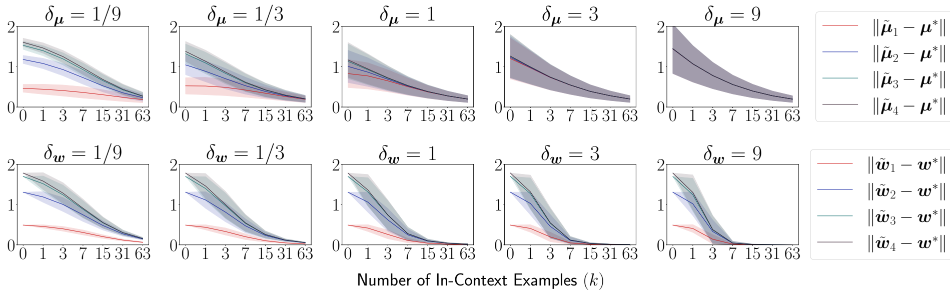


Figure 12: Numerical computations of $\|\tilde{\boldsymbol{\mu}}_\beta - \boldsymbol{\mu}^*\|$, $\|\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}^*\|$ for Component Shifting (CS).

Analysis of $\tilde{\boldsymbol{\mu}}_\beta$. We provide the derivation of $\tilde{\boldsymbol{\mu}}_\beta$ in Eq. 11 (see Appendix J.5.1 for details):

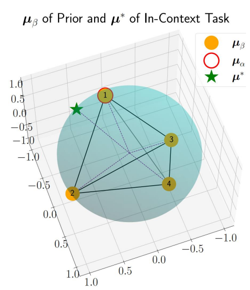
$$\tilde{\boldsymbol{\mu}}_\beta = (\boldsymbol{\mu}_\beta + k\delta_\mu \bar{\boldsymbol{\mu}})/(1 + (k+1)\delta_\mu). \tag{13}$$

Thus, when k increases, $\tilde{\boldsymbol{\mu}}_\beta$ moves close to the value of $\frac{\sum_{i=1}^{k} \boldsymbol{x}_i}{k}$ and $\lim_{k\to\infty} \tilde{\boldsymbol{\mu}}_\beta = \boldsymbol{\mu}^*$. We also show the numerical computation of the distance between shifted $\tilde{\boldsymbol{\mu}}_\beta$ and $\boldsymbol{\mu}^*$ in the first row of Fig. 12.

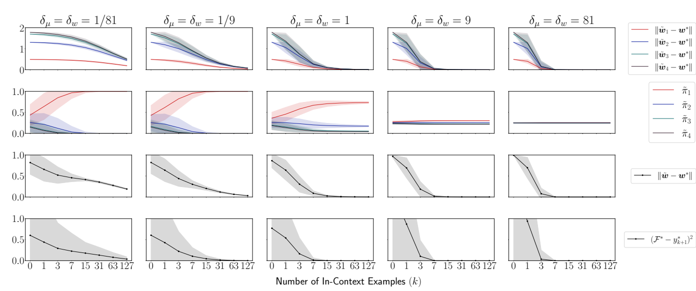
Analysis of $\tilde{\boldsymbol{w}}_\beta$. We provide the derivation of $\tilde{\boldsymbol{w}}_\beta$ in Eq. 12 (see Appendix J.5.2 for details):

$$\tilde{\boldsymbol{w}}_\beta = (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*) + \boldsymbol{w}^*. \tag{14}$$

Notice when $k \to \infty$, $k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = k\delta_w \frac{\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k} \to k\delta_w(\tau_x^2 \boldsymbol{I} + \boldsymbol{w}^* \boldsymbol{w}^{*\top})$, thus $\lambda_d(k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) \to \infty$, and $\lim_{k\to\infty} \tilde{\boldsymbol{w}}_\beta = \boldsymbol{w}^*$, where $\lambda_d(\boldsymbol{A})$ indicates the minimum eigenvalue of $\boldsymbol{A}$. We also show the numerical computation of the distance between shifted $\tilde{\boldsymbol{w}}_\beta$ and $\boldsymbol{w}^*$ in the second row of Fig. 12.



(a) **The Tetrahedron setting.** An illustration of the in-context task and the prior centers. $\forall m \in \{1, 2, 3, 4\}$, We set $\boldsymbol{\mu}_m = \boldsymbol{w}_m$.



(b) **CR, CS, and risks under the Tetrahedron setting.** In the first two rows, we show the effects of CS and CR with an increasing number of in-context examples. In the third row, we show how far the in-context predicted function $\tilde{\boldsymbol{w}}$ is from the target function $\boldsymbol{w}^*$. In the fourth row, we show the ICL risk.

Figure 13: **Numerical experiments.** (left) An illustration of the pretraining priors (right) The numerical computational results

J.3 Prior Task Noises, CS, CR, and ICL Prediction

We numerically compute how $\tilde{\pi}_m$, $\tilde{\boldsymbol{w}}_m$, and the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ evolve with increasing k under different prior task noise conditions. The numerical computation is based on the Tetrahedron setting with four prior mixture components as illustrated in Fig. 13(a). See Appendix D.1 for more setting details. Fig. 13(b) shows the computational results. The first row shows the CS effect, demonstrating the impact of increasing k on $\tilde{\boldsymbol{w}}_m$. The second row shows the CR effect, illustrating the impact of increasing k on $\tilde{\pi}_m$. The third and fourth rows depict how increasing k influences the risk of learning the function $\boldsymbol{w}^*$. We observe that with low task noises and a small k value, the CR effect initially prevails, significantly boosting the mixture weight of component α over others. Then, as k increases further, the CS effect aligns all component centers with $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$.

J.4 Derivation Collection of $\Psi_{\mu}(\alpha,\beta)$ and $\Psi_{w}(\alpha,\beta)$

This section collects derivations for $\Psi_{\boldsymbol{\mu}}(\alpha,\beta)$ and $\Psi_{\boldsymbol{w}}(\alpha,\beta)$. The derivation of $\Psi_{\boldsymbol{\mu}}(\alpha,\beta)$ is collected in Sec J.4.1 and the derivation of $\Psi_{\boldsymbol{w}}(\alpha,\beta)$ is collected in Sec J.4.2.

J.4.1 Derivation of $\Psi_{\mu}(\alpha,\beta)$

This section collects the derivation of $\Psi_{\boldsymbol{\mu}}(\alpha,\beta)$ in Eq. 6 of Sec. J.1:

$$
\begin{aligned}
&\Psi_{\boldsymbol{\mu}}(\alpha,\beta)\\
=&\log(\exp(-\frac{\|\boldsymbol{\mu}_\beta\|^2-\|\boldsymbol{\mu}_\beta+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}}{2\sigma_\mu^2})/\exp(-\frac{\|\boldsymbol{\mu}_\alpha\|^2-\|\boldsymbol{\mu}_\alpha+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}}{2\sigma_\mu^2}))\\
=&\frac{(1+(k+1)\delta_\mu)\|\boldsymbol{\mu}_\beta\|^2-\|\boldsymbol{\mu}_\beta+\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}-\frac{(1+(k+1)\delta_\mu)\|\boldsymbol{\mu}_\alpha\|^2-\|\boldsymbol{\mu}_\alpha+\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}\\
=&\frac{-\|\boldsymbol{\mu}_\beta+\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}-\frac{-\|\boldsymbol{\mu}_\alpha+\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}\\
=&\frac{-\|\boldsymbol{\mu}_\beta\|^2-2\boldsymbol{\mu}_\beta^\top(\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i)-\|\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}-\frac{-\|\boldsymbol{\mu}_\alpha\|^2-2\boldsymbol{\mu}_\alpha^\top(\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i)-\|\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}\\
=&\frac{(k+1)\delta_\mu\|\boldsymbol{\mu}_\beta\|^2-2\boldsymbol{\mu}_\beta^\top(\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i)+\delta_\mu\sum_{i=1}^{k+1}\|\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}-\frac{(k+1)\delta_\mu\|\boldsymbol{\mu}_\alpha\|^2-2\boldsymbol{\mu}_\alpha^\top(\delta_\mu\sum_{i=1}^{k+1}\boldsymbol{x}_i)+\delta_\mu\sum_{i=1}^{k+1}\|\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}\\
=&\frac{\sum_{i=1}^{k+1}\delta_\mu\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}-\frac{\sum_{i=1}^{k+1}\delta_\mu\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_\mu^2(1+(k+1)\delta_\mu)}\\
=&\frac{\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2-\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}.
\end{aligned}
$$

