Norm-Bounded Low-Rank Adaptation

Ruigang Wang
The University of Sydney
ruigang.wang@sydney.edu

Krishnamurthy (Dj) Dvijotham ServiceNow dvij@cs.washington.edu Ian R. Manchester
The University of Sydney
ian.manchester@sydney.edu.au

Abstract

In this work, we propose norm-bounded low-rank adaptation (NB-LoRA) for parameter-efficient fine tuning. We introduce two parameterizations that allow explicit bounds on each singular value of the weight adaptation matrix, which can therefore satisfy any prescribed unitarily invariant norm bound, including the Schatten norms (e.g., nuclear, Frobenius, spectral norm). The proposed parameterizations are unconstrained and complete, i.e. they cover all matrices satisfying the prescribed rank and norm constraints. Experiments on vision fine-tuning benchmarks show that the proposed approach can achieve good adaptation performance while avoiding model catastrophic forgetting and also substantially improve robustness to a wide range of hyper-parameters, including adaptation rank, learning rate and number of training epochs. We also explore applications in privacy-preserving model merging and low-rank matrix completion.

1 Introduction

Pre-trained vision and language models have demonstrated impressive generalization capability across a wide variety of tasks; see, e.g. Achiam et al. (2023); Touvron et al. (2023). When a specific target task is identified, however, it has been observed that parameter-efficient fine-tuning (PEFT) techniques, e.g. Houlsby et al. (2019); Hu et al. (2022), can improve performance via quick model adaption with low computation and data requirements. The primary goal for an effective PEFT method is to achieve good adaptation performance with high training efficiency, i.e., dramatically fewer trainable parameters and a number of training epochs. Alongside this primary goal, it is often also desirable to maintain the generalization per-

formance of the original pre-trained model as much as possible Qiu et al. (2023).

A common PEFT strategy is to learn an additive or multiplicative perturbation to the pre-trained model weights, and the key step is to construct a parameter-efficient representation of this adaptation matrix. Low-rank adaption (LoRA) Hu et al. (2022) is a widely applied PEFT method, which parameterizes an additive adaptation matrix via low-rank decomposition. Specifically, the fine-tuned weight W_{ft} is

$$W_{ft} = W_{pt} + W := W_{pt} + A^{\top} B,$$
 (1)

where the pre-trained weight $W_{pt} \in \mathbb{R}^{m \times n}$ is the frozen during fine-tuning, and $A \in \mathbb{R}^{r \times m}$, $B \in \mathbb{R}^{r \times n}$ are the learnable parameters, so that W has rank $r \ll \min(m, n)$.

The rank is a one way to quantify the "size" of a matrix, corresponding to its underlying dimensionality. But matrix norms – such as nuclear, Frobenius, or spectral norms – provide another notion of size, quantifying the magnitude of a matrix's elements and of its operation on vectors.

In this paper, we consider the Schatten p-norms as a general family which includes the nuclear norm, Frobenius norm, and spectral norm $(p=1,2,\infty)$, respectively). The Schatten p-norm of a matrix W is defined as follows: let $\sigma_1(W) \geq \cdots \geq \sigma_r(W) \geq 0$ be the singular values of W, then

$$||W||_{S_p} = \left(\sum_{i=1}^r \sigma_i^p(W)\right)^{1/p},$$

for $1 \le p < \infty$ and $||W||_{S_p} = \sigma_1(W)$ for $p = \infty$. I.e. it is the vector p-norm of the sequence of singular values.

It is important to note that a matrix can have low rank but arbitrarily high Schatten p-norm for any p, and vice-versa. Therefore LoRA, which only controls rank,

can still suffer from typical issues associated with large parameter norms such training instability and sensitivity to hyper-parameters including learning rate Biderman et al. (2024); Bini et al. (2024) and number of training epochs Qiu et al. (2023).

In this work, we argue that both rank and norm of the weight adaptation matrix should be controlled for PEFT, and our main technical contributions are flexible parameterizations that enable this. To be precise, we introduce two options for parameterizing low-rank updates with Schatten norm constraints on the weight update: $W \in \mathbb{W}_p^{r,\delta}$ with

$$\mathbb{W}_p^{r,\delta} = \{ W \in \mathbb{R}^{m \times n} \mid \text{rank}(W) \le r, \ \|W\|_{S_p} \le \delta \},$$
(2)

where $r \in \mathbb{N}$ and $\delta > 0$ are the rank and norm budgets, respectively.

Beyond PEFT, we argue that controlling both rank and norm of a matrix is a basic capability that appears in a variety of problems, including classical problems such as low-rank matrix completion Recht et al. (2010); Mishra et al. (2013) as well as emerging concerns such as privacy-preserving model merging. We briefly discuss these applications in Section 4.

Contributions. We propose norm-bounded low-rank adaptation (NB-LoRA), an approach parameter-efficient fine-tuning and related problems.

We provide two parameterizations that automatically satisfy the following rank and norm constraints:

$$rank(W) \le r, \ \|W\|_{S_n} \le \delta, \tag{3}$$

with $r \in \mathbb{N}$ and $\delta > 0$..

- Our parameterizations are unconstrained, i.e. mappings $W : \mathbb{R}^N \mapsto \mathbb{R}^{m \times n}$, and are *complete*, i.e., for any $W \in \mathbb{R}^{m \times n}$ satisfying (3), there exists a (not necessarily unique) $\theta \in \mathbb{R}^N$ such that $W = \mathcal{W}(\theta)$.
- Numerical experiments on PEFT of a vision transformer illustrate that NB-LoRA can achieve equivalent adaptation performance to LoRA and other existing methods while exhibiting less "forgetting" of the source model. Also, norm bounds appear to significantly reduce sensitivity to hyperparameter variation.
- We illustrate the broader utility of NB-LoRA via simple privacy-preserving model merging and lowrank matrix completion applications.

2 Related Work

Weight distance regularization for LoRA. Recent work has shown that LoRA can be highly sensitive to learning rate Bini et al. (2024); Biderman et al. (2024) and it is susceptible to over-training Qiu et al. (2023). To mitigate these effects, several recent works have proposed to control various measures of "distance" of the weight adaption. For example, Gouk et al. (2021); Chen et al. (2023) propose an approach that preserves the Euclidean weight distances between pre-trained and fine-tuned models. Liu et al. (2024a) investigate the vector-wise norm of the adaption matrix.

Orthogonal Fine-Tuning (OFT). A recent approach based on multiplicate adaption Qiu et al. (2023) has received much attention. Specially, it modifies the pre-trained weight via $W_{ft} = W_{pt}R$, where R is learnable orthogonal adaption matrix which can be efficiently parameterized by via block-wise Cayley transformation and butterfly factorization Liu et al. (2024b). The equivalent additive update $W = W_{pt}(I - R)$ has bounded Frobenius norm as $||I-R||_F$ is bounded. However, W is often high-rank and its Frobenius bound increases with the dimension. Low-rank structure and small bound can be enforced by using the Householder transformation Dong et al. (2024). One way to increase the rank is a series of Householder transformation Yuan et al. (2024) or its variant based on hyperplane reflections Bini et al. (2024). However, the computational cost and norm bound increase with the number of such transformations.

SVD-based LoRA. Zhang et al. (2023) takes SVD-like parameterization $W = U\Sigma V^{\top}$ with adaptive learning of Σ and imposing the orthogonality of U,V via penalties. In Lingam et al. (2024), the U,V from SVD decomposition of W_{pt} are re-used and a small square matrix Σ is learned during fine-tuning. No norm bounds or constraint on singular values were considered in those works.

3 Parameterizations for Norm-Bounded Low-Rank Adaptation

Here we construct our two parameterizations. In each case, we first construct a parameterization for matrices with bounds one individual singular value. Then, by further parameterizing the singular value bounds,

we can parameterize matrices satisfying any unitarily invariant norm bound.

3.1 Preliminaries

We will be comparing singular values of matrices of potentially different ranks and sizes, so for convenience for any $j \in \mathbb{N}$ we set the jth singular value of W to be $\sigma_j(W) = 0$ if $j > \operatorname{rank}(W)$. We can now introduce the relation \preceq_{σ} .

Definition 3.1. Let A, B be two matrices. We say $A \leq_{\sigma} B$ if $\sigma_j(A) \leq \sigma_j(B), \forall j \in \mathbb{N}$.

Note the \leq_{σ} is reflexive $(A \leq_{\sigma} A)$ and transitive $(A \leq_{\sigma} B, B \leq_{\sigma} C \Rightarrow A \leq_{\sigma} C)$. But it is not antisymmetric (i.e., $A \leq_{\sigma} B, B \leq_{\sigma} A \not\Rightarrow A = B$), e.g., A, B are two orthogonal matrices. Moreover, we have $||A||_{S_p} \leq ||B||_{S_p}$ for any p if $A \leq_{\sigma} B$.

Let $s \in \mathbb{R}^r_{++}$ with $\mathbb{R}_{++} = (0, \infty)$ and $S = \operatorname{diag}(s)$ be diagonal matrix with $S_{jj} = s_j$. We define the set of matrices whose singular values are bounded by S by

$$\mathbb{W}_S := \{ W \in \mathbb{R}^{m \times n} \mid W \preceq_{\sigma} S \}. \tag{4}$$

Note that for any $W \in \mathbb{W}_S$, we have $\operatorname{rank}(W) \leq r$ and $\|W\|_{S_p} \leq \|s\|_p$ with $\|\cdot\|_p$ as the vector *p*-norm.

Definition 3.2. A mapping $W : \mathbb{R}^N \to \mathbb{W}_S$ is said to be a *complete parameterization* for \mathbb{W}_S if $W(\mathbb{R}^N) = \mathbb{W}_S$.

3.2 NB-LoRA I

Our first complete parameterization is based on a mild modification to the singular value decomposition. It takes $F_u \in \mathbb{R}^{m \times r}, F_v \in \mathbb{R}^{n \times r}$ and $d \in \mathbb{R}^r$ as learnable parameters and then produces W as follows:

1) Compute intermediate variables

$$U = \text{Cayley}(F_u), \quad V = \text{Cayley}(F_v),$$

where the Cayley transformation for a tall rectangular matrix $F \in \mathbb{R}^{q \times r}$ with $q \geq r$ is defined by

Cayley
$$\left(F := \begin{bmatrix} X \\ Y \end{bmatrix}\right) := \begin{bmatrix} (I-Z)(I+Z)^{-1} \\ -2Y(I+Z)^{-1} \end{bmatrix},$$

where $Z = X - X^{\top} + Y^{\top}Y$ with $X \in \mathbb{R}^{r \times r}$ and $Y \in \mathbb{R}^{(q-r) \times r}$:

2) Project d onto the interval [-s, s], i.e.,

$$\hat{d} = \frac{s \odot d}{\max(|d|, s)},\tag{6}$$

where $|\cdot|$, \odot are the element-wise operations taking absolute value and multiplication, respectively;

3) Take $D = \operatorname{diag}(\hat{d})$ and compute W as follows

$$W = UDV^{\top}. (7)$$

Note that by construction U and V have orthonormal columns, however (7) is not exactly the standard singular value decomposition as D can have negative elements on the diagonal. However this is the a key feature to make the above parameterization complete.

Theorem 3.3. The parameterization W: $(F_u, F_v, d) \mapsto W$ defined in (7) is complete for the set W_S .

Remark 3.4. The above parameterization can be extended to a nonlinear layer

$$f(x) = U D_1 \phi (D_2 V^{\top} x), \tag{8}$$

where D_1, D_2 are diagonal matrices, and ϕ is a scalar activation with slope-restricted in [0, 1]. If $|D_1D_2| \leq S$, then the Jacobian $\partial f/\partial x$ has controlled rank and norm bound, i.e. $\partial f/\partial x \in \mathbb{W}_S$ for all $x \in \mathbb{R}^n$.

Proof. It is obvious that $W \in \mathbb{W}_S$. Now we show the converse part: for any $W \in \mathbb{W}_S$, there exist $\theta = (F_u, F_v, d)$ such that $W = \mathcal{W}(\theta)$. We consider the SVD decomposition $W = U_w \Sigma_w V_w^{\top}$. The difficulty is that Cayley transformation (5) does not cover the set of all orthogonal matrices: it omits orthogonal matrices with determinant -1. The combination of Lemmas A.1 and A.2 in Appendix A provide the solution: for any orthogonal matrix Q, one can find a diagonal matrix P with $P_{ii} \in \{-1,1\}$ such that QP can be covered by the Cayley transformation. Then, we rewrite W as

$$W = U_w P_w (P_w \Sigma_w P_v) (V_w P_v)^{\top} = U D V^{\top}, \quad (9)$$

where U, D, V can be used to recover the feasible θ . \square

Imposing the Norm Bound for NB-LoRA I. Based on (7), we construct a complete parameterization for $\mathbb{W}_{p}^{r,\delta}$ by

$$D = \operatorname{diag}\left(\frac{\delta d}{\max(\|d\|_p, 1)}\right). \tag{10}$$

Note that the singular values of W are completely determined by d. Despite the simplicity of this representation, it could restrict "flexibility" for some applications as the singular values only depend on a small set of parameters.

3.3 NB-LoRA II

We now present the second complete parameterization $W : \mathbb{R}^N \mapsto \mathbb{W}_S$, i.e., $W(\mathbb{R}^N) = \mathbb{W}_S$. It takes $F \in \mathbb{R}^{(m+n)\times r}$ as the learnable parameter and produces W as follows:

- 1) Apply the Cayley transformation G = Cayley(F) and then take matrix partition $G = [A B]^{\top}$ with $A \in \mathbb{R}^{r \times m}, B \in \mathbb{R}^{r \times n}$;
- 2) Construct W as follows

$$W = 2A^{\top}SB. \tag{11}$$

Theorem 3.5. If $r \leq \max(m, n)$, then (11) is a complete parameterization for W_S .

Remark 3.6. A special case of the above theorem is $S = I_m$, which is a complete parameterization of all 1-Lipschitz linear layer, i.e. f(x) = Wx with $||W||_2 \le 1$, see Proposition 3.3 of Wang and Manchester (2023).

Remark 3.7. It is worth to notice that $\max(m, n)$ is the exact upper bound of r for a complete \mathcal{W} . For example, we consider m = 1, n = 1, r = 2 and $S = I_r$. It is obvious that $W = 2A^{\top}SB \equiv 0$ since $\{A, B\}$ is an orthogonal basis of R^2 due to the Cayley transformation (5).

Remark 3.8. One can also apply (11) to construct a nonlinear layer with low-rank and norm-bounded Jacobian, i.e., $f(x) = 2A^{\top}D_1\phi(D_2Bx)$ where D_1, D_2 are diagonal. If $0 \leq D_1D_2 \leq S$, then $\partial f/\partial x \in \mathbb{W}_S$ for all $x \in \mathbb{R}^n$. However, unlike (8), large hidden dimension $r > \max(m, n)$ could restrict the expressivity of f.

Proof. Our proof consists of two parts: I) we show that $\mathcal{W}(F) \in \mathbb{W}_S$ for any $F \in \mathbb{R}^{(m+n)\times r}$; II) conversely, we show that for $W \in \mathbb{W}_S$, there exists an $F \in \mathbb{R}^{(m+n)\times r}$ such that $W = \mathcal{W}(F)$.

For Part I, we rewrite (11) as $W = 2Q^{\top}K$ with $Q = S^{1/2}A$ and $K = S^{1/2}B$, where $S^{1/2} = \operatorname{diag}(\sqrt{S_{jj}}) \in \mathbb{R}^{r \times r}$. Then, W has maximally r non-zero singular values. For any $1 \leq j \leq r$, we have

$$\sigma_j(W) = 2\sigma_j(Q^\top K) \le \sigma_j \left(QQ^\top + KK^\top\right)$$
$$= \sigma_j \left(S^{1/2}G^\top GS^{1/2}\right) = \sigma_j(S), \tag{12}$$

where the inequality is the matrix arithmetic-geometric mean inequality Bhatia and Kittaneh (1990); Bhatia (2013), and the last equality follows by $G^{\top}G = I$ which is ensured by the Cayley transformation (5).

The poof of Part II is constructive, i.e., find feasible A, B from W and then compute the corresponding free

parameters X, Y by "inverting" the Cayley transformation. Without loss of generality, we assume that the diagonal elements of S is in descending order, i.e., $\sigma_j(S) = S_{jj}$ for $j = 1, \ldots, r$. We first consider the low-rank case, i.e., $r \leq \min(m, n)$.

We take the reduced SVD decomposition $W = U_w \Sigma_w V_w^{\top}$ where $U_w \in \mathbb{R}^{m \times r}, V_w \in \mathbb{R}^{n \times r}$ are semi-orthogonal, and the positive diagonal matrix $S_w \in \mathbb{R}^{r \times r}$ satisfies $S_w \preceq_{\sigma} S$ by assumption. We consider the following candidates:

$$A = P\Sigma_a U_w^{\top}, \quad B = P\Sigma_b V_w^{\top}, \tag{13}$$

where $P \in \mathbb{R}^{r \times r}$ is a diagonal matrix with $P_{jj} = \pm 1$, and $\Sigma_a, \Sigma_b \in \mathbb{R}^{r \times r}$ are positive diagonal matrices. Since $G = \begin{bmatrix} A & B \end{bmatrix}^{\top}$ is semi-orthogonal, we have

$$I = AA^{\top} + BB^{\top} = P(\Sigma_a^2 + \Sigma_b^2)P^{\top} = \Sigma_a^2 + \Sigma_b^2$$
. (14)

And furthermore, $W = 2A^{T}SB$ implies that

$$U_w \Sigma_w V_w^{\top} = 2U_w \Sigma_a P^{\top} S P \Sigma_b V_w^{\top} = U_w (2\Sigma_a \Sigma_b S) V_w^{\top}.$$

By defining $J = \Sigma_w/S$, we obtain

$$2\Sigma_a \Sigma_b = J. \tag{15}$$

Since $J_{kk} \in [0, 1]$, Eq. (14) and (15) yield a solution of

$$\Sigma_a = \frac{\sqrt{I+J} + \sqrt{I-J}}{2}, \ \Sigma_b = \frac{\sqrt{I+J} - \sqrt{I-J}}{2}.$$

To determine the matrix P in (13), we first define the matrices $U \in R^{r \times r}, V \in \mathbb{R}^{(m+n-r) \times r}$ based on the partition $G = \begin{bmatrix} U^\top & V^\top \end{bmatrix}^\top$. We then pick up P such that I+U is invertible. Note that U depends on P through A. Lemma A.1 shows that such choice always exists. Finally, we compute X,Y via

$$X = \frac{1}{2}Z, \quad Y = -\frac{1}{2}V(I+Z),$$

with $Z = (I + U)^{-1}(I - U)$. From Lemma A.2 we can conclude that (5) holds, which further leads to $\mathcal{W}(F) = W$ with $F = \begin{bmatrix} X^\top & Y^\top \end{bmatrix}^\top$.

For the full-rank case, without loss of generality we assume that $m \geq r \geq n$. Given an SVD decomposition $W = U\Sigma V^{\top}$ with $U \in \mathbb{R}^{m \times n}$ and $\Sigma, V \in \mathbb{R}^{n \times n}$, we construct an "augmented" SVD decomposition as

$$W = U_w \Sigma_w V_w^\top = \begin{bmatrix} U & \hat{U} \end{bmatrix} \begin{bmatrix} \Sigma & \\ & \mathbf{0}_{r-n} \end{bmatrix} \begin{bmatrix} V^\top \\ \mathbf{0}_{(r-n) \times n} \end{bmatrix},$$

where $\hat{U} \in \mathbb{R}^{m \times (r-n)}$ satisfies $U_w^{\top} U_w = I_r$. We take the A, B candidates in (13) with $\Sigma_a = \operatorname{diag}(\hat{\Sigma}_a, I_{r-n})$

and $\Sigma_b = \operatorname{diag}(\hat{\Sigma}_b, \mathbf{0}_{r-n})$, where $\hat{\Sigma}_a, \hat{\Sigma}_b \in \mathbb{R}^{n \times n}$ are positive diagonal matrices. By following the similar steps in the low-rank case, we can obtain

$$\hat{\Sigma}_a^2 + \hat{\Sigma}_b^2 = I_n, \quad 2\hat{\Sigma}_a\hat{\Sigma}_b = \hat{J} := \Sigma/\hat{S},$$

where \hat{S} is the upper-left $n \times n$ -block of S. By assumption we have $\hat{J}_{jj} \in [0,1]$. We then follow the same procedure in the low-rank case to recover F such that $\mathcal{W}(F) = W$.

Imposing the Norm Bound for NB-LoRA II. By following a similar procedure as NB-LoRA I, we construct a complete parameterization for $\mathbb{W}_p^{r,\delta}$ based on (11) with

$$S = \operatorname{diag}\left(\frac{\delta |d|}{\max(\|d\|_p, 1)}\right).$$

Here $d \in \mathbb{R}^r$ is a learnable parameter. In contrast to (10), S is a non-negative diagonal matrix.

Comparison of NB-LoRA I and II. Both parameterizations are complete. The major difference is in NB-LoRA I the parameterization is intuitively-related to SVD in the sense that U and V have orthonormal columns and the singular values are entirely determined by D. On the other hand, with NB-LoRA II, the singular values of W are determined by both the weights A, B and the bound S, see (13) - (15). This offer us extra "flexibility" to adapt the singular values, particularly when small norm bound is imposed.

Choice of Norm Bound. In PEFT we typically bound the adaptation weight norm by normalizing with the pre-trained weight, i.e. $\|W\|_{S_p}/\|W_{pt}\|_{S_p} \leq \gamma$ where $\gamma > 0$ is a hyperparameter. I.e. we set the norm bound as take the norm bound δ as $\delta = \gamma \|W_{pt}\|_{S_p}$.

4 Experiments

To illustrate the potential applications of NB-LoRA, we conducted a series of experiments in PEFT of a vision transformer (ViT) model. In particular, we explore adaptation performance vs forgetting and hyperparameter robustness, and compare to standard LoRA and other PEFT methods. Then we investigate the application of NB-LoRA to privacy-preserving model merging, and finally discuss some effects of the parameterizations via a simple low-rank matrix completion problem.

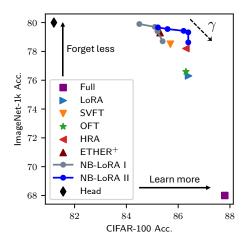


Figure 1: Parameter Efficient Fine Tuning of a ViT: adaption performance on CIFAR-100 vs model forgetting on ImageNet-1k. Compared with other methods, the proposed NB-LoRA II lies on the frontier of topright corner. Our model uses rank of 8 and nuclear norm with ratio bound of $\gamma = 0.005, 0.01, 0.03, 0.2$.

4.1 PEFT: Adaptation vs Forgetting

The main goal of this experiment is to explore the utility of norm bounds in preventing catastrophic model forgetting McCloskev and Cohen (1989); French (1999); Wang et al. (2024). Our hypothesis is that tight control of the adaption norm will prevent loss of performance on the pre-trained model, while still enabling good adaptation performance. We perform experiments on ViT-B/16 model Dosovitskiy et al. (2020), which is pre-trained on ImageNet-21k Deng et al. (2009) and then fine-tuned to ImageNet-1k. Similar to the setup in Kopiczko et al. (2024), we adapt Q, V matrices and learn the classification head for other vision benchmarks, including CIFAR-100 Krizhevsky et al. (2009), Food-101 Bossard et al. (2014) and CUB-200-2011 Wah et al. (2011). The **metric for model** forgetting is the test accuracy of the fine-tuned model on ImageNet-1k.

We compare our approach with **Head** learning which only learns the head, **Full** fine-tuning which updates Q, V matrices, the standard **LoRA** Hu et al. (2022), as well as other methods with the capability of controlling certain weight distance, including **SVFT** Lingam et al. (2024), **OFT** Qiu et al. (2023), **HRA** Yuan et al. (2024) and **ETHER**⁺ Bini et al. (2024). We denote our model with rank r and Schatten-p norm ratio bound γ by **NB-LoRA** $_p^{r,\gamma}$, where p,r or γ is omitted if it can be inferred from the context.

Figure 1 shows that there is generally a trade-off between good adaptation to the target (CIFAR-100) and preventing forgetting of pre-training (ImageNet-1k). Fixing the Q,V matrices and only updating the head forgets little, but has poor adaptation performance, whereas a full update of all Q,V weights has excellent adaptation performance but with significant forgetting. Among existing PEFT methods, LoRA, OFT, HRA, achieve quite good adaptation performance but with substantial forgetting (less severe for HRA), whereas ETHER⁺ and SVFT forget less but have diminished adaptation performance.

For both of the proposed NB-LoRA parameterizations, we can see that increasing γ allows one to trade off adaptation performance against greater forgetting. For this setup, NB-LoRA II appears to uniformly outperform NB-LoRA I in terms of this trade-off, and so we use NB-LoRA II in the remaining PEFT experiments. When the norm bound γ is near 0.03, NB-LoRA II achieves similar adaptation performance to LoRA, OFT, and HRA, but forgets substantially less. As γ is varied, it generally remains on the upper-right frontier of the performance trade-off.

Figure 2 shows in detail how the adaptation performance, forgetting, and adaptation norm evolve vs training epochs for the different methods. It can be seen that the norm of NB-LoRA is substantially lower than the other methods and remains below the prescribed bound. The figure also illustrates all the singular values at a particular layer, where our model lies in the bottom left. Similar figures for Food-101 and CUB-200-2011 datasets can be found in the appendix (Figure 6).

Table 1 provides a more extensive set of results including the Food-101 and CUB-200-2011 data sets as targets, and the performance of NB-LoRA with Frobenius and spectral norm bounds (Schatten 2 and ∞ -norm) in addition to nuclear norm (Schatten 1-norm) used above. It can be seen that the above observations are consistent across target datasets, and that the choice of norm has minimal effect, although slightly better adaptation was achieved with the nuclear norm and slightly lower forgetting with the Frobenius norm.

4.2 PEFT: Hyper-Parameter Robustness

In the next set of experiments, we explore the potential benefits of norm bounds for reducing the dependence on model and training hyperparameters. Again we are adapting from ImageNet-1k to CIFAR-100. We compared NB-LoRA (with nuclear, Frobenius, and spectral

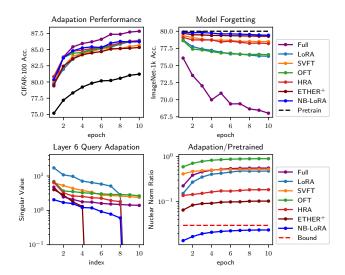


Figure 2: Test accuracy of PEFT on CIFAR-100 and ImageNet-1k, as well as the nuclear norm ratio (adaption/pre-trained) during the training.

norm) to standard LoRA in terms of the chosen rank of the adaptation, the learning rate used during training, and the number of epochs used in training.

Table 2 shows the results. It can be seen that LoRA is very sensitive to hyperparameters: increasing the rank resulted in greater forgetting and eventually unstable training. Increasing the learning rate had a similar effect. Furthermore, training for longer actually decreased performance on both the target dataset (CIFAR-100) and the source dataset (ImageNet-1k), most significantly on the latter.

In contrast, the performance of NB-LoRA is remarkably consistent for a wide range of ranks, learning rates, and number of training epochs, with results on both source and target generally varying less than 1% across the different combinations and always outperforming LoRA on one or (usually) both of source and target datasets. This consistency was observed with each of nuclear, Frobenius, and spectral norm.

The appendix contains further results on the Food-101 and CUB-200-2011 datasets (Table 3), for which similar fragility of LoRA and robustness of NB-LoRA can be observed.

4.3 Differentially-Private Model Merging

There has been recent interest in using PEFT-derived methods such as LoRA for merging multiple models trained on separate data sets, see e.g. Stoica et al.

Table 1: Performance on image classification tasks - CIFAR-100 (C100), Food101 (F101), and CUB-200-2011 (C200). We only adapt Q, V matrices for all methods, following prior work Kopiczko et al. (2024). We report the adaption performance (test accuracy on the target dataset), model forgetting (test accuracy on the source dataset ImageNet-1k, denoted by IN1k), and the maximum norm ratio over all adaption blocks. We observe that the adaptation performance of our approach is close to the standard LoRA Hu et al. (2022) while NB-LoRA forgets much less.

Target Data	CIFAR-100							Food10	1		CUB-200-2011						
Model	Model Size		IN1k	1k Max Ratio $S_1/S_2/S_{\infty}$			F101	IN1k	Max Ratio $S_1/S_2/S_{\infty}$			C200	IN1k	Max B	atio $S_1/S_2/S_\infty$		
Head Full	- 14.2M	81.2 87.8	80.0 68.0	0.553	0.512	1.470	77.1 86.8	80.0 64.2	0.790	0.626	1.432	73.0 80.8	80.0 66.7	0.582	0.518	1.367	
SVFT	98K 166K	85.7 86.3	78.5 76.6	0.521 0.891	0.715 0.825	1.993 0.868	82.8 83.5	77.9 75.7	0.540 1.006	0.912 0.926	2.139 0.922	76.7 77.5	78.6 77.5	0.407	0.619 0.569	1.643 0.614	
HRA	147k	86.3	78.2	0.184	0.426	1.245	83.9	76.8	0.149	0.364	1.126	77.4	78.8	0.134	0.287	0.942	
${ m ETHER^+} \ { m LoRA}$	74K 295K	85.3 86.4	$79.3 \\ 76.3$	$0.103 \\ 0.475$	$0.322 \\ 2.254$	0.697 4.987	82.8 84.8	78.7 72.8	0.088 0.682	0.306 2.189	$0.641 \\ 6.019$	76.3 78.5	$79.3 \\ 77.4$	0.053	$0.250 \\ 0.692$	$0.491 \\ 3.075$	
NB-LoRA $_{p=1}^{\gamma=0.03}$ NB-LoRA $_{p=2}^{\gamma=0.08}$	295K 295K	86.2 85.5	79.4 79.6	0.024 0.027	0.176 0.072	0.822 0.388	84.2 83.5	77.8 78.7	0.029 0.025	0.215 0.078	1.068 0.425	78.1 77.3	79.1 79.5	0.018	0.128 0.045	0.721 0.266	
NB-LoRA $_{p=\infty}^{\gamma=0.3}$	295K	85.8	79.6	0.094	0.215	0.277	83.8	78.6	0.090	0.268	0.280	77.5	79.4	0.036	0.109	0.178	

(2024). A variant of this setting is one in which multiple small *private* data-sets are to be used to train models for merging, but under the requirement that these individual contributions to the merged model can not be inferred from the final model weights.

In this section we explore the benefits of norm-bounded LoRA for privacy preserving model merging. The main intuition is that privacy can be preserved by adding noise proportional to an a-priori bound on the norm of the adaptation weights. Hence if good adaptation performance can be achieved with constrained norm, then a smaller noise signal needs to be added (degrading performance less) to achieve a given privacy budget.

To illustrate this we split the CIFAR-100 training set into K=10 "private" sets with disjoint categories, and fine-tune ViT models separately on these. We then merge the models like so:

$$W_{ft} = W_{pt} + \frac{1}{K} \sum_{k=1}^{K} W_k + \sigma B \eta_k,$$

where η_k are i.i.d. Gaussian random matrices of the same dimension as W_k , B is an a-priori bound on the Frobenius norm of the weights: $\|W_k\|_{S_2} \leq B$, and σ is a chosen scalar. In this setting it can be shown that the fine-tuned weight W_{ft} is $\frac{1}{\sigma}$ -Gaussian differentially private with respect to the individual W_k Dong et al. (2022). We then test the resulting merged model on the full CIFAR-100 test set.

We trained each of the W_k via standard LoRA and via NB-LoRA II with a bound of $B = 0.01 \|W_{pt}\|_{S_2}$. For standard LoRA, we project the resulting weights

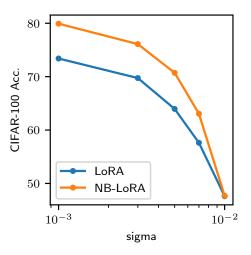


Figure 3: Accuracy of privacy-preserving model merge on the target dataset (CIFAR-100). NB-LoRA uniformly outperforms LoRA at each noise level σ , corresponding to a given privacy budget.

onto the ball $||W_k||_{S_2} \leq B$ to achieve the same privacy guarantees.

The results in Figure 3 show that both methods exhibit deteriorating performance as the noise level σ increases, as expected, but that NB-LoRA has uniformly outperforms LoRA, with more than 5% performance gap for lower noise levels. This indicates that for a given privacy budget, substantially better downstream performance can be achieved by incorporating norm bounds during training, rather than projecting after training.

Table 2: Robustness comparison for different hyper-parameters (H.P.): rank, learning rate (LR), and number of training epochs. We take Rank=16, LR=0.005 and Epochs=10 as the default setup and change one parameter in each experiment. The norm ratio bound γ in our model is chosen to achieve similar adaptation performance as LoRA in the default setup. We report the adaption performance on CIFAR-100 (C100), model forgetting on ImageNet-1k (IN1k), as well as the maximum norm ratio over all LoRA blocks. The item "failed" in the table means that the model training failed to converge. Due to the norm-bound constraint, our model can learns weight adaptations with much smaller norm than LoRA. Moreover, our models achieve more consistent results on both the source and target datasets than LoRA across a wide range of hyper-parameters.

Н. Р.			NB-LoRA $_{p=1}^{\gamma=0.03}$						$NB-LoRA_{p=2}^{\gamma=0.08}$						NB-LoRA $_{p=\infty}^{\gamma=0.3}$					
Rank	C 100 IN1k Max Ratio $S_1/S_2/S_{\infty}$			C100	IN1k	Norm	Ratio S_1	$/S_2/S_{\infty}$	C100	IN1k	Max F	tatio S_1	$/S_2/S_{\infty}$	C100	IN1k Max Ratio $S_1/S_2/S$			$/S_2/S_{\infty}$		
8	86.4	76.3	0.475	2.254	4.987	86.2	79.4	0.024	0.176	0.822	85.8	79.6	0.027	0.072	0.388	85.8	79.6	0.094	0.215	0.277
16	86.9	74.8	0.590	2.479	4.982	86.5	79.5	0.024	0.138	0.660	86.2	79.6	0.035	0.073	0.364	86.3	79.4	0.146	0.301	0.275
32	86.7	71.2	1.010	2.996	5.284	86.3	79.5	0.026	0.106	0.353	86.2	79.5	0.044	0.075	0.250	86.7	79.1	0.235	0.384	0.271
64	failed	failed	failed	failed	failed	86.3	79.4	0.027	0.078	0.284	86.3	79.4	0.053	0.077	0.269	86.9	78.8	0.351	0.507	0.256
384	failed	failed	failed	failed	failed	85.8	79.6	0.023	0.038	0.074	86.1	79.5	0.093	0.073	0.174	87.4	77.3	1.440	1.044	0.257
L. R.	C100	IN1k	Max R	tatio S_1	S_2/S_{∞}	C100	IN1k	Max l	Ratio S_1	$/S_2/S_{\infty}$	C100	IN1k	Max F	tatio S_1	$/S_2/S_{\infty}$	C100	IN1k	Max R	Ratio S_{1}	$/S_2/S_{\infty}$
0.002	87.1	77.0	0.262	1.363	3.532	86.1	79.3	0.019	0.104	0.544	86.2	79.6	0.026	0.063	0.309	86.1	79.4	0.107	0.206	0.244
0.005	86.9	74.8	0.590	2.479	4.982	86.5	79.5	0.024	0.138	0.660	86.2	79.6	0.035	0.073	0.364	86.3	79.4	0.146	0.301	0.275
0.010	failed	failed	failed	failed	failed	86.5	79.3	0.027	0.150	0.661	86.3	79.4	0.039	0.077	0.376	86.4	79.1	0.180	0.364	0.286
0.020	failed	failed	failed	failed	failed	86.6	79.0	0.028	0.167	0.732	86.3	79.3	0.041	0.079	0.493	86.4	78.7	0.220	0.400	0.297
Epochs	C100	C100 IN1k Max Ratio $S_1/S_2/S_\infty$			C100	IN1k	Max l	Ratio S_1	$/S_2/S_{\infty}$	C100	C100 IN1k Max Ratio $S_1/S_2/S_{\infty}$				C100	IN1k	N1k Max Ratio $S_1/S_2/S$			
10	86.9	74.8	0.590	2.479	4.982	86.5	79.5	0.024	0.138	0.660	86.2	79.6	0.035	0.073	0.364	86.3	79.4	0.146	0.301	0.275
20	86.8	72.1	0.926	3.094	5.652	86.5	79.0	0.026	0.147	0.598	86.0	79.3	0.035	0.076	0.375	86.4	78.9	0.166	0.347	0.284
30	86.5	70.1	1.100	3.167	5.768	85.9	78.7	0.026	0.145	0.619	85.5	79.1	0.036	0.076	0.372	86.1	78.5	0.168	0.345	0.284
40	86.1	69.1	1.140	3.094	6.458	86.4	78.6	0.025	0.142	0.593	85.6	79.0	0.036	0.075	0.440	85.7	78.2	0.163	0.339	0.282

4.4 Low-Rank Matrix Completion with Nuclear Norm Bound or Penalty

Low rank matrix completion is a basic problem that appears in many applications, see e.g. Recht et al. (2010) and references therein. Commonly, the nuclear norm is minimized as a proxy for rank, subject to constraints on the known entries. This can be expressed as a semidefinite program Candes and Recht (2012); however, for very large matrices this can be intractible, and first-order methods based on low-rank parameterizations can be applied; see, e.g., Cai et al. (2010); Ma et al. (2011); Mishra et al. (2013). In particular, Mishra et al. (2013) considers the situation in which the rank is expected to be low but its precise value is not known. The matrix is parameterized to have "moderately" low rank and a nuclear norm penalty is applied to further reduce the rank.

To be precise, suppose certain elements of W are known (possibly with some noise): $w_{ij} = \tilde{w}_{ij}$ for $(i, j) \in \mathcal{I}$ where I is some subset of indices. Then one optimizes

$$\min_{W} \sum_{(i,j)\in\mathcal{I}} |w_{ij} - \tilde{w}_{ij}|^2 + \gamma ||W||_{S_1}$$
 (16)

over low-rank W. Alternatively, $||W||_{S_1}$ can be bounded at some chosen level, and fit minimized.

The advantage of our parameterizations is that (16) becomes an unconstrained minimization over \mathbb{R}^N which can be tackled via standard gradient methods, and does not require any projections or SVD computations at each step, unlike (Cai et al., 2010; Ma et al., 2011; Mishra et al., 2013).

We solve Problem (16) with NB-LoRA I and II to illustrate some properties of these paramaterizations. The matrix W was 150×100 , the true rank was 10 and nuclear norm of 390, and 20% of entries were unknown, and the known entries have additive noise. The chosen rank of the parameterization was 20.

Starting with a norm bound, in Figure 4 it can be seen that the bound increases, the fit error of known entries decreases smoothly until it reaches a minimum at approximately the true norm bound, after which it remains flat. This indicates that regularization with a norm penalty will be effective, as illustrated in the middle-left plot. In the upper-right we show that the test error has a clear minimum at the true value of the nuclear norm (390), and the numerical rank increases from the true value of 10 at around the same point.

The two paramaterizations for NB-LoRA generally perform similarly. However, we can observe some differences: the test error is significantly better for NB-LoRA II when the norm bound is overestimated. In

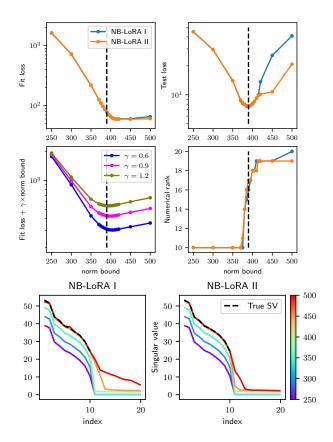


Figure 4: Behavior of NB-LoRA parameterizations (I and II) on the low-rank matrix completion problem with nuclear norm bounds. See Section 4.4 for discussion.

the lower two plots, it can be observed that the "tail" singular values (i.e. $\sigma_{11}, ..., \sigma_{20}$) remain lower with NB-LoRA II than NB-LoRA I for higher values of the norm bound, i.e. parameterization II seems to have a greater bias towards low rank solutions.

Imposing a nuclear norm penalty (with $\gamma=0.9$) rather than a bound, we see in Figure 5 that both parameterizations easily find a solution with low fit and test error, and with nuclear norm approximately equal to the true value. The found solutions also match the true singular values well. In this example, NB-LoRA II appears to converge slightly faster, but the difference is minor.

5 Conclusion

In this paper we propose that low rank and normbounded weight perturbations be used for model fine tuning. To this end, we introduce two model parameterizations (NB-LoRA I and II) which are complete, i.e.

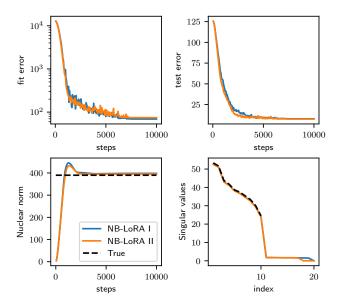


Figure 5: Performance of NB-LoRA I and II on the lowrank matrix completion problem with a nuclear norm penalty. Both fit error and test error smoothly converge. The nuclear norm converges to a value close to its true value. At the final step, all singular values closely match their true values.

they cover all matrices of a specified rank and Schatten *p*-norm bound.

In experiments on fine tuning, we compare to standard LoRA and other existing methods and demonstrate that norm bounds can reduce source model forgetting without impacting target adaptation. We also observed significantly improved robustness to hyperparameters such as rank, learning rate, and number of training epochs.

We also argue that controlling both matrix rank and norm is a basic problem in learning and optimization. To illustrate this, we also applied NB-LoRA to a privacy-preserving model merging problem, in which substantial increases in accuracy were observed for a given privacy budget, and a low-rank matrix completion problem, in which the proposed parameterization proved easy to apply and effective.

References

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

Bhatia, R. (2013). Matrix analysis, volume 169.

- Springer Science & Business Media.
- Bhatia, R. and Kittaneh, F. (1990). On the singular values of a product of operators. SIAM Journal on Matrix Analysis and Applications, 11(2):272–277.
- Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. (2024). LoRA learns less and forgets less. Transactions on Machine Learning Research.
- Bini, M., Roth, K., Akata, Z., and Khoreva, A. (2024). Ether: Efficient finetuning of large-scale models with hyperplane reflections. In *ICML*.
- Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101-mining discriminative components with random forests. In Computer vision-ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13, pages 446-461. Springer.
- Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.
- Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.
- Chen, J., Zhang, A., Shi, X., Li, M., Smola, A., and Yang, D. (2023). Parameter-efficient fine-tuning design spaces. In *International Conference on Learning Representations*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. IEEE.
- Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37.
- Dong, W., Sun, Y., Yang, Y., Zhang, X., Lin, Z., Yan, Q., Zhang, H., Wang, P., Yang, Y., and Shen, H. T. (2024). Efficient adaptation of pre-trained vision transformer via householder transformation. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani,

- M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.
- Gouk, H., Hospedales, T., et al. (2021). Distance-based regularisation of deep networks for fine-tuning. In *International Conference on Learning Representations*.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on ma*chine learning, pages 2790–2799. PMLR.
- Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. In *ICLR*.
- Kopiczko, D. J., Blankevoort, T., and Asano, Y. M. (2024). Elora: Efficient low-rank adaptation with random matrices. In *The Twelfth International Con*ference on Learning Representations.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Lingam, V., Neerkaje, A. T., Vavre, A., Shetty, A., Gudur, G. K., Ghosh, J., Choi, E., Dimakis, A., Bojchevski, A., and sujay sanghavi (2024). SVFT: Parameter-efficient fine-tuning with singular vectors. In 2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024).
- Liu, S.-y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. (2024a). Dora: Weight-decomposed low-rank adaptation. In Forty-first International Conference on Machine Learning.
- Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., Feng, H., Liu, Z., Heo, J., Peng, S., et al. (2024b). Parameter-efficient orthogonal finetuning via butterfly factorization. In *The Twelfth International Con*ference on Learning Representations.
- Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

- Ma, S., Goldfarb, D., and Chen, L. (2011). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353.
- McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Mishra, B., Meyer, G., Bach, F., and Sepulchre, R. (2013). Low-rank optimization with trace norm penalty. SIAM Journal on Optimization, 23(4):2124–2149.
- Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., and Schölkopf, B. (2023). Controlling text-to-image diffusion by orthogonal finetuning. Advances in Neural Information Processing Systems, 36:79320-79362.
- Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.
- Stoica, G., Ramesh, P., Ecsedi, B., Choshen, L., and Hoffman, J. (2024). Model merging with svd to tie the knots. arXiv preprint arXiv:2410.19735.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.
- Wang, L., Zhang, X., Su, H., and Zhu, J. (2024). A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pat*tern Analysis and Machine Intelligence.
- Wang, R. and Manchester, I. (2023). Direct parameterization of lipschitz-bounded deep networks. In *ICML*.
- Winston, E. and Kolter, J. Z. (2020). Monotone operator equilibrium networks. *Advances in neural information processing systems*, 33:10718–10728.
- Yuan, S., Liu, H., and Xu, H. (2024). Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. (2023). Adaptive budget allocation for parameter-efficient fine-tuning. In The Eleventh International Conference on Learning Representations.

A Key Technical Lemmas

Here we present two key lemmas which are used in the proof of the main theorems.

Lemma A.1. For any $Q \in \mathbb{R}^{r \times r}$, there exists a diagonal matrix P with $P_{jj} \in \{-1,1\}$ such that $I + PQ^{\top}$ is invertible.

Proof. We prove it by construction. Let e_k, q_k be the kth column of I and Q, respectively. We set $A_1 = I$ and construct A_2, \ldots, A_{n+1} via

$$A_{k+1}^{-1} = A_k^{-1} - \frac{s_k A_k^{-1} e_k q_k^{\top} A_k^{-1}}{1 + s_k q_k^{\top} A_k^{-1} e_k}, \tag{17}$$

where $s_k = \operatorname{sgn}(v_k^\top A_k^{-1} e_k)$ with $\operatorname{sgn}(0) = 1$. From Sherman-Morrison formula, A_{k+1} is well-defined (i.e., invertible) and satisfies $A_{k+1} = A_k + s_k e_k q_k^\top$. Thus,

$$A_{n+1} = I + \sum_{k=1}^{n} s_k e_k q_k^{\top} = I + PQ^{\top}, \tag{18}$$

is also invertible, where $P = \operatorname{diag}(s_1, \ldots, s_n)$.

Lemma A.2. Let $G \in \mathbb{R}^{(r+s)\times r}$ be a semi-orthogonal matrix with partition $G = \begin{bmatrix} U \\ V \end{bmatrix}$ with $U \in \mathbb{R}^{r\times r}$ and $V \in \mathbb{R}^{s\times r}$. Then, G = Cayley(F) for some $F \in \mathbb{R}^{(r+s)\times r}$ if and only if I + U is invertible.

Proof. From the Cayley transformation (5) we have the following relationships:

$$U = (I - Z)(I + Z)^{-1}, \quad V = -2Y(I + Z)^{-1}, \quad Z = X - X^{\top} + Y^{\top}Y.$$
(19)

(if). From the above equation we have $I + U = (I + Z)^{-1}$ invertible.

(**only if**). The proof is constructive, i.e., finding $X, Z \in \mathbb{R}^{k \times k}$ and $Y \in \mathbb{R}^{(n-k) \times k}$ satisfying (19). We consider the following candidate solution

$$Z = (I+U)^{-1}(I-U), \quad Y = -\frac{1}{2}V(I+Z), \quad X = \frac{1}{2}Z.$$
 (20)

Since the first two equations in (19) holds, we then verify the last equation via

$$\begin{split} Z + X^\top - X - Y^\top Y &= \frac{1}{2} (Z + Z^\top) - Y^\top Y \\ &= \frac{1}{2} [(I + U)^{-1} (I - U) + (I - U^\top) (I + U^\top)^{-1}] - (I + u^\top)^{-1} V^\top V (I + U)^{-1} \\ &= \frac{1}{2} [(I - U) (I + U)^{-1} + (I + U^\top)^{-1} (I - U^\top)] - (I + U^\top)^{-1} V^\top V (I + U)^{-1} \\ &= \frac{1}{2} (I + U^\top)^{-1} [(I + U^\top) (I - U) + (I - U^\top) (I + U) - 2 V^\top V] (I + U)^{-1} \\ &= (I + U^\top)^{-1} [I - U^\top U - V^\top V] (I + U)^{-1} = 0, \end{split}$$

where the second line is due to that $(I+U)^{-1}$ and (I-U) are commutative, the last line follows by $U^{\top}U+V^{\top}V=I$.

B Training Details

ViT adaption. We take the ViT-B/16 model Dosovitskiy et al. (2020) and insert adaption blocks into the Q, V matrices, following the setup in Kopiczko et al. (2024). For the model **HRA**, we take 8 Householder transformation. In **OFT**, we parameterize the orthogonal matrix with 96 diagonal blocks where each block is 8×8 matrix. In **ETHER**⁺, we both left- and right-multiply the weight with matrices based on hyperplane reflection. The effective rank is 4. For **SVFT**, we use the parameterization $W = U\Sigma V^{\top}$ with $\Sigma \in \mathbb{R}^{p \times p}$ where U, V are frozen parameters obtained from the SVD of W_{pt} . We trained the model with p = 8, 16, 32, 64 and found that p = 64 yields similar performance as other methods when p = 64. For our model, we re-parameterize X, Y in $Z = X - X^{\top} + Y^{\top}Y$ by $g\frac{X}{\|X\|_F}$ and $h\frac{Y}{\|Y\|_F}$ with learnable scalars g, h. The main reason is that the Cayley transform in (5) involves both linear and quadratic terms, see details in Winston and Kolter (2020). We choose AdamW Loshchilov and Hutter (2019) as the optimizer with default learning rate of 5e-3 and weight decay of 0.01. For the full fine-tuning, we reduce the learning rate to 5e-4. We take one-cycle learning rate scheduler with warm-up ratio of 0.1.

Model merging. We split the CIFAR-100 dataset into 10 disjoint subset, where each subset is used to learn a fine-tuned ViT-base model through rank 16 adaptation. We use Frobenius norm with ratio bound of 0.01 for our model. To perform DP model merging on LoRA, we project the weight update if its Frobenius exceeds the norm budget in our model. We use the same training setup as the previous ViT adaption experiment.

Matrix completion. We construct $\tilde{W} \in \mathbb{R}^{m \times n}$ with m = 150, n = 100 via $\tilde{W} = W_t + W_n := A^{\top}B + W_n$ where $A \in \mathbb{R}^{r \times m}, B \in \mathbb{R}^{r \times n}$ with r = 10 are random matrices drawn from uniform distribution. W_n is is a random matrix drawn from Gaussian distribution with zero mean and std of 0.1. Note that W_n often has full rank but its singular values are relatively small compared with W_t . We then obtain the dataset by randomly dropping 20% elements of \tilde{W} . For reconstruction, we parameterize our model with rank of 2r. We use AdamW with learning rate of 0.1 as the default optimizer. We implement one-cycle rate scheduler with warm-up ratio of 0.1.

C Additional Results

Table 3: Hyper-parameter robustness for Food-101 and CUB-200-2011 datasets. Simiar to CIFAR-100, we take Rank=16, LR=0.005 and Epochs=10 as the default setup and change one parameter in each experiment. We report the adaption performance on the target dataset, model forgetting on the source dataset, as well as the maximum norm ratio over all LoRA blocks. