Dual Operating Modes of In-Context Learning

Ziqian Lin [1] Kangwook Lee [2]

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 a generalized probabilistic model for pretraining data, obtaining a quantitative understanding of the two operating modes of ICL. Leveraging our analysis, we provide the first explanation of an unexplained phenomenon observed with real-world large language models (LLMs). Under some settings, the ICL risk initially increases and then decreases with more in-context examples. Our analysis offers 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, which will eventually diminish as task learning takes effect with more in-context samples. 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 Transformers and LLMs. The code is available at: `https://github.com/UW-Madison-Lee-Lab/Dual_Operating_Modes_of_ICL`.

1. Introduction

Large language models (LLMs) exhibit a significant improvement in predictive performance when provided with in-context examples (Brown et al., 2020). This emergent ability of LLMs, known as in-context learning (ICL), **operates in two distinct modes: task learning and task retrieval** (Pan et al., 2023). Large language models exemplify this duality. 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 and utilize a *pretrained* skill. A clear evidence of the task retrieval mode is presented by Min et al. (2022), where the authors show ICL performance remains largely unaffected even when in-context examples are annotated with random labels. This suggests that LLMs simply retrieve a pretrained skill rather than learn it from in-context examples.

The dual nature of ICL can be explained as follows. LLMs are a next-token predictor that is pretrained on a large pretraining set, consisting of diverse data from diverse domains/tasks. To predict the next token optimally in such a scenario, the model must first learn the task prior from pretraining data and then implicitly perform Bayesian inference at the test time (Xie et al., 2022; Raventos et al., 2023). Optimal prediction on multitask pretraining data requires adherence to the learned prior (over the tasks present in the pretraining data) and making predictions based on the posterior. The ability to learn and apply this prior during test-time inference enables task retrieval–if in-context examples align closely with a task encountered during pretraining, the model can swiftly adjust its posterior and predict without learning a new skill. Simultaneously, the model can learn a novel or uncommon skill given sufficient in-context samples and a non-zero prior probability for that skill.

Although the link between pretraining and ICL's dual modes is conceptually straightforward, formally establishing this connection is an unresolved challenge. 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 under which the retrieval mode is a dominant one and vice versa?*

A New Model for Pretraining Data To find the answers to these questions, we first propose a new probabilistic model for pretraining data by assuming the pretraining data has a latent clustered structure. In particular, we consider in-context learning of linear functions following the recent

[1] Department of Computer Science, University of Wisconsin-Madison, Madison, Wisconsin, USA [2] Department of Electrical & Computer Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA. Correspondence to: Kangwook Lee <kangwook.lee@wisc.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

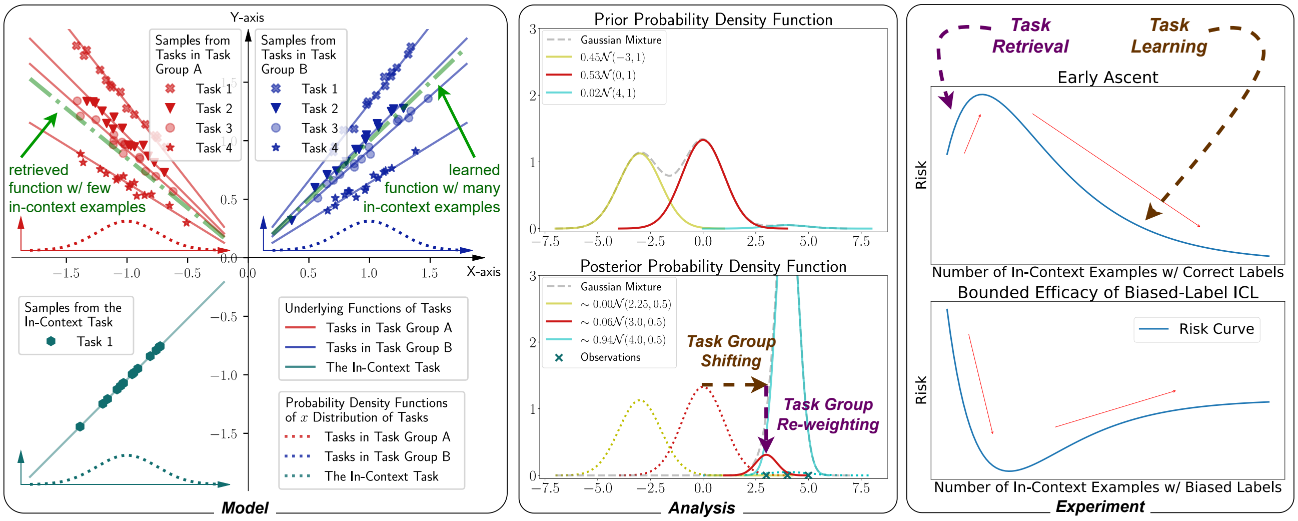


Figure 1: **A summary of our contributions.** We first propose a probabilistic model for pretraining data and in-context examples. By analyzing our model, we obtain a quantitative understanding of the dual operating modes of ICL, and explain two real-world phenomena observed with LLMs.

work (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). A next-token prediction model is prompted with (1) a sequence of $(\boldsymbol{x}, y)$ pairs, which come from a common linear function, and (2) one test input $\boldsymbol{x}_{\text{test}}$. An ideal model capable of in-context learning linear models should internally fit a linear function (say $y = \widehat{\boldsymbol{w}}^T \boldsymbol{x}$) using the in-context examples and then generate the predicted label $y_{\text{test}} = \widehat{\boldsymbol{w}}^T \boldsymbol{x}_{\text{test}}$ as the next token. The recent work (Raventos et al., 2023; Wu et al., 2024) show that such in-context learning is feasible by training a next-token prediction model on a large pretraining dataset, consisting of sequences of labeled samples drawn from diverse linear functions.

We extend the existing model for pretraining data (Raventos et al., 2023) by introducing multiple task groups and task-dependent input distributions. When one generates pretraining data, one must specify a probability distribution of linear functions (equivalently, that of the linear coefficient $\boldsymbol{w}$). While most of the prior work assumes that $\boldsymbol{w}$ is drawn from a single Gaussian distribution, we will model it as drawn from a Gaussian *mixture* model, where each Gaussian component models a *task group*. This model better reflects real-world data that exhibits a clustered structure (Xie et al., 2022). Furthermore, we also allow each mixture component to have its own distribution for input $\boldsymbol{x}$. Shown on the left-most panel in Fig. 1 is a simple visualization of our model. The blue task group is modeled as the distribution of linear functions with positive coefficients ($\boldsymbol{w} \approx 1$) with the input distribution centered around $\mathbb{E}[\boldsymbol{x}] = +1$. The red lines represent the other task group – linear functions with negative coefficients ($\boldsymbol{w} \approx -1$) with the input distribution centered at $\mathbb{E}[\boldsymbol{x}] = -1$. See Sec. 3 for more details.

Analysis With our new model for pretraining data, we analyze the optimal pretrained model under the squared loss, *i.e.*, the MMSE estimator of the label given input with in-context examples. Here, the pretraining distribution (of linear functions) is the prior, and in-context examples are the observations. Leveraging the fact that the Gaussian mixture is a conjugate prior to the Gaussian likelihood function, we obtain a closed-form expression of the posterior distribution. By fully quantifying the posterior distribution of $\boldsymbol{w}$ in the form of a Gaussian mixture, we characterize how in-context examples are used to update each component's posterior mean and posterior mixture probability. We will call updates of mixture probabilities as *task group (component) re-weighting* and updates of component means 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. In particular, we show that, under some mild assumptions, task group re-weighting is the dominant factor when provided with few in-context samples, rendering the task retrieval mode. With many in-context samples, task group shifting occurs, resulting in the task learning mode.

Explanation of Two Real-World Phenomena To demonstrate the practical value of the new insights we have gained from our model, we will leverage our analysis to explain and predict two phenomena observed with LLMs in practice.

- **The *early ascent* phenomenon** refers to the observation that, under certain conditions, the ICL risk initially increases and then decreases when more in-context examples are introduced (Brown et al., 2020; Xie et al., 2022). See the right-most panel of Fig. 1 for visualization. Based on our analysis, 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, which will eventually diminish as task learning takes effect with more in-context samples.

- **Bounded efficacy of biased-label ICL** is predicted by our model. ICL performs well even with in-context examples that are annotated with biased labels (Lyu et al., 2023; Min et al., 2022). Our model provides a rigorous justification of this approach: If in-context examples with biased labels carry sufficient information for retrieving a correct pretrained task, then this approach would work. At the same time, our analysis suggests that the operating mode of ICL will make a transition from task retrieval to task learning with more in-context examples. When the learning mode starts taking place, the test risks of such methods will start increasing as the pretrained model will start fitting the biased labels. See the right-most panel of Fig. 1 for visualization. This bounded efficacy has not been reported in the literature (Min et al., 2022; Pan et al., 2023). We found that this was due to the small number of examples tested. With more in-context samples, we observe the predicted bounded efficacy phenomenon with real-world LLMs such as Mistral 7B (Jiang et al., 2023), Mixtral 8×7B (Jiang et al., 2024), Llama 2 (Touvron et al., 2023), and GPT-4 (OpenAI, 2023).

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. To illustrate, in the context of sentence sentiment classification using ICL, Pan et al. (2023) explore three labeling schemes for in-context examples: (i) correct semantic labels, (ii) correct but abstract labels ("0" and "1"), and (iii) random semantic labels ("positive" or "negative"). Pan et al. (2023) claim that ICL is in the task recognition mode when the model is provided with randomly labeled in-context data, and observe that its efficacy does not correlate with model size or the quantity of demonstrations. In fact, later, we will show that via our analysis, an increasing number of demonstrations will eventually decrease the ICL accuracy. Conversely, ICL with correct but abstract labels, classified as task learning, shows improved performance in proportion to model size and in-context example count. ICL with correct labels yields the highest accuracy since both task recognition and task learning benefit it.

Explaining ICL via Bayesian Inference. Xie et al. (2022) use a Hidden Markov Model (HMM) (Ghahramani & Jordan, 1995; Rabiner, 1989) to model the pretraining data. That is, each sequence in pretraining data is generated by an HMM, whose parameters are randomly drawn from a particular distribution. During pretraining, a next-token prediction model is trained to predict tokens in pretraining sequences, which requires the inference of the latent HMM parameters. While this model accurately reflects real-world pretraining data characteristics, such as long-range dependencies, the absence of a closed-form solution for optimal prediction makes detailed analysis of ICL infeasible. 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 consisting of $(\boldsymbol{x}, y)$ pairs in the form of $(\boldsymbol{x}_1, y_1, \boldsymbol{x}_2, y_2, \ldots)$. The pretraining objective is to predict only the tokens at odd positions, *i.e.*, to predict y, but not $\boldsymbol{x}$. Garg et al. (2022) empirically evaluate the Transformer architecture (Vaswani et al., 2017), while the authors of Raventos et al. (2023) proposed a probabilistic model to generate sequences according to noisy linear regression. More specifically, $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle + \epsilon_i$, where $\boldsymbol{w}^*$ is the coefficient shared within the same sequence and ϵ_i is noise. While this linear regression model facilitates a tractable analysis and elucidates certain aspects of the dual operating modes of ICL, it falls short in modeling the clustered characteristic of nature language. Han et al. (2023) show that ICL asymptotically approaches kernel regression as the in-context samples increases. Jeon et al. (2024) introduce information-theoretic tools to show that the ICL risk should decay in both the number and sequence lengths of in-context examples. On the other hand, our proposed model allows for tractable analysis and captures the clustered characteristic of pretraining data.

Explaining ICL via Gradient Descent. Garg et al. (2022) hint that the pretrained Transformer might implicitly execute gradient descent under ICL. Akyürek et al. (2023); von Oswald et al. (2023); Dai et al. (2023) expand this notion by theoretically showing that one attention layer can be exactly constructed to perform gradient descent, and empirically finding similarities between in-context inference and gradient descent algorithm. Further, Ahn et al. (2023); Mahankali et al. (2024); Zhang et al. (2023) dive into the training process of Transformers. Ahn et al. (2023); Mahankali et al. (2024) theoretically show that under certain conditions, Transformers with one or more attention layers trained on noisy linear regression task minimizing the pretraining loss will implement gradient descent algorithm. Zhang et al. (2023) show that a single linear self-attention layer trained by gradient flow with a suitable random initialization finds a global minimum of the objective function, where ICL of the Transformer achieves prediction error competitive with the best linear predictor.

Others. Wu et al. (2024) studies the sample complexity required for pretraining a linear attention model and presents a statistical bound. In our work, we do not consider a particular model architecture nor the statistical aspects of pretraining – we assume a pretrained model is optimally trained on infinitely large pretraining data, similar to the previous work (Xie et al., 2022; Raventos et al., 2023; Han et al., 2023). Giannou et al. (2023) show a looped Transformer can emulate any algorithms, such as SGD. Bai et al. (2023) show Transformers can perform in-context algorithm selection, *i.e.*, adaptively selecting different ICL algorithms such as gradient descent, least square, or ridge regression. (Li et al., 2023) study the generalization bounds for ICL with Transformers.

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. Consider pretraining this model on sequences consisting of $(\boldsymbol{x}, y)$[1] pairs in the form of $(\boldsymbol{x}_1, y_1, \boldsymbol{x}_2, y_2, \ldots)$, with the model trained to predict only the y values, thereby skipping the prediction of x. Here, we assume odd-numbered tokens represent d-dimension real-valued vectors, and even-numbered tokens represent scalars. During inference, the model receives a sequence of $2k+1$ tokens. The first $2k$ tokens are k labeled samples $(\boldsymbol{x}_i, y_i), i \in \{1, \ldots, k\} =: [K]$, and the last token is unlabeled $\boldsymbol{x}_{k+1}$. Ideally, the model should predict the correct next token, y_{k+1}.

3.1. Data Generative Model

In the pretraining phase, we assume the next-token predictor is pretrained on diverse tasks, each representing a continuous joint distribution of $(\boldsymbol{x}, y)$. Before we move on to the exact pretraining data generative model proposed in this paper, we first provide a general setting for the data generation process. A task is defined by a joint distribution $\mathcal{D}_{\boldsymbol{x},y}$, which specifies the likelihood of obtaining a sample $(\boldsymbol{x}, y)$ from this task. Each task is sampled from the task prior $\mathcal{D}^{\text{prior}}$, meaning $\mathcal{D}^{\text{prior}}$ represents a distribution over distributions. The pretraining data comprises numerous sequences, each containing K labeled samples i.i.d. drawn from a distribution $\mathcal{D}_{\boldsymbol{x},y}$. We formally describe our pretraining data generative model in Assumption 1.

Assumption 1 (Pretraining Data Generative Model). Given an integer $K > 0$, a pretraining **task prior** $\mathcal{D}^{\text{prior}}$, we generate a sequence $\mathcal{S}_K$ as follows:
(a) Sample a task from the task prior: $\mathcal{D}_{\boldsymbol{x},y} \sim \mathcal{D}^{\text{prior}}$;
(b) Sample K labeled samples from the chosen task: $\forall i \in \{1, 2, \ldots, K\}$, $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}_{\boldsymbol{x},y}$;
(c) Define a sequence $\mathcal{S}_K$: $\mathcal{S}_K = [\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_K, y_K]$.

In the sequence, 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

Let $\mathcal{L}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_K}\left[\frac{1}{K}\sum_{k=0}^{K-1}(\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1})^2\right]$ as the pretraining objective, 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 optimization problem $\mathcal{F}^* = \operatorname{argmin}_{\mathcal{F}} \mathcal{L}(\mathcal{F})$ can be decomposed into K separate optimization problems for $k \in \{0, \ldots, K-1\}$:

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

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:

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

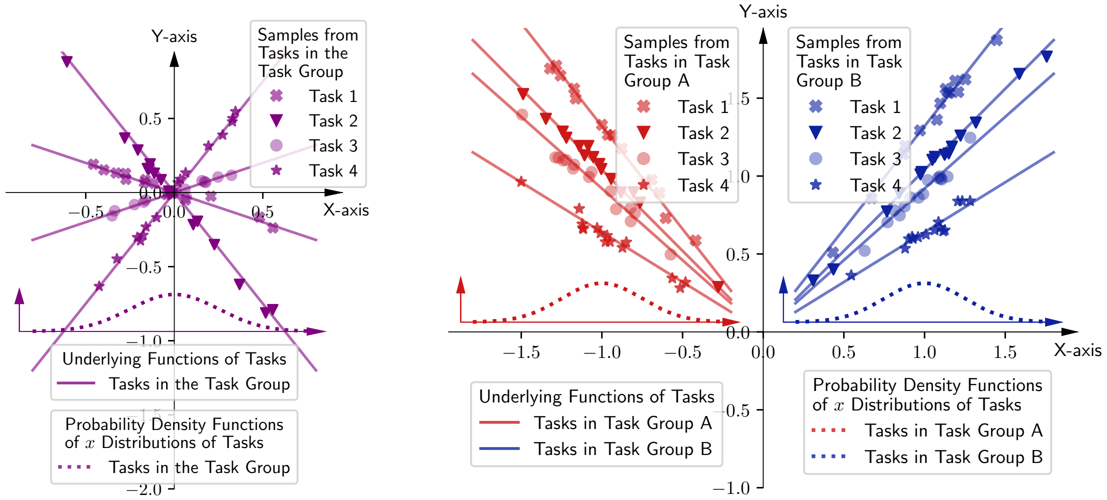
Thus, $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$ is the expectation (over task posterior) of $\mathop{\mathbb{E}}_{y_{k+1}}[y_{k+1}|\mathcal{D}_{\boldsymbol{x},y}, \boldsymbol{x}_{k+1}]$ regarding $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ as observation. We show that a pretrained Transformer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

3.3. Gaussian/Linear Assumptions on Pretraining Data Generative Model

Let us now elaborate further assumptions on $\mathcal{D}^{\text{prior}}$ and $\mathcal{D}_{\boldsymbol{x},y}$ in the Assumption 1 for a tractable posterior, extending beyond the scope of Raventos et al. (2023), who propose the data generative model that each task is a noisy linear regression task, the function $\boldsymbol{w}$ for each task is drawn from the same Gaussian distribution, and different tasks share the same $\boldsymbol{x}$ distribution. In contrast, our model posits that task functions are derived from a Gaussian mixture distribution, and tasks employ varying $\boldsymbol{x}$ distributions, as illustrated in Fig. 2. We formally formulate this setting in Assumption 6.

Assumption 2 (Gaussian/Linear Assumptions for Pretraining Data Generative Model).

[1] It is more rigorous to represent the vector $\boldsymbol{x}$ as multiple tokens. However, viewing it as a high-dimensional "token" simplifies our notation while not affecting our analysis. Thus, with a slight abuse of notation, we will treat both $\boldsymbol{x}_i$ and y_i as tokens for simplicity.



(a) Pretraining data Raventos et al. (2023).

(b) Our pretaining data with 2 task groups.

Figure 2: Pretraining data model of Raventos et al. (2023) and ours.

(a) $(\boldsymbol{\mu}, \boldsymbol{w}) \sim \mathcal{D}^{\text{prior}} : P(\boldsymbol{\mu}, \boldsymbol{w}) = \sum_{m=1}^{M} \pi_m P(\boldsymbol{\mu}, \boldsymbol{w}|T_m)$, where T_m is the m^{th} mixture component[2] 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 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) input: $\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu})$, $P(\boldsymbol{x}|\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sigma_x^2 \boldsymbol{I})$;
(c) label: $y|\boldsymbol{x} \sim \mathcal{D}_{y|\boldsymbol{x}}(\boldsymbol{w}) : P(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma_y^2)$;
(d) $\|\boldsymbol{\mu}_m\| = \|\boldsymbol{w}_m\| = 1, \forall m \in [M]$;
(e) $\exists r > 1$ that $\forall \alpha, \beta \in [M], \frac{1}{r} \leq \frac{\pi_\alpha}{\pi_\beta} \leq r$;
(f) $\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\mu}_m, \boldsymbol{w}, \boldsymbol{w}_m \in \mathbb{R}^d, \boldsymbol{I} \in \mathbb{R}^{d \times d}$.

Remark 3.1. Based on Assumptions 6(b) and 6(c), we define the probability of observing a sample $(\boldsymbol{x}, y)$ within a task $(\boldsymbol{\mu}, \boldsymbol{w})$ as the "noisy linear regression" likelihood.

Assumption 6(a) indicates that the pretraining dataset of an LLM consists of M different task groups. Assumption 6(b) posits that tasks have varying $\boldsymbol{x}$ distribution with varying mean but share the same covariance matrix. Assumption 6(c) assumes tasks as noisy linear regressions with the same noise scale in labels. Assumption 2(e) posits comparable mixture weights π across different task groups.

4. Inference and Dual Operating Modes

The previous Sec. 3.2 shows that performing ICL with the optimally pretrained next-token predictor is equivalent to computing the posterior mean of the label. In Sec. 4.1, we give the generation process of in-context examples. In Sec. 4.2, under Assumption 6 and treating $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ as observation, we derive a closed-form expression for the task posterior $\mathcal{D}^{\text{post}}$, and identify two factors in the transition from prior to posterior: Component Shifting and Component Re-weighting. In Sec. 4.3, we derive a closed-form expression of the ICL prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$. Further, Sec. 4.4 presents the results of numerical computation conducted under the tetrahedron setting, as illustrated in Fig. 3(a). The numerical computation results demonstrate the effects of component shifting and re-weighting. Finally, Sec. 4.5 raises the definitions of the dual operating modes with component shifting and re-weighting.

4.1. In-Context Task and In-Context Function

We introduce Assumption 3 for the in-context task and the in-context function of in-context examples:

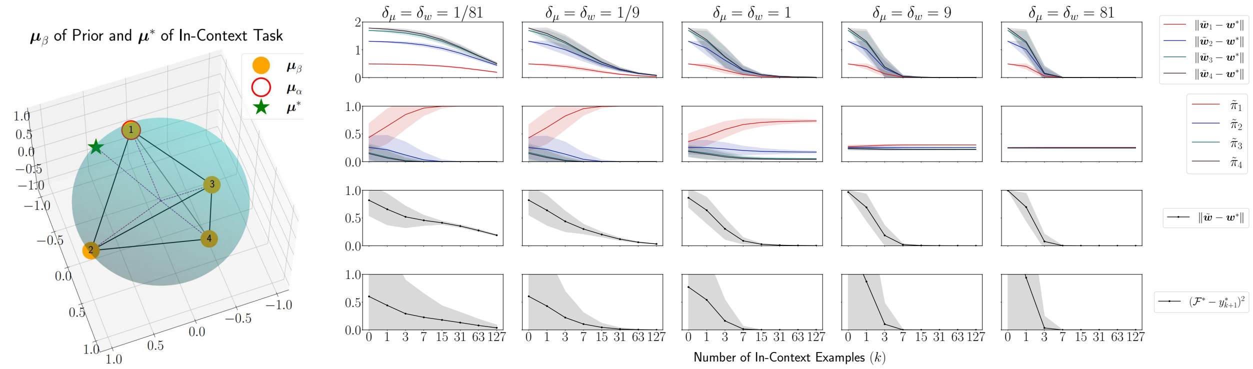
Assumption 3 (Gaussian/Linear Assumptions for In-Context Examples).
(a) The input sequence $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$ of ICL satisfies, $\forall i$, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I})$, $y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^* \rangle$;
(b) $\|\boldsymbol{\mu}^*\| = \|\boldsymbol{w}^*\| = 1$.

Assumption 3(a) states that each in-context example $(\boldsymbol{x}_i, y_i)$ is drawn from the in-context task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$, with $\boldsymbol{w}^*$ representing the specific in-context function and the labels being free from noise.

4.2. Closed-Form Expression of Posterior

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

[2] The concept "mixture component" is derived from Gaussian mixture models in the statistical literature and is analogous to the term "Task Group" depicted in the left-most panel of Fig. 1.



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

Lemma 4.1 (Conjugate Distributions with Noisy Linear Regression Likelihood)**.** *Under Assumption 6, 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 means $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$:

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

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 G for the proof.

Remark 4.2. Gaussian mixture is known to be a conjugate prior to the Gaussian likelihood. The outlined conjugate distributions in this lemma extend the Gaussian mixture conjugate distributions by substituting the Gaussian likelihood with the "noisy linear regression" likelihood in Remark 3.1.

Lemma 4.1 states that the task posterior remains a Gaussian mixture, with its mixture components shifted and re-weighted from the task prior. Therefore, understanding the impact of in-context examples on the posterior requires understanding how in-context examples affect the two factors:

- **Component Shifting (CS).** The component center is shifted from $(\boldsymbol{\mu}_m, \boldsymbol{w}_m)$ to $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$.
- **Component Re-weighting (CR).** The component weight is re-weighted from π to $\tilde{\pi}$.

Remark 4.3. The term "component" comes from the literature on Gaussian mixtures. It serves as an alternative to "Task Group" as shown in Fig. 2. The terminology "Component Shifting" and "Component Re-weighting" can be viewed as "Task Group Shifting" and "Task Group Re-weighting". We will abbreviate "mixture component center" to simply "center" when there is no ambiguity.

Leveraging Assumption 3, we collected mathematical analyses of CS and CR in Appendix H. The analysis explores the impacts of pretraining task noises and the number of in-context examples on $\tilde{\boldsymbol{\mu}}_m$, $\tilde{\boldsymbol{w}}_m$, and $\tilde{\pi}_m$, and examines the convergence of $\tilde{\boldsymbol{\mu}}_m$, $\tilde{\boldsymbol{w}}_m$, and $\tilde{\pi}_m$, as k approaches infinity.