J.4.2 Derivation of $\Psi_{w}(\alpha,\beta)$

This section collects the derivation of $\Psi_{\boldsymbol{\mu}}(\alpha,\beta)$ in Eq. 8 of Sec. J.1:

$$
\begin{aligned}
&\Psi_{\boldsymbol{w}}(\alpha,\beta)\\
=&\log(\exp(-\frac{\|\boldsymbol{w}_\alpha\|^2-\|\boldsymbol{w}_\alpha+k\delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2})/\exp(-\frac{\|\boldsymbol{w}_\beta\|^2-\|\boldsymbol{w}_\beta+k\delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}))\\
=&\frac{\|\boldsymbol{w}_\beta\|^2-\|\boldsymbol{w}_\beta+k\delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}-\frac{\|\boldsymbol{w}_\alpha\|^2-\|\boldsymbol{w}_\alpha+k\delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\\
&(\text{Note } k\delta_w\bar{\boldsymbol{w}}=\delta_w\sum_{i=1}^{k}\boldsymbol{x}_iy_i=\delta_w\sum_{i=1}^{k}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{w}^*=k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\boldsymbol{w}^*)\\
=&\frac{\|\boldsymbol{w}_\beta\|^2-\|\boldsymbol{w}_\beta+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}-\frac{\|\boldsymbol{w}_\alpha\|-\|\boldsymbol{w}_\alpha+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\\
=&\frac{\|\boldsymbol{w}_\beta\|^2-\|(\boldsymbol{w}_\beta-\boldsymbol{w}^*)+(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}-\frac{\|\boldsymbol{w}_\alpha\|^2-\|(\boldsymbol{w}_\alpha-\boldsymbol{w}^*)+(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\\
=&\frac{\|\boldsymbol{w}_\beta\|^2-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}-2(\boldsymbol{w}_\beta-\boldsymbol{w}^*)^\top\boldsymbol{w}^*}{2\sigma_w^2}-\frac{\|\boldsymbol{w}_\alpha\|^2-\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}-2(\boldsymbol{w}_\alpha-\boldsymbol{w}^*)^\top\boldsymbol{w}^*}{2\sigma_w^2}\\
=&\frac{\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}-\frac{\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\\
=&\frac{\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}-\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}
\end{aligned}
$$

J.5 Derivation Collection of $\tilde{\boldsymbol{\mu}}_\beta$ and $\tilde{\boldsymbol{w}}_\beta$

This section collects derivations for $\tilde{\boldsymbol{\mu}}_\beta$ and $\tilde{\boldsymbol{w}}_\beta$. The derivation of $\tilde{\boldsymbol{\mu}}_\beta$ is collected in Appendix J.5.1 and the derivation of $\Psi_{\boldsymbol{w}}$ is collected in Appendix J.5.2.

J.5.1 Derivation of $\tilde{\boldsymbol{\mu}}_\beta$

This section collects the derivation of $\boldsymbol{\mu}_\beta$ in Eq. 13 of Sec. J.1:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_\beta &= (\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_\beta + (k+1)\delta_\mu \bar{\boldsymbol{\mu}}) \\ &= (\boldsymbol{I} + (k+1)\delta_\mu \boldsymbol{I})^{-1}(\boldsymbol{\mu}_\beta + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i) \\ &= \frac{\boldsymbol{\mu}_\beta + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i}{1 + (k+1)\delta_\mu}\end{aligned}$$

J.5.2 Derivation of $\tilde{\boldsymbol{w}}_\beta$

This section collects the derivation of $\boldsymbol{w}_\beta$ in Eq. 14 of Sec. J.1:

$$\begin{aligned}\tilde{\boldsymbol{w}}_\beta &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}) \\ &(\text{recall } k\delta_w \bar{\boldsymbol{w}} = \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i = \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w}^* = k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \boldsymbol{w}^*) \\ &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \boldsymbol{w}^*) \\ &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^* + (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})\boldsymbol{w}^*) \\ &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*) + \boldsymbol{w}^* \end{aligned} \tag{15}$$

K Proof of ICL Bounds

K.1 Proof Tools

We use the following inequalities in our proofs:

K.1.1 Gaussian Tail Bound

If $Z_i \sim \mathcal{N}(0,1)$, then for $t > 0$ we have:

$$P\left(\frac{\sum_{i=1}^k Z_i}{k} > t\right) \leq \exp\left(-\frac{kt^2}{2}\right)$$
$$P\left(\frac{\sum_{i=1}^k Z_i}{k} < -t\right) \leq \exp\left(-\frac{kt^2}{2}\right)$$

K.1.2 Chi-squared Tail Bound

If $X \sim \chi(k)$, *i.e.*, $X = \sum_{i=1}^k Z_i^2$ where $Z_i \sim \mathcal{N}(0,1)$ then:

$$P\left(\frac{X}{k} - 1 > 2\sqrt{t_1} + 2t_1\right) \leq \exp\left(-kt_1^2\right)$$
$$P\left(\frac{X}{k} - 1 < -2\sqrt{t_1}\right) \leq \exp\left(-kt_1^2\right)$$

As a looser but symmetric bound, for $t > 0$, we have:

$$P\left(\frac{X}{k} - 1 > t\right) \leq \exp\left(-\frac{kt^2}{8}\right)$$
$$P\left(\frac{X}{k} - 1 < -t\right) \leq \exp\left(-\frac{kt^2}{8}\right)$$

(See Chi-square Tail Bound.)

K.1.3 Norm Tail Bound

If $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \tau_x^2 \boldsymbol{I})$, $\boldsymbol{\epsilon}_i \in \mathbb{R}^d$, $\boldsymbol{I} \in \mathbb{R}^{d\times d}$, then for $t > 0$ we have:

$$P\left(\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\| > \sqrt{\frac{\tau_x^2 d}{k}(1+t)}\right) \leq \exp\left(-\frac{kt^2}{8}\right)$$

Proof.

$$\begin{aligned}
&\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\|^2 \\
=&\sum_{j=1}^d \left(\frac{\sum_{i=1}^k \epsilon_{i,j}}{k}\right)^2 \\
=&\frac{\tau_x^2}{k}\sum_{j=1}^d \left(\frac{\sum_{i=1}^k \epsilon_{i,j}}{\tau_x\sqrt{k}}\right)^2 \\
&(\text{Notice } \epsilon_{i,j} \sim \mathcal{N}(0, \tau_x^2) \text{ and let } Z_j = \frac{\sum_{i=1}^k \epsilon_{i,j}}{\tau_x\sqrt{k}} \sim \mathcal{N}(0,1)) \\
=&\frac{\tau_x^2 d}{k}\frac{\sum_{i=1}^d Z_i^2}{d}
\end{aligned}$$

therefore by Appendix K.1.2 we have:

$$P\left(\frac{\tau_x^2 d}{k}\frac{\sum_{i=1}^d Z_i^2}{d} > \frac{\tau_x^2 d}{k}(1+t)\right) \leq \exp\left(-\frac{kt^2}{8}\right)$$

□

K.1.4 Eigenvalue Concentration Bound

Lemma 4. *If* $\forall i$, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \tau_x^2 \boldsymbol{I})$, $\boldsymbol{A} = \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$, *and* $\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k} = \frac{\sum_{i=1}^k (\boldsymbol{x}_i - \boldsymbol{\mu})}{k}$, *we have* $\forall t > 0$*:*

$$P\left(L \leq \lambda_d(\boldsymbol{A}) \leq \lambda_1(\boldsymbol{A}) \leq U \text{ and } \|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\| < \tau_x\sqrt{\gamma(1+t)}\right) > 1 - 3\exp\left(-\frac{kt^2}{8}\right)$$

where $L = \tau_x^2(1 - \frac{t}{2} - \gamma)^2 - 2\tau_x\gamma\sqrt{1+t}$, $U = 1 + \tau_x^2(1 + \frac{t}{2} + \gamma)^2 + 2\tau_x\gamma\sqrt{1+t}$ *and* $\lambda_i(\boldsymbol{A})$ *is the* i^{th} *biggest eigenvalue of the matrix* $\boldsymbol{A}$ *and* $\gamma = \sqrt{\frac{d}{k}}$.

We begin with decomposing $\boldsymbol{A}$ to three components $\boldsymbol{A} = \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \frac{\sum_{i=1}^k (\boldsymbol{\mu}\boldsymbol{\epsilon}_i^\top + \boldsymbol{\epsilon}_i \boldsymbol{\mu}^\top)}{k} + \frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^\top}{k}$, where $\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$, then consider the eigenvalues of them.

