S^2FT: Efficient, Scalable and Generalizable LLM Fine-tuning by Structured Sparsity

Xinyu Yang[1], **Jixuan Leng**[1], **Geyang Guo**[2], **Jiawei Zhao**[3], **Ryumei Nakada**[4], **Linjun Zhang**[4], **Huaxiu Yao**[5], **Beidi Chen**[1]

[1]CMU, [2]Georgia Tech, [3]Caltech, [4]Rutgers, [5]UNC-Chapel Hill

`xinyuya2, beidic@andrew.cmu.edu`

https://infini-ai-lab.github.io/S2FT

Abstract

Current PEFT methods for LLMs can achieve either high quality, efficient training, or scalable serving, but not all three simultaneously. To address this limitation, we investigate sparse fine-tuning and observe a remarkable improvement in generalization ability. Utilizing this key insight, we propose a family of Structured Sparse Fine-Tuning (**S^2FT**) methods for LLMs, which *concurrently achieve state-of-the-art fine-tuning performance, training efficiency, and inference scalability*. S^2FT accomplishes this by "selecting sparsely and computing densely". It selects a few heads and channels in the MHA and FFN modules for each Transformer Block, respectively. Next, it co-permutes weight matrices on both sides of the coupled structures in LLMs to connect the selected components in each layer into a dense submatrix. Finally, S^2FT performs in-place gradient updates on all submatrices. Through theoretical analysis and empirical results, our method prevents overfitting and forgetting, delivers SOTA performance on both commonsense and arithmetic reasoning with 4.6% and 1.3% average improvements compared to LoRA, and outperforms full FT by 11.5% when generalize to various domains after instruction tuning. By integrating our partial back-propagation algorithm, S^2FT saves the fine-tuning memory up to $3\times$ and improves the latency by 1.5-2.7$\times$ compared to full FT, while delivering an average 10% improvement over LoRA on both metrics. We further demonstrate that S^2FT can be decoupled into adapters, enabling effective fusion, fast switch, and efficient parallelism for serving multiple fine-tuned models.

1 Introduction

Recently, Large Language Models (LLMs) have achieved significant success [16, 1, 64]. With these models being applied in diverse domains, full fine-tuning (FT) is commonly employed to enhance their downstream capabilities [54, 6, 71]. However, retraining all parameters comes with three drawbacks: (i) Full FT suffers from catastrophic forgetting, where a model forgets pre-trained knowledge while acquiring new information [44, 8]. (ii) As the model and dataset sizes grow at scale, full FT becomes increasingly computation-demanding and memory-intensive [68]. (iii) It is impractical to store and serve thousands of fine-tuned LLMs on modern GPUs if each requires full parameter storage [77, 58].

Parameter-efficient fine-tuning (PEFT) methods propose to address these bottlenecks by updating a small fraction of parameters [21]. Rather than merely reducing the number of learnable parameters, an ideal PEFT method should possess three key properties to be practically effective and efficient:

High Quality: It should exhibit both memorization and generalization capabilities, balancing the acquisition of new information from fine-tuning tasks with the retention of pre-trained knowledge.

Efficient Training: It should minimize the memory footprint for model gradient and optimization states, and further translate such memory efficiency into less computation and fine-tuning speedup.

Scalable Serving: It should avoid adding inference overhead when serving a single PEFT model. For multiple models, new parameters should be partially stored as adapters to save memory, and allows for effective fusion [75], fast switch [33], and efficient parallelism [58] among thousands of adapters.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

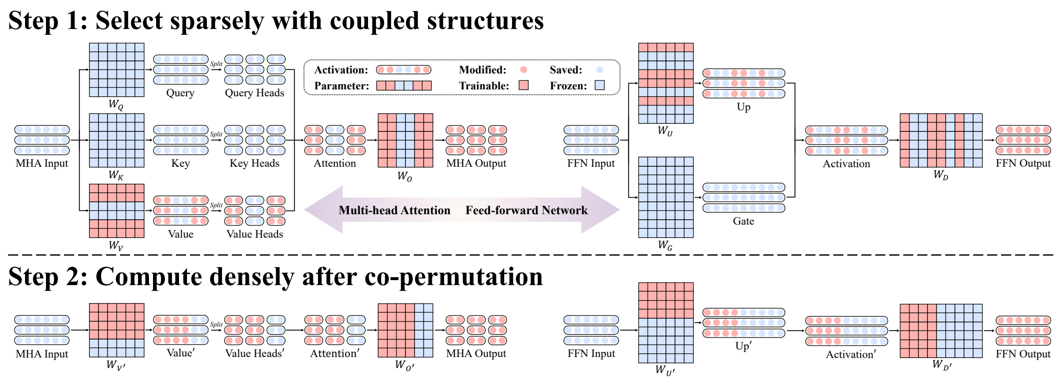


Figure 1: **An Overview of the S^{2}FT Family for LLMs**: First, we perform sparse selection of specific attention heads and channels within the coupled structures of the MHA and FFN modules. Next, we apply co-permutation to the weight matrices on both sides of these structures, enabling dense gradient computation only for the selected components. While we demonstrate S^{2}FT by selecting the same heads/channels on both sides for clarity, our approach also supports asymmetric selection strategies.

However, achieving all the aforementioned goals simultaneously is challenging. Common PEFT approaches, such as LoRA [27], DoRA [38], and Galore [76], project the model's weights or gradients onto a low-rank subspace. While this significantly reduces memory footprint, their performance lags behind full fine-tuning in most large-scale scenarios. Recent state-of-the-art PEFT methods have aimed to improve performance but at the cost of serving efficiency. ReFT operates on a frozen base model and learns task-specific interventions on hidden representations that cannot be merged into the original model, leading to a $2.2\times$ increase in inference latency. LISA [48] employs a coarse-grained selective method by randomly freezing most Transformer blocks during optimization, which requires significantly more trainable parameters. Consequently, in scaled serving settings like S-LoRA [58], LISA can only serve at most $\frac{1}{10}$ as many fine-tuned models as LoRA under the same memory budget.

Prior to the era of LLMs, PEFT methods based on unstructured sparse fine-tuning (SpFT) have shown a strong trade-off between low number of parameters and high model performance without sacrificing serving efficiency [61, 3, 69]. We hypothesize that SpFT, which selectively updates a small subset of model parameters, can outperform LoRA and its variants in generalization capabilities. In Figure 2, our findings across various generalization tasks support this hypothesis. However, the unstructured nature of SpFT necessitates sparse operations in computation, hindering its efficient training and scalable serving on modern hardware. This makes SpFT less practical for adapting LLMs at scale.

In this work, we propose a family of Structured Sparse Fine-Tuning (**S^{2}FT**) methods to "select sparsely and compute densely" (See Figure 1), thereby closing the efficiency gap in SpFT. Inspired by structured weight pruning techniques [45, 42], we first identify several coupled structures inherent in LLMs that are connected by intermediate activations. For example, in the multi-head attention (MHA) module, each attention head in the query, key, and value projections is linked to only a few rows in the output projection. Similarly, in the feed-forward network (FFN) module, each column in the up and gate projections corresponds to a single row in the down projection. By co-permuting the matrices on both sides of these coupled structures, we can preserve the original output of these structures, with only the order of the intermediate activations changed. Exploiting this property, our S^{2}FT strategically selects a subset of attention heads for the MHA module and a subset of channels for the FFN module. We then permute the coupled structures to connect the selected components within each linear layer into a dense submatrix. Finally, through our partial back-propagation algorithm with only two-line code modification, S^{2}FT performs in-place gradient updates exclusively for all selected submatrices, boosting training efficiency by eliminating redundant forward activations and backward calculation.

Through our theoretical analysis, we demonstrate that S^{2}FT mitigates overfitting and forgetting under distribution shifts. Empirically, S^{2}FT outperforms other PEFT methods on LLaMA and Mistral family models, improving 1.2-4.1% on commonsense reasoning tasks and 0.6-1.9% on arithmetic reasoning ones. It also surpasses full FT by 11.5% when generalize to various domains after instruction tuning.

Finally, we conduct a comprehensive analysis to verify the training efficiency and serving scalability of S^{2}FT. Compared to existing PEFT methods, S^{2}FT not only saves 1.4-3.0$\times$ memory, but also increases latency by 1.5 to 2.7$\times$, making LLM fine-tuning more accessible. Additionally, S^{2}FT's parameter updates can be decomposed into adapters, enabling adapter fusion with smaller performance drop than LoRA. Our method also results in more scalable and efficient adapter switch and parallelism through reduced matrix multiplications, showcasing strong potential for large-scale LLM serving scenarios.

2 Memorization or Generalization?

In this section, we evaluate the memorization and generalization capabilities of various fine-tuning methods, including full FT, LoRA, and SpFT. We hypothesize that SpFT can generalize better to downstream tasks. To support this hypothesis, we present detailed observations and analyses. Further theoretical analysis about the generalization capabilities of the S^2FT family can be found in Section 4.

Hypothesis. We hypothesize that SpFT offers superior generalization than both full FT and LoRA, while maintaining comparable memorization to LoRA with the same number of trainable parameters.

Experimental Setup. We fine-tune the `Llama3-8B` on the Math10K data [28] using SpFT, LoRA, and full FT. In addition to training losses, accuracies are measured on downstream tasks in LLM-Adapters, including near out-of-distribution (OOD) generalization on both easy (i.e, MultiArith, AddSub, SingleEq, MAWPS) and hard (i.e, GSM8K, AQuA, SVAMP) arithmetic reasoning tasks, and far OOD generalization on commonsense reasoning ones. For PEFT methods, we set three ratios of trainable parameters ($p = 10\%, 1\%, 0.1\%$) and search for the optimal hyperparameters on the valid set. In SpFT, trainable parameters are selected randomly with given ratios. See details in Appendix C.

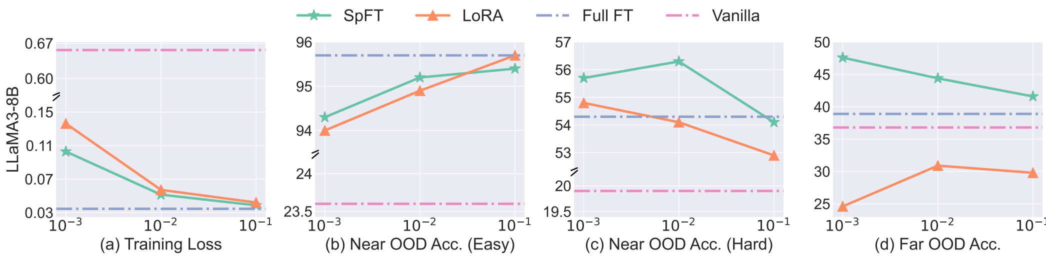


Figure 2: Accuracy comparison of SpFT, LoRA and Full FT at varying ratios of trainable parameters in various settings. SpFT exhibits strong generalization ability while full FT excels in memorization.

Observations. Figure 2 indicates several key findings. First, SpFT achieves lower training losses than LoRA when using the same ratio of trainable parameters, especially at very small ratios. This gap arises from the more complex optimization process in LoRA, which requires the simultaneous updating of two matrices [23]. Second, we observe both elevated training loss and reduced average accuracy on easier math tasks as the ratio decreases, suggesting a positive correlation between memorization abilities and trainable parameters. Notably, with only 10% of the parameters updated, PEFT methods learn comparable memorization abilities to full FT when trained on a 10k-sample dataset.

When generalizing to complex mathematical problems or commonsense reasoning tasks, the performance ranking emerges as: SpFT $>$ Full FT $>$ LoRA. SpFT effectively transfers reasoning abilities to commonsense domains, while LoRA exhibits significant performance drops in far OOD generalization. This indicates (i) freezing a larger fraction of the parameters can retain more pre-trained abilities, and (ii) approximating high-dimensional gradients with low-rank decomposition may overfit fine-tuned data and hinder the model from generalization. Since LLMs are pre-trained on high-quality data, SpFT emerges as the preferred choice for fine-tuning on task-specific data of varying quality.

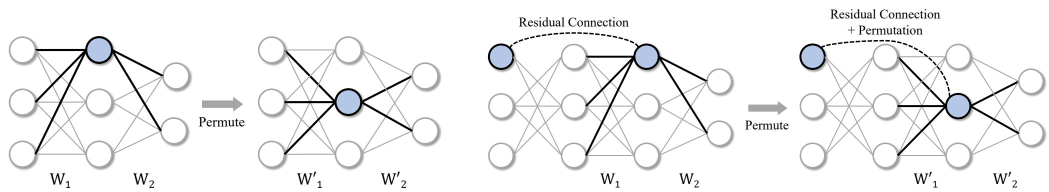
3 The S^2FT family of methods

While SpFT demonstrates strong generalization ability and good overall performance in Section 2, its unstructured nature poses challenges for efficient training and scalable serving on modern hardware (e.g., GPU). This is because of the need for sparse operations when storing and computing weights, gradients, and optimization states, which are significantly slower than their dense variants on GPU. This motivates our investigation into structured sparsity approaches that utilize only dense operations: *Can structured sparsity improve hardware efficiency while preserving performance by selecting sparsely but computing densely? If so, how far can the flexibility of selection be pushed in this context?*

To answer this question, we design a family of Structured Sparse Fine-Tuning (**S^2FT**) methods with dense-only computations, making PEFT effective, efficient and scalable. We begin by discovering the coupled structure in LLMs in Section 3.1. Leveraging this property, Section 3.2 introduce the selection and permutation strategies of S^2FT, with overall pipeline illustrated in Figure 1b. In Section 3.3, we present our partial back-propagation algorithm that enables end-to-end training latency reduction.

3.1 Discover Coupled Structures in LLMs

We initiate our pursuit of flexible structured sparsity by examining the coupled structures in LLMs.



(a) Basic Structure (b) Residual Structure

Figure 3: Grouped model weights with basic structure and residual structure. All highlighted weights must be permuted simultaneously. Residual structures require additional permutation during runtime.

Structure Dependency in LLMs. Inspired by prior work on structured pruning [45, 17], our study start by building the dependencies between activations and weights for LLMs. Let A denote an activation and W denote a weight in the model. We define $\mathrm{In}(A)$ as the set of parameters that directly contribute to the computation of A, and $\mathrm{Out}(A)$ as the set of parameters that depend on A in the computation of subsequent activations. The dependency between structures can be defined as follows:

$$W_1 \in \mathrm{In}(A) \wedge \mathrm{Deg}^+(W_1) = 1 \Rightarrow A \text{ is dependent on } W_1 \tag{1}$$

$$W_2 \in \mathrm{Out}(A) \wedge \mathrm{Deg}^-(W_2) = 1 \Rightarrow W_2 \text{ is dependent on } A \tag{2}$$

where $\mathrm{Deg}^+(W_1)$ represents the out-degree of weight W_1, and $\mathrm{Deg}^-(W_2)$ represents the in-degree of weight W_2. Each equation represents a unqiue directional dependency between activations and weights. When both equations hold simultaneously, a coupled structure exists between W_1 and W_2. In Figure 3, we employ deep linear networks to illustrate two types of coupled structures in LLMs:

Basic Structures: In Figure 3a, these structures exist in both the multi-head attention (MHA) and feed-forward network (FFN) modules. Taking LLaMA as an example, in the MHA module, we consider the Query ($\boldsymbol{Q}$), Key ($\boldsymbol{K}$), and Value ($\boldsymbol{V}$) projections as W_1, and the Output ($\boldsymbol{O}$) projection as W_2, while $\texttt{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top)\boldsymbol{V}(x)$ acting as the activation between weight matrices. Similarly, in the FFN module, the Up ($\boldsymbol{U}$) and Gate ($\boldsymbol{G}$) projections function as W_1, with the Down ($\boldsymbol{D}$) projection corresponding to W_2. Here, $\boldsymbol{U}(x) \cdot \texttt{SwiGLU}(\boldsymbol{G}(x))$ serves as the activations connecting W_1 and W_2.

Residual Structures: In Figure 3b, this type of coupled structures exists between the MHA and FFN modules. We further consider how residual connections influence the activations in these structures.

Permutation Invariance of Coupled Structures. Figure 3 demonstrates that W_1 and W_2 can be co-permuted using the same order, which only affects the order of activations between them while preserving the original output from the coupled structure. Since residual dependencies require an additional runtime step to permute the residuals, we will focus on basic dependencies in our method.

3.2 Sparse Selection and Permutation

At this point, all coupled structures within the model have been identified. The subsequent sparse selection and permutation processes are straightforward, with overall pipeline illustrated in Figure 1b.

MHA Module: There are four linear layers in a MHA module: $Q, K, V, O \in \mathbb{R}^{d\times d}$. For a model with h attention heads, each head $i \in [h]$ has its own projections denoted as $Q_i \in \mathbb{R}^{d\times d_h}$, $K_i \in \mathbb{R}^{d\times d_h}$, $V_i \in \mathbb{R}^{d\times d_h}$, and $O_i \in \mathbb{R}^{d_h\times d}$, where $d_h = d/h$ is the dimension per head. Let $S_{\text{MHA}} \subseteq [h]$ denotes a small subset of attention heads. By permuting S_{MHA} to the beginning of each weight matrix, we are able to update these selected heads using dense-only operations, while keeping the other ones frozen.

FFN Module: There are three linear layers in a FFN module: $U, G \in \mathbb{R}^{k\times d}$ and $D \in \mathbb{R}^{d\times k}$. In S^{2}FT, only a few channels of them require gradient updates. Let $S_{\text{FFN}} \subseteq [d]$ denote the selected channels. We can permute S_{FFN} to the beginning of each weight matrix and only fine-tune this compact subset.

Next, we provide several strategies for identifying and selecting important subsets in each module.

1. **S^{2}FT-R (S^{2}FT)**: In this strategy, a subset of channels is randomly selected and set to be trainable.
2. **S^{2}FT-W**: This variant selects subsets based on the magnitude of activations on a calibration set.
3. **S^{2}FT-A**: This variant selects subsets based on the magnitude of activations on a calibration set.
4. **S^{2}FT-S**: Top-K subsets are ranked and selected by the product of weight and activation magnitudes.
5. **S^{2}FT-G**: This variant selects subsets based on the magnitude of gradients on a calibration set.

Here, 1 and 2 can be directly applied without preprocessing. 3 and 4 only require a forward pass on a small calibration dataset. While 5 necessitates a backward pass on this dataset, it does not store optimization states and can mitigate memory footprints for activations through gradient checkpointing [18]. By default, we use S^{2}FT-R for a fair comparison and discuss other variants in Section 5.4.

3.3 Partial Back-propagation Algorithm

Finally, we introduce our partial back-propagation algorithm with only two line modifications in PyTorch. our algorithm stores trainable channels based on their start and end positions, thereby improving training efficiency by eliminating redundant forward activations and backward calculations.

```
def setup_context(ctx, inputs, output):
   activation, weight, bias, start, end = inputs
   # only save partial input tensors for gradient calculation in forward
   ctx.save_for_backward(activation[:, start:end], weight, bias, start, end)

def gradient_update(parameter, gradient, start, end):
    # only modify the assigned positions of weight matrices during optimization
    parameter[:, start:end].add_(gradient)
```

4 Theoretical Analysis

In this section, we theoretically explore why S^{2}FT demonstrates stronger generalization capabilities compared to LoRA. We consider a pre-trained L-layer deep linear networks, which has been widely used to facilitate the theoretical analysis of complex DNNs [57, 30, 43, 22, 34, 5]. Let $f^{\text{pre}}(x) := W_L^{\text{pre}} W_{L-1}^{\text{pre}} \dots W_1^{\text{pre}} x$ be the pre-trained deep linear network, where $W_\ell^{\text{pre}} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, with $d_0 = p$ and $d_L = q$. We fine-tune the ℓ-th layer with low-rankness level $r \le \min\{d_\ell, d_{\ell-1}\}$ or sparsity level $s = \lfloor r \cdot \frac{d_\ell + d_{\ell-1}}{d_{\ell-1}} \rfloor$. Denote a class of adaptation with parameters $U \in \mathbb{R}^{d_\ell \times d}$ and $V \in \mathbb{R}^{d_{\ell-1} \times d}$ as

$$f_{\ell,U,V}(x) := \overline{W}_{\ell+1}^{\text{pre}} (W_\ell^{\text{pre}} + UV^\top) \underline{W}_{\ell-1}^{\text{pre}} x, \tag{3}$$

where $\overline{W}_\ell^{\text{pre}} := W_L^{\text{pre}} W_{L-1}^{\text{pre}} \dots W_\ell^{\text{pre}} \in \mathbb{R}^{d_L \times d_{\ell-1}}$ and $\underline{W}_\ell^{\text{pre}} := W_\ell^{\text{pre}} W_{\ell-1}^{\text{pre}} \dots W_1^{\text{pre}} \in \mathbb{R}^{d_\ell \times d_0}$ with $\underline{W}_0^{\text{pre}} = I_p$ and $\overline{W}_L^{\text{pre}} = I_q$. In a transformer-based LLM, each row of W_ℓ can represents either the parameters in a single head for the MHA module or that in a single channel for the FFN module.