4.3. Closed-form Expression of ICL Prediction

With Assumption 6 and Lemma 4.1, we have the following corollary for the prediction $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$:

Corollary 4.4. *Let $\tilde{\boldsymbol{w}} = \sum_{m=1}^{M} \tilde{\pi}_m \tilde{\boldsymbol{w}}_m$. With pretraining data generative model 1 and Assumption 6, 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. Apply Assumption 1 to Eq. 1, $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{w}) \sim \mathcal{D}^{\text{prior}}}[\langle \boldsymbol{x}_{k+1}, \boldsymbol{w} \rangle | \mathcal{S}_k \oplus \boldsymbol{x}_{k+1}]$. Using Lemma 4.1, this reduces to $\sum_{m=1}^{M} \tilde{\pi}_m \underset{(\boldsymbol{\mu},\boldsymbol{w}) \sim \widetilde{T}_m}{\mathbb{E}} [\langle \boldsymbol{x}_{k+1}, \boldsymbol{w} \rangle]$. Due to the linearity of expectation and inner product, the prediction can be simplified as $\langle \boldsymbol{x}_{k+1}, \sum_{m=1}^{M} \tilde{\pi}_m \tilde{\boldsymbol{w}}_m \rangle = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle$. □

Thus, the prediction is a convex combination of predictions by the centers of those shifted and re-weighted mixture components in the posterior. We are interested in how π_m and $\boldsymbol{w}_m$ change to $\tilde{\pi}_m$ and $\tilde{\boldsymbol{w}}_m$ with increasing k and how the pretraining prior distribution properties affect these changes.

4.4. 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 as k increases under different prior task noise conditions. The numerical computation is based on the tetrahedron setting with four prior mixture components as illustrated in Fig. 3(a). See Appendix B.1 for details of the tetrahedron setting. Fig. 3(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 1 over others. Then, as k increases further, the CS effect aligns all component centers with $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$.

4.5. Dual Operating Modes

The ***"task retrieval"*** mode describes a scenario where the impact of component re-weighting surpasses that of component shifting, leading to the prediction that is primarily influenced by the interplay between pretraining priors and in-context examples. An illustration of this is shown in the first column of Fig. 3(b), where the re-weighting of $\tilde{\pi}_m$ is more pronounced than the shifting of $\tilde{\boldsymbol{w}}_m$, indicating that CR plays a pivotal role in altering the prediction. In contrast, the ***"task learning"*** mode refers to situations where component shifting dominates over component re-weighting, resulting in the prediction almost depending on in-context examples and neglecting the pretraining priors.

5. Early Ascent

We now explain the early ascent phenomenon by analyzing a finegrained risk bound of ICL. (See Appendix C Theorem C.1 for the coarser bound.)

5.1. Finegrained Upper Bound

The finegrained upper bound for ICL risk is shown below:

Theorem 5.1 (Finegrained Upper Bound for ICL Risk). *Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, ICL risk is upper bounded by:*

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

where $\mathcal{L}_k^* = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^*)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle)^2$, $\|\boldsymbol{w}_m - \boldsymbol{w}^*\|$ is the distance between the in-context function $\boldsymbol{w}^*$ and the function $\boldsymbol{w}_m$ of center 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 L and Eq. 15 for proof details. In Appendix L.1, we further refine the bound for cases when in-context $\boldsymbol{x}_i$ only spans in a subspace of $\mathbb{R}^d$, resulting in $\lambda_1(\boldsymbol{A}) = 1$ constantly.

In-context examples affect the upper bound by affecting the two factors $\tilde{\pi}_\beta$ and $\lambda_1(\boldsymbol{A})$, corresponding to CR and CS introduced in Sec. 4.2. When ignoring the CR effect and only considering CS, the finegrained upper bound degrades to the general coarse bound in Appendix C Theorem C.1.

5.2. The Effect of Dual Operating Modes on ICL Risk

We numerically compute ICL risk under varied settings to explore the effect of the dual operating modes on the risk in Fig. 4. When pretraining task noises are low, *i.e.*, δ_μ and δ_w are small, the task retrieval mode happens with a small number of in-context examples, and the upper bound is affected by how $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$ is close to a prior center. Specifically, the task prior boosts the learning process of ICL if the in-context task is close to a prior center, due to the task retrieval mode quickly retrieving the task of the nearest prior center.

5.3. Early Ascent with Biased x Distribution

However, the task retrieval mode may not always benefit ICL. We notice a weird phenomenon is observed by Brown et al. (2020) and Xie et al. (2022). As the number of in-context samples increased, the performance of ICL first decreased and then increased. Brown et al. (2020) reports 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) also replicated this phenomenon with their synthetic dataset. Xie et al. (2022) explains this by "the few-shot setting introduces the distracting prompt structure, which can initially lower accuracy."

To obtain some insights, we present a simple scenario where $\boldsymbol{x}$ misleads the prediction by an LLM. Consider the following one-shot prompt for English-to-Korean translation: "What is the color of apple? 사과의 색깔은 무엇인가?[3] What is the color of banana?" The correct answer should be "바나나의 색깔은 무엇인가?"[4] However, GPT-3.5 generates "바나나의 색깔은 노란색 입니다," which means "The color of bananas is yellow." This shows that pretrained LLMs could retrieve an incorrect skill (question answering in this example) by observing misleading input ($\boldsymbol{x}$).

[3] "What is the color of apple?" in Korean.

[4] "What is the color of banana?" in Korean.

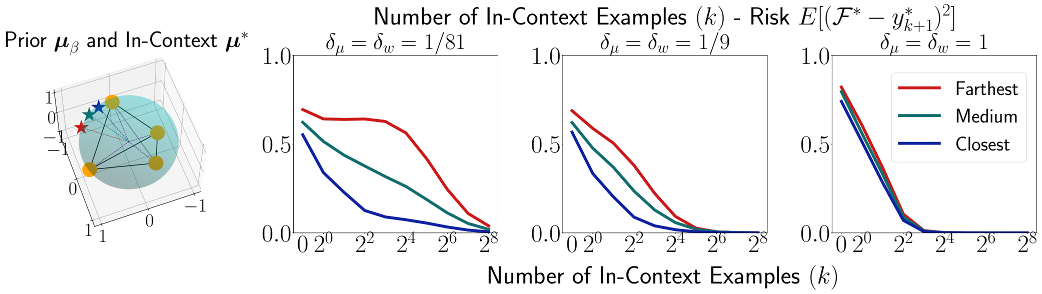
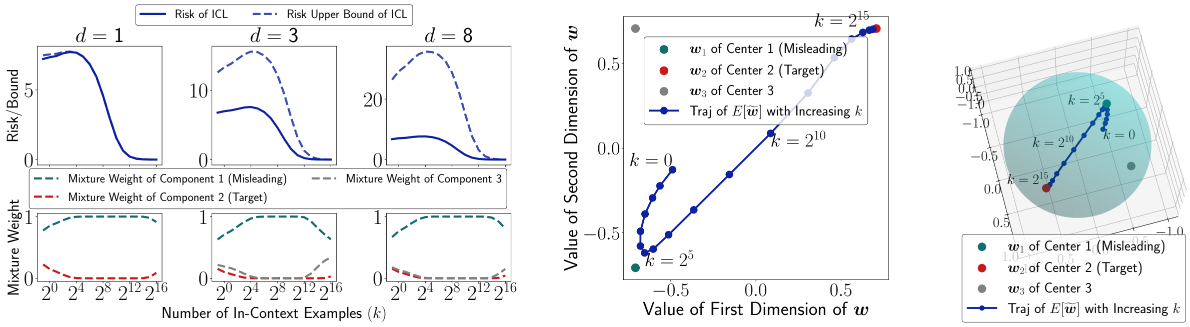


Figure 4: **Distance to the closest prior vs ICL risk.** We compute ICL risks of three target tasks colored red (farthest), green, and blue (closest), under the tetrahedron setting, illustrated in the left-most figure. The red target task has the longest distance to the closest prior center, and the blue target task has the shortest distance to the closest prior center. We can observe that the target task is easier to learn when the distance to the closest prior is smaller.



(a) Risk and π_m as k increases under $d \in \{1, 3, 8\}$.

(b) Expectation of $\tilde{\boldsymbol{w}}$ as k increases under $d \in \{2, 3\}$.

Figure 5: **The early ascent phenomenon.** Fig. 8(a) and Fig. 8(b) show that the task retrieval mode is dominant up to $k = 32$, and component 1's mixture weight increases ($\mathbb{E}[\tilde{\boldsymbol{w}}]$ approaches $\boldsymbol{w}_1$). Since this component is farther than the other one, the risk starts increasing. At larger k values, the risk starts decreasing ($\mathbb{E}[\tilde{\boldsymbol{w}}]$ approaches $\boldsymbol{w}_2$) via task learning. See Appendix B.3 for setting details. We further examine the early ascent phenomenon under linear regression with varied levels of label noises in Appendix I.1, and under non-linear regression and discrete token prediction in Appendix I.2.

Based on our analysis, we further show that the early ascent phenomenon provably occurs under a certain assumption Appendix J.1. We also reproduce early ascent in Fig. 8(a), where the upper bound and the risk initially increase due to the misleading task (of center 1) is retrieved first. Fig. 8(b) further demonstrates the relative locations of the retrieved functions to functions of prior centers. Finally, we give the formal theorem on the early ascent phenomenon:

Theorem 5.2 (Early Ascent)**.** *Assume* $\alpha = \arg\min_m \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}^*\|^2}{2\sigma_x^2} + \frac{\|(\boldsymbol{w}_m - \boldsymbol{w}^*)^\top \boldsymbol{\mu}^*\|^2 + d\tau_x^2\|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2}{2\sigma_y^2}$ *is the most misleading task and the task* α *satisfies* $\mathbb{E}_{\boldsymbol{x}_1}\left[(\mathcal{F}^*(\boldsymbol{x}_1) - \langle \boldsymbol{w}^*, \boldsymbol{x}_1\rangle)^2\right] < \mathbb{E}_{\boldsymbol{x}_1}\left[\langle \boldsymbol{x}_1, \boldsymbol{w}_\alpha - \boldsymbol{w}^*\rangle^2\right]$. *Then, when* δ_μ *and* δ_w *are small enough,* $\exists k \geq 1$ *s.t.:*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_1}\left[(\mathcal{F}^*(\boldsymbol{x}_1) - \langle \boldsymbol{w}^*, \boldsymbol{x}_1\rangle)^2\right]\\ &< \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{w}^*, \boldsymbol{x}_{k+1}\rangle)^2\right],\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{x}_1}\left[\langle \boldsymbol{x}_1, \boldsymbol{w}_\alpha - \boldsymbol{w}^*\rangle^2\right]$ *equals to the risk when the prediction fully depends on the misleading task function* $\boldsymbol{w}_\alpha$ *of prior center* α*. See Appendix J.2 for proof details.*

Theorem 5.2 shows that, if the misleading task α has a higher risk than the zero-shot risk, then when δ_μ and δ_w are small enough, the early ascent phenomenon happens.

6. Bounded Efficacy of Biased-Label ICL

We further predict the bounded efficacy phenomenon by examining the bound of ICL with biased labels. The assumption for ICL with biased labels is described as follows:

Assumption 4 (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 $\beta \neq \alpha$: $\forall \beta \neq \alpha$, $\|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 \geq d^2_{\boldsymbol{\mu}}$, $\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq d^2_{\boldsymbol{w}}$, and $\tau_x^2\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - (1+\tau_x^2)\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq \tau_x^2 u^2_{\boldsymbol{w}}$.

Assumption 4 depicts that to retrieve $\boldsymbol{w}_\alpha$ associated with the prior center α, the in-context task is selected based on its proximity to center α, ensuring it is closer to center α.

6.1. Upper Bound for ICL Risk with Biased Labels

The following theorem shows an upper bound for ICL risk with biased labels to retrieve a task:

Theorem 6.1 (Upper Bound for ICL Risk with Biased Labels)**.** *Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, ICL risk with biased labels is upper bounded by:*

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1 + d\tau_x^2) + \frac{C_1}{k\delta_w}\exp\left(C_2 k^{\frac{\delta}{2}-\frac{3}{4}}\right) + O(k^{-2})$$

where $\mathcal{L}_k^\alpha = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^\alpha)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle)^2$. When δ_μ and δ_w are sufficiently small, exists a particular interval for k s.t.:

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1 + d\tau_x^2)\min\{1, 4k^2{\delta_w}^2(1+\tau_x^2)^2\} + C_3\exp\left(-k\left(\frac{d_{\boldsymbol{\mu}}^2}{8\sigma_x^2} + \frac{u_{\boldsymbol{w}}^2\tau_x^2}{8\sigma_y^2}\right)\right) + C_4\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right).$$

As k increases, the second and third terms dominate and exponential decay when k is small, and the first term dominates and increases when k is large. C_1, C_2, C_3, and C_4 are constants depending on the prior setting, τ_x, and $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$. See Appendix M for proof details.

k	0	1	2	4	8	16
$+$	75.0%	36.2%	**33.9%**	49.3%	79.3%	85.1%
Biased $+$	100.0%	98.3%	95.9%	60.5%	24.4%	**16.8%**

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

6.2. Bounded Efficacy of Biased-Label ICL in GPT-4

This section further shows that the bounded efficacy phenomenon exists in GPT-4 in Table 1. With the task "biased addition $(+)$" as the in-context task corresponding to $\boldsymbol{w}^*$, as the number of in-context examples increases, ICL will first retrieve the skill "addition $(+)$" corresponding to $\boldsymbol{w}_\alpha$ which has a strong pretraining prior. Later, it will learn the "biased $+$" task, leading to the bounded efficacy phenomenon.

6.3. Bounded Efficacy for Zero-Shot ICL

We further introduce Lemma 6.2, a variation of the previous Theorem 6.1, to explain zero-shot ICL, an ICL algorithm capable of functioning with random labels (Lyu et al., 2023).

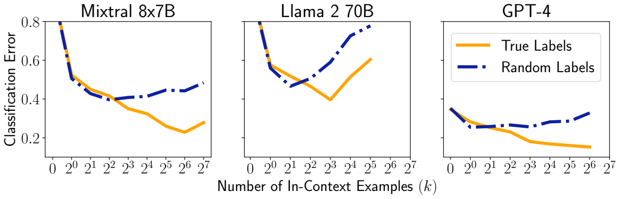


Figure 6: **Bounded efficacy.** The error rates of ICL with random labels start increasing at large k. See Appendix F for more experimental results.

Lemma 6.2 ((informal) Upper Bound for Zero-Shot ICL)**.** *Assume a next-token predictor attains the optimal pretraining risk, the risk of ICL with random labels (provide no information) will reveal a bounded efficacy phenomenon. See Appendix N for proof details.*

Lemma 6.2 says that as the number of in-context examples increases, the loss curve of zero-shot ICL with random 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 for the number of in-context examples ranging from 1 to 32. 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. 6 highlights the bounded efficacy phenomenon in the error curve associated with random labels. Compared with true labels, the error rate of ICL with random labels increases at a much smaller k value, clearly exhibiting the bounded efficacy phenomenon we predicted.

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 large language models. Our work lays the groundwork for future research in further exploration and improvement of ICL.

We conclude our paper with the limitations of our current framework: (i) the gap between our assumed pretraining linear regression tasks and complex, non-linear, categorical, real-world pretraining tasks of LLMs; (ii) the labels of in-context samples are assumed to be noiseless.

Acknowledgements

This work was supported by the NSF Award DMS-2023239, NSF CAREER Award CCF-2339978, Amazon Research Award, and a grant from FuriosaAI.

We would like to express our sincere gratitude to Kartik Sreenivasan for his invaluable discussions for this research. His insights and expertise have been instrumental in shaping this study. Additionally, we sincerely thank Andrew Geng for his contributions to coding for the initial experimental setup. His skills and dedication have been pivotal in the early stages of our research.

Impact Statement

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

References

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

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

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

Barbieri, F., Camacho-Collados, J., Anke, L. E., and Neves, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dagan, I., Glickman, O., and Magnini, B. The PASCAL recognising textual entailment challenge. In *PASCAL Machine Learning Challenges Workshop (MLCW)*, 2005.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. 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.

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

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

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

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped Transformers as programmable computers. In *International Conference on Machine Learning (ICML)*, 2023.

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

Jeon, H. J., Lee, J. D., Lei, Q., and Van Roy, B. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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

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

Lyu, X., Min, S., Beltagy, I., Zettlemoyer, L., and Hajishirzi, H. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

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

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. A SICK cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation (LREC)*, 2014.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. 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.

Pan, J., Gao, T., Chen, H., and Chen, D. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

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

Raventos, A., Paul, M., Chen, F., and Ganguli, S. 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.

Razeghi, Y., IV, R. L. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

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

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

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

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

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

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

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

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

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

A. Notations

This section collects all notations used in the main paper.

Notations 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 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 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, $[\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_K, y_K]$.
- $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$: $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} = [\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1}]$, which is a sequence of k in-context examples appended with $\boldsymbol{x}_{k+1}$.
- $\boldsymbol{\mu}$ and $\boldsymbol{w}$: the parameters that jointly specify a task. $\boldsymbol{\mu}$ specifies the distribution of $\boldsymbol{x}$, and $\boldsymbol{w}$ specifies 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 specified 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 $\mathcal{D}^{\text{prior}}$.
- $P(\boldsymbol{x}|\boldsymbol{\mu})$: the probability of $\boldsymbol{x}$ in $\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu})$.
- $P(y|\boldsymbol{x}, \boldsymbol{w})$: 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 the pretraining data generative model 1.
- M: the number of mixture components in a Gaussian mixture prior.
- $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$: the probability of $\boldsymbol{x}$ in the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- m, α, and β: the indices of mixture components in a Gaussian mixture prior.
- T_m: the m^{the} mixture component in a Gaussian mixture prior.
- π_m: the mixture weight of the m^{th} mixture component in a Gaussian mixture prior.
- $\boldsymbol{\mu}_m$ and $\boldsymbol{w}_m$: $(\boldsymbol{\mu}_m, \boldsymbol{w}_m)$ is the center of the m^{th} 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 introduced in Sec. 4:

- $\mathcal{D}^{\text{post}}$: The posterior distribution of the pretraining prior $\mathcal{D}^{\text{prior}}$ after observing $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$.
- $\|\cdot\|$: the L_2 norm.
- $\|\boldsymbol{x}\|^2$: 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 $\boldsymbol{A}$, $\|\boldsymbol{x}\|^2_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$.
- $P(\boldsymbol{\mu}, \boldsymbol{w}|\mathcal{S}_k \oplus \boldsymbol{x}_{k+1})$: the probability of task $(\boldsymbol{\mu}, \boldsymbol{w})$ in the posterior after observing $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$.
- $\widetilde{T}_m$: the m^{th} mixture component in the Gaussian mixture posterior.
- $\tilde{\pi}_m$: the mixture weight of the m^{th} mixture component in the Gaussian mixture posterior.
- $\tilde{\boldsymbol{\mu}}_m$ and $\tilde{\boldsymbol{w}}_m$: $(\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{w}}_m)$ is the center of the m^{th} mixture component in the Gaussian mixture posterior.
- $P(\boldsymbol{\mu}, \boldsymbol{w}|\widetilde{T}_m)$: the probability of task $(\boldsymbol{\mu}, \boldsymbol{w})$ in the m^{th} mixture component of posterior.
- δ_μ and δ_w: the ratios of squared task noises over squared sample noises. $\delta_\mu = \frac{\sigma_\mu^2}{\sigma_x^2}$, 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. $\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) = \langle \boldsymbol{x}_{k+1}, \tilde{\boldsymbol{w}} \rangle = \left\langle \boldsymbol{x}_{k+1}, \sum_{m=1}^{M} \tilde{\pi}_m \tilde{\boldsymbol{w}}_m \right\rangle$.
- $c_m^{\boldsymbol{\mu}}$ and $c_m^{\boldsymbol{w}}$: 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 d^{th} 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 $\boldsymbol{A}$.
- $\lambda_1(\boldsymbol{A})$: the 1st, the largest eigenvalue of matrix $\boldsymbol{A}$.
- y_{k+1}^*: the label of learning the function $\boldsymbol{w}^*$. $y_{k+1}^* = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle$.

Notations introduced in Sec. 5:

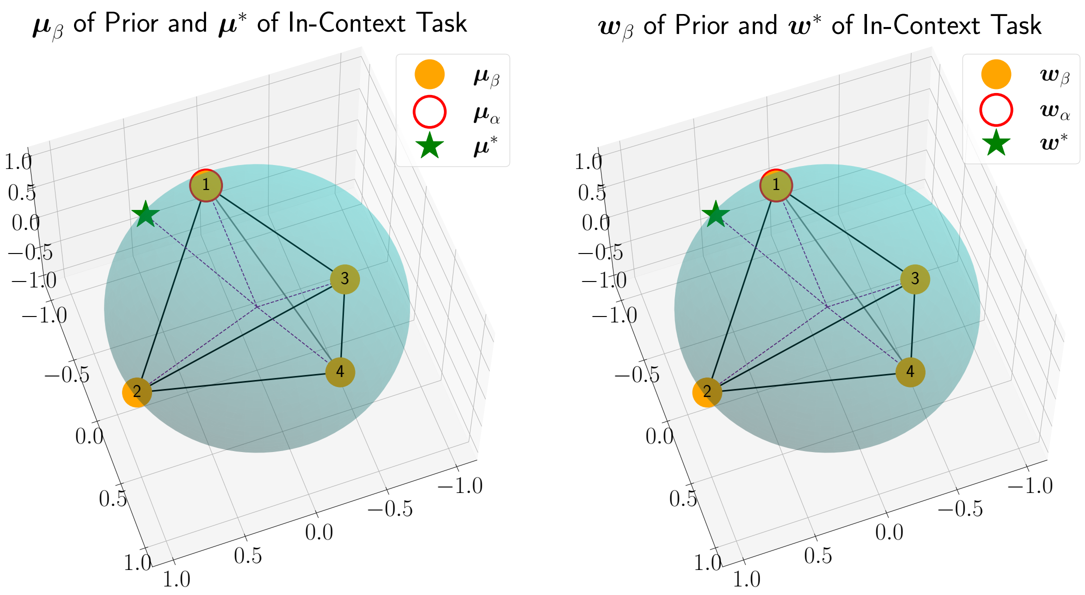
- The L2 loss of ICL learning to learn the function $\boldsymbol{w}^*$. $\mathcal{L}_k^* = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^*)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle)^2$.

Notations introduced in Sec. 6:

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

B. Prior Examples

This section outlines our configurations of prior settings in numerical computations and preliminary Transformer experiments, focusing on the geometrical arrangement of the centers in the priors. Specifically, we detail the configurations where the centers form shapes of 3-dimensional regular polyhedra in Sec. B.1, extend to configurations in d-dimensional spaces in Sec. B.2, and discuss a unique setup related to the early ascent phenomenon in Sec. B.3.


$\boldsymbol{\mu}_\beta$ of Prior and $\boldsymbol{\mu}^*$ of In-Context Task

$\boldsymbol{w}_\beta$ of Prior and $\boldsymbol{w}^*$ of In-Context Task

Figure 7: **Visualization of the tetrahedron setting.** The figure shows the pretraining prior centers and the in-context task. For $\beta \in \{1, 2, 3, 4\}$, $(\boldsymbol{\mu}_\beta, \boldsymbol{w}_\beta)$ is a 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 denoted by $\boldsymbol{\mu}$ or $\boldsymbol{w}$.

B.1. Regular Polyhedrons

Taking into account the centers of the mixture components from the pretraining prior, which manifest as distinct points forming the vertices of various shapes, we examine 3-dimensional regular polyhedrons. These include 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.

The configuration of a regular polyhedron with M centers is established in accordance with the parameters outlined in Assumption 6, as detailed below:

- Dimension $d = 3$, the number of mixture components equals to M;
- The centers of mixture components form a regular polyhedron with M vertices;
- All components' mixture weights are the same, $\pi_m = 1/M$, and $\boldsymbol{\mu}_m = \boldsymbol{w}_m$, for all $m \in [M]$;
- 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 the in-context task, $\boldsymbol{\mu}^* = \frac{2\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{\|2\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|}$ and $\boldsymbol{w}^* = \frac{2\boldsymbol{w}_1 + \boldsymbol{w}_2}{\|2\boldsymbol{w}_1 + \boldsymbol{w}_2\|}$ if not specified, where $\boldsymbol{\mu}_2$ is one of the the closest centers to $\boldsymbol{\mu}_1$.

We 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 in the prior and the in-context task are shown in Fig. 7. The parameters are noted as follows:

- Dimension $d = 3$, number of mixture components $M = 4$;
- The centers of topics form a tetrahedron as shown in Fig. 7. $\boldsymbol{\mu}_1 = \boldsymbol{w}_1 = [0, 0, -1]^\top$, $\boldsymbol{\mu}_2 = \boldsymbol{w}_2 = [\sqrt{\frac{8}{9}}, 0, \frac{1}{3}]^\top$, $\boldsymbol{\mu}_3 = \boldsymbol{w}_3 = [-\sqrt{\frac{2}{9}}, +\sqrt{\frac{2}{3}}, \frac{1}{3}]^\top$, and $\boldsymbol{\mu}_4 = \boldsymbol{w}_4 = [-\sqrt{\frac{2}{9}}, -\sqrt{\frac{2}{3}}, \frac{1}{3}]^\top$;
- All components' mixture weights are the same, $\pi_m = 1/4$, and $\boldsymbol{\mu}_m = \boldsymbol{w}_m$, for all $m \in \{1, 2, 3, 4\}$;

Table 2: **Prior settings for early ascent.** The pretraining task prior comprises two components for one dimension and three for two or more dimensions. ICL aims 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 $\boldsymbol{x}$ distribution to the task of prior center 1 but a closer $\boldsymbol{x} \to y$ mapping to the prior center 2. The parameters for all cases are set to $\sigma_\mu = \sigma_w = 0.05$, $\sigma_x = \tau_x = 1$, and $\sigma_y = 2$. Refer to Fig. 8(b) for visualization of the prior centers under dimension $d \in \{1, 2, 3\}$.

Case	Component /Task	Mixture Weight	$\boldsymbol{\mu}$	$\boldsymbol{w}$
$d = 1$	Component 1	½	$\boldsymbol{\mu}_1 = [+1]$	$\boldsymbol{w}_1 = [-1]$
	Component 2	½	$\boldsymbol{\mu}_2 = [-1]$	$\boldsymbol{w}_2 = [+1]$
	Component 3	/	/	/
	In-context Task	/	$\boldsymbol{\mu}^* = [+1]$	$\boldsymbol{w}^* = [+1]$
$d = 2$	Component 1	⅓	$\boldsymbol{\mu}_1 = [+1, +1]$	$\boldsymbol{w}_1 = [-1, -1]$
	Component 2	⅓	$\boldsymbol{\mu}_2 = [-1, -1]$	$\boldsymbol{w}_2 = [+1, +1]$
	Component 3	⅓	$\boldsymbol{\mu}_3 = [+1, -1]$	$\boldsymbol{w}_3 = [-1, +1]$
	In-context Task	/	$\boldsymbol{\mu}^* = [+1, +1]$	$\boldsymbol{w}^* = [+1, +1]$
$d \geq 2$	Component 1	⅓	$\boldsymbol{\mu}_1 = [+1] + [+1] \times (d-1)$	$\boldsymbol{w}_1 = [-1] + [-1] \times (d-1)$
	Component 2	⅓	$\boldsymbol{\mu}_2 = [-1] + [-1] \times (d-1)$	$\boldsymbol{w}_2 = [+1] + [+1] \times (d-1)$
	Component 3	⅓	$\boldsymbol{\mu}_3 = [+1] + [-1] \times (d-1)$	$\boldsymbol{w}_3 = [-1] + [+1] \times (d-1)$
	In-context Task	/	$\boldsymbol{\mu}^* = [+1] \times d$	$\boldsymbol{w}^* = [+1] \times d$

- 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 $m = 3$ and $m = 4$ produce slightly different curves.

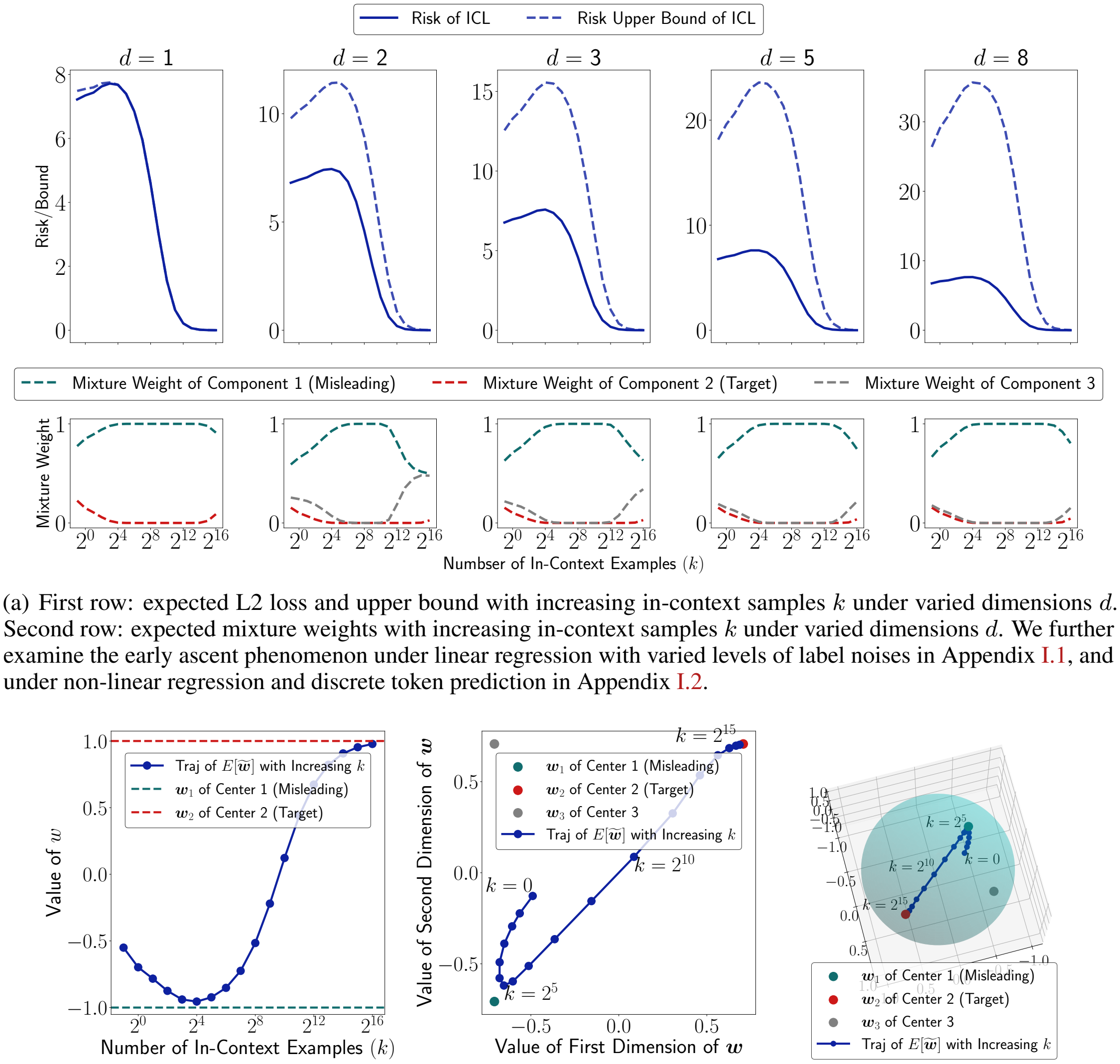
B.2. d-Dimensional Examples

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 equals to d, number of mixture component $M = d$;
- For all $m \in [M]$, $\boldsymbol{\mu}_m = \boldsymbol{e}_m$ and $\boldsymbol{\mu}_{m,i} = \begin{cases} 1 & \text{if } i = m \\ 0 & \text{if } i \neq m \end{cases}$, *i.e.*, $\boldsymbol{\mu}_m$ is the m^{th} vector in the standard basis of $\mathbb{R}^m$, characterized by having all elements equal to 0 except for the m^{th} element, which is 1.
- All components' mixture weights are the same, $\pi_m = 1/d$, and $\boldsymbol{\mu}_m = \boldsymbol{w}_m$, for all $m \in [M]$;
- 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 $\sigma_\mu = \sigma_w = 0.25$;
- For the in-context task, we have $\boldsymbol{\mu}^* = \frac{2\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{\|2\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|}$ and $\boldsymbol{w}^* = \frac{2\boldsymbol{w}_1 + \boldsymbol{w}_2}{\|2\boldsymbol{w}_1 + \boldsymbol{w}_2\|}$.

B.3. Early Ascent Examples

Table 2 outlines the prior configuration used to produce the early ascent phenomenon, where the in-context task is designed with a distribution of $\boldsymbol{x}$ close to a misleading task. The full results are shown in Fig. 8.



(a) First row: expected L2 loss and upper bound with increasing in-context samples k under varied dimensions d. Second row: expected mixture weights with increasing in-context samples k under varied dimensions d. We further examine the early ascent phenomenon under linear regression with varied levels of label noises in Appendix I.1, and under non-linear regression and discrete token prediction in Appendix I.2.

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

Figure 8: **The early ascent phenomenon.** Fig. 8(a) displays the trends of expected losses, upper bounds, and mixture weights, while Fig. 8(b) presents the trend of the expectation of $\tilde{\boldsymbol{w}}$. We can see that the task retrieval mode is dominant up to $k = 32$, and component 1's mixture weight increases ($\mathbb{E}[\tilde{\boldsymbol{w}}]$ approaches $\boldsymbol{w}_1$). Since this misleading component 1 is far from the target component 2, the risk starts increasing. At larger k values, the risk starts decreasing ($\mathbb{E}[\tilde{\boldsymbol{w}}]$ approaches $\boldsymbol{w}_2$) via task learning.

C. Coarse Upper Bound for ICL Risk

The following theorem shows a coarse upper bound of the ICL risk parallel to Theorem 5.1:

Theorem C.1 (Coarse Upper Bound for ICL Risk)**.** *Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, the 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 L 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_w and k are colored for easier observation.

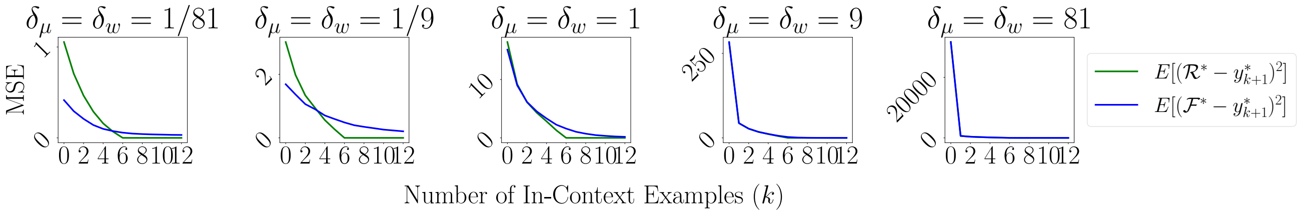


Figure 9: **In-context learning vs ridge regression.** $\mathcal{R}^*$ indicates the prediction by ridge regression, $\mathcal{F}^*$ indicates the prediction by ICL with a Bayes-optimal next-token predictor, and $y^*_{k+1} = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle$. Let the k samples draw from a task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$, which is drawn from the pretraining prior distribution. The dimension d of $\boldsymbol{x}$ equals 6. We observe that ICL performs better than ridge regression when k is small, 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. 9 shows the experiment results. Under certain settings for the task prior $\mathcal{D}_{\boldsymbol{\mu},\boldsymbol{w}}$, when the task prior has low task variances, ICL performs better than ridge regression with a fixed regularization parameter under small k.

D. Transformer Performance in Approximating Bayesian Inference

We examine if a Transformer network 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 noises*, *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 HuggingFace[5]. 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.

Experiment Results. Fig. 10, 11, and 12 show the experimental results, where $\hat{\mathcal{F}}$ denotes the prediction of the Transformer network, $\mathcal{F}^*$ denotes the prediction of Bayesian inference, and $y^*_{k+1} = \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}^* \rangle$ is the label of learning the in-context function. In Fig. 10, we consider the tetrahedron setting (see Apendix B.1 for setting details) under varied task noises ($\delta_\mu = \delta_w \in \{1/256, 1/64, 1/16, 1/4, 1\}$). In Fig. 11, we consider settings of regular shapes (see Appendix B.1 for setting details) with different numbers of vertices/components ($M \in \{4, 6, 8, 12, 20\}$). In Fig. 12, we consider settings with varied dimensions (see Appendix B.2 for setting details, $d \in \{2, 4, 8, 16, 32\}$). We observe that the trained Transformer network can approximate the Bayes-optimal predictor under varied settings, and the larger the number of dimensions and the number of mixture components, the harder it is for the Transformer network to approximate Bayesian prediction.

E. Additional Information for Bounded Efficacy in GPT-4

E.1. Experimental Setting

Table 3 introduces the experiment setting of GPT-4, including the system message, the prompt, the in-context task, the “biased +” task, and the “addition (+)” task. Designating the “biased +” task as the in-context task, *i.e.*, $c_i = a_i + b_i + 1$, we measure the performances on two goals, including learning the “biased +” task and retrieving the “addition (+)” task.

[5] https://huggingface.co/

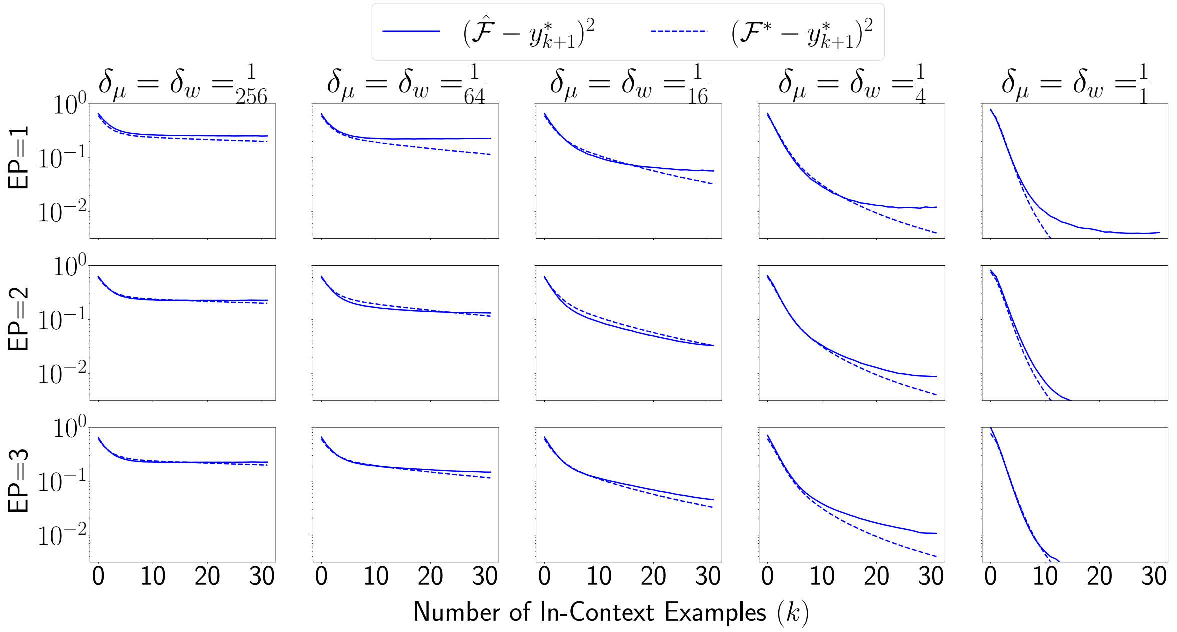


Figure 10: **Prior task noises.** The figure shows the experiment results under varied noise levels. δ_μ and δ_w indicate the noise levels of the pretraining task prior. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer network. The results show that the trained Transformer network's performance can approach the performance of Bayesian inference.

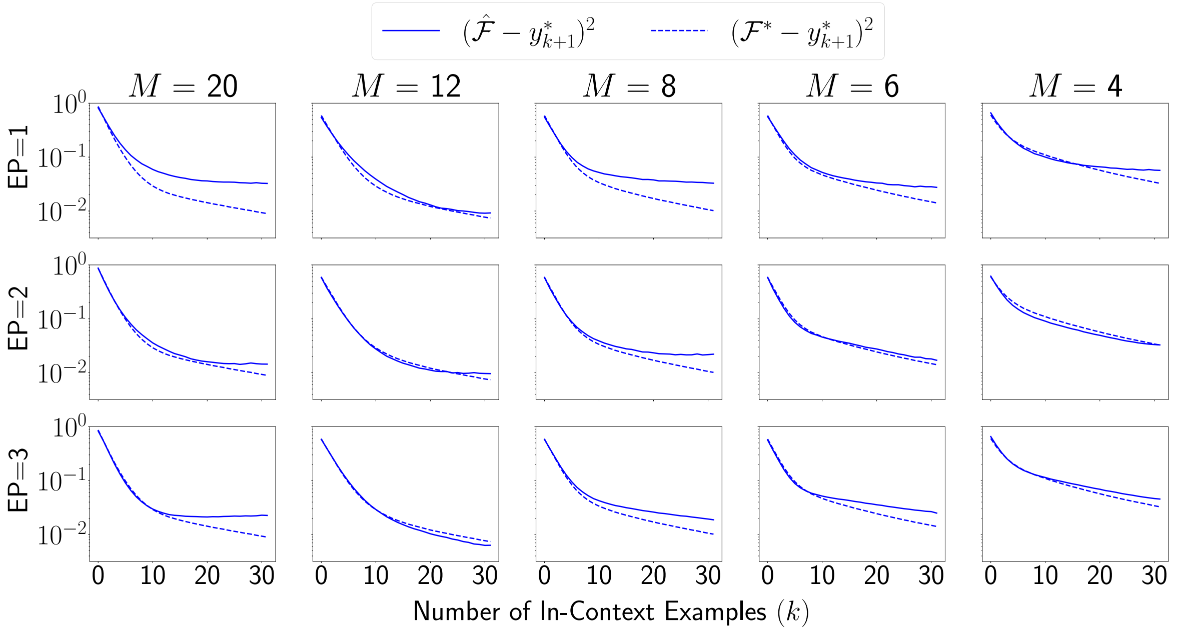


Figure 11: **Number of components.** The figure shows the experiment results under varied component densities. M indicates the number of mixture components corresponding to different 3D regular polyhedrons described in Appendix B.1, and $\delta_\mu = \delta_w = \frac{1}{16}$. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer network. The higher the component density is, the harder it is for the Transformer network to approach Bayesian inference.

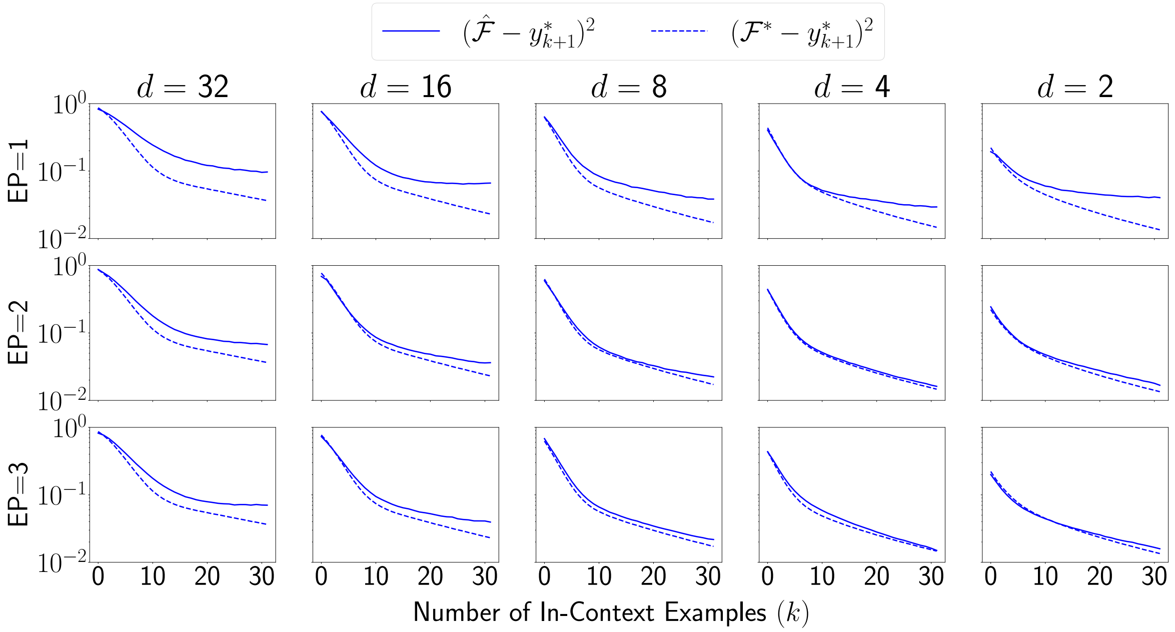


Figure 12: **Feature dimension.** The figure shows the experiment results under varied dimensions. d indicates the dimension and the number of mixture components (see Appendix B.2 for setting details), and $\delta_\mu = \delta_w = \frac{1}{16}$. $\mathcal{F}^*$ indicates the prediction of Bayesian inference while $\hat{\mathcal{F}}$ indicates the prediction of the trained Transformer network. The higher the feature dimension is, the harder it is for the Transformer network to approach Bayesian inference.

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

Setting	Desciption
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 “biased +” Task with True Labels	Aiming to learn the “biased +” task, a(?)b=(a+b+1), with in-context examples following the same “biased +” task, a(?)b=(a+b+1).
Goal of Retrieving the “addition (+)” Task with Biased Labels	Aiming to retrieve the “addition (+)” task, a(?)b=(a+b). However, the in-context examples are provided with a slightly different task “biased +”, a(?)b=(a+b+1).

E.2. Additional Results

This section collects four pairs of prompts and predictions for $k = 0, 2, 8$ in Tables 4, 5, and 6. The results show that ICL with biased labels will initially retrieve a commonsense pretraining task due to task retrieval, and finally learn the in-context task because of task learning.

Table 4: Zero in-context example, $k = 0$. 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 ($a(?)b = (a + b + 1)$). "..." denotes the hidden part of the prompt. Please refer to Table 3 for the whole prompt.

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 5: Two 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 ($a(?)b = (a + b + 1)$). "..." denotes the hidden part of the prompt. Please refer to Table 3 for the whole prompt.

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: Eight 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 ($a(?)b = (a + b + 1)$). "..." denotes the hidden part of the prompt. Please refer to Table 3 for the whole prompt.

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

F. Bounded Efficacy in Zero-shot ICL

This section introduces the experiment setting of Fig. 6. We start by introducing the experiment results in Fig. 13 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 6.2), Fig. 13 seems to imply a conflict phenomenon. Thus, we further extend the number of

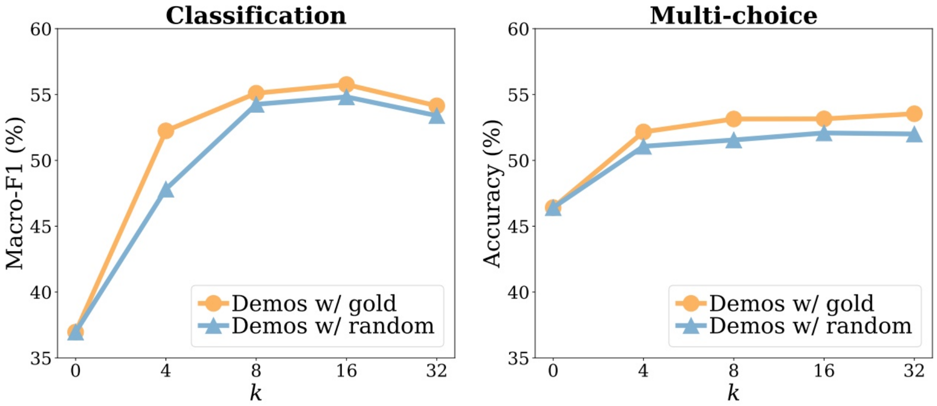


Figure 13: Ablations on varying numbers of examples in the demonstrations (k). Models that are the best under 13B in each task category (Channel MetaICL and Direct GPT-J, respectively) are used.

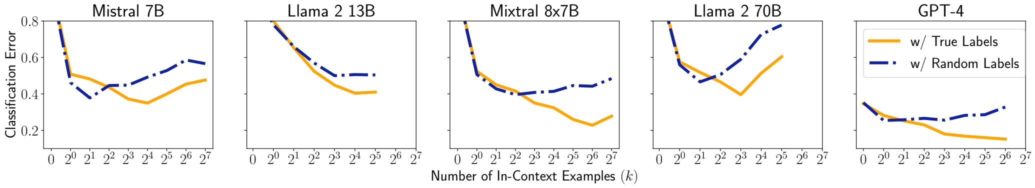


Figure 14: As k increases, the classification error curve of ICL with random labels exhibits the bounded efficacy phenomenon. The curve with true labels further confirms that this phenomenon is not due to models tending to perform worse on long sequences.

in-context examples in Fig. 13 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[6] 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), Mixtral 8×7B (32768), Llama2 13B (4096), Llama2 70B (4096), and GPT-4 (8192) 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[7].

G. The Derivation of Posterior

This section provides detailed derivations for Lemma 4.1. We begin by showing the posterior is potentially still a Gaussian mixture in Sec. G.1. Then, in Sec. G.2, we show how Eq. 2 is proportion to Eq. 3, which is precisely a Gaussian mixture.

G.1. Prior to Posterior

We start by showing the posterior is potentially still a Gaussian mixture. For fixed $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$:

$$\begin{aligned}
&P(\boldsymbol{\mu}, \boldsymbol{w} | \mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) \\
&\propto P(\boldsymbol{\mu}, \boldsymbol{w} | \mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) \\
&= P(\boldsymbol{\mu}, \boldsymbol{w}, \mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) \\
&= P(\boldsymbol{\mu}, \boldsymbol{w}) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} | \boldsymbol{\mu}, \boldsymbol{w}) \\
&= \left(\sum_{m=1}^{M} \pi_m P(\boldsymbol{\mu}, \boldsymbol{w} | T_m) \right) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} | \boldsymbol{\mu}, \boldsymbol{w})
\end{aligned}$$

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

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

$$= \sum_{m=1}^{M} \pi_m P(\boldsymbol{\mu}, \boldsymbol{w} | T_m) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} | \boldsymbol{\mu}, \boldsymbol{w}) \tag{2}$$

$$\propto \sum_{m=1}^{M} \tilde{\pi}_m P(\boldsymbol{\mu}, \boldsymbol{w} | \widetilde{T}_m). \tag{3}$$

We give the derivation from Eq. 2 to Eq. 3 in the next section.

G.2. Closed-form Solution from Eq. 2 to Eq. 3

We analyze each component (indicated by a specific m) in Eq. 2. Given fixed $\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}$, for all $m \in [M]$ and all $(\boldsymbol{\mu}, \boldsymbol{w})$, we have:

$$\log(P(\boldsymbol{\mu}, \boldsymbol{w} | T_m) P(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1} | \boldsymbol{\mu}, \boldsymbol{w}))$$

$$= -\frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}\|^2}{2\sigma_\mu^2} - \frac{\|\boldsymbol{w}_m - \boldsymbol{w}\|^2}{2\sigma_w^2} - \frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu} - \boldsymbol{x}_i\|^2}{2\sigma_x^2} - \frac{\sum_{i=1}^{k} \|\boldsymbol{x}_i^\top \boldsymbol{w} - y_i\|^2}{2\sigma_y^2}$$

$$+ \log\left(\frac{(2\pi)^{-d/2}}{\sigma_\mu^d}\right) + \log\left(\frac{(2\pi)^{-d/2}}{\sigma_w^d}\right) + (k+1)\log\left(\frac{(2\pi)^{-d/2}}{\sigma_x^d}\right) + k\log\left(\frac{(2\pi)^{-1/2}}{\sigma_y}\right)$$

(Let $C_3 = \log\left(\frac{(2\pi)^{-d/2}}{\sigma_\mu^d}\right) + \log\left(\frac{(2\pi)^{-d/2}}{\sigma_w^d}\right) + (k+1)\log\left(\frac{(2\pi)^{-d/2}}{\sigma_x^d}\right) + k\log\left(\frac{(2\pi)^{-1/2}}{\sigma_y}\right)$.)

$$= C_3 - \frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}\|^2}{2\sigma_\mu^2} - \frac{\|\boldsymbol{w}_m - \boldsymbol{w}\|^2}{2\sigma_w^2} - \frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu} - \boldsymbol{x}_i\|^2}{2\sigma_x^2} - \frac{\sum_{i=1}^{k} \|\boldsymbol{x}_i^\top \boldsymbol{w} - y_i\|^2}{2\sigma_y^2}$$

$$= C_3 - \left(\frac{\|\boldsymbol{\mu}_m - \boldsymbol{\mu}\|^2}{2\sigma_\mu^2} + \frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu} - \boldsymbol{x}_i\|^2}{2\sigma_x^2}\right) - \left(\frac{\|\boldsymbol{w}_m - \boldsymbol{w}\|^2}{2\sigma_w^2} + \frac{\sum_{i=1}^{k} \|\boldsymbol{x}_i^\top \boldsymbol{w} - y_i\|^2}{2\sigma_y^2}\right)$$

(Let $\delta_\mu = \frac{\sigma_\mu^2}{\sigma_x^2}$ and $\delta_w = \frac{\sigma_w^2}{\sigma_y^2}$.)

$$= C_3 - \frac{1}{2\sigma_\mu^2}\left((\|\boldsymbol{\mu}_m\|^2 - 2\boldsymbol{\mu}_m^\top \boldsymbol{\mu} + \|\boldsymbol{\mu}\|^2) + \delta_\mu\left((k+1)\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}^\top \sum_{i=1}^{k+1} \boldsymbol{x}_i + \sum_{i=1}^{k+1} \|\boldsymbol{x}_i\|^2\right)\right)$$

$$- \frac{1}{2\sigma_\mu^2}\left((\|\boldsymbol{w}_m\|^2 - 2\boldsymbol{w}_m^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2) + \delta_w\left(\sum_{i=1}^{k} \boldsymbol{w}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w} - 2\boldsymbol{w}^\top \sum_{i=1}^{k} \boldsymbol{x}_i y_i + \sum_{i=1}^{k} y_i^2\right)\right)$$

$$= C_3 - \frac{1}{2\sigma_\mu^2}\left(\|\boldsymbol{\mu}_m\|^2 + (1 + (k+1)\delta_\mu)\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}\left(\boldsymbol{\mu}_m + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i\right) + \delta_\mu \sum_{i=1}^{k+1} \|\boldsymbol{x}_i\|^2\right)$$

$$- \frac{1}{2\sigma_w^2}\left(\|\boldsymbol{w}_m\|^2 + \boldsymbol{w}^\top\left(\boldsymbol{I} + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\boldsymbol{w} - 2\boldsymbol{w}\left(\boldsymbol{w}_m + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i\right) + \delta_w \sum_{i=1}^{k} y_i^2\right)$$

(Let $C_4 = C_3 - \frac{\delta_\mu}{2\sigma_\mu^2} \sum_{i=1}^{k+1} \|\boldsymbol{x}_i\|^2 - \frac{\delta_w}{2\sigma_w^2} \sum_{i=1}^{k} y_i^2$.)

$$= C_4 - \frac{1}{2\sigma_\mu^2}\left(\|\boldsymbol{\mu}_m\|^2 + (1 + (k+1)\delta_\mu)\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}\left(\boldsymbol{\mu}_m + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i\right)\right)$$

$$- \frac{1}{2\sigma_w^2}\left(\|\boldsymbol{w}_m\|^2 + \boldsymbol{w}^\top\left(\boldsymbol{I} + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\boldsymbol{w} - 2\boldsymbol{w}\left(\boldsymbol{w}_m + \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i\right)\right)$$

(Let $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = \boldsymbol{I}$ and $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$.)

$$= C_4 - \frac{1}{2\sigma_\mu^2}\left(\|\boldsymbol{\mu}_m\|^2 + \|\boldsymbol{\mu}\|^2_{\boldsymbol{I} + (k+1)\delta_\mu \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}} - 2\boldsymbol{\mu}^\top\left(\boldsymbol{\mu}_m + \delta_\mu \sum_{i=1}^{k+1} \boldsymbol{x}_i\right)\right)$$

$$
\begin{aligned}
&\quad -\frac{1}{2\sigma_w^2}\left(\|\boldsymbol{w}_m\|^2+\|\boldsymbol{w}\|^2_{\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}-2\boldsymbol{w}^\top\left(\boldsymbol{w}_m+\delta_w\sum_{i=1}^{k}\boldsymbol{x}_iy_i\right)\right)\\
&\text{(Let } \bar{\boldsymbol{\mu}}=\sum_{i=1}^{k+1}\boldsymbol{x}_i \text{ and } \bar{\boldsymbol{w}}=\frac{\sum_{i=1}^{k}\boldsymbol{x}_iy_i}{k}.\text{)}\\
&=C_4-\frac{1}{2\sigma_\mu^2}(\|\boldsymbol{\mu}_m\|^2+\|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}-2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_m+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}))\\
&\quad -\frac{1}{2\sigma_w^2}(\|\boldsymbol{w}_m\|^2+\|\boldsymbol{w}\|^2_{\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}-2\boldsymbol{w}^\top(\boldsymbol{w}_m+k\delta_w\bar{\boldsymbol{w}}))\\
&\text{(Let } \Delta_\mu=(k+1)\delta_\mu \text{ and } \Delta_w=k\delta_w.\text{)}\\
&=C_4-\frac{1}{2\sigma_\mu^2}(\|\boldsymbol{\mu}_m\|^2+\|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}-2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}}))\\
&\quad -\frac{1}{2\sigma_w^2}(\|\boldsymbol{w}_m\|^2+\|\boldsymbol{w}\|^2_{\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}-2\boldsymbol{w}^\top(\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}}))\\
&=C_4-\Big(\|\boldsymbol{\mu}_m\|^2+\Big(\|\boldsymbol{\mu}\|^2_{\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}-2\boldsymbol{\mu}^\top(\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}})+\|\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}\Big)-\|\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}\Big)/2\sigma_\mu^2\\
&\quad -\Big(\|\boldsymbol{w}_m\|^2+\Big(\|\boldsymbol{w}\|^2_{\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}-2\boldsymbol{w}^\top(\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}})+\|\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\Big)-\|\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\Big)/2\sigma_w^2\\
&=C_4-\frac{1}{2\sigma_\mu^2}\bigg(\Big(\|\boldsymbol{\mu}_m\|^2-\|\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}\Big)+\|\boldsymbol{\mu}-(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}})\|^2_{\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}\bigg)\\
&\quad -\frac{1}{2\sigma_w^2}\bigg(\Big(\|\boldsymbol{w}_m\|^2-\|\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\Big)+\|\boldsymbol{w}-(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}})\|^2_{\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}\bigg).
\end{aligned}
$$

Notice C_4 is independent to m, $\boldsymbol{\mu}$, and $\boldsymbol{w}$, thus we have:

$$
\begin{aligned}
&P(\boldsymbol{\mu},\boldsymbol{w}|T_m)P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w})\\
&\propto\exp\left(-\frac{1}{2\sigma_\mu^2}\bigg(\Big(\|\boldsymbol{\mu}_m\|^2-\|\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}}\|^2_{(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}}\Big)+\|\boldsymbol{\mu}-(\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m+\Delta_\mu\bar{\boldsymbol{\mu}})\|^2_{\boldsymbol{I}+\Delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}}\bigg)\right)\\
&\quad\cdot\exp\left(-\frac{1}{2\sigma_w^2}\bigg(\Big(\|\boldsymbol{w}_m\|^2-\|\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\Big)+\|\boldsymbol{w}-(\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m+\Delta_w\bar{\boldsymbol{w}})\|^2_{\boldsymbol{I}+\Delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}\bigg)\right)\\
&\propto\underbrace{\exp\left(-\frac{\|\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{\mu}}}\underbrace{\exp\left(-\frac{\|\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)}_{c_m^{\boldsymbol{w}}}\\
&\quad\cdot\mathcal{N}(\boldsymbol{\mu}|(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}),\sigma_\mu^2(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1})\\
&\quad\cdot\mathcal{N}(\boldsymbol{w}|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m+k\delta_w\bar{\boldsymbol{w}}),\sigma_w^2(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}).
\end{aligned}
$$

By defining $P(\boldsymbol{\mu},\boldsymbol{w}|\widetilde{T})=\mathcal{N}(\boldsymbol{\mu}|(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}),\sigma_\mu^2(\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1})\cdot\mathcal{N}(\boldsymbol{w}|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m+k\delta_w\bar{\boldsymbol{w}}),\sigma_w^2(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})$ and $\tilde{\pi}_m=\pi_m c_m^{\boldsymbol{\mu}}c_m^{\boldsymbol{w}}$. We have:

$$\pi_m P(\boldsymbol{\mu},\boldsymbol{w}|T_m)P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w})\propto\tilde{\pi}_m P(\boldsymbol{\mu},\boldsymbol{w}|\widetilde{T}_m).$$

Therefore,

$$\sum_{m=1}^{M}\pi_m P(\boldsymbol{\mu},\boldsymbol{w}|T_m)P(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}|\boldsymbol{\mu},\boldsymbol{w})\propto\sum_{m=1}^{M}\tilde{\pi}_m P(\boldsymbol{\mu},\boldsymbol{w}|\widetilde{T}_m).$$

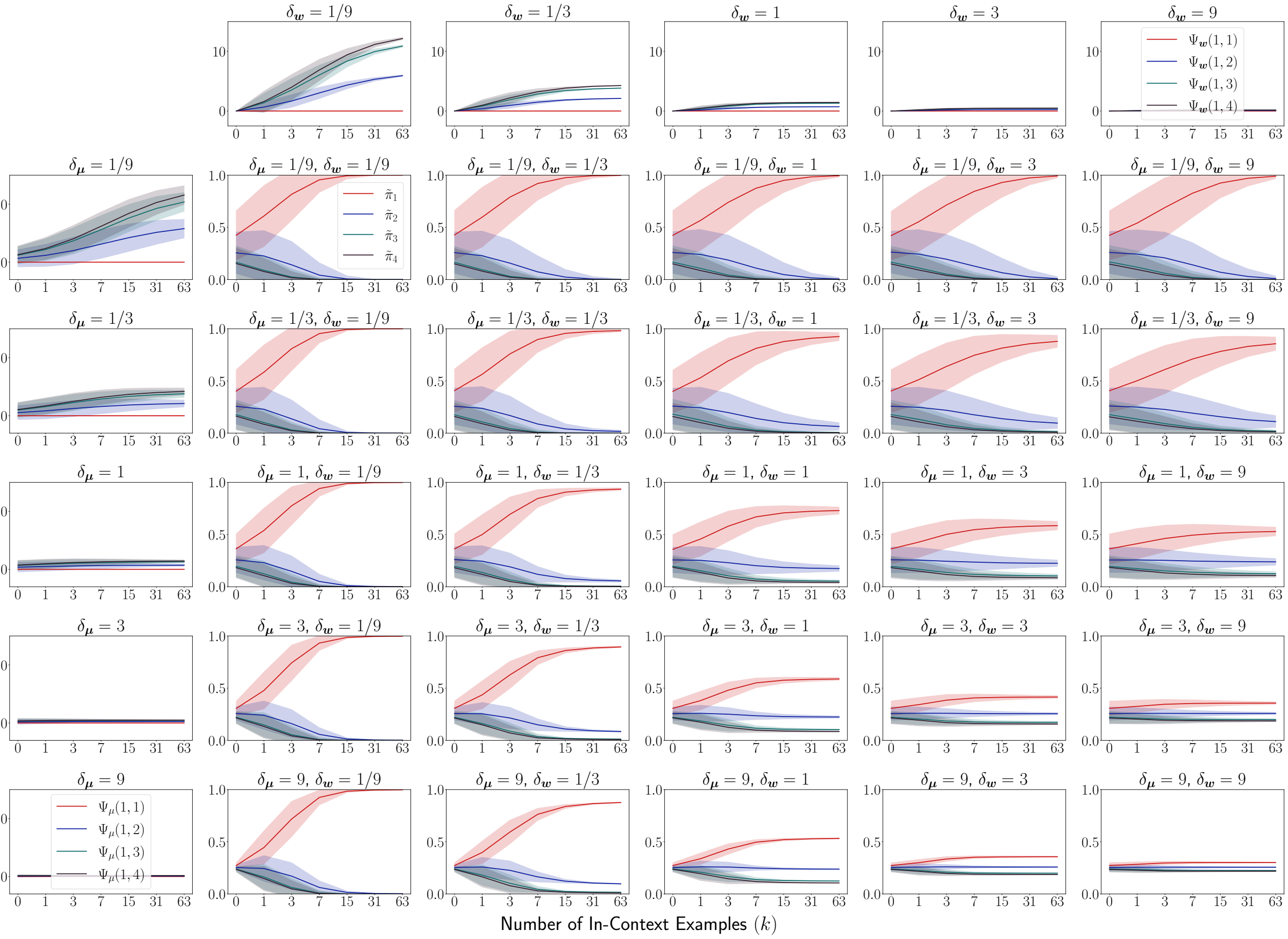


Figure 15: **Numerical analysis on component re-weighting.** The trends of $\Psi_{\boldsymbol\mu}$, $\Psi_{\boldsymbol w}$, and π_m for CR with increasing k under varying task noise parameters.

H. Detailed Analysis of Component Shifting and Re-weighting

H.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 prior center to the in-context task as described in Assumption 3. We assess this via the ratio $r(\alpha,\beta)$ of $\tilde\pi_\alpha$ to $\tilde\pi_\beta$:

$$r(\alpha,\beta)=\frac{\tilde\pi_\alpha}{\tilde\pi_\beta}=\frac{\pi_\alpha C_0 c^{\boldsymbol\mu}_\alpha c^{\boldsymbol w}_\alpha}{\pi_\beta C_0 c^{\boldsymbol\mu}_\beta c^{\boldsymbol w}_\beta}=\frac{\pi_\alpha}{\pi_\beta}\exp(\Psi_{\boldsymbol\mu}(\alpha,\beta)+\Psi_{\boldsymbol w}(\alpha,\beta)), \tag{4}$$

where we define two functions $\Psi_{\boldsymbol\mu}(\alpha,\beta)=\log(c^{\boldsymbol\mu}_\alpha/c^{\boldsymbol\mu}_\beta)$ and $\Psi_{\boldsymbol w}(\alpha,\beta)=\log(c^{\boldsymbol w}_\alpha/c^{\boldsymbol w}_\beta)$ 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:

$$\Psi_{\boldsymbol\mu}(\alpha,\beta)=(\sum_{i=1}^{k+1}\|\boldsymbol\mu_\beta-\boldsymbol x_i\|^2-\sum_{i=1}^{k+1}\|\boldsymbol\mu_\alpha-\boldsymbol x_i\|^2)/(2\sigma_x^2(1+(k+1)\delta_\mu)). \tag{5}$$

(See Appendix H.3.1 for derivation.) Since $\boldsymbol x_i\sim\mathcal N(\boldsymbol\mu^*,\tau_x^2\boldsymbol I)$, choosing $\boldsymbol\mu^*$ closer to $\boldsymbol\mu_\alpha$ tends to make $\Psi_{\boldsymbol\mu}(\alpha,\beta)$ positive and increase 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. 15 shows the numerical computation of $\Psi_{\boldsymbol{\mu}}(\alpha,\beta)$ with varied task noises under the tetrahedron setting (see Appendix B.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 increase as k increases.

Meanwhile, we also have:

$$\lim_{\sigma_\mu\to 0}\Psi_{\boldsymbol{\mu}}(\alpha,\beta) = (\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)/(2\sigma_x^2) \tag{6}$$

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

$$\Psi_{\boldsymbol{w}}(\alpha,\beta) = (\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}})/(2\sigma_w^2). \tag{7}$$

(See Appendix H.3.2 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 4.1) is semi-positive definite, thus choosing $\boldsymbol{w}^*$ closer to $\boldsymbol{w}_\alpha$ tends to make $\Psi_{\boldsymbol{w}}(\alpha,\beta)$ positive and increase faster as k increases. 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. 15 shows the numerical computation of $\Psi_{\boldsymbol{w}}(\alpha,\beta)$ with varied task noises under the tetrahedron setting (see Appendix B.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 increase as k increases. However, one should note 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:

$$\begin{aligned}\lim_{\sigma_w\to 0}\Psi_{\boldsymbol{w}}(\alpha,\beta) &= (\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}})/(2\sigma_w^2)\\ &= (\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2_{k\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}} - \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2_{k\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}})/(2\sigma_y^2)\\ &= (\sum_{i=1}^{k}\|y_i^\beta - y_i^*\|^2 - \sum_{i=1}^{k}\|y_i^\alpha - y_i^*\|^2)/(2\sigma_y^2),\end{aligned} \tag{8}$$

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. 6 and 8 and we have:

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

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. 15. 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.

H.2. Analysis of Component Shifting

The Component Shifting effect in Lemma 4.1 involves shifting the variables $\tilde{\boldsymbol{\mu}}_m$ and $\tilde{\boldsymbol{w}}_m$:

$$\tilde{\boldsymbol{\mu}}_m = (\boldsymbol{I}+(k+1)\delta_\mu\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})^{-1}(\boldsymbol{\mu}_m+(k+1)\delta_\mu\bar{\boldsymbol{\mu}}), \tag{10}$$

$$\tilde{\boldsymbol{w}}_m = (\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m+k\delta_w\bar{\boldsymbol{w}}). \tag{11}$$

The following analyses examine these two variables with increasing k.

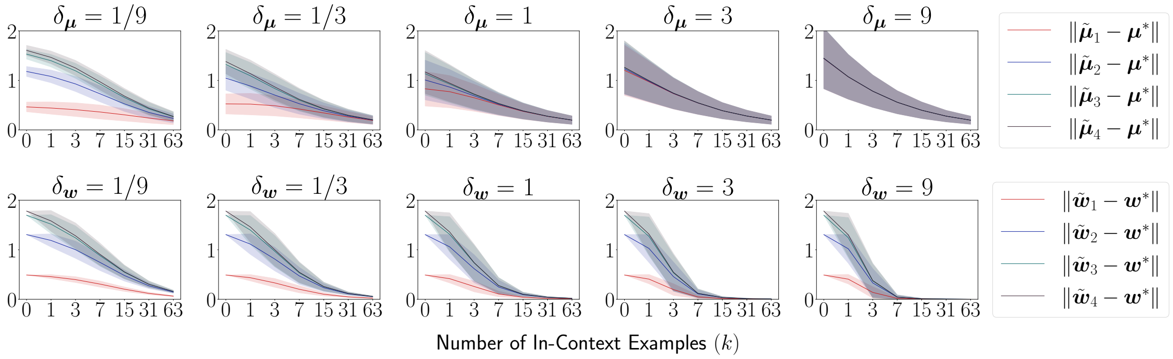


Figure 16: Numerical computations of $\|\tilde{\boldsymbol{\mu}}_m - \boldsymbol{\mu}^*\|$, $\|\tilde{\boldsymbol{w}}_m - \boldsymbol{w}^*\|$ for Component Shifting (CS).

Analysis of $\tilde{\boldsymbol{\mu}}_m$. We provide the derivation of $\tilde{\boldsymbol{\mu}}_m$ in Eq. 10 (see Appendix H.4.1 for details):

$$\tilde{\boldsymbol{\mu}}_m = (\boldsymbol{\mu}_m + k\delta_\mu\bar{\boldsymbol{\mu}})/(1+(k+1)\delta_\mu). \tag{12}$$

Thus, when k increases, $\tilde{\boldsymbol{\mu}}_m$ moves close to the value of $\frac{\sum_{i=1}^k \boldsymbol{x}_i}{k}$ and $\lim_{k\to\infty} \tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}^*$. We also show the numerical computation of the distance between shifted $\tilde{\boldsymbol{\mu}}_m$ and $\boldsymbol{\mu}^*$ in the first row of Fig. 16.

Analysis of $\tilde{\boldsymbol{w}}_m$. We provide the derivation of $\tilde{\boldsymbol{w}}_m$ in Eq. 11 (see Appendix H.4.2 for details):

$$\tilde{\boldsymbol{w}}_m = (\boldsymbol{I} + k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m - \boldsymbol{w}^*) + \boldsymbol{w}^*. \tag{13}$$

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 $\lambda_d(k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) \to \infty$, $\lambda_1((\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}) \to 0$, $\lim_{k\to\infty}(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m - \boldsymbol{w}^*) \leq \lim_{k\to\infty}\lambda_1((\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}) \cdot \|\boldsymbol{w}_m - \boldsymbol{w}^*\| = 0$ and $\lim_{k\to\infty}\tilde{\boldsymbol{w}}_m = \boldsymbol{w}^*$, where $\lambda_d(\boldsymbol{A})$ indicates the minimum eigenvalue of $\boldsymbol{A}$. We also show the numerical computed distance between $\tilde{\boldsymbol{w}}_m$ and $\boldsymbol{w}^*$ in the second row of Fig. 16.

H.3. 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 H.3.1 and the derivation of $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ is collected in Sec H.3.2.

H.3.1. Derivation of $\Psi_\mu(\alpha, \beta)$

This section collects the derivation of $\Psi_{\boldsymbol{\mu}}(\alpha, \beta)$ in Eq. 5 of Sec. H.1:

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

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

H.3.2. Derivation of $\Psi_w(\alpha, \beta)$

This section collects the derivation of $\Psi_{\boldsymbol{w}}(\alpha, \beta)$ in Eq. 7 of Sec. H.1:

$$
\begin{aligned}
&\Psi_{\boldsymbol{w}}(\alpha, \beta) \\
&= \log(c_\alpha^{\boldsymbol{w}} / c_\beta^{\boldsymbol{w}}) \\
&= \log \left(\frac{\exp\left(-\frac{\|\boldsymbol{w}_\alpha\|^2 - \|\boldsymbol{w}_\alpha + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2} \right)}{\exp\left(-\frac{\|\boldsymbol{w}_\beta\|^2 - \|\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2} \right)} \right) \\
&= \frac{\|\boldsymbol{w}_\beta\|^2 - \|\boldsymbol{w}_\beta + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2} - \frac{\|\boldsymbol{w}_\alpha\|^2 - \|\boldsymbol{w}_\alpha + k\delta_w \bar{\boldsymbol{w}}\|^2_{(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}
\end{aligned}
$$

(Note $k\delta_w \bar{\boldsymbol{w}} = \delta_w \sum\limits_{i=1}^{k} \boldsymbol{x}_i y_i = \delta_w \sum\limits_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w}^* = k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \boldsymbol{w}^*$.)

$$
\begin{aligned}
&= \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}
$$

H.4. Derivation Collection of $\tilde{\boldsymbol{\mu}}_m$ and $\tilde{\boldsymbol{w}}_m$

This section collects derivations for $\tilde{\boldsymbol{\mu}}_m$ and $\tilde{\boldsymbol{w}}_m$. The derivation of $\tilde{\boldsymbol{\mu}}_m$ is collected in Appendix H.4.1, and the derivation of $\tilde{\boldsymbol{w}}_m$ is collected in Appendix H.4.2.

H.4.1. Derivation of $\tilde{\boldsymbol{\mu}}_m$

This section collects the derivation of $\tilde{\boldsymbol{\mu}}_m$ in Eq. 12 of Sec. H.1:

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

H.4.2. DERIVATION OF $\tilde{\boldsymbol{w}}_m$

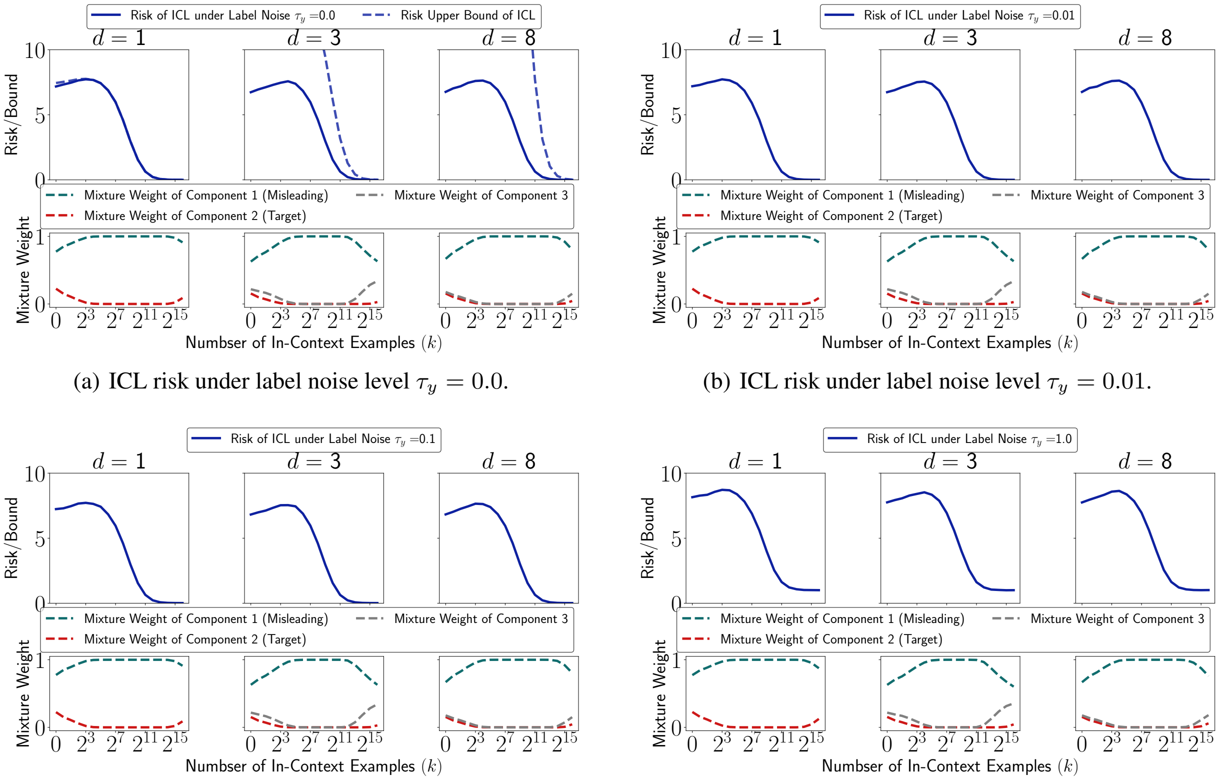
This section collects the derivation of $\tilde{\boldsymbol{w}}_m$ in Eq. 13 of Sec. H.1:

$$
\begin{aligned}
\tilde{\boldsymbol{w}}_m &= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m + k\delta_w \bar{\boldsymbol{w}}) \\
&\quad (\text{Recall } k\delta_w \bar{\boldsymbol{w}} = \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i y_i = \delta_w \sum_{i=1}^{k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{w}^* = k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \boldsymbol{w}^*.) \\
&= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \boldsymbol{w}^*) \\
&= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m - \boldsymbol{w}^* + (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})\boldsymbol{w}^*) \\
&= (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m - \boldsymbol{w}^*) + \boldsymbol{w}^*. \qquad (14)
\end{aligned}
$$

I. Additional Experiments for Early Ascent

I.1. Early Ascent and Bounded Efficacy under Noisy Labels

We further examine phenomena of early ascent and bounded efficacy with noisy labels under varied noise levels. The results show that these two phenomena are robust to label noises to some extend.



(a) ICL risk under label noise level $\tau_y = 0.0$.

(b) ICL risk under label noise level $\tau_y = 0.01$.

(c) ICL risk under label noise level $\tau_y = 0.1$.

(d) ICL risk under label noise level $\tau_y = 1.0$.

Figure 17: **Early ascent under varied label noises.** Results show that the early ascent phenomenon maintains for noise level $\tau_y \in [0, 1.0]$. Label noise level $\sigma_y = 1.0$ is used for pretraining.

I.2. Early Ascent under Non-Linear Regression and Discrete Token Prediction

This section uses Fig. 19 to show the existence of the early ascent phenomenon on non-linear regression and discrete token prediction with our designed distributions of pretraining and in-context samples. Fig. 19(a) shows that the early ascent

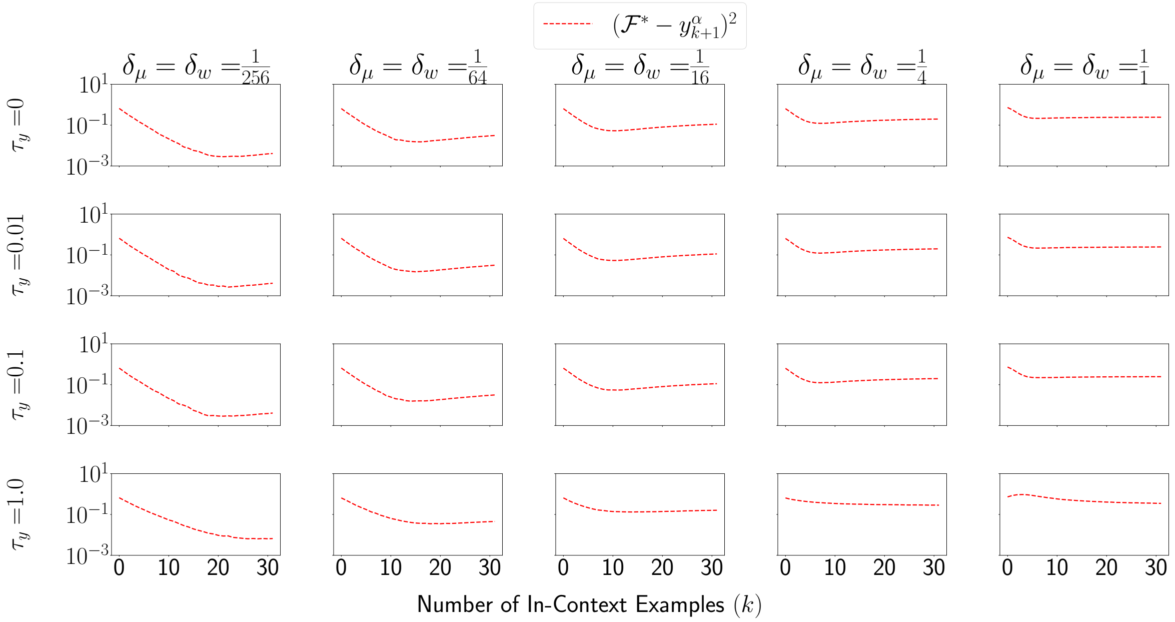
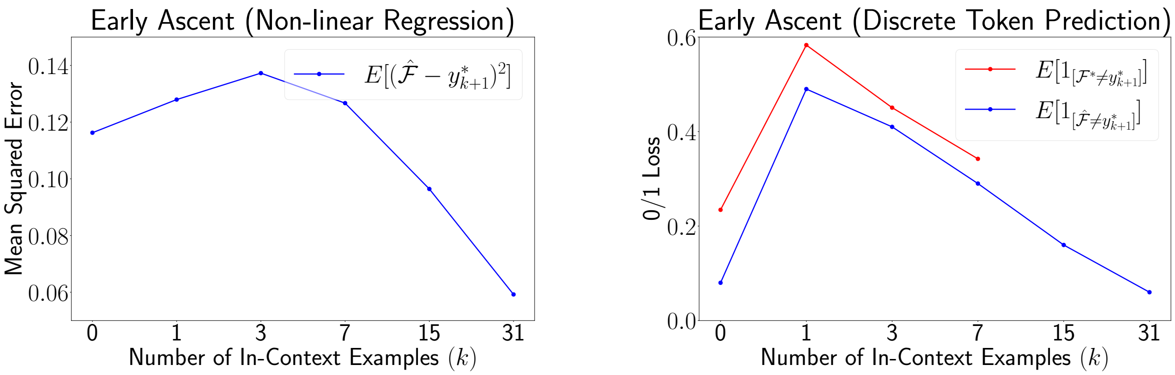


Figure 18: **Bounded efficacy under varied label noises.** Results show that the bounded efficacy phenomenon maintains for noise level $\tau_y \in [0, 0.1]$. Label noise level $\sigma_y = 1.0$ is used for pretraining.

phenomenon exists when a 2-layer neural network with Tanh Activation function serves as the non-linear function, and Fig. 19(b) shows that the early ascent phenomenon exists when the dataset consists of sequences of tokens with discrete values rather than sequences of vectors with continuous values. For the details of experiments including our designed distributions of pretraining and in-context samples, please refer to Sec. I.2.1 for the experiment with non-linear regression and Sec. I.2.2 for the experiment with discrete token prediction.



(a) Experiment under non-linear regressions.

(b) Experiment under discrete token prediction.

Figure 19: $\hat{\mathcal{F}}$ indicates the prediction by a pretrained Transformer model and $\mathcal{F}^*$ indicates the prediction by numerical computation following a Bayes optimal predictor. While we cannot derive the optimal predictor under non-linear regression, we can derive the optimal predictor under discrete token prediction.

I.2.1. Experiment Design for Non-Linear Regression

The following assumption shows the data generation model to generate a non-linear sequence $[\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_K, y_K]$, where $\boldsymbol{x}_i$ is a vector and y_i is a scalar. The non-linear function mapping $\boldsymbol{x}$ to y is highlighted in red in the assumption.

Assumption 5 (Pretraining Data Generative Model for Non-linear Regression).
(a) sample a task from the task distribution: $(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{v}) \sim \mathcal{D}^{\text{prior}}$, $P(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{v}) = \sum_{m=1}^{M} \pi_m P(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{v}|T_m)$, where T_m is the m^{th} mixture component, *i.e.*, $P(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{v}|T_m) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\mu}_m, \sigma_\mu^2 \boldsymbol{I}) \cdot \frac{1}{\sqrt{(2\pi)^{d^2}\sigma_W^{d^2}}} \exp(\frac{\|\boldsymbol{W}-\boldsymbol{W}_m\|_F^2}{2}) \cdot \mathcal{N}(\boldsymbol{v}; \boldsymbol{v}_m, \sigma_v^2 \boldsymbol{I})$, and π_m is the mixture weight. $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability of $\boldsymbol{x}$ in the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\|\cdot\|_F$ indicates the Frobenius norm, $\sum_{m=1}^{M} \pi_m = 1$, $0 < \pi_m < 1$, $(\boldsymbol{\mu}_m, \boldsymbol{w}_m)$ is the center of the mixture component T_m, and all components share the same covariance matrix controlled by σ_μ, σ_W, and σ_v;
(b) input variable distribution: within a sequence, $\forall i \in [K]$, $\boldsymbol{x}_i \sim \mathcal{D}_{\boldsymbol{x}}(\boldsymbol{\mu})$, $P(\boldsymbol{x}|\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sigma_x^2 \boldsymbol{I})$;
(c) label distribution: within a sequence, $\forall i \in [K]$, $y_i|\boldsymbol{x}_i \sim \mathcal{D}_{y|\boldsymbol{x}_i}(\boldsymbol{W}, \boldsymbol{v})$, $P(y_i|\boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{v}) = \mathcal{N}(y_i|\langle \tanh(\boldsymbol{W}\boldsymbol{x}_i), \boldsymbol{v}\rangle, \sigma_y^2)$, where $\tanh()$ is a Tanh Activation function;
(d) $\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\mu}_m, \boldsymbol{v}, \boldsymbol{v}_m \in \mathbb{R}^d$, and $\boldsymbol{W}, \boldsymbol{W}_m \in \mathbb{R}^{d\times d}$.

For experimental setting of Fig. 19(a), we set $d = 2$, $\sigma_\mu = 1, \sigma_W = \sigma_v = 0.5$, $\sigma_x = \sigma_y = 1$, $M = 2$, $\pi_1 = 0.1, \pi_2 = 0.9$, $\boldsymbol{\mu}_1 = [1, 0]^\top, \boldsymbol{\mu}_2 = [0, 1]^\top$, $\boldsymbol{W}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{W}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, and $\boldsymbol{v}_1 = [1, 0]^\top, \boldsymbol{v}_2 = [0, 1]^\top$. In-context samples follows task $(\boldsymbol{\mu}^*, \boldsymbol{W}^*, \boldsymbol{v}^*)$, where $\boldsymbol{\mu}^* = \boldsymbol{\mu}_1$, $\boldsymbol{W}^* = \boldsymbol{W}_2$, $\boldsymbol{v}^* = \boldsymbol{v}_2$, and $\sigma_y = 1$. Notice that although we add label noise to in-context samples, when evaluating the prediction, we calculate error/loss based on the clean label.

I.2.2. Experiment Design for Discrete Token Prediction

The following assumption shows the data generation model to generate a non-linear sequence $[x_1, y_1, \dots, x_K, y_K]$, where x_i and y_i are both integers (discrete tokens).

Assumption 6 (Pretraining Data Generative Model for Discrete Token Prediction).
(a) sample a task from the task distribution: $(\mu, w) \sim \mathcal{D}^{\text{prior}}$, $\mu \in [M], w \in [M]$, $P(\mu, w) = \sum_{m=1}^{M} \pi_m P(\mu, w|T_m)$, where T_m is the m^{th} mixture component, *i.e.*, $P(\mu, w|T_m) = 1_{[w=w_m]}((1-(M-1)\sigma_\mu)1_{[\mu=\mu_m]} + \sigma_\mu 1_{[\mu\neq\mu_m]})$, and π_m is the mixture weight.
(b) input variable distribution: within a sequence, $\forall i \in [K]$, $x_i \sim \mathcal{D}_x(\mu)$, $P(x_i|\mu) = (1-(M-1)\sigma_x)1_{[x=\mu]} + \sigma_x 1_{[x\neq\mu]}$;
(c) label distribution: within a sequence, $\forall i \in [K]$, $y_i|x_i \sim \mathcal{D}_{y|x_i}(w)$, $P(y_i|x_i, w) = (1-(M-1)\sigma_y)1_{[y_i=x_i+w \mod M]} + \sigma_y 1_{[y_i\neq x_i+w \mod M]}$.

For experimental setting of Fig. 19(b), we set $M = 6$,$\pi_1 = 0.04, \pi_3 = 0.481, \pi_5 = 0.479, \pi_2 = \pi_4 = \pi_6 = 0$, $\sigma_\mu = 0.05$, $\sigma_x = 0.04$, $\sigma_y = 0.13$, $\mu_1 = w_1 = 1, \mu_3 = w_3 = 3, \mu_5 = w_5 = 5$. In-context samples follows task (μ^*, w^*), where $\mu^* = \mu_1$, $w^* = w_3$, and $\sigma_y = 0.13$. Notice that although we add label noise to in-context samples, when evaluating the prediction, we calculate error/loss based on the clean label.

J. Mathematical Derivation for Early Ascent

We show that the early ascent phenomenon occurs under a specific setting in Sec. J.1. Then, we give formal theory with proof to show when early ascent happens in Sec. J.2.

J.1. A Specific Setting of Early Ascent

To have a cleaner mathematical understanding of this phenomenon, this section uses the setting of $d = 1$, the first row, in Table 2 to show the mathematical logic. (Some parameter settings are described in Table 2's caption.) Following Theorem 5.1, the upper bound of ICL risk is as follows:

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

$$
\begin{aligned}
&= 4\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\frac{\tilde{\pi}_1}{\tilde{\pi}_1 + \tilde{\pi}_2} \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2 \right] \\
&\text{(Recall } \frac{\tilde{\pi}_1}{\tilde{\pi}_2} = r(1,2) \text{ as Eq. 4.)} \\
&= 4\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\frac{r(1,2)}{1 + r(1,2)} \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2 \right].
\end{aligned}
$$

Noticing $\delta_\mu = \frac{0.05^2}{1^2}$ and $\delta_w = \frac{0.05^2}{2^2}$ are 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 $\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\frac{r(1,2)}{1+r(1,2)} \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2 \right] \approx \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\frac{r(1,2)}{1+r(1,2)} \|\boldsymbol{x}_{k+1}\|^2 \right]$ and a larger $r(1,2)$ means a larger upper bound. In the following, we will examine whether the increase of k leads to the increase of $r(1,2)$.

Following Eq. 4:

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

We first analyze $\Psi_{\boldsymbol{\mu}}(1,2)$, following Eq. 5:

$$
\begin{aligned}
\mathbb{E}[\Psi_{\boldsymbol{\mu}}(1,2)] &= \mathbb{E}\left[\frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu}_2 - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1} \|\boldsymbol{\mu}_1 - \boldsymbol{x}_i\|^2}{2\sigma_x^2 (1 + (k+1)\delta_\mu)} \right] \\
&\text{(Since } \delta_\mu \approx 0, \text{ thus when } k \text{ is small, we have:)} \\
&\approx \mathbb{E}\left[\frac{\sum_{i=1}^{k+1} \|\boldsymbol{\mu}_2 - \boldsymbol{x}_i\|^2 - \sum_{i=1}^{k+1} \|\boldsymbol{\mu}_1 - \boldsymbol{x}_i\|^2}{2\sigma_x^2} \right] \\
&= \frac{k+1}{2\sigma_x^2} \mathbb{E}\left[\|\boldsymbol{\mu}_2 - \boldsymbol{x}_1\|^2 - \|\boldsymbol{\mu}_1 - \boldsymbol{x}_1\|^2 \right] \\
&= \frac{k+1}{2\sigma_x^2} (\mathbb{E}[\|\boldsymbol{\mu}_2 - \boldsymbol{x}_1\|^2] - \mathbb{E}[\|\boldsymbol{\mu}_1 - \boldsymbol{x}_1\|^2]) \\
&= \frac{k+1}{2\sigma_x^2} (\mathbb{E}[\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}^*\|^2] + \tau_x^2) - (\mathbb{E}[\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}^*\|^2] + \tau_x^2) \\
&(\boldsymbol{\mu}^* \text{ is the same as } \boldsymbol{\mu}_1, \text{ but different from } \boldsymbol{\mu}_2.) \\
&= \frac{k+1}{2\sigma_x^2} (\mathbb{E}[\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}^*\|^2] - 0) \\
&= \frac{k+1}{2 \times 1^2} \times 2^2 \\
&= 2(k+1).
\end{aligned}
$$

We then analyze $\Psi_{\boldsymbol{w}}(1,2)$, following Eq. 7:

$$
\begin{aligned}
\mathbb{E}[\Psi_{\boldsymbol{w}}(1,2)] &= \mathbb{E}\left[-\frac{\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2_{\boldsymbol{I} - (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2} \right] \\
&\text{(Since } \delta_w \approx 0, \text{ thus when } k \text{ is small, we have:)} \\
&\approx -\mathbb{E}\left[\frac{(\boldsymbol{w}_1 - \boldsymbol{w}^*)^\top k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} (\boldsymbol{w}_1 - \boldsymbol{w}^*)}{2\sigma_w^2} \right] \\
&\text{(Notice the feature dimension } d = 1, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^k \|\boldsymbol{x}_i\|^2}{k}.) \\
&\approx -\mathbb{E}\left[\frac{\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 k\delta_w \sum_{i=1}^k \|\boldsymbol{x}_i\|^2}{2\sigma_w^2} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= -\mathbb{E}\left[\frac{2\sum_{i=1}^{k}\|\boldsymbol{x}_i\|^2}{\sigma_y^2}\right] \\
&= -\frac{2k}{\sigma_y^2}\mathbb{E}\left[\|\boldsymbol{x}_1\|^2\right] \\
&= -\frac{2k}{\sigma_y^2}(\|\boldsymbol{\mu}^*\|^2+\tau_x^2) \\
&= -\frac{2k}{2^2}\times(1+1) = -k.
\end{aligned}
$$

Therefore, when k is small, $r(1,2) = \Psi_{\boldsymbol{\mu}}(1,2)+\Psi_{\boldsymbol{w}}(1,2) \approx \exp(k+2)$, and the upper bound is approximately equal to:

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

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

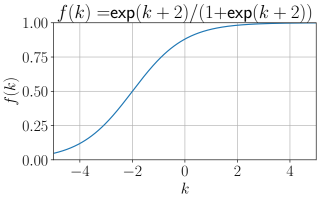


Figure 20: Illustration of the function $\exp(k+2)/(1+\exp(k+2))$

J.2. Theorem of Early Ascent

Theorem 5.2 (Early Ascent). Assume $\mathbb{E}_{\boldsymbol{x}_1}\left[(\mathcal{F}^*(\boldsymbol{x}_1)-\langle\boldsymbol{w}^*,\boldsymbol{x}_1\rangle)^2\right] < \mathbb{E}_{\boldsymbol{x}_1}\left[\langle\boldsymbol{x}_1,\boldsymbol{w}_\alpha-\boldsymbol{w}^*\rangle^2\right]$, 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:

$$
\exists k \geq 1 \text{ s.t. } \mathbb{E}_{\boldsymbol{x}_1}\left[(\mathcal{F}^*(\boldsymbol{x}_1)-\langle\boldsymbol{w}^*,\boldsymbol{x}_1\rangle)^2\right] < \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[(\mathcal{F}^*(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-\langle\boldsymbol{w}^*,\boldsymbol{x}_{k+1}\rangle)^2\right].
$$

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:

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

$$
\begin{aligned}
&(\text{Following Eq. 9, we have } \lim_{(\sigma_\mu,\sigma_w)\to(0,0)}\Psi_{\boldsymbol{\mu}}(m,1)+\Psi_{\boldsymbol{w}}(m,1) = \frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2-\|\boldsymbol{\mu}_1-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2} \\
&\qquad+\sum_{i=1}^{k}\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2-\|\boldsymbol{\mu}_1-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2-\|y_i^1-y_i^*\|^2}{2\sigma_y^2}\right))
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{k\to\infty}\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left\langle\frac{\sum_{m=1}^{M}\pi_m\exp\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2}+\sum_{i=1}^{k}(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2}{2\sigma_y^2})\right)(\boldsymbol{w}_m-\boldsymbol{w}^*)}{\sum_{m=1}^{M}\pi_m\exp\left(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2}+\sum_{i=1}^{k}(\frac{\|\boldsymbol{\mu}_m-\boldsymbol{x}_i\|^2}{2\sigma_x^2}+\frac{\|y_i^m-y_i^*\|^2}{2\sigma_y^2})\right)},\boldsymbol{x}_{k+1}\right\rangle^2\right] \\
&= \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\langle\boldsymbol{w}_\alpha-\boldsymbol{w}^*,\boldsymbol{x}_{k+1}\rangle^2] \\
&= \mathbb{E}_{\boldsymbol{x}_1}[\langle\boldsymbol{w}_\alpha-\boldsymbol{w}^*,\boldsymbol{x}_1\rangle^2],
\end{aligned}
$$

where $\alpha = \arg\min_m \frac{\|\boldsymbol{\mu}_m-\boldsymbol{\mu}^*\|^2}{2\sigma_x^2}+\frac{\|(\boldsymbol{w}_m-\boldsymbol{w}^*)^\top\boldsymbol{\mu}^*\|^2+d\tau_x^2\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2}{2\sigma_y^2}$. □

K. Proof Tools

This section introduces the inequalities used in our proofs for Theorems 5.1 (finegrained upper bound for ICL risk), 6.1 (upper bound for ICL with biased labels), C.1 (coarse upper bound for ICL risk) and Lemma 6.2 ((informal) upper bound for zero-shot ICL):

K.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.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 (Boucheron et al., 2013):

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

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

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

K.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:

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

where $\|\cdot\|$ indicates the L_2 norm.

Proof.

$$\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\|^2$$
$$= \sum_{j=1}^d \left(\frac{\sum_{i=1}^k \epsilon_{i,j}}{k}\right)^2$$
$$= \frac{\tau_x^2}{k} \sum_{j=1}^d \left(\frac{\sum_{i=1}^k \epsilon_{i,j}}{\tau_x \sqrt{k}}\right)^2$$

(Notice $\epsilon_{i,j} \sim \mathcal{N}(0, \tau_x^2)$ and let $Z_j = \dfrac{\sum_{i=1}^k \epsilon_{i,j}}{\tau_x\sqrt{k}} \sim \mathcal{N}(0,1)$.)

$$= \frac{\tau_x^2 d}{k} \frac{\sum_{i=1}^d Z_i^2}{d}.$$

Therefore, by applying Appendix K.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.4. Eigenvalue Concentration Bound

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

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

where $L = \tau_x^2(1-\frac{t}{2}-\gamma)^2 - 2\tau_x\gamma\sqrt{1+t}$, $U = 1 + \tau_x^2(1+\frac{t}{2}+\gamma)^2 + 2\tau_x\gamma\sqrt{1+t}$, $\lambda_i(\boldsymbol{A})$ 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 $\boldsymbol{A} = \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k} = \frac{\sum_{i=1}^k (\boldsymbol{\mu}+\boldsymbol{\epsilon}_i)(\boldsymbol{\mu}+\boldsymbol{\epsilon}_i)^\top}{k} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top}{k} + \frac{\sum_{i=1}^k (\boldsymbol{\mu}\boldsymbol{\epsilon}_i^\top + \boldsymbol{\epsilon}_i\boldsymbol{\mu}^\top)}{k}$, then consider the eigenvalue bound of each of them.