For the first component, we have:

$$0 \leq \lambda_d(\boldsymbol{\mu}\boldsymbol{\mu}^\top) < \lambda_1(\boldsymbol{\mu}\boldsymbol{\mu}^\top) \leq 1$$

Then, we analyze the second component with Eqs. (8.8) and (8.9) in Covariance Matrix Estimation: Gaussian Data. We have for all $s > 0$:

$$P\left(\left(1-s-\sqrt{\frac{d}{k}}\right)^2 \le \frac{1}{\tau_x^2}\lambda_d\left(\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top}{k}\right) < \frac{1}{\tau_x^2}\lambda_1\left(\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top}{k}\right) \le \left(1+s+\sqrt{\frac{d}{k}}\right)^2\right) > 1-2\exp\left(-\frac{ks^2}{2}\right)$$

Finally for the third component, we examine $\frac{\sum_{i=1}^k(\boldsymbol{\mu}\boldsymbol{\epsilon}_i^\top+\boldsymbol{\epsilon}_i\boldsymbol{\mu}^\top)}{k}$. We have for all $\|\boldsymbol{a}\|=1$:

$$\left\|\boldsymbol{a}^\top\frac{\sum_{i=1}^k(\boldsymbol{\mu}\boldsymbol{\epsilon}_i^\top+\boldsymbol{\epsilon}_i\boldsymbol{\mu}^\top)}{k}\boldsymbol{a}\right\| = 2\left\|\boldsymbol{a}^\top\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\boldsymbol{\mu}^\top\boldsymbol{a}\right\| \le 2\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\|$$

(Notice by Norm Tail Bound in Appendix K.1.3, we have $P\left(\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\| > \sqrt{\frac{\tau_x^2 d}{k}}(1+t)\right) \le \exp\left(-\frac{kt^2}{8}\right)$)

$$\Longrightarrow P\left(-2\sqrt{\frac{\tau_x^2 d}{k}}(1+t) \le \lambda_d\left(2\boldsymbol{\mu}\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i^\top}{k}\right) \le \lambda_1\left(2\boldsymbol{\mu}\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i^\top}{k}\right) \le 2\sqrt{\frac{\tau_x^2 d}{k}}(1+t)\right) > 1-\exp\left(-\frac{kt^2}{8}\right)$$

Let $\gamma = \sqrt{\frac{d}{k}}$, $s = t/2$, and summarize three components we have:

$$P\left(\tau_x^2\left(1-\frac{t}{2}-\gamma\right)^2 - 2\tau_x\gamma\sqrt{1+t} \le \lambda_d(\boldsymbol{A}) \le \lambda_1(\boldsymbol{A}) \le 1+\tau_x^2\left(1+\frac{t}{2}+\gamma\right)^2 + 2\tau_x\gamma\sqrt{1+t}\right) > 1-3\exp\left(-\frac{kt^2}{8}\right)$$

As a summary, we have:

$$P\left(\mathrm{L} \le \lambda_d(\boldsymbol{A}) \le \lambda_1(\boldsymbol{A}) \le \mathrm{U} \text{ and } \|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\| < \tau_x\sqrt{\gamma(1+t)}\right) > 1-3\exp\left(-\frac{kt^2}{8}\right)$$

where $\gamma = \sqrt{\frac{d}{k}}$, $\mathrm{L} = \tau_x^2(1-\frac{t}{2}-\gamma)^2 - 2\tau_x\gamma\sqrt{1+t}$, $\mathrm{U} = 1+\tau_x^2\left(1+\frac{t}{2}+\gamma\right)^2 + 2\tau_x\gamma\sqrt{1+t}$, and $\lambda_i(\boldsymbol{A})$ is the ith biggest eigenvalue of the matrix $\boldsymbol{A}$.

K.2 ICL WITH CORRECT LABELS TO LEARN A TASK

This section introduces the proof of Theorem 5.

Proof. Assuming we are using in-context examples following Assumption 2(a), *i.e.*, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I}), y_i = \langle\boldsymbol{x}_i, \boldsymbol{w}^*\rangle$, and we aim to have the prediction of $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ to be $\langle\boldsymbol{x}_{k+1}, \boldsymbol{w}^*\rangle$, *i.e.*, to learn the function of the in-context task. Let $\mathcal{L}_k^*$ indicate the squared loss $(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle\boldsymbol{x}_{k+1}, \boldsymbol{w}^*\rangle)^2$, where $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ is the prediction of $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ by the Bayes-optimal next-token predictor $\mathcal{F}^*$. We derive the upper bound of the expected squared loss as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\mathcal{L}_k^*] \\
=&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle\boldsymbol{w}^*, \boldsymbol{x}_{k+1}\rangle)^2\right] \\
&\text{(By Corollary 1)} \\
=&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left(\sum_{\beta=1}^M \tilde{\pi}_\beta\langle\tilde{\boldsymbol{w}}_\beta, \boldsymbol{x}_{k+1}\rangle - \langle\boldsymbol{w}^*, \boldsymbol{x}_{k+1}\rangle\right)^2\right] \\
=&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left(\left\langle\sum_{\beta=1}^M \tilde{\pi}_\beta(\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}^*), \boldsymbol{x}_{k+1}\right\rangle\right)^2\right] \\
&\text{(See Eq. 15 for the derivation of } \tilde{\boldsymbol{w}}_\beta\text{)} \\
=&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left(\left\langle\sum_{\beta=1}^M \tilde{\pi}_\beta((\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*) + \boldsymbol{w}^* - \boldsymbol{w}^*), \boldsymbol{x}_{k+1}\right\rangle\right)^2\right] \\
&\text{(Let } \boldsymbol{A} = (\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}\text{, and notice } \boldsymbol{A} \text{ is symmetric positive definite.)}
\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\left(\left\langle \sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta (\boldsymbol{A}(\boldsymbol{w}_\beta - \boldsymbol{w}^*)), \boldsymbol{x}_{k+1} \right\rangle\right)^2\right] \\
&\leq \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \langle (\boldsymbol{A}(\boldsymbol{w}_\beta - \boldsymbol{w}^*)), \boldsymbol{x}_{k+1}\rangle^2] \\
&= \sum\nolimits_{\beta=1}^{M} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta ((\boldsymbol{w}_\beta - \boldsymbol{w}^*)^\top \boldsymbol{A}\boldsymbol{x}_{k+1})^2] \\
&\leq \sum\nolimits_{\beta=1}^{M} \|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \qquad (16)\\
&\leq 4\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \\
&= 4\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2] \\
&\quad \text{(Notice } \boldsymbol{A} \text{ is a random matrix only depends on } \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k, \text{ but not } \boldsymbol{x}_{k+1}.) \\
&= 4\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2]\mathbb{E}_{\mathcal{S}_k}[\lambda_1^2(\boldsymbol{A})] \\
&= 4(1 + d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}[\lambda_1^2(\boldsymbol{A})]
\end{aligned}
$$

We further have the upper bound on the expected squared loss with Lemma 4:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^*] \\
&< 4(1 + d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}[\lambda_1^2(\boldsymbol{A})] \\
&< 4(1 + d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}\left[\left(\frac{1}{1 + k\delta_w \lambda_d(\frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k})}\right)^2\right] \\
&\quad \text{(Apply Lemma 4 to } \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}.) \\
&< 4(1 + d\tau_x^2)\left(\left(\frac{1}{1 + k\delta_w (\tau_x^2(1 - \frac{t}{2} - \gamma)^2 - 2\tau_x\gamma\sqrt{1+t})}\right)^2 + 3\exp\left(-\frac{kt^2}{8}\right)\right)
\end{aligned}
$$

Let $t = k^{\delta - \frac{1}{2}}$, where $\frac{1}{2} > \delta > 0$ and δ is arbitrary small. We have:

$$
\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^*] < \frac{4(1 + d\tau_x^2)}{\tau_x^4 \delta_w^2 k^2} + O(k^{\delta - \frac{5}{2}})
$$

□

We further validate the expected loss with numerical computations in Fig. 14.

K.3 ICL WITH BIASED LABELS TO RETRIEVE A TASK

We start with the Assumption of biased labels:

Assumption 3 (ICL with Biased Labels)**.** *The function $\boldsymbol{w}^*$ of ICL with biased labels is different from the target function $\boldsymbol{w}_\alpha$, i.e., $\boldsymbol{w}^* \neq \boldsymbol{w}_\alpha$ where $\boldsymbol{w}_\alpha$ is a function of a pretraining task prior center. The in-context task is closer to the prior center α compared to all the other prior centers $\beta \neq \alpha$: $\forall \beta \neq \alpha, \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 \geq d_{\boldsymbol{\mu}}^2, \|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq d_{\boldsymbol{w}}^2$, and $\tau_x^2\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - (1 + \tau_x^2)\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq \tau_x^2 u_{\boldsymbol{w}}^2$.*

This section further details the proof of Theorem 3, with Fig.15 serving as a visual guide. The non-asymptotic and asymptotic bound share the same foundational elements in the proof. However, they are different in handling the components marked in pink. Fig. 15 is thus provided to offer a clearer understanding of its overall framework and assist readers in navigating through the proof.

