Understanding Overadaptation in Supervised Fine-Tuning: The Role of Ensemble Methods

Yifan Hao [*1] **Xingyuan Pan** [*1] **Hanning Zhang** [*1] **Chenlu Ye** [1] **Rui Pan** [1] **Tong Zhang** [1]

Abstract

Supervised fine-tuning (SFT) on domain-specific data is the dominant approach for adapting foundation models to specialized tasks. However, it has been observed that SFT models tend to forget knowledge acquired during pretraining. In vision models, ensembling a pretrained model with its fine-tuned counterpart has been shown to mitigate this issue (Wortsman et al., 2022b). In this work, we demonstrate that the same holds for language models, and, more strikingly, we observe an *overadaptation* phenomenon: the ensemble model not only retains general knowledge from the foundation model but also outperforms the fine-tuned model even on the fine-tuning domain itself. Despite the empirical success of ensembling, a theoretical understanding of its benefits remains underexplored. We develop a formal theoretical analysis of the overadaptation phenomenon. Ensembling mitigates this by balancing two primary sources of error: bias, caused by insufficient fine-tuning, and variance, introduced by overfitting to fine-tuning data. While regularization techniques aim to address this trade-off, we show that ensembling provides a more effective solution. We analyze this phenomenon in over-parameterized linear settings and demonstrate that interpolating between pretrained and fine-tuned weights significantly improves performance. These findings offer theoretical justification for the observed advantages of model ensembling, supported by empirical experiments consistent with our analysis.

[*]Equal contribution [1]University of Illinois Urbana-Champaign, Illinois. Correspondence to: Tong Zhang <tongzhang@tongzhang-ml.org>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

With the remarkable success of large language models (LLMs) such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), and Claude (Anthropic, 2023), the pretrain-finetune paradigm has gained significant attention for its outstanding performance. Supervised fine-tuning (SFT) is a widely adopted approach for adapting foundation models to specific downstream tasks. However, a well-known challenge with SFT models is the tendency to forget information acquired during pre-training (McCloskey & Cohen, 1989; Goodfellow et al., 2013). Model ensembling, also known as model averaging, has emerged as one of the most effective strategies to address this issue. Its benefits have been empirically demonstrated in vision models (Wortsman et al., 2022b) and in reinforcement learning from human feedback (RLHF) (Lin et al., 2024). By simply interpolating the weights of pre-trained and fine-tuned models, ensembling has shown competitive performance in mitigating forgetting compared to other approaches. In this work, we observe that the same advantage extends to supervised fine-tuning of LLMs. Moreover, beyond its effectiveness in mitigating forgetting on upstream tasks, we also observe a surprising phenomenon of *overadaptation*, which reveals that model ensembling can outperform the fine-tuned model even on downstream tasks, where the fine-tuned model is expected to excel, which also aligns with previous results in Wortsman et al. (2022b); Lin et al. (2024).

However, despite the impressive empirical effectiveness of ensemble methods, the corresponding theoretical insights are still limited, especially in the context of modern over-parameterized model. Most theoretical studies on ensembling have focused on traditional under-parameterized settings (Brown et al., 2005; Lin et al., 2023; 2024), which do not align with the current use of over-parameterized large neural networks. While some works (Allen-Zhu & Li, 2020; Hao et al., 2024) have demonstrated the benefits of ensembling independently trained models, these only address improvements in out-of-distribution (OOD) robustness. To the best of our knowledge, no existing work has addressed the central question:

Why does ensembling achieve such remarkable efficiency, enhancing both generalization on downstream tasks and

mitigating forgetting on upstream tasks?

In this work, we address this question by investigating the role of ensembling in addressing the *overadaptation* phenomenon. During supervised fine-tuning, specialized fine-tuned models overly focus on downstream tasks, leading to a loss of valuable information retained in the pre-trained model. The effectiveness of ensembling can be attributed to its ability to mitigate overadaptation. By simply averaging the weights of the pre-trained and fine-tuned models, the ensemble recovers the information lost during fine-tuning while preserving the knowledge gained from the fine-tuning process.

Specifically, we start with presenting empirical evidence in Section 3 that highlights the harmful effects of overadaptation and demonstrates the efficiency benefits of ensembling in both improving fine-tuning performance and mitigating forgetting. Building on these observations, we develop a formal theoretical framework in Section 4 and Section 5 to analyze the effectiveness of model ensembling, focusing on its impact on fine-tuning tasks and pre-training task retention. We attribute the improved performance of ensembling to its ability to better balance the "bias-variance" trade-off in test error. To simplify our explanation, we model the pre-training and fine-tuning processes using an over-parameterized linear setup. Within the context of canonical linear regression, we represent the pre-trained model as the "ridgeless" estimator on Task 1. For fine-tuning on Task 2, total overfitting (overadaptation) is characterized by "ridgeless" regression, while non-overfitting approaches, such as early stopping, are captured through ridge regression (Lin & Rosasco, 2017; Lu et al., 2022). Our main theoretical findings can be summarized as follows.

1.1. High-level theoretical insights

On the theoretical side, based on an over-parameterized regression setup, after pre-training on Task 1, we fine-tune the model on a specific Task 2. The results are stated on two aspects as follows.

Focusing on the performance on Task 2, we prove that

1. (Poor Performance of pre-trained Model): The pre-trained model exhibits a high "bias" term in test error due to its inability to capture task-specific information for Task 2;
2. (Limitations of fine-tuning without regularizer): Fine-tuning without any regularization, i.e., using a "ridgeless" estimator, results in a high "variance" term in test error due to overfitting (overadaptation) on noisy data, leading to poor performance;
3. (Impact of regularization in fine-tuning): Applying ridge regression during fine-tuning helps mitigate overfitting by achieving a better balance in the "bias-variance" trade-off, thereby reducing test error;
4. (Ensemble for improved trade-off management): Combining the pre-trained model with either the ridge or "ridgeless" fine-tuned model enables a more effective balance of the "bias-variance" trade-off in test error, leading to improved performance on Task 2.

And for the forgetting phenomenon, we consider the performances on both Task 1 and Task 2, establishing that

1. (Impact of regularization in forgetting): Without hurting the performance on Task 2, ridge regression mitigates forgetting on Task 1 by balancing the "bias-variance" trade-off, effectively managing both pre-trained and fine-tuned errors simultaneously.
2. (Ensemble for enhanced forgetting mitigation): Leveraging the pre-trained and fine-tuned models, ensembling further improves such "bias-variance" balance, providing an additional reduction in forgetting.

1.2. Empirical validation overview

Our theoretical analysis is mainly inspired by the "magical" empirical results, which are deferred to Section 3. Here we highlight the main results, to show the consistency between our theoretical results and the empirical phenomenon. Specifically, we give empirical evidences showing that:

1. (Harmful overadaptation in fine-tuning): When the training process extends into the "overfitting" regime (without applying early stopping), fine-tuning performance deteriorates;
2. (Enhanced performance through ensemble): Model ensemble consistently improves model performance across various experiment settings, both achieving better performance on fine-tuning tasks and mitigating forgetting on pre-training tasks efficiently.

2. Related Works

There exists a substantial body of work on model ensemble, pre-training and fine-tuning. In this section, we review the most relevant works to ours.

Model ensemble. Model ensemble has been a popular technique to enhances generalization performance, as documented in the existing literature (Hansen & Salamon, 1990; Krogh & Vedelsby, 1994; Perrone & Cooper, 1995; Opitz & Maclin, 1999; Dietterich, 2000; Zhou et al., 2002; Polikar, 2006; Rokach, 2010; Rame et al., 2022; Arpit et al., 2022; Kumar et al., 2022a; Wortsman et al., 2022a; Lin et al.,

2023). Recently, ensemble the pre-trained and fine-tuned models has been verified to benefit the out-of-distribution robustness (Wortsman et al., 2022b), as well as decreasing forgetting in reinforcement learning from human feedback (Lin et al., 2024).

Understanding model ensemble. On the theoretical side, works explaining the good performance of ensemble are limited, especially in the context of overparameterized models. In traditional underparameterized settings, Brown et al. (2005) decomposes the prediction error of ensemble models into bias, variance and a covariance term between individual models, proposing algorithms to encourage the diversity of individual models to reduce the covariance term; similarly, Lin et al. (2023) showed that ensembling two independently trained models increases feature diversity, thereby improving out-of-distribution (OOD) robustness, while Lin et al. (2024) extended this framework to demonstrate that feature diversity also mitigates forgetting on upstream tasks. In the case of over-parameterized models, recent works such as Allen-Zhu & Li (2020) and Hao et al. (2024) have shown that ensembling independently trained models improves OOD robustness. A comprehensive explanation for the improved model efficiency remains lacking.

Pre-training and fine-tuning. The pre-training and fine-tuning paradigm have become a cornerstone in developing high-performance models across various domains, particularly in large language models (LLMs). Recent advancements have focused on enhancing this framework through various techniques. The LP-FT technique, introduced by Kumar et al. (2022b), involves initializing the pre-trained feature extractor with a reasonably good classifier; Huang et al. (2021) proposed low-rank adaptation (LoRA) to reduce the number of trainable parameters during fine-tuning, which benefits parameter-efficient training; Tian et al. (2023) presented a trainable projected gradient method aimed at enhancing out-of-distribution (OOD) generalization; model ensemble has also demonstrated effectiveness in improving performance, as evidenced by many studies (Cha et al., 2021; Chu et al., 2022; Wortsman et al., 2022a; Lin et al., 2024).

3. Motivated by the Intriguing Phenomenon

We start our observation on instruction-following fine-tuning tasks, highlighting the empirical findings from three key aspects: (i) the impact of using a regularizer versus overfitting (overadaptation), the performance of ensembling pre-trained and fine-tuned models on (ii) fine-tuning tasks and (iii) pre-training tasks.

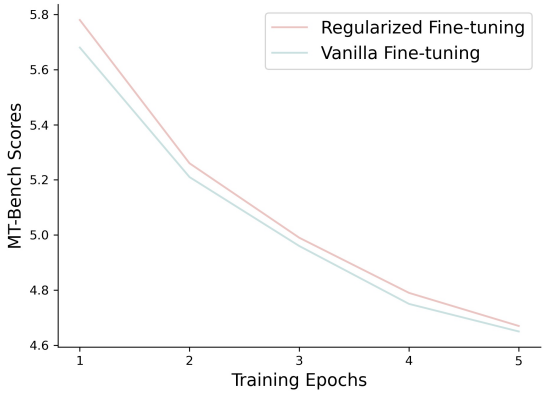


Figure 1. **Early Stop Experiments:** The performance on MT-Bench when the training epoch increases. We conduct vanilla fine-tuning and DiffNorm-Penalty fine-tuning with **Llama-3-8B** on Dolly dataset.

3.1. Datasets and Benchmarks

Our experiments utilize the Dolly dataset (Conover et al., 2023), a popular instruction-following dataset that covers a wide range of tasks, including Creative Writing, Closed QA, Open QA, Summarization, Information Extraction, Classification and Brainstorming, ensuring its high-diversity.

The LLMs' instruction-following ability is evaluated on MT-Bench (Zheng et al., 2023) with single-answer grading. This benchmark prompts conversational assistants with challenging multi-turn open-ended questions and utilizes "LLM-as-a-judge" for evaluation, which comprises 80 questions, evenly distributed across 8 categories: Writing, Roleplay, Extraction, Reasoning, Math, Coding, Knowledge I (STEM), and Knowledge II (humanities/social science). For each response, the LLM judge of GPT-4 will provide a score on a scale of 1 to 10, indicating the overall instruction-following ability of the evaluated conversational assistant.

We also assess LLMs' general ability on MMLU (Hendrycks et al., 2021) and Commonsense-QA (Talmor et al., 2019). MMLU is a dataset containing 57 tasks including mathematics, chemistry, computer science, law, and more, which measures multi-task ability and requires extensive world knowledge and problem-solving ability. Commonsense-QA contains more than 10K real-world common sense questions. It requires LLMs to identify related real-world knowledge and distinguish the distracted answers.

The goal of the experiments is to highlight the strengths of regularization and ensembling in improving the tradeoff between instruction following and general abilities, reveal-

ing that these common generalization techniques not only enhance downstream task performance, but also alleviate forgetting issues in LLMs.

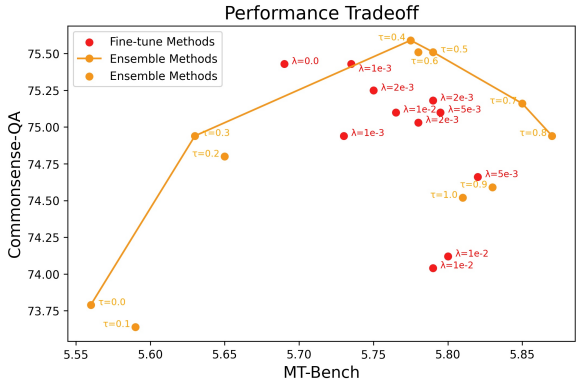


Figure 2. **Experiments on Commonsense-QA and MT-bench:** The performance tradeoff on MT-Bench and Commonsense-QA for ensemble methods and fine-tune methods. The ensemble methods are based on the pre-trained model and the DiffNorm-Penalty model, where the results of different τ values and seeds are presented here. We use **Llama-3-8B** in our experiments.

3.2. Experiment Settings

Our experiments are based on three well-known open-source base models: Llama-3-8B (Dubey et al., 2024), Qwen2-7B (Yang et al., 2024), and Gemma-2-9B (Team et al., 2024), where for each model, we compare the vanilla fine-tuning approach to variants with generalization techniques applied. Depending on whether or not ensembling is used, and which types of regularization are adopted, totaling $2 \times 2 = 4$ variants are proposed to be compared with the vanilla fine-tuning baseline.

For regularization, an additional penalty term of $\|\hat{\theta}\|_2^2$ or $\|\hat{\theta} - \hat{\theta}_1\|_2^2$ is added to the fine-tuning loss, where $\hat{\theta}$ means the fine-tuned model parameters and $\hat{\theta}_1$ represents the pre-trained weights. We denote those variants as Normal-Penalty and DiffNorm-Penalty separately. For ensembling, the parameters of fine-tuned models are weighted-averaged with the pre-trained model. Those variants are denoted as Avg-Norm-Penalty and Avg-DiffNorm-Penalty, respectively. We make our implementation publicly available [1]. For more experimental details, please refer to Appendix A.

3.3. Results

Our empirical results uncover three key issues in the standard fine-tuning process, as shown in Figure 1, Table 1, Figure 2 and Figure 3:

[1]https://github.com/xypan0/LLMForgetting

Overfitting (overadaptation) is harmful during the fine-tuning process. In Figure 1, we show the training process on the Dolly dataset and performance on MT-bench without applying early-stopping. It becomes evident that performance deteriorates quickly with additional epochs, indicating harmful overfitting. Even with early stopping, as seen in Table 1, the use of regularizers, such as Norm-Penalty and DiffNorm-Penalty, improves generalization performance compared to non-regularized fine-tuning (Vanilla-FT).