For the first component $\boldsymbol{\mu}\boldsymbol{\mu}^\top$, 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 $\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top}{k}$. Following Vershynin (2018, Theorem 4.6.1, p. 97), we have for any $1 - \sqrt{\frac{d}{k}} > s > 0$:

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

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

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

(Notice by Norm Tail Bound in Appendix K.3, we have $P\left(\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\| > \sqrt{\frac{\tau_x^2 d}{k}(1+t)}\right) \leq \exp\left(-\frac{kt^2}{8}\right)$.)

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

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

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

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

As a summary, we have:

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

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 i^{th} biggest eigenvalue of the matrix $\boldsymbol{A}$.

L. ICL to Learn the In-Context Function

This section introduces the proof of Theorem C.1 (coarse upper bound for ICL risk) and Theorem 5.1 (finegrained upper bound for ICL risk). The upper bound of Theorem 5.1 is derived at Eq. 15.

Proof. Assuming we are using in-context examples following Assumption 3, *i.e.*, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I}), y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^*\rangle$, $\|\boldsymbol{\mu}^*\| = \|\boldsymbol{w}^*\| = 1$, 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 $(\boldsymbol{w}^*)$ of the in-context task $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$. 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 $\mathcal{F}^*$ under Assumption 6 for pretraining data generation. 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[\left(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{w}^*, \boldsymbol{x}_{k+1}\rangle\right)^2\right] \\
&\text{(By Corollary 4.4.)} \\
&= \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left(\sum_{m=1}^M \tilde{\pi}_m \langle \tilde{\boldsymbol{w}}_m, \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_{m=1}^M \tilde{\pi}_m (\tilde{\boldsymbol{w}}_m - \boldsymbol{w}^*), \boldsymbol{x}_{k+1}\right\rangle\right)^2\right] \\
&\text{(See Eq. 14 for the derivation of } \tilde{\boldsymbol{w}}_m.\text{)} \\
&= \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left(\left\langle \sum_{m=1}^M \tilde{\pi}_m ((\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_m - \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.)} \\
&= \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\left\langle \sum_{m=1}^M \tilde{\pi}_m \boldsymbol{A}(\boldsymbol{w}_m - \boldsymbol{w}^*), \boldsymbol{x}_{k+1}\right\rangle^2\right] \\
&\text{(Notice } \left(\sum_{\beta=1}^M \tilde{\pi}_\beta a_\beta\right)^2 \leq \sum_{\beta=1}^M \tilde{\pi}_\beta a_\beta^2\text{, since } \mathbb{E}[a]^2 \leq \mathbb{E}[a^2].\text{)} \\
&\leq \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{m=1}^M \tilde{\pi}_m \langle \boldsymbol{A}(\boldsymbol{w}_m - \boldsymbol{w}^*), \boldsymbol{x}_{k+1}\rangle^2\right] \\
&= \sum_{m=1}^M \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_m ((\boldsymbol{w}_m - \boldsymbol{w}^*)^\top \boldsymbol{A}\boldsymbol{x}_{k+1})^2\right] \\
&\leq \sum_{m=1}^M \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_m \|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2 \lambda_1(\boldsymbol{A})^2 \|\boldsymbol{x}_{k+1}\|^2\right] \\
&= \sum_{m=1}^M \|\boldsymbol{w}_m - \boldsymbol{w}^*\|^2 \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_m \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2\right] \qquad (15) \\
&\leq 4\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{m=1}^M \tilde{\pi}_m \|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2\right] \\
&= 4\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2 \lambda_1(\boldsymbol{A})^2\right] \\
&\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}.\text{)} \\
&= 4\mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2\right] \mathbb{E}_{\mathcal{S}_k}\left[\lambda_1^2(\boldsymbol{A})\right] \\
&= 4(1 + d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}\left[\lambda_1^2(\boldsymbol{A})\right].
\end{aligned}
$$

We further simplify $\mathbb{E}_{\mathcal{S}_k}\left[\lambda_1^2(\boldsymbol{A})\right]$ using Lemma K.1:

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

$$
\begin{aligned}
&\leq 4(1+d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}\left[\lambda_1^2(\boldsymbol{A})\right]\\
&\leq 4(1+d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}\left[\left(\frac{1}{1+k\delta_w\lambda_d(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k})}\right)^2\right]
\end{aligned}
$$

(By applying Lemma K.1 to $\dfrac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}$.)

$$
\begin{aligned}
&\leq 4(1+d\tau_x^2)\mathbb{E}_{\mathcal{S}_k}\left[\left(\frac{1}{1+k\delta_w \mathrm{L}}\right)^2\right]\\
&\leq 4(1+d\tau_x^2)\left(\left(\frac{1}{1+k\delta_w(\tau_x^2(1-\frac{t}{2}-\gamma)^2-2\tau_x\gamma\sqrt{1+t})}\right)^2+3\exp\left(-\frac{kt^2}{8}\right)\right).
\end{aligned}
$$

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

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

□

We further validate our analysis with numerical computations in Fig. 21, including the trend of $\tilde{\pi}_m$ for $m\in[M]$, $\lambda_j\left(\delta_w\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right)$ for $j\in[d]$, $\lambda_j\left(\boldsymbol{I}+\delta_w\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)$ for $j\in[d]$, $1/\|\tilde{\boldsymbol{w}}-\boldsymbol{w}^*\|$, $1/\mathbb{E}[\mathcal{F}^*(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-y_{k+1}^*]$, and $1/\mathbb{E}[(\mathcal{F}^*(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-y_{k+1}^*)^2]$ as k increases.

L.1. Case When In-context Input Variable Spans in Subspace

In this section, we refine Eq. 15 for the finegrained bound in Theorem 5.1. Specifically, we refine the following inequality for case when in-context input variable $\boldsymbol{x}_i$ only spans in the subspace of $\mathbb{R}^d$, resulting in $\lambda_1(\boldsymbol{A})=1$ constantly as mentioend in Theorem 5.1:

$$
\begin{aligned}
&\sum\nolimits_{m=1}^M \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_m((\boldsymbol{w}_m-\boldsymbol{w}^*)^\top\boldsymbol{A}\boldsymbol{x}_{k+1})^2\right]\\
&\leq\sum\nolimits_{m=1}^M \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_m\|\boldsymbol{w}_m-\boldsymbol{w}^*\|^2\lambda_1(\boldsymbol{A})^2\|\boldsymbol{x}_{k+1}\|^2\right],
\end{aligned}
$$

where $\boldsymbol{A}=(\boldsymbol{I}+\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}$ is derived in Lemma 4.1. Violating Assumption 3(a), in this section we consider the case that $\boldsymbol{x}_i\sim\mathcal{N}(\boldsymbol{\mu},\mathrm{diag}(\underbrace{1,\dots,1}_{d'},0,\dots,0))$, where $\boldsymbol{\mu}=[p,\underbrace{0,\dots,0}_{d'-1},q,0,\dots,0]^\top$. (If $\boldsymbol{\mu}$ does not follows the format $[p,\underbrace{0,\dots,0}_{d'-1},q,0,\dots,0]^\top$, we can always rotate the coordinates so $\boldsymbol{\mu}$ has this format.) Therefore, we have matrix $\boldsymbol{A}$ (after rotation) with the following format:

$$
\boldsymbol{A}=\begin{cases}
\begin{bmatrix}\boldsymbol{I}_{d'\times d'}+\sum_{i=1}^k \boldsymbol{x}_{i,1:d'}\boldsymbol{x}_{i,1:d'}^\top & \boldsymbol{0}_{d'\times(d-d')}\\ \boldsymbol{0}_{(d-d')\times d'} & \boldsymbol{I}_{(d-d')\times(d-d')}\end{bmatrix}^{-1}, \text{ if } q=0\\
\begin{bmatrix}\boldsymbol{I}_{(d'+1)\times(d'+1)}+\sum_{i=1}^k \boldsymbol{x}_{i,1:(d'+1)}\boldsymbol{x}_{i,1:(d'+1)}^\top & \boldsymbol{0}_{(d'+1)\times(d-d'-1)}\\ \boldsymbol{0}_{(d-d'-1)\times(d'+1)} & \boldsymbol{I}_{(d-d'-1)\times(d-d'-1)}\end{bmatrix}^{-1}, \text{ if } q>0
\end{cases}
$$

where $\boldsymbol{x}_{i,1:d'}=[\boldsymbol{x}_{i,1},\boldsymbol{x}_{i,2},\dots,\boldsymbol{x}_{i,d'}]^\top$, $\boldsymbol{I}_{a\times a}$ indicates an identity matrix with shape a by a, and $\boldsymbol{0}_{a\times b}$ indicates a zero matrix with shape a by b. Finally, we can revise the upper bound for the case when $\boldsymbol{x}_i$ only spans in a subspace of $\mathbb{R}^d$ using the new format of $\boldsymbol{A}$ as follows:

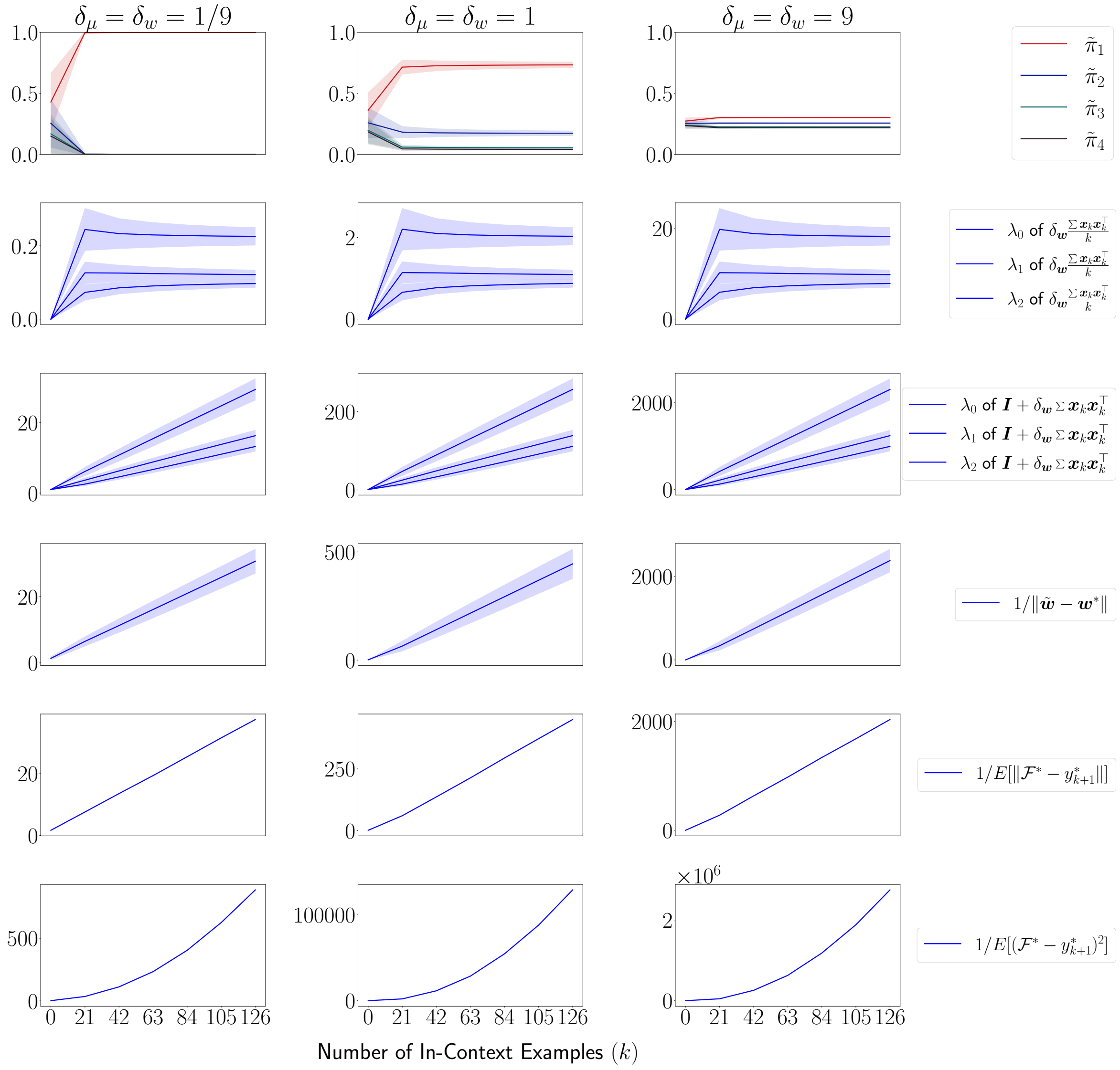


Figure 21: The numerical computation of the task learning. The second and third rows show the eigenvalues of the matrices $\delta_w \frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}$ and $\boldsymbol{I} + \delta_w \sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top$. The fourth row shows 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.

When $q = 0$, we have:

$$
\begin{aligned}
&\sum_{m=1}^{M} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\tilde{\pi}_m ((\boldsymbol{w}_m - \boldsymbol{w}^*)^\top \boldsymbol{A} \boldsymbol{x}_{k+1})^2\right] \\
&\leq \sum_{m=1}^{M} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\tilde{\pi}_m ((\boldsymbol{w}_m - \boldsymbol{w}^*)_{1:d'}^\top \boldsymbol{A}_{1:d',1:d'} \boldsymbol{x}_{k+1,1:d'} + (\boldsymbol{w}_m - \boldsymbol{w}^*)_{(d'+1):d}^\top \boldsymbol{I}_{(d-d')\times(d-d')} \boldsymbol{x}_{k+1,(d'+1):d})^2\right] \\
&\leq \sum_{m=1}^{M} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\tilde{\pi}_m (\|(\boldsymbol{w}_m - \boldsymbol{w}^*)_{1:d'}\|^2 \lambda_1(\boldsymbol{A}_{1:d',1:d'})^2 \|\boldsymbol{x}_{k+1,1:d'}\|^2 + \|(\boldsymbol{w}_m - \boldsymbol{w}^*)_{(d'+1):d}\|^2 \|\boldsymbol{x}_{k+1,(d'+1):d}\|^2)\right], \\
&(\text{Notice } \|\boldsymbol{x}_{k+1,(d'+1):d}\|^2 = 0) \\
&= \sum_{m=1}^{M} \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}} \left[\tilde{\pi}_m \|(\boldsymbol{w}_m - \boldsymbol{w}^*)_{1:d'}\|^2 \lambda_1(\boldsymbol{A}_{1:d',1:d'})^2 \|\boldsymbol{x}_{k+1,1:d'}\|^2\right],
\end{aligned}
$$

When $q > 0$, we skip the analysis since the analysis for $q > 0$ is the same as the analysis for $q = 0$. The only difference is that d' for $q > 0$ is one bigger than d' for $q = 0$.

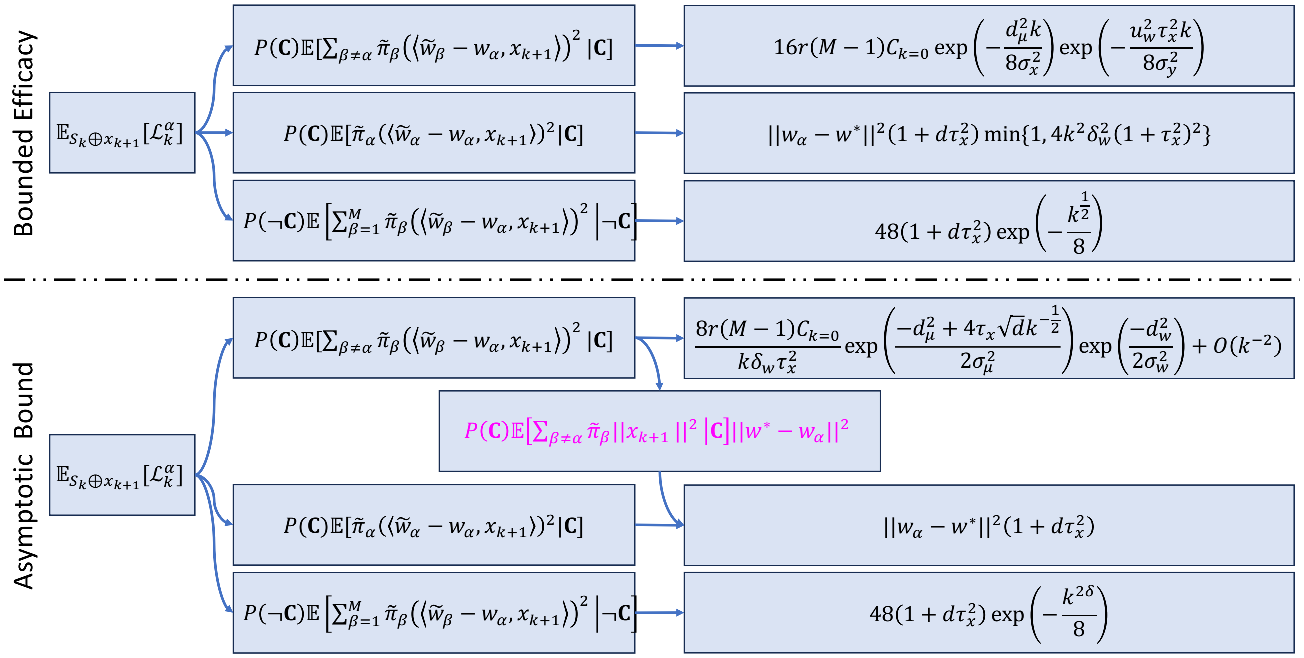

Bounded Efficacy
$\mathbb{E}_{S_k \oplus x_{k+1}}[\mathcal{L}_k^\alpha]$
$P(\mathbf{C})\mathbb{E}[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta(\langle\tilde{w}_\beta - w_\alpha, x_{k+1}\rangle)^2\,|\mathbf{C}]$
$16r(M-1)C_{k=0}\exp\left(-\frac{d_\mu^2 k}{8\sigma_x^2}\right)\exp\left(-\frac{u_w^2\tau_x^2 k}{8\sigma_y^2}\right)$
$P(\mathbf{C})\mathbb{E}[\tilde{\pi}_\alpha(\langle\tilde{w}_\alpha - w_\alpha, x_{k+1}\rangle)^2|\mathbf{C}]$
$||w_\alpha - w^*||^2(1+d\tau_x^2)\min\{1,4k^2\delta_w^2(1+\tau_x^2)^2\}$
$P(\neg\mathbf{C})\mathbb{E}\left[\sum_{\beta=1}^M\tilde{\pi}_\beta(\langle\tilde{w}_\beta - w_\alpha, x_{k+1}\rangle)^2\,\middle|\neg\mathbf{C}\right]$
$48(1+d\tau_x^2)\exp\left(-\frac{k^{\frac12}}{8}\right)$
Asymptotic Bound
$\mathbb{E}_{S_k \oplus x_{k+1}}[\mathcal{L}_k^\alpha]$
$P(\mathbf{C})\mathbb{E}[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta(\langle\tilde{w}_\beta - w_\alpha, x_{k+1}\rangle)^2\,|\mathbf{C}]$
$\frac{8r(M-1)C_{k=0}}{k\delta_w\tau_x^2}\exp\left(\frac{-d_\mu^2+4\tau_x\sqrt{d}k^{-\frac12}}{2\sigma_\mu^2}\right)\exp\left(\frac{-d_w^2}{2\sigma_w^2}\right)+O(k^{-2})$
$P(\mathbf{C})\mathbb{E}[\sum_{\beta\neq\alpha}\tilde{\pi}_\beta||x_{k+1}||^2\,|\mathbf{C}]||w^*-w_\alpha||^2$
$P(\mathbf{C})\mathbb{E}[\tilde{\pi}_\alpha(\langle\tilde{w}_\alpha - w_\alpha, x_{k+1}\rangle)^2|\mathbf{C}]$
$||w_\alpha - w^*||^2(1+d\tau_x^2)$
$P(\neg\mathbf{C})\mathbb{E}\left[\sum_{\beta=1}^M\tilde{\pi}_\beta(\langle\tilde{w}_\beta - w_\alpha, x_{k+1}\rangle)^2\,\middle|\neg\mathbf{C}\right]$
$48(1+d\tau_x^2)\exp\left(-\frac{k^{2\delta}}{8}\right)$

Figure 22: Proof roadmap of ICL with biased labels, Theorem. 6.1.

M. ICL with Biased Labels to Retrieve A Task

This section details the proof of Theorem 6.1, with Fig.22 serving as a visual guide. The non-asymptotic bound for the bounded efficacy phenomenon and the asymptotic bound share the same foundational elements in the proof. However, they are different in handling the components marked in pink. Fig. 22 is thus provided to offer a clearer understanding of its overall framework and assist readers in navigating through the proof. In the following sections, Sec. M.1 introduces the non-asymptotic bound revealing the bounded efficacy phenomenon, and Sec. M.2 introduces the asymptotic bound.

M.1. Non-Asymptotic Bound for the Bounded Efficacy Phenomenon

This section proves the non-asymptotic bound in Theorem 6.1: Consider a next-token predictor attaining the optimal pretraining risk. When δ_μ and δ_w are sufficiently small, there exists a particular interval (refer to Sec.M.1.5 for the interval) for k such that ICL risk with biased labels is upper bounded by:

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < C_3\exp\left(-k\left(\frac{d_{\boldsymbol{\mu}}^2}{8\sigma_x^2}+\frac{u_{\boldsymbol{w}}^2\tau_x^2}{8\sigma_y^2}\right)\right)+48(1+d\tau_x^2)\exp\left(-\frac{k^{\frac12}}{8}\right)$$
$$+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2(1+d\tau_x^2)\min\{1,4k^2{\delta_w}^2(1+\tau_x^2)^2\}.$$

where $\mathcal{L}_k^\alpha=(\mathcal{F}(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-y_{k+1}^\alpha)^2=(\mathcal{F}(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-\langle\boldsymbol{x}_{k+1},\boldsymbol{w}_\alpha\rangle)^2$ C_3 is a constant depending on the prior setting, τ_x, and $(\boldsymbol{\mu}^*,\boldsymbol{w}^*)$. With small k, the first and second terms dominate and exponential decay. With large k, the third term dominates and increases. Thus, the upper bound reveals a bounded efficacy phenomenon.

Proof. Assuming we are using in-context examples following Assumptions 3 and 4, *i.e.*, $\boldsymbol{x}_i\sim\mathcal{N}(\boldsymbol{\mu}^*,\tau_x^2\boldsymbol{I}), y_i=\langle\boldsymbol{x}_i,\boldsymbol{w}^*\rangle$, $\|\boldsymbol{\mu}^*\|=\|\boldsymbol{w}^*\|=1$, and we aim to retrieve the function $\boldsymbol{w}_\alpha$ of the prior center $(\boldsymbol{\mu}_\alpha,\boldsymbol{w}_\alpha)$ which is close to the in-context task. Let $\mathcal{L}_k^\alpha$ indicate the squared risk $(\mathcal{F}^*(\mathcal{S}_k\oplus\boldsymbol{x}_{k+1})-\langle\boldsymbol{x}_{k+1},\boldsymbol{w}_\alpha\rangle)^2$, 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 risk, we consider $\boldsymbol{x}_i\sim\mathcal{N}(\boldsymbol{\mu}^*,\tau_x^2\boldsymbol{I})$ in two cases: (1) $\mathbf{C}$: L $<\lambda_d\left(\frac{\sum_{i=1}^k\boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right)\leq\lambda_1\left(\frac{\sum_{i=1}^k\boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right)<$ U and $\left\|\frac{\sum_{i=1}^k\boldsymbol{\epsilon}_i}{k}\right\|<\tau_x\sqrt{\gamma(1+t)}$ (see Lemma K.1 for t, γ, L and U) and (2) $\neg\mathbf{C}$: at least one of the previous inequalities does not hold. Following Lemma K.1, 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 risk by splitting the risk 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}}\left[(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2\right] \\
&\text{(By Corollary 4.4.)} \\
&= \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\left(\sum\nolimits_{\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\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta = 1.\text{)} \\
&= \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\left(\sum\nolimits_{\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] \\
&\text{(Notice } \left(\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta a_\beta\right)^2 \leq \sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta a_\beta^2\text{, since } \mathbb{E}[a]^2 \leq \mathbb{E}[a^2].\text{)} \\
&\leq \mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\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\nolimits_{\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\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \mathbf{C}\right] \\
&\quad + P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\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\nolimits_{\beta \neq \alpha} \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \mathbf{C}\right] && \text{(Part } A\text{)} \\
&\quad + 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}] && \text{(Part } B\text{)} \\
&\quad + P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\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]. && \text{(Part } C\text{)}
\end{aligned}
$$

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

M.1.1. Bounded Efficacy - Part A

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

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\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\nolimits_{\beta \neq \alpha} \tilde{\pi}_\beta \|\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha\|^2 \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \\
&\text{(See Eq. 14 for the derivation of } \tilde{\boldsymbol{w}}_\beta.\text{)} \\
&= P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\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}.\text{)} \\
&= 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] \\
&\leq 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] \\
&\text{(Notice } \|\boldsymbol{w}_\beta - \boldsymbol{w}^*\| \leq 2.\text{)} \\
&\leq 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] \\
&\text{(Notice } \boldsymbol{A} = (\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1} \text{ and conditioned on } \mathbf{C} \text{ we have } \mathrm{L} < \lambda_d(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \lambda_1(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \mathrm{U}.\text{)}
\end{aligned}
$$

$$\le P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\left(\frac{2}{1+k\delta_w\mathrm{L}}+\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|\right)^2$$

(Notice $\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|\le 2$.)

$$\le 16P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right].$$

(By applying Eqs. 4, 5, 7, and Assumption 2(e) on $\dfrac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}$:)

$$< 16P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\Bigg[\sum\nolimits_{\beta\neq\alpha} r\exp\left(\frac{-\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\sum_{i=1}^{k+1}\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)$$
$$\cdot\exp\left(\frac{-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\right)\|\boldsymbol{x}_{k+1}\|^2\Bigg|\mathbf{C}\Bigg]$$

(In the first exponential term, by splitting $\sum_{i=1}^{k+1}$ to $\sum_{i=1}^{k}$ and $i=k+1$:)

$$< 16P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\Bigg[\sum\nolimits_{\beta\neq\alpha}\underbrace{r\exp\left(\frac{-\sum_{i=1}^{k}\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\sum_{i=1}^{k}\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)}_{\text{Part }A\text{-}1}$$
$$\underbrace{\cdot\exp\left(\frac{-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\right)}_{\text{Part }A\text{-}2}$$
$$\underbrace{\cdot\exp\left(\frac{-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_{k+1}\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2}_{\text{Part }A\text{-}3}\Bigg|\mathbf{C}\Bigg]$$

(Note that $\boldsymbol{x}_1,\dots,\boldsymbol{x}_k$ are dependent on $\mathbf{C}$ but $\boldsymbol{x}_{k+1}$ is not. Thus, we split them for further analysis.)