Proof. Assuming we are using in-context examples following Assumption 2(a), *i.e.*, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I}), y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle$, and we aim to have the prediction of $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ to be $\langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle$, *i.e.*, to retrieve the prediction of the clean task α. Let $\mathcal{L}_k^\alpha$ indicate the squared loss $(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) -$

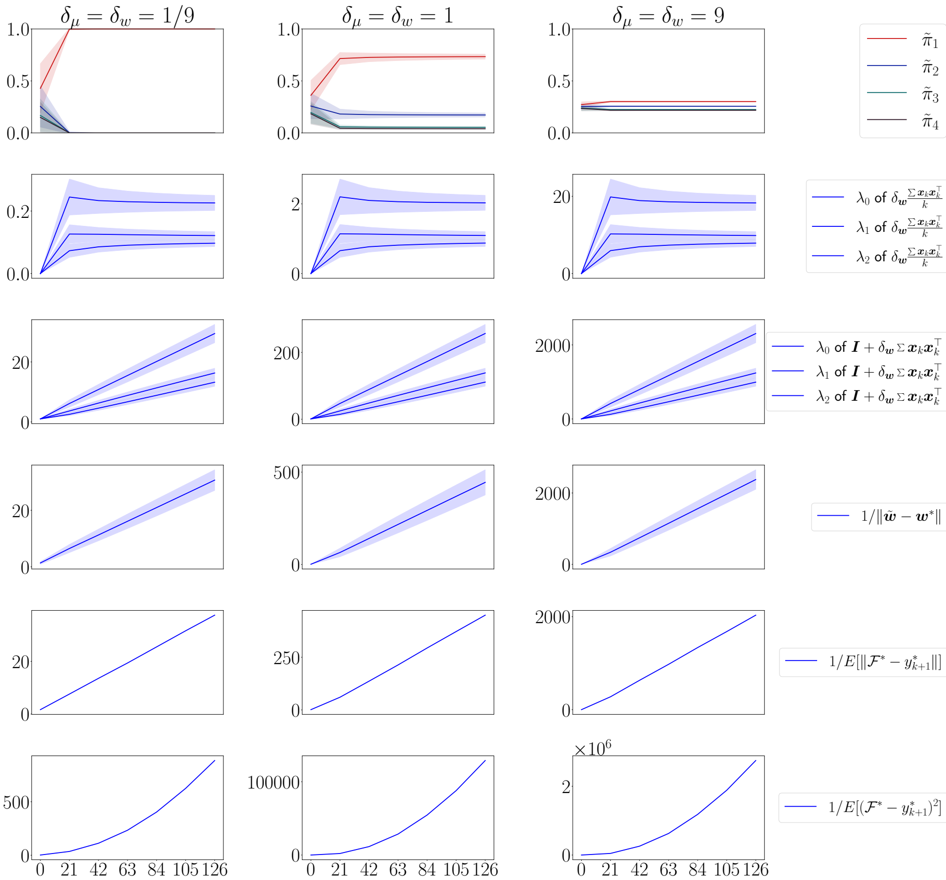


Figure 14: The numerical computation of the task learning. The second and third rows show the eigenvalues of the matrices $\delta_w \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$ and $\boldsymbol{I} + \delta_w \sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top$. The fourth row shows the distance between the predicted $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}^*$ has a reciprocal decreasing rate with respect to k. The fifth and sixth rows indicate the expected squared loss follows a quadratic decreasing rate with respect to k.

$\langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle)^2$, where $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ is the prediction of $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ by the Bayes-optimal next-token predictor $\mathcal{F}^*$. In order to have an upper bound on the loss, we consider $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I})$ in two cases: (1) $\mathbf{C}$: $\mathrm{L} < \lambda_d(\frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}) \leq \lambda_1(\frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}) < \mathrm{U}$ and $\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\| < \tau_x \sqrt{\gamma(1+t)}$ (see Lemma 4 for t, γ, L and U) and (2) $\neg\mathbf{C}$: at least one of the previous inequalities does not hold. (the probability of $\neg\mathbf{C}$ is bounded by: $P(\neg\mathbf{C}) \leq 3\exp(-\frac{kt^2}{8})$).

We start our upper bound analysis on the expected squared loss by splitting the loss into three parts:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha] \\
=&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle)^2] \\
&\text{(By Corollary 1)} \\
=&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\left(\sum_{\beta=1}^M \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta, \boldsymbol{x}_{k+1} \rangle - \langle \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle\right)^2\right] \\
&\text{(Notice } \sum_{\beta=1}^M \pi_\beta = 1\text{)} \\
=&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\left(\sum_{\beta=1}^M \tilde{\pi}_\beta \left(\langle \tilde{\boldsymbol{w}}_\beta, \boldsymbol{x}_{k+1} \rangle - \langle \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle\right)\right)^2\right]
\end{aligned}
$$

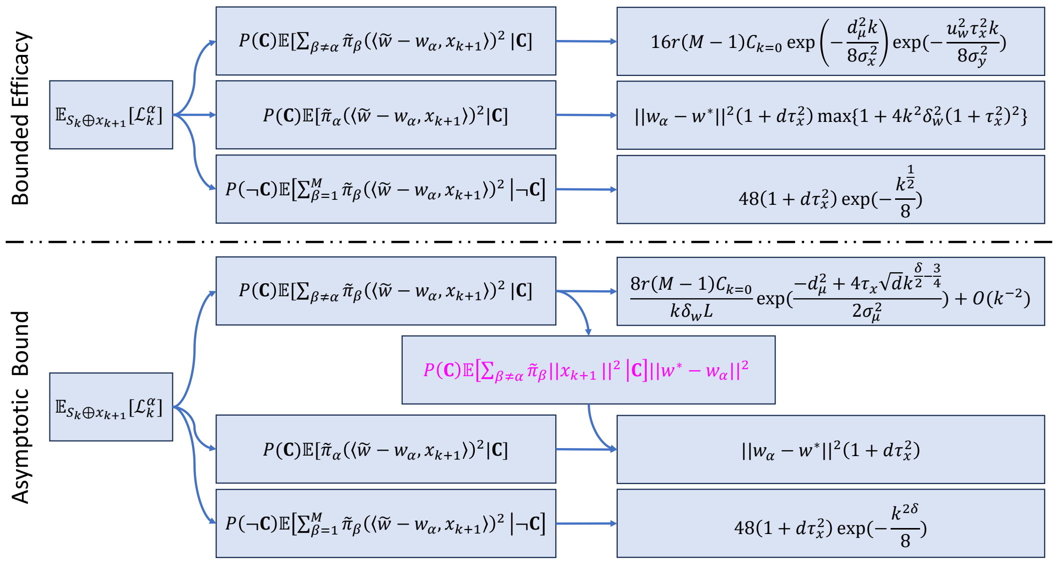


Figure 15: Proof roadmap of ICL with biased labels, Theorem. 3.

$$
\begin{aligned}
&(\text{Notice } (\textstyle\sum_{\beta=1}^{M}\tilde{\pi}_\beta a_\beta)^2 \le \sum_{\beta=1}^{M}\tilde{\pi}_\beta a_\beta^2, \text{ since } \mathbb{E}[a]^2 \le \mathbb{E}[a^2])\\
\le&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta=1}^{M}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta,\boldsymbol{x}_{k+1}\rangle-\langle\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\right]\\
=&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta=1}^{M}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\right]\\
=&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta=1}^{M}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\mathbf{C}\right]+\\
&P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta=1}^{M}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\neg\mathbf{C}\right]\\
=&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\mathbf{C}\right]+ &(17)\\
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2|\mathbf{C}]+ &(18)\\
&P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta=1}^{M}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\neg\mathbf{C}\right] &(19)
\end{aligned}
$$

We will analyze three parts one by one in the following three sections respectively. □

K.3.1 ICL WITH BIASED LABELS - PART 1

Proof. We firstly analyze the term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$ (Part. 17):

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\mathbf{C}\right]\\
<&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\\
&\quad(\text{See Eq. 15 for the derivation of } \tilde{\boldsymbol{w}}_\beta)\\
=&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta-\boldsymbol{w}^*)+\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\\
&(\text{Let } \boldsymbol{A}=(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}, \text{ and } \lambda_1(\boldsymbol{A}) \text{ is the largest eigenvalue of matrix } \boldsymbol{A}))\\
=&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{A}(\boldsymbol{w}_\beta-\boldsymbol{w}^*)+\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]
\end{aligned}
$$

$$
\begin{aligned}
\le & P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \tilde{\pi}_\beta (\|\boldsymbol{A}(\boldsymbol{w}_\beta - \boldsymbol{w}^*)\| + \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|)^2 \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \\
\le & P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 (2\lambda_1(\boldsymbol{A}) + \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|)^2 \middle| \mathbf{C}\right] \\
\le & P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \left(\frac{2}{1+k\delta_w L} + \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|\right)^2 \\
\le & P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha} \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \left(\frac{4}{(1+k\delta_w L)^2} + \frac{8}{1+k\delta_w L}\right) + \\
& P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2
\end{aligned} \tag{20}
$$

The magenta-colored term will be used for the asymptotic bound in section K.3.2 and the bound for the bounded efficacy phenomenon in this section. Apply Eqs. 5, 6, and 8 and Assumption 1(e) to $\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}$, we have:

$$
\begin{aligned}
& P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha} \frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha} \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \\
< & P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum\nolimits_{\beta\neq\alpha} r \exp(\frac{-\sum_{i=1}^k \|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \sum_{i=1}^k \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}) \cdot \\
& \exp(\frac{-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}) \cdot \\
& \exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}) \|\boldsymbol{x}_{k+1}\|^2 [\mathbf{C}] \\
= & rP(\mathbf{C}) \sum\nolimits_{\beta\neq\alpha} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{\sum_{i=1}^k (-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)}{2\sigma_x^2(1+(k+1)\delta_\mu)}) \cdot \\
& \exp(\frac{-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}) \cdot \\
& \exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}) \|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]
\end{aligned}
$$

In the following, we analyze the three-colored terms separately.

Recall in case $\mathbf{C}$ we have:

$$
\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\| < \tau_x \gamma \sqrt{1+t}
$$

Therefore, when conditioned on case $\mathbf{C}$ we have:

$$
\begin{aligned}
& \frac{\sum_{i=1}^k (-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)}{1+(k+1)\delta_\mu} \\
& \text{(Let } \boldsymbol{x}_i = \boldsymbol{\mu}^* + \boldsymbol{\epsilon}_i) \\
= & k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + \frac{\sum_{i=1}^k 2\langle \boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\alpha, \boldsymbol{\epsilon}_i\rangle}{k}}{1+(k+1)\delta_\mu} \\
= & k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + \langle 2(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\alpha), \frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\rangle}{1+(k+1)\delta_\mu} \\
\le & k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + 2\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\alpha\| \|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\|}{1+(k+1)\delta_\mu} \\
\le & k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + 4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu}
\end{aligned}
$$

(Branch to purple for asymptotic bound or to orange for the bound for the U-shaped pattern)

$$
\begin{aligned}
&\text{(Let } t = k^{\delta-\frac{1}{2}} \text{ and } \delta \text{ is arbitrarily small. See Assumption 3 for definition of } d_{\boldsymbol{\mu}}^2\text{)} \\
\leq & -\frac{d_{\boldsymbol{\mu}}^2}{\delta_\mu} + \frac{4\tau_x\sqrt{d}}{\delta_\mu} k^{\frac{\delta}{2}-\frac{3}{4}} + O(k^{-1}) \\
&\text{(let } t = k^{-\frac{1}{4}}, \text{When } \delta_\mu \ll 1, \text{ such that } \exists k \leq \frac{1}{\delta_\mu} - 1, \text{ s.t. } \frac{d_{\boldsymbol{\mu}}^2}{2} > 4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}\text{)} \\
< & -k\frac{d_{\boldsymbol{\mu}}^2}{4}
\end{aligned}
$$

Recall in case **C** we have:

$$
\mathrm{L} < \lambda_d\left(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right) < \lambda_1\left(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right) < \mathrm{U}
$$

Therefore when conditioned on case **C** we also have:

$$
\begin{aligned}
& -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} \\
< & -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\lambda_d(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}) + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\lambda_1(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}) \\
&\text{(where } \lambda_1(\boldsymbol{A}) \text{ and } \lambda_d(\boldsymbol{A}) \text{ indicate the maximal and minimal eigenvalues of the matrix } \boldsymbol{A} \in \mathbb{R}^{d\times d}\text{)} \\
< & -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\left(1 - \frac{1}{1+k\delta_w\mathrm{L}}\right) + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\left(1 - \frac{1}{1+k\delta_w\mathrm{U}}\right) \\
&\text{(Branch to purple for asymptotic bound or to orange for the bound for the U-shaped pattern.)} \\
= & (-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2) + \left(+\frac{\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2}{1+k\delta_w\mathrm{L}} - \frac{\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2}{1+k\delta_w\mathrm{U}}\right) \\
&\text{(Let } t = k^{\delta-\frac{1}{2}} \text{ and } \delta \text{ is arbitrarily small. See Assumption 3 for definition of } d_{\boldsymbol{w}}^2\text{)} \\
= & -(\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2) + \left(\frac{\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2}{k\delta_w\tau_x^2} - \frac{\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2}{k\delta_w(1+\tau_x^2)}\right) + O(k^{\delta-\frac{3}{2}}) \\
< & -d_{\boldsymbol{w}}^2 + \frac{4}{\delta_w\tau_x^2}k^{-1} + O(k^{\delta-\frac{3}{2}}) \\
= & -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{L}}{1+k\delta_w\mathrm{L}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{U}}{1+k\delta_w\mathrm{U}} \\
< & -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{L}}{1+k\delta_w\tau_x^2} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{U}}{1+k\delta_w\tau_x^2} \\
&\text{(Let } t = k^{-\frac{1}{4}}, \text{ when } \delta_w \ll 1, \text{ such that } \exists k \leq \frac{1}{\delta_w\tau_x^2}, \text{ s.t. } \mathrm{L}\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \mathrm{U}\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 > \tau_x^2u_{\boldsymbol{w}}^2/2\text{)} \\
&\text{(See Assumption 3 for definition of } u_{\boldsymbol{w}}^2\text{)} \\
< & -k\delta_w\frac{\tau_x^2u_{\boldsymbol{w}}^2}{4}
\end{aligned}
$$

Further, we have:

$$
\begin{aligned}
& P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}\right] \\
< & \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2\right] \\
&\text{(Let } \boldsymbol{x}_{k+1} = \boldsymbol{\mu}^* + \boldsymbol{\epsilon}\text{)} \\
= & \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^* - \boldsymbol{\epsilon}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^* - \boldsymbol{\epsilon}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2\right] \\
= & \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 + \langle 2(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\alpha), \boldsymbol{\epsilon}\rangle}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2\right] \\
&\text{(Let } -\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 = -D, 2\sigma_x^2(1+(k+1)\delta_\mu) = E, \boldsymbol{b} = 2(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\alpha)\text{)}
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-D+\boldsymbol{b}^\top \boldsymbol{\epsilon}}{E})\|\boldsymbol{x}_{k+1}\|^2] \\
\leq&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-D+\boldsymbol{b}^\top \boldsymbol{\epsilon}}{E})(2\|\boldsymbol{\mu}^*\|^2+2\|\boldsymbol{\epsilon}\|^2)] \\
=&2(\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-D+\boldsymbol{b}^\top \boldsymbol{\epsilon}}{E})]+\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-D+\boldsymbol{b}^\top \boldsymbol{\epsilon}}{E})\|\boldsymbol{\epsilon}\|^2]) \\
=&2(\exp(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2}-\frac{D}{E})+\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-D+\boldsymbol{b}^\top \boldsymbol{\epsilon}}{E})\|\boldsymbol{\epsilon}\|^2]) \\
=&2(\exp(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2}-\frac{D}{E})+\tau_x^2(1+\frac{\tau_x^2\|\boldsymbol{b}\|^2}{E^2})\exp(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2}-\frac{D}{E})+(d-1)\tau_x^2\exp(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2}-\frac{D}{E})) \\
=&2(1+\tau_x^2(d+\frac{\tau_x^2\|\boldsymbol{b}\|^2}{E^2}))\exp(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2}-\frac{D}{E}) \\
=&C_{k=0}
\end{aligned} \tag{21}
$$

Thus, for the asymptotic bound, we have (notice we will not use the magenta-colored term 20 for the asymptotic bound in this section):

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}\right]\left(\frac{4}{(1+k\delta_w L)^2}+\frac{8}{1+k\delta_w L}\right) \\
<&r\sum_{\beta\neq\alpha}P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\sum_{i=1}^k\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\cdot \\
&\exp(\frac{-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2})\cdot \\
&\exp(\frac{-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]\cdot \\
&\left(\frac{4}{(1+k\delta_w L)^2}+\frac{8}{1+k\delta_w L}\right) \\
<&r\sum_{\beta\neq\alpha}\exp(\frac{-\frac{d_{\boldsymbol{\mu}}^2}{\delta_\mu}+\frac{4\tau_x\sqrt{d}}{\delta_\mu}k^{\frac{\delta}{2}-\frac{3}{4}}+O(k^{-1})}{2\sigma_x^2})\exp(\frac{-d_{\boldsymbol{w}}^2+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{\delta-\frac{3}{2}})}{2\sigma_w^2})C_{k=0}\left(\frac{8}{k\delta_w L}+O(k^{-2})\right) \\
=&r(M-1)C_{k=0}\exp(\frac{-d_{\boldsymbol{\mu}}^2+4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}+O(k^{-1})}{2\sigma_\mu^2})\exp(\frac{-d_{\boldsymbol{w}}^2+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{\delta-\frac{3}{2}})}{2\sigma_w^2})\left(\frac{8}{k\delta_w L}+O(k^{-2})\right) \\
=&\frac{8r(M-1)C_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2+4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}+O(k^{-1})}{2\sigma_\mu^2})\exp(\frac{-d_{\boldsymbol{w}}^2+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{\delta-\frac{3}{2}})}{2\sigma_w^2})+O(k^{-2}) \\
=&\frac{8r(M-1)C_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2+4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}}{2\sigma_\mu^2})\exp(\frac{-d_{\boldsymbol{w}}^2}{2\sigma_w^2})+O(k^{-2})
\end{aligned}
$$

Thus, for the U-shaped bound, we have (notice we will use the magenta magenta-colored term for the U-shaped bound in this section):

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}\right]\left(\frac{4}{(1+k\delta_w L)^2}+\frac{8}{1+k\delta_w L}\right)+ \\
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\Big|\mathbf{C}\right]\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|^2 \\
\leq&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}\right]\cdot 16 \\
<&16r\sum_{\beta\neq\alpha}P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\sum_{i=1}^k\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\cdot
\end{aligned}
$$

$$\exp(\frac{-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}).$$
$$\exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$<16r(M-1)C_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2}\right)\exp\left(-\frac{u_{\boldsymbol{w}}^2\tau_x^2 k}{8\sigma_y^2}\right)$$

□

K.3.2 ICL WITH BIASED LABELS - PART 2

Proof. We then deal with the second term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$, the part 18:

$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$$
$$<P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\| -(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})(\boldsymbol{w}_\alpha - \boldsymbol{w}^*) + (\boldsymbol{w}_\alpha - \boldsymbol{w}_\alpha)\|^2\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_K}[\tilde{\pi}_\alpha\lambda_1^2(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$

(Let $\lambda_1(\boldsymbol{A})$ be the maximal eigenvalue of the matrix $\boldsymbol{A}$)

$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\lambda_1^2(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(1-\frac{1}{1+k\delta_w \mathrm{U}})^2\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}](1-\frac{1}{1+k\delta_w \mathrm{U}})^2$$

Thus, for the asymptotic bound, we have (notice we will use the magenta-colored term 20 for the asymptotic bound in this section):

Adding the magenta-colored term 20 in section K.3.1:

$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]+$$
$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2$$
$$=\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}](1-\frac{1}{1+k\delta_w \mathrm{U}})^2$$
$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2$$
$$\leq\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]+$$
$$\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]$$
$$=\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$\leq\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2\right]$$
$$=\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)$$

Thus, for the U-shaped bound, we have (notice we will not use the magenta magenta-colored term for the U-shaped bound in this section):

$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$$
$$\leq\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}](1-\frac{1}{1+k\delta_w \mathrm{U}})^2$$
$$\leq\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2\right](1-\frac{1}{1+k\delta_w \mathrm{U}})^2$$
$$=\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)(1-\frac{1}{1+k\delta_w \mathrm{U}})^2$$

$$(\text{Let } t = k^{-\frac{1}{4}},\ \text{and assuming } \delta_w \ll 1,\ \text{such that } \exists k,\ \text{s.t. } \mathrm{U} < 2(1+\tau_x^2))$$

$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)(\frac{k\delta_w \mathrm{U}}{1+k\delta_w \mathrm{U}})^2$$
$$<\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)\max\{1, 4k^2\delta_w^2(1+\tau_x^2)^2\}$$

□

K.3.3 ICL WITH BIASED LABELS - PART 3

Proof. Finally for the third term $P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_K}[\sum_{\beta=1}^M \tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \neg\mathbf{C}]$, the part 19:

$$P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta=1}^M \tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \neg\mathbf{C}]$$
$$=P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta=1}^M \tilde{\pi}_\beta\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*) + \boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta=1}^M \tilde{\pi}_\beta(2\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*)\|^2 + 2\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2)\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta=1}^M \tilde{\pi}_\beta(2\cdot 4 + 2\cdot 4)\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$=16P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\beta=1}^M \tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$<16P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
(Notice $\mathbf{C}$ is defined on $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$)
$$<16P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2]$$
$$<16(1+d\tau_x^2)P(\neg\mathbf{C})$$
$$<48(1+d\tau_x^2)\exp(-\frac{k^{2\delta}}{8})$$

□

K.3.4 ICL WITH BIASED LABELS - SUMMARY

Proof. Summarizing three terms, we have:

$$\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha]$$
(Branch to purple for asymptotic bound or to orange for the bound for the U-shaped pattern.)
$$<\frac{8r(M-1)C_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}}{2\sigma_\mu^2})\exp(\frac{-d_{\boldsymbol{w}}^2}{2\sigma_w^2}) + O(k^{-2}) +$$
$$\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + 48(1+d\tau_x^2)\exp(-\frac{k^{2\delta}}{8})$$
$$<16r(M-1)C_{k=0}\exp(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2})\exp(-\frac{u_{\boldsymbol{w}}^2\tau_x^2 k}{8\sigma_y^2}) +$$
$$\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)\max\{1, 4k^2\delta_w^2(1+\tau_x^2)^2\} + 48(1+d\tau_x^2)\exp(-\frac{k^{\frac{1}{2}}}{8})$$

The region for the orange formula are:

$$k \le \min\{\frac{1}{\delta_\mu} - 1, \frac{1}{\delta_w\tau_x^2}\}$$
$$4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}) < \frac{d_{\boldsymbol{\mu}}^2}{2}$$
$$\mathrm{L}\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \mathrm{U}\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 > \tau_x^2 u_{\boldsymbol{w}}^2/2$$
$$\mathrm{U} < 2(1+\tau_x^2)$$

□

L Proof of Lemma 3

In this subsection, we introduce the proof of Lemma 3. We first give the full version of the lemma:

Lemma 5 ((informal) Upper Bound for Zero-Shot ICL)**.** *Assume a next-token predictor attains the optimal pretraining risk, and Assumption 1 has only two components α and β, with centers $(\boldsymbol{\mu}_\alpha, \boldsymbol{w}_\alpha) = (-\boldsymbol{\mu}_\beta, -\boldsymbol{w}_\beta)$. When performing ICL with in-context examples following $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^* \mid \tau_x^2 \boldsymbol{I})$ and $y_i = 0$, i.e., y_i has the same preference to prior component α as β, ICL risk is upper bounded by:*

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < \frac{8rC_{k=0}}{k\delta_w L}\exp\left(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}}{2\sigma_\mu^2}\right) + O(k^{-2}) + \\ \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + 48(1+d\tau_x^2)\exp\left(-\frac{k^{2\delta}}{8}\right),$$

where δ is an arbitrarily small positive number and $C_{k=0}$ is a constant depends on the setting (see Eq. 21). When δ_μ and δ_w are sufficiently small, there is a special region for k that:

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < 16rC_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2}\right) + (1+d\tau_x^2)\max\{1, k^2{\delta_w}^2(1+\tau_x^2)^2\} + 12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right).$$

See Appendix L for proof details. We observe that when k is small in this region, the first and third terms dominate and exponential decay, and when k is large, the second term dominates.

The proof techniques are very similar to the proof techniques for task retrieval in Sec. K.3. We are using in-context examples following $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I}), y_i = 0$, *i.e.*, $\boldsymbol{w}^* = \boldsymbol{0}$, and we aim to have the prediction on $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ as $\langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha\rangle$, *i.e.*, to retrieve the prediction of the clean task α. In order to have an upper bound on the loss, we consider $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I})$ in two regions: (1) $\mathbf{C}$: $\mathrm{L} < \lambda_d(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}) \le \lambda_1(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}) < \mathrm{U}$ (see Lemma 4 for L and U) and (2) $\neg\mathbf{C}$: either the previous inequality does not hold. The probability of $\neg\mathbf{C}$ is bounded by:

$$P(\neg\mathbf{C}) < 3\exp\left(-\frac{kt^2}{8}\right).$$

Let $\mathcal{L}_k^{\mathrm{R}}$ indicate the squared loss $(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha\rangle)^2$ on $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$. With the help of Lemma 1 and Corollary 1, we can derive the expected squared loss on the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$, and then based on $\mathbf{C}$ and the target task $\alpha = 1$ (meanwhile we assume another task is indexed $\beta = 2$), we split the expected squared loss into three parts similar to Sec. K.3:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\mathcal{L}_k^{\mathrm{R}}] \\
<&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}] + &(22)\\
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}] + &(23)\\
&P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa(\langle\tilde{\boldsymbol{w}}_\kappa - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \neg\mathbf{C}] &(24)
\end{aligned}$$

L.1 Proof of Lemma 3: Part 1

We firstly analyze the first term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$ in Part. 22. Similar to Sec. K.3, we have:

$$\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta(\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}] \\
\le&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\left(\frac{4}{(1+k\delta_w L)^2} + \frac{8}{1+k\delta_w L}\right) + \\
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2
\end{aligned}$$

The magenta-colored term will not be used for the asymptotic bound for the above term 22 (will be merged with the term 23) and will be used for the U-shaped bound for the above term 22 (will not be merged with the term 23). Apply Eqs. 5, 6, and 8 and Assumption 1(e) to $\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}$, we have a different results from Sec. K.3 since we have $\boldsymbol{w}_\beta = -\boldsymbol{w}_\alpha$ and $\boldsymbol{w}^* = \mathbf{0}$:

$$\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}] \\
<&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[r\exp(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \sum_{i=1}^k\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\cdot \\
&\exp(\frac{-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2})\cdot \\
&\exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}] \\
&(\text{Notice } \boldsymbol{w}^* = \mathbf{0}, \boldsymbol{w}_\beta = -\boldsymbol{w}_\alpha) \\
=&rP(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \sum_{i=1}^k\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\cdot \\
&\exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]
\end{aligned}$$

Same to Sec. K.3, when conditioned on case $\mathbf{C}$, we have:

$$\begin{aligned}
&\frac{\sum_{i=1}^k(-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)}{1+(k+1)\delta_\mu} \\
&(\text{Branch to purple for asymptotic bound or to orange for the bound for the U-shaped pattern.}) \\
&(\text{Let } t = k^{\delta - \frac{1}{2}} \text{ and } \delta \text{ is small.}) \\
<&-\frac{d_{\boldsymbol{\mu}}^2}{\delta_\mu} + \frac{4\tau_x\sqrt{d}}{\delta_\mu}k^{\frac{\delta}{2}-\frac{3}{4}} + O(k^{-1}) \\
&(\text{let } t = k^{-\frac{1}{4}}, \text{When } \delta_\mu \ll 1, \text{ such that } \exists k \leq \frac{1}{\delta_\mu}, \text{ s.t. } \frac{d_{\boldsymbol{\mu}}^2}{2} > 4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}) \\
<&-\frac{d_{\boldsymbol{\mu}}^2}{4}
\end{aligned}$$

Same to Sec. K.3, when conditioned on case $\mathbf{C}$, we have:

$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\exp(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)})\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}] = C_{k=0}$$

As a summary of the above analysis, for the asymptotic bound, we have:

$$\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right]\left(\frac{4}{(1+k\delta_w L)^2} + \frac{8}{1+k\delta_w L}\right) \\
<&r\exp(\frac{-\frac{d_{\boldsymbol{\mu}}^2}{\delta_\mu} + \frac{4\tau_x\sqrt{d}}{\delta_\mu}k^{\frac{\delta}{2}-\frac{3}{4}} + O(k^{-1})}{2\sigma_x^2})C_{k=0}(\frac{8}{k\delta_w L} + O(k^{-2})) \\
=&\frac{8rC_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}} + O(k^{-1})}{2\sigma_\mu^2}) + O(k^{-2}) \\
=&\frac{8rC_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}}{2\sigma_\mu^2}) + O(k^{-2})
\end{aligned}$$

As a summary of the above analysis, for the U-shaped bound, we have:

$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\left(\frac{4}{(1+k\delta_w L)^2}+\frac{8}{1+k\delta_w L}\right)+$$
$$P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2$$
$$<16P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2 \mid \mathbf{C}]$$
$$<16rC_{k=0}\exp(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2})$$

L.2 Proof of Lemma 3: Part 2

The analysis for the second term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \mathbf{C}]$, the part 23 is the same as Sec. K.3.

L.3 Proof of Lemma 3: Part 3

Finally for the third term $P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_K}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa(\langle\tilde{\boldsymbol{w}}_\kappa - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \neg\mathbf{C}]$, the part 24:

$$P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa(\langle\tilde{\boldsymbol{w}}_\kappa - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 \mid \neg\mathbf{C}]$$
$$<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa(2\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\kappa - \boldsymbol{w}^*)\|^2 + 2\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2)\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$

(Recall $\boldsymbol{w}^* = \mathbf{0}$)

$$<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa(2\cdot 1 + 2\cdot 1)\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$=4P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\sum_{\kappa=1}^2 \tilde{\pi}_\kappa\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$
$$<4P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 \mid \neg\mathbf{C}]$$

(Notice $\mathbf{C}$ is defined on $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$)

$$<4P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2]$$
$$<4(1+d\tau_x^2)P(\neg\mathbf{C})$$
$$<12(1+d\tau_x^2)\exp(-\frac{k^{2\delta}}{8})$$

L.4 Proof of Lemma 3: Summary

Similar to Sec. K.3, summarizing three terms, we have:

$$\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^{\mathrm{R}}]$$

(Branch to purple for asymptotic bound or to orange for the bound for the U-shaped pattern.)

$$<\frac{8rC_{k=0}}{k\delta_w L}\exp(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{\frac{\delta}{2}-\frac{3}{4}}}{2\sigma_{\boldsymbol{\mu}}^2}) + O(k^{-2}) +$$
$$\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + 48(1+d\tau_x^2)\exp(-\frac{k^{2\delta}}{8})$$
$$<16rC_{k=0}\exp(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2})\exp(-\frac{u_{\boldsymbol{w}}^2\tau_x^2 k}{8\sigma_y^2}) +$$
$$\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)\max\{1, 4k^2\delta_w^2(1+\tau_x^2)^2\} + 12(1+d\tau_x^2)\exp(-\frac{k^{\frac{1}{2}}}{8})$$
$$=16rC_{k=0}\exp(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2})\exp(-\frac{u_{\boldsymbol{w}}^2\tau_x^2 k}{8\sigma_y^2}) +$$

$$(1 + d\tau_x^2) \max\{1, 4k^2\delta_w^2(1 + \tau_x^2)^2\} + 12(1 + d\tau_x^2) \exp(-\frac{k^{\frac{1}{2}}}{8})$$

M Demo Section as a Warmup

We study how in-context examples affect the prediction of ICL by a pretrained Bayes-optimal next-token predictor and how the pretraining distribution affects this phenomenon. Assume the next-token predictor f is initially pretrained on a dataset distribution to produce the minimum risk minimizer f^*, and then the pretrained f^* is used to predict the next value y of the value x. Instead of directly inference via $f^*(x)$, we consider inference with additional k in-context examples $\{x_i\}_{i=1}^k$ via the format $f^*([x_1, \ldots, x_k, x])$. We aim to theoretically examine the effect of in-context examples $\{x_i\}_{i=1}^k$ on the prediction $f^*([x_1, \ldots, x_k, x])$. While the formal problem setting may involve heavy math, this demo section illustrates the basic phenomenon for better delivering our work.

The following demo subsections are organized as follows. We first introduce the problem setting in Sec. M.1. We then connect ICL with Bayesian inference in Sec. M.2. Further, we introduce the assumptions for the pretraining dataset in Sec. M.3. Finally, we derive a closed-form posterior and introduce two phenomena, "Component Shifting" and "Component Re-weighting" in Sec. M.4.

M.1 Demo: Pretraing Data Generative Modela

ICL involves two important components: the pretraining dataset, and the LM supporting varied input lengths. We assume the LM $f : \cup_{k \in \{0,\ldots,K-1\}} \mathcal{R}^{k\times 1} \to \mathcal{R}^{1\times 1}$ can fit the pretraining distribution exactly with enough data and expressivity. To generate a training sample, we first sample a task μ from underlying task distribution $\mathcal{D}_\mu$, and then we generate values of the sequence from a distribution $\mathcal{D}_x(\mu)$ based on the task $\boldsymbol{\mu}$. The sample generation process is described below:

Assumption 4 (Demo: Pretraining Data Generative Model)**.** *Given a task prior distribution $\mathcal{D}_\mu$, and a conditioned x sampler $\mathcal{D}_x(\mu)$ conditioned on task μ, the process of generating a sequence $S_K = [x_1, x_2, \ldots, x_K]$ with length K follows:*
(a) Sample a task μ from the Prior: $\mu \sim \mathcal{D}_\mu$, and the probability of μ is indicated by $P(\mu)$;
(b) Sample K samples, x from the chosen task: For $i \in \{1, 2, \ldots, K\}$, $x_i \sim \mathcal{D}_x(\mu)$, and the probability of $x_i = x$ is indicated by $P(x|\mu)$;
(c) Define a Sequence S_k: For capital K, $S_K = [x_1, \ldots, x_K]$; and for lowercase k, the sequence of the first k demonstrations of S_K is indicated by $S_k = [x_1, \ldots, x_k]$, e.g., $S_2 = [x_1, x_2]$.

The generation process is related to real-world scenarios via two points: (i) For sampling step 4(a), the LM is trained on varied tasks; (ii) For sampling step 4(b), when one person/agent produces texts for one task, the generated text could be noisy. For instance, given a task such as describing a football game, one person has multiple ways to describe it.

M.2 Demo: Bayes-Optimal Next-Token Predictor

Now we consider training $f(\cdot)$ using sample S_K generated via above generation process 4 via:

$$\mathcal{L}(f) = \mathop{\mathbb{E}}_{S_K}\left[\frac{1}{K}\sum_{k=0}^{K-1}(f(S_k) - x_{k+1})^2\right] = \mathop{\mathbb{E}}_{\mu\sim\mathcal{D}_\mu}\left[\mathop{\mathbb{E}}_{\substack{x_i\sim\mathcal{D}(\mu),\\ i\in\{1,\ldots,K\}}}\left[\frac{1}{K}\sum_{k=0}^{K-1}(f(S_k) - x_{k+1})^2\middle|\mu\right]\right].$$

A highly expressive f can be viewed as K separate models $f_0, \ldots, f_{K-1}$, where f_k takes a sequence of k values as input. Thus when the model f has enough expressivity, the optimization problem $\text{argmin}_f\, \mathcal{L}(f)$ of minimization of the loss function $\mathcal{L}(f)$ could be regarded as K different minimization tasks:

$$f_k^* = \mathop{\text{argmin}}_{f_k} \mathop{\mathbb{E}}_{S_K}[(f(S_k) - x_{k+1})^2], \forall k \in \{0, \ldots, K-1\}.$$