Given n observations $(x_i^{(\text{i})}, y_i^{(\text{i})}) \subset \mathbb{R}^p \times \mathbb{R}^q$, we fine-tune f^{pre} by minimizing the empirical risk $\mathcal{R}_n^{(\text{i})}(f_{\ell,U,V}) := (1/n) \sum_{i \in [n]} \|y_i^{(\text{i})} - f_{\ell,U,V}(x_i^{(\text{i})})\|^2$ via gradient descent. For LoRA, we train both low-rank matrices (U, V) in Equation (3) with $d \leftarrow r$. For S^{2}FT, we train only V in Equation (3) with $d \leftarrow s$ and fixed $U \leftarrow U_S^{\text{S}^2\text{FT}} := [e_{a_1}; e_{a_2}; \dots; e_{a_s}]$, which specifies s rows to fine-tune, where $S = \{a_1, \dots, a_s\} \subset [d_\ell]$ and e_a is the a-th standard basis. Motivated from the results that gradient descent has implicit regularization [74, 19, 5], we directly consider the minimum norm solutions.

We consider a multiple linear regression setting. Assume that the in-distribution training data $(x^{(\text{i})}, y^{(\text{i})}) \in \mathbb{R}^{p+q}$ and out-of-distribution test data $(x^{(\text{o})}, y^{(\text{o})}) \in \mathbb{R}^{p+q}$ are generated i.i.d. according to

$$y^{(k)} = B^{(k)} x^{(k)} + \epsilon^{(k)}, \quad k \in \{\text{i}, \text{o}\},$$

where $B^{(k)} \in \mathbb{R}^{q \times p}$ is the coefficient matrix, $x^{(k)}$ and $\epsilon^{(k)}$ are mean zero sub-Gaussian signal and noise with covariance matrices $\Sigma_x^{(k)}$ and $\Sigma_\epsilon^{(k)}$, respectively. The generalization capacity is measured by the fine-tuned model's excess risk $\mathcal{E}(f) := \mathbb{E}[\|y^{(o)} - f(x^{(o)})\|^2] - \inf_{f'} \mathbb{E}[\|y^{(o)} - f'(x^{(o)})\|^2]$.

For these OOD data, LoRA suffers from forgetting, while S^{2}FT can maintain pre-training knowledge.

Assumption 4.1 (Distribution Shift)**.** Assume that $\Sigma_x^{(\text{i})} = \Sigma_x^{(\text{o})} = \Sigma_x$ for some $\Sigma_x \in \mathbb{R}^{p \times p}$, and $\|(\overline{W}_{\ell+1}^{\text{pre}} U_S^{\text{S}^2\text{FT}})(\overline{W}_{\ell+1}^{\text{pre}} U_S^{\text{S}^2\text{FT}})^\dagger (B^{(\text{o})} - B^{(\text{i})}) \Sigma_x^{1/2}\|_{\text{F}}^2 \le \varepsilon^2 \mathcal{E}^{(\text{o})}(f^{\text{pre}})$ for some $\varepsilon > 0$.

Assumption 4.1 states that while the covariate distribution remains unchanged, the label distribution conditioned on covariates may shift, but not exceeding a factor of ϵ^2 of the OOD risk of f^{pre}. This holds for fine-tuning with proper channel selection, where primarily the output distribution is changed.

Theorem 4.2 (Out-of-distribution Excess Risk, Informal)**.** *Suppose Assumption 4.1 holds. Consider $n \to \infty$. If $B^{(\text{i})} = \overline{W}_{\ell+1}^{\text{pre}} \tilde{B}^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}}$ holds for some $\tilde{B}^{(\text{i})} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, and $s \le \operatorname{rank}(\Sigma_f^{(\text{i})})$, then,*

$$\mathcal{E}^{(\text{o})}(f_{\ell, U_S^{\text{S}^2\text{FT}}, V^{\text{S}^2\text{FT}}}) \le (1 + 3\varepsilon^2) \mathcal{E}^{(\text{o})}(f^{\text{pre}}), \quad \mathcal{E}^{(\text{o})}(f_{\ell, U^{\text{LoRA}}, V^{\text{LoRA}}}) \ge \|(B^{(\text{o})} - B^{(\text{i})}) \Sigma_x^{1/2}\|_{\text{F}}^2.$$

Theorem 4.2 indicates that the OOD risk of S^{2}FT is bounded above by that of f^{pre}, while that of LoRA is bounded below by the label shift magnitude. If f^{pre} already has low risk for OOD tasks, and the label shift is significant, S^{2}FT is expected to outperform LoRA. Essentially, when the OOD task deviates significantly from the FT distribution, LoRA may forget the pre-trained knowledge and overfit to the FT data, compromising its generalization capabilities. See formal statements in Theorem E.8.

5 Experiments

In this section, we conduct a series of experiments across three diverse benchmarks covering more than 20 datasets. Our goal is to provide a rich picture of how S^{2}FT performs in different scenarios. Here, we compare our method with different fine-tuning strategies and categories including: (i) Full fine-tuning (FT), (ii) *reparameterized fine-tuning*: LoRA [27], DoRA [38], and Galore [76], (iii) *adapter-based fine-tuning*: Series Adapter [26], Parallel Adapter [24], and LoReFT [67], (iv) *prompt-based fine-tuning*: Prefix-Tuning [36], (v) *sparse fine-tuning*: LISA [48]. For a fair comparison, we keep a comparable number of trainable parameters in S^{2}FT to that of LoRA. The design choices for trainable parameter allocations in S^{2}FT will be detailed in Section 5.4. All other hyperparameters are selected via cross-validation. Detailed setups and baseline descriptions are provided in Appendix D.

5.1 Commonsense Reasoning

Dataset Descriptions. The commonsense reasoning dataset comprise eight subsets: BoolQ [12], PIQA [9], SocialQA [56], HellaSwag [73], WinoGrande [55], ARC-challenge [13], ARC-easy [13], and OpenbookQA [46]. Following the experimental setup of LLM-Adapters [28], we split each dataset into training and test sets. Subsequently, we combine the training data from all eight tasks into a single fine-tuning dataset and evaluate performance on the individual test dataset for each task.

Results. Table 1 showcases that S^{2}FT consistently outperforms existing PEFT methods across the `LLaMA-7B/13B`, `LLaMA2-7B` and `LLaMA3-8B` models. When compared to LoRA and DoRA, it achieves average performance gains of 4.6% and 2.8%, respectively. Additionally, S^{2}FT also demonstrates superior performance against recent approaches including Galore, LoReFT, and LISA, with accuracy improvements of at least 1.0%. Remarkably, despite using less than 1% of trainable parameters, our method surpasses full FT by 0.5%. The 3.0% improvement observed on the `LLaMA3-8B` suggests that maintaining most pre-trained parameters frozen enables better generalization to test distributions.

Table 1: Comparison among various fine-tuning methods for the `LLaMA-7B/13B`, `LLaMA2-7B`, and `LLaMA3-8B` models on eight commonsense reasoning tasks. Non-PEFT methods are marked in gray. ([1]: from DoRA paper, [2]: from ReFT paper, [3]: reproduced by us, [†]: projected trainable parameters)

Model	Method	# Param(%)	BoolQ	PIQA	SIQA	HellaSwag	Wino	ARC-e	ARC-c	OBQA	Avg. ↑
ChatGPT[1]	-	-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA-7B	Full FT[3]	100	70.3	84.2	80.1	92.3	85.4	86.6	72.8	83.4	81.9
	Prefix [36][1]	0.11	64.3	76.8	73.9	42.1	72.1	72.9	54.0	60.6	64.6
	Series [26][1]	0.99	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
	Parallel [24][1]	3.54	67.9	76.4	78.8	69.8	78.9	73.7	57.3	75.2	72.2
	LoRA [27][3]	0.83	69.2	81.7	78.4	83.4	80.8	79.0	62.4	78.4	76.7
	DoRA [38][1]	0.84	68.5	82.9	79.6	84.8	80.8	81.4	65.8	81.0	78.1
	Galore [76][3]	0.83[†]	68.6	79.0	78.5	84.7	80.1	80.3	62.1	77.3	76.3
	LoReFT [67][2]	0.03	69.3	84.4	80.3	93.1	84.2	83.2	68.2	78.9	80.2
	LISA [48][3]	9.91	70.4	82.1	78.7	92.4	82.9	84.9	70.2	78.4	80.0
	S^{2}FT (Ours)	0.81	**72.7**	**83.7**	**79.6**	**93.4**	**83.5**	**86.1**	**72.2**	**83.4**	**81.8**
LLaMA-13B	Full FT[3]	100	74.5	86.3	81.3	94.4	86.9	89.7	77.9	88.8	85.0
	Prefix [36][1]	0.03	65.3	75.4	72.1	55.2	68.6	79.5	62.9	68.0	68.4
	Series [26][1]	0.80	71.8	83.0	79.2	88.1	82.4	82.5	67.3	81.8	79.5
	Parallel [24][1]	2.89	72.5	84.9	79.8	92.1	84.7	84.2	71.2	82.4	81.4
	LoRA [27][1]	0.67	72.1	83.5	80.5	90.5	83.7	82.8	68.3	82.4	80.5
	DoRA [38][1]	0.68	72.4	84.9	81.5	92.4	84.2	84.2	69.6	82.8	81.5
	LoReFT [67][2]	0.03	72.1	**86.3**	**81.8**	**95.1**	**87.2**	86.2	73.7	84.2	83.3
	S^{2}FT (Ours)	0.65	**74.2**	85.7	80.7	94.9	86.4	**88.4**	**76.3**	**87.8**	**84.3**
LLaMA2-7B	Full FT[3]	100	74.7	84.9	78.7	93.7	84.1	87.5	75.2	85.0	83.0
	LoRA [27][1]	0.83	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA [38][1]	0.84	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	S^{2}FT (Ours)	0.81	**72.9**	**86.1**	**80.2**	**94.3**	**85.5**	**87.2**	**74.6**	**83.4**	**83.0**
LLaMA3-8B	Full FT[3]	100	73.9	86.2	79.1	93.1	85.8	88.1	78.2	84.0	83.6
	LoRA [27][1]	0.70	70.8	85.2	79.7	92.5	84.9	88.9	78.7	84.4	82.5
	DoRA [38][1]	0.71	74.6	**89.3**	79.9	95.5	85.6	90.5	80.4	85.8	85.2
	S^{2}FT (Ours)	0.70	**75.0**	89.0	**80.7**	**96.5**	**88.0**	**92.5**	**83.4**	**87.8**	**86.6**

Table 2: Comparison among various fine-tuning methods for different models on seven math reasoning tasks. Non-PEFT methods are marked in gray. ([1]: from LLM-Adapters paper, [2]: reproduced by us)

Model	Method	# Param(%)	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	MAWPS	Avg. ↑
GPT-3.5[1]	-	-	83.8	56.4	85.3	38.9	88.1	69.9	87.4	72.8
LLaMA-7B	Full FT[2]	100	98.8	43.1	91.1	20.9	94.3	60.6	88.2	71.0
	Prefix [36][1]	0.11	63.2	24.4	57.0	14.2	55.3	38.1	63.4	45.1
	Series [26][1]	0.99	92.8	33.3	80.0	15.0	83.5	52.3	77.7	62.1
	Parallel [24][1]	3.54	94.5	35.3	86.6	18.1	86.0	49.6	82.4	64.6
	LoRA [27][2]	0.83	98.0	40.0	91.2	21.7	93.1	56.7	85.3	69.7
	DoRA [38][2]	0.84	97.3	38.9	89.6	**22.4**	**93.9**	**58.4**	85.3	69.4
	S^2FT (Ours)	0.81	**98.8**	**41.3**	**91.4**	21.3	93.5	**58.4**	**86.1**	**70.1**
LLaMA-13B	Full FT[2]	100	98.3	47.6	92.9	26.0	95.1	65.7	88.7	73.5
	Prefix [36][1]	0.03	72.2	31.1	56.0	15.7	62.8	41.4	66.8	49.4
	Series [26][1]	0.80	93.0	44.0	80.5	22.0	87.6	50.8	78.6	65.2
	Parallel [24][1]	2.89	94.3	43.3	83.0	20.5	89.6	55.7	81.1	66.8
	LoRA [27][2]	0.67	97.5	47.8	89.9	20.5	94.3	61.2	87.4	71.2
	DoRA [38][2]	0.68	97.2	48.1	**90.6**	20.9	93.9	63.8	**88.2**	71.8
	S^2FT (Ours)	0.65	**97.7**	**48.4**	90.4	**22.8**	**95.5**	**63.9**	87.8	**72.4**
LLaMA2-7B	Full FT[2]	100	99.3	47.5	91.1	24.4	96.7	62.5	89.1	72.9
	LoRA [27][2]	0.83	97.5	44.0	91.2	20.9	94.1	59.2	85.7	70.4
	DoRA [38][2]	0.84	98.2	43.8	90.1	24.4	94.5	59.1	**89.1**	71.3
	S^2FT (Ours)	0.81	**98.5**	**44.3**	**91.1**	**25.2**	**94.7**	**61.8**	88.2	**72.0**
LLaMA3-8B	Full FT[2]	100	99.2	62.0	93.9	26.8	96.7	74.0	91.2	77.7
	LoRA [27][2]	0.70	99.5	61.6	92.7	25.6	96.3	73.8	90.8	77.2
	DoRA [38][2]	0.71	98.8	62.7	92.2	26.8	96.9	74.0	91.2	77.5
	S^2FT (Ours)	0.70	**99.7**	**65.8**	**93.7**	**31.5**	**97.8**	**76.0**	**92.4**	**79.6**

5.2 Arithmetic Reasoning

Dataset Descriptions. We followed Hu et al. [28] and evaluated S^2FT on seven math reasoning tasks, including MultiArith [53], GSM8K [14], AddSub [25], AQuA [37], SingleEq [31], SVAMP [50] and MAWPS [32]. Our fine-tuning employed the Math10K dataset [28], which combines training sets from GSM8K, MAWPS, and AQuA, augmented with LM-generated chain-of-thought steps. Therefore, these three tasks are considered ID, while the remaining four are classified as OOD tasks.

Results. As showcased in Table 2, S^2FT consistently outperforms other PEFT methods for different models. On average, it achieves improvements of 1.3% and 0.9% over LoRA and DoRA, respectively. These results highlight the versatility and effectiveness of our approach across a diverse range of tasks. Additionally, we observe substantial improvements even when compared to Full FT for the `LLaMA3-8B` model, particularly on complex tasks such as GSM8K and AQuA. This suggests that S^2FT better preserves the original reasoning capabilities of this stronger model while acquiring new skills from the fine-tuning data, thereby validating the enhanced generalization ability of our method.

5.3 Instruction Following

Dataset Descriptions. To further showcase S^2FT's superior generalization ability, we employ the instruction-following fine-tuning task with Alpaca GPT-4 dataset, which comprises 52k samples generated by GPT-4 [2] based on inputs from Alpaca [63]. Performance is measured on MT-Bench [78], featuring 80 high-quality, multi-turn questions designed to assess LLMs on eight different aspects.

Results. Table 3 offers a comprehensive evaluation of Full FT, LoRA, LISA, and S^2FT across various tasks in the MT-Bench benchmark, including Writing, Roleplay, Reasoning, Code, Math, Extraction, STEM, and Humanities. It is observed that S^2FT $>$ LISA $>$ Full FT $>$ LoRA/Galore $\geq$ Vanilla for both `Mistral-7B` and `LLama2-7B`. This is because sparse FT methods like S^2FT and LISA retain more pre-trained knowledge while acquiring new skills on the FT dataset, thereby generalizing better to diverse tasks in MT-Bench. Moreover, our method outperforms LISA due to its fine-grained and flexible selection strategy, enabling all layers to learn to follow instructions on the full fine-tuning set.

5.4 Design Choices for Trainable Parameter Allocations

Finally, we detail how S^2FT distribute trainable parameters across layers, modules, and channels.

Uniform across Layers: Following Chen et al. [10], we allocate parameters to each layer uniformly.

Table 3: Performance comparison of LLM fine-tuning methods trained on the Alpaca GPT-4 dataset. We report the MT-Bench score as the evaluation metric. All baseline results are cited from LISA [48].

Model	Method	Writing	Roleplay	Reasoning	Code	Math	Extraction	STEM	Humanities	Avg. ↑
Mistral-7B	Vanilla	5.25	3.20	4.50	1.60	2.70	**6.50**	6.17	4.65	4.32
	Full FT	5.50	**4.45**	5.45	2.50	3.25	5.78	4.75	5.45	4.64
	LoRA	5.30	4.40	4.65	2.35	3.30	5.50	5.55	4.30	4.41
	Galore	5.05	5.27	4.45	1.70	2.50	5.21	5.52	5.20	4.36
	LISA	6.84	3.65	5.45	2.20	2.75	5.65	5.95	6.35	4.85
	S^2FT (Ours)	**6.95**	4.40	**5.50**	**2.70**	**3.55**	5.95	**6.35**	**6.75**	**5.27**
LLaMA2-7B	Vanilla	2.75	4.40	2.80	1.55	1.80	3.20	5.25	4.60	3.29
	Full FT	5.55	6.45	3.60	1.75	2.00	4.70	6.45	7.50	4.75
	LoRA	6.30	5.65	4.05	1.60	1.45	4.17	6.20	6.20	4.45
	Galore	5.60	6.40	3.20	1.25	1.95	**5.05**	6.57	7.00	4.63
	LISA	6.55	**6.90**	3.45	1.60	**2.16**	4.50	6.75	7.65	4.94
	S^2FT (Ours)	**6.75**	6.60	**4.15**	**1.65**	1.85	4.75	**7.45**	**8.38**	**5.20**

Fine-tune Important Modules: Figure 4 analyzes the effectiveness of different components in a LLaMA-like Transformer Block for fine-tuning, including Query, Key, Value, Output, Up, Gate, and Down projections. To ensure a fair comparison, we maintain a fixed number of trainable parameters when fine-tuning each component. The results show that the effectiveness of components in fine-tuning follows the order: Query/Key $\ll$ Value/Up/Gate $<$ Output/Down. This is because Query/Key are only used to measure token similarities, while others serve as persistent memories of training data. Based on this finding, we allocate our parameter budget fairly to the Output and Down projections. For the `LLama3-8B` and `Mistral-7B` models, we only fine-tune the Down projection due to the inflexible selection in multi-query attention. Further analysis of this setting is left for future research.

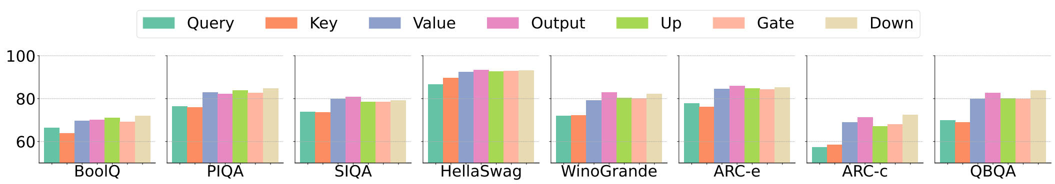


Figure 4: The impact of different components in fine-tuning, including Query, Key, Value, Output, Up, Gate, and Down projection. We fix the trainable parameter budget and only fine-tune one component.

Table 4: Comparison of various channel selection strategies on the commonsense and arithmetic reasoning datasets for the `LLama3-8B`. We report the average accuracy (%) as the evaluation metric.

Task	S^2FT-R	S^2FT-W		S^2FT-A		S^2FT-S		S^2FT-G	
		Large	Small	Large	Small	Large	Small	Large	Small
Commonsense	86.6	85.9 (-0.7)	85.3 (-1.3)	84.7 (-1.9)	87.3 (+0.7)	85.1 (-1.5)	87.2 (+0.6)	85.4 (-1.2)	86.2 (-0.4)
Arithmetic	79.6	78.4 (-1.2)	78.4 (-1.2)	77.1 (-2.5)	80.0 (+0.4)	76.8 (-2.8)	79.8 (+0.2)	77.8 (-1.8)	79.5 (-0.1)

Selection across Channels: In Section 3.2, we discuss several strategies for channel selection. In our main experiments, we employ random selection to ensure fair comparisons with baseline methods, as these approaches treat all channels with equal importance. However, the sparse structure of S^2FT offers controllability during fine-tuning, allowing us to prioritize important channels in the selection process to further boost performance. Table 4 compared nine different strategies, incorporating five varying selection metrics (i.e., random, weight, activation, weight-activation product, and gradient), each choosing either the largest or smallest values. For S^2FT-A, S^2FT-S, and S^2FT-G, we employ 1% of the fine-tuning data as a calibration set, introducing only negligible overhead during inference.

Our results demonstrate that random selection serves as a strong baseline due to its unbiased nature. Among heuristic metrics, selecting channels with the smallest activations (i.e., S^2FT-A and S^2FT-S) outperforms random selection. This indicates that these channels contain less task-specific information, enabling us to inject new knowledge through fine-tuning while preserving pre-trained capabilities in other channels. In contrast, other strategies introduce bias that compromises model performance. Notably, the counterintuitive accuracy decrease in S^2FT-G (Large) suggests that channels with large gradients contain task-related pre-trained knowledge, and modifying them will disrupt these abilities.

6 Analysis

Having demonstrated the strong generalization capability and overall performance of S^{2}FT, we now further explore its training efficiency and serving scalability compared to other fine-tuning techniques.

6.1 Training Efficiency

To evaluate training efficiency, we examine two crucial metrics: peak memory footprint and average training latency. These numbers are measured on a single Nvidia A100 (80G) SXM GPU. We keep a comparable number of parameters for all methods. To obtain the average latency, we fine-tune the model for 50 runs, each run including 200 iterations, with 10 warmup runs excluded in measurement.