In the following, we separately analyze the three terms, Part A-1, Part A-2, and Part A-3. The high-level idea is that, as k increases, due to the concentration of Part A-1 and Part A-2, they can be upper bounded by a function of k. Then, regarding Part A-1 and Part A-2 as constant values (their upper bounds), the expectation of Part A-3 can be upper bounded.

Part A-1. We first deal with Part A-1. When conditioned on case $\mathbf{C}$, we have:

$$\frac{\sum_{i=1}^{k}(-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2)}{1+(k+1)\delta_\mu}$$

(Let $\boldsymbol{x}_i=\boldsymbol{\mu}^*+\boldsymbol{\epsilon}_i$)

$$=k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+\frac{\sum_{i=1}^{k}2\langle\boldsymbol{\mu}_\beta-\boldsymbol{\mu}_\alpha,\boldsymbol{\epsilon}_i\rangle}{k}}{1+(k+1)\delta_\mu}$$

$$=k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+\left\langle 2(\boldsymbol{\mu}_\beta-\boldsymbol{\mu}_\alpha),\frac{\sum_{i=1}^{k}\boldsymbol{\epsilon}_i}{k}\right\rangle}{1+(k+1)\delta_\mu}$$

$$\le k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+2\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}_\alpha\|\left\|\frac{\sum_{i=1}^{k}\boldsymbol{\epsilon}_i}{k}\right\|}{1+(k+1)\delta_\mu}$$

(Recall we have $\forall\beta\in[M],\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}_\alpha\|\le 2,$ and in case $\mathbf{C}$ we have: $\left\|\dfrac{\sum_{i=1}^{k}\boldsymbol{\epsilon}_i}{k}\right\|<\tau_x\gamma\sqrt{1+t}$.)

$$<k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu}.$$

Let $t = k^{-\frac{1}{4}}$. Recall in Assumption 4, we have $\forall \beta \neq \alpha, \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 \geq d_{\boldsymbol{\mu}}^2$. If $\delta_\mu \ll 1$ s.t. $I_{\boldsymbol{\mu}} = \{k|(k+1)\delta_\mu \leq 1 \text{ and } \frac{d_{\boldsymbol{\mu}}^2}{2} > 4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}\} \neq \varnothing$, then when $k \in I_{\boldsymbol{\mu}}$ we have:

$$k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + 4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu} < k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + \frac{d_{\boldsymbol{\mu}}^2}{2}}{2} = -k\frac{d_{\boldsymbol{\mu}}^2}{4}.$$

Part A-2. We then deal with Part A-2. When conditioned on case **C**, we have:

$$\begin{aligned}
&-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\\
&(\lambda_1(\boldsymbol{A}) \text{ and } \lambda_d(\boldsymbol{A}) \text{ indicate the largest and smallest eigenvalues of the matrix } \boldsymbol{A} \in \mathbb{R}^{d\times d}.)\\
&< -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\lambda_d(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}) + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\lambda_1(\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1})\\
&(\text{Recall in case } \mathbf{C} \text{ we have: } \mathrm{L} < \lambda_d(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \lambda_1(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \mathrm{U}.)\\
&< -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\left(1-\frac{1}{1+k\delta_w\mathrm{L}}\right) + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\left(1-\frac{1}{1+k\delta_w\mathrm{U}}\right)\\
&= -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{L}}{1+k\delta_w\mathrm{L}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{U}}{1+k\delta_w\mathrm{U}}\\
&< -\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{L}}{1+k\delta_w\tau_x^2} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{U}}{1+k\delta_w\tau_x^2}
\end{aligned}$$

Let $t = k^{-\frac{1}{4}}$. If $\delta_w \ll 1$ s.t. $I_{\boldsymbol{w}} = \{k|k\delta_w\tau_x^2 \leq 1 \text{ and } \mathrm{L}\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \mathrm{U}\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 > \frac{\tau_x^2 u_{\boldsymbol{w}}^2}{2}\} \neq \varnothing$, (note $\lim_{k\to\infty} \mathrm{L}\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - \mathrm{U}\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 = \tau_x^2\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 - (1+\tau_x^2)\|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \geq \tau_x^2 u_{\boldsymbol{w}}^2$) then when $k \in I_{\boldsymbol{w}}$, we have:

$$-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{L}}{1+k\delta_w\tau_x^2} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\frac{k\delta_w\mathrm{U}}{1+k\delta_w\tau_x^2} < -\frac{\tau_x^2 u_{\boldsymbol{w}}^2}{2}\frac{k\delta_w}{1+k\delta_w\tau_x^2} < -k\delta_w\frac{\tau_x^2 u_{\boldsymbol{w}}^2}{4}.$$

Part A-3. We finally deal with Part A-3. Part A-3 is independent to case **C**, and 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\middle|\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}.)\\
&= \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).)\\
&= \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-D+\boldsymbol{b}^\top\boldsymbol{\epsilon}}{E}\right)\|\boldsymbol{x}_{k+1}\|^2\right]\\
&(\text{Notice } \|\boldsymbol{x}_{k+1}\|^2 = \|\boldsymbol{\mu}^* + \boldsymbol{\epsilon}\|^2 \leq 2\|\boldsymbol{\mu}^*\|^2 + 2\|\boldsymbol{\epsilon}\|^2.)\\
&\leq \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-D+\boldsymbol{b}^\top\boldsymbol{\epsilon}}{E}\right)(2\|\boldsymbol{\mu}^*\|^2 + 2\|\boldsymbol{\epsilon}\|^2)\right]\\
&(\text{Notice } \|\boldsymbol{\mu}^* + \boldsymbol{\epsilon}\|^2 = 1.)\\
&= 2\left(\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-D+\boldsymbol{b}^\top\boldsymbol{\epsilon}}{E}\right)\right] + \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-D+\boldsymbol{b}^\top\boldsymbol{\epsilon}}{E}\right)\|\boldsymbol{\epsilon}\|^2\right]\right)\\
&= 2\left(\exp\left(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2} - \frac{D}{E}\right) + \mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-D+\boldsymbol{b}^\top\boldsymbol{\epsilon}}{E}\right)\|\boldsymbol{\epsilon}\|^2\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= 2\left(\exp\left(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2} - \frac{D}{E}\right) + \tau_x^2\left(1 + \frac{\tau_x^2\|\boldsymbol{b}\|^2}{E^2}\right)\exp\left(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2} - \frac{D}{E}\right) + (d-1)\tau_x^2\exp\left(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2} - \frac{D}{E}\right)\right) \\
&= 2\left(1 + \tau_x^2\left(d + \frac{\tau_x^2\|\boldsymbol{b}\|^2}{E^2}\right)\right)\exp\left(\frac{\tau_x^2\|\boldsymbol{b}\|^2}{2E^2} - \frac{D}{E}\right) \\
&= C_{k=0}.
\end{aligned}
$$

Summary of Part A. Thus, summarizing Part A-1, Part A-2, and Part A-3, we have:

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\tilde{\pi}_\beta\langle\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2\middle|\mathbf{C}\right] \\
&< 16P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\Bigg[\sum\nolimits_{\beta\neq\alpha} r\underbrace{\exp\left(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \sum_{i=1}^k\|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)}_{\text{Part } A\text{-1}} \\
&\quad\cdot\underbrace{\exp\left(\frac{-\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}} + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\right)}_{\text{Part } A\text{-2}} \\
&\quad\cdot\underbrace{\exp\left(\frac{-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_{k+1}\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2}_{\text{Part } A\text{-3}}\Bigg|\mathbf{C}\Bigg] \\
&< 16r(M-1)C_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2}\right)\exp\left(-\frac{u_{\boldsymbol{w}}^2\tau_x^2 k}{8\sigma_y^2}\right) \\
&= 16r(M-1)C_{k=0}\exp\left(-k(\frac{d_{\boldsymbol{\mu}}^2}{8\sigma_x^2} + \frac{u_{\boldsymbol{w}}^2\tau_x^2}{8\sigma_y^2})\right)
\end{aligned}
$$

□

M.1.2. BOUNDED EFFICACY - PART B

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

$$
\begin{aligned}
&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}] \\
&\leq P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|\tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2|\mathbf{C}] \\
&\text{(See Eq. 14 for the derivation of } \tilde{\boldsymbol{w}}_\alpha.\text{)} \\
&= P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\alpha - \boldsymbol{w}^*) + \boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2\|\boldsymbol{x}_{k+1}\|^2|\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}^* - \boldsymbol{w}_\alpha)\|^2\|\boldsymbol{x}_{k+1}\|^2|\mathbf{C}] \\
&\text{(Let } \lambda_1(\boldsymbol{A}) \text{ be the maximal eigenvalue of the matrix } \boldsymbol{A}.\text{)} \\
&\leq \|\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|\mathbf{C}] \\
&\text{(Recall that conditioned on } \mathbf{C} \text{ we have } \mathrm{L} < \lambda_d(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \lambda_1(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}}) < \mathrm{U}.\text{)} \\
&< \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\tilde{\pi}_\alpha\left(1 - \frac{1}{1+k\delta_w\mathrm{U}}\right)^2\|\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}}[\tilde{\pi}_\alpha\|\boldsymbol{x}_{k+1}\|^2|\mathbf{C}]\left(1 - \frac{1}{1+k\delta_w\mathrm{U}}\right)^2 \\
&< \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2\mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2\right]\left(1 - \frac{1}{1+k\delta_w\mathrm{U}}\right)^2 \\
&= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)\left(1 - \frac{1}{1+k\delta_w\mathrm{U}}\right)^2
\end{aligned}
$$

$$= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1 + d\tau_x^2)\left(\frac{k\delta_w \mathrm{U}}{1 + k\delta_w \mathrm{U}}\right)^2.$$

Let $t = k^{-\frac{1}{4}}$. if $\delta_w \ll 1$ s.t. $I_{\mathrm{U}} = \{k|\mathrm{U} < 2(1 + \tau_x^2)\} \neq \varnothing$, then when $k \in I_{\mathrm{U}}$ we have:

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

□

M.1.3. Bounded Efficacy - Part C

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

$$P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \neg\mathbf{C}\right]$$

$$\leq P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \|\tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha\|^2 \|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

(See Eq. 14 for the derivation of $\tilde{\boldsymbol{w}}_\beta$.)

$$= P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \|(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*) + \boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2 \|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

$$< P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta (2\|(\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\beta - \boldsymbol{w}^*)\|^2 + 2\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2)\|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

$$< P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \left(2\|\boldsymbol{w}_\beta - \boldsymbol{w}^*\|^2 \lambda_1^2\left((\boldsymbol{I} + k\delta_w \bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}\right) + 2\|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2\right)\|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

$$< P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta (2 \cdot 4 \cdot 1 + 2 \cdot 4)\|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

$$= 16P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \neg\mathbf{C}\right]$$

$$< 16P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 | \neg\mathbf{C}]$$

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

$$< 16P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2]$$

$$< 16(1 + d\tau_x^2)P(\neg\mathbf{C})$$

(Let $t = k^{-\frac{1}{4}}$.)

$$< 48(1 + d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right).$$

□

M.1.4. Bounded Efficacy - Summary

Proof. Summarizing Part A, Part B, and Part C, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha] \\
&< 16r(M-1)C_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2}\right)\exp\left(-\frac{u_{\boldsymbol{w}}^2 \tau_x^2 k}{8\sigma_y^2}\right) \\
&\quad + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1 + d\tau_x^2)\min\{1, 4k^2\delta_w^2(1 + \tau_x^2)^2\} + 48(1 + d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right)
\end{aligned}$$

$$\begin{aligned}
&= C_3 \exp\left(-k\left(\frac{d_{\boldsymbol{\mu}}^2}{8\sigma_x^2} + \frac{u_{\boldsymbol{w}}^2\tau_x^2}{8\sigma_y^2}\right)\right) + 48(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right) \\
&\quad + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2)\min\{1, 4k^2\delta_w^2(1+\tau_x^2)^2\}.
\end{aligned}$$

□

M.1.5. The Particular Interval

The particular interval for the non-asymptotic bound is the union of $I_{\boldsymbol{\mu}}$, $I_{\boldsymbol{w}}$, and I_{U}:

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

M.2. Asymptotic Bound

This section proves the non-asymptotic bound in Theorem 6.1: Consider a next-token predictor attaining the optimal pretraining risk. As $k \to \infty$, ICL risk with biased labels is upper bounded by:

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + \frac{C_1}{k}\exp\left(C_2 k^{-\frac{1}{2}}\right) + O(k^{-2}),$$

where $\mathcal{L}_k^\alpha = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^\alpha)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha\rangle)^2$, and C_1 and C_2 are constants depending on the prior setting, τ_x, and $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$.

The proof of the asymptotic bound is heavily overlapped with the proof of the non-asymptotic bound. **We will hide the overlapped derivations with "$(\ldots)$".**

Proof. Assuming we are using in-context examples following Assumptions 3 and 4, *i.e.*, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I}), y_i = \langle \boldsymbol{x}_i, \boldsymbol{w}^*\rangle$, $\|\boldsymbol{\mu}^*\| = \|\boldsymbol{w}^*\| = 1$, and we aim to retrieve the function $\boldsymbol{w}_\alpha$ of the prior center $(\boldsymbol{\mu}_\alpha, \boldsymbol{w}_\alpha)$ which is close to the in-context task. Let $\mathcal{L}_k^\alpha$ indicate the squared risk $(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha\rangle)^2$, 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 risk, we consider $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2\boldsymbol{I})$ in two cases: (1) **C**: $\mathrm{L} < \lambda_d\left(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right) \le \lambda_1\left(\frac{\sum_{i=1}^k \boldsymbol{x}_i\boldsymbol{x}_i^\top}{k}\right) < \mathrm{U}$ and $\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\| < \tau_x\sqrt{\gamma(1+t)}$ (see Lemma K.1 for t, γ, L and U) and (2) $\neg$**C**: at least one of the previous inequalities does not hold. Following Lemma K.1, the probability of $\neg$**C** is bounded by: $P(\neg\mathbf{C}) \le 3\exp(-\frac{kt^2}{8})$).

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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha] \\
&(\ldots) \\
&= P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta \ne \alpha} \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \mathbf{C}\right] && \text{(Part } A'\text{)} \\
&\quad + P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \langle \tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 | \mathbf{C}] && \text{(Part } B'\text{)} \\
&\quad + P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^M \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \neg\mathbf{C}\right]. && \text{(Part } C'\text{)}
\end{aligned}$$

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

M.2.1. Asymptotic Bound - Part A'

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

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\tilde{\pi}_\beta\langle\tilde{\boldsymbol{w}}_\beta-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle^2\middle|\mathbf{C}\right]\\
&(\dots)\\
&<P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\left(\frac{2}{1+k\delta_w\mathrm{L}}+\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|\right)^2\\
&(\text{Notice } \|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|\leq 2.)\\
&\leq P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\left(\frac{4}{(1+k\delta_w\mathrm{L})^2}+\frac{8}{1+k\delta_w\mathrm{L}}\right) &&(16)\\
&\quad+P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\tilde{\pi}_\beta\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|^2. &&(17)
\end{aligned}
$$

Line 17 will be merged with Part B' and analyzed in Sec. M.2.2. The current section will analyze the line 16. We start by analyzing the term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]$. By Eqs. 4, 5, 7, and Assumption 2(e) on $\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}$, we have:

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha}\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\\
&(\dots)\\
&<P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\Bigg[\sum\nolimits_{\beta\neq\alpha}r\underbrace{\exp\left(\frac{-\sum_{i=1}^k\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\sum_{i=1}^k\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)}_{\text{Part } A'\text{-1}}\\
&\quad\underbrace{\cdot\exp\left(\frac{-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}}{2\sigma_w^2}\right)}_{\text{Part } A'\text{-2}}\\
&\quad\underbrace{\cdot\exp\left(\frac{-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_{k+1}\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2}_{\text{Part } A'\text{-3}}\Bigg|\mathbf{C}\Bigg]
\end{aligned}
$$

(Note that $x_1,\dots,x_k$ are dependent on $\mathbf{C}$ but x_{k+1} is not. Thus, we break them for further analysis.)

In the following, we separately analyze the three terms, Part A'-1, Part A'-2, and Part A'-3. The high-level idea is that, as k increases, due to the concentration of Part A'-1 and Part A'-2, they can be upper bounded by a function of k. Then, regarding Part A'-1 and Part A'-2 as constant values (their upper bounds), the expectation of Part A'-3 can be upper bounded.

Part A'-1. We first deal with Part A-1. When conditioned on case $\mathbf{C}$, we have:

$$
\begin{aligned}
&\frac{\sum_{i=1}^k(-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_i\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_i\|^2)}{1+(k+1)\delta_\mu}\\
&(\dots)\\
&<k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu}.
\end{aligned}
$$

With Assumption 4, we have $d_{\boldsymbol{\mu}}^2\leq\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2$. With Lemma K.1, we have $\gamma=\sqrt{\frac{d}{k}}$. Let $t=k^{\delta-\frac{1}{2}}$ and $0<\delta<\frac{1}{2}$, we have:

$$
k\frac{\|\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}^*\|^2-\|\boldsymbol{\mu}_\beta-\boldsymbol{\mu}^*\|^2+4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu}=-\frac{d_{\boldsymbol{\mu}}^2}{\delta_\mu}+\frac{4\tau_x\sqrt{d}}{\delta_\mu}k^{-\frac{1}{2}}+O(k^{-1}).
$$

Part A'-2. We then deal with Part A'-2. When conditioned on case **C**, we have:

$$\begin{aligned}
&-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2_{\boldsymbol{I}-(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}}\\
&(\dots)\\
&<-\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2\left(1-\frac{1}{1+k\delta_w\mathrm{L}}\right)+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2\left(1-\frac{1}{1+k\delta_w\mathrm{U}}\right)\\
&=-(\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2)+\left(\frac{\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2}{1+k\delta_w\mathrm{L}}-\frac{\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2}{1+k\delta_w\mathrm{U}}\right).
\end{aligned}$$

With Assumption 4, we have $d^2_{\boldsymbol{w}}\le\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2$. Lemma K.1 gives the definitions of L and U. Let $t=k^{\delta-\frac{1}{2}}$ and $0<\delta<\frac{1}{2}$, we have:

$$\begin{aligned}
&=-d^2_{\boldsymbol{w}}+\left(\frac{\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2}{k\delta_w\tau_x^2}-\frac{\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2}{k\delta_w(1+\tau_x^2)}\right)+O(k^{-2})\\
&<-d^2_{\boldsymbol{w}}+\frac{\|\boldsymbol{w}_\beta-\boldsymbol{w}^*\|^2}{k\delta_w\tau_x^2}+O(k^{-2})\\
&<-d^2_{\boldsymbol{w}}+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{-2}).
\end{aligned}$$

Part A'-3. We finally deal with Part A'-3. Part A'-3 is independent to case **C**, and we have:

$$\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\exp\left(\frac{-\|\boldsymbol{\mu}_\beta-\boldsymbol{x}_{k+1}\|^2+\|\boldsymbol{\mu}_\alpha-\boldsymbol{x}_{k+1}\|^2}{2\sigma_x^2(1+(k+1)\delta_\mu)}\right)\|\boldsymbol{x}_{k+1}\|^2\middle|\mathbf{C}\right]\\
&(\dots)\\
&=C_{k=0}.
\end{aligned}$$

Summary of Part A'. Thus, summarizing Part A'-1, Part A'-2, and Part A'-3, 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]\left(\frac{4}{(1+k\delta_w\mathrm{L})^2}+\frac{8}{1+k\delta_w\mathrm{L}}\right)\\
&<P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta\neq\alpha}r\underbrace{\exp\left(\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)}\right)}_{\text{Part }A'\text{-1}}\right.\\
&\quad\underbrace{\cdot\exp\left(\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}\right)}_{\text{Part }A'\text{-2}}\\
&\quad\left.\underbrace{\cdot\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}_{\text{Part }A'\text{-3}}\middle|\mathbf{C}\right]\\
&\quad\cdot\left(\frac{4}{(1+k\delta_w\mathrm{L})^2}+\frac{8}{1+k\delta_w\mathrm{L}}\right)\\
&\text{(Notice }\lim_{k\to\infty}\mathrm{L}=\lim_{k\to\infty}\tau_x^2\left(1-\frac{t}{2}-\gamma\right)^2-2\tau_x\gamma\sqrt{1+t}=\tau_x^2.\text{)}\\
&<r\sum_{\beta\neq\alpha}\exp\left(\frac{-\frac{d^2_{\boldsymbol{\mu}}}{\delta_\mu}+\frac{4\tau_x\sqrt{d}}{\delta_\mu}k^{-\frac{1}{2}}+O(k^{-1})}{2\sigma_x^2}\right)\exp\left(\frac{-d^2_{\boldsymbol{w}}+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{-2})}{2\sigma_w^2}\right)C_{k=0}\left(\frac{8}{k\delta_w\tau_x^2}+O(k^{-2})\right)\\
&=r(M-1)C_{k=0}\exp\left(\frac{-d^2_{\boldsymbol{\mu}}+4\tau_x\sqrt{d}k^{-\frac{1}{2}}+O(k^{-1})}{2\sigma_\mu^2}\right)\exp\left(\frac{-d^2_{\boldsymbol{w}}+\frac{4}{\delta_w\tau_x^2}k^{-1}+O(k^{-2})}{2\sigma_w^2}\right)\left(\frac{8}{k\delta_w\tau_x^2}+O(k^{-2})\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{8r(M-1)C_{k=0}}{k\delta_w\tau_x^2} \exp\left(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{-\frac{1}{2}} + O(k^{-1})}{2\sigma_\mu^2}\right) \exp\left(\frac{-d_{\boldsymbol{w}}^2 + \frac{4}{\delta_w\tau_x^2}k^{-1} + O(k^{-2})}{2\sigma_w^2}\right) + O(k^{-2}) \\
&= \frac{8r(M-1)C_{k=0}}{k\delta_w\tau_x^2} \exp\left(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{-\frac{1}{2}}}{2\sigma_\mu^2}\right) \exp\left(\frac{-d_{\boldsymbol{w}}^2}{2\sigma_w^2}\right) + O(k^{-2})
\end{aligned}
$$

□

M.2.2. Asymptotic Bound - Part B'

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

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \langle \tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 | \mathbf{C}] \\
&(\dots) \\
&< \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \|\boldsymbol{x}_{k+1}\|^2 | \mathbf{C}] \left(1 - \frac{1}{1 + k\delta_w \mathrm{U}}\right)^2 .
\end{aligned}
$$

We add the line 17 in Sec. M.2.1 back:

$$
\begin{aligned}
&P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha (\langle \tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle)^2 | \mathbf{C}] + \underbrace{P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta \neq \alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2}_{\text{line 17 in Sec. M.2.1}} \\
&< \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \|\boldsymbol{x}_{k+1}\|^2 | \mathbf{C}] \left(1 - \frac{1}{1 + k\delta_w \mathrm{U}}\right)^2 \\
&\quad + P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta \neq \alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|^2 \\
&\leq \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \|\boldsymbol{x}_{k+1}\|^2 | \mathbf{C}] + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta \neq \alpha} \tilde{\pi}_\beta \|\boldsymbol{x}_{k+1}\|^2 \middle| \mathbf{C}\right] \\
&(\text{Notice } \sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta = 1) \\
&= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2 | \mathbf{C}] \\
&< \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 \mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1}\|^2\right] \\
&= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2 (1 + d\tau_x^2)
\end{aligned}
$$

□

M.2.3. Asymptotic Bound - Part C'

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

$$
\begin{aligned}
&P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\beta=1}^{M} \tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 \middle| \neg\mathbf{C}\right] \\
&(\dots) \\
&< 16(1 + d\tau_x^2) P(\neg\mathbf{C}) \\
&(\text{Let } t = k^{\delta - \frac{1}{2}}.) \\
&< 48(1 + d\tau_x^2) \exp\left(-\frac{k^{2\delta}}{8}\right).
\end{aligned}
$$

□

M.2.4. Asymptotic Bound - Summary

Proof. Summarizing Part A', Part B', and Part C', we have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha] \\
&< \frac{8r(M-1)C_{k=0}}{k\delta_w \tau_x^2} \exp\left(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{-\frac{1}{2}}}{2\sigma_\mu^2}\right) \exp\left(\frac{-d_{\boldsymbol{w}}^2}{2\sigma_w^2}\right) + O(k^{-2}) \\
&\quad + \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + 48(1+d\tau_x^2)\exp\left(-\frac{k^{2\delta}}{8}\right) \\
&= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + \frac{8r(M-1)C_{k=0}}{k\delta_w \tau_x^2} \exp\left(\frac{-d_{\boldsymbol{\mu}}^2 + 4\tau_x\sqrt{d}k^{-\frac{1}{2}}}{2\sigma_\mu^2}\right) \exp\left(\frac{-d_{\boldsymbol{w}}^2}{2\sigma_w^2}\right) + O(k^{-2}) \\
&= \|\boldsymbol{w}_\alpha - \boldsymbol{w}^*\|^2(1+d\tau_x^2) + \frac{C_1}{k}\exp(C_2 k^{-\frac{1}{2}}) + O(k^{-2})
\end{aligned}$$

□

N. Proof of Lemma 6.2

In this subsection, we introduce the proof of Lemma 6.2. We first give the full version of the lemma:

Lemma 6.2 (Upper Bound for Zero-Shot ICL). Assume a next-token predictor attains the optimal pretraining risk, and Assumption 6 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 $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^* | \tau_x^2 \boldsymbol{I})$, assume $\|\boldsymbol{\mu}^*\| = 1$, and $y_i = 0$, *i.e.*, y_i has the same preference to prior component α as β. When δ_μ and δ_w are sufficiently small, there is a particular interval for k that ICL risk is upper bounded by:

$$\mathbb{E}_{\mathcal{S}_k}[\mathcal{L}_k^\alpha] < C_4 \exp\left(-\frac{d_{\boldsymbol{\mu}}^2 k}{8\sigma_x^2}\right) + 12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right) + (1+d\tau_x^2)\min\{1, k^2 {\delta_w}^2(1+\tau_x^2)^2\},$$

where $\mathcal{L}_k^\alpha = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - y_{k+1}^\alpha)^2 = (\mathcal{F}(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle)^2$, C_4 is a constant depending on the prior, τ_x, and $(\boldsymbol{\mu}^*, \boldsymbol{w}^*)$. When k is small, the first and second terms dominate and exponential decay. When k is large, the third term dominates and increases.