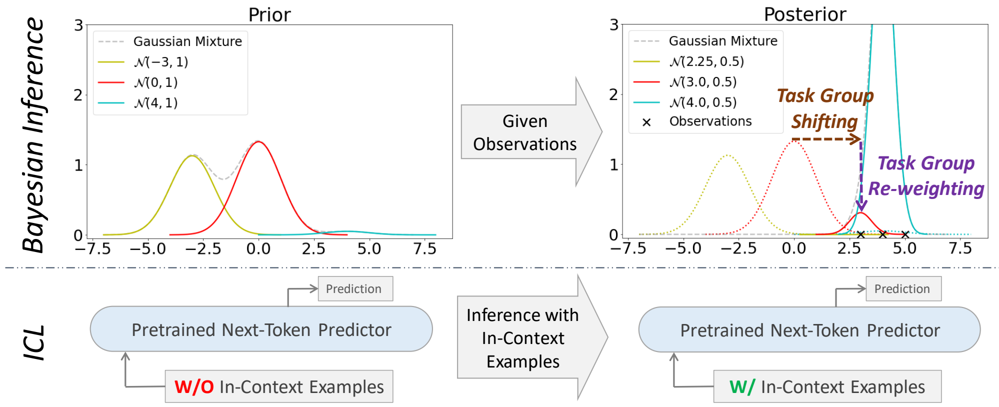


Figure 16: The left part of the figure indicates the pretrained next-token predictor is pretrained on the task prior distribution according to Assumption 5, and the prediction is based on the prior without in-context examples. The right part of the figure indicates that with in-context samples, the prediction is based on posterior, regarding the in-context examples as observed samples.

Thus, the solution f_k^* for each k is a minimum mean square error (MMSE) estimator (Van Trees, 2004, page 63), and the prediction of $f^*(S_k)$ satisfies:

$$f^*(S_k) = \mathop{\mathbb{E}}_{S_K}[x_{k+1}|S_k] = \mathop{\mathbb{E}}_{\mu \sim \mathcal{D}_\mu}[\mathop{\mathbb{E}}_{\substack{x_i \sim \mathcal{D}(\mu),\\ i \in \{1,\dots,K\}}}[x_{k+1}|\mu, S_k]|S_k] = \mathop{\mathbb{E}}_{\mu \sim \mathcal{D}_\mu}[\mathop{\mathbb{E}}_{x_{k+1} \sim \mathcal{D}(\mu)}[x_{k+1}|\mu]|S_k]. \tag{25}$$

The prediction $f^*(S_k)$ is the expectation of $\mathop{\mathbb{E}}_{x_{k+1} \sim \mathcal{D}(\mu)}[x_{k+1}|\mu]$ on the task posterior observing S_k.

M.3 Demo: Gaussian Assumptions on Pretraining Data Generative Model

In Sec. M.2 we connect ICL with Bayesian inference, and in Eq. 25 we observe that the prediction $f^*(S_k)$ depends on the posterior. We are interested in how the in-context examples affect the prediction and the posterior. We make assumptions on the pretraining dataset to have a closed-form expression of the posterior facilitating further analyses:

Assumption 5 (Demo: Gaussian Generative Model for Pretraining Data)**.**
(a) $\mu \sim \mathcal{D}_\mu : P(\mu) = \sum_{\beta=1}^{M} \pi_\beta P(\mu|T_\beta)$*, where* T_β *is the* β^{th} *mixture component of the Gaussian mixture, i.e.,* $P(\mu|T_\beta) = \mathcal{N}(\mu|\mu_\beta, \sigma^2)$*, and* π_β *is the corresponding mixture weight.* $\sum_{\beta=1}^{M} \pi_\beta = 1$*,* $0 < \pi_\beta < 1$*,* μ_β *is the center of the mixture component* T_β*, and all components share the same covariance matrix controlled by* σ*;*
(b) $x \sim \mathcal{D}_x(\mu)$*:* $P(x|\mu) = \mathcal{N}(x|\mu_\beta, \tau^2)$*.*

Under our setting, we train the next-token predictor on M tasks, mirroring real-world LM pretrained on varied topics including environment, market, movie, sports, etc. These tasks have text sequences from diverse sources like individuals, agents, and websites. Given that each source interprets tasks uniquely, they provide "noisy" versions of the same task. We model this using a Gaussian mixture for the task prior. The center of each component represents a specific task, while its variance captures the interpretive noises. Consequently, sequences of values are generated based on these "noisy" tasks.

M.4 Demo: Posterior Analysis

With further Assumption 5 on the prior, we can derive closed-form expression on the posterior:

$$P(\mu|S_k) \propto \sum_{m=1}^{M} \tilde{\pi}_\beta \mathcal{N}(\mu|\tilde{\mu}_\beta, \tilde{\sigma}^2) \tag{26}$$

$$(\tilde{\pi}_\beta = \pi_\beta \exp(\frac{(\mu_\beta - \frac{\sum_{i=1}^k x_i}{k})^2}{2(\tau^2 + k\sigma^2)}), \tilde{\mu}_\beta = \frac{\tau^2\mu_\beta + \sigma^2 \sum_{i=1}^k x_i}{\tau^2 + k\sigma^2})^2, \tilde{\sigma}^2 = \frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2})$$

From Eq. 26, we observe two factors when comparing the posterior with the prior in Assumption 5: (i) Component-Shifting: after observing $S_k = [x_1, x_2, \ldots, x_k]$, the center of each mixture component is shifted to $\frac{\tau^2\mu_\beta + \sigma^2 \sum_{i=1}^k x_i}{\tau^2 + k\sigma^2}$; (ii) Component Re-weighting: the mixture weight π_β of each mixture component is re-weighted by multiplying $\exp(\frac{(\mu_\beta - \frac{\sum_{i=1}^k x_i}{k})^2}{2(\tau^2 + k\sigma^2)})$ (which needs to be further normalized so that re-weighted mixture weights sum to 1). Fig. 16 illustrates the phenomena of Component Shifting and Component Re-weighting by observing in-context examples.

N Proof of Posterior Derivation in Demo

In this section, we give a detailed derivation of the posterior in Eq. 26 in Sec. M.4:

$$\begin{aligned}
P(\mu \mid S_k) &\propto P(\mu, S_k) \\
&= P(S_k \mid \mu) P(\mu) \\
&= (\Pi_{i=1}^k P(x_i \mid \mu)) P(\mu) \\
&= (\Pi_{i=1}^k \mathcal{N}(x_i \mid \mu, \tau^2)) \sum_{m=1}^M \pi_\beta \mathcal{N}(\mu \mid \mu_\beta, \sigma^2) \\
&\propto (\Pi_{i=1}^k \exp(-\frac{(x_i - \mu)^2}{2\tau^2})) \sum_{m=1}^M \pi_\beta \exp(-\frac{(\mu - \mu_\beta)^2}{2\sigma^2}) \\
&= \exp(-\frac{\sum_{i=1}^k (x_i - \mu)^2}{2\tau^2}) \sum_{m=1}^M \pi_\beta \exp(-\frac{(\mu - \mu_\beta)^2}{2\sigma^2}) \\
&= \sum_{\beta=1}^M \pi_\beta \exp(-\frac{\tau^2(\mu - \mu_\beta)^2 + \sigma^2 \sum_{i=1}^k (x_i - \mu)^2}{2\tau^2\sigma^2}) \\
&= \sum_{m=1}^M \pi_\beta \exp(-\frac{\mu^2(\tau^2 + k\sigma^2) - 2\mu(\tau^2\mu_\beta + \sigma^2 \sum x_i) + (\tau^2\mu_\beta^2 + \sigma^2 \sum x_i^2)}{2\tau^2\sigma^2}) \\
&= \sum_{m=1}^M \pi_\beta \exp(-\frac{(\mu - \frac{\tau^2\mu_\beta + \sigma^2 \sum x_i}{\tau^2 + k\sigma^2})^2 + \frac{\tau^2\mu_\beta^2 + \sigma^2 \sum x_i^2}{\tau^2 + k\sigma^2} - (\frac{\tau^2\mu_\beta + \sigma^2 \sum x_i}{\tau^2 + k\sigma^2})^2}{2\frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2}}) \\
&\propto \sum_{m=1}^M \pi_\beta \exp(\frac{(\mu_\beta - \frac{\sum_{i=1}^k x_i}{k})^2}{2(\tau^2 + k\sigma^2)}) \exp(-\frac{(\mu - \frac{\tau^2\mu_\beta + \sigma^2 \sum_{i=1}^k x_i}{\tau^2 + k\sigma^2})^2}{2\frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2}}) \\
&\propto \sum_{m=1}^M \tilde{\pi}_\beta \mathcal{N}(\mu \mid \tilde{\mu}_\beta, \tilde{\sigma}^2)
\end{aligned}$$

$$(\tilde{\pi}_\beta = \pi_\beta \exp(\frac{(\mu_\beta - \frac{\sum_{i=1}^k x_i}{k})^2}{2(\tau^2 + k\sigma^2)}), \tilde{\mu}_\beta = \frac{\tau^2\mu_\beta + \sigma^2 \sum_{i=1}^k x_i}{\tau^2 + k\sigma^2})^2, \tilde{\sigma}^2 = \frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2})$$