Ensemble enhances generalization performance. As shown in Table 1, the ensemble of pre-trained and fine-tuned models consistently outperforms the individually fine-tuned models on fine-tuning tasks. This improved performance highlights the effectiveness of model ensembling, aligning with the empirical findings in Lin et al. (2024). Although both methods, Norm-Penalty and DiffNorm-Penalty, use early-stopping to prevent overfitting, they apply different penalties additionally in the training process. Interestingly, in our experiments, Norm-Penalty consistently outperforms DiffNorm-Penalty in almost all settings, suggesting that the choice of regularizer plays a crucial role and warrants further exploration.

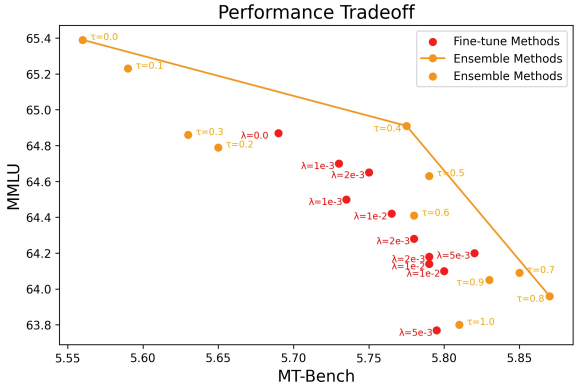


Figure 3. **Experiments on MMLU and MT-bench:** The performance tradeoff on MT-Bench and MMLU for ensemble methods and fine-tune methods. The ensemble methods are based on the pre-trained model and the DiffNorm-Penalty model, where results of different τ values and seeds are presented here. We use **Llama-3-8B** in our experiments.

Ensemble improves trade-off between pre-training and fine-tuning tasks. We also evaluate the trade-off between pre-training tasks (Commonsense-QA and MMLU) and the downstream task (MT-bench). The results are shown in Figure 2 and Figure 3. Compared to individually fine-tuned models, ensemble models achieve better trade-offs on these tasks especially when τ is larger than 0.5. The ensemble models generally have a high MT-bench score while maintaining good performance on Commonsense-QA and

MMLU. The fine-tuned models would suffer from more forgetting if they achieve high MT-bench scores. The results suggest that ensemble methods could achieve a good balance of instruction-following ability and generalization ability.

4. Problem Setup and Notations

Notation. For any matrix A, we use $\|A\|_2$ to denote its ℓ_2 operator norm and use $\mathrm{tr}\{A\}$ to denote its trace. The i-th largest eigenvalue of A is denoted as $\mu_i(A)$. The transposed matrix of A is denoted as A^T. And the inverse matrix of A is denoted as A^{-1}. The notation $a = o(b)$ and $a \ll b$ mean that $a/b \to 0$, $a = \omega(b)$ means that $a/b \to \infty$, $a = O(b)$ means that a/b is bounded, and $a \asymp b$ means $a = O(b)$ as well as $b = O(a)$.

Given the significant performance improvements achieved through ensembling, we seek to understand its benefits in this section. To this end, we analyze its effects within the framework of over-parameterized linear regression.

To be specific, the pre-training process is taken on Task 1, where n i.i.d. training examples $(x_1, y_1), \ldots, (x_n, y_n)$ from distribution $\mathcal{D}$ take values in $\mathbb{R}^p \times \mathbb{R}$ and obey the following linear model with parameter $\theta \in \mathbb{R}^p$:

$$\mathbb{E}[y_i \mid x_i] = x_i^T \theta. \tag{1}$$

To capture the strong performance of the pre-trained model, we consider the min-norm ("ridgeless") estimator on Task 1:

$$\hat{\theta}_1 := X^T (X X^T)^{-1} Y, \tag{2}$$

where $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$ and $Y = [y_1, \ldots y_n]^T \in \mathbb{R}^n$. Starting with the pre-trained estimator $\hat{\theta}_1$, fine-tuning process is taken on Task 2, where n i.i.d. training examples $(\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)$ are sampled from another distribution $\widetilde{\mathcal{D}}$ [2], as well as following another linear model:

$$\mathbb{E}[\widetilde{y}_i \mid \widetilde{x}_i] = \widetilde{x}_i^T \widetilde{\theta}. \tag{3}$$

During the fine-tuning process, we consider two scenarios. The first is "ridgeless" regression, defined by the following objective function:

$$\arg\min_{\theta} \|\theta - \hat{\theta}_1\|_2^2, \quad \text{s.t.} \quad \widetilde{X}\theta = \widetilde{Y},$$

where $\widetilde{X} = [\widetilde{x}_1, \ldots, \widetilde{x}_n]^T \in \mathbb{R}^{n \times p}$ and $\widetilde{Y} = [\widetilde{y}_1, \ldots, \widetilde{y}_n]^T \in \mathbb{R}^n$. Accordingly, the estimator is as

$$\hat{\theta}_2 := \hat{\theta}_1 + \widetilde{X}^T (\widetilde{X}\widetilde{X}^T)^{-1} (\widetilde{Y} - \widetilde{X}\hat{\theta}_1). \tag{4}$$

[2] For simplicity, we assume that the sample sizes for pre-training and fine-tuning are the same. However, our results remain valid even when using a larger dataset for pre-training compared to fine-tuning.

The second objective function incorporates a regularizer with parameter λ to avoid overfitting, as in ridge regression:

$$\arg\min_{\theta} \frac{1}{n} \left\{ \|\widetilde{X}\theta - \widetilde{Y}\|_2^2 + \lambda \|\theta - \hat{\theta}_1\|_2^2 \right\}$$

which implies a solution as

$$\hat{\theta}_\lambda = \widetilde{X}^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1} (\widetilde{Y} - \widetilde{X}\hat{\theta}_1). \tag{5}$$

With the pre-trained estimator and fine-tuning estimator, we also consider the ensemble (weighted averaging) estimator:

$$\hat{\theta}_\lambda^\tau = (1-\tau)\hat{\theta}_1 + \tau\hat{\theta}, \quad \text{where} \quad \hat{\theta} = \hat{\theta}_2, \hat{\theta}_\lambda, \tag{6}$$

with an averaging coefficient $0 \le \tau \le 1$. In our settings, the performance measures for an estimator $\hat{\theta}$ are the excess mean squared errors on Task 1:

$$\begin{aligned}\mathcal{L}_{\mathrm{pre}}(\hat{\theta}) &:= \mathbb{E}_{(x_\star, y_\star, \theta)}\left[(x_\star^T\hat{\theta} - y_\star)^2\right] - \mathbb{E}_{(x_\star, y_\star)}\left[(x_\star^T\theta - y_\star)^2\right] \\ &= \mathbb{E}_{x_\star}\left[(x_\star^T\hat{\theta} - x_\star^T\theta)^2\right],\end{aligned}$$

and Task 2:

$$\begin{aligned}\mathcal{L}_{\mathrm{ft}}(\hat{\theta}) &:= \mathbb{E}_{(\widetilde{x}_\star, \widetilde{y}_\star, \widetilde{\theta})}\left[(\widetilde{x}_\star^T\hat{\theta} - \widetilde{y}_\star)^2\right] - \mathbb{E}_{(\widetilde{x}_\star, \widetilde{y}_\star)}\left[(\widetilde{x}_\star^T\widetilde{\theta} - \widetilde{y}_\star)^2\right] \\ &= \mathbb{E}_{\widetilde{x}_\star}\left[(\widetilde{x}_\star^T\hat{\theta} - \widetilde{x}_\star^T\widetilde{\theta})^2\right],\end{aligned}$$

where the variables $(x_\star, y_\star)$ and $(\widetilde{x}_\star, \widetilde{y}_\star)$ are independent copies of (x_1, y_1) and $(\widetilde{x}_1, \widetilde{y}_1)$ respectively.

Motivation for the objective function. The two objective functions used in the fine-tuning process serve to characterize the presence or absence of early stopping. Specifically, if early stopping is not employed, the overadaptation on downstream tasks can be described as "ridgeless" regression. Conversely, if early stopping is applied, it can be viewed as utilizing a ridge regularizer (Lin & Rosasco, 2017; Lu et al., 2022).

In further analysis, we adopt the following assumptions on settings above:

1. On Task 1, $x_i = \Sigma^{1/2}\eta_i$, where $\Sigma := \mathbb{E}[x_i x_i^T] = \mathrm{diag}[\lambda_1, \ldots, \lambda_p]$, and the components of η_i are independent σ_x-subgaussian random variables with mean zero and unit variance;

2. On Task 2, $\widetilde{x}_i = \widetilde{\Sigma}^{1/2}\widetilde{\eta}_i$, where $\widetilde{\Sigma} := \mathbb{E}[\widetilde{x}_i \widetilde{x}_i^T] = \mathrm{diag}[\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_p]$, and the components of $\widetilde{\eta}_i$ are also independent σ_x-subgaussian random variables with mean zero and unit variance;

3. $\mathbb{E}[y_i - x_i^T\theta | x_i, \theta]^2 = \mathbb{E}[\epsilon_i]^2 = \sigma^2 > 0$, $\mathbb{E}[\widetilde{y}_i - \widetilde{x}_i^T\widetilde{\theta} | \widetilde{x}_i, \widetilde{\theta}]^2 = \mathbb{E}[\widetilde{\epsilon}_i]^2 = \widetilde{\sigma}^2 > 0$;

Table 1. MT-Bench scores of models fine-tuned on Dolly (Conover et al., 2023).

METHODS	REGULARIZER	ENSEMBLING	LLAMA-3-8B	QWEN2-7B	GEMMA-2-9B
VANILLA-FT	-	✗	5.68	6.57	6.52
NORM-PENALTY	$\|\hat{\theta}\|_2^2$	✗	5.84	6.81	6.59
DIFFNORM-PENALTY	$\|\hat{\theta}-\hat{\theta}_1\|_2^2$	✗	5.78	6.68	6.65
AVG-NORM-PENALTY	$\|\hat{\theta}\|_2^2$	✓	**5.96**	**7.10**	6.83
AVG-DIFFNORM-PENALTY	$\|\hat{\theta}-\hat{\theta}_1\|_2^2$	✓	5.85	6.84	**6.89**

4. The true model parameters $\theta, \widetilde{\theta}$ are independent of samples, and could be decomposed as

$$\theta = \theta_c + \alpha_1, \quad \widetilde{\theta} = \theta_c + \alpha_2,$$

where $\theta_c, \alpha_1, \alpha_2$ are independent with each other, as well as holding $\|\theta_c\|_2^2 < \infty$, $\mathbb{E}[\alpha_1\alpha_1^T] = \zeta_1 I_p$ and $\mathbb{E}[\alpha_2\alpha_2^T] = \zeta_2 I_p$, where $\zeta_1, \zeta_2 > 0$.

Discussion on the setting. Pre-trained models often achieve reasonably good generalization across various tasks, and fine-tuning serves to enhance their performance on some specific tasks. Thus, it is reasonable to assume that, despite the differences between Task 1 and Task 2, they share many similarities, which makes the "benign overfitting" estimator $\hat{\theta}_1$ a good "initial point" in fine-tuning process. This leads us to posit that the true model parameters θ and $\widetilde{\theta}$ share significant information, represented by θ_c, while also exhibiting some differences characterized by α_1 and α_2. Additionally, we assume that Σ and $\widetilde{\Sigma}$ share a large amount of same eigenvectors (see Condition 1), further reflecting the similarity between Task 1 and Task 2.

The connection between theoretical and empirical results. The theory aims to provide explanations for the benefits of ensemble, adopting a linear setting for intuitive insights. Such simplifications have been widely used in prior works (Mallinar et al., 2022; Kumar et al., 2022b). The connection between theoretical and empirical results can be established by adopting an NTK explanation, as fine-tuning results in parameters close to pretraining points. Specifically, for a nonlinear neural network $f(x, \vartheta)$ in the NTK regime, we approximate it using a first-order Taylor expansion $f(x, \vartheta) \approx f(x, \vartheta_0) + \nabla_\vartheta f(x, \vartheta_0)^T(\vartheta - \vartheta_0)$. Comparing this with the linear setting $y = x^T\theta_*$, we can interpret the "features" in neural networks, i.e., $\nabla_\vartheta f(x, \vartheta_0)$, as the input x in the linear model, and the trainable parameter $\vartheta - \vartheta_0$ as the parameter θ_* in $y = x^T\theta_*$. Since $f(x, \vartheta_0)$ remains unchanged during training, its effect can be disregarded in this simplification. And in our linear setup, the high-dimensional assumption on x (see Condition 2) can characterize the "features" $\nabla_\vartheta f(x, \vartheta_0)$ in overparameterized neural networks.

5. Main Theorems and Interpretations

In this section, we present the test performance of various estimators and provide explanations for the improvement in both generalization on fine-tuning tasks and forgetting mitigation on pre-training tasks achieved through model ensemble, highlighting the "bias-variance" trade-off phenomenon. To simplify our explanations, we consider the covariance matrices for $x, \widetilde{x}$ as follows,

Condition 1. *Denoting* $\Sigma := \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ *and* $\widetilde{\Sigma} := \mathrm{diag}\{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_p\}$*, we have*

$$\lambda_i = \begin{cases} 1, & i = 1, \ldots, k^*, \\ \gamma, & i > k^*, \end{cases} \quad \widetilde{\lambda}_i = \begin{cases} 1, & i = 1, \ldots, k^*, \\ \gamma, & k^* < i \le \widetilde{p}, \\ 0, & i > \widetilde{p}. \end{cases}$$

And our main theorem is also based on the following condition:

Condition 2. *We consider the following three items:*

1. *(good performance of pre-trained model) For Task 1, there exists a constant* $0 < \xi < 1$*, such that*

$$k^* = O(1), \quad p = \omega(n), \quad p = o(n^{1+\xi}),$$

2. *(sparsity of Task 2) For Task 2, the following conditions hold*

$$\widetilde{p} > n, \quad \widetilde{p} \asymp n, \quad \widetilde{p}\gamma \asymp 1, \quad \widetilde{p}\gamma > 2c_1\widetilde{\sigma}^2/\zeta_2,$$

where $c_1 > 0$ *is a constant only depending on* σ_x*.*

3. *(non-negligible data noise and task difference) For the noise level and the "bias" parameter* α_1, α_2*, we have*

$$\zeta_1 = O(n^{-\xi}), \quad , \zeta_2 \asymp \sigma^2 \asymp \widetilde{\sigma}^2, \quad \zeta_2 = o(1),$$
$$\zeta_2 = \omega(\max\{n^{-(1-\xi)/2}, n^{-\xi}\}).$$

Discussion on the conditions. Condition 1 describes a high-dimensional eigenvalue structure characterized by several "large" eigenvalues alongside many "small" ones. This

structure aligns with the eigenvalue decay observed in general kernel matrices within deep neural networks (Fan & Wang, 2020; Li et al., 2024). We can further validate this assumption by analyzing the eigenvalue distribution of the Hessian matrix in practical models using PyHessian (Yao et al., 2020) and variants of Lanczos algorithms(Zhang et al., 2024), which confirms the presence of several "large" eigenvalues and many "small" eigenvalues. In Condition 2, the first item reflects the "benign" overfitting scenario discussed in Bartlett et al. (2019), which helps to characterize the good performance of the pre-trained model. The second item delineates the "sparse" structure of Task 2, i.e, there are many zero eigenvalues in $\widetilde{\Sigma}$. Such sparse structure observed in fine-tuning tasks reflects the nature of knowledge specialization across different inputs. While pretraining involves diverse inputs encompassing broad knowledge, fine-tuning is performed on specific tasks with a narrower scope, leading to a "sparse" structure in our theoretical formulation. The third item outlines certain conditions that address the significance of data noise and the differences between the two tasks. Since the specific order relationships are technical assumptions intended to clarify our theoretical results more clearly, we believe that relaxing them will not compromise our theoretical intuition. We consider this a possible direction for further exploration.

Before delving into our main results in Theorem 5.1, we introduce two notations:

$$\lambda' := \frac{\widetilde{\sigma}^2}{n\zeta_2},$$

$$\begin{aligned}\tau'(\lambda) := \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\ \Big(\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\ + \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}\Big)^{-1},\end{aligned}$$

The main theorem is then stated as follows:

Theorem 5.1. *For any σ_x, ξ defined above, as Condition 1 and Condition 2 are satisfied, there exists a constant $c > 1$ such that for $\delta \in (0,1)$ and $\ln(1/\delta) < n^{\xi}/c$, with probability at least $1-\delta$ over $X, \widetilde{X}$,*

1. *(the effectiveness of regularization.) for any $0 < \lambda \leq 2\lambda'$, we have*

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_1).$$

2. *(forgetting mitigation with model ensemble.) for any $0 < \lambda \leq 2\lambda'$ and $\tau'(\lambda)/2 \leq \tau < 1$, we have*

$$\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau) < \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) < \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_2) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2).$$

3. *(improving performance on ensemble.) for any $0 \leq \lambda < \lambda'$ and $\tau'(\lambda) \leq \tau < 1$, we have*

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda).$$

The detailed proof is in Appendix C. The results highlight three key insights: (i) selecting an appropriate regularizer during fine-tuning helps reduce overadaptation on noisy samples, leading to improved generalization on fine-tuning task; ensembling the pre-trained and fine-tuned models can decrease overadaptation further, then (ii) enhances performance on fine-tuning task, as well as (iii) mitigating forgetting phenomenon on pre-training task. These benefits can be understood through a *"bias-variance" trade-off phenomenon*:

1. Both $\mathcal{L}_{\mathrm{pre}}$ and $\mathcal{L}_{\mathrm{ft}}$ contain "bias" term and "variance" term.

2. The pre-trained estimator $\hat{\theta}_1$ is mainly dominated by "bias" terms, as it is induced from a sufficiently high-dimensional distribution (see Condition 2). It performs poorly on Task 2 and achieves good performance on Task 1, because it only contains the information in pre-training process, and lacks information specific to Task 2, resulting in a small "bias" term in $\mathcal{L}_{\mathrm{pre}}$, as well as a large "bias" term in $\mathcal{L}_{\mathrm{ft}}$.

3. On the other hand, the "ridgeless" estimator $\hat{\theta}_2$, though it minimizes "bias" error, overfits the noisy training data during fine-tuning, causing a significant "variance" term in $\mathcal{L}_{\mathrm{ft}}$ and both large "bias" term and large "variance" term in $\mathcal{L}_{\mathrm{pre}}$.

4. Introducing a proper regularizer has the ability to achieve better performance on $\mathcal{L}_{\mathrm{ft}}$ and $\mathcal{L}_{\mathrm{pre}} + \mathcal{L}_{\mathrm{ft}}$, by balancing the "bias" and "variance" errors more effectively.

5. The improved generalization on both $\mathcal{L}_{\mathrm{ft}}$ and $\mathcal{L}_{\mathrm{pre}}+\mathcal{L}_{\mathrm{ft}}$ from model ensemble results from further balancing these error terms with a properly chosen weight τ, which applies to both the "ridgeless" estimator $\hat{\theta}_2$ and the ridge-regularized estimator $\hat{\theta}_\lambda$.

Comparing with previous viewpoints. From a traditional statistical perspective, which mainly focuses on limited model complexity, increasing the model complexity typically results in a higher "variance" error and a lower "bias" error (Zhang, 2023). This trade-off suggests that overfitting noisy training data leads to poor generalization and high test error. However, recent advancements have introduced the concept of "benign overfitting" (Bartlett et al., 2019), which suggests that sufficiently large models can achieve superior performance despite overfitting. In our analysis, the pre-training process mainly operates on a high-dimensional distribution $\mathcal{D}$, facilitating strong performance on Task 1 and aligning with the principles of "benign overfitting". Conversely, the fine-tuning phase focuses on a "sparse" structure,

i.e, $\widetilde{\mathcal{D}}$, associated with limited model complexity. This limited complexity explains the observed harmful overfitting during fine-tuning.

Empirical validation. We also conduct simulations across diverse settings to validate Theorem 5.1. The details and results, summarized in Appendix B, demonstrate the strong performance of the ensemble model on both pre-training and fine-tuning tasks, aligning well with our theoretical findings.

6. Proof Sketches of Theorem 5.1

In this section, we summarize the proof sketch of Theorem 5.1, which mainly contains two steps, and the detailed proof is in Appendix C. For simplification, we take the following two notations in further analysis:

$$P_{\widetilde{X},\lambda} := \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}, \quad P_{\widetilde{X}} := \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}.$$

6.1. Excess Risks Approximation

First, we show that with a high probability, the excess risks corresponding to estimators in Equation (2), Equation (4), Equation (5) and Equation (6) could be expressed as:

Lemma 6.1. *As Condition 1 and Condition 2 are satisfied, there exist a constant $c > 1$ such that for $\delta \in (0,1)$ and $\ln(1/\delta) < n^{\xi}/c$, with probability at least $1-\delta$ over $X, \widetilde{X}$, if $0 \le \lambda \le \widetilde{\sigma}^2/(n\zeta_2)$, we have*

$$\begin{aligned}
\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_1) &\approx \zeta_2 \mathrm{tr}\{\widetilde{\Sigma}\},\\
\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2) &\approx \zeta_2 \mathrm{tr}\{(I-P_{\widetilde{X}})^2\widetilde{\Sigma}\} + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},\\
\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) &\approx \zeta_2 \mathrm{tr}\{(I-P_{\widetilde{X},\lambda})^2\widetilde{\Sigma}\}\\
&\quad + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},\\
\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau) &\approx \zeta_2 \mathrm{tr}\{(I-\tau P_{\widetilde{X},\lambda})^2\widetilde{\Sigma}\}\\
&\quad + \tau^2\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_1) \ll \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_2), \quad \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_1) \ll \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda),\\
&\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_1) \ll \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau),\\
&\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_2) \approx \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\}\\
&\qquad + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\},\\
&\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda) \approx \zeta_2 \mathrm{tr}\{P_{\widetilde{X},\lambda}^2\Sigma\} + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\},\\
&\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau) \approx \zeta_2\tau^2 \mathrm{tr}\{P_{\widetilde{X},\lambda}^2\Sigma\}\\
&\qquad + \tau^2\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\}.
\end{aligned}$$

Proof. See details in Appendix C.2 and C.3. □

Using Lemma 6.1, the terms $\mathcal{L}_{\mathrm{ft}}$ and $\mathcal{L}_{\mathrm{pre}} + \mathcal{L}_{\mathrm{ft}}$ in Theorem 5.1 can be primarily determined by two key factors: the terms related to ζ_2 (the difference between the pre-training and fine-tuning tasks), which can be denoted as the "bias" terms, and the terms related to $\widetilde{\sigma}$ (the variance of data noise in the fine-tuning task), which can be denoted as the "variance" term. Insufficient fine-tuning leads to a large "bias" and small "variance", while overadaptation in fine-tuning results in large "variance" and small "bias". By effectively balancing this trade-off, model ensembling can achieve superior performance.

6.2. Estimator Performances Comparison

The effectiveness of regular. After obtaining the results in Lemma 6.1, we could compare the excess risk on different estimators. The results

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_1) \tag{7}$$

could be induced directly from Condition 2. For the excess risk of ridge estimator, taking derivative with respect to λ on its approximation

$$\zeta_2\mathrm{tr}\{(I-P_{\widetilde{X},\lambda})^2\widetilde{\Sigma}\} + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},$$

we find that such excess risk will decrease with the increase of λ within the range $0 \le \lambda \le \widetilde{\sigma}^2/(n\zeta_2)$, which implies

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2), \quad \forall 0 < \lambda \le 2\widetilde{\sigma}^2/(n\zeta_2). \tag{8}$$

Forgetting mitigation with model ensemble. The analysis is similar. With the results in Lemma 6.1, $\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda)$ is mainly dominated by

$$\begin{aligned}
&\zeta_2\mathrm{tr}\{(I-P_{\widetilde{X},\lambda})^2\widetilde{\Sigma}\} + \zeta_2\mathrm{tr}\{P_{\widetilde{X},\lambda}^2\widetilde{\Sigma}\}\\
&+ 2\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}.
\end{aligned}$$

Taking derivative with respect to λ, we find that such term will decrease while $0 \le \lambda \le 2\widetilde{\sigma}^2/(n\zeta_2)$, which implies that

$$\begin{aligned}
\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) < \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_2) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2),\\
\forall 0 < \lambda \le 2\widetilde{\sigma}^2/(n\zeta_2).
\end{aligned} \tag{9}$$

And considering the benefits of model ensemble, for any fixed $\lambda \in (0, 2\widetilde{\sigma}^2/(n\zeta_2)]$, $\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau)$ is mainly dominated by

$$\begin{aligned}
&\zeta_2\mathrm{tr}\{(I-\tau P_{\widetilde{X},\lambda})^2\widetilde{\Sigma}\}\\
&+ \tau^2\zeta_2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}\\
&+ 2\tau^2\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}.
\end{aligned}$$

While we take derivative with respect to τ, such term will increase with the increasing of τ as $\tau'(\lambda)/2 \le \tau \le 1$. So we have

$$\begin{aligned}
\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau) < \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda),\\
\forall \tau'(\lambda)/2 \le \tau < 1.
\end{aligned} \tag{10}$$

Improved fine-tuning performance on ensemble. Finally, for any fixed λ with range $[0, \widetilde{\sigma}^2/(n\zeta_2))$, we could take derivative with respect to τ on the approximated excess risk of ensemble estimator:

$$\zeta_2 \mathrm{tr}\{(I-\tau P_{\widetilde{X},\lambda})^2 \widetilde{\Sigma}\}+\tau^2\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},$$

we found such excess risk will increase with the increasing of τ while $\tau'(\lambda) \leq \tau \leq 1$, so we have

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau) < \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda), \tag{11}$$

while $0 \leq \lambda < \widetilde{\sigma}^2/(n\zeta_2)$ and $\tau'(\lambda) \leq \tau \leq 1$.

Combing all of the results in Equation (7), Equation (8), Equation (9), Equation (10) and Equation (11), we could finish the proof of Theorem 5.1.

7. Conclusion and Discussion

In this work, we bridge the gap in understanding how ensembling pre-trained and fine-tuned models controls overadaptation, as well as enhancing both downstream performance and mitigating forgetting on upstream tasks. Motivated by surprising empirical findings showing that ensembling not only improves fine-tuning outcomes but also preserves pre-trained knowledge, we provide a formal theoretical analysis within an over-parameterized linear setting. Our results reveal that ensembling mitigates overadaptation by effectively balancing the trade-off between "bias" and "variance" errors in excess risk—an issue that regularization alone may not fully resolve. This theoretical insight is further supported by experiments and simulations, which closely align with our predictions.

Our results not only offer a deeper theoretical understanding of ensembling in the context of pre-trained models but also provide practical guidance for enhancing the performance of fine-tuning strategies. This work lays a foundation for future research into refining ensembling methods and exploring their application to broader machine learning tasks.

Acknowledgements

This work was partially supported by NSF grant No. 2416897 and ONR grant No. N000142512318.

Impact Statement

This paper contributes foundational research in the areas of model ensemble within the machine learning community. Our primary goal is to advance the theoretical understanding for the efficient performance of ensemble methods. Given the scope of this research, we do not anticipate immediate ethical concerns or direct societal consequences. Therefore, we believe there are no specific ethical considerations or immediate societal impacts to be emphasized in the context of this work.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Anthropic, A. Introducing claude, 2023.

Arpit, D., Wang, H., Zhou, Y., and Xiong, C. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300v3*, 2019.

Brown, G., Wyatt, J. L., Tino, P., and Bengio, Y. Managing diversity in regression ensembles. *Journal of machine learning research*, 6(9), 2005.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Chu, X., Jin, Y., Zhu, W., Wang, Y., Wang, X., Zhang, S., and Mei, H. Dna: Domain generalization with diversified neural averaging. In *International conference on machine learning*, pp. 4010–4034. PMLR, 2022.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret,

C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damlaj, I., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Fan, Z. and Wang, Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33: 7710–7721, 2020.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Hansen, L. K. and Salamon, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

Hao, Y., Lin, Y., Zou, D., and Zhang, T. On the benefits of over-parameterization for out-of-distribution generalization. *arXiv preprint arXiv:2403.17592*, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Huang, Y., Zhang, Y., Chen, J., Wang, X., and Yang, D. Continual learning for text classification with information disentanglement based regularization. *arXiv preprint arXiv:2104.05489*, 2021.

Koltchinskii, V. and Lounici, K. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Krogh, A. and Vedelsby, J. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Kumar, A., Ma, T., Liang, P., and Raghunathan, A. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence*, pp. 1041–1051. PMLR, 2022a.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022b.

Li, Y., Yu, Z., Chen, G., and Lin, Q. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

Lin, J. and Rosasco, L. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Lin, Y., Tan, L., Hao, Y., Wong, H., Dong, H., Zhang, W., Yang, Y., and Zhang, T. Spurious feature diversification improves out-of-distribution generalization. *arXiv preprint arXiv:2309.17230*, 2023.

Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., Pan, R., Wang, H., Hu, W., Zhang, H., Dong, H., Pi, R., Zhao, H., Jiang, N., Ji, H., Yao, Y., and Zhang, T. Mitigating the alignment tax of RLHF. In *The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://arxiv.org/abs/2309.06256.

Lu, Y., Blanchet, J., and Ying, L. Sobolev acceleration and statistical optimality for learning elliptic equations via gradient descent. *Advances in Neural Information Processing Systems*, 35:33233–33247, 2022.

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., and Nakkiran, P. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Opitz, D. and Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

Perrone, M. P. and Cooper, L. N. When networks disagree: Ensemble methods for hybrid neural networks. In *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*, pp. 342–358. World Scientific, 1995.

Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

Rame, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., and Cord, M. Diverse weight averaging for out-of-distribution generalization. *arXiv preprint arXiv:2205.09739*, 2022.

Rokach, L. Ensemble-based classifiers. *Artificial intelligence review*, 33:1–39, 2010.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Tian, J., He, Z., Dai, X., Ma, C.-Y., Liu, Y.-C., and Kira, Z. Trainable projected gradient method for robust fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7836–7845, 2023.

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

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022a.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022b.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. doi: 10.1017/9781009093057.

Zhang, Y., Chen, C., Ding, T., Li, Z., Sun, R., and Luo, Z. Why transformers need adam: A hessian perspective. *Advances in Neural Information Processing Systems*, 37: 131786–131823, 2024.

Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.-C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zhou, Z.-H., Wu, J., and Tang, W. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1-2):239–263, 2002.

A. Experimental Details

A.1. Hyperparameter Searching

We have conducted hyper-parameter searching on the fine-tuning process and the ensemble process. We fine-tune the models with a global batch size of 16, and an epoch of 1 using Adam optimizer on 8 GPUs. To select a suitable learning rate and penalty, we search the learning rate on $\{5 \times 10^{-6}, 2 \times 10^{-6}, 10^{-6}\}$, and penalty coefficient λ on $\{10^{-2}, 5 \times 10^{-3}, 2 \times 10^{-3}, 10^{-3}\}$. We also search the ensemble weight τ uniformly on $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

To preliminarily validate the performance and choose the hyper-parameter, we have a carefully curated instruction-following dataset. The validation dataset consists of multi-turn conversations between the user and the assistant, covering writing, reasoning, coding, math, STEM, and humanities topics. We prompt GPT4 using the prompt *"Help me generate 3 sets of 2-turn instructions to evaluate the {category} ability of LLMs. The instructions for the second turn need to be highly relevant to the first turn. The following is an example.\n\n\n EXAMPLE:{example}\n TURN1:{turn1}\n TURN2:{turn2}\n"*, where *{category}* corresponds to one of the 8 categories in MT-Bench and *{example}* is one example from MT-Bench. In this way, we obtain a validation dataset that is highly similar to MT-Bench. Specifically, our validation dataset contains 600 samples, evenly distributed across the 8 categories in MT-Bench. We then represent the performance using the loss calculated on the validation dataset.

A.2. Implementation

We implemented our fine-tuning code based on Huggingface Transformers[3] and Accelerate[4] libraries, where Fully Sharded Data Parallel (Zhao et al., 2023) is utilized for model parallel training and acceleration. Our training and evaluation are conducted on 8 NVIDIA H100 GPUs.

A.3. Extension to LoRA

Methods	MT-Bench	Commonsense-QA	MMLU
DiffNorm-Penalty + Ensemble	5.85	**74.49**	63.97
LoRA	5.83	73.71	65.29
LoRA + Ensemble	**5.99**	73.87	**65.31**

Table 2. Performance comparison of different LoRA-based methods.

We also conduct experiments with LoRA (Hu et al., 2021). Specifcally, we set $r = 32$, $\alpha = 32$, $dropout = 0.01$, and target modules to *q-projection, v-projection*. The results are shown in Table 2. We first observe that LoRA can mitigate overadaptation as well but tends to forget more in certain benchmarks, such as Commonsense-QA in comparison with DiffNorm-Penalty. On top of that, it is observed that further ensembling with the pre-trained model yields additional performance improvement in all benchmarks. Such results highlight the benefits of ensemble methods.

A.4. Variance of MT-Bench

We also examine the variance of the MT-Bench by fine-tuning the model 5 trials with different seeds, under Norm-Penalty and ensembling with $\tau = 0.8$. The results are shown in Table 3. Overall, we observe a standard variance of 0.06, which is sufficiently small in comparison to the score gaps in Table 1.

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Standard Deviation
5.84	5.91	5.95	5.96	6.01	**0.06**

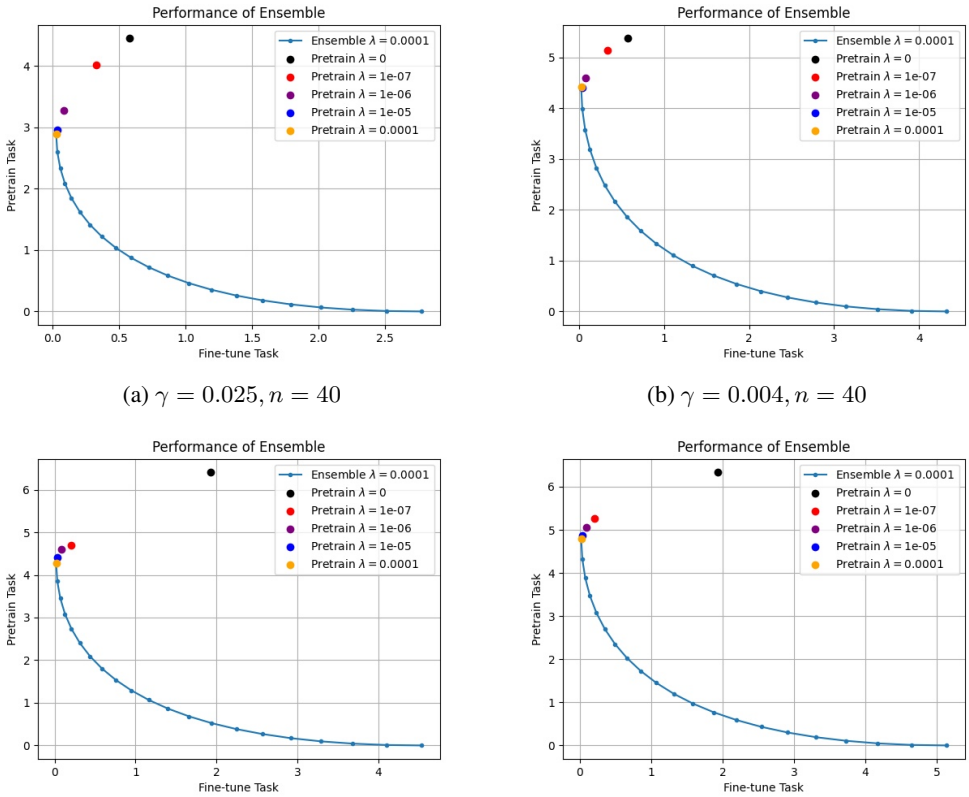
Table 3. Variance estimation of ensembled Norm-Penalty with $\tau = 0.8$ on LLaMA-3-8b.

[3]https://github.com/huggingface/transformers
[4]https://github.com/huggingface/accelerate

B. Empirical Validation for Theorem 5.1

To validate Theorem 5.1, we first utilize artificial datasets, where we construct pre-trained and fine-tuned datasets based on 4 diverse groups of parameters, respectively. Specifically, consider 4 cases for different eigen-value parameter γ and the size of pre-trained set n: (a) $\gamma = n^{-1.0},\ n = 40$; (b) $\gamma = n^{-1.5},\ n = 40$; (c) $\gamma = n^{-1.0},\ n = 40$; (d) $\gamma = n^{-1.5},\ n = 60$. For each case, we set data dimension as $p = 10000$, the size of test data as 1000. We generate a pre-train dataset and a fine-tune dateset from two normal distributions $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$, respectively, where Σ_1 has eigenvalues $\lambda_1 = 1, \lambda_2 = \ldots = \lambda_p = n^{-1.5}$, and Σ_2 has eigenvalues $\lambda_1 = 1, \lambda_2 = \ldots = \lambda_n = n^{-1}, \lambda_{n+1} = \ldots = \lambda_p = 0$. The ground-truth parameter for the pre-train and fine-tune $\theta_c + \alpha_1$ and $\theta_c + \alpha_2$, where $\|\theta_c\|_2 = 1$, $\alpha_1 \sim \mathcal{N}(0, 0.01^2 I_p)$, and $\alpha_2 \sim \mathcal{N}(0, 0.1^2 I_p)$. The variance of data noise is 0.1^2. After obtaining the pre-trained estimator $\hat{\theta}_1$, we fine-tune the estimator on the other dataset to compute the "min-norm" estimator $\hat{\theta}_2$ according to Equation (4), and the estimator with regularizer $\hat{\theta}_\lambda$. We tune the hyper-parameter λ within a range and choose the λ that achieves the best excess risk, which is $\lambda = 0.0001$. Finally, we calculate a group of ensemble estimators $\hat{\theta}_\lambda^\tau$ with τ ranging from 0 to 1, and plot the curve of the error on the pretrain task versus the error in the fine-tuning task for the group of $\hat{\theta}_\lambda^\tau$ in four cases in comparison with the fine-tuned estimator with difference λ, the "min-norm" estimator and the $\hat{\theta}_\lambda$ in Figure 4. According to Figure 4, the performance curve for the ensemble estimator $\hat{\theta}_\lambda^\tau$ achieves better trade-off on the two tasks compared to fine-tuning estimators, which aligns with Theorem 5.1.

Additionally, to validate the performance only on the fine-tuned task, we also consider the four settings mentioned above. To simulate the realistic situation that it is difficult to find the best λ, and we can only tune it into a small range, we take $\lambda = 1e - 7$. Finally, we calculate a group of ensemble estimators $\hat{\theta}_\lambda^\tau$ with τ ranging from 0 to 1, and plot the curve of excess risk for the group of $\hat{\theta}_\lambda^\tau$ in four cases in comparison with the pre-trained estimator, the "min-norm" estimator and the $\hat{\theta}_\lambda$ in Figure 5. The figure implies that if we tune the ensemble parameter τ to the optimal, the ensemble estimator $\hat{\theta}_\lambda^\tau$ performs best, and then the performance decreases in the order of the estimator with ridge regularization $\hat{\theta}_\lambda$, the "min-norm" estimator $\hat{\theta}_2$ and the pre-trained estimator $\hat{\theta}_1$, which aligns with Theorem 5.1.



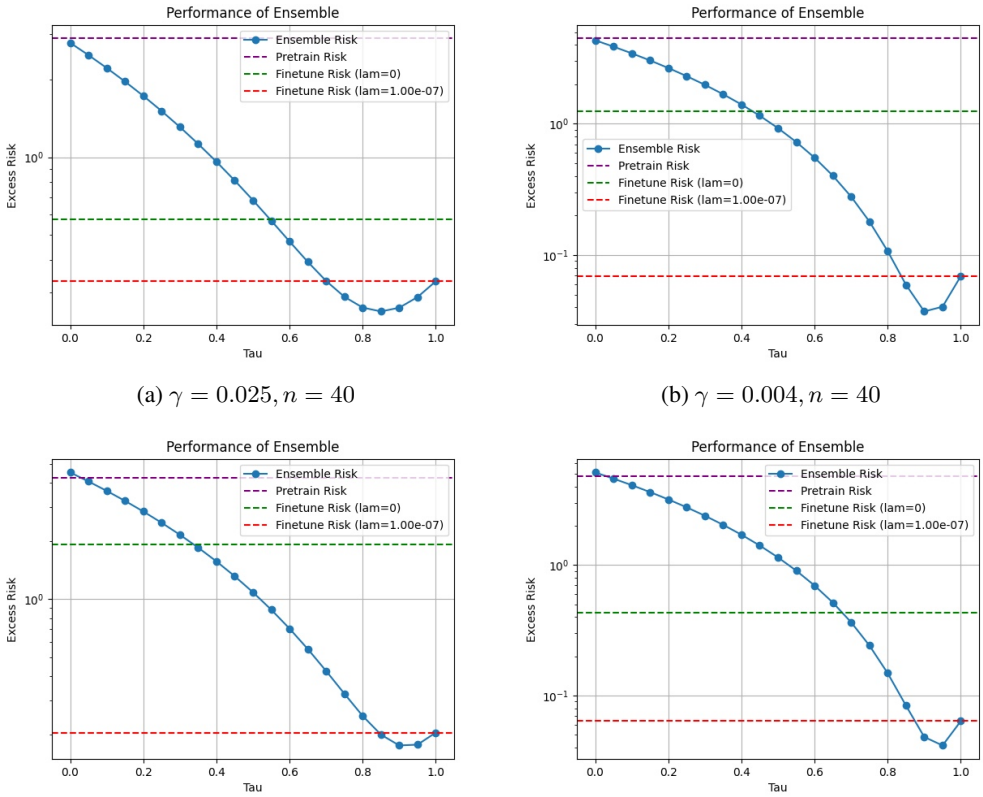
(a) $\gamma = 0.025, n = 40$

(b) $\gamma = 0.004, n = 40$

(c) $\gamma = 0.017, n = 60$

(d) $\gamma = 0.0022, n = 60$

Figure 4. Performance of Ensemble with dimension $p = 10^4$



(a) $\gamma = 0.025, n = 40$ (b) $\gamma = 0.004, n = 40$

(c) $\gamma = 0.017, n = 60$ (d) $\gamma = 0.0022, n = 60$

Figure 5. Performance of Ensemble with dimension $p = 10^4$

C. Proofs

C.1. Notation and Constant List

Before the main proof process, we denote several corresponding constants and notations in Table 4:

Symbol	Value / Expression
c'	$\max\{2, (1+16\ln 3\cdot\sigma_x^2\cdot 54e)32\ln 3\cdot\sigma_x^2\cdot 54e\}$
c	$> 256\cdot(162e)^4\sigma_x^4$
c_1	$\max\{2c', (1/c'-c')^{-1}\}$
c_2	$8(162e)^2\sigma_x^2$
c_3	2
λ_k	the k-th eigenvalue of matrix Σ, i.e, $\mu_k(\Sigma)$
$\widetilde{\lambda}_k$	the k-th eigenvalue of matrix $\widetilde{\Sigma}$, i.e, $\mu_k(\widetilde{\Sigma})$
r_k	$\frac{\sum_{j>k}\lambda_j}{\lambda_{k+1}}$

Table 4. Constant and Notation List

The definition of r_k is the same as the definition in Bartlett et al. (2019). In Bartlett et al. (2019), the critical index $s^*(b)$ for a given $b>0$ is defined as

$$s^*(b) := \inf\{k\geq 0 : r_k \geq bn\}. \tag{12}$$

In our data settings, without lose of generality, we choose $b=1$, and obtain the critical index $s^*(1)=k^*$ in Σ as well as $\widetilde{\Sigma}$.

C.2. Excess Risk Decomposition

The detailed analysis is start with a composition for excess risks. First, the estimators mentioned in main text could be expressed as:

$$\begin{aligned}
\hat{\theta}_1 &= X^T(XX^T)^{-1}(X\theta+\epsilon) = X^T(XX^T)^{-1}(X\theta_c + X\alpha_1 + \epsilon),\\
\hat{\theta}_2 &= \hat{\theta}_1 + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}(\widetilde{X}\widetilde{\theta} - \widetilde{X}\hat{\theta}_1 + \widetilde{\epsilon})\\
&= [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T(XX^T)^{-1}(X\theta_c + X\alpha_1+\epsilon) + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}(\theta_c+\alpha_2) + \widetilde{X}(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{\epsilon}\\
&= [X^T(XX^T)^{-1}X + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X} - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\theta_c\\
&\quad + [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T(XX^T)^{-1}X\alpha_1\\
&\quad + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\alpha_2 + \widetilde{X}(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{\epsilon} + [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T(XX^T)^{-1}\epsilon,\\
\hat{\theta}_\lambda &= \hat{\theta}_1 + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}(\widetilde{X}\widetilde{\theta} - \widetilde{X}\hat{\theta}_1 + \widetilde{\epsilon})\\
&= [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T(XX^T)^{-1}(X\theta_c + X\alpha_1+\epsilon)\\
&\quad + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}(\theta_c+\alpha_2) + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{\epsilon}\\
&= [X^T(XX^T)^{-1}X + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X} - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\theta_c\\
&\quad + [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T(XX^T)^{-1}X\alpha_1 + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}\alpha_2\\
&\quad + [I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T(XX^T)^{-1}\epsilon + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{\epsilon},\\
\hat{\theta}_\lambda^\tau &= (1-\tau)\hat{\theta}_1 + \tau\hat{\theta}_\lambda\\
&= [X^T(XX^T)^{-1}X + \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X} - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\theta_c\\
&\quad + [I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T(XX^T)^{-1}X\alpha_1 + \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}\alpha_2\\
&\quad + [I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T(XX^T)^{-1}\epsilon + \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{\epsilon}.
\end{aligned}$$

Then focusing on the excess risks on Task 1 and Task 2, i.e,

$$\mathcal{L}_{\rm pre}(\hat{\theta}) := \mathbb{E}_{x,\epsilon,\widetilde{\epsilon},\alpha_1,\alpha_2}(\hat{\theta}-\theta)^T\Sigma(\hat{\theta}-\theta), \quad \mathcal{L}_{\rm ft}(\hat{\theta}) := \mathbb{E}_{x,\epsilon,\widetilde{\epsilon},\alpha_1,\alpha_2}(\hat{\theta}-\widetilde{\theta})^T\widetilde{\Sigma}(\hat{\theta}-\widetilde{\theta}),$$

we have the following results:

$$\begin{aligned}
\mathcal{L}_{\rm pre}(\hat{\theta}_1) &= \theta_c^T[I-X^T(XX^T)^{-1}X]\Sigma[I-X^T(XX^T)^{-1}X]\theta_c + \zeta_1{\rm tr}\{[I-X^T(XX^T)^{-1}X]\Sigma\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X\Sigma X^T\} \\
\mathcal{L}_{\rm pre}(\hat{\theta}_2\} &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\Sigma[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{[I-X^T(XX^T)^{-1}X+X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\Sigma \\
&\qquad\qquad [I-X^T(XX^T)^{-1}X+\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}XX^T(XX^T)^{-1}X]\} \\
&\quad + \zeta_2{\rm tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\Sigma[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T\} + \widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \\
\mathcal{L}_{\rm pre}(\hat{\theta}_\lambda) &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{[I-X^T(XX^T)^{-1}X+X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma\} \\
&\qquad\qquad [I-X^T(XX^T)^{-1}X+\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}XX^T(XX^T)^{-1}X]\} \\
&\quad + \zeta_2{\rm tr}\{[\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]^2\Sigma\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \\
\mathcal{L}_{\rm pre}(\hat{\theta}_\lambda^\tau) &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{[I-X^T(XX^T)^{-1}X+\tau X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma\} \\
&\qquad\qquad [I-X^T(XX^T)^{-1}X+\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}XX^T(XX^T)^{-1}X]\} \\
&\quad + \zeta_2\tau^2{\rm tr}\{(\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X})^2\Sigma\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\Sigma[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \tau^2\widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}_{\rm ft}(\hat{\theta}_1) &= \theta_c^T[I-X^T(XX^T)^{-1}X]\widetilde{\Sigma}[I-X^T(XX^T)^{-1}X]\theta_c + \zeta_1{\rm tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\} + \zeta_2{\rm tr}\{\widetilde{\Sigma}\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X\widetilde{\Sigma}X^T\} \\
\mathcal{L}_{\rm ft}(\hat{\theta}_2\} &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{(XX^T)^{-1}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T\} + \zeta_2{\rm tr}\{(I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T\} + \widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
\mathcal{L}_{\rm ft}(\hat{\theta}_\lambda) &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{(XX^T)^{-1}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \zeta_2{\rm tr}\{(I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
\mathcal{L}_{\rm ft}(\hat{\theta}_\lambda^\tau) &= \theta_c^T[I-X^T(XX^T)^{-1}X][I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}][I-X^T(XX^T)^{-1}X]\theta_c \\
&\quad + \zeta_1{\rm tr}\{(XX^T)^{-1}X[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \zeta_2{\rm tr}\{(I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} \\
&\quad + \sigma^2{\rm tr}\{(XX^T)^{-2}X[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I-\tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\quad + \tau^2\widetilde{\sigma}^2{\rm tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}.
\end{aligned}$$

C.3. Term Bounds Estimation

After obtaining the expressions about different terms within the excess risk, we could estimate the related upper and lower bounds now.

C.3.1. Terms corresponding to θ_c

All of the excess risks about $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_\lambda$ and $\hat{\theta}_\lambda^\tau$ contain a term related to θ_c. Here we could obtain their upper bounds for Task 1 as

$$
\begin{aligned}
&\theta_c^T[I - X^T(XX^T)^{-1}X]\Sigma[I - X^T(XX^T)^{-1}X]\theta_c \\
&= \theta_c^T[I - X^T(XX^T)^{-1}X]\left(\Sigma - \frac{1}{n}X^TX\right)[I - X^T(XX^T)^{-1}X]\theta_c] \leq \|\theta_c\|_2^2\|\Sigma - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\Sigma[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\Sigma[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\Sigma - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\Sigma[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\Sigma[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\Sigma - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\Sigma[I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\Sigma[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\Sigma - \frac{1}{n}X^TX\|_2,
\end{aligned}
$$

and for Task 2 as:

$$
\begin{aligned}
&\theta_c^T[I - X^T(XX^T)^{-1}X]\widetilde{\Sigma}[I - X^T(XX^T)^{-1}X]\theta_c \\
&= \theta_c^T[I - X^T(XX^T)^{-1}X]\left(\widetilde{\Sigma} - \frac{1}{n}X^TX\right)[I - X^T(XX^T)^{-1}X]\theta_c] \leq \|\theta_c\|_2^2\|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\widetilde{\Sigma}[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\widetilde{\Sigma}[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2 \\
&\theta_c^T[I - X^T(XX^T)^{-1}X][I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}][I - X^T(XX^T)^{-1}X]\theta_c \\
&\leq \theta_c^T[I - X^T(XX^T)^{-1}X]\widetilde{\Sigma}[I - X^T(XX^T)^{-1}X]\theta_c \leq \|\theta_c\|_2^2\|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2,
\end{aligned}
$$

which implies that we just need to upper bound the following two terms

$$
\|\Sigma - \frac{1}{n}X^TX\|_2, \quad \|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2 \leq \|\widetilde{\Sigma} - \Sigma\|_2 + \|\Sigma - \frac{1}{n}X^TX\|_2.
$$

From Condition 1 and 2, we have

$$
\|\widetilde{\Sigma} - \Sigma\|_2 \leq \max\{\gamma, \gamma\},
$$

and induced by Lemma D.7, with probability at least $1 - e^{-n^\xi}$, we have

$$
\|\Sigma - \frac{1}{n}X^TX\|_2 \leq n^{-\frac{1-\xi}{2}}. \tag{13}
$$

Combining the two results above, we can also obtain

$$
\|\widetilde{\Sigma} - \frac{1}{n}X^TX\|_2 \leq \max\{\gamma, \gamma\} + n^{-\frac{1-\xi}{2}}. \tag{14}
$$

C.3.2. TERMS CORRESPONDING TO ζ_1

For these terms corresponding to ζ_1 in $\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_2), \mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda)$ and $\mathcal{L}_{\mathrm{pre}}(\hat{\theta}_\lambda^\tau)$, we could approximate their upper bounds as:

$$\begin{aligned}
&\zeta_1 \mathrm{tr}\{[I - X^T(XX^T)^{-1}X + X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\Sigma \\
&\qquad [I - X^T(XX^T)^{-1}X + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\} \le \zeta_1 \mathrm{tr}\{\Sigma\}, \\
&\zeta_1 \mathrm{tr}\{[I - X^T(XX^T)^{-1}X + X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\Sigma\} \\
&\qquad [I - X^T(XX^T)^{-1}X + \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\} \le \zeta_1 \mathrm{tr}\{\Sigma\}, \\
&\zeta_1 \mathrm{tr}\{[I - X^T(XX^T)^{-1}X + \tau X^T(XX^T)^{-1}X\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\Sigma\} \\
&\qquad [I - X^T(XX^T)^{-1}X + \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}X^T(XX^T)^{-1}X]\} \le \zeta_1 \mathrm{tr}\{\Sigma\}.
\end{aligned}$$

Similarly, for the terms in $\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_2), \mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda)$ and $\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda^\tau)$, we can obtain their upper bounds as:

$$\begin{aligned}
&\zeta_1 \mathrm{tr}\{(XX^T)^{-1}X[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}]X^T\} \\
&\le \zeta_1 \mathrm{tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\}, \\
&\zeta_1 \mathrm{tr}\{(XX^T)^{-1}X[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\le \zeta_1 \mathrm{tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\}, \\
&\zeta_1 \mathrm{tr}\{(XX^T)^{-1}X[I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]\widetilde{\Sigma}[I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]X^T\} \\
&\le \zeta_1 \mathrm{tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\},
\end{aligned}$$

which implies that for estimating the upper bounds of these terms, we just need to upper bound the following two terms:

$$\zeta_1 \mathrm{tr}\{\Sigma\}, \quad \zeta_1 \mathrm{tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\}.$$

The first term could be expressed as

$$\zeta_1 \mathrm{tr}\{\Sigma\} = \zeta_1 \left(k^* + p\gamma\right), \tag{15}$$

and we just need to approximate the second term. First, recalling the decomposition $\Sigma = \sum_i \lambda_i e_i e_i^T$, we have

$$XX^T = \sum_i \lambda_i z_i z_i^T, \quad X\Sigma X^T = \sum_i \lambda_i^2 z_i z_i^T, \tag{16}$$

in which

$$z_i := \frac{1}{\sqrt{\lambda_i}} X e_i \tag{17}$$

are independent σ_x-subgaussian random vectors in R^n with mean 0 and covariance I. Then we will take the following notations in further analysis:

$$A = XX^T, \quad A_k = \sum_{i>k} \lambda_i z_i z_i^T, \quad A_{-k} = \sum_{i \neq k} \lambda_i z_i z_i^T. \tag{18}$$

Using Woodbury identity, we have

$$\begin{aligned}
\zeta_1 \mathrm{tr}\{(XX^T)^{-1}X\widetilde{\Sigma}X^T\} &= \zeta_1 \sum_i \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-1} z_i \\
&= \zeta_1 \left(\sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i z_i^T A_{-i}^{-1} z_i}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i} + \sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-1} z_i \right).
\end{aligned} \tag{19}$$

For any $i = 1, \ldots, k^*$, we have

$$z_i^T A_{-i}^{-1} z_i \le \frac{\|z_i\|_2^2}{\mu_n(A_{-i})}, \quad z_i^T A_{-i}^{-1} z_i \ge (\Pi_{\mathscr{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathscr{L}_i} z_i) \ge \frac{\| \Pi_{\mathscr{L}_i} z_i \|_2^2}{\mu_{k^*+1}(A_{-i})}, \tag{20}$$

where $\mathscr{L}_i$ is denoted as the subspace in $\mathbb{R}^n$, related to the $n-k^*$ eigenvalues of A_{-i}. Considering Lemma D.1 and Lemma D.2, with probability at least $1-5e^{-n/c}$, we have

$$\frac{1}{c_1}\lambda_{k^*+1}r_{k^*} \le \mu_n(A_{-i}) \le \mu_{k^*+1}(A_{-i}) \le c_1\lambda_{k^*+1}r_{k^*}, \quad \|z_i\|_2^2 \le c_2 n, \quad \|\Pi_{\mathscr{L}_i} z_i\|_2^2 \ge n/c_3,$$

where c_1, c_2, c_3 are constants only depending on σ_x. The results above imply that

$$z_i^T A_{-i}^{-1} z_i \le \frac{c_1 c_2 n}{\lambda_{k^*+1}r_{k^*}}, \quad z_i^T A_{-i}^{-1} z_i \ge \frac{n}{c_1 c_3 \lambda_{k^*+1} r_{k^*}},$$

so with probability at least $1-5e^{-n/c}$, we have

$$\sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i z_i^T A_{-1}^{-1} z_i}{1+\lambda_i z_i^T A_{-1}^{-1} z_i} \le \sum_{i=1}^{k^*} \widetilde{\lambda}_i \frac{c_1 c_2 n \lambda_i/(\lambda_{k^*+1}r_{k^*})}{1+\lambda_i c_1 c_2 n/(\lambda_{k^*+1}r_{k^*})} \le \sum_{i=1}^{k^*} \widetilde{\lambda}_i = k^*, \tag{21}$$

where the last equality is from Condition 1. For the remaining part, considering Lemma D.1, with probability at least $1-2e^{-n/c}$, we have

$$\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-1} z_i \le \frac{\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2}{\mu_n(XX^T)} \le c_1^2 \frac{\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2}{\lambda_{k^*+1}r_{k^*}} \le c_1^2 \frac{\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2}{\lambda_{k^*+1}r_{k^*}},$$

and further considering Lemma D.6, with probability at least $1-2e^{-n/c}$, we have

$$\begin{aligned}\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2 &\le n \sum_{i>k^*} \widetilde{\lambda}_i \lambda_i + 2\sigma_x \max\left\{ \frac{n\widetilde{\lambda}_{k^*+1}\lambda_{k^*+1}}{c}, \sqrt{n \sum_{i>k^*} \widetilde{\lambda}_i^2 \lambda_i^2 / c} \right\} \\ &= n\widetilde{p}\gamma\gamma + 2\sigma_x \max\left\{ \frac{n\gamma\gamma}{c}, \gamma\gamma\sqrt{\frac{n\widetilde{p}}{c}} \right\} \le 2n\widetilde{p}\gamma\gamma,\end{aligned}$$

where the second equality is from Condition 1 and the last inequality is from Condition 2. This result implies that with probability at least $1-4e^{-n/c}$, we have

$$\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-1} z_i \le \frac{c_1^2 2n\widetilde{p}\gamma\gamma}{\lambda_{k^*+1}r_{k^*}} = \frac{c_1^2 2n\widetilde{p}\gamma\gamma}{p\gamma} = \frac{2c_1^2 n\widetilde{p}\gamma}{p}. \tag{22}$$

Combing the results in Eq. Equation (19), Equation (21) and Equation (22), with probability at least $1-10e^{-n/2c}$, we could obtain that

$$\zeta_1 \mathrm{tr}\{(XX^T)^{-1} X\widetilde{\Sigma} X^T\} \le \zeta_1 \left(k^* + \frac{2c_1^2 n\widetilde{p}\gamma}{p} \right). \tag{23}$$

C.3.3. Terms corresponding to σ

Similar to the analysis above, these terms corresponding to σ in excess risks on Task 1 and Task 2 can be upper bounded by

$$\sigma^2 \mathrm{tr}\{(XX^T)^{-2} X\Sigma X^T\}, \quad \sigma^2 \mathrm{tr}\{(XX^T)^{-2} X\widetilde{\Sigma} X^T\}$$

respectively, and the estimation of its upper bound is similar to the previous item. Here we first consider the second item, by Woodbury identity, we have

$$\begin{aligned}\sigma^2 \mathrm{tr}\{(XX^T)^{-2} X\widetilde{\Sigma} X^T\} &= \sum_i \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-2} z_i \\ &= \sigma^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i z_i^T A_{-i}^2 z_i}{[1+\lambda_i z_i^T A_{-i}^{-1} z_i]^2} + \sigma^2 \sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-2} z_i.\end{aligned} \tag{24}$$

From the results in Equation (20), with probability at least $1-5e^{-n/c}$, for any $i=1,\ldots,k^*$, we have

$$z_i^T A_{-i}^{-2} z_i \leq \frac{c_1^2 c_2 n}{(\lambda_{k^*+1} r_{k^*})^2}, \quad z_i^T A_{-i}^{-1} z_i \geq \frac{n}{c_1 c_3 \lambda_{k^*+1} r_{k^*}},$$

which implies that

$$\begin{aligned}
\sigma^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i z_i^T A_{-i}^2 z_i}{[1+\lambda_i z_i^T A_{-i}^{-1} z_i]^2} &\leq \sigma^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i c_1^2 c_2 n/(\lambda_{k^*+1} r_{k^*})^2}{[1+\lambda_i n/(c_1 c_3 \lambda_{k^*+1} r_{k^*})]^2} \\
&\leq \sigma^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i c_1^4 c_2 c_3^2 n}{n^2 \lambda_i^2 + c_1^2 c_3^2 \lambda_{k^*+1}^2 r_{k^*}^2} \\
&\leq \sigma^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i \lambda_i c_1^4 c_2 c_3^2 n}{n^2 \lambda_i^2} = \sigma^2 c_1^4 c_2 c_3^2 \frac{1}{n} \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i}{\lambda_i} = \sigma^2 c_1^4 c_2 c_3^2 \frac{k^*}{n},
\end{aligned} \tag{25}$$

where the second inequality is from $(a+b)^2 \geq a^2+b^2$ while $a,b>0$, the third inequality is from $a^2+b^2 \geq a^2$, and the last equality is from Condition 1. And for another part, considering Lemma D.1, with probability at least $1-2e^{-n/c}$, we have

$$\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-2} z_i \leq \frac{\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2}{\mu_n (XX^T)^2} \leq c_1^2 \frac{\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2}{\lambda_{k^*+1}^2 r_{k^*}^2},$$

and further considering Lemma D.6, with probability at least $1-2e^{-n/c}$, we have

$$\begin{aligned}
\sum_{i>k^*} \widetilde{\lambda}_i \lambda_i \|z_i\|_2^2 &\leq n \sum_{i>k^*} \widetilde{\lambda}_i \lambda_i + 2\sigma_x \max\left\{ \frac{n\widetilde{\lambda}_{k^*+1} \lambda_{k^*+1}}{c}, \sqrt{n \sum_{i>k^*} \widetilde{\lambda}_i^2 \lambda_i^2 / c} \right\} \\
&= n\widetilde{p}\gamma\gamma + 2\sigma_x \max\left\{ \frac{n\gamma^2}{c}, \gamma^2 \sqrt{\frac{n\widetilde{p}}{c}} \right\} \leq 2n\widetilde{p}\gamma^2,
\end{aligned}$$

where the second equality is from Condition 1 and the last inequality is from Condition 2. which implies that with probability at least $1-4e^{-n/c}$, we have

$$\sigma^2 \sum_{i>k^*} \widetilde{\lambda}_i \lambda_i z_i^T (XX^T)^{-2} z_i \leq \sigma^2 2c_1^2 \frac{n\widetilde{p}\gamma\gamma}{\lambda_{k^*+1}^2 r_{k^*}^2} = \sigma^2 2c_1^2 \frac{n\widetilde{p}}{p^2}, \tag{26}$$

where the last equality is from Condition 1. Combining the results in Equation (24) Equation (25) and Equation (26), with probability at least $1-10e^{-n/2c}$, we have

$$\sigma^2 \mathrm{tr}\{(XX^T)^{-2} X \widetilde{\Sigma} X^T\} \leq \sigma^2 \left(c_1^4 c_2 c_3^2 \frac{k^*}{n} + 2c_1^2 \frac{n\widetilde{p}}{p^2} \right). \tag{27}$$

The analysis for the first item is similar, which implies that with probability at least $1-10e^{-n/2c}$, we could obtain that

$$\sigma^2 \mathrm{tr}\{(XX^T)^{-2} X \Sigma X^T\} \leq \sigma^2 \left(c_1^4 c_2 c_3^2 \frac{k^*}{n} + 2c_1^2 \frac{n}{p} \right). \tag{28}$$

C.3.4. TERMS CORRESPONDING TO ζ_2

In $\mathcal{L}_{\mathrm{pre}}$, to approximate the upper and lower bounds of terms related to ζ_2, we need to estimate:

$$\zeta_2 \mathrm{tr}\{[\widetilde{X}^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1} \widetilde{X}]^2 \Sigma\} = \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2} \widetilde{X}\widetilde{X}^T \widetilde{X} \Sigma \widetilde{X}^T\},$$

for any $\lambda \geq 0$. And in $\mathcal{L}_{\text{ft}}$, for the terms related to ζ_2, we have the following results:

$$\begin{aligned}
\zeta_2 \text{tr}\{(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} &= \zeta_2 \text{tr}\{\widetilde{\Sigma}\} - \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}, \\
\zeta_2 \text{tr}\{(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} &\leq \zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})\} \\
&= \zeta_2 \text{tr}\{\widetilde{\Sigma}\} - \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}, \\
\zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\} &\geq \zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X})^2\} \\
&= \zeta_2 \text{tr}\{\widetilde{\Sigma}\} - \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}, \\
\zeta_2 \text{tr}\{(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} &\leq \zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})\} \\
&= \zeta_2 \text{tr}\{\widetilde{\Sigma}\} - \zeta_2\tau \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}, \\
\zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\} &\geq \zeta_2 \text{tr}\{\widetilde{\Sigma}(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X})^2\} \\
&= \zeta_2 \text{tr}\{\widetilde{\Sigma}\} - \zeta_2(2\tau - \tau^2)\text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},
\end{aligned}$$

so here we need to estimate the upper and lower bounds for term

$$\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}$$

for any $\lambda \geq 0$. The analysis is similar to the analysis above, recalling the decomposition $\widetilde{\Sigma} = \sum_i \widetilde{\lambda}_i e_i e_i^T$, we have

$$\widetilde{X}\widetilde{X}^T = \sum_i \widetilde{\lambda}_i \widetilde{z}_i \widetilde{z}_i^T, \quad \widetilde{X}\widetilde{\Sigma}\widetilde{X}^T = \sum_i \widetilde{\lambda}_i^2 \widetilde{z}_i \widetilde{z}_i^T,$$

in which

$$\widetilde{z} := \frac{1}{\sqrt{\widetilde{\lambda}_i}}\widetilde{X}e_i,$$

are independent σ_x-subgaussian random vectors in $\mathbb{R}^n$ with mean zero and covariance I. So we take the following notations in further analysis:

$$\widetilde{A} = \widetilde{X}\widetilde{X}^T, \quad \widetilde{A}_k = \sum_{i>k} \widetilde{\lambda}_i \widetilde{z}_i \widetilde{z}_i^T, \quad \widetilde{A}_{-k} = \sum_{i \neq k} \widetilde{\lambda}_i \widetilde{z}_i \widetilde{z}_i^T \tag{29}$$

Starting with the analysis of $\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}$, we consider its upper bound firstly. Using Woodbury identity, we have

$$\begin{aligned}
&\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&= \zeta_2 \sum_i \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{z}_i \\
&= \zeta_2 \left(\sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i}{1 + \widetilde{\lambda}_i \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i} + \sum_{i>k^*} \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{z}_i \right).
\end{aligned} \tag{30}$$

For any $i = 1, \ldots, k^*$, we have

$$\begin{aligned}
&\widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i \leq \frac{\|\widetilde{z}_i\|_2^2}{\mu_n(\widetilde{A}_{-i}) + n\lambda}, \\
&\widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i \geq (\Pi_{\widetilde{\mathscr{L}_i}}\widetilde{z}_i)^T (\widetilde{A}_{-i} + n\lambda I)^{-1}(\Pi_{\widetilde{\mathscr{L}_i}}\widetilde{z}_i) \geq \frac{\| \Pi_{\widetilde{\mathscr{L}_i}} z_i \|_2^2}{\mu_{k^*+1}(\widetilde{A}_{-i})},
\end{aligned} \tag{31}$$

where $\widetilde{\mathscr{L}_i}$ is denoted as the subspace in $\mathbb{R}^n$, related to the $n - k^*$ eigenvalues of $\widetilde{A}_{-i}$. Considering Lemma D.1 and Lemma D.2, with probability at least $1 - 5e^{-n/c}$, we have

$$\frac{1}{c_1}\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} \leq \mu_n(\widetilde{A}_{-i}) \leq \mu_{k^*+1}(\widetilde{A}_{-i}) \leq c_1 \widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}, \quad \|\widetilde{z}_i\|_2^2 \leq c_2 n, \quad \|\Pi_{\widetilde{\mathscr{L}_i}}\widetilde{z}_i\|_2^2 \geq n/c_3, \tag{32}$$

where c_1, c_2, c_3 are constants only depending on b, σ_x. The results above imply that

$$\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i \le \frac{c_1c_2n}{\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda}, \quad \widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda)^{-1}z_i \ge \frac{n}{c_1c_3(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}.$$

so with probability at least $1-5e^{-n/c}$, we have

$$\begin{aligned}\sum_{i=1}^{k^*}\frac{\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i}{1+\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i} &\le \sum_{i=1}^{k^*}\widetilde{\lambda}_i\frac{c_1c_2n\widetilde{\lambda}_i/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}{1+\widetilde{\lambda}_ic_1c_2n/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}\\ &= k^*\frac{c_1c_2n/(\widetilde{p}\gamma+n\lambda)}{1+c_1c_2n/(\widetilde{p}\gamma+n\lambda)}\\ &\le k^*\min\{\frac{c_1c_2n}{c_1c_2n+\widetilde{p}\gamma},\frac{c_1c_2}{c_1c_2+\lambda}\},\end{aligned} \tag{33}$$

where the last inequality is from $a+b\ge\max\{a,b\}$. For the remaining part, considering Lemma D.1, with probability at least $1-2e^{-n/c}$, we have

$$\sum_{i>k^*}\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{z}_i \le \frac{\sum_{i>k^*}\widetilde{\lambda}_i^2\|\widetilde{z}_i\|_2^2}{\mu_n(\widetilde{X}\widetilde{X}^T)+n\lambda} \le c_1^2\frac{\sum_{i>k^*}\widetilde{\lambda}_i^2\|\widetilde{z}_i\|_2^2}{\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda},$$

and further considering Lemma D.6, with probability at least $1-2e^{-n/c}$, we have

$$\begin{aligned}\sum_{i>k^*}\widetilde{\lambda}_i^2\|\widetilde{z}_i\|_2^2 &\le n\sum_{i>k^*}\widetilde{\lambda}_i^2+2\sigma_x\max\left\{\frac{n\widetilde{\lambda}_{k^*+1}^2}{c},\sqrt{n\sum_{i>k^*}\widetilde{\lambda}_i^4/c}\right\}\\ &= n\widetilde{p}\gamma^2+2\sigma_x\max\left\{\frac{n\gamma^2}{c},\gamma^2\sqrt{\frac{n\widetilde{p}}{c}}\right\}\le 2n\widetilde{p}\gamma^2,\end{aligned}$$

where the second equality is from Condition 1 and the last inequality is from Condition 2. This result implies that with probability at least $1-4e^{-n/c}$, we have

$$\sum_{i>k^*}\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{z}_i \le \frac{c_1^2 2n\widetilde{p}\gamma^2}{\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda} = \frac{2c_1^2n\widetilde{p}\gamma^2}{\widetilde{p}\gamma+n\lambda} \le 2c_1^2\min\left\{n\gamma,\frac{\widetilde{p}\gamma^2}{\lambda}\right\}. \tag{34}$$

Combing the results in Eq. Equation (30), Equation (33) and Equation (34), with probability at least $1-10e^{-n/2c}$, we could obtain that

$$\zeta_2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}\le\zeta_2\left(k^*\min\{\frac{c_1c_2n}{c_1c_2n+\widetilde{p}\gamma},\frac{c_1c_2}{c_1c_2+\lambda}\}+2c_1^2\min\left\{n\gamma,\frac{\widetilde{p}\gamma^2}{\lambda}\right\}\right). \tag{35}$$

Then we turn to its lower bound. Considering Equation (31) for any index $i=1,\dots,\infty$, as

$$\mu_{k^*+1}(\widetilde{A}_{-i})\le\mu_{k^*+1}(\widetilde{X}\widetilde{X}^T)\le c_1\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}$$

is always satisfied, with probability at least $1-5e^{-n/c}$ we can get a lower bound as

$$\begin{aligned}\frac{\widetilde{\lambda}_i^2\widetilde{x}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i}{1+\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i} &\ge \frac{\widetilde{\lambda}_i^2n/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}{c_1c_3+\widetilde{\lambda}_in/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}\\ &\ge\frac{1}{c_1c_3}\frac{\widetilde{\lambda}_i^2n/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}{1+\widetilde{\lambda}_in/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)}>0,\end{aligned} \tag{36}$$

and for the whole trace term $\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}$, due to Lemma D.5, with probability at least $1 - 10e^{-n/c}$, we have

$$\begin{aligned}
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &= \zeta_2 \sum_i \frac{\widetilde{\lambda}_i^2 \widetilde{x}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i}{1 + \widetilde{\lambda}_i \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i} \\
&\geq \frac{\zeta_2}{2c_1c_3} \sum_i \frac{\widetilde{\lambda}_i^2 n/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + n\lambda)}{1 + \widetilde{\lambda}_i n/(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + n\lambda)} \\
&\geq \frac{\zeta_2}{6c_1c_3} \sum_i \min\left\{\frac{\widetilde{\lambda}_i^2}{\lambda}, \widetilde{\lambda}_i, \frac{n\widetilde{\lambda}_i^2}{\widetilde{\lambda}_{k^*+1} r_{k^*}}\right\} \\
&= \frac{\zeta_2}{6c_1c_3}\left(k^* \min\{\frac{1}{\lambda}, 1\} + (\widetilde{p} - k^*)\min\{\frac{\gamma^2}{\lambda}, \frac{n\gamma}{\widetilde{p}}\}\right),
\end{aligned} \tag{37}$$

where the first equality is from Woodbury identity, the first inequality is from Equation (36), the second inequality is from the fact

$$(a + b + c)^{-1} \geq (3\max\{a, b, c\})^{-1} = \frac{1}{3}\min\{a^{-1}, b^{-1}, c^{-1}\}, \quad a, b, c > 0,$$

and the last equality is induced from Condition 1 and Condition 2. And combing Equation (35) and Equation (37), we could consider different situations with respect to the value of λ.

While $\lambda \leq \widetilde{p}\gamma/n$, we have

$$\begin{aligned}
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\leq \zeta_2\left(\frac{c_1c_2k^*n}{c_1c_2n + \widetilde{p}\gamma} + 2c_1^2 n\gamma\right), \\
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\geq \frac{\zeta_2}{6c_1c_3}\left(k^* + \frac{n\gamma(\widetilde{p} - k^*)}{\widetilde{p}}\right).
\end{aligned} \tag{38}$$

And while $\lambda \geq \widetilde{p}\gamma/n$, with a high probability, we have

$$\begin{aligned}
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\leq \zeta_2\left(\frac{c_1c_2k^*}{c_1c_2 + \lambda} + \frac{2c_1^2\widetilde{p}\gamma^2}{\lambda}\right), \\
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\geq \frac{\zeta_2}{6c_1c_3}\left(k^* \min\{\frac{1}{\lambda}, 1\} + \frac{\gamma^2(\widetilde{p} - k^*)}{\lambda}\right).
\end{aligned} \tag{39}$$

Then we turn to the term $\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\Sigma\widetilde{X}^T\}$. While we have $n\lambda \leq \widetilde{p}\gamma/c_1$, with probability at least $1 - 5e^{-n/c}$, we have

$$n\lambda \leq \mu_n(\widetilde{X}\widetilde{X}^T) \Longrightarrow n\lambda I \preceq \widetilde{X}\widetilde{X}^T \Longleftarrow \widetilde{X}\widetilde{X}^T + n\lambda I \preceq 2\widetilde{X}\widetilde{X}^T,$$

which implies that

$$\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \geq \frac{\zeta_2}{4}\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\}, \tag{40}$$

as we also have

$$\widetilde{X}\widetilde{X}^T + \lambda I \succeq \widetilde{X}\widetilde{X}^T \Longrightarrow \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \leq \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\}. \tag{41}$$

Combing both of the two results above, we just need to estimate the term $\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\}$. Using Woodbury identity, we have

$$\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\} = \zeta_2 \sum_i \widetilde{\lambda}_i\lambda_i \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{z}_i,$$

and recalling the identity

$$\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} = \zeta_2 \sum_i \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{z}_i,$$

comparing such two terms and Condition 1, we have

$$\widetilde{\lambda}_i\lambda_i = \widetilde{\lambda}_i^2, \quad \forall i = 1, \ldots, p,$$

which implies that

$$\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\Sigma\widetilde{X}^T\} = \zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}. \tag{42}$$

Then considering Equation (38), Equation (40) and Equation (41), while $n\lambda \leq \widetilde{p}\gamma/c_1$, we have

$$\begin{aligned}
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\leq \zeta_2 \left(\frac{c_1 c_2 k^* n}{c_1 c_2 n + \widetilde{p}\gamma} + 2c_1^2 n\gamma\right), \\
\zeta_2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} &\geq \frac{\zeta_2}{24 c_1 c_3}\left(k^* + \frac{n\gamma(\widetilde{p} - k^*)}{\widetilde{p}}\right).
\end{aligned} \tag{43}$$

C.3.5. Terms corresponding to $\widetilde{\sigma}$

Finally, for the terms related to $\widetilde{\sigma}$, we need to estimate the bounds for the terms $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}^T\Sigma\widetilde{X}\}$ and $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}^T\Sigma\widetilde{X}\}$ in $\mathcal{L}_{\mathrm{pre}}$, and the terms $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}^T\widetilde{\Sigma}\widetilde{X}\}$ and $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}^T\widetilde{\Sigma}\widetilde{X}\}$ in $\mathcal{L}_{\mathrm{ft}}$, i.e, for the terms $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\}$ and $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}$ in which $\lambda \geq 0$.

Firstly, we consider the approximation on $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}$. For its upper bound, we could take Woodbury identity as follows:

$$\begin{aligned}
&\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&= \widetilde{\sigma}^2 \sum_i \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{z}_i \\
&= \widetilde{\sigma}^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-2}\widetilde{z}_i}{[1 + \widetilde{\lambda}_i \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-2}\widetilde{z}_i]^2} + \widetilde{\sigma}^2 \sum_{i>k^*} \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{z}_i.
\end{aligned} \tag{44}$$

According to Equation (31) and Equation (32), with probability at least $1 - 5e^{-n/c}$, for any index $i = 1, \ldots, k^*$, we have

$$\widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-2}\widetilde{z}_i \leq \frac{c - 1^2 c_2 n}{(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + n\lambda)^2}, \quad \widetilde{z}_i (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i \geq \frac{n}{c_1 c_3 (\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + n\lambda)}, \tag{45}$$

which implies that

$$\begin{aligned}
\widetilde{\sigma}^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-2}\widetilde{z}_i}{[1 + \widetilde{\lambda}_i \widetilde{z}_i^T (\widetilde{A}_{-i} + n\lambda I)^{-1}\widetilde{z}_i]^2} &\leq \widetilde{\sigma}^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i^2 c_1^2 c_2 n / (\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + n\lambda)^2}{[1 + \widetilde{\lambda}_i n / (c_1 c_3 \widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*} + c_1 c_3 n\lambda)]^2} \\
&\leq \widetilde{\sigma}^2 \sum_{i=1}^{k^*} \frac{\widetilde{\lambda}_i^2 c_1^4 c_2 c_3^2 n}{n^2 \widetilde{\lambda}_i^2 + c_1^2 c_3^2 \widetilde{\lambda}_{k^*+1}^2 \widetilde{r}_{k^*}^2 + c_1^2 c_3^2 n^2 \lambda^2} \\
&\leq \widetilde{\sigma}^2 c_1^4 c_2 c_3^2 \sum_{i=1}^{k^*} \min\left\{\frac{1}{n}, \frac{n\widetilde{\lambda}_i^2}{\widetilde{\lambda}_{k^*+1}^2 \widetilde{r}_{k^*}^2}, \frac{\widetilde{\lambda}_i^2}{\lambda^2}\right\} \\
&= \widetilde{\sigma}^2 c_1^4 c_2 c_3^2 k^* \min\left\{\frac{1}{n}, \frac{1}{\lambda^2}\right\},
\end{aligned} \tag{46}$$

where the third inequality is from $a^2 + b^2 + c^2 \geq \max\{a^2, b^2, c^2\}$, and the last equality is from Condition 1 and Condition 2. As for the remaining part, considering Lemma D.1, with probability at least $1 - 2e^{-n/c}$, we have

$$\sum_{i>k^*} \widetilde{\lambda}_i^2 \widetilde{z}_i^T (\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{z}_i \leq \frac{\sum_{i>k^*} \widetilde{\lambda}_i^2 \|\widetilde{z}_i\|_2^2}{\mu_n(\widetilde{X}\widetilde{X}^T)^2 + n^2\lambda^2} \leq c_1^2 \frac{\sum_{i>k^*} \widetilde{\lambda}_i^2 \|\widetilde{z}_i\|_2^2}{\widetilde{\lambda}_{k^*+1}^2 \widetilde{r}_{k^*}^2 + n^2\lambda^2} = c_1^2 \frac{\sum_{i>k^*} \widetilde{\lambda}_i^2 \|\widetilde{z}_i\|_2^2}{\widetilde{p}^2\gamma^2 + n^2\lambda^2},$$

and further considering Lemma D.6, with probability at least $1 - 2e^{-n/c}$, we have

$$\begin{aligned}
\sum_{i>k^*} \widetilde{\lambda}_i^2 \|\widetilde{z}_i\|_2^2 &\leq n \sum_{i>k^*} \widetilde{\lambda}_i^2 + 2\sigma_x \max\left\{\frac{n\widetilde{\lambda}_{k^*+1}^2}{c}, \sqrt{n \sum_{i>k^*} \widetilde{\lambda}_i^4 / c}\right\} \\
&= n\widetilde{p}\gamma^2 + 2\sigma_x \max\left\{\frac{n\gamma^2}{c}, \gamma^2\sqrt{\frac{n\widetilde{p}}{c}}\right\} \leq 2n\widetilde{p}\gamma^2,
\end{aligned}$$

where the second inequality is from Condition 1 and the last inequality is from Condition 2. It implies that with probability at least $1-4e^{-n/c}$, we have

$$\widetilde{\sigma}^2\sum_{i>k^*}\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{z}_i\le\widetilde{\sigma}^2 2c_1^2\frac{n\widetilde{p}\gamma^2}{\widetilde{p}^2\gamma^2+n^2\lambda^2}\le 2c_1^2\widetilde{\sigma}^2\min\left\{\frac{n}{\widetilde{p}},\frac{\widetilde{p}\gamma^2}{n\lambda^2}\right\}. \tag{47}$$

Combing the results in Equation (44), Equation (46) and Equation (47), with probability at least $1-10e^{-n/2c}$, we have

$$\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\le\widetilde{\sigma}^2\left(c_1^4c_2c_3^2k^*\min\left\{\frac{1}{n},\frac{1}{\lambda^2}\right\}+2c_1^2\min\left\{\frac{n}{\widetilde{p}},\frac{\widetilde{p}\gamma^2}{n\lambda^2}\right\}\right). \tag{48}$$

Then for the lower bound, on each index $i=1,\ldots,\infty$, we have

$$\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-2}\widetilde{z}_i\ge\frac{1}{\|\widetilde{z}_i\|_2^2}\left(\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i\right)^2,$$

which implies that with probability at least $1-5e^{-n/c}$,

$$\begin{aligned}\frac{\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-2}\widetilde{z}_i}{[1+\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-2}\widetilde{z}_i]^2}&\ge\frac{1}{\|\widetilde{z}_i\|_2^2}\left(\frac{\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i}{1+\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-1}\widetilde{z}_i}\right)^2\\&\ge\frac{1}{c_2n}\left(\frac{n\widetilde{\lambda}_i}{c_1c_3(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*})+n\widetilde{\lambda}_i}\right)^2>0,\end{aligned} \tag{49}$$

where the last inequality is from Equation (31) and Equation (45). Then for the whole term, due to Lemma D.5, with probability at least $1-10e^{-n/c}$, we have

$$\begin{aligned}\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}&=\widetilde{\sigma}^2\sum_i\frac{\widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-2}\widetilde{z}_i}{[1+\widetilde{\lambda}_i\widetilde{z}_i^T(\widetilde{A}_{-i}+n\lambda I)^{-2}\widetilde{z}_i]^2}\\&\ge\frac{\widetilde{\sigma}^2}{2c_2n}\sum_i\left(\frac{n\widetilde{\lambda}_i}{c_1c_3(\widetilde{\lambda}_{k^*+1}\widetilde{r}_{k^*}+n\lambda)+n\widetilde{\lambda}_i}\right)^2\\&\ge\frac{\widetilde{\sigma}^2}{18c_1^2c_2c_3^2n}\sum_i\min\left\{1,\frac{\widetilde{\lambda}_i^2}{\lambda^2},\frac{n^2\widetilde{\lambda}_i^2}{\widetilde{\lambda}_{k^*+1}^2\widetilde{r}_{k^*}^2}\right\}\\&=\frac{\widetilde{\sigma}^2}{18c_1^2c_2c_3^2n}\left(k^*\min\{1,\frac{1}{\lambda^2},\frac{n^2}{\widetilde{p}^2\gamma^2}\}+(\widetilde{p}-k^*)\min\{1,\frac{\gamma^2}{\lambda^2},\frac{n^2}{\widetilde{p}^2}\}\right)\\&=\frac{\widetilde{\sigma}^2}{18c_1^2c_2c_3^2n}\left(k^*\min\{1,\frac{1}{\lambda^2}\}+(\widetilde{p}-k^*)\min\{\frac{\gamma^2}{\lambda^2},\frac{n^2}{\widetilde{p}^2}\}\right),\end{aligned} \tag{50}$$

where the first inequality is from Equation (49), the second inequality is from the fact that

$$(a+b+c)^{-2}\ge(3\max\{a,b,c\})^{-2}=\frac{1}{9}\min\{a^{-2},b^{-2},c^{-2}\},\quad\forall a,b,c>0,$$

and the last two equality is from Condition 1 and Condition 2. Combing both Equation (48) and Equation (50), we could also consider the following two cases.

While $\lambda\le\widetilde{p}\gamma/n$, with a high probability, we have

$$\begin{aligned}\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}&\le\widetilde{\sigma}^2\left(\frac{c_1^4c_2c_3^2k^*}{n}+2c_1^2\frac{n}{\widetilde{p}}\right),\\\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T+n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}&\ge\widetilde{\sigma}^2\frac{1}{18c_1^2c_2c_3^2}\left(\frac{k^*}{n}+\frac{n(\widetilde{p}-k^*)}{\widetilde{p}^2}\right).\end{aligned} \tag{51}$$

And while $\lambda \geq \widetilde{p}\gamma/n$, we have

$$\begin{aligned}
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \leq \widetilde{\sigma}^2\left(c_1^4 c_2 c_3^2 k^* \min\{\frac{1}{n}, \frac{1}{\lambda^2}\} + 2c_1^2\frac{\widetilde{p}^2\gamma^2}{n\lambda^2}\right),\\
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \geq \widetilde{\sigma}^2\frac{1}{18c_1^2c_2c_3^2}\left(\frac{k^*}{n}\min\{1, \frac{1}{\lambda^2}\} + \frac{\gamma^2(\widetilde{p}-k^*)}{n\lambda^2}\right).
\end{aligned} \tag{52}$$

Then we turn to the term $\widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\}$. Similarly, using Woodbury identity, we have

$$\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} = \widetilde{\sigma}^2\sum_i \widetilde{\lambda}_i\lambda_i\widetilde{z}_i^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{z}_i,$$

as well as

$$\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} = \widetilde{\sigma}^2\sum_i \widetilde{\lambda}_i^2\widetilde{z}_i^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{z}_i,$$

comparing such two terms and Condition 1, we have

$$\widetilde{\lambda}_i\lambda_i = \widetilde{\lambda}_i^2, \quad \forall i = 1, \ldots, p,$$

which implies that

$$\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} = \widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}. \tag{53}$$

The result above implies that while $\lambda \leq \widetilde{p}\gamma/n$, with a high probability, we have

$$\begin{aligned}
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \leq \widetilde{\sigma}^2\left(\frac{c_1^4c_2c_3^2k^*}{n} + 2c_1^2\frac{n}{\widetilde{p}}\right),\\
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \geq \widetilde{\sigma}^2\frac{1}{18c_1^2c_2c_3^2}\left(\frac{k^*}{n} + \frac{n(\widetilde{p}-k^*)}{\widetilde{p}^2}\right).
\end{aligned} \tag{54}$$

And while $\lambda \geq \widetilde{p}\gamma/n$, we have

$$\begin{aligned}
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \leq \widetilde{\sigma}^2\left(c_1^4 c_2 c_3^2 k^* \min\{\frac{1}{n}, \frac{1}{\lambda^2}\} + 2c_1^2\frac{\widetilde{p}^2\gamma^2}{n\lambda^2}\right),\\
&\widetilde{\sigma}^2 \mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \geq \widetilde{\sigma}^2\frac{1}{18c_1^2c_2c_3^2}\left(\frac{k^*}{n}\min\{1, \frac{1}{\lambda^2}\} + \frac{\gamma^2(\widetilde{p}-k^*)}{n\lambda^2}\right).
\end{aligned} \tag{55}$$

C.4. Analysis on $\mathcal{L}_{\mathrm{ft}}$

From Condition 2, we have the following results:

$$\begin{aligned}
&O\left(\max\{\gamma, \gamma\} + n^{-(1-\xi)/2}\right) \ll O(\zeta_2\widetilde{p}\gamma), \quad O\left(\zeta_1(1 + n\widetilde{p}\gamma/p)\right) \ll O(\zeta_2\widetilde{p}\gamma),\\
&O\left(\sigma^2(n^{-1}, n\widetilde{p}\gamma/(p^2\gamma))\right) \ll O(\zeta_2\widetilde{p}\gamma),\\
&O\left(\zeta_2(1 + n\gamma)\right) \asymp O(\zeta_2\widetilde{p}\gamma), \quad O\left(\widetilde{\sigma}^2(n^{-1} + n/\widetilde{p})\right) \asymp O(\zeta_2\widetilde{p}\gamma).
\end{aligned}$$

Then comparing Equation (14), Equation (23), Equation (27), Equation (38), Equation (39), Equation (51) and Equation (52), we could obtain that the excess risks on different estimators always dominated by terms related to ζ_2 and $\widetilde{\sigma}^2$. So for the ridge regression, we have

$$\mathcal{L}_{\mathrm{ft}}(\hat{\theta}_\lambda) \approx \zeta_2\mathrm{tr}\{(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} + \widetilde{\sigma}^2\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} := f(\lambda),$$

then take derivative with respect to λ, we could further obtain that

$$f'(\lambda) = 2n(\zeta_2 n\lambda - \widetilde{\sigma}^2)\mathrm{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-3}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},$$

which implies that the optimal choice of λ nearly equals to

$$\lambda' = \frac{\widetilde{\sigma}^2}{n\zeta_2} = O(n^{-1}).$$

As $\hat{\theta}_\lambda = \hat{\theta}_2$ if $\lambda = 0$ and the optinal value $\lambda' > 0$, we have

$$\mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda) \leq \mathcal{L}_{\text{ft}}(\hat{\theta}_2), \quad \forall 0 \leq \lambda \leq 2\lambda'. \tag{56}$$

And according to Lemma D.1, considering Condition 2, with probability at least $1 - 2e^{-n/c}$, we have

$$\mu_n(\widetilde{X}\widetilde{X}^T) \geq \frac{1}{c_1}(\widetilde{p} - k^*)\gamma > \frac{\widetilde{\sigma}^2}{\zeta_2},$$

which implies that

$$\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} - \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} = -\text{tr}\{(\widetilde{X}\widetilde{X}^T)^{-2}(\zeta_2\widetilde{X}\widetilde{X}^T - \widetilde{\sigma}^2 I)\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} < 0,$$

so we could obtain that

$$\mathcal{L}_{\text{ft}}(\hat{\theta}_2) < \mathcal{L}_{\text{ft}}(\hat{\theta}_1). \tag{57}$$

Combing the results in Equation (56) and Equation (57), we could finish the proof of the first item in Theorem 5.1.

Finally, while choosing $0 \leq \lambda < \lambda'$, according to Equation (38), Equation (39), Equation (51) and Equation (52), we have

$$\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \asymp \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \asymp \zeta_2 \text{tr}\{\widetilde{\Sigma}\}.$$

And for the excess risk of ensemble estimator, we have

$$\mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda^\tau) \approx \zeta_2 \text{tr}\{(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} + \tau^2\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} := g(\tau),$$

then take derivative with respect to τ, we could obtain that

$$\begin{aligned} g'(\tau) = {} & 2\tau\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} + 2\tau\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\ & - 2\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}, \end{aligned}$$

which implies the optimal choice of τ is as

$$\tau'(\lambda) = \frac{\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}}{\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} + \widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\}}.$$

While $0 \leq \lambda < \lambda' = \widetilde{\sigma}^2/(n\zeta_2)$, we can obtain that $0 \leq \tau'(\lambda) \leq 1$, due to the fact that

$$\begin{aligned} & \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} + \widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\ & - \zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\ = {} & \text{tr}\left\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\left(\widetilde{\sigma}^2 I + \zeta_2(\widetilde{X}\widetilde{X}^T + n\lambda I)\widetilde{X}\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1} - \zeta_2(\widetilde{X}\widetilde{X}^T + n\lambda I)\right)\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\right\} \\ = {} & \text{tr}\left\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\left(\widetilde{\sigma}^2 I + \zeta_2\widetilde{X}\widetilde{X}^T - \zeta_2(\widetilde{X}\widetilde{X}^T + n\lambda I)\right)\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\right\} \\ = {} & (\widetilde{\sigma}^2 - \zeta_2 n\lambda)\text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \geq 0. \end{aligned}$$

So we could draw the conclusion that for any $\tau'(\lambda) \leq \tau < 1$, we have

$$\mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda^\tau) < \mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda), \tag{58}$$

which finishes the proof of the third item in Theorem 5.1.

C.5. Analysis on $\mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{ft}}$

Similar to the analysis on $\mathcal{L}_{\text{ft}}$, based on Condition 2, we could compare Equation (14), Equation (13), Equation (23), Equation (15), Equation (27), Equation (28), Equation (38), Equation (39), Equation (43), Equation (51), Equation (52), Equation (54) and Equation (55), and obtain that

$$\begin{aligned}
\mathcal{L}_{\text{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda) &\approx \zeta_2 \text{tr}\{(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} + \widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&\quad + \zeta_2 \text{tr}\{[\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]^2\Sigma\} + \widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\Sigma\widetilde{X}^T\} \\
&= \zeta_2 \text{tr}\{(I - \widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} + 2\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&\quad + \zeta_2 \text{tr}\{[\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}]^2\widetilde{\Sigma}\} \\
&:= h(\lambda),
\end{aligned}$$

where the second equality is from Equation (42) and Equation (53). Taking derivative with respect to λ, we have

$$h'(\lambda) = 2n\text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-3}[(\zeta_2 n\lambda - 2\widetilde{\sigma}^2)I - \zeta_2\widetilde{X}\widetilde{X}^T]\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},$$

which implies that while $\lambda \leq 2\lambda' = 2\widetilde{\sigma}^2/(n\zeta_2)$, we could always obtain

$$h'(\lambda) < 0.$$

So for any $0 < \lambda \leq 2\lambda'$, we have

$$\mathcal{L}_{\text{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda) < \mathcal{L}_{\text{pre}}(\hat{\theta}_2) + \mathcal{L}_{\text{ft}}(\hat{\theta}_2). \tag{59}$$

And with this range of λ, we could obtain the similar result as:

$$\begin{aligned}
\mathcal{L}_{\text{pre}}(\hat{\theta}^\tau_\lambda) + \mathcal{L}_{\text{ft}}(\hat{\theta}^\tau_\lambda) &\approx \zeta_2 \text{tr}\{(I - \tau\widetilde{X}^T(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X})^2\widetilde{\Sigma}\} + 2\tau^2\widetilde{\sigma}^2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&\quad + \tau^2\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&:= J(\tau),
\end{aligned}$$

and taking derivative with respect to τ, we have

$$\begin{aligned}
J'(\tau) &= 4\tau\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} + 4\tau\widetilde{\sigma}^3 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&\quad - 2\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\},
\end{aligned}$$

which implies that the optimal choice of τ is $\tau'(\lambda)/2$. And while $\lambda \leq 2\lambda'$, we have

$$\begin{aligned}
&4\tau\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{X}^T\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} + 4\tau\widetilde{\sigma}^3 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&- 2\zeta_2 \text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-1}\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} \\
&= 2\text{tr}\{(\widetilde{X}\widetilde{X}^T + n\lambda I)^{-2}[\zeta_2\widetilde{X}\widetilde{X}^T + (2\widetilde{\sigma}^2 - \zeta_2 n\lambda)I]\widetilde{X}\widetilde{\Sigma}\widetilde{X}^T\} > 0,
\end{aligned}$$

which implies that $0 \leq \tau'(\lambda)/2 \leq 1$. So we could further obtain that for any $\tau'(\lambda)/2 \leq \tau < 1$, we have

$$\mathcal{L}_{\text{pre}}(\hat{\theta}^\tau_\lambda) + \mathcal{L}_{\text{ft}}(\hat{\theta}^\tau_\lambda) < \mathcal{L}_{\text{pre}}(\hat{\theta}_\lambda) + \mathcal{L}_{\text{ft}}(\hat{\theta}_\lambda). \tag{60}$$

Combing the results in Equation (59) and Equation (60), we could finish the proof of the second item in Theorem 5.1.

D. Auxiliary Lemmas

Lemma D.1 (Lemma 10 in Bartlett et al., 2019)**.** *There are constants $b, c \geq 1$ such that, for any $k \geq 0$, with probability at least $1 - 2e^{-\frac{n}{c}}$,*

1. *for all $i \geq 1$,*

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c_1(\sum_{j>k}\lambda_j + \lambda_{k+1}n);$$

2. *for all* $1 \le i \le k$,

$$\mu_n(A) \ge \mu_n(A_{-i}) \ge \mu_n(A_k) \ge \frac{1}{c_1}\sum_{j>k}\lambda_j - c_1\lambda_{k+1}n;$$

3. *if* $r_k \ge bn$, *then*

$$\frac{1}{c_1}\lambda_{k+1}r_k \le \mu_n(A_k) \le \mu_1(A_k) \le c_1\lambda_{k+1}r_k,$$

where $c_1 > 1$ *is a constant only depending on* σ_x.

Lemma D.2 (Corollary 24 in Bartlett et al., 2019)**.** *For any centered random vector* $\boldsymbol{z} \in \mathbb{R}^n$ *with independent* σ_x^2 *sub-Gaussian coordinates with unit variances, any* k *dimensional random subspace* $\mathscr{L}$ *of* $\mathbb{R}^n$ *that is independent of* $\boldsymbol{z}$*, and any* $t > 0$*, with probability at least* $1 - 3e^{-t}$,

$$\| z \|^2 \le n + 2(162e)^2\sigma_x^2(t + \sqrt{nt}),$$
$$\| \Pi_{\mathscr{L}} z \|^2 \ge n - 2(162e)^2\sigma_x^2(k + t + \sqrt{nt}),$$

where $\Pi_{\mathscr{L}}$ *is the orthogonal projection on* $\mathscr{L}$.

Lemma D.3. *There are constants* $b, c \ge 1$ *such that, for any* $k \ge 0$*, with probability at least* $1 - 2e^{-\frac{n}{c}}$*:*

1. *for all* $i \ge 1$,

$$\mu_{k+1}(A_{-i} + \lambda I) \le \mu_{k+1}(A + \lambda I) \le \mu_1(A_k + \lambda I) \le c_1(\sum_{j>k}\lambda_j + \lambda_{k+1}n) + \lambda;$$

2. *for all* $1 \le i \le k$,

$$\mu_n(A + \lambda I) \ge \mu_n(A_{-i} + \lambda I) \ge \mu_n(A_k + \lambda I) \ge \frac{1}{c_1}\sum_{j>k}\lambda_j - c_1\lambda_{k+1}n + \lambda;$$

3. *if* $r_k \ge bn$, *then*

$$\frac{1}{c_1}\lambda_{k+1}r_k + \lambda \le \mu_n(A_k + \lambda I) \le \mu_1(A_k + \lambda I) \le c_1\lambda_{k+1}r_k + \lambda.$$

Proof. With Lemma D.1, the first two claims follow immediately. For the third claim: if $r_k \ge bn$, we have that $bn\lambda_{k+1} \le \sum_{j>k}\lambda_j$, so

$$\mu_1(A_k + \lambda I) \le c_1\lambda_{k+1}r_k(\Sigma) + \lambda \le \lambda + c_1\lambda_{k+1}r_k,$$
$$\mu_n(A_k + \lambda I) \ge \frac{1}{c_1}\lambda_{k+1}r_k + \lambda \ge \frac{1}{c_1}\lambda_{k+1}r_k(\Sigma) + \lambda,$$

for the same constant $c_1 > 1$ as in Lemma D.1. □

Lemma D.4 (Proposition 2.7.1 in Vershynin, 2018)**.** *For any random variable* ξ *that is centered,* σ^2*-subgaussian, and unit variance,* $\xi^2 - 1$ *is a centered* $162e\sigma^2$*-subexponential random variable, that is,*

$$\mathbb{E}\exp(\lambda(\xi^2 - 1)) \le \exp((162e\lambda\sigma^2)^2),$$

for all such λ *that* $|\lambda| \le 1/(162e\sigma^2)$.

Lemma D.5 (Lemma 15 in Bartlett et al., 2019)**.** *Suppose that* $\{\eta_i\}$ *is a sequence of non-negative random variables, and that* $\{t_i\}$ *is a sequence of non-negative real numbers (at least one of which is strictly positive) such that, for some* $\delta \in (0, 1)$ *and any* $i \ge 1$, $\Pr(\eta_i > t_i) \ge 1 - \delta$*. Then,*

$$\Pr\left(\sum_i \eta_i \ge \frac{1}{2}\sum_i t_i\right) \ge 1 - 2\delta.$$

Lemma D.6 (Lemma 2.7.6 in Vershynin, 2018)**.** *For any non-increasing sequence $\{\lambda_i\}_{i=1}^{\infty}$ of non-negative numbers such that $\sum_i \lambda_i < \infty$, and any independent, centered, $\sigma-$subexponential random variables $\{\xi_i\}_{i=1}^{\infty}$, and any $x > 0$, with probability at least $1 - 2e^{-x}$*

$$|\sum_i \lambda_i \xi_i| \leq 2\sigma \max\left(x\lambda_1, \sqrt{x\sum_i \lambda_i^2}\right).$$

Lemma D.7 (Theorem 9 in Koltchinskii & Lounici (2017))**.** *Let $z_1, \ldots, z_n$ be i.i.d. sub-gaussian random variables with zero mean, then with probability at least $1 - 2e^{-t}$,*

$$\|\mathbb{E}zz^T - \frac{1}{n}\sum_{i=1}^{n} z_i z_i^T\|_2 \leq \|\mathbb{E}zz^T\|_2 \max\{\sqrt{\frac{trace(\mathbb{E}zz^T)}{n}}, \frac{trace(\mathbb{E}zz^T)}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\}.$$