Proof. The proof techniques are very similar to the proof techniques used in Sec. M.1. Assuming we are using in-context examples following $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}^*, \tau_x^2 \boldsymbol{I}), \|\boldsymbol{\mu}^*\| = 1, y_i = 0$, *i.e.*, $\boldsymbol{w}^* = \boldsymbol{0}$, and we aim to retrieve the function $\boldsymbol{w}_\alpha$ of the prior center $(\boldsymbol{\mu}_\alpha, \boldsymbol{w}_\alpha)$ which is close to the in-context task. Let $\mathcal{L}_k^\alpha$ indicate the squared loss $(\mathcal{F}^*(\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}) - \langle \boldsymbol{x}_{k+1}, \boldsymbol{w}_\alpha \rangle)^2$, 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) **C**: $\mathrm{L} < \lambda_d \left(\frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}\right) \leq \lambda_1 \left(\frac{\sum_{i=1}^k \boldsymbol{x}_i \boldsymbol{x}_i^\top}{k}\right) < \mathrm{U}$ and $\left\|\frac{\sum_{i=1}^k \boldsymbol{\epsilon}_i}{k}\right\| < \tau_x \sqrt{\gamma(1+t)}$ (see Lemma K.1 for t, γ, L and U) and (2) $\neg$**C**: at least one of the previous inequalities does not hold. Following Lemma K.1, the probability of $\neg$**C** is bounded by: $P(\neg\mathbf{C}) \leq 3\exp(-\frac{kt^2}{8})$).

Similar to Sec. M.1, we split the expected squared loss into three parts:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha] \\
&< P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle^2 | \mathbf{C}] && \text{(Part } A''\text{)} \\
&\quad + P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha \langle \tilde{\boldsymbol{w}}_\alpha - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle^2 | \mathbf{C}] && \text{(Part } B''\text{)} \\
&\quad + P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\kappa \in \{\alpha,\beta\}} \tilde{\pi}_\kappa \langle \tilde{\boldsymbol{w}}_\kappa - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1} \rangle^2 | \neg\mathbf{C}\right]. && \text{(Part } C''\text{)}
\end{aligned}$$

□

N.1. Proof of Lemma 6.2: Part A''

Proof. We first analyze the term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 | \mathbf{C}]$, Part A''. Similar to Sec. M.1, 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 | \mathbf{C}] \\
&< P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\frac{\tilde{\pi}_\beta}{\tilde{\pi}_\alpha} \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 | \mathbf{C}] \cdot \left(\frac{2}{1+k\delta_w \mathrm{L}} + \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|\right)^2 \\
&< P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\Bigg[r \exp\left(\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)}\right) \\
&\quad \cdot \exp\left(\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}\right) \\
&\quad \cdot \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 \Bigg| \mathbf{C}\Bigg] \cdot \left(\frac{2}{1+k\delta_w \mathrm{L}} + \|\boldsymbol{w}^* - \boldsymbol{w}_\alpha\|\right)^2 \\
&\text{(Notice } \boldsymbol{w}^* = \mathbf{0}, \boldsymbol{w}_\beta = -\boldsymbol{w}_\alpha.\text{)} \\
&= rP(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\Bigg[\exp\left(\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)}\right) \\
&\quad \cdot \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 \Bigg| \mathbf{C}\Bigg] \cdot 3^2 \\
&= 9rP(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}\Bigg[\underbrace{\exp\left(\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)}\right)}_{A''\text{-}1} \\
&\quad \underbrace{\cdot \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}_{A''\text{-}3} \Bigg| \mathbf{C}\Bigg].
\end{aligned}
$$

Same to Sec. M.1.1, when conditioned on case $\mathbf{C}$, for Part A''-1 we have:

$$
\frac{\sum_{i=1}^k (-\|\boldsymbol{\mu}_\beta - \boldsymbol{x}_i\|^2 + \|\boldsymbol{\mu}_\alpha - \boldsymbol{x}_i\|^2)}{1+(k+1)\delta_\mu} < k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + 4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu}.
$$

Let $t = k^{-\frac{1}{4}}$. Recall in Assumption 4, we have $\forall \beta \neq \alpha, \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 \geq d^2_{\boldsymbol{\mu}}$. If $\delta_\mu \ll 1$ s.t. $I_{\boldsymbol{\mu}} = \{k | (k+1)\delta_\mu \leq 1 \text{ and } \frac{d^2_{\boldsymbol{\mu}}}{2} > 4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}\} \neq \varnothing$, then when $k \in I_{\boldsymbol{\mu}}$ we have:

$$
k\frac{\|\boldsymbol{\mu}_\alpha - \boldsymbol{\mu}^*\|^2 - \|\boldsymbol{\mu}_\beta - \boldsymbol{\mu}^*\|^2 + 4\tau_x\gamma\sqrt{1+t}}{1+(k+1)\delta_\mu} < -\frac{d^2_{\boldsymbol{\mu}}}{4}.
$$

Same to Sec. M.1.1, when conditioned on case $\mathbf{C}$, for Part A''-3 we have:

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

As a summary of the above analysis, we have:

$$
P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k \oplus \boldsymbol{x}_{k+1}}[\tilde{\pi}_\beta \langle \tilde{\boldsymbol{w}}_\beta - \boldsymbol{w}_\alpha, \boldsymbol{x}_{k+1}\rangle^2 | \mathbf{C}] < 9rC_{k=0}\exp\left(-\frac{d^2_{\boldsymbol{\mu}}k}{8\sigma_x^2}\right).
$$

□

N.2. Proof of Lemma 6.2: Part B''

Proof. We then deal with the second term $P(\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\tilde{\pi}_\alpha(\langle\tilde{\boldsymbol{w}}_\alpha-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2|\mathbf{C}]$, Part B''. The analysis is exactly the same as Sec. M.1.2, and we have:

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

Let $t=k^{-\frac{1}{4}}$. if $\delta_w\ll 1$ s.t. $I_{\mathrm{U}}=\{k|\mathrm{U}<2(1+\tau_x^2)\}\neq\varnothing$, then when $k\in I_{\mathrm{U}}$ we have:

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

□

N.3. Proof of Lemma 6.2: Part C''

Proof. Finally, for the third term $P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\sum_{\kappa\in\{\alpha,\beta\}}\tilde{\pi}_\kappa\langle\tilde{\boldsymbol{w}}_\kappa-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle^2|\neg\mathbf{C}]$, Part C''. Similar to Sec. M.1.3, we have:

$$
\begin{aligned}
&P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\kappa\in\{\alpha,\beta\}}\tilde{\pi}_\kappa(\langle\tilde{\boldsymbol{w}}_\kappa-\boldsymbol{w}_\alpha,\boldsymbol{x}_{k+1}\rangle)^2\middle|\neg\mathbf{C}\right]\\
&<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\kappa\in\{\alpha,\beta\}}\tilde{\pi}_\kappa\left(2\|(\boldsymbol{I}+k\delta_w\bar{\boldsymbol{\Sigma}}_{\boldsymbol{w}})^{-1}(\boldsymbol{w}_\kappa-\boldsymbol{w}^*)\|^2+2\|\boldsymbol{w}^*-\boldsymbol{w}_\alpha\|^2\right)\|\boldsymbol{x}_{k+1}\|^2\middle|\neg\mathbf{C}\right]\\
&\text{(Recall } \boldsymbol{w}^*=\mathbf{0}.)\\
&<P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\kappa\in\{\alpha,\beta\}}\tilde{\pi}_\kappa(2\cdot1\cdot1+2\cdot1)\|\boldsymbol{x}_{k+1}\|^2\middle|\neg\mathbf{C}\right]\\
&=4P(\neg\mathbf{C})\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}\left[\sum\nolimits_{\kappa\in\{\alpha,\beta\}}\tilde{\pi}_\kappa\|\boldsymbol{x}_{k+1}\|^2\middle|\neg\mathbf{C}\right]\\
&<4P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2|\neg\mathbf{C}]\\
&\text{(Notice } \mathbf{C} \text{ is defined on } \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k\}.)\\
&<4P(\neg\mathbf{C})\mathbb{E}_{\boldsymbol{x}_{k+1}}[\|\boldsymbol{x}_{k+1}\|^2]\\
&<4(1+d\tau_x^2)P(\neg\mathbf{C})\\
&\text{(Let } t=k^{-\frac{1}{4}}.)\\
&<12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right).
\end{aligned}
$$

□

N.4. Proof of Lemma 6.2: Summary

Proof. Summarizing Part A'', Part B'', and Part C'', we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k\oplus\boldsymbol{x}_{k+1}}[\mathcal{L}_k^\alpha]\\
&<9rC_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2k}{8\sigma_x^2}\right)+\|\boldsymbol{w}_\alpha-\boldsymbol{w}^*\|^2(1+d\tau_x^2)\min\{1,4k^2\delta_w^2(1+\tau_x^2)^2\}+12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right)\\
&=9rC_{k=0}\exp\left(-\frac{d_{\boldsymbol{\mu}}^2k}{8\sigma_x^2}\right)+(1+d\tau_x^2)\min\{1,4k^2\delta_w^2(1+\tau_x^2)^2\}+12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right)\\
&=C_4\exp\left(-\frac{d_{\boldsymbol{\mu}}^2k}{8\sigma_x^2}\right)+12(1+d\tau_x^2)\exp\left(-\frac{k^{\frac{1}{2}}}{8}\right)+(1+d\tau_x^2)\min\{1,4k^2\delta_w^2(1+\tau_x^2)^2\}.
\end{aligned}
$$

□

N.5. The Particular Interval

The particular interval for the risk bound revealing bounded efficacy is the union of $I_{\boldsymbol{\mu}}$ and I_{U}:

$$\begin{aligned} k &\leq \frac{1}{\delta_{\mu}} - 1 \\ 4\tau_x\gamma\sqrt{1+k^{-\frac{1}{4}}}) &< \frac{d_{\boldsymbol{\mu}}^2}{2} \\ \mathrm{U} &< 2(1+\tau_x^2). \end{aligned}$$

O. Toy Example for Component Shifting and Component Re-weighting

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 token y of the token x. Instead of direct 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 verbose 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. O.1. We then connect ICL with Bayesian inference in Sec. O.2. Further, we introduce the assumptions for the pretraining dataset in Sec. O.3. Finally, we derive a closed-form posterior and introduce two phenomena, “Component Shifting” and “Component Re-weighting” in Sec. O.4.

O.1. Toy Example: Pretraing Data Generative Modela

ICL involves two important components: the pretraining dataset, and the next-token predictor supporting varied input lengths. We assume the next-token predictor $f : \cup_{k\in\{0,\ldots,K-1\}}\mathcal{R}^{k\times 1} \to \mathcal{R}^{1\times 1}$ 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 tokens of the sequence from a distribution $\mathcal{D}_x(\mu)$ based on the task $\boldsymbol{\mu}$. The sample generation process is described as follows:

Assumption 7 (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 task prior: $\mu \sim \mathcal{D}_\mu$, and the probability of μ is indicated by $P(\mu)$;
(b) Sample K samples, each denoted by x_i, 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 7(a), the LM is trained on varied tasks; (ii) For sampling step 7(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.

O.2. Toy Example: Bayes-Optimal Next-Token Predictor

Now we consider training $f(\cdot)$ using sample S_K generated via the above generation process 7:

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

f can be viewed as K separate models $f_0, \ldots, f_{K-1}$, where f_k takes a sequence of k tokens as input. Therefore, when the model f has enough expressivity, the optimization problem $f^* = \mathrm{argmin}_f\, \mathcal{L}(f)$ could be regarded as K different optimization problems:

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

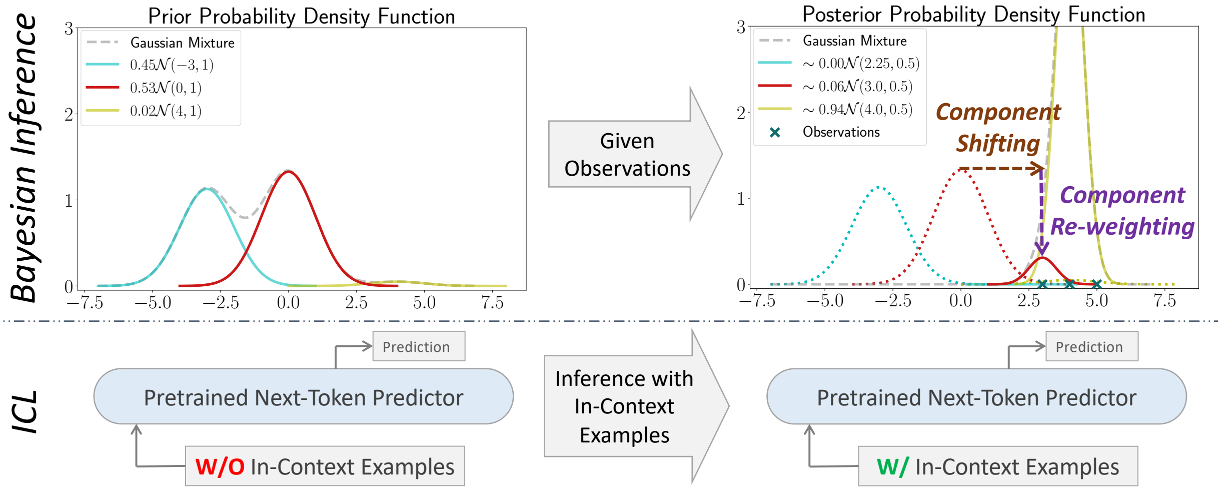


Figure 23: The left part of the figure indicates the pretrained next-token predictor is pretrained on the task prior distribution according to Assumption 8, and the prediction is based on the prior without in-context examples. The right part of the figure indicates that with in-context samples, the prediction 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:

$$f^*(S_k) = \mathop{\mathbb{E}}_{S_K}[x_{k+1}|S_k] = \mathop{\mathbb{E}}_{\mu\sim\mathcal{D}_\mu}\Big[\mathop{\mathbb{E}}_{\substack{x_i\sim\mathcal{D}(\mu),\\ i\in\{1,\dots,K\}}}[x_{k+1}|\mu,S_k]|S_k\Big] = \mathop{\mathbb{E}}_{\mu\sim\mathcal{D}_\mu}\Big[\mathop{\mathbb{E}}_{x_{k+1}\sim\mathcal{D}(\mu)}[x_{k+1}|\mu]|S_k\Big]. \tag{18}$$

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.

O.3. Toy Example: Gaussian Assumptions on Pretraining Data Generative Model

In Sec. O.2, we connect ICL with Bayesian inference, and in Eq. 18, 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 8 (Demo: Gaussian Assumptions for Generative Model for Pretraining Data).
(a) Task distribution: $\mu \sim \mathcal{D}_\mu, P(\mu) = \sum_{m=1}^M \pi_m P(\mu|T_m)$, where T_m is the m^{th} mixture component of the Gaussian mixture, *i.e.*, $P(\mu|T_m) = \mathcal{N}(\mu|\mu_m, \sigma^2)$, and π_m is the corresponding mixture weight. $\sum_{m=1}^M \pi_m = 1, 0 < \pi_m < 1, \mu_m$ is the center of the mixture component T_m, and all components share the same covariance matrix controlled by σ;
(b) Token distribution: $x \sim \mathcal{D}_x(\mu)$, $P(x|\mu) = \mathcal{N}(x|\mu_m, \tau^2)$.

O.4. Toy Example: Posterior Analysis

With Assumption 8, we derive the closed-form expression of the posterior as follows:

$$P(\mu|S_k) \propto \sum_{m=1}^M \tilde{\pi}_m \mathcal{N}(\mu|\tilde{\mu}_m, \tilde{\sigma}^2). \tag{19}$$

$$(\tilde{\pi}_m = \pi_m \exp\left(\frac{k\left(\mu_m - \frac{\sum_{i=1}^k x_i}{k}\right)^2}{2(\tau^2 + k\sigma^2)}\right), \tilde{\mu}_m = \frac{\tau^2\mu_m + \sigma^2\sum_{i=1}^k x_i}{\tau^2 + k\sigma^2}, \tilde{\sigma}^2 = \frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2})$$

See Sec. O.5 for proof details. From Eq. 19, we observe two factors when comparing the posterior with the prior in Assumption 8: (i) Component Shifting: after observing $S_k = [x_1, x_2, \dots, x_k]$, the center of each mixture component is shifted to $\frac{\tau^2\mu_m+\sigma^2\sum_{i=1}^k x_i}{\tau^2+k\sigma^2}$; (ii) Component Re-weighting: the mixture weight π_m of each mixture component is re-weighted by multiplying $\exp\left(\frac{k\left(\mu_m-\frac{\sum_{i=1}^k x_i}{k}\right)^2}{2(\tau^2+k\sigma^2)}\right)$ (which needs to be further normalized so that re-weighted mixture weights sum to 1). Fig. 23 illustrates the phenomena of Component Shifting and Component Re-weighting by observing in-context examples.

O.5. Proof of Posterior Derivation in Toy Example

In this section, we give a detailed derivation of the posterior in Eq. 19 of Sec. O.4:

$$\begin{aligned}
P(\mu|S_k) &\propto P(\mu,S_k)\\
&= P(S_k|\mu)P(\mu)\\
&= (\Pi_{i=1}^k P(x_i|\mu))P(\mu)\\
&= \sum_{m=1}^M \pi_m\mathcal{N}(\mu|\mu_m,\sigma^2)(\Pi_{i=1}^k\mathcal{N}(x_i|\mu,\tau^2)).
\end{aligned}$$

We then show $\mathcal{N}(\mu|\mu_m,\sigma^2)(\Pi_{i=1}^k\mathcal{N}(x_i|\mu,\tau^2))$ is proportional to a Gaussian distribution:

$$\log\left(\mathcal{N}(\mu|\mu_m,\sigma^2)\cdot\Pi_{i=1}^k\mathcal{N}(x_i|\mu,\tau^2)\right)$$

$$=\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{(\mu-\mu_m)^2}{2\sigma^2}\right)+\sum_{i=1}^k\left(\log\left(\frac{1}{\sqrt{2\pi}\tau}\right)-\frac{(x_i-\mu)^2}{2\tau^2}\right)$$

(Let $C_{10}=\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)+k\log\left(\frac{1}{\sqrt{2\pi}\tau}\right)$)

$$=C_{10}-\frac{(\mu-\mu_m)^2}{2\sigma^2}-\sum_{i=1}^k\frac{(x_i-\mu)^2}{2\tau^2}$$

$$=C_{10}-\frac{1}{2\tau^2\sigma^2}\left(\tau^2(\mu-\mu_m)^2+\sigma^2\sum_{i=1}^k(x_i-\mu)^2\right)$$

(Abbreviate $\sum_{i=1}^k$ as $\sum$ for simplicity.)

$$=C_{10}-\frac{1}{2\tau^2\sigma^2}\left(\mu^2(\tau^2+k\sigma^2)-2\mu\left(\tau^2\mu_m+\sigma^2\sum x_i\right)+\left(\tau^2\mu_m^2+\sigma^2\sum x_i^2\right)\right)$$

$$=C_{10}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\left(\left(\mu-\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2+\frac{\tau^2\mu_m^2+\sigma^2\sum x_i^2}{\tau^2+k\sigma^2}-\left(\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2\right)$$

$$=C_{10}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\left(\left(\mu-\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2+\frac{(\tau^2\mu_m^2+\sigma^2\sum x_i^2)(\tau^2+k\sigma^2)-(\tau^2\mu_m+\sigma^2\sum x_i)^2}{(\tau^2+k\sigma^2)^2}\right)$$

$$=C_{10}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\left(\left(\mu-\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2+\frac{k\sigma^2\tau^2\mu_m^2+\sigma^2\sum x_i^2(\tau^2+k\sigma^2)-2\mu_m\tau^2\sigma^2\sum x_i-(\sigma^2\sum x_i)^2}{(\tau^2+k\sigma^2)^2}\right)$$

(Let $C_{11}=C_{10}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\cdot\frac{\sigma^2\sum x_i^2(\tau^2+k\sigma^2)-(\sigma^2\sum x_i)^2-\tau^2\sigma^2(\sum x_i)^2/k}{(\tau^2+k\sigma^2)^2}$.)

$$=C_{11}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\left(\left(\mu-\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2+\frac{k\sigma^2\tau^2\mu_m^2-2\mu_m\tau^2\sigma^2\sum x_i+\tau^2\sigma^2(\sum x_i)^2/k}{(\tau^2+k\sigma^2)^2}\right)$$

$$=C_{11}-\frac{\tau^2+k\sigma^2}{2\tau^2\sigma^2}\left(\left(\mu-\frac{\tau^2\mu_m+\sigma^2\sum x_i}{\tau^2+k\sigma^2}\right)^2+\frac{k\tau^2\sigma^2}{(\tau^2+k\sigma^2)^2}\cdot\left(\mu_m-\frac{\sum x_i}{k}\right)^2\right)$$

$$= C_{11} - \frac{k\left(\mu_m - \frac{\sum_{i=1}^k x_i}{k}\right)^2}{2(\tau^2 + k\sigma^2)} - \frac{\left(\mu - \frac{\tau^2\mu_m + \sigma^2\sum_{i=1}^k x_i}{\tau^2 + k\sigma^2}\right)^2}{2 \cdot \frac{\tau^2\sigma^2}{\tau^2 + k\sigma^2}}.$$

Notice C_{11} is independent to $m, \forall m \in [M]$ and μ. Therefore, we have:

$$\pi_m \cdot \mathcal{N}(\mu|\mu_m, \sigma^2) \cdot \Pi_{i=1}^k \mathcal{N}(x_i|\mu, \tau^2) \propto \tilde{\pi}_m \cdot \mathcal{N}(\mu|\tilde{\mu}_m, \tilde{\sigma}^2),$$

where $\tilde{\pi}_m = \pi_m \exp\left(-\frac{k\left(\mu_m - \frac{\sum_{i=1}^k x_i}{k}\right)^2}{2(\tau^2+k\sigma^2)}\right)$, $\tilde{\mu}_m = \frac{\tau^2\mu_m + \sigma^2\sum_{i=1}^k x_i}{\tau^2 + k\sigma^2}$, and $\tilde{\sigma}^2 = \frac{\tau^2\sigma^2}{\tau^2+k\sigma^2}$. Thus:

$$\begin{aligned} P(\mu|S_k) &\propto \sum_{m=1}^M \pi_m \mathcal{N}(\mu|\mu_m, \sigma^2)(\Pi_{i=1}^k \mathcal{N}(x_i|\mu, \tau^2)) \\ &\propto \tilde{\pi}_m \mathcal{N}(\mu|\tilde{\mu}_m, \tilde{\sigma}^2). \end{aligned}$$