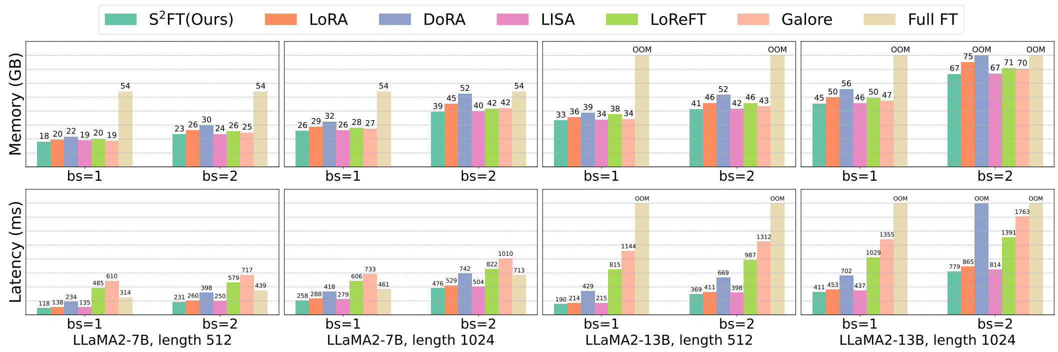


Figure 5: Comparison of memory and computation efficiency during training on the `LLaMA2-7B/13B` with varying sequence lengths and batch sizes. Average latency and peak memory usage are reported. S^{2}FT significantly improves training latency while reducing memory footprint compared to baselines.

In Figure 5, we thoughtfully profile S^{2}FT on various model sizes, sequence lengths, and batch sizes. Compared to Full FT, S^{2}FT saves 1.4-3.0$\times$ memory, and speedups fine-tuning by 1.5-2.7 times. When benchmarked against other PEFT methods, S^{2}FT establishes new standards for efficiency, offering average reductions of 2% in memory usage and 9% in latency. Notably, S^{2}FT outperforms the widely adopted LoRA, achieving about 10% improvement in both matrics by avoiding the need to store new parameters and perform additional calculations. Our partial back-propagation algorithm further improves efficiency by saving unnecessary forward activations and backward calculations.

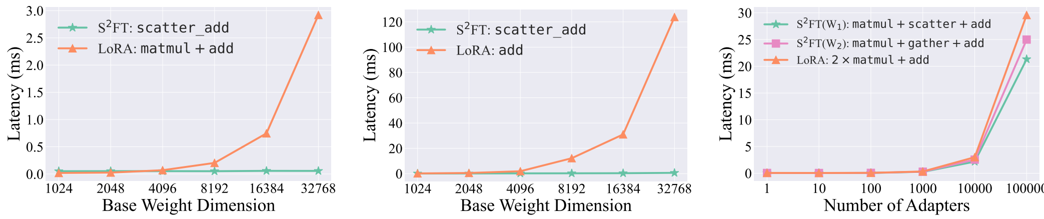
6.2 Serving Scalability

While S^{2}FT avoids additional inference overhead for a single fine-tuned model through in-place gradient updates, we will now discuss its scalability for serving thousands of fine-tuned models. To begin, we introduce the unmerged computation paradigm of S^{2}FT: Given a pre-trained weight matrix $W^{pre} \in \mathbb{R}^{d\times k}$ and its corresponding fine-tuned weight matrix W with sparsity level s, we define the weight difference as $\Delta W = W - W^{\text{pre}}$. Similar to Section 4, ΔW can be decomposed into the product of a weight matrix $V \in \mathbb{R}^{k\times s}$ and a permutation matrix $U \in \mathbb{R}^{d\times s}$. This decomposition allows us to "unmerge" an adapter $\Delta W = UV^\top$ from W, thereby sharing similarities with other adapters during inference. Following Zhong et al. [79], we consider three different adapter composition scenarios:

Adapter Fusion. To combine knowledge from multiple trained adapters, we employ weighted fusion when fine-tuning is impractical due to limited data access or computational resources. However, this approach degrades performance. In Table 5, we compare the effectiveness of LoRA and S^{2}FT when combining adapters trained separately on commonsense and arithmetic reasoning tasks, where we consider both fine-tuning overlapped and non-overlapped parameters for different adapters in S^{2}FT. Our results show that S^{2}FT with non-overlapped parameters achieves the best performance, while the overlapped variant shows inferior results. This is because S^{2}FT (non-overlap) modifies orthogonal low-rank spaces for different tasks. Similarly, LoRA largely retains task-specific capabilities during adapter fusion by optimizing low-rank projection matrices to create separate spaces for each adapter.

Table 5: Adapter Fusion Results for LoRA and S^{2}FT trained on the commonsense and arithmetic reasoning datasets using the `LLama3-8B`. We report the average accuracy (%) as the evaluation metric.

Task	**LoRA**			**S^{2}FT**			
	Commonsense	Arithmetic	Fused	Commonsense	Arithmetic	Fused (overlap)	Fused (non-overlap)
Commonsense	83.1	32.1	79.8 (-3.3)	86.6	42.3	82.0 (-4.6)	84.0 (-2.6)
Arithmetic	12.0	77.2	71.6 (-5.6)	12.8	79.6	72.2 (-7.4)	75.3 (-4.3)



(a) Switch Time on GPU (b) Switch Time on CPU (c) Parallelism Time on GPU

Figure 6: Comparison of latency for adapter switch and parallelism on a single linear layer. S^{2}FT improves scalability for switch on GPU and CPU, while saving 22% time during parallelism on GPU.

Adapter Switch. Another way to leveraging multiple adapters is to dynamically switch between them. This process involves four steps: unfusing the old adapter, unloading it from memory, loading the new adapter, and fusing it into the model. In such setting, LoRA needs two matrix multiplications (`matmul`) and two additions (`add`) on GPU whereas S^{2}FT only requires two sparse addition (`scatter_add`). In Figure 6a, we increase the base weight dimension while maintaining a sparsity of 32 for S^{2}FT and a low-rankness of 16 for LoRA. Notably, we observe that LoRA's switching time scales quadratically, while S^{2}FT remains nearly constant. Moreover, in I/O-constrained scenarios such as deployment on CPU, S^{2}FT further accelerates adapter switch by only updating a small fraction of the original weights, reducing the volume of I/O transfers, as time compared between `scatter_add` and `add` in Figure 6b.

Adapter Parallelism. To serve thousands of adapters in parallel, we decompose the computation into separate batched computations for W^{pre} and ΔW following S-LoRA [58]. While LoRA requires two `matmul` and one `add` on GPU, S^{2}FT reduces this to a `matmul`, an `add`, and either a `scatter` or `gather` for W_1 and W_2 in Section 3.1. Figure 6c shows that S^{2}FT achieves up to 22% faster inference than LoRA under the same memory constraints, with more speedup as the number of adapters scales.

7 Related Work

PEFT methods reduce the fine-tuning cost for large models, which can be categorized into 4 groups:

Adapter-based Fine-tuning introduces additional trainable module into the original model. Series Adapters insert components between MHA or FFN layers [51, 26], while parallel adapters add modules alongside existing components [24]. Recently, ReFT [67] was introduced to directly learn interventions on hidden representations. However, they introduce additional latency during inference.

Prompt-based Fine-tuning adds randomly-initialized soft tokens to the input (usually as a prefix) and train their embeddings while freezing the model weights [36, 40, 35]. These approaches result in poor performance compared to other groups, while come at the cost of significant inference overhead.

Reparameterized Fine-tuning utilizes low-rank projections to reduce trainable parameters while allowing operations with high-dimensional matrices. LoRA[27] and its recent variants like DoRA[38], AsyLoRA [80], and FLoRA [59], use low-rank matrices to approximate additive weight updates during training. To alleviate the limitations of low-rank structure, other work also add or multiply orthogonal matrices to enable high-rank updating, including MoRA [29], OFT [52], and BOFT [39]. These methods require no additional inference cost as the weight updates can be merged into models.

Sparse Fine-tuning aims to reduce the number of fine-tuned parameters by selecting a subset of pre-trained parameters that are critical to downstream tasks while discarding unimportant ones. This kind of methods are commonly used in the pre-LLM era [20, 72, 62]. However, they cannot reduce the memory footprints due to their unstructured nature. Recent approaches address this limitation through three directions: (1) developing structured variants that sacrifice selection flexibility for better hardware efficiency [48, 81], (2) incorporating sparsity into LoRA [66, 15, 41] but yield limited efficiency gains, or (3) using sparse operators for lower memory cost but slow down training [4, 49, 7].

Our work is based on the last category but achieving better performance and efficiency simultaneously. Additionally, we focus on scalable inference of PEFT methods, with S^{2}FT being the only approach that enables effective fusion, rapid switching, and efficient parallelism when serving multiple adapters.

8 Conclusion

This paper introduces S^{2}FT, a novel PEFT family that simultaneously achieves high quality, efficient training, and scalable serving for LLM fine-tuning. S^{2}FT accomplishes this by selecting sparsely and compute densely. It selects a subset of heads and channels to be trainable for the MHA and FFN modules, respectively. The weight matrices from the two sides of the coupled structures in LLMs are co-permuted to connect the selected components into dense matrices, and only these parameters are updated using dense operations. We hope S^{2}FT can be considered as a successor to LoRA for PEFT.

9 Acknowledgement

We would like to thank Songlin Yang, Tianqi Chen, Hanshi Sun, and Chris De Sa for their helpful discussions, and the authors of LLM-Adapters, ReFT, and DoRA for providing detailed results.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7

[3] Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. *arXiv preprint arXiv:2110.07560*, 2021. 2

[4] Alan Ansell, Ivan Vulić, Hannah Sterz, Anna Korhonen, and Edoardo M Ponti. Scaling sparse fine-tuning to large language models. *arXiv preprint arXiv:2401.16405*, 2024. 10

[5] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 17

[6] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023. 1

[7] Kartikeya Bhardwaj, Nilesh Prasad Pandey, Sweta Priyadarshi, Viswanath Ganapathy, Rafael Esteves, Shreya Kadambi, Shubhankar Borse, Paul Whatmough, Risheek Garrepalli, Mart Van Baalen, et al. Rapid switching and multi-adapter fusion via sparse high rank adapters. *arXiv preprint arXiv:2407.16712*, 2024. 10

[8] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024. 1

[9] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020. 6, 16

[10] Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:2301.01821*, 2023. 7

[11] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021. 30

[12] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019. 6, 16

[13] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 6, 16

[14] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 7, 16

[15] Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. *arXiv preprint arXiv:2311.11696*, 2023. 10

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[17] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16091–16101, 2023. 4

[18] Jianwei Feng and Dong Huang. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11442, 2021. 4

[19] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017. 5, 17

[20] Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*, 2020. 10

[21] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 1

[22] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016. 5, 17

[23] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024. 3

[24] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 6, 7, 10

[25] Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, 2014. 7, 16

[26] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 6, 7, 10

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 6, 7, 10

[28] Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*, 2023. 3, 6, 7, 16

[29] Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, et al. Mora: High-rank updating for parameter-efficient fine-tuning. *arXiv preprint arXiv:2405.12130*, 2024. 10

[30] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016. 5, 17

[31] Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597, 2015. 7, 16

[32] Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. Mawps: A math word problem repository. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157, 2016. 7, 16

[33] Rui Kong, Qiyang Li, Xinyu Fang, Qingtian Feng, Qingfeng He, Yazhu Dong, Weijun Wang, Yuanchun Li, Linghe Kong, and Yunxin Liu. Lora-switch: Boosting the efficiency of dynamic llm adapters via system-algorithm co-design. *arXiv preprint arXiv:2405.17741*, 2024. 1

[34] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018. 5, 17

[35] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 10

[36] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 6, 7, 10

[37] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017. 7, 16

[38] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 2, 6, 7, 10

[39] Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. *arXiv preprint arXiv:2311.06243*, 2023. 10

[40] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *AI Open*, 2023. 10

[41] Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. Alora: Allocating low-rank adaptation for fine-tuning large language models. *arXiv preprint arXiv:2403.16187*, 2024. 10

[42] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR, 2023. 2

[43] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017. 5, 17

[44] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023. 1

[45] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023. 2, 4

[46] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018. 6, 16

[47] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. *arXiv preprint arXiv:2302.06232*, 2023. 31

[48] Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning. *arXiv preprint arXiv:2403.17919*, 2024. 2, 6, 8, 10

[49] Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms. *arXiv preprint arXiv:2406.16797*, 2024. 10

[50] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, June 2021. Association for Computational Linguistics. 7, 16

[51] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020. 10

[52] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 10

[53] Subhro Roy and Dan Roth. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*, 2016. 7, 16

[54] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023. 1

[55] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. 6, 16

[56] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. 6, 16

[57] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. 5, 17

[58] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-lora: Serving thousands of concurrent lora adapters. *arXiv preprint arXiv:2311.03285*, 2023. 1, 2, 10

[59] Chongjie Si, Xuehui Wang, Xue Yang, Zhengqin Xu, Qingyun Li, Jifeng Dai, Yu Qiao, Xiaokang Yang, and Wei Shen. Flora: Low-rank core space for n-dimension. *arXiv preprint arXiv:2405.14739*, 2024. 10

[60] GW Stewart. On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, 17(1):33–45, 1969. 32

[61] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 2

[62] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 10

[63] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023. 7

[64] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[65] J Leo van Hemmen and Tsuneya Ando. An inequality for trace ideals. *Communications in Mathematical Physics*, 76:143–148, 1980. 32

[66] Haoyu Wang, Tianci Liu, Tuo Zhao, and Jing Gao. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. *arXiv preprint arXiv:2406.10777*, 2024. 10

[67] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024. 6, 10

[68] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023. 1

[69] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*, 2021. 2

[70] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. 30

[71] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023. 1

[72] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 10

[73] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. 6, 16

[74] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 5, 17

[75] Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023. 1

[76] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024. 2, 6

[77] Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. *arXiv preprint arXiv:2405.00732*, 2024. 1

[78] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[79] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024. 9

[80] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. *arXiv preprint arXiv:2402.16842*, 2024. 10

[81] Ligeng Zhu, Lanxiang Hu, Ji Lin, and Song Han. Lift: Efficient layer-wise fine-tuning for large model models. 2023. 10

A Limitations

While our work demonstrates the effectiveness of S^{2}FT for LLM fine-tuning, several promising directions remain unexplored. First, extending S^{2}FT to other architectures with coupled structures, such as CNNs and RNNs, can broaden its applicability. Second, verifying our approach beyond language tasks, particularly in large vision/multi-modal models, will enhance its versatility. Third, exploring more selection strategies can provide deeper insights into optimal fine-tuning protocols due to S^{2}FT's controllability. Finally, although our work confirms the feasibility of scalable deployment, developing a practical and efficient serving system for S^{2}FT remains an important next step.

B Broader Impacts

Since our work focuses on PEFT, it leads to a reduction in hardware resource and energy consumption. Given the growing adoption of LLMs across diverse domains and the corresponding surge in fine-tuning demands, S^{2}FT should represent an important step toward more sustainable AI development.

C Detailed Experimental Setups for Section 2

In this study, we use SpFT, LoRA, and Full FT to fine-tune the `LLaMA-3-8B` model on Math10K dataset [28] for 3 epochs. The Math10K dataset combines training sets from GSM8K [14], MAWPS [32], and AQuA [37], augmented with language model-generated chain-of-thought steps. We train the model for 3 epochs with a batch size of 64. For both PEFT methods–SpFT and LoRA–we tune with three ratios of trainable parameters (p) in each linear layer: 10%, 1%, and 0.1%. We evaluate the model's performance on both arithmetic and commonsense reasoning tasks, representing near out-of-distribution (OOD) and far OOD generalization scenarios, respectively. The arithmetic tasks comprise seven subtasks: MultiArith [53], GSM8K [14], AddSub [25], AQuA [37], SingleEq [31], SVAMP [50], and MAWPS [32]. The commonsense reasoning evaluation includes eight subtasks: BoolQ [12], PIQA [9], SocialQA [56], HellaSwag [73], WinoGrande [55], ARC-challenge [13], ARC-easy [13], and OpenbookQA [46]. Based on task complexity within arithmetic reasoning, we categorize MultiArith, AddSub, SingleEq, and MAWPS as easy subtasks, while the remaining arithmetic tasks are classified as hard subtasks.

D Detailed Experimental Setups and Hyperparamters for Section 5

The detailed selection strategies and number of trainable parameters are presented in Section 5. Additional hyperparameter configurations for all tasks are provided in Table 6. We maintain the same hyperparameter settings across the `LLaMA-7B`, `LLaMA-13B`, `LLaMA2-7B`, and `LLaMA3-8B` models.

Table 6: Hyperparameter configurations of S^{2}FT on various base models across three tasks.

Hyperparameters	**Commonsense Reasoning**	**Arithmetic Reasoning**	**Instruction Following**
Optimizer	AdamW	AdamW	AdamW
LR	2e-4	1e-3	2e-5
LR Scheduler	linear	linear	cosine
Batch size	16×4	16×4	16×4
Warmup Steps	100	100	0
Epochs	3	3	1

E Proofs for Theoretical Results in Section 4

Here we provide proofs for the results in Section 4.

E.1 Notation

For a vector a, let $\|a\|$ be the ℓ_2 norm of a. For $d_1 \geq d_2$, denote a set of orthogonal matrices by $\mathbb{O}_{d_1,d_2} := \{R \in \mathbb{R}^{d_1 \times d_2} : R^\top R = I_{d_2}\}$. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\|A\|_{\mathrm{F}}$ and $\|A\|_{\mathrm{op}}$

be the Frobenius norm and spectral norm of A, respectively. Denote the condition number of A by $\kappa_*(A) := \|A\|_{\mathrm{op}}/\lambda_*(A)$. Let $A^\dagger$ be Moore-Penrose inverse of A. For a symmetric matrix A, denote its effective rank by $r_e(A) := \mathrm{tr}(A)/\|A\|_{\mathrm{op}}$. Note that $r_e(A) \le \mathrm{rank}(A)$ always holds. For $a, b \in \mathbb{R}$, we let $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\mathrm{SVD}_r(A) := \Phi_r(A)\Lambda_r(A)\Psi_r^\top(A)$ be the top-r singular value decomposition of A, where $\Phi_r(A) \in \mathbb{O}_{d_1,r}$ and $\Psi_r(A) \in \mathbb{O}_{d_2,r}$ are top-r left and right singular vectors of A, respectively, and $\Lambda_r(A) = \mathrm{diag}(\lambda_1(A), \dots, \lambda_r(A)) \in \mathbb{R}^{r\times r}$ is a diagonal matrix of singular values of A, where $\lambda_j(A)$ denotes the j-th largest singular value of A. Define $\Phi_*(A) := \Phi_{\mathrm{rank}(A)}(A)$ and $\Psi_*(A) := \Psi_{\mathrm{rank}(A)}(A)$ as the left and right singular vectors of A corresponding to non-zero singular values, respectively. Define the smallest *positive* singular value of A as $\lambda_*(A) = \lambda_{\mathrm{rank}(A)}(A)$ and let $\Lambda_*(A) = \Lambda_{\mathrm{rank}(A)}(A)$. For a deep learning model fine-tuned on n i.i.d. samples $(x_i^{(\mathrm{i})}, y_i^{(\mathrm{i})}) \subset \mathbb{R}^p \times \mathbb{R}^q$, we say an event $\mathcal{F}$ occurs with high probability when $\mathbb{P}(\mathcal{F}) = 1 - \exp\big(-\Omega(\log^2(n + p + q))\big)$.

E.2 Setup

We consider multivariate regression task. Using n i.i.d. samples $(x_i^{(\mathrm{i})}, y_i^{(\mathrm{i})}) \subset \mathbb{R}^p \times \mathbb{R}^q$ from in-distribution task, we fine-tune a pre-trained network $f^{\mathrm{pre}} : \mathbb{R}^p \to \mathbb{R}^q$ for better prediction.

Deep Linear Networks We consider deep linear networks of the form $x \mapsto W_L W_{L-1} \dots W_1 x : \mathbb{R}^d \to \mathbb{R}^p$, where $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, with $d_L = q$ and $d_0 = p$. In comparison to multi-head attention transformers, each row of W_ℓ can be viewed as corresponding to the parameters in a single head. Let $f^{\mathrm{pre}}(x) = W_L^{\mathrm{pre}} W_{L-1}^{\mathrm{pre}} \dots W_1^{\mathrm{pre}} x : \mathbb{R}^p \to \mathbb{R}^q$ represent a pre-trained neural network. We denote $\overline{W}_\ell^{\mathrm{pre}} := W_L^{\mathrm{pre}} W_{L-1}^{\mathrm{pre}} \dots W_\ell^{\mathrm{pre}} \in \mathbb{R}^{d_L \times d_{\ell-1}}$ as the weights up to the ℓ-th layer, and $\underline{W}_\ell^{\mathrm{pre}} := W_\ell^{\mathrm{pre}} W_{\ell-1}^{\mathrm{pre}} \dots W_1^{\mathrm{pre}} \in \mathbb{R}^{d_\ell \times d_0}$ as the weights above the ℓ-th layer, with the promise that $\underline{W}_0^{\mathrm{pre}} = I$. Deep linear networks have been widely used to facilitate the theoretical analysis of modern complex deep neural networks [57, 30, 43, 22, 34, 5].

Fine-Tuning We employ ℓ_2 distance as the error metric. Given a pre-trained network f^{pre}, we fine-tune its ℓ-th layer by minimizing the empirical in-distribution risk $\mathcal{R}_n^{(\mathrm{i})}(f) := (1/n)\sum_{i\in[n]} \|y_i^{(\mathrm{i})} - f(x_i^{(\mathrm{i})})\|^2$, where $(x_i^{(\mathrm{i})}, y_i^{(\mathrm{i})}) \subset \mathbb{R}^p \times \mathbb{R}^q$ are n i.i.d. observations from in-distribution task. More specifically, we consider a class of rank-d adaptation defined as

$$f_{\ell,U,V}(x) := \overline{W}_{\ell+1}^{\mathrm{pre}}(W_\ell^{\mathrm{pre}} + UV^\top)\underline{W}_{\ell-1}^{\mathrm{pre}} x, \tag{4}$$

where $U \in \mathbb{R}^{d_\ell \times d}$ and $V \in \mathbb{R}^{d_{\ell-1}\times d}$ are parameters to fine-tune. Note that by regarding multiple consecutive layers as a single layer, our settings can be extended to multi-layer fine-tuning.

We specifically compare two fine-tuning methods: LoRA and S^{2}FT.

- **LoRA.** For a fixed $\ell \in [L]$, and low-rankness level $1 \le r \le \min\{d_\ell, d_{\ell-1}\}$, we train the low-rank matrices (U, V) in (4) by minimizing the empirical in-distribution risk via gradient descent. Motivated from the previous results that gradient descent has implicit regularization [74, 19, 5], we directly consider the minimum norm solutions:

$$(U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}) \in \operatorname*{arg\,min}_{U,V} \|(U, V)\|_{\mathrm{F}}^2 \quad \text{s.t. } (U, V) \text{ minimizes } \mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U,V}). \tag{5}$$

- **S^{2}FT.** For a fixed $\ell \in [L]$, and a sparsity level $s = \lfloor r \cdot \frac{d_\ell + d_{\ell-1}}{d_{\ell-1}} \rfloor$, we train only V in (4) with the fixed choice of $U \leftarrow U_S^{\mathrm{S^2FT}} := [e_{a_1}; e_{a_2}; \dots; e_{a_s}]$, which specifies s channels to fine-tune, where $S = \{a_1, a_2, \dots, a_s\} \subset [d_\ell]$. Here e_a is the standard basis vector with the a-th entry being 1. We minimize the empirical in-distribution risk via gradient descent. Similar to LoRA, we consider the following minimum norm solution:

$$V^{\mathrm{S^2FT}} = \operatorname*{arg\,min}_{V} \|V\|_{\mathrm{F}}^2 \quad \text{s.t. } V \text{ minimizes } \mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V}). \tag{6}$$

Data Generating Process As a simplification of the data generating process, we consider multiple linear regression. Assume that the in-distribution data $(x^{(\mathrm{i})}, y^{(\mathrm{i})}) \in \mathbb{R}^{p+q}$ and out-of-distribution data

$(x^{(\mathrm{o})}, y^{(\mathrm{o})}) \in \mathbb{R}^{p+q}$ are generated according to

$$y^{(k)} = B^{(k)} x^{(k)} + \epsilon^{(k)}, \ \ k \in \{\mathrm{i}, \mathrm{o}\}, \tag{7}$$

where $B^{(k)} \in \mathbb{R}^{q\times p}$, and $\epsilon^{(k)} \in \mathbb{R}^q$ is the error term satisfying $\mathbb{E}[\epsilon^{(k)}|x^{(k)}] = 0$. Assume that $\Sigma_\epsilon^{(k)} := \mathbb{E}[\epsilon^{(k)}\epsilon^{(k)\top}] \in \mathbb{R}^{q\times q}$ exists and $\mathbb{E}[x^{(k)}] = 0$. The signal covariance matrix is denoted by $\Sigma_x^{(k)} := \mathbb{E}[x^{(k)}x^{(k)\top}] \in \mathbb{R}^{p\times p}$.

We define the in-distribution and out-of-distribution risks of $f : \mathbb{R}^p \to \mathbb{R}^q$ as:

$$\mathcal{R}^{(k)}(f) = \mathbb{E}[\|y^{(k)} - f(x^{(k)})\|], \ \ k \in \{\mathrm{i}, \mathrm{o}\}.$$

For notational brevity, we can write $W^{\mathrm{pre}} = \underline{W}_L^{\mathrm{pre}} \in \mathbb{R}^{q\times p}$. Let $X^{(\mathrm{i})} := (x_1^{(\mathrm{i})}, \dots, x_n^{(\mathrm{i})}) \in \mathbb{R}^{p\times n}$, $Y^{(\mathrm{i})} := (y_1^{(\mathrm{i})}, \dots, y_n^{(\mathrm{i})}) \in \mathbb{R}^{q\times n}$, and $E^{(\mathrm{i})} = (\epsilon_1^{(\mathrm{i})}, \dots, \epsilon_n^{(\mathrm{i})}) := Y^{(\mathrm{i})} - B^{(\mathrm{i})}X^{(\mathrm{i})} \in \mathbb{R}^{q\times n}$. Denote the in-distribution sample covariance matrices by $\hat{\Sigma}_x^{(\mathrm{i})} := (1/n)X^{(\mathrm{i})}X^{(\mathrm{i})\top}$, $\hat{\Sigma}_\epsilon^{(\mathrm{i})} := (1/n)E^{(\mathrm{i})}E^{(\mathrm{i})\top}$, $\hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})} := (1/n)X^{(\mathrm{i})}E^{(\mathrm{i})\top}$, $\hat{\Sigma}_{\epsilon,x}^{(\mathrm{i})} = \hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})\top}$. Define $\check{\Sigma}_{x,\epsilon}^{(k)} = (X^{(\mathrm{i})\top})^\dagger E^{(\mathrm{i})\top}$, $\hat{A} := (\underline{W}_{\ell-1}^{\mathrm{pre}}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top})^{1/2}$, $A := (\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top})^{1/2}$, $\Phi' := \Phi_*(\overline{W}_{\ell+1}^{\mathrm{pre}})$, $\Phi_S'' := \Phi_*(\overline{W}_{\ell+1}^{\mathrm{pre}}U_S^{\mathrm{S^2FT}})$, $D = B^{(\mathrm{i})} - W^{\mathrm{pre}}$, $\hat{D} := B^{(\mathrm{i})} - W^{\mathrm{pre}} + \check{\Sigma}_{\epsilon,x}^{(\mathrm{i})}$. Also define $M := \Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger$ and $\hat{M} := \Phi'^\top\hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\hat{A}^\dagger$. Let $\hat{\Psi}' := \Psi_*(\hat{A})$, and $G_\ell^{(\mathrm{i},\mathrm{o})} := (\underline{W}_\ell^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2})^\dagger\underline{W}_\ell^{\mathrm{pre}}\Sigma_x^{(\mathrm{o})1/2}$ be a matrix that captures the covariate shift at the ℓ-th layer.

We consider fine-tuning the ℓ-th ($\ell \in [L]$) layer of the pre-trained deep linear network $f^{\mathrm{pre}}(x) = W_L^{\mathrm{pre}}W_{L-1}^{\mathrm{pre}}\dots W_1^{\mathrm{pre}}x$ using in-distribution observations $(x_i^{(\mathrm{i})}, y_i^{(\mathrm{i})})_{i\in[n]}$.

To measure the performance of models, we define the excess risks of f for the task $k \in \{\mathrm{i}, \mathrm{o}\}$ as

$$\mathcal{E}^{(k)}(f) := \mathbb{E}[\|y^{(k)} - f(x^{(k)})\|^2] - \inf_{f'}\mathbb{E}[\|y^{(k)} - f'(x^{(k)})\|^2],$$

where the infimum is taken over all square integrable functions.

E.3 Assumptions

We assume that $\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top} \neq 0$, since otherwise $\underline{W}_{\ell-1}^{\mathrm{pre}}x^{(\mathrm{i})} = 0$ almost surely and fine-tuning the ℓ-th layer does not improve the performance of pre-trained model. Define the in-distribution prediction residuals for pre-trained model f^{pre} by $\Sigma_f^{(\mathrm{i})} := \mathbb{E}[(B^{(\mathrm{i})}x^{(\mathrm{i})} - W^{\mathrm{pre}}x^{(\mathrm{i})})(B^{(\mathrm{i})}x^{(\mathrm{i})} - W^{\mathrm{pre}}x^{(\mathrm{i})})^\top]$. Note that $\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}) = \mathrm{tr}\Big(\Sigma_f^{(\mathrm{i})}\Big)$. We also assume that $\|\Sigma_f^{(\mathrm{i})}\|_{\mathrm{op}} > 0$. Since otherwise $\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}) = \|\Sigma_f^{(\mathrm{i})}\|_{\mathrm{F}}^2 = 0$ and there is no room for improvement from pre-trained model.

We introduce several assumptions.

Assumption E.1 (Sub-Gaussianity)**.** Assume that there exist some constants $c_1, c_2 \in (0, \infty)$ such that $(x^{(\mathrm{i})}, \epsilon^{(\mathrm{i})})$ in the model 7 satisfies

$$\gamma^\top\Sigma_x^{(\mathrm{i})}\gamma \geq c_1\|\gamma^\top x^{(\mathrm{i})}\|_{\psi_2}^2, \quad \text{and} \quad \gamma'^\top\Sigma_\epsilon^{(\mathrm{i})}\gamma' \geq c_2\|\gamma'^\top\epsilon^{(\mathrm{i})}\|_{\psi_2}^2,$$

for any $\gamma \in \mathbb{R}^d$ and $\gamma' \in \mathbb{R}^p$, where $\|y\|_{\psi_2}$ is the sub-Gaussian norm defined as

$$\|y\|_{\psi_2} := \inf\{\upsilon > 0 : \mathbb{E}[\exp(y^2/\upsilon^2)] \leq 2\}$$

for a random variable y taking values in $\mathbb{R}$.

Assumption E.2 (Sufficiently Many Observations)**.** Assume that

$$n \gg (\kappa_*^4(A)r_e(A^2) + \kappa_*^2(\Sigma_x^{(\mathrm{i})})r_e(\Sigma_x^{(\mathrm{i})}) + r_e(D\Sigma_x^{(\mathrm{i})}D^\top))\log^2(n+d+p),$$

$$n \gg \frac{\|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}}{\|D\Sigma_x^{(\mathrm{i})}D^\top\|_{\mathrm{op}}}(r_e(\Sigma_\epsilon^{(\mathrm{i})}) + r_e(A^2))\log^2(n+d+p),$$

and

$$n \gg \kappa_*^4(\Sigma_x^{(\mathrm{i})})\frac{r_e(\Sigma_x^{(\mathrm{i})})(r_e(\Sigma_\epsilon^{(\mathrm{i})}) + r_e(\Sigma_x^{(\mathrm{i})}))}{r_e(A^2)}\log^2(n+d+p).$$

Assumption E.3 (Eigengap Condition)**.** Assume that there exists some constant $C_{\mathrm{g}} > 0$ such that

$$\frac{\lambda_s(\Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} A^{\dagger})}{\lambda_s(\Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} A^{\dagger}) - \lambda_{s+1}(\Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} A^{\dagger})} \lesssim C_{\mathrm{g}}$$

holds.

Assumption E.3 is necessary to identify the rank-s approximation of M, which is used to derive the risk of LoRA.

Assumption E.4 (Approximate Sparsity of Heads)**.** Assume that there exists some $S_0 \subset [d_\ell]$ with $|S_0| \le s$ and $\delta > 0$ such that

$$\sum_{a \in [d_\ell] \setminus S_0} \| e_a^{\top} (\overline{W}_{\ell+1}^{\mathrm{pre}})^{\dagger} (B^{(\mathrm{i})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{i})1/2} \|^2 \le \delta^2 \| (\overline{W}_{\ell+1}^{\mathrm{pre}})^{\dagger} (B^{(\mathrm{i})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{i})1/2} \|_{\mathrm{F}}^2$$

holds.

Assumption E.5 (Distribution Shift)**.** Assume that $\Sigma_x^{(\mathrm{i})} = \Sigma_x^{(\mathrm{o})} = \Sigma_x$ for some $\Sigma_x \in \mathbb{R}^{d \times d}$ and that $\|\Phi_*^{\top} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}}) (B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{1/2} \|_{\mathrm{F}}^2 \le \varepsilon^2 \mathcal{E}^{(\mathrm{o})}(f^{\mathrm{pre}})$ for some $\varepsilon > 0$.

Assumption E.6 (Condition Number)**.** Assume that $\kappa_*(M) \lesssim 1$, $\kappa_*(\overline{W}_{\ell+1}^{\mathrm{pre}}) \lesssim 1$, $\kappa_*(\Sigma_f^{(\mathrm{i})}) \lesssim 1$ and $\kappa_*(\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top}) \lesssim 1$.

Note that Assumption E.6 is not essential to our analysis.

E.4 Main Results

We first demonstrate that LoRA and S^{2}FT exhibit comparable memorization abilities. Next, we present a formal restatement of 4.2 that combine Theorems E.10, E.11, E.13, E.15, and Lemma E.14.

Theorem E.7. *Suppose that Assumptions E.1, E.2, E.3, E.4, and E.6 hold. Choose S such that $S \supset S_0$ holds. Let $U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}$ be the LoRA adaptation matrices defined in (5). Let $V^{\mathrm{S^2FT}}$ be the S^{2}FT adaptation matrices given $U_S^{\mathrm{S^2FT}}$ defined in (6). Then, for all sufficiently large n, the following holds with probability $1 - \exp\big(-\Omega(\log^2(n + d + p))\big)$: for any $\eta > 0$,*

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V^{\mathrm{S^2FT}}}) \le (1 + \eta)(T_{\mathrm{bias}}^{\mathrm{S^2FT}})^2 + (1 + \eta^{-1})(T_{\mathrm{variance}}^{\mathrm{S^2FT}})^2,$$
$$\mathcal{E}^{(\mathrm{i})}(f_{\ell, U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}}) \le (1 + \eta)(T_{\mathrm{bias}}^{\mathrm{LoRA}})^2 + (1 + \eta^{-1})(T_{\mathrm{variance}}^{\mathrm{LoRA}})^2,$$

where

$$0 \le (T_{\mathrm{bias}}^{\mathrm{LoRA}})^2 - \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) \le (T_{\mathrm{bias}}^{\mathrm{S^2FT}})^2 - \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) \lesssim \delta^2 \mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}),$$
$$(T_{\mathrm{variance}}^{\mathrm{S^2FT}})^2 \lesssim (\|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}} + \|\Sigma_f^{(\mathrm{i})}\|_{\mathrm{op}}) \frac{s d_{\ell-1} \log^2(n + p + q)}{n},$$
$$(T_{\mathrm{variance}}^{\mathrm{LoRA}})^2 \lesssim (\|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}} + \|\Sigma_f^{(\mathrm{i})}\|_{\mathrm{op}}) \frac{r(d_\ell + d_{\ell-1}) \log^2(n + p + q)}{n}.$$

Theorem E.8 (Restatement of Theorem 4.2)**.** *Consider the limit $n \to \infty$. Suppose that Assumption E.5 holds. Let $U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}$ be the LoRA adaptation matrices defined in (15). Let $V^{\mathrm{S^2FT}}$ be the S^{2}FT adaptation matrices given $U_S^{\mathrm{S^2FT}}$ defined in (25). If $B^{(\mathrm{i})} = \overline{W}_{\ell+1}^{\mathrm{pre}} \tilde{B} \underline{W}_{\ell-1}^{\mathrm{pre}}$ holds for some $\tilde{B}^{(\mathrm{i})} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, and $s \le \mathrm{rank}(\Sigma_f^{(\mathrm{i})})$, then,*

$$\mathcal{E}^{(\mathrm{o})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V^{\mathrm{S^2FT}}}) \le (1 + 3\varepsilon^2) \mathcal{E}^{(\mathrm{o})}(f^{\mathrm{pre}}),$$
$$\mathcal{E}^{(\mathrm{o})}(f_{\ell, U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}}) \ge \|(B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{1/2}\|_{\mathrm{F}}^2.$$

Intuition of the proof of Theorem E.8. LoRA forgets pre-trained tasks due to its model complexity. Consider the simplest low-rank adaptation to a single-layer linear network:

$$\Delta_1 \in \mathop{\arg\min}_{\substack{\Delta_1' \in \mathbb{R}^{d_1 \times d_0} \\ \mathrm{rank}(\Delta_1') = s}} \mathbb{E}[\|y^{(\mathrm{i})} - (W_1^{\mathrm{pre}} + \Delta_1') x^{(\mathrm{i})}\|^2].$$

Assume that $\Sigma_x^{(\mathrm{i})} = I$, then we can show that the solution is $\Delta_1 = \mathrm{SVD}_s(B^{(\mathrm{i})} - W_1^{\mathrm{pre}})$. Under the condition that the rank of $B^{(\mathrm{i})} - W_1^{\mathrm{pre}}$ is smaller than, or comparable to s, LoRA fine-tuned model can learn the in-distribution best regressor in ℓ_2 sense, since $(W_1^{\mathrm{pre}} + \Delta_1)x \approx B^{(\mathrm{i})}x = \mathbb{E}[y^{(\mathrm{i})}|x^{(\mathrm{i})} = x]$. Hence it makes LoRA fine-tuned model vlunerable to distribution shift.

On the other hand, we model S²FT as fine-tuning only a few heads:

$$\Delta_1 \in \operatorname*{arg\,min}_{\Delta_1' = \sum_{a\in S} e_a v_a^\top, v_a \in \mathbb{R}^{d_0}} \mathbb{E}[\|y^{(\mathrm{i})} - (W_1^{\mathrm{pre}} + \Delta_1')x^{(\mathrm{i})}\|^2].$$

Although S²FT is a special case of LoRA, the constraint on the direction of low-rank matrix prevents overfitting to the in-distribution task. To see this, note that a sparse fine-tuned model can be written as

$$(W_1^{\mathrm{pre}} + \Delta_1)x = W_1^{\mathrm{pre}}x + \sum_{a\in S} e_a e_a^\top (B^{(\mathrm{i})} - W_1^{\mathrm{pre}})x = \sum_{a\in S^c} e_a e_a^\top W_1^{\mathrm{pre}}x + \sum_{a \in S} e_a e_a^\top B^{(\mathrm{i})}x,$$

where $S \subset [d_1]$ is a set of channels/heads with cardinality s. Since sparse fine-tuned model keeps parameters from the pre-trained model, except for rows specified by S, the model less forget pre-training tasks. $\square$

E.5 Proofs for LoRA

E.5.1 Excess Risk of LoRA

Lemma E.9 (Excess Risk)**.** *Consider the minimum norm solution*

$$(U^{\mathrm{LoRA}}, V^{\mathrm{LoRA}}) \in \operatorname*{arg\,min}_{(U,V)\in\mathbb{R}^{d_\ell \times s}\times\mathbb{R}^{d_{\ell-1}\times s}} \|(U,V)\|_{\mathrm{F}}^2 \quad \textit{s.t. } (U,V) \textit{ minimizes } \mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U,V}).$$

Then, the low-rank adaptation matrix satisfies

$$U^{\mathrm{LoRA}}V^{\mathrm{LoRA}\top} = (\overline{W}_{\ell+1}^{\mathrm{pre}})^\dagger \mathit{SVD}_s(\overline{W}_{\ell+1}^{\mathrm{pre}}(\overline{W}_{\ell+1}^{\mathrm{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\hat{A}^\dagger)\hat{A}^\dagger,$$

and

$$\mathcal{E}^{(k)}(f_{\ell,U^{\mathrm{LoRA}},V^{\mathrm{LoRA}}}) = \mathrm{tr}\bigg(\Big(B^{(k)} - W^{\mathrm{pre}} - \mathit{SVD}_s(\overline{W}_{\ell+1}^{\mathrm{pre}}(\overline{W}_{\ell+1}^{\mathrm{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\hat{A}^\dagger)\hat{A}^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Big)\Sigma_x^{(k)} \cdot \Big(B^{(k)} - W^{\mathrm{pre}} - \mathit{SVD}_s(\overline{W}_{\ell+1}^{\mathrm{pre}}(\overline{W}_{\ell+1}^{\mathrm{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\hat{A}^\dagger)\hat{A}^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Big)^\top\bigg)$$

for $k \in \{\mathrm{i}, \mathrm{o}\}$.

Proof of Lemma E.9. The empirical risk of $f_{\ell,U,V}$ for the in-distribution task can be written as

$$\begin{aligned}
\mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U,V}) &= \frac{1}{n}\sum_{i\in[n]} \|(B^{(\mathrm{i})} - W^{\mathrm{pre}})x_i^{(\mathrm{i})} + \epsilon_i^{(\mathrm{i})} - \overline{W}_{\ell+1}^{\mathrm{pre}}UV^\top\underline{W}_{\ell-1}^{\mathrm{pre}}x_i^{(\mathrm{i})}\|^2 \\
&= \mathrm{tr}\Big((B^{(\mathrm{i})} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}}UV^\top\underline{W}_{\ell-1}^{\mathrm{pre}})\hat{\Sigma}_x^{(\mathrm{i})}(B^{(\mathrm{i})} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}}UV^\top\underline{W}_{\ell-1}^{\mathrm{pre}})^\top\Big) \\
&\quad + 2\,\mathrm{tr}\Big((B^{(\mathrm{i})} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}}UV^\top\underline{W}_{\ell-1}^{\mathrm{pre}})\hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})}\Big) + \mathrm{tr}\Big(\hat{\Sigma}_\epsilon^{(\mathrm{i})}\Big) \\
&= \mathrm{tr}\Big(V^\top\underline{W}_{\ell-1}^{\mathrm{pre}}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}VU^\top\overline{W}_{\ell+1}^{\mathrm{pre}\top}\overline{W}_{\ell+1}^{\mathrm{pre}}U\Big) \\
&\quad - 2\,\mathrm{tr}\Big(\overline{W}_{\ell+1}^{\mathrm{pre}}UV^\top\underline{W}_{\ell-1}^{\mathrm{pre}}\Big\{\hat{\Sigma}_x^{(\mathrm{i})}(B^{(\mathrm{i})} - W^{\mathrm{pre}})^\top + \hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})}\Big\}\Big) \\
&\quad + \mathrm{tr}\Big((B^{(\mathrm{i})} - W^{\mathrm{pre}})\hat{\Sigma}_x^{(\mathrm{i})}(B^{(\mathrm{i})} - W^{\mathrm{pre}})^\top\Big) + 2\,\mathrm{tr}\Big((B^{(\mathrm{i})} - W^{\mathrm{pre}})\hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})}\Big) + \mathrm{tr}\Big(\hat{\Sigma}_\epsilon^{(\mathrm{i})}\Big).
\end{aligned} \tag{8}$$

Since $\hat{\Sigma}^{(\mathrm{i})}_{x,\epsilon} = \hat{\Sigma}^{(\mathrm{i})}_x (X^{(\mathrm{i})\top})^\dagger E^{(\mathrm{i})\top} = \hat{\Sigma}^{(\mathrm{i})}_x \check{\Sigma}^{(\mathrm{i})}_{x,\epsilon}$,

$$\begin{aligned}
\mathcal{R}^{(\mathrm{i})}_n(f_{\ell,U,V}) &= \operatorname{tr}\Big(\hat{A}VU^\top \overline{W}^{\mathrm{pre}\top}_{\ell+1}\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top\hat{A}\Big) - 2\operatorname{tr}\Big(\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top\hat{A}\hat{A}^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\hat{\Sigma}^{(\mathrm{i})}_x\hat{D}^\top\Big)\\
&\quad - 2\operatorname{tr}\Big(\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top(I-\hat{A}\hat{A}^\dagger)\underline{W}^{\mathrm{pre}}_{\ell-1}\hat{\Sigma}^{(\mathrm{i})}_x\hat{D}^\top\Big)\\
&\quad + \operatorname{tr}\Big(D\hat{\Sigma}^{(\mathrm{i})}_x D^\top\Big) + 2\operatorname{tr}\Big(D\hat{\Sigma}^{(\mathrm{i})}_{x,\epsilon}\Big) + \operatorname{tr}\Big(\hat{\Sigma}^{(\mathrm{i})}_\epsilon\Big)\\
&= \|\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top\hat{A} - \hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger\|_{\mathrm{F}}^2 - \|\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger\|_{\mathrm{F}}^2\\
&\quad + \operatorname{tr}\Big(D\hat{\Sigma}^{(\mathrm{i})}_x D^\top\Big) + 2\operatorname{tr}\Big(D\hat{\Sigma}^{(\mathrm{i})}_{x,\epsilon}\Big) + \operatorname{tr}\Big(\hat{\Sigma}^{(\mathrm{i})}_\epsilon\Big),
\end{aligned} \tag{9}$$

where we used $(I-\hat{A}\hat{A}^\dagger)\underline{W}^{\mathrm{pre}}_{\ell-1}\hat{\Sigma}^{(\mathrm{i})1/2}_x = 0$. From (9), minimizing $\mathcal{R}^{(\mathrm{i})}_n(f_{\ell,U,V})$ is equivalent to minimizing the norm:

$$\begin{aligned}
\|\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top\hat{A} - \hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger\|_{\mathrm{F}}^2 &= \|\overline{W}^{\mathrm{pre}}_{\ell+1}UV^\top\hat{A} - \overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger\|_{\mathrm{F}}^2\\
&\quad + \|(I-\overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger)\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger\|_{\mathrm{F}}^2.
\end{aligned}$$

This is minimized by (U', V') satisfying

$$\begin{aligned}
U'V'^\top &= (\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger \mathrm{SVD}_s(\overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger)\hat{A}^\dagger\\
&\quad + (I-(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\overline{W}^{\mathrm{pre}}_{\ell+1})A_1 + A_2(I-\hat{\Psi}'\hat{\Psi}'^\top),
\end{aligned} \tag{10}$$

where $A_1, A_2 \in \mathbb{R}^{d_\ell\times d_{\ell-1}}$ are arbitrary matrices. Since we particularly consider the minimum norm solution, we must have $A_1 = 0$ and $A_2 = 0$. Hence

$$\overline{W}^{\mathrm{pre}}_{\ell+1}U^{\mathrm{LoRA}}V^{\mathrm{LoRA}\top}\underline{W}^{\mathrm{pre}}_{\ell-1} = \mathrm{SVD}_s(\overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger)\hat{A}^\dagger\underline{W}^{\mathrm{pre}}_{\ell-1}.$$

Therefore, the excess risk for $k\in\{\mathrm{i},\mathrm{o}\}$ becomes

$$\begin{aligned}
\mathcal{E}^{(k)}(f_{\ell,U^{\mathrm{LoRA}},V^{\mathrm{LoRA}}}) &= \mathbb{E}\bigg[\Big(B^{(k)}x^{(k)} - \overline{W}^{\mathrm{pre}}_{\ell+1}(W^{\mathrm{pre}}_\ell + U^{\mathrm{LoRA}}V^{\mathrm{LoRA}\top})\underline{W}^{\mathrm{pre}}_{\ell-1}x^{(k)}\Big)^2\bigg]\\
&= \operatorname{tr}\bigg(\Big(B^{(k)} - W^{\mathrm{pre}} - \mathrm{SVD}_s(\overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger)\hat{A}^\dagger\underline{W}^{\mathrm{pre}}_{\ell-1}\Big)\Sigma^{(k)}_x\\
&\qquad\cdot\Big(B^{(k)} - W^{\mathrm{pre}} - \mathrm{SVD}_s(\overline{W}^{\mathrm{pre}}_{\ell+1}(\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\hat{A}^\dagger)\hat{A}^\dagger\underline{W}^{\mathrm{pre}}_{\ell-1}\Big)^\top\bigg).
\end{aligned}$$

This concludes the proof. □

E.5.2 In-distribution Excess Risk of LoRA

Let $\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{full}}_\ell)$ denote the excess risk of f^{pre} after fine-tuning all the parameters of the ℓ-th layer under *population* in-distribution risk.

Theorem E.10 (Restatement of Theorem E.7: LoRA Part)**.** *Suppose that Assumptions E.1, E.2 and E.3 hold. Then, the following holds with probability* $1-\exp\big(-\Omega(\log^2(n+d+p))\big)$*. For any* $\eta>0$*,*

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell,U^{\mathrm{LoRA}},V^{\mathrm{LoRA}}}) \le (1+\eta)(T^{\mathrm{LoRA}}_{\mathrm{bias}})^2 + (1+\eta^{-1})(T^{\mathrm{LoRA}}_{\mathrm{variance}})^2,$$

where

$$(T^{\mathrm{LoRA}}_{\mathrm{bias}})^2 \le \frac{0\vee(\operatorname{rank}(D\Sigma^{(\mathrm{i})}_x D^\top)-s)}{\operatorname{rank}(D\Sigma^{(\mathrm{i})}_x D^\top)}\kappa_*^2(D\Sigma^{(\mathrm{i})}_x D^\top)\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}) + \mathcal{E}^{(\mathrm{i})}(f^{\mathrm{full}}_\ell), \tag{11}$$

$$\begin{aligned}
(T^{\mathrm{LoRA}}_{\mathrm{variance}})^2 &\lesssim C^2\kappa_*^4(M)\|\Sigma^{(\mathrm{i})}_\epsilon\|_{\mathrm{op}}\kappa_*^2(A)\frac{s(r_e(\Phi'^\top\Sigma^{(\mathrm{i})}_\epsilon\Phi') + r_e(A^2))\log^2(n+d+p)}{n}\\
&\quad + C^2\kappa_*^4(M)\|D\Sigma^{(\mathrm{i})}_x D^\top\|_{\mathrm{op}}\frac{s(\kappa_*^2(A)r_e(\Phi'^\top D\Sigma^{(\mathrm{i})}_x D^\top\Phi') + \kappa_*^6(A)r_e(A^2))\log^2(n+d+p)}{n}.
\end{aligned}$$

Note that the first term on the right hand side of (11) depends on the rank of residual matrix $\Sigma^{(\mathrm{i})}_f = D\Sigma^{(\mathrm{i})}_x D^\top$. It becomes zero when $\operatorname{rank}(\Sigma^{(\mathrm{i})}_f)\le s$ and small when $s/\operatorname{rank}(\Sigma^{(\mathrm{i})}_f)\approx 1$.

Proof of Theorem E.10. Let $\overline{W}_\ell^{\text{LoRA}} := \overline{W}_{\ell+1}^{\text{pre}} U^{\text{LoRA}} V^{\text{LoRA}\top}$. From Lemma E.9, we have

$$\begin{aligned}\mathcal{E}^{(\text{i})}(f_{\ell, U^{\text{LoRA}}, V^{\text{LoRA}}}) &= \operatorname{tr}\Big((D - \overline{W}_\ell^{\text{LoRA}} \underline{W}_{\ell-1}^{\text{pre}})\Sigma_x^{(\text{i})}(D - \overline{W}_\ell^{\text{LoRA}} \underline{W}_{\ell-1}^{\text{pre}})^\top\Big) \\ &= \|(\overline{W}_\ell^{\text{LoRA}} A A^\dagger \underline{W}_{\ell-1}^{\text{pre}} - D)\Sigma_x^{(\text{i})1/2}\|_{\text{F}}^2,\end{aligned}$$

where we used $(I - AA^\dagger)\underline{W}_{\ell-1}^{\text{pre}} \Sigma_x^{(\text{i})1/2} = 0$. From Lemma E.9

$$\overline{W}_\ell^{\text{LoRA}} A = \text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} \hat{A}^\dagger)\hat{A}^\dagger A.$$

This gives

$$\begin{aligned}\|(\overline{W}_\ell^{\text{LoRA}} A A^\dagger \underline{W}_{\ell-1}^{\text{pre}} - D)\Sigma_x^{(\text{i})1/2}\|_{\text{F}} &\le \|(\overline{W}_\ell^{\text{LoRA}} A - \text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger))A^\dagger \underline{W}_{\ell-1}^{\text{pre}} \Sigma_x^{(\text{i})1/2}\|_{\text{F}} \\ &\quad + \|\text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger)A^\dagger \underline{W}_{\ell-1}^{\text{pre}} \Sigma_x^{(\text{i})1/2} - D\Sigma_x^{(\text{i})1/2}\|_{\text{F}} \\ &=: T_{\text{variance}}^{\text{LoRA}} + T_{\text{bias}}^{\text{LoRA}}.\end{aligned}$$

We bound $T_{\text{variance}}^{\text{LoRA}}$ and $T_{\text{bias}}^{\text{LoRA}}$ separately.

Bound $T_{\text{variance}}^{\text{LoRA}}$. For the term $T_{\text{variance}}^{\text{LoRA}}$, since $A^\dagger \underline{W}_{\ell-1}^{\text{pre}} \Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger = A^\dagger A^2 A^\dagger$,

$$T_{\text{variance}}^{\text{LoRA}} = \|\text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} \hat{A}^\dagger)\hat{A}^\dagger A - \text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger)A^\dagger A\|_{\text{F}}.$$

Therefore,

$$\begin{aligned}T_{\text{variance}}^{\text{LoRA}} &\le \|\text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger)A^\dagger A - \text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} \hat{A}^\dagger)A^\dagger A\|_{\text{F}} \\ &\quad + \|\text{SVD}_s(\overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} \hat{A}^\dagger)(\hat{A}^\dagger A - A^\dagger A)\|_{\text{F}} \\ &=: T_{\text{variance},1}^{\text{LoRA}} + T_{\text{variance},2}^{\text{LoRA}},\end{aligned}$$

We first bound $T_{\text{variance},1}^{\text{LoRA}}$. From Lemma F.1 and Assumption E.3, we have

$$\begin{aligned}T_{\text{variance},1}^{\text{LoRA}} &\le \|\text{SVD}_s(\hat{M}) - \text{SVD}_s(M)\|_{\text{F}} \\ &\le \kappa_*^2(M)\frac{\lambda_s(M)}{\lambda_s(M) - \lambda_{s+1}(M)}\sqrt{s}\|\hat{M} - M\|_{\text{op}} \\ &\le \kappa_*^2(M) C\sqrt{s}\|\hat{M} - M\|_{\text{op}},\end{aligned}$$

where $\hat{M} = \overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} \hat{A}^\dagger$ and $M = \overline{W}_{\ell+1}^{\text{pre}} (\overline{W}_{\ell+1}^{\text{pre}})^\dagger D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} A^\dagger$. From Lemma F.3,

$$\begin{aligned}\|\hat{M} - M\|_{\text{op}} &\le \|\Phi'^\top \hat{D}\hat{\Sigma}_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top} - \Phi'^\top D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top}\|_{\text{op}} \|\hat{A}^\dagger\|_{\text{op}} \\ &\quad + \|D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top}\|_{\text{op}} \|\hat{A}^\dagger - A^\dagger\|_{\text{op}} \\ &\lesssim \|\Sigma_\epsilon^{(\text{i})}\|_{\text{op}}^{1/2} \kappa_*(A)\sqrt{\frac{(r_e(\Phi'^\top \Sigma_\epsilon^{(\text{i})}\Phi') + r_e(A^2))\log^2(n+d+p)}{n}} \\ &\quad + \|D\Sigma_x^{(\text{i})} D^\top\|_{\text{op}}^{1/2} \kappa_*(A)\sqrt{\frac{(r_e(\Phi'^\top D\Sigma_x^{(\text{i})} D^\top \Phi') + r_e(A^2))\log^2(n+d+p)}{n}} \\ &\quad + \|D\Sigma_x^{(\text{i})} \underline{W}_{\ell-1}^{\text{pre}\top}\|_{\text{op}} \frac{\kappa_*(A)}{\lambda_*(A)}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}} \\ &\lesssim \|\Sigma_\epsilon^{(\text{i})}\|_{\text{op}}^{1/2} \kappa_*(A)\sqrt{\frac{(r_e(\Phi'^\top \Sigma_\epsilon^{(\text{i})}\Phi') + r_e(A^2))\log^2(n+d+p)}{n}} \\ &\quad + \|D\Sigma_x^{(\text{i})} D^\top\|_{\text{op}}^{1/2} \sqrt{\frac{(\kappa_*^2(A) r_e(\Phi'^\top D\Sigma_x^{(\text{i})} D^\top \Phi') + \kappa_*^4(A) r_e(A^2))\log^2(n+d+p)}{n}}\end{aligned}$$

holds on the event $\mathcal{F}$, where we used $\|D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\|_{\mathrm{op}} \le \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}}\|A\|_{\mathrm{op}}$. Hence

$$T_{\text{variance},1}^{\text{LoRA}} \lesssim C_{\mathrm{g}}\kappa_*^2(M)\|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}^{1/2}\kappa_*(A)\sqrt{\frac{s(r_e(\Phi'^\top\Sigma_\epsilon^{(\mathrm{i})}\Phi') + r_e(A^2))\log^2(n+d+p)}{n}}$$
$$+ C_{\mathrm{g}}\kappa_*^2(M)\|D\Sigma_x^{(\mathrm{i})}D^\top\|_{\mathrm{op}}^{1/2}\sqrt{\frac{s(\kappa_*^2(A)r_e(\Phi'^\top D\Sigma_x^{(\mathrm{i})}D^\top\Phi') + \kappa_*^4(A)r_e(A^2))\log^2(n+d+p)}{n}}.$$

Next we bound $T_{\text{variance},2}^{\text{LoRA}}$. Again from Lemma F.3,

$$\begin{aligned}
T_{\text{variance},2}^{\text{LoRA}} &\le \sqrt{s}\|\hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\|_{\mathrm{op}}\|\hat{A}^\dagger\|_{\mathrm{op}}\|\hat{A}^\dagger - A^\dagger\|_{\mathrm{op}}\|A\|_{\mathrm{op}} \\
&\lesssim \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}}\|\Sigma_x^{(\mathrm{i})1/2}\underline{W}_{\ell-1}^{\mathrm{pre}\top}\|_{\mathrm{op}}\frac{\kappa_*^2(A)}{\lambda_*(A)}\sqrt{\frac{sr_e(A^2)\log^2(n+d+p)}{n}} \\
&= \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}}\kappa_*^3(A)\sqrt{\frac{sr_e(A^2)\log^2(n+d+p)}{n}}
\end{aligned}$$

holds on the event $\mathcal{F}$. Therefore,

$$T_{\text{variance}}^{\text{LoRA}} \lesssim C_{\mathrm{g}}\kappa_*^2(M)\|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}^{1/2}\kappa_*(A)\sqrt{\frac{s(r_e(\Phi'^\top\Sigma_\epsilon^{(\mathrm{i})}\Phi') + r_e(A^2))\log^2(n+d+p)}{n}}$$
$$+ C_{\mathrm{g}}\kappa_*^2(M)\|D\Sigma_x^{(\mathrm{i})}D^\top\|_{\mathrm{op}}^{1/2}\sqrt{\frac{s(\kappa_*^2(A)r_e(\Phi'^\top D\Sigma_x^{(\mathrm{i})}D^\top\Phi') + \kappa_*^6(A)r_e(A^2))\log^2(n+d+p)}{n}} \tag{12}$$

hold with high probability.

Bound $T_{\text{bias}}^{\text{LoRA}}$. Note that

$$\begin{aligned}
(T_{\text{bias}}^{\text{LoRA}})^2 &= \|\mathrm{SVD}_s(M)A^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2} - D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 \\
&= \|\underbrace{\mathrm{SVD}_s(M)A^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2} - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}(A^2)^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2}}_{=:T_1}\|_{\mathrm{F}}^2 \\
&\quad + \|\underbrace{D\Sigma_x^{(\mathrm{i})1/2}(I - \Sigma_x^{(\mathrm{i})1/2}\underline{W}_{\ell-1}^{\mathrm{pre}\top}(A^2)^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2})}_{=:T_2}\|_{\mathrm{F}}^2 \\
&\quad + \|\underbrace{(I - \Phi'\Phi'^\top)D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}(A^2)^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2}}_{=:T_3}\|_{\mathrm{F}}^2
\end{aligned}$$

where the second equality follows from the fact that cross terms are zero, i.e., $\mathrm{tr}(T_1T_2^\top) = \mathrm{tr}(T_2T_3^\top) = \mathrm{tr}(T_3T_1^\top) = 0$ since $\Psi_*(\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2})\Psi_*^\top(\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2}) = \Sigma_x^{(\mathrm{i})1/2}\underline{W}_{\ell-1}^{\mathrm{pre}\top}(A^2)^\dagger\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2}$ and

$$(I - \Phi'\Phi'^\top)\Phi_*(\mathrm{SVD}_s(M)) = 0,\ \ \underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2}(I - \Psi_*(\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2})\Psi_*^\top(\underline{W}_{\ell-1}^{\mathrm{pre}}\Sigma_x^{(\mathrm{i})1/2})) = 0$$

hold. Thus from Lemma E.17,

$$(T_{\text{bias}}^{\text{LoRA}})^2 = \|\mathrm{SVD}_s(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger) - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger\|_{\mathrm{F}}^2 + \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}). \tag{13}$$

Notice that

$$\begin{aligned}
&\|\mathrm{SVD}_s(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger) - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger\|_{\mathrm{F}}^2 \\
&\quad \le \{0 \vee (\mathrm{rank}(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger) - s)\}\|\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger\|_{\mathrm{op}}^2 \\
&\quad \le \{0 \vee (\mathrm{rank}(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\mathrm{pre}\top}A^\dagger) - s)\}\|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}}^2 \\
&\quad \le \frac{0 \vee (\mathrm{rank}(D\Sigma_x^{(\mathrm{i})}D^\top) - s)}{\mathrm{rank}(D\Sigma_x^{(\mathrm{i})1/2})}\kappa_*^2(D\Sigma_x^{(\mathrm{i})}D^\top)\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}),
\end{aligned} \tag{14}$$

where the last inequality follows since

$$\|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 = \|\Lambda_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{F}}^2 \ge \mathrm{rank}(D\Sigma_x^{(\mathrm{i})1/2})\lambda_*^2(D\Sigma_x^{(\mathrm{i})1/2}) = \frac{\mathrm{rank}(D\Sigma_x^{(\mathrm{i})1/2})}{\kappa_*^2(D\Sigma_x^{(\mathrm{i})1/2})}\|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}}^2.$$

Summary Note that for any $\eta > 0$, $(T^{\text{LoRA}}_{\text{variance}} + T^{\text{LoRA}}_{\text{bias}})^2 \le (1+\eta)(T^{\text{LoRA}}_{\text{bias}})^2 + (1+1/\eta)(T^{\text{LoRA}}_{\text{variance}})^2$ holds. Therefore,

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell,U^{\text{LoRA}},V^{\text{LoRA}}}) \le (1+\eta)(T^{\text{LoRA}}_{\text{bias}})^2 + (1+\eta^{-1})(T^{\text{LoRA}}_{\text{variance}})^2.$$

Combined with (12), (13), and (14), this concludes the proof. $\square$

E.5.3 Out-of-distribution Excess Risk of LoRA

We define the low-rank matrix obtained by LoRA under population in-distribution risk as

$$(U^{\text{LoRA}}_\infty, V^{\text{LoRA}}_\infty) \in \operatorname*{arg\,min}_{U,V} \|(U,V)\|_{\mathrm{F}}^2 \quad \text{s.t. } (U,V) \text{ minimizes } \mathcal{R}^{(\mathrm{i})}(f_{\ell,U,V}). \tag{15}$$

Theorem E.11 (Restatement of Theorem E.8: LoRA Part)**.** *For $(U^{\text{LoRA}}_\infty, V^{\text{LoRA}}_\infty)$, defined in (15)*

$$\begin{aligned}\mathcal{E}^{(\mathrm{o})}(f_{\ell,U^{\text{LoRA}}_\infty,V^{\text{LoRA}}_\infty}) &\lesssim \|(I - \Phi'\Phi'^\top)B^{(\mathrm{o})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}}^2 + \|(B^{(\mathrm{o})} - B^{(\mathrm{i})})\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2\|G^{(\mathrm{i},\mathrm{o})}_{\ell-1}\|_{\mathrm{op}}^2 \\ &+ \|(B^{(\mathrm{o})} - W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2}G^{(\mathrm{i},\mathrm{o})}_{\ell-1})\|_{\mathrm{F}} \\ &+ \frac{0 \vee (\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D^\top) - s)}{\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D^\top)}\kappa_*^2(D\Sigma_x^{(\mathrm{i})}D^\top)\|G^{(\mathrm{i},\mathrm{o})}_{\ell-1}\|_{\mathrm{op}}^2\mathcal{E}^{(\mathrm{i})}(f^{\text{pre}}).\end{aligned}$$

Furthermore, for any $\eta \in (0,1)$,

$$\begin{aligned}\mathcal{E}^{(\mathrm{o})}(f_{\ell,U^{\text{LoRA}}_\infty,V^{\text{LoRA}}_\infty}) \ge (1-\eta)&\left\|(B^{(\mathrm{o})} - B^{(\mathrm{i})})\Sigma_x^{(\mathrm{o})1/2}\right\|_{\mathrm{F}}^2 - 3(\eta^{-1}-1)\|(I - \Phi'\Phi'^\top)B^{(\mathrm{i})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}}^2 \\ &- 3(\eta^{-1}-1)\|(B^{(\mathrm{i})} - W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2}G^{(\mathrm{i},\mathrm{o})}_{\ell-1})\|_{\mathrm{F}}^2 \\ &- 3(\eta^{-1}-1)\frac{0 \vee (\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D^\top) - s)}{\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D)}\kappa_*^2(D\Sigma_x^{(\mathrm{i})}D^\top)\|G^{(\mathrm{i},\mathrm{o})}_{\ell-1}\|_{\mathrm{op}}\mathcal{E}^{(\mathrm{i})}(f^{\text{pre}}).\end{aligned} \tag{16}$$

Proof of Theorem E.11. With a slight modification to the proof of Lemma E.9, it follows that

$$\begin{aligned}\mathcal{E}^{(\mathrm{o})}(f_{\ell,U^{\text{LoRA}}_\infty,V^{\text{LoRA}}_\infty}) &= \operatorname{tr}\Bigg(\Big(B^{(\mathrm{o})} - W^{\text{pre}} - \mathrm{SVD}_s(\overline{W}^{\text{pre}}_{\ell+1}(\overline{W}^{\text{pre}}_{\ell+1})^\dagger D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}A^\dagger)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Big)\Sigma_x^{(\mathrm{o})} \\ &\qquad \cdot \Big(B^{(\mathrm{o})} - W^{\text{pre}} - \mathrm{SVD}_s(\overline{W}^{\text{pre}}_{\ell+1}(\overline{W}^{\text{pre}}_{\ell+1})^\dagger D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}A^\dagger)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Big)^\top\Bigg) \\ &= \left\|(B^{(\mathrm{o})} - W^{\text{pre}})\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}A^\dagger)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\right\|_{\mathrm{F}}^2.\end{aligned} \tag{17}$$

Recall that $M := \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}A^\dagger$. Then,

$$\begin{aligned}&\left\|(B^{(\mathrm{o})} - W^{\text{pre}})\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}A^\dagger)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\right\|_{\mathrm{F}} \\ &\le \left\|(B^{(\mathrm{o})} - W^{\text{pre}})\Sigma_x^{(\mathrm{o})1/2} - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}^{\text{pre}\top}_{\ell-1}(A^2)^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\right\|_{\mathrm{F}} \\ &\quad + \|MA^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(M)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} \\ &= \left\|(B^{(\mathrm{o})} - W^{\text{pre}})\Sigma_x^{(\mathrm{o})1/2} - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})1/2}(\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2})^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\right\|_{\mathrm{F}} \\ &\quad + \|MA^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(M)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} \\ &\le \|(I - \Phi'\Phi'^\top)B^{(\mathrm{o})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} + \|\Phi'\Phi'^\top(B^{(\mathrm{o})} - B^{(\mathrm{i})})\Sigma_x^{(\mathrm{i})1/2}G^{(\mathrm{i},\mathrm{o})}_{\ell-1}\|_{\mathrm{F}} \\ &\quad + \|\Phi'\Phi'^\top(B^{(\mathrm{o})} - W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2}G^{(\mathrm{i},\mathrm{o})}_{\ell-1})\|_{\mathrm{F}} \\ &\quad + \|MA^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(M)A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} \\ &\le \|(I - \Phi'\Phi'^\top)B^{(\mathrm{o})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} + \|(B^{(\mathrm{o})} - B^{(\mathrm{i})})\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}\|G^{(\mathrm{i},\mathrm{o})}_{\ell-1}\|_{\mathrm{op}} \\ &\quad + \|(B^{(\mathrm{o})} - W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2}G^{(\mathrm{i},\mathrm{o})}_{\ell-1})\|_{\mathrm{F}} + \|M - \mathrm{SVD}_s(M)\|_{\mathrm{F}}\|A^\dagger\underline{W}^{\text{pre}}_{\ell-1}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{op}},\end{aligned}$$

where we used $\Phi'\Phi'^\top W^{\text{pre}} = W^{\text{pre}}$. From (14), we have

$$
\begin{aligned}
\{\mathcal{E}^{(\mathrm{o})}(f_{\ell,U_\infty^{\text{LoRA}},V_\infty^{\text{LoRA}}})\}^{1/2} \le{}& \|(I-\Phi'\Phi'^\top)B^{(\mathrm{o})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}} + \|(B^{(\mathrm{o})}-B^{(\mathrm{i})})\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}\|G_{\ell-1}^{(\mathrm{i,o})}\|_{\mathrm{op}} \\
&+ \|(B^{(\mathrm{o})}-W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2}-\Sigma_x^{(\mathrm{i})1/2}G_{\ell-1}^{(\mathrm{i,o})})\|_{\mathrm{F}} \\
&+ \|G_{\ell-1}^{(\mathrm{i,o})}\|_{\mathrm{op}}\kappa_*(D\Sigma_x^{(\mathrm{i})}D^\top)\sqrt{\frac{0\vee(\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D^\top)-s)}{\operatorname{rank}(D\Sigma_x^{(\mathrm{i})1/2})}\mathcal{E}^{(\mathrm{i})}(f^{\text{pre}})},
\end{aligned}
$$

where we used $\|A^\dagger \underline{W}_{\ell-1}^{\text{pre}}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{op}} = \|G_{\ell-1}^{(\mathrm{i,o})}\|_{\mathrm{op}}$. This gives the first claim.

Using $2\operatorname{tr}(AB^\top) \ge -\eta\|A\|_{\mathrm{F}}^2 - (1/\eta)\|B\|_{\mathrm{F}}^2$ for any $\eta>0$ and any matrices A, B of the same shape, (17) can be rewritten as

$$
\begin{aligned}
\mathcal{E}^{(\mathrm{o})}(f_{\ell,U_\infty^{\text{LoRA}},V_\infty^{\text{LoRA}}}) ={}& \Big\|(B^{(\mathrm{o})}-B^{(\mathrm{i})})\Sigma_x^{(\mathrm{o})1/2} + \underbrace{(I-\Phi'\Phi'^\top)(B^{(\mathrm{i})}-W^{\text{pre}})\Sigma_x^{(\mathrm{o})1/2}}_{=:T_1} \\
&+ \underbrace{\Phi'\Phi'^\top(B^{(\mathrm{i})}-W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2}-\Sigma_x^{(\mathrm{i})1/2}G_{\ell-1}^{(\mathrm{i,o})})}_{=:T_2} \\
&+ \underbrace{MA^\dagger\underline{W}_{\ell-1}^{\text{pre}}\Sigma_x^{(\mathrm{o})1/2} - \mathrm{SVD}_s(M)A^\dagger\underline{W}_{\ell-1}^{\text{pre}}\Sigma_x^{(\mathrm{o})1/2}}_{=:T_3}\Big\|_{\mathrm{F}}^2 \\
={}& \Big\|(B^{(\mathrm{o})}-B^{(\mathrm{i})})\Sigma_x^{(\mathrm{o})1/2}\Big\|_{\mathrm{F}}^2 + 2\operatorname{tr}\Big((B^{(\mathrm{o})}-B^{(\mathrm{i})})\Sigma_x^{(\mathrm{o})1/2}(T_1+T_2+T_3)^\top\Big) \\
&+ \|T_1+T_2+T_3\|_{\mathrm{F}}^2 \\
\ge{}& (1-\eta)\Big\|(B^{(\mathrm{o})}-B^{(\mathrm{i})})\Sigma_x^{(\mathrm{o})1/2}\Big\|_{\mathrm{F}}^2 + (1-\eta^{-1})\|T_1+T_2+T_3\|_{\mathrm{F}}^2. \qquad (18)
\end{aligned}
$$

Choose $\eta\in(0,1)$. By a similar argument as above, and using $\Phi'\Phi'^\top W^{\text{pre}} = W^{\text{pre}}$, we can show that

$$
\begin{aligned}
\|T_1+T_2+T_3\|_{\mathrm{F}}^2 \le{}& 3\|T_1\|_{\mathrm{F}}^2 + 3\|T_2\|_{\mathrm{F}}^2 + 3\|T_3\|_{\mathrm{F}}^2 \\
\le{}& 3\|(I-\Phi'\Phi'^\top)B^{(\mathrm{i})}\Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}}^2 + 3\|(B^{(\mathrm{i})}-W^{\text{pre}})(\Sigma_x^{(\mathrm{o})1/2}-\Sigma_x^{(\mathrm{i})1/2}G_{\ell-1}^{(\mathrm{i,o})})\|_{\mathrm{F}}^2 \\
&+ 3\frac{0\vee(\operatorname{rank}(D\Sigma_x^{(\mathrm{i})}D^\top)-s)}{\operatorname{rank}(D\Sigma_x^{(\mathrm{i})1/2})}\kappa_*^2(D\Sigma_x^{(\mathrm{i})}D^\top)\|G_{\ell-1}^{(\mathrm{i,o})}\|_{\mathrm{op}}\mathcal{E}^{(\mathrm{i})}(f^{\text{pre}}),
\end{aligned}
$$

where we used (14) again. This concludes the proof. □

E.6 Proofs for Structured Sparse Fine-tuning

E.6.1 Excess Risk of Structured Sparse Fine-tuning

Lemma E.12 (Excess Risk). *Given $S\subset[d_\ell]$, consider the minimum norm solution*

$$
V^{\mathrm{S^2FT}} \in \operatorname*{arg\,min}_{V\in\mathbb{R}^{d_{\ell-1}\times s}} \|V\|_{\mathrm{F}}^2 \quad \textit{s.t. } V \textit{ minimizes } \mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V}).
$$

Then, the low-rank adaptation matrix satisfies

$$
U_S^{\mathrm{S^2FT}}V^{\mathrm{S^2FT}\top} = U_S^{\mathrm{S^2FT}}(\overline{W}_{\ell+1}^{\text{pre}}U_S^{\mathrm{S^2FT}})^\dagger\hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\text{pre}\top}(\hat{A}^\dagger)^2, \qquad (19)
$$

and

$$
\begin{aligned}
\mathcal{E}^{(k)}(f_{\ell,U_S^{\mathrm{S^2FT}},V^{\mathrm{S^2FT}}}) = \operatorname{tr}\Big(&\Big(B^{(k)} - W^{\text{pre}} - \overline{W}_{\ell+1}^{\text{pre}}U_S^{\mathrm{S^2FT}}(\overline{W}_{\ell+1}^{\text{pre}}U_S^{\mathrm{S^2FT}})^\dagger\hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\text{pre}\top}(\hat{A}^\dagger)^2\underline{W}_{\ell-1}^{\text{pre}}\Big)\Sigma_x^{(k)} \\
&\cdot\Big(B^{(k)} - W^{\text{pre}} - \overline{W}_{\ell+1}^{\text{pre}}U_S^{\mathrm{S^2FT}}(\overline{W}_{\ell+1}^{\text{pre}}U_S^{\mathrm{S^2FT}})^\dagger\hat{D}\hat{\Sigma}_x^{(\mathrm{i})}\underline{W}_{\ell-1}^{\text{pre}\top}(\hat{A}^\dagger)^2\underline{W}_{\ell-1}^{\text{pre}}\Big)^\top\Big)
\end{aligned}
$$

for $k\in\{\mathrm{i},\mathrm{o}\}$.

Proof. Since $\hat{\Sigma}_{x,\epsilon}^{(k)} = (1/n)X^{(k)}E^{(k)\top}$ and $\hat{\Sigma}_x^{(k)} = (1/n)X^{(k)}X^{(k)\top}$, we have $\hat{\Sigma}_{x,\epsilon}^{(k)} = \hat{\Sigma}_x^{(k)}(X^{(k)\top})^\dagger E^{(k)\top} =: \hat{\Sigma}_x^{(k)}\check{\Sigma}_{x,\epsilon}^{(k)}$. Similar to (9), we have

$$\mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V}) = \|\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} V^\top \hat{A} - \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} \hat{A}^\dagger\|_{\mathrm{F}}^2 - \|\hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} \hat{A}^\dagger\|_{\mathrm{F}}^2$$
$$+ \operatorname{tr}\Big(D\hat{\Sigma}_x^{(\mathrm{i})} D^\top\Big) + 2\operatorname{tr}\Big(D\hat{\Sigma}_{x,\epsilon}^{(\mathrm{i})}\Big) + \operatorname{tr}\Big(\hat{\Sigma}_\epsilon^{(\mathrm{i})}\Big).$$

Thus minimizing $\mathcal{R}_n^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V})$ is equivalent to minimizing the norm

$$\begin{aligned} &\|\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} V^\top \hat{A} - \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} \hat{A}^\dagger\|_{\mathrm{F}}^2 \\ &\quad = \|\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} V^\top \hat{A} - \overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} \hat{A}^\dagger\|_{\mathrm{F}}^2 \\ &\quad\quad + \|(I - (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})(\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger)\hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} \hat{A}^\dagger\|_{\mathrm{F}}^2. \end{aligned} \tag{20}$$

By the same argument as in the proof of Lemma E.9, the minimum norm solution $V^{\mathrm{S^2FT}}$ is obtained by

$$V^{\mathrm{S^2FT}} = (\hat{A}^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}} \hat{\Sigma}_x^{(\mathrm{i})} \hat{D}^\top (U_S^{\mathrm{S^2FT}\top} \overline{W}_{\ell+1}^{\mathrm{pre}\top})^\dagger.$$

The excess risk for $k \in \{\mathrm{i}, \mathrm{o}\}$ becomes

$$\begin{aligned} \mathcal{E}^{(k)}(f_{\ell,U_S^{\mathrm{S^2FT}},V^{\mathrm{S^2FT}}}) &= \mathbb{E}\left[\left(B^{(k)}x^{(k)} - \overline{W}_{\ell+1}^{\mathrm{pre}}(W_\ell^{\mathrm{pre}} + U_S^{\mathrm{S^2FT}} V^{\mathrm{S^2FT}\top})\underline{W}_{\ell-1}^{\mathrm{pre}} x^{(k)}\right)^2\right] \\ &= \operatorname{tr}\Bigg(\Big(B^{(k)} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (\hat{A}^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}}\Big)\Sigma_x^{(k)} \\ &\quad \cdot \Big(B^{(k)} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (\hat{A}^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}}\Big)^\top\Bigg). \end{aligned}$$

This concludes the proof. ☐

E.6.2 In-distribution Excess Risk of Structured Sparse Fine-tuning

Theorem E.13 (Restatement of Theorem E.7: S²FT Part)**.** *Suppose that Assumptions E.1 and E.2 hold. Fix $S \subset [d_\ell]$ with $|S| = s$. Then, the following holds with probability $1 - \exp\big(-\Omega(\log^2(n+d+p))\big)$. For any $\eta > 0$,*

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V^{\mathrm{S^2FT}}}) \le (1+\eta)(T_{\mathrm{bias}}^{\mathrm{S^2FT}})^2 + (1+\eta^{-1})(T_{\mathrm{variance}}^{\mathrm{S^2FT}})^2,$$

where $T_{\mathrm{bias}}^{\mathrm{S^2FT}} \ge T_{\mathrm{bias}}^{\mathrm{LoRA}}$ and

$$\begin{aligned} (T_{\mathrm{bias}}^{\mathrm{S^2FT}})^2 &\le \|(\Phi'\Phi'^\top - \Phi''_S\Phi''^\top_S)\Phi_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}^2 \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{pre}}) + \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}), \\ (T_{\mathrm{variance}}^{\mathrm{S^2FT}})^2 &\lesssim \|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}} \kappa_*^2(A) \frac{s(r_e(\Phi''^\top_S \Sigma_\epsilon^{(\mathrm{i})} \Phi''_S) + r_e(A^2))\log^2(n+d+p)}{n} \\ &\quad + \|D\Sigma_x^{(\mathrm{i})} D^\top\|_{\mathrm{op}} \frac{s(\kappa_*^2(A) r_e(\Phi''^\top_S D\Sigma_x^{(\mathrm{i})} D^\top \Phi''_S) + \kappa_*^8(A) r_e(A^2))\log^2(n+d+p)}{n}. \end{aligned} \tag{21}$$

Note that the term $\|(\Phi'\Phi'^\top - \Phi''_S\Phi''^\top_S)\Phi_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}$ in (21) measures the distance between subspaces spanned by Φ' and Φ''_S in a label space, weighted by $\Phi_*(\Sigma_f^{(\mathrm{i})})$. In high level, this quantity shows the closeness between the ℓ-th layer full fine-tuning and S²FT. It takes small values when the important 'heads' for residual prediction are sparsely distributed among all heads. This aligns with the intuition that S²FT only selectively fine-tunes small number of coordinates, and thus relying on the information contained in those coordinates.

Proof of Theorem E.13. By the same argument as in the proof of Theorem E.10 combined with Lemma E.12, we have

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell,U_S^{\mathrm{S^2FT}},V^{\mathrm{S^2FT}}}) = \|(\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} V^{\mathrm{S^2FT}\top} A A^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} - D)\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2,$$

and

$$
\begin{aligned}
&\|(\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S V^{\mathrm{S^2FT}\top} A A^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1} - D)\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}} \\
&\quad \le \|\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger (\hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} (\hat{A}^2)^\dagger - D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} (A^2)^\dagger) A\|_{\mathrm{F}} \\
&\quad\quad + \|\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} (A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1} \Sigma_x^{(\mathrm{i})1/2} - D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}} \\
&\quad =: T^{\mathrm{S^2FT}}_{\mathrm{variance}} + T^{\mathrm{S^2FT}}_{\mathrm{bias}}.
\end{aligned}
$$

We bound $T^{\mathrm{S^2FT}}_{\mathrm{variance}}$ and $T^{\mathrm{S^2FT}}_{\mathrm{bias}}$ separately.

Bound $T^{\mathrm{S^2FT}}_{\mathrm{variance}}$. Note that

$$
\begin{aligned}
T^{\mathrm{S^2FT}}_{\mathrm{variance}} &= \|\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} (\hat{A}^\dagger)^2 A - \overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger\|_{\mathrm{F}} \\
&\le \|\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger - \overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger\|_{\mathrm{F}} \\
&\quad + \|\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S (\overline{W}^{\mathrm{pre}}_{\ell+1} U^{\mathrm{S^2FT}}_S)^\dagger \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} ((\hat{A}^\dagger)^2 - (A^\dagger)^2) A\|_{\mathrm{F}} \\
&=: T^{\mathrm{S^2FT}}_{\mathrm{variance},1} + T^{\mathrm{S^2FT}}_{\mathrm{variance},2}.
\end{aligned}
$$

For the term $T^{\mathrm{S^2FT}}_{\mathrm{variance},1}$, using Lemma F.3,

$$
\begin{aligned}
T^{\mathrm{S^2FT}}_{\mathrm{variance},1} &\le 2\sqrt{s}\|\Phi''^\top_S D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} - \Phi''^\top_S \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \|A^\dagger\|_{\mathrm{op}} \\
&\lesssim \|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}^{1/2} \kappa_*(A) \sqrt{\frac{s(r_e(\Phi''^\top_S \Sigma_\epsilon^{(\mathrm{i})} \Phi''_S) + r_e(A^2)) \log^2(n+d+p)}{n}} \\
&\quad + \|D\Sigma_x^{(\mathrm{i})} D^\top\|_{\mathrm{op}}^{1/2} \kappa_*(A) \sqrt{\frac{s(r_e(\Phi''^\top_S D\Sigma_x^{(\mathrm{i})} D^\top \Phi''_S) + r_e(A^2)) \log^2(n+d+p)}{n}}
\end{aligned}
$$

holds on the event $\mathcal{F}$, where the first inequality follows since the term inside the norm is at most rank-$2s$. Again from Lemma F.3,

$$
\begin{aligned}
T^{\mathrm{S^2FT}}_{\mathrm{variance},2} &\le \sqrt{s}\|\Phi''^\top_S \hat{D}\hat{\Sigma}_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \|(\hat{A}^\dagger)^2 - (A^\dagger)^2\|_{\mathrm{op}} \|A\|_{\mathrm{op}} \\
&\lesssim \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}} \|\Sigma_x^{(\mathrm{i})1/2} \underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \frac{\kappa_*^3(A)}{\lambda_*(A)} \sqrt{\frac{s r_e(A^2) \log^2(n+d+p)}{n}} \\
&= \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}} \kappa_*^4(A) \sqrt{\frac{s r_e(A^2) \log^2(n+d+p)}{n}}
\end{aligned}
$$

holds on the event $\mathcal{F}$. Therefore,

$$
\begin{aligned}
T^{\mathrm{S^2FT}}_{\mathrm{variance}} &\lesssim \|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}^{1/2} \kappa_*(A) \sqrt{\frac{s(r_e(\Phi''^\top_S \Sigma_\epsilon^{(\mathrm{i})} \Phi''_S) + r_e(A^2)) \log^2(n+d+p)}{n}} \\
&\quad + \|D\Sigma_x^{(\mathrm{i})} D^\top\|_{\mathrm{op}}^{1/2} \sqrt{\frac{s(\kappa_*^2(A) r_e(\Phi''^\top_S D\Sigma_x^{(\mathrm{i})} D^\top \Phi''_S) + \kappa_*^8(A) r_e(A^2)) \log^2(n+d+p)}{n}}.
\end{aligned} \tag{22}
$$

Bound $T^{\mathrm{S^2FT}}_{\mathrm{bias}}$. By the same argument as in the proof of Theorem E.10,

$$
\begin{aligned}
(T^{\mathrm{S^2FT}}_{\mathrm{bias}})^2 &= \|\Phi''_S \Phi''^\top_S D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger\|_{\mathrm{F}}^2 + \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) \qquad (23) \\
&\le \|(\Phi''_S \Phi''^\top_S - \Phi'\Phi'^\top)\Phi_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}^2 \|D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger\|_{\mathrm{F}}^2 + \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) \\
&= \|(\Phi''_S \Phi''^\top_S - \Phi'\Phi'^\top)\Phi_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}^2 \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{pre}}) + \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}), \qquad (24)
\end{aligned}
$$

where we used $\|D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger\|_{\mathrm{F}}^2 \le \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 = \mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{pre}})$. We can verify $T^{\mathrm{S^2FT}}_{\mathrm{bias}} \ge T^{\mathrm{LoRA}}_{\mathrm{bias}}$ by comparing (13) and (23), since $\mathrm{SVD}_s(\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger)$ is the best rank-s approximation of $\Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger$ and $\Phi''_S \Phi''^\top_S D\Sigma_x^{(\mathrm{i})} \underline{W}^{\mathrm{pre}\top}_{\ell-1} A^\dagger$ is at most rank-s.

Summary Note that for any $\eta > 0$, $(T^{\mathrm{S^2FT}}_{\text{variance}} + T^{\mathrm{S^2FT}}_{\text{bias}})^2 \le (1+\eta)(T^{\mathrm{S^2FT}}_{\text{bias}})^2 + (1+1/\eta)(T^{\mathrm{S^2FT}}_{\text{variance}})^2$ holds. Thus

$$\mathcal{E}^{(\mathrm{i})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V^{\mathrm{S^2FT}}}) \le (1+\eta)(T^{\mathrm{S^2FT}}_{\text{bias}})^2 + (1+\eta^{-1})(T^{\mathrm{S^2FT}}_{\text{variance}})^2.$$

Combined with (22) and (24), this concludes the proof. $\square$

Next we characterize the bias terms $T^{\mathrm{LoRA}}_{\text{bias}}$ and $T^{\mathrm{S^2FT}}_{\text{bias}}$ under sparsity assumption.

Lemma E.14. *Suppose that Assumption E.4 holds. Then, for a sparse fine-tuned network with the choice $S \supset S_0$, it follows that*

$$\mathcal{E}^{(\mathrm{i})}(f_\ell^{\text{full}}) \le (T^{\mathrm{LoRA}}_{\text{bias}})^2 \le (T^{\mathrm{S^2FT}}_{\text{bias}})^2 \le \mathcal{E}^{(\mathrm{i})}(f_\ell^{\text{full}}) + \delta^2 \kappa_*^2(\overline{W}^{\text{pre}}_{\ell+1}) \mathcal{E}^{(\mathrm{i})}(f^{\text{pre}}).$$

Proof. Note that $\Phi''_S \Phi''^{\top}_S$ is a projection into a subspace, which is contained in a subspace projected by $\Phi' \Phi'^{\top}$. Thus

$$\begin{aligned}
&\|\Phi''_S \Phi''^{\top}_S D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger - \Phi' \Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2 \\
&\quad = \|(\Phi''_S \Phi''^{\top}_S - I)\Phi' \Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2 \\
&\quad = \|(\Phi''_S \Phi''^{\top}_S - I)\overline{W}^{\text{pre}}_{\ell+1} (\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2 \\
&\quad = \|(\Phi''_S \Phi''^{\top}_S - I)\overline{W}^{\text{pre}}_{\ell+1} ((I - U_S^{\mathrm{S^2FT}} U_S^{\mathrm{S^2FT}\top}) + U_S^{\mathrm{S^2FT}} U_S^{\mathrm{S^2FT}\top}) (\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2 \\
&\quad = \|(\Phi''_S \Phi''^{\top}_S - I)\overline{W}^{\text{pre}}_{\ell+1} (I - U_S^{\mathrm{S^2FT}} U_S^{\mathrm{S^2FT}\top}) (\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2,
\end{aligned}$$

where the last equality follows since $(\Phi''_S \Phi''^{\top}_S - I)\overline{W}^{\text{pre}}_{\ell+1} U_S^{\mathrm{S^2FT}} = 0$ by definition of $\Phi''_S = \Phi_*(\overline{W}^{\text{pre}}_{\ell+1} U_S^{\mathrm{S^2FT}})$. Thus

$$\begin{aligned}
&\|\Phi''_S \Phi''^{\top}_S D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger - \Phi' \Phi'^{\top} D \Sigma_x^{(\mathrm{i})} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger \|_{\mathrm{F}}^2 \\
&\quad \le \|\overline{W}^{\text{pre}}_{\ell+1}\|_{\text{op}}^2 \|(I - U_S^{\mathrm{S^2FT}} U_S^{\mathrm{S^2FT}\top})(\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 \|\Sigma_x^{(\mathrm{i})1/2} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger\|_{\text{op}}^2 \\
&\quad = \|\overline{W}^{\text{pre}}_{\ell+1}\|_{\text{op}}^2 \|\Sigma_x^{(\mathrm{i})1/2} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger\|_{\text{op}}^2 \sum_{a \in [d_\ell] \setminus S} \|e_a^\top (\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})1/2}\|^2 \\
&\quad \le \delta^2 \|\overline{W}^{\text{pre}}_{\ell+1}\|_{\text{op}}^2 \|(\overline{W}^{\text{pre}}_{\ell+1})^\dagger D \Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 \\
&\quad \le \delta^2 \kappa_*^2(\overline{W}^{\text{pre}}_{\ell+1}) \|D \Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the second inequality follows from $\|\Sigma_x^{(\mathrm{i})1/2} \underline{W}^{\text{pre}\top}_{\ell-1} A^\dagger\|_{\text{op}} \le 1$, Assumption E.4 and $S \supset S_0$. The conclusion follows from (13) and (23). $\square$

E.6.3 Out-of-distribution Excess Risk of Structured Sparse Fine-tuning

Given $S \subset [d_\ell]$ with $|S| = s$, we define the low-rank adaptation matrix obtained by S^{2}FT under population in-distribution risk as

$$V_\infty^{\mathrm{S^2FT}} = \arg\min_V \|V\|_{\mathrm{F}}^2 \quad \text{s.t. } V \text{ minimizes } \mathcal{R}^{(\mathrm{i})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V}). \tag{25}$$

Theorem E.15 (Restatement of Theorem E.8: S^{2}FT Part)**.** *Fix $S \subset [d_\ell]$ with $|S| = s$. For $V_\infty^{\mathrm{S^2FT}}$ defined in* (25)*,*

$$\begin{aligned}
\mathcal{E}^{(\mathrm{o})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V_\infty^{\mathrm{S^2FT}}}) \le \mathcal{E}^{(\mathrm{o})}(f^{\text{pre}}) &+ 3\big\|\Phi''_S \Phi''^{\top}_S (B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{(\mathrm{o})1/2}\big\|_{\mathrm{F}}^2 \\
&+ 3\|B^{(\mathrm{i})}(\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i},\mathrm{o})})\|_{\mathrm{F}}^2 \\
&+ 3\|\overline{W}^{\text{pre}}_\ell\|_{\text{op}}^2 \|\underline{W}^{\text{pre}}_{\ell-1} \Sigma_x^{(\mathrm{o})1/2} - \underline{W}^{\text{pre}}_{\ell-1} \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i},\mathrm{o})}\|_{\mathrm{F}}^2.
\end{aligned}$$

Remark E.16. If there is no covariate shift, i.e., $\Sigma_x^{(\mathrm{i})} = \Sigma_x^{(\mathrm{o})} = \Sigma_x$ for some Σ_x, Theorem E.15 further gives the bound

$$\begin{aligned}
\mathcal{E}^{(\mathrm{o})}(f_{\ell, U_S^{\mathrm{S^2FT}}, V_\infty^{\mathrm{S^2FT}}}) \le \mathcal{E}^{(\mathrm{o})}(f^{\text{pre}}) &+ 3\big\|\Phi''_S \Phi''^{\top}_S (B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{1/2}\big\|_{\mathrm{F}}^2 \\
&+ 3\|B^{(\mathrm{i})} \Sigma_x^{1/2}(I - (\underline{W}^{\text{pre}}_{\ell-1} \Sigma_x^{1/2})^\dagger \underline{W}^{\text{pre}}_{\ell-1} \Sigma_x^{1/2}))\|_{\mathrm{F}}^2.
\end{aligned}$$

Proof of Theorem E.15. With a slight modification to Lemma E.12, we obtain

$$\begin{aligned}
\mathcal{E}^{(\mathrm{o})}(f_{\ell,U_S^{\mathrm{S^2FT}},V_\infty^{\mathrm{S^2FT}}}) &= \mathrm{tr}\bigg(\Big(B^{(\mathrm{o})} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}}\Big) \Sigma_x^{(\mathrm{o})} \\
&\qquad \cdot \Big(B^{(\mathrm{o})} - W^{\mathrm{pre}} - \overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}} (\overline{W}_{\ell+1}^{\mathrm{pre}} U_S^{\mathrm{S^2FT}})^\dagger D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}}\Big)^\top\bigg) \\
&= \Big\|(B^{(\mathrm{o})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{o})1/2} - \Phi_S'' \Phi_S''^\top D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2}\Big\|_{\mathrm{F}}^2 \\
&= \|(I - \Phi_S'' \Phi_S''^\top)(B^{(\mathrm{o})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}}^2 \\
&\quad + \|\underbrace{\Phi_S'' \Phi_S''^\top \Big\{(B^{(\mathrm{o})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{o})1/2} - D \Sigma_x^{(\mathrm{i})1/2} (\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2})^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2}\Big\}}_{=:T}\|_{\mathrm{F}}^2,
\end{aligned}$$

where we used $\Sigma_x^{(\mathrm{i})1/2} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^\dagger)^2 \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2} = (\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2})^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2}$. Note that

$$\begin{aligned}
\|T\|_{\mathrm{F}} &\le \Big\|\Phi_S'' \Phi_S''^\top \Big\{B^{(\mathrm{o})} \Sigma_x^{(\mathrm{o})1/2} - B^{(\mathrm{i})} \Sigma_x^{(\mathrm{i})1/2} (\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2})^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2}\Big\}\Big\|_{\mathrm{F}} \\
&\quad + \Big\|\Phi_S'' \Phi_S''^\top \overline{W}_\ell^{\mathrm{pre}} \Big\{\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2} - \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2} (\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2})^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2}\Big\}\Big\|_{\mathrm{F}} \\
&\le \Big\|\Phi_S'' \Phi_S''^\top (B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{(\mathrm{o})1/2}\Big\|_{\mathrm{F}} \\
&\quad + \Big\|\Phi_S'' \Phi_S''^\top B^{(\mathrm{i})} (\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i,o})})\Big\|_{\mathrm{F}} \\
&\quad + \Big\|\Phi_S'' \Phi_S''^\top \overline{W}_\ell^{\mathrm{pre}} (\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2} - \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i,o})})\Big\|_{\mathrm{F}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{E}^{(\mathrm{o})}(f_{\ell,U_S^{\mathrm{S^2FT}},V_\infty^{\mathrm{S^2FT}}}) &= \|(I - \Phi_S'' \Phi_S''^\top)(B^{(\mathrm{o})} - W^{\mathrm{pre}}) \Sigma_x^{(\mathrm{o})1/2}\|_{\mathrm{F}}^2 + \|T\|_{\mathrm{F}}^2 \\
&\le \mathcal{E}^{(\mathrm{o})}(f^{\mathrm{pre}}) + 3\Big\|\Phi_S'' \Phi_S''^\top (B^{(\mathrm{o})} - B^{(\mathrm{i})}) \Sigma_x^{(\mathrm{o})1/2}\Big\|_{\mathrm{F}}^2 \\
&\quad + 3\|B^{(\mathrm{i})} (\Sigma_x^{(\mathrm{o})1/2} - \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i,o})})\|_{\mathrm{F}}^2 \\
&\quad + 3\|\overline{W}_\ell^{\mathrm{pre}}\|_{\mathrm{op}}^2 \|\underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{o})1/2} - \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2} G_{\ell-1}^{(\mathrm{i,o})}\|_{\mathrm{F}}^2,
\end{aligned}$$

where we used $x + y + z \le 3x^2 + 3y^2 + 3z^2$. This concludes the proof. □

E.7 Proofs for Full Fine-tuning

Define $f_\ell^{\mathrm{full}}(x) = \overline{W}_{\ell+1}^{\mathrm{pre}} (W_\ell^{\mathrm{pre}} + \Delta_\ell^{\mathrm{full}}) \underline{W}_{\ell-1}^{\mathrm{pre}} x$ as a fine-tuned network with full fine-tuning applied to the ℓ-th layer, evaluated under the population in-distribution risk, where $\Delta_\ell^{\mathrm{full}}$ is obtained by

$$\Delta_\ell^{\mathrm{full}} \in \operatorname*{arg\,min}_{\Delta' \in \mathbb{R}^{d_\ell \times d_{\ell-1}}} \mathbb{E}\bigg[\Big(B^{(\mathrm{i})} x^{(\mathrm{i})} - \overline{W}_{\ell+1}^{\mathrm{pre}} (W_\ell^{\mathrm{pre}} + \Delta') \underline{W}_{\ell-1}^{\mathrm{pre}} x^{(\mathrm{i})}\Big)^2\bigg].$$

Lemma E.17 (In-distribution Excess Risk). *For f_ℓ^{full}, it holds that*

$$\begin{aligned}
\mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) &= \|D \Sigma_x^{(\mathrm{i})1/2} (I - \Sigma_x^{(\mathrm{i})1/2} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^2)^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{F}}^2 \\
&\quad + \|(I - \Phi' \Phi'^\top) D \Sigma_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top} (A^2)^\dagger \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2.
\end{aligned}$$

Proof of Lemma E.17. Similar to the proof of Theorem E.10, we have

$$\begin{aligned}
\mathcal{E}^{(\mathrm{i})}(f_\ell^{\mathrm{full}}) &= \min_{\Delta \in \mathbb{R}^{d_\ell \times d_{\ell-1}}} \mathbb{E}\bigg[\Big(B^{(\mathrm{i})} x^{(\mathrm{i})} - \overline{W}_{\ell+1}^{\mathrm{pre}} (W_\ell^{\mathrm{pre}} + \Delta) \underline{W}_{\ell-1}^{\mathrm{pre}} x^{(\mathrm{i})}\Big)^2\bigg] \\
&= \min_{\Delta \in \mathbb{R}^{d_\ell \times d_{\ell-1}}} \|D \Sigma_x^{(\mathrm{i})1/2} - \overline{W}_{\ell+1}^{\mathrm{pre}} \Delta \underline{W}_{\ell-1}^{\mathrm{pre}} \Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2,
\end{aligned}$$

and

$$\|D\Sigma_x^{(\mathrm{i})1/2} - \overline{W}^{\mathrm{pre}}_{\ell+1}\Delta \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2 = \|\underbrace{\overline{W}^{\mathrm{pre}}_{\ell+1}\Delta \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2} - \Phi'\Phi'^\top D\Sigma_x^{(\mathrm{i})}\underline{W}^{\mathrm{pre}\top}_{\ell-1}A^\dagger}_{=:T_1}\|_{\mathrm{F}}^2 + \|\underbrace{D\Sigma_x^{(\mathrm{i})1/2}(I - \Sigma_x^{(\mathrm{i})1/2}\underline{W}^{\mathrm{pre}\top}_{\ell-1}(A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2})}_{=:T_2}\|_{\mathrm{F}}^2 + \|\underbrace{(I - \Phi'\Phi'^\top)D\Sigma_x^{(\mathrm{i})}\underline{W}^{\mathrm{pre}\top}_{\ell-1}(A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2}}_{=:T_3}\|_{\mathrm{F}}^2, \tag{26}$$

where we used the fact that the inner products $\mathrm{tr}(T_1T_2^\top) = \mathrm{tr}(T_2T_3^\top) = \mathrm{tr}(T_3T_1^\top) = 0$. By choosing $\Delta = (\overline{W}^{\mathrm{pre}}_{\ell+1})^\dagger D\Sigma_x^{(\mathrm{i})}\underline{W}^{\mathrm{pre}\top}_{\ell-1}A^\dagger$ for example, the term T_1 becomes 0. Thus

$$\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{full}}_\ell) = \|D\Sigma_x^{(\mathrm{i})1/2}(I - \Sigma_x^{(\mathrm{i})1/2}\underline{W}^{\mathrm{pre}\top}_{\ell-1}(A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{F}}^2 + \|(I - \Phi'\Phi'^\top)D\Sigma_x^{(\mathrm{i})}\underline{W}^{\mathrm{pre}\top}_{\ell-1}(A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{F}}^2.$$

This gives the desired result. □

We obtain the following corollary as a direct consequence of Lemma E.17.

Corollary E.18. *For f^{full}_ℓ, it holds that*

$$\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{full}}_\ell) \le \|\Psi_*^\top(D\Sigma_x^{(\mathrm{i})1/2})(I - \Sigma_x^{(\mathrm{i})1/2}\underline{W}^{\mathrm{pre}\top}_{\ell-1}(A^2)^\dagger \underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}) + \|(I - \Phi'\Phi'^\top)\Phi_*(D\Sigma_x^{(\mathrm{i})1/2})\|_{\mathrm{op}}\mathcal{E}^{(\mathrm{i})}(f^{\mathrm{pre}}). \tag{27}$$

The first term on the right hand side of (27) measures the distance between two subspaces spanned by $\Psi_*(D\Sigma_x^{(\mathrm{i})1/2})$ and $\Psi_*(\underline{W}^{\mathrm{pre}}_{\ell-1}\Sigma_x^{(\mathrm{i})1/2})$. Intuitively speaking, this quantifies the information coded at the ℓ-th layer, and the necessary information to predict residuals. Thus, it bounds the maximum improvement by the ℓ-th layer fine-tuning. The second term measures the subspace distance between the subspace where prediction residuals reside, and the subspace predictable by the ℓ-th layer fine-tuning.

F Auxiliary Results for Proofs

Lemma F.1. *Fix $s, d_1, d_2 \in \mathbb{N}^+$. For any $A, B \in \mathbb{R}^{d_1\times d_2}$, if $\|B - A\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}}$ and $\lambda_s(A) > \lambda_{s+1}(A)$ hold, then,*

$$\|SVD_s(B) - SVD_s(A)\|_{\mathrm{F}} \lesssim \kappa_*^2(A)\frac{\lambda_s(A)}{\lambda_s(A) - \lambda_{s+1}(A)}\big(\sqrt{s}\|B - A\|_{\mathrm{op}} \wedge \|B - A\|_{\mathrm{F}}\big).$$

Proof. By triangle inequality,

$$\begin{aligned}\|\mathrm{SVD}_s(B) - \mathrm{SVD}_s(A)\|_{\mathrm{F}} &= \|\Phi_s(B)\Phi_s^\top(B)B - \Phi_s(A)\Phi_s^\top(A)A\|_{\mathrm{F}}\\ &\le \|\Phi_s(B)\Phi_s^\top(B)(B - A)\|_{\mathrm{F}} + \|(\Phi_s(B)\Phi_s^\top(B) - \Phi_s(A)\Phi_s^\top(A))A\|_{\mathrm{F}}\\ &\le \sqrt{s}\|B - A\|_{\mathrm{op}} + \|\Phi_s(B)\Phi_s^\top(B) - \Phi_s(A)\Phi_s^\top(A)\|_{\mathrm{F}}\|A\|_{\mathrm{op}}.\end{aligned}$$

Using Davis-Kahan theorem (Theorem 4 from [70]), and Lemma 2.6 from [11],

$$\|\Phi_s(B)\Phi_s^\top(B) - \Phi_s(A)\Phi_s^\top(A)\|_{\mathrm{F}} \le \frac{6\sqrt{2}\|A\|_{\mathrm{op}}(\sqrt{s}\|B - A\|_{\mathrm{op}} \wedge \|B - A\|_{\mathrm{F}})}{\lambda_s^2(A) - \lambda_{s+1}^2(A)}.$$

Thus

$$\begin{aligned}\|\mathrm{SVD}_s(B) - \mathrm{SVD}_s(A)\|_{\mathrm{F}} &\lesssim \frac{\|A\|_{\mathrm{op}}^2}{\lambda_s^2(A)}\frac{\lambda_s^2(A)}{\lambda_s^2(A) - \lambda_{s+1}^2(A)}(\sqrt{s}\|B - A\|_{\mathrm{op}} \wedge \|B - A\|_{\mathrm{F}})\\ &\lesssim \frac{\|A\|_{\mathrm{op}}^2}{\lambda_s^2(A)}\frac{\lambda_s(A)}{\lambda_s(A) - \lambda_{s+1}(A)}(\sqrt{s}\|B - A\|_{\mathrm{op}} \wedge \|B - A\|_{\mathrm{F}}).\end{aligned}$$

This concludes the proof. □

We cite the concentration inequality for cross-covariance matrices from [47].

Lemma F.2 (Proposition 9.1 from [47])**.** *Let Z and $\tilde{Z}$ be mean zero random vectors taking values in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Denote covariance matrices of Z and $\tilde{Z}$ by Σ_Z and $\Sigma_{\tilde{Z}}$, respectively. Fix any $t > 0$. Assume that there exist constants $c_1, c_2 > 0$ such that*

$$\gamma^\top \Sigma_Z \gamma \geq c_1 \|\gamma^\top Z\|_{\psi_2}^2 \quad \text{and} \quad \gamma'^\top \Sigma_{\tilde{Z}} \gamma' \geq c_2 \|\gamma'^\top \tilde{Z}\|_{\psi_2}^2 \tag{28}$$

holds for any $\gamma \in \mathbb{R}^{d_1}$ and $\gamma' \in \mathbb{R}^{d_2}$. Choose $n \gg (r_e(\Sigma_Z) \wedge r_e(\Sigma_{\tilde{Z}}))(t + \log(d_1 + d_2))$. Let $(Z_i, \tilde{Z}_i)_{i\in[n]}$ be n independent copies of $(Z, \tilde{Z})$. Then, there exists a constant $C = C(c_1, c_2) > 0$ such that with probability at least $1 - e^{-t}$,

$$\left\| \frac{1}{n} \sum_{i\in[n]} Z_i \tilde{Z}_i^\top - \mathbb{E}[Z\tilde{Z}^\top] \right\|_{\mathrm{op}} \leq C \|\Sigma_Z\|_{\mathrm{op}}^{1/2} \|\Sigma_{\tilde{Z}}\|_{\mathrm{op}}^{1/2} \sqrt{\frac{(r_e(\Sigma_Z) + r_e(\Sigma_{\tilde{Z}})(t + \log(d_1 + d_2))}{n}}$$

hold.

Note that if a random variable Z taking values in $\mathbb{R}^d$ satisfies $\gamma^\top \Sigma_Z \gamma \geq c\|\gamma^\top Z\|_{\psi_2}^2$ for any $\gamma \in \mathbb{R}^d$ with some constant $c > 0$, AZ also satisfies $\gamma'^\top \Sigma_{AZ} \gamma' \geq c\|\gamma'^\top AZ\|_{\psi_2}^2$ for any $\gamma' \in \mathbb{R}^{d'}$ and any matrix $A \in \mathbb{R}^{d' \times d}$ and arbitrary $d' \in \mathbb{N}^+$, where $\Sigma_{AZ} = A\Sigma_Z A^\top$.

We then prove the following lemma to show the existance of a 'good' high probability event to bound multiple inequalities.

Lemma F.3. *Suppose that Assumptions E.1 and E.2 hold. Fix any $S \subset [d_\ell]$. Then, there exists an event $\mathcal{F}$ with $\mathbb{P}(\mathcal{F}) = 1 - \exp\big(-\Omega(\log^2(n + d + p))\big)$ such that on the event $\mathcal{F}$, for $\Phi \in \{\Phi', \Phi''_S\}$,*

$$\|\Phi^\top \hat{D} \hat{\Sigma}_x^{(\mathrm{i})} \underline{W}_{\ell-1}^{\mathrm{pre}\top}\|_{\mathrm{op}} \lesssim \|D\Sigma_x^{(\mathrm{i})1/2}\|_{\mathrm{op}} \|A\|_{\mathrm{op}}, \quad \|\hat{A}^\dagger\|_{\mathrm{op}} \lesssim \|A^\dagger\|_{\mathrm{op}}, \tag{29}$$

and

$$\|(\hat{A}^2)^\dagger - (A^2)^\dagger\|_{\mathrm{op}} \lesssim \frac{\kappa_*^2(A)}{\lambda_*^2(A)} \sqrt{\frac{r_e(A^2)\log^2(n + d + p)}{n}}, \tag{30}$$

$$\|\hat{A} - A\|_{\mathrm{op}} \lesssim \kappa_*^2(A) \|A\|_{\mathrm{op}} \sqrt{\frac{r_e(A^2)\log^2(n + d + p)}{n}}, \tag{31}$$

$$\|\hat{A}^\dagger - A^\dagger\|_{\mathrm{op}} \lesssim \frac{\kappa_*(A)}{\lambda_*(A)} \sqrt{\frac{r_e(A^2)\log^2(n + d + p)}{n}} \tag{32}$$

hold. Furthermore,

$$\begin{aligned} &\|\Phi^\top (\hat{D}\hat{\Sigma}_x^{(\mathrm{i})1/2} - D\Sigma_x^{(\mathrm{i})1/2}) \underline{W}_{\ell-1}^{\mathrm{pre}\top}\|_{\mathrm{op}} \\ &\quad \lesssim \|\Sigma_\epsilon^{(\mathrm{i})}\|_{\mathrm{op}}^{1/2} \|A\|_{\mathrm{op}} \sqrt{\frac{(r_e(\Phi^\top \Sigma_\epsilon^{(\mathrm{i})} \Phi) + r_e(A^2))\log^2(n + d + p)}{n}} \\ &\qquad + \|D\Sigma_x^{(\mathrm{i})} D^\top\|_{\mathrm{op}}^{1/2} \|A\|_{\mathrm{op}} \sqrt{\frac{(r_e(\Phi^\top D\Sigma_x^{(\mathrm{i})} D^\top \Phi) + r_e(A^2))\log^2(n + d + p)}{n}} \end{aligned} \tag{33}$$

holds on the event $\mathcal{F}$.

Proof. We only prove for $\Phi = \Phi'$ without loss of generality. Before proving Lemma F.3, we first derive several concentration inequalities. Assumption E.2 implies

$$\begin{aligned} &n \gg r_e(A^2)\log^2(n + d + p), \\ &n \gg r_e(\Sigma_x^{(\mathrm{i})})\log^2(n + d + p), \\ &n \gg (r_e(\Sigma_\epsilon^{(\mathrm{i})}) \wedge r_e(\Sigma_x^{(\mathrm{i})}))\log^2(n + d + p), \\ &n \gg (r_e(\Phi^\top \Sigma_\epsilon^{(\mathrm{i})} \Phi) \wedge r_e(A^2))\log^2(n + d + p), \\ &n \gg (r_e(\Phi^\top D\Sigma_x^{(\mathrm{i})} D^\top \Phi) \wedge r_e(A^2))\log^2(n + d + p). \end{aligned}$$

Using Lemma F.2, we obtain

$$\|\hat{A}^2 - A^2\|_{\text{op}} = \|\underline{W}^{\text{pre}}_{\ell-1}\hat{\Sigma}^{(\text{i})}_x\underline{W}^{\text{pre}\top}_{\ell-1} - \underline{W}^{\text{pre}}_{\ell-1}\Sigma^{(\text{i})}_x\underline{W}^{\text{pre}\top}_{\ell-1}\|_{\text{op}}$$
$$\lesssim \|A\|^2_{\text{op}}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}}, \tag{34}$$

and

$$\|\hat{\Sigma}^{(\text{i})}_{\epsilon,x}\|_{\text{op}} \lesssim \|\Sigma^{(\text{i})}_\epsilon\|^{1/2}_{\text{op}}\|\Sigma^{(\text{i})}_x\|^{1/2}_{\text{op}}\sqrt{\frac{(r_e(\Sigma^{(\text{i})}_\epsilon)+r_e(\Sigma^{(\text{i})}_x))\log^2(n+d+p)}{n}}, \tag{35}$$

$$\|\hat{\Sigma}^{(\text{i})}_x - \Sigma^{(\text{i})}_x\|_{\text{op}} \lesssim \|\Sigma^{(\text{i})}_x\|_{\text{op}}\sqrt{\frac{r_e(\Sigma^{(\text{i})}_x)\log^2(n+d+p)}{n}}, \tag{36}$$

$$\left\|\Phi^\top\hat{\Sigma}^{(\text{i})}_{\epsilon,x}(\Sigma^{(\text{i})}_x)^\dagger\Sigma^{(\text{i})}_x\underline{W}^{\text{pre}\top}_{\ell-1}\right\|_{\text{op}} \lesssim \|\Sigma^{(\text{i})}_\epsilon\|^{1/2}_{\text{op}}\|A\|_{\text{op}}\sqrt{\frac{(r_e(\Phi^\top\Sigma^{(\text{i})}_\epsilon\Phi)+r_e(A^2))\log^2(n+d+p)}{n}}, \tag{37}$$

$$\left\|\Phi^\top D(\hat{\Sigma}^{(\text{i})}_x - \Sigma^{(\text{i})}_x)\underline{W}^{\text{pre}\top}_{\ell-1}\right\|_{\text{op}} \lesssim \|D\Sigma^{(\text{i})}_x D^\top\|^{1/2}_{\text{op}}\|A\|_{\text{op}}\sqrt{\frac{(r_e(\Phi^\top D\Sigma^{(\text{i})}_x D^\top\Phi)+r_e(A^2))\log^2(n+d+p)}{n}}, \tag{38}$$

with high probability. Hereafter we only focus on the event $\mathcal{F}$ where these inequalities hold. We divide the proof into 2 parts.

Part 1. In this part we derive (30), (31) and (32). Note that $\|\hat{A}^2 - A^2\|_{\text{op}} \le \lambda_*(A^2)/2$ holds on the event $\mathcal{F}$ since $n \gg \kappa^4_*(A)r_e(A^2)\log^2(n+d+p)$ by Assumption E.2, and hence $\text{rank}(\hat{A}^2) = \text{rank}(A^2)$. Using Theorem 5.2 from [60],

$$\frac{\|(\hat{A}^2)^\dagger - (A^2)^\dagger\|_{\text{op}}}{\|(A^2)^\dagger\|_{\text{op}}} \lesssim \left(1 - \frac{\kappa_*(A^2)\|\hat{A}^2 - A^2\|_{\text{op}}}{\|A\|^2_{\text{op}}}\right)^{-1}\frac{\kappa_*(A^2)\|\hat{A}^2 - A^2\|_{\text{op}}}{\|A\|^2_{\text{op}}}.$$

Again from Assumption E.2, (34) gives

$$\|(\hat{A}^2)^\dagger - (A^2)^\dagger\|_{\text{op}} \lesssim \frac{\kappa_*(A^2)}{\lambda_*(A^2)}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}}.$$

This yields (30). Proposition 3.2 from [65] and (34) yield,

$$\|(\Phi'''^\top\hat{A}^2\Phi''')^{1/2} - (\Phi'''^\top A^2\Phi''')^{1/2}\|_{\text{op}} \le \frac{\|\Phi'''^\top(\hat{A}^2 - A^2)\Phi'''\|_{\text{op}}}{\lambda_*^{1/2}(\Phi'''^\top A^2\Phi''')} \lesssim \frac{\|A\|^2_{\text{op}}}{\lambda_*(A)}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}},$$

where $\Phi''' := \Phi_*(A^2)$, and we used $\lambda_*(\Phi'''^\top A^2\Phi''') \ge \lambda_*(A^2)$. Since $\hat{A} = \Phi'''(\Phi'''^\top\hat{A}^2\Phi''')^{1/2}\Phi'''^\top$ and $A^{1/2} = \Phi'''(\Phi'''^\top A^2\Phi''')^{1/2}\Phi'''^\top$, we obtain (31) as

$$\|\hat{A} - A\|_{\text{op}} \lesssim \kappa_*(A)\|A\|_{\text{op}}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}}. \tag{39}$$

Again using Theorem 5.2 from [60] combined with Assumption E.2, we obtain (32) as

$$\|\hat{A}^\dagger - A^\dagger\|_{\text{op}} \lesssim \frac{\kappa^2_*(A)}{\lambda_*(A)}\sqrt{\frac{r_e(A^2)\log^2(n+d+p)}{n}}.$$

This yields $\|\hat{A}^\dagger\|_{\text{op}} \lesssim \|A^\dagger\|_{\text{op}}$.

Part 2. Next we derive (33). By a similar argument as Part 1, (36) and Assumption E.2,

$$\|(\hat{\Sigma}^{(\text{i})}_x)^\dagger - (\Sigma^{(\text{i})}_x)^\dagger\|_{\text{op}} \lesssim \frac{\|\Sigma^{(\text{i})}_x\|_{\text{op}}}{\lambda^2_*(\Sigma^{(\text{i})}_x)}\sqrt{\frac{r_e(\Sigma^{(\text{i})}_x)\log^2(n+d+p)}{n}}. \tag{40}$$

Since $\hat{D} - D = \check{\Sigma}^{(\mathrm{i})}_{\epsilon,x} = \hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger$,

$$
\begin{aligned}
&\|\Phi^\top(\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x - D\Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \\
&\quad\le \left\|\Phi^\top(\hat{D}-D)\Sigma^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top D(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top(\hat{D}-D)(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} \\
&\quad= \left\|\Phi^\top\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger\Sigma^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top D(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} \\
&\quad\le \left\|\Phi^\top\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}(\Sigma^{(\mathrm{i})}_x)^\dagger\Sigma^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}\left((\Sigma^{(\mathrm{i})}_x)^\dagger\Sigma^{(\mathrm{i})}_x - (\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger\Sigma^{(\mathrm{i})}_x\right)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} \\
&\qquad+ \left\|\Phi^\top D(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} + \left\|\Phi^\top\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger(\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\right\|_{\mathrm{op}} \\
&\quad=: Q_1 + R_1 + Q_2 + R_2.
\end{aligned}
$$

We bound Q_1, Q_2, R_1 and R_2 separately. For the terms Q_1 and Q_2, (37) and (38) give

$$
Q_1 \lesssim \|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\sqrt{\frac{(r_e(\Phi^\top\Sigma^{(\mathrm{i})}_\epsilon\Phi) + r_e(A^2))\log^2(n+d+p)}{n}}, \tag{41}
$$

$$
Q_2 \lesssim \|D\Sigma^{(\mathrm{i})}_x D^\top\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\sqrt{\frac{(r_e(\Phi^\top D\Sigma^{(\mathrm{i})}_x D^\top\Phi) + r_e(A^2))\log^2(n+d+p)}{n}}. \tag{42}
$$

For the term R_1, using (35) and (40),

$$
\begin{aligned}
R_1 &\le \|\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}\|_{\mathrm{op}}\|(\Sigma^{(\mathrm{i})}_x)^\dagger - (\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger\|_{\mathrm{op}}\|\Sigma^{(\mathrm{i})}_x\|^{1/2}_{\mathrm{op}}\|\Sigma^{(\mathrm{i})1/2}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \\
&\lesssim \frac{\|\Sigma^{(\mathrm{i})}_x\|^2_{\mathrm{op}}\|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}}{\lambda^2_*(\Sigma^{(\mathrm{i})}_x)}\sqrt{\frac{(r_e(\Sigma^{(\mathrm{i})}_\epsilon) + r_e(\Sigma^{(\mathrm{i})}_x))\log^2(n+d+p)}{n}}\sqrt{\frac{r_e(\Sigma^{(\mathrm{i})}_x)\log^2(n+d+p)}{n}} \\
&\lesssim \kappa^2_*(\Sigma^{(\mathrm{i})}_x)\|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\frac{\sqrt{r_e(\Sigma^{(\mathrm{i})}_x)(r_e(\Sigma^{(\mathrm{i})}_\epsilon) + r_e(\Sigma^{(\mathrm{i})}_x))}\log^2(n+d+p)}{n}
\end{aligned}
$$

For the term R_2, using (35) and (36),

$$
\begin{aligned}
R_2 &\le \|\hat{\Sigma}^{(\mathrm{i})}_{\epsilon,x}\|_{\mathrm{op}}\|(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger\|_{\mathrm{op}}\|\hat{\Sigma}^{(\mathrm{i})}_x - \Sigma^{(\mathrm{i})}_x\|_{\mathrm{op}}\|(\Sigma^{(\mathrm{i})}_x)^\dagger\|^{1/2}_{\mathrm{op}}\|\Sigma^{(\mathrm{i})1/2}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \\
&\lesssim \|(\Sigma^{(\mathrm{i})}_x)^\dagger\|^{3/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|\Sigma^{(\mathrm{i})}_x\|^{3/2}_{\mathrm{op}}\sqrt{\frac{(r_e(\Sigma^{(\mathrm{i})}_\epsilon) + r_e(\Sigma^{(\mathrm{i})}_x))\log^2(n+d+p)}{n}}\sqrt{\frac{r_e(\Sigma^{(\mathrm{i})}_x)\log^2(n+d+p)}{n}} \\
&\lesssim \kappa^{3/2}_*(\Sigma^{(\mathrm{i})}_x)\|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\frac{\sqrt{r_e(\Sigma^{(\mathrm{i})}_x)(r_e(\Sigma^{(\mathrm{i})}_\epsilon) + r_e(\Sigma^{(\mathrm{i})}_x))}\log^2(n+d+p)}{n},
\end{aligned}
$$

where we used $\|(\hat{\Sigma}^{(\mathrm{i})}_x)^\dagger\|_{\mathrm{op}} \lesssim \|(\Sigma^{(\mathrm{i})}_x)^\dagger\|_{\mathrm{op}}$ by Assumption E.2 combined with (40). Again from Assumption E.2, $R_1 + R_2$ is bounded by the right hand side of (41). Therefore,

$$
\begin{aligned}
&\|\Phi^\top(\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x - D\Sigma^{(\mathrm{i})}_x)\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \\
&\quad\lesssim \|\Sigma^{(\mathrm{i})}_\epsilon\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\sqrt{\frac{(r_e(\Phi^\top\Sigma^{(\mathrm{i})}_\epsilon\Phi) + r_e(A^2))\log^2(n+d+p)}{n}} \\
&\qquad+ \|D\Sigma^{(\mathrm{i})}_x D^\top\|^{1/2}_{\mathrm{op}}\|A\|_{\mathrm{op}}\sqrt{\frac{(r_e(\Phi^\top D\Sigma^{(\mathrm{i})}_x D^\top\Phi) + r_e(A^2))\log^2(n+d+p)}{n}}.
\end{aligned}
$$

Finally, from Assumption E.2, we obtain $\|\Phi^\top\hat{D}\hat{\Sigma}^{(\mathrm{i})}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}} \lesssim \|D\Sigma^{(\mathrm{i})1/2}_x\|_{\mathrm{op}}\|\Sigma^{(\mathrm{i})1/2}_x\underline{W}^{\mathrm{pre}\top}_{\ell-1}\|_{\mathrm{op}}$. This concludes the proof. $\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions of this work.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the Conclusion Section (section 8).

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main assumptions and theorems are provided in Section 4, while additional details and complete proofs can be found in Appendix E.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: The paper has disclosed all the information in the method and experiment part.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA] .

   Justification: We have the code required to reproduce our experimental results and are working towards making our code available in a public GitHub repository.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: The experimental setting is clearly described in Section 2, Section 5 and Section 6, and we will make our code available in a public GitHub repository.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: All statistics and results included in the paper are accompanied by confidence intervals.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

Justification: Information for the resources required to reproduce the experiments are included in the oaoer. All experiments are run with 4 x A100 (80G). For the efficiency analysis, a single A100 GPU was used.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: The research conducted in the paper fully conforms with the NeurIPS Code of Ethics in every respect.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: We discuss the broader impacts of our work in Appendix.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: Our paper does not introduce any assets that have a high risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] .

    Justification: We have explicitly mentioned the citations for the datasets and have ensured that all conditions are fully respected.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: Upon acceptance, we will make our codebase publicly available and complete documentation for our assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: We do not include any experiments with human subjects or crowdsourcing.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: We do not include any experiments with human subjects or crowdsourcing.