SC-LoRA: Balancing Efficient Fine-tuning and Knowledge Preservation via Subspace-Constrained LoRA

Anonymous ACL submission

Abstract

Parameter-Efficient Fine-Tuning (PEFT) methods, particularly Low-Rank Adaptation (LoRA), are indispensable for efficiently customizing Large Language Models (LLMs). However, vanilla LoRA suffers from slow convergence speed and knowledge forgetting problems. Recent studies have leveraged the power of designed LoRA initialization, to enhance the fine-tuning efficiency, or to preserve knowledge in the pre-trained LLM. However, none of these works can address the two cases at the same time. To this end, we introduce **S**ubspace-**C**onstrained LoRA (**SC-LoRA**), a novel LoRA initialization framework engineered to navigate the trade-off between efficient fine-tuning and knowledge preservation. We achieve this by constraining the output of trainable LoRA adapters in a low-rank subspace, where the context information of fine-tuning data is most preserved while the context information of preserved knowledge is least retained, in a balanced way. Such constraint enables the trainable weights to primarily focus on the main features of fine-tuning data while avoiding damaging the preserved knowledge features. We provide theoretical analysis on our method, and conduct extensive experiments including *safety preservation* and *world knowledge preservation*, on various downstream tasks. In our experiments, SC-LoRA succeeds in delivering superior fine-tuning performance while markedly diminishing knowledge forgetting, surpassing contemporary LoRA initialization methods.

1 Introduction

Fine-tuning effectively adapts large language models to downstream tasks (Luo et al., 2025; Yu et al., 2024). Due to the high computational cost of full fine-tuning, parameter-efficient fine-tuning (PEFT) methods (Xu et al., 2023; Han et al., 2024) have been proposed to reduce the number of trainable parameters while maintaining good fine-tuning performance. Among various PEFT methods, LoRA (Hu et al., 2022) is a simple yet efficient approach that introduces trainable low-rank adaptation modules for tuning. While LoRA offers significant parameter efficiency, it has two important problems: (1) the convergence speed of the fine-tuning process is relatively slow due to the noise and zero initialization of adapter modules; (2) it potentially leads to catastrophic forgetting problem (Goodfellow et al., 2015) as other fine-tuning methods do, such as harming the world knowledge stored in pre-trained LLMs (Yang et al., 2024), and degrading the safety of aligned LLMs (Qi et al., 2024).

Recent works have found that carefully designed initialization on LoRA adapters can solve these problems. Meng et al. (2024) initializes LoRA adapters by parts of Singular Value Decomposition (SVD) of original weight W_0, leading to faster convergence and improved performance by encapsulating the most significant information stored in W_0. Later works (Yang et al., 2024; Paischer et al., 2024) initialize LoRA weights based on semantic information stored in the activations of each layer on the target fine-tuning dataset. These data-driven approaches successfully enhance the fine-tuning speed and performance. Towards catastrophic forgetting problem in LoRA fine-tuning, Yang et al. (2024) proposes to initialize LoRA weights by the least principal directions of world knowledge data features, successfully alleviating the forgetting problem. However, these works can only solve either side of the two problems, but do not consider the trade-off between enhancing fine-tuning performance and preserving pre-trained knowledge, which is a common need when doing parameter-efficient fine-tuning.

In this paper, we introduce **S**ubspace-**C**onstrained **LoRA**, a balanced LoRA scheme that achieves both better fine-tuning results and good preservation of knowledge in LLMs. Specifically, we compute directions of linear layer output that align with the principal directions of fine-tuning data and at the same time are orthogonal to the principal directions of preserved knowledge. These directions are then used to initialize the adapter weights, constraining the output vectors (of each adapter layer) in a subspace spanned by these directions. This constraint intuitively makes the updating terms to focus on the fine-tuning data information, while avoiding affecting the preserved knowledge. By extensive experiments, we verify that by such constraint on balanced directions, our method achieves both efficient fine-tuning and excellent knowledge preservation, solving the problems that previous methods cannot address. In conclusion, our contribution includes:

1. We propose SC-LoRA, a balanced LoRA scheme that can achieve efficient fine-tuning and knowledge preservation at the same time, which previous methods cannot handle.

2. We provide theoretical proofs to explain our strategies, including analysis on subspace selection and initialization setting.

3. We conduct extensive experiments regarding both *safety preservation* and *world knowledge preservation* on various downstream tasks, verifying the effectiveness of our method.

2 Related Work

Parameter-Efficient Fine-Tuning (PEFT). Modern large language models (LLMs) with billions of parameters face significant computational and memory challenges during full-parameter fine-tuning on downstream tasks, motivating the development of Parameter-Efficient Fine-Tuning (PEFT) methods that optimize only a small amount of parameters while maintaining model performance (Xu et al., 2023; Han et al., 2024).

Common PEFT approaches include partial fine-tuning (Ben Zaken et al., 2022; Bu et al., 2024) that only tune part of the parameters; soft prompt fine-tuning (Hambardzumyan et al., 2021; Lester et al., 2021), where trainable prompts are appended to inputs with model parameters frozen; adapter tuning (Houlsby et al., 2019; Lin et al., 2020; Rücklé et al., 2021; Karimi Mahabadi et al., 2021; Pfeiffer et al., 2021; He et al., 2022; Wang et al., 2022; Lei et al., 2023) which inserts additional trainable layers into LLMs and fix the base model parameters; and LoRA(Hu et al., 2022; Aghajanyan et al., 2021), which decomposes weight updates into low-rank matrices. Different from other approaches, LoRA does not change the original model architecture or incurring extra computational cost during inference since the extra adapters can be merged into original parameters.

LoRA Initialization. Multiple LoRA initialization methods have been proposed, with the aim of improving training efficiency or obtaining other abilities.

PiSSA (Meng et al., 2024) argued that the default initialization of "Gaussian noise (He et al., 2015) and zero" to the adapters can lead to slow convergence. Hence they propose to apply singular value decomposition to original weight matrices and utilizes the top components to initialize LoRA, encapsulating the most significant information stored in original weights. CorDA (Yang et al., 2024) utilizes covariance matrices of data context, and takes the first (or last) singular vectors after context-oriented decomposition as initialization of LoRA adapters. They propose two different modes, one for improving fine-tuning performance and the other for mitigating world knowledge forgetting. Similar to CorDA, EVA (Paischer et al., 2024) feeds fine-tuning data into the model, applies sigular value decomposition to activation covariance matrices, and takes top singular vectors as initialization weights. LoRA-GA (Wang et al., 2024b) also utilizes data context but applies decomposition on the gradient. Hayou et al. (2024) analyze the initialization of LoRA adapters, and have shown how the asymmetry of two low rank matrices affects training dynamics.

Harmful Finetuning Attack and Defense strategies. To prevent potential misuse, LLMs usually undergo specific training to align them with human values before deployment (Ouyang et al., 2022; Bai et al., 2022). Nevertheless, jailbreak attacks employ carefully designed inputs to circumvent this alignment, with prominent methods including Greedy Coordinate Gradient (GCG) (Zou et al., 2023), AutoDAN (Liu et al., 2023), and PAIR (Chao et al., 2023). Beyond these direct attacks, fine-tuning can also undermine a model's safety alignment, even when non-harmful data is used (Qi et al., 2024; He et al., 2024).

Consequently, researchers have developed various defense strategies against such fine-tuning risks, generally falling into following approaches: enhancing the original safety alignment (Huang et al., 2024c,a; Li et al., 2025a), restricting the gradient of fine-tuning parameters or the scope of trained residuals (Wei et al., 2024; Li et al., 2025b), mixing additional safety data (Wang et al., 2024a; Huang et al., 2024b), modifying the loss function (Qi et al., 2025) and post-fine-tuning processing (Yia et al., 2024; Hsu et al., 2024). Different from previous works, our method focuses on an alternative approach of mitigating safety risks during fine-tuning by only modifying initialization, without mixing safety data during fine-tuning, appending prefix during inference time, or adding extra high-rank modules to the model - which would incur computation overhead either in training or inference time.

World Knowledge Forgetting. Catastrophic forgetting (McCloskey and Cohen, 1989) is a phenomenon when models lose previously acquired knowledge when adapting to new tasks, and has been extensively studied in deep learning. Early approaches to solve the problem include knowledge distillation (Li and Hoiem, 2018), rehearsal (Riemer et al., 2019) and dynamic architectures (Yan et al., 2021). For large language models, preserving world knowledge remains challenging due to massive pre-training data and model size. Recent efforts mitigate forgetting by freezing pre-trained layers while introducing new adapters (Wu et al., 2024; Dou et al., 2024). Recently Yang et al. (2024) proposed CorDA with Knowledge-Preserved Adaptation (KPA) mode, addressing world knowledge forgetting through LoRA initialization.

3 Method

Below, we first review the vanilla LoRA, and describe our proposed SC-LoRA method.

3.1 LoRA

Following the hypothesis that the update of weight matrices presents a low rank structure (Aghajanyan et al., 2021), LoRA (Hu et al., 2022) uses the product of two trainable low-rank matrices to learn the weight change while keeping the original weight matrices frozen. To express in mathematical form, LoRA adds low-rank adapters A, B to original weight matrix W_0 by $W' = W_0 + BA$, where $W', W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $r \ll \min(d_{\text{in}}, d_{\text{out}})$. When fine-tuning, W_0 is kept frozen, and A, B are trainable parameters.

From the default initialization scheme of LoRA, A is initialized by Kaiming Initialization (He et al., 2015) while B is initialized by zero matrix. Consequently, the adapter term $BA = O$ and $W' = W_0$ at the start of fine-tuning, ensuring the coherence with the model before fine-tuning. For initializations with non-zero adapter BA (Meng et al., 2024; Yang et al., 2024; Wang et al., 2024b), the frozen weights are adjusted to the residual term $W_{\text{res}} = W_0 - B_{\text{init}}A_{\text{init}}$. Then the adapted weight is $W' = W_0 - B_{\text{init}}A_{\text{init}} + BA = W_{\text{res}} + BA$. In transformer-based LLMs, LoRA adapters are applied to weight matrices within the self-attention and multilayer perceptron (MLP) layers.

3.2 SC-LoRA

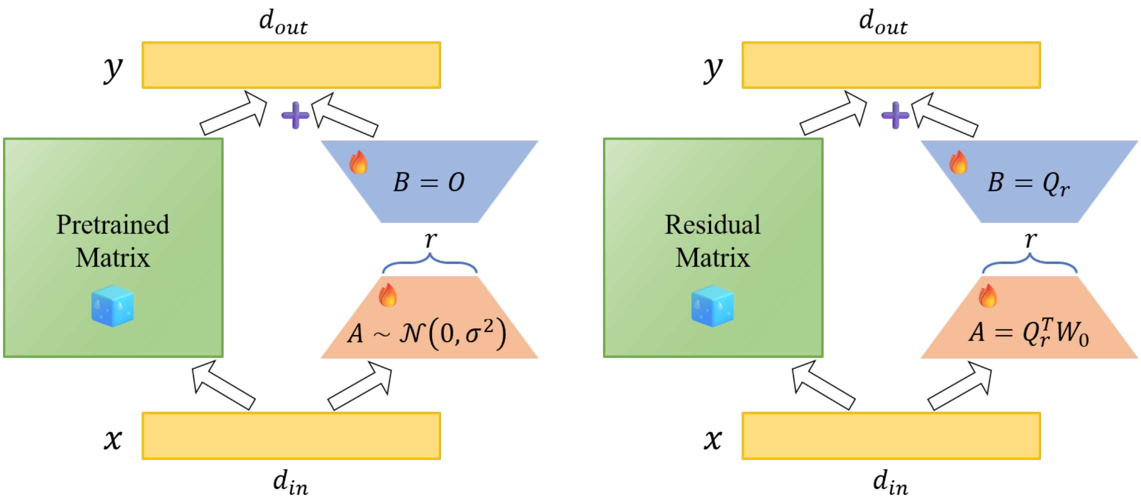
Known as catastrophic forgetting problem (Chen et al., 2020), a large language model often performs worse on its pre-trained knowledge after fine-tuning on a downstream task. To this end, we consider fine-tuning a large language model on downstream task T_+, while preserving its ability on the other task T_-. Consider the output of a linear layer $h = W_0 x = W_{\text{res}}x + B_{\text{init}}A_{\text{init}}x$. We denote $\mathcal{P}_+$ and $\mathcal{P}_-$ the distribution of h when the model is fed with data from T_+ and T_-, respectively. Our aim is to initialize A, B within the r-rank constraint so that BAx preserves the most of $\mathcal{P}_+$ and the least of $\mathcal{P}_-$, so that after initialization, the trainable term BAx is constrained to primarily focus on $\mathcal{P}_+$ while avoiding modifying $\mathcal{P}_-$. This is equivalent to identify a low-dimensional subspace $S \subset \mathbb{R}^{d_{\text{out}}}$ with rank r, on which the projection of $\mathcal{P}_+$ is mostly preserved and the projection of $\mathcal{P}_-$ is mostly eliminated. To evaluate such property of subspace S, we define the following reward:

Definition 1. *For a subspace $S \subset \mathbb{R}^{d_{\text{out}}}$ of dimension r, define the reward $R(S)$ over $\mathcal{P}_\pm$ as:*

$$R(S) = (1-\beta)\mathbb{E}_{X_+\sim\mathcal{P}_+}\left[\|\Pi_S(X_+)\|_2^2\right] - \beta\mathbb{E}_{X_-\sim\mathcal{P}_-}\left[\|\Pi_S(X_-)\|_2^2\right], \tag{1}$$

where $\beta \in [0,1]$ is a hyperparameter to tune. Here $\Pi_S : \mathbb{R}^{d_{\text{out}}} \to S$ denote the orthogonal projection operator onto S. See Appendix A.1 for mathematical definition of Π_S.

The first term of $R(S)$ quantifies the context information of T_+ contained in subspace S, while the second penalizes that of T_-. We use β to balance the trade-off between focusing on T_+ and preservation on T_-. Given the objective to maximize $R(S)$, in the following we provide Theorem 1 to



(a) LoRA (b) SC-LoRA

Figure 1: Comparison of LoRA with default Kaiming initialization and our proposed SC-LoRA. (a) LoRA initializes down-projection matrix A by Gaussian noise and up-projection matrix B by zero matrix. (b) Our SC-LoRA initializes A by $Q_r^\top W_0$ and B by Q_r, where Q_r consists of r orthonormal vectors as columns obtained by Algorithm 1.

compute the optimal subspace and then use it to set our LoRA initialization scheme.

Theorem 1. *Let* $\mathrm{Cov}_+, \mathrm{Cov}_-$ *be the covariance matrices of random vectors* $X_+ \sim \mathcal{P}_+$ *and* $X_- \sim \mathcal{P}_-$*, respectively:*

$$\mathrm{Cov}_+ = \mathbb{E}\left[X_+ X_+^\top\right], \tag{2}$$

$$\mathrm{Cov}_- = \mathbb{E}\left[X_- X_-^\top\right]. \tag{3}$$

And let

$$\Delta\mathrm{Cov} = (1-\beta)\mathrm{Cov}_+ - \beta\mathrm{Cov}_-. \tag{4}$$

Then do eigenvalue decomposition of $\Delta\mathrm{Cov}$ *and take the first* r *eigenvectors* $\{q_i\}_{i\in[r]}$ *with the largest eigenvalues. Then, if following condition holds, the reward* $R(S)$ *is maximized:*

$$S = \mathrm{span}\left(\{q_i\}_{i\in[r]}\right). \tag{5}$$

Proof. See Appendix A.2. □

Theorem 1 shows the steps to compute the optimal subspace that maximized $R(S)$. Then, to constrain the updating output term BAx in the subspace S, we propose our LoRA initialization method:

$$B_{\mathrm{init}} = (q_1\, q_2\, \cdots\, q_r), \tag{6}$$

$$A_{\mathrm{init}} = (q_1\, q_2\, \cdots\, q_r)^\top W_0, \tag{7}$$

$$W_{\mathrm{res}} = W_0 - B_{\mathrm{init}} A_{\mathrm{init}}, \tag{8}$$

as illustrated in Figure 1b. To explain the initialization setting, we provide the following theorem:

Theorem 2. *Let* h, x *be the output and input of the original linear layer* W_0*, satisfying* $h = W_0 x$*. When* A, B *are initialized by Equations 7, 8, the following property holds:*

$$B_{\mathrm{init}} A_{\mathrm{init}} x = \Pi_S(h) \in S,\ \forall x \in \mathbb{R}^{d_{\mathrm{in}}}. \tag{9}$$

Proof. See Appendix A.3. □

Together with Theorem 1, our initialization method has the following properties: When $\beta = 0$ and the model is fed with data from task T_+, h follows distribution $\mathcal{P}_+$, then the norm of the updating term BAx is maximized, providing the most context information of T_+ for training; When $\beta = 1$ and the model is fed with data from task T_-, h follows distribution $\mathcal{P}_-$, then the norm of BAx is minimized, passing the least context information of T_- to trainable parameters. When $\beta \in (0, 1)$, it is the balance between the two cases. The property indicates that, during fine-tuning, the trainable weights are updating more on features related to T_+ and less on features related to T_-, and hence enhancing learning T_+ while avoiding damaging information related to T_-.

The pseudo-code of our initialization algorithm is shown in Algorithm 1. In practice, it is hard to format the true distribution and covariance of output vectors, so we approximate them by feeding hundreds of samples into the model, and use the collection of output vectors to approximate the distribution.

Algorithm 1 SC-LoRA initialization.

Require: Datasets $\mathcal{D}_+, \mathcal{D}_-$ from tasks T_+, T_-, respectively.

1: Let $B_+ = |\mathcal{D}_+|, B_- = |\mathcal{D}_-|$, L be the length of each sample (clipped to same length).

2: Separately feed samples in $\mathcal{D}_+, \mathcal{D}_-$ into the pre-trained model, collect batched output $\hat{X}_+ \in \mathbb{R}^{d_{out} \times B_+ L}, \hat{X}_- \in \mathbb{R}^{d_{out} \times B_- L}$ of each linear layer. Within each sample, the output vector is summed over all tokens.

3: $\mathrm{Cov}_+ \leftarrow \frac{1}{B_+} \hat{X}_+ \hat{X}_+^\top$.

4: $\mathrm{Cov}_- \leftarrow \frac{1}{B_-} \hat{X}_- \hat{X}_-^\top$.

5: Do eigenvalue decomposition on $\Delta\mathrm{Cov} = (1-\beta)\mathrm{Cov}_+ - \beta\mathrm{Cov}_-$, and take the first r eigenvectors $\{q_i\}_{i \in r}$ with the largest eigenvalues.

6: $Q_r \leftarrow (q_1\, q_2\, \cdots\, q_r)$.

7: $B_{\text{init}} \leftarrow Q_r$.

8: $A_{\text{init}} \leftarrow Q_r^\top W_0$.

9: $W_{\text{res}} \leftarrow W_0 - B_{\text{init}} A_{\text{init}}$.

4 Experiments

In the experiments below, we compare SC-LoRA with 5 baselines:

(1) Full fine-tuning. Fine-tune on all parameters of the model;

(2) Vanilla LoRA (Hu et al., 2022). Fine-tune only on LoRA adapters, with B initialized with Gaussian noise (He et al., 2015), and A initialized by zero;

(3) PiSSA (Meng et al., 2024), for efficient fine-tuning. It applies SVD on pre-trained weight W_0 and initializes LoRA adapters by the main parts of decomposition;

(4) CorDA (Yang et al., 2024) Instruction-Previewed Adaptation (IPA) mode, for efficient fine-tuning. It feeds fine-tuning data into the model to get the covariance of activations, applies self-defined context-oriented decomposition, and initializes LoRA adapters with principal directions obtained in decomposition;

(5) CorDA Knowledge-Preserved Adaptation (KPA) mode, for knowledge preservation. The initialization algorithm is basically the same as IPA mode except that it feeds preserved knowledge data and take the least principal directions for initialization.

For initialization of CorDA IPA and KPA mode, we calculate the covariance matrices with 256 samples from fine-tuning dataset and preserved knowledge dataset, respectively with 256 samples. We use AdamW optimizer (Loshchilov and Hutter, 2019) with the following hyper-parameters: batch size 128, learning rate 2e-5 (except for experiment in Section 4.3, where we tune the learning rate of baselines for better performance), cosine annealing learning rate schedule, warm-up ratio 0.03, and no weight decay. The rank of LoRA and its variants are all set to 128 for comparison. For SC-LoRA, we tune the hyperparameter β to find a good balanced result. All experiment results are obtained by running on only one seed.

Below we discuss results in three settings: (1) Preservation of world knowledge when fine-tuning on math task; (2) Preservation of safety when fine-tuning on benign data; (3) Preservation of safety when fine-tuning on poisoned data.

4.1 World Knowledge Preservation

Pre-trained LLMs also have other pre-trained knowledge that is easy to lose after fine-tuning on downstream tasks, such as world knowledge (Yang et al., 2024). In this setting, we aim to preserve the intrinsic world knowledge (e.g., common sense) within the pre-trained LLM while providing efficient fine-tuning on downstream tasks. We fine-tune the Llama-2-7b model (Touvron et al., 2023) on math task and evaluate its math ability (utility) and world knowledge performance. We train on 100000 samples of MetaMathQA (Yu et al., 2024) for 1 epoch and evaluate its math ability on GSM8k (Cobbe et al., 2021) and MATH (Yu et al., 2024) validation sets. World knowledge is evaluated by the exact matching score on TriviaQA (Joshi et al., 2017), NQ-open (Lee et al., 2019), and WebQS (Berant et al., 2013) through Evaluation-Harness (Gao et al., 2024). We select 256 random samples from NQ-open as world knowledge samples used for the initialization of SC-LoRA and CorDA KPA, and 256 random samples from MetaMathQA as fine-tuning dataset for initializing SC-LoRA and CorDA IPA. Note that samples used in initialization are separate from those in evaluation.

As shown in Table 1, the results of full fine-tuning and LoRA show the degradation on world knowledge when fine-tuning on downstream task MetaMathQA. SC-LoRA achieves best math ability (surpassing full fine-tuning), and preserves world knowledge relatively well. When $\beta = 0.8$, it

Method		#Params	TriviaQA↑	NQ-open↑	WebQS↑	Avg↑	GSM8k↑	MATH↑	Avg↑
Llama-2-7b		-	52.52	18.86	5.86	25.75	-	-	-
Full fine-tuning		6738M	47.42	4.16	6.64	19.41	50.27	6.94	28.60
LoRA		320M	46.81	1.05	7.04	18.30	41.77	5.46	23.62
PiSSA		320M	47.44	3.32	6.84	19.20	51.63	7.70	29.67
CorDA IPA		320M	30.20	9.83	5.41	15.15	51.40	8.34	29.87
CorDA KPA		320M	46.21	**10.64**	**7.33**	21.39	45.03	6.54	25.79
SC-LoRA	$\beta = 0$	320M	44.26	5.18	7.19	18.88	**53.53**	**8.98**	**31.25**
	$\beta = 0.5$	320M	48.91	7.70	6.89	21.17	53.37	8.62	31.00
	$\beta = 0.8$	320M	**50.52**	**10.64**	7.04	**22.73**	52.46	7.62	30.04

Table 1: Results of world knowledge preservation and math ability after fine-tuning on MetaMATH.

surpasses all baselines on both utility and knowledge preservation. Also, from the results of SC-LoRA, we can see a clear trend when increasing β, that the knowledge preservation ability is increasing while the utility is decreasing, which aligns with our design methodology for β in Section 3. More details will be shown in Section 4.4 to analyze this trend.

4.2 Safety Preservation on Benign Finetuning

Qi et al. (2024) has shown that fine-tuning on benign data can compromise the safety of aligned LLMs. In this setting, we aim to preserve the safety of aligned LLM while providing efficient fine-tuning on downstream tasks. Following the experimental settings by Qi et al. (2025), we fine-tune Llama-2-7b-Chat model with safety alignment (Touvron et al., 2023) on Samsum (Gliwa et al., 2019) for 1 epoch. Samsum is a dataset for conversation summarization task, containing 14732 training samples and 819 testing samples.

To initialize our SC-LoRA model, we randomly select 256 samples from training set of Samsum ($\mathcal{D}_+$) to compute covariance matrix Cov_+ for each linear layer, then use 256 harmful-question&refusal-answer pairs (as the safety dataset $\mathcal{D}_-$) provided by Qi et al. (2025) to compute Cov_-. These two collections of samples are also used to compute the covariance matrices of CorDA IPA and CorDA KPA respectively.

For utility evaluation, we employ the standard ROUGE-1 score (Lin, 2004) for testing set of Samsum. For safety evaluation, we let the fine-tuned models to generate answers for 330 malicious questions provided by Qi et al. (2024) (distinct from malicious questions for initialization) and employ DeepSeek-V3 (DeepSeek-AI et al., 2025) API to judge the harmfulness, assigning each answer an integer score from 1 (safe) to 5 (most harmful). We report the average score as **harmfulness score** of the model and the fraction of maximum-risk responses (score = 5) as **harmfulness rate**. Lower values for both metrics indicate stronger safety of the model.

As shown in Table 2, SC-LoRA achieves high utility, even surpassing full fine-tuning on Samsum dataset when $\beta = 0.5$. At the same time, SC-LoRA shows almost no safety degradation compared to the model before fine-tuning, while all baselines except CorDA KPA present notable safety degradation, since they are not designed for knowledge preservation. However, the utilities of all fine-tuning methods (except for CorDA IPA) are generally close. We hypothesize that the task of summarization is quite simple, so training for only 1 epoch is enough for utility convergence. Also, the results of SC-LoRA shows that when β is increasing, the safety preservation becomes better while utility is decreasing. This aligns with our design of β to balance the trade-off.

4.3 Safety Preservation on Data Poisoning Attack

Harmful data injection is a common attack method to degrade the safety of LLMs during fine-tuning (Huang et al., 2024a,b,c). In this experiment, we aim to preserve safety in poisoned data scenarios. To construct the poisoned dataset, we first take 25600 data samples from training set of MetaMathQA (Yu et al., 2024), then replace 1% of the data by harmful question-answer pairs provided by (Qi et al., 2024). We train each method for 1 epoch on the poisoned dataset. For the initialization of SC-LoRA and CorDA IPA, we use 256 samples from training set of MetaMathQA. The safety samples used for the initialization of SC-LoRA and CorDA KPA are the same with the previous experiment (Section 4.2).

For utility evaluation, we compute the answer accuracy on the validation set of GSM8k (Cobbe

Method		#Params	HS↓	HR(%)↓	Utility↑
Llama-2-7b-Chat		-	1.100	1.212	24.13
Full fine-tuning		6738M	1.364	5.455	51.41
LoRA		320M	1.176	2.424	50.32
PiSSA		320M	1.252	4.242	51.87
CorDA IPA		320M	1.209	3.333	44.61
CorDA KPA		320M	1.106	0.606	50.89
SC-LoRA	$\beta = 0.5$	320M	1.161	1.818	**52.54**
	$\beta = 0.7$	320M	1.148	1.818	52.07
	$\beta = 0.9$	320M	**1.097**	**0.000**	51.67

Table 2: Results of Safety preservation and fine-tuning performance when training on benign dataset Samsum. #Params is the number of trainable parameters. HS and HR denote harmfulness score and harmfulness rate respectively.

Method		#Params	HS↓	HR(%)↓	Utility↑
Llama-2-7b-Chat		-	1.100	1.212	-
Full fine-tuning		6738M	2.248	23.94	41.47
LoRA	lr=2e-5	320M	**1.118**	**1.212**	31.69
	lr=5e-5	320M	2.276	23.64	37.68
	lr=1e-4	320M	3.155	41.52	41.93
PiSSA		320M	2.379	29.39	41.77
CorDA IPA		320M	4.239	67.27	43.75
CorDA KPA		320M	1.127	**1.212**	40.33
SC-LoRA	$\beta = 0.5$	320M	1.630	10.91	**45.56**
	$\beta = 0.7$	320M	1.224	3.030	45.26
	$\beta = 0.9$	320M	1.136	**1.212**	45.26

Table 3: Results of safety preservation and fine-tuning performance when training on poisoned dataset MetaMathQA with 1% malicious question-answer pairs.

et al., 2021). Safety evaluation follows the setting in the previous section 4.2. For better comparability, we tune the learning rate of LoRA to 2e-5, 5e-5 and 1e-4. The learning rate for other methods is fixed to 2e-5.

From the results in Table 3, we can observe that the data points exhibit a wider spread among these methods, both in utility and safety metric. Compared to the original model, SC-LoRA ($\beta = 0.9$) exhibits almost no safety degradation, and achieves best utility, even surpassing full fine-tuning by 3.79 points. When increasing the learning rate, LoRA shows a sharp decline in safety alignment while math ability is increasing. LoRA (lr=2e-5) and CorDA KPA, though preserving safety well, are insufficient in fine-tuning performance compared to our method. PiSSA and CorDA IPA, though showing their capacity in better fine-tuning, heavily degrades the safety of the model. This again shows the potential of our method to enhance the utility of the model and preserve safety at the same time, even when the fine-tuning dataset contains a small fraction of harmful content. Also, the utility and safety of SC-LoRA follows the same trend as in fine-tuning on benign data when β is increasing, supporting the sedign of our method.

4.4 Experimental Analysis on the Functionality of Hyper-Parameter β

As explained in Section 3, the value of β balance the trade-off between knowledge preservation and fine-tuning efficiency. Intuitively, when increasing β, there exists a trend that the fine-tuning performance will drop and the knowledge preservation ability will increase. While we have observed this trend in the previous section, we illustrate the trend more explicitly in Figure 2 and 3. In Figure 2, both two curves shows knowledge preservation improvement when β is increasing: one for safety increasing, and the other for world knowledge preservation improvement. In Figure 3, the math ability decreases when β is increasing, aligning with our

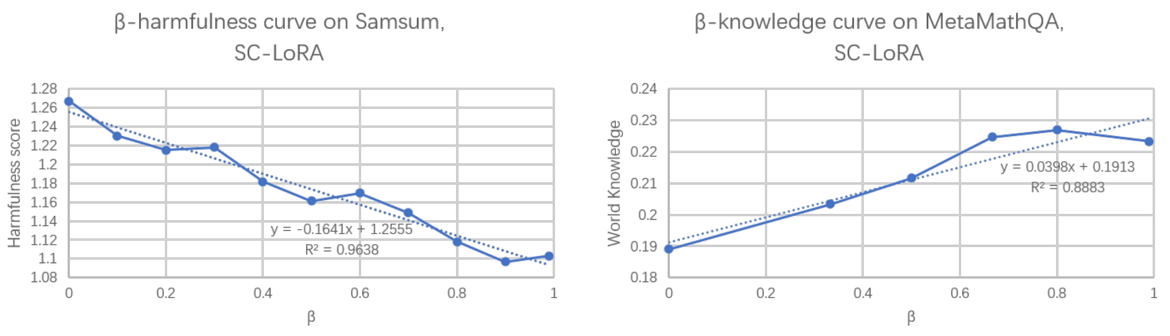


Figure 2: Relations between β and knowledge preservation performance. The experiment setting of the left figure is described in Section 4.2, while that of the right figure is described in Section 4.1. Lower harmfulness score or higher world knowledge score indicates better performance on knowledge preservation.

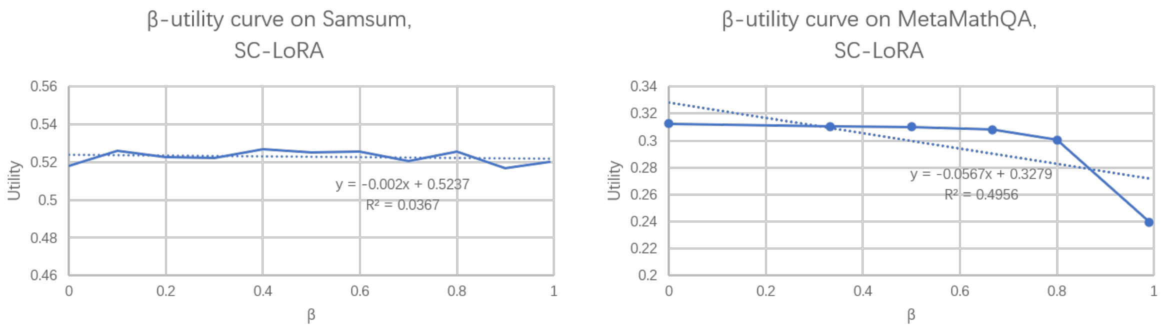


Figure 3: Relations between β and fine-tuning performance. The experiment setting of the left figure is described in Section 4.2, while that of the right figure is conducted in Section 4.1. The right figure shows clear monotonicity with β, while such trend does not occur in the left figure.

expectations. The utility on Samsum, however, does not show evident trend as β varies, but fluctuating around 0.52. We hypothesize that the task of summerization is quite simple, so whatever the value of β, it is sufficient for utility convergence during fine-tuning.

These trends give experimental support to our method design, that by adjusting β we can balance the trade-off. Interestingly, a linear relationship was observed between β values and knowledge preservation in some experimental settings.

5 Conclusion

Aimed to balance the trade-off between efficient fine-tuning and knowledge preservation, this paper presents a data-driven LoRA initialization that utilizes the subspace constrain, in order to strengthen the target knowledge while downgrading its influence on preserved knowledge. Theoretical analysis are provided to support our method, including the choice of subspace and the initialization setting. We conduct extensive experiments regrading safety preservation and world knowledge preservation, during fine-tuning on various downstream tasks such as math and summarization. The results of experiments strongly demonstrate that our method can not only promote fine-tuning performance on downstream tasks, but also preserve the intrinsic knowledge stored in pre-trained model, surpassing contemporary LoRA initialization methods.

6 Limitations

First, SC-LoRA is just a LoRA initialization method, and does not strongly constrain the updates during fine-tuning process. Hence after fine-tuning on more complex tasks and with more steps, the knowledge preservation ability can also drop (see the preservation drop of NQ-open in Table 1 for example). Second, its application on preserving other types of knowledge remains unexplored. Future work may consider applying SC-LoRA to preserving multimodal large language model's performance on pre-training tasks (Zhai et al., 2024) or large language model's reasoning ability.

These aspects provide promising directions for future researches.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2024. Differentially private bias-term fine-tuning of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4730–4751. PMLR.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Preprint*, arXiv:1312.6211.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. The impact of initialization on lora finetuning dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pages 117015–117040. Curran Associates, Inc.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What is in your safe data? identifying benign data that breaks safety. In *First Conference on Language Modeling*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Safe loRA: The silver lining of reducing safety risks when finetuning large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024a. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024b. Lazy safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2405.18641*, 2.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024c. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *arXiv preprint arXiv:2402.01109*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, volume 34, pages 1022–1035. Curran Associates, Inc.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Zhao, Yuexin Wu, Bo Li, Yu Zhang, and Ming-Wei Chang. 2023. Conditional adapters: Parameter-efficient transfer learning with fast inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 8152–8172. Curran Associates, Inc.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025a. Salora: Safety-alignment preserved low-rank adaptation. In *International Conference on Representation Learning*, volume 2025, pages 90827–90843.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025b. Safety layers in aligned large language models: The key to LLM security. In *The Thirteenth International Conference on Learning Representations*.

Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online. Association for Computational Linguistics.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*.

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

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121038–121072. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. 2024. One initialization to rule them all: Fine-tuning via explained variance adaptation. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, , and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. 2024a. Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shaowen Wang, Linxi Yu, and Jian Li. 2024b. Lora-ga: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, pages 54905–54931. Curran Associates, Inc.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52588–52610. PMLR.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. 2024. LLaMA pro: Progressive LLaMA with block expansion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6518–6537, Bangkok, Thailand. Association for Computational Linguistics.

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

Shipeng Yan, Jiangwei Xie, and Xuming He. 2021. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. 2024. Corda: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 71768–71791. Curran Associates, Inc.

Xin Yia, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. 2024. A safety realignment framework via subspace-oriented model fusion for large language models.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pages 202–227. PMLR.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Definitions and Proofs

A.1 Mathematical Definition of Π_S

Definition 2. *Suppose S is a subspace of $\mathbb{R}^n$ of dimension r, and let $\{q_i\}_{i\in[r]}$ be an orthonormal basis of S, then the orthogonal projection operator onto S, denoted Π_S, is defined as:*

$$\begin{aligned}\Pi_S(x) &= \sum_{i=1}^{r}(q_i^\top x)q_i \\ &= \sum_{i=1}^{r}(q_i q_i^\top)x, \forall x \in \mathbb{R}^n.\end{aligned} \tag{10}$$

Note: the selection of the orthonormal basis does not affect Π_S.

A.2 Proof for Theorem 1

Proof. Suppose for some subspace $S \subset \mathbb{R}^d$ (ignore the subscript of d_{out} for simplicity) of dimension r, there exists an orthonormal basis $\{v_i\}_{i\in[r]}$ that spans S, that is $S = \text{span}\left(\{v_i\}_{i\in[r]}\right)$.

For simplicity, denote

$$\tilde{I}_r = \sum_{i\in[r]} v_i v_i^\top, \tag{11}$$

then the following equality holds:

$$\begin{aligned}\tilde{I}_r^\top \tilde{I}_r &= \sum_{i\in[r]}\sum_{j\in[r]} v_i v_i^\top v_j v_j^\top \\ &= \sum_{i\in[r]}\sum_{j\in[r]} v_i \langle v_i, v_j\rangle v_j^\top \\ &= \sum_{i\in[r]}\sum_{j\in[r]} \delta_{ij} v_i v_j^\top \\ &= \sum_{i\in[r]} v_i v_i^\top = \tilde{I}_r.\end{aligned} \tag{12}$$

From property of projection,

$$\begin{aligned}\Pi_S\left(X_\pm\right) &= \sum_{i=1}^{r}\langle X_\pm, v_i\rangle v_i = \sum_{i=1}^{r} v_i v_i^\top X_\pm \\ &= \left(\sum_{i=1}^{r} v_i v_i^\top\right) X_\pm = \tilde{I}_r X_\pm.\end{aligned} \tag{13}$$

Thus

$$\begin{aligned}&\mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[\left\|\Pi_S\left(X_\pm\right)\right\|_2^2\right] \\ =&\mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[\left\|\tilde{I}_r X_\pm\right\|_2^2\right] \\ =&\mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[\text{tr}\left(X_\pm^\top \tilde{I}_r^\top \tilde{I}_r X_\pm\right)\right] \\ =&\mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[\text{tr}\left(X_\pm^\top \tilde{I}_r X_\pm\right)\right] \\ =&\mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[\text{tr}\left(\tilde{I}_r X_\pm X_\pm^\top\right)\right] \\ =&\text{tr}\left(\tilde{I}_r \mathbb{E}_{X_\pm\sim\mathcal{P}_\pm}\left[X_\pm X_\pm^\top\right]\right) \\ =&\text{tr}\left(\tilde{I}_r \text{Cov}_\pm\right).\end{aligned} \tag{14}$$

Suppose the spectral decomposition of $(1-\beta)\text{Cov}(X_+) - \beta\text{Cov}(X_-)$ is $Q\Sigma Q^\top$, where $Q = (q_1\, q_2\, \cdots\, q_d)$, Σ is diagonal with eigenvalues sorted in descending order. Then we have

$$\begin{aligned}R(S) &= (1-\beta)\mathbb{E}_{X_+\sim\mathcal{P}_+}\left[\left\|\Pi_S\left(X_+\right)\right\|_2^2\right] \\ &\quad - \beta\mathbb{E}_{X_-\sim\mathcal{P}_-}\left[\left\|\Pi_S(X_-)\right\|_2^2\right] \\ &= (1-\beta)\text{tr}\left(\tilde{I}_r\text{Cov}_+\right) - \beta\text{tr}\left(\tilde{I}_r\text{Cov}_-\right) \\ &= \text{tr}\left(\tilde{I}_r\Delta\text{Cov}\right) \\ &= \sum_{i\in[r]}\text{tr}\left(v_i v_i^\top Q\Sigma Q^\top\right) \\ &= \sum_{i\in[r]} v_i^\top Q\Sigma Q^\top v_i.\end{aligned} \tag{15}$$

Extend $\{v_i\}_{i\in[r]}$ to a complete orthonormal basis $\{v_i\}_{i\in[d]}$ for $\mathbb{R}^d$, and denote $u_i = Q_i^\top v_i$. Since Q is an orthogonal matrix, $\{u_i\}_{i\in[d]}$ is also an orthonormal basis for $\mathbb{R}^d$. From Ky Fan's theorem on eigenvalues, $\max\left(\sum_{i\in[r]} v_i^\top Q\Sigma Q^\top v_i\right) = \sum_{i\in[r]}\Sigma_{ii}$, and one can easily verify that the condition above achieves the maximum.

For the if and only if part (adding the condition of eigenvalue gap): suppose $U = (u_1\, u_2\, \cdots\, u_d)^\top$ as an orthogonal matrix, then

$$\begin{aligned}R(\{v_i\}_{i\in[r]}) &= \sum_{i\in[r]} u_i^\top \Sigma u_i \\ &= \sum_{i\in[r]}\sum_{j=1}^{d}\Sigma_{jj}U_{ij}^2 \\ &= \sum_{j=1}^{d}\left(\Sigma_{jj}\sum_{i\in[r]}U_{ij}^2\right).\end{aligned} \tag{16}$$

From property of orthogonal matrix, $\sum_{i\in[r]} U_{ij}^2 \leq 1$ and $\sum_{j=1}^{d}\sum_{i\in[r]} U_{ij}^2 = r$, then to maximize R, from the additional assumption we need $\sum_{i\in[r]} U_{ij}^2 = \begin{cases} 1, & 1\leq j\leq r \\ 0, & r+1\leq j\leq d \end{cases}$, which is equivalent to

$$U_{1:r,1:d}^{\top} U_{1:r,1:d} = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix}. \tag{17}$$

From $U_{1:r,1:d} = (v_1\, v_2\, \cdots\, v_r)^{\top} Q$, we know that this is also equivalent to

$$(v_1\, v_2\, \cdots\, v_r)(v_1\, v_2\, \cdots\, v_r)^{\top} = Q \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} Q^{\top}, \tag{18}$$

which is also written as

$$\sum_{i=1}^{r} v_i v_i^{\top} = \sum_{i=1}^{r} q_i q_i^{\top}. \tag{19}$$

Indicating $S = \mathrm{span}\left(\{q_i\}_{i\in[r]}\right)$.

$\square$

A.3 Proof of Theorem 2

Proof. Denote $Q_r = (q_1\, q_2\, \cdots\, q_r)$.

Since $\{q_i\}_{i\in[r]}$ is a orthonormal basis that spans S, from definition of orthogonal projection we have

$$\Pi_S(h) = \sum_{i=1}^{r} q_i q_i^{\top} h = Q_r Q_r^{\top} h. \tag{20}$$

Thus $\forall x \in \mathbb{R}^{d_{\text{in}}}$, we have

$$B_{\text{init}} A_{\text{init}} x = Q_r Q_r^{\top} W_0 x = Q_r Q_r^{\top} h = \Pi_S(h), \tag{21}$$

which completes the proof.

$\square$

B Numerical instability in sparse sample setting

When the sample size is much larger than the output activation dimension, $\min(|\mathcal{D}_+|L, |\mathcal{D}_-|L) \gg d_{\text{out}}$, setting $\beta \in [0,1]$ causes no issue. However, when samples are sparse (specifically, when the number of negative-task sample $B_- < (d_{\text{out}} - r)/L$, setting $\beta = 1$ introduces multiple valid solutions in the spectral decomposition step due to high-dimensional freedom in the null space of Cov_-. Mathematically, the rank of Cov_- is at most $B_- L$, resulting in a null space of dimension $d_{\text{out}} - \mathrm{rank}(\mathrm{Cov}_-) \geq d_{\text{out}} - B_- L > r$. Consequently, **any arbitrary set of r orthonormal vectors** in this null space can satisfy the decomposition criterion, leading to non-unique initialization of parameters A and B. Even when $B_- \sim (d_{\text{out}} - r)/L$, the decomposition results may also be affected significantly by data selection and clipping.

To mitigate this instability, we recommend setting $1-\beta$ to a small positive value (rather than exactly zero). This retains the regularization from Cov_+ in the objective function, which constrains the null space ambiguity and stabilizes the spectral decomposition, empirically improves fine-tuning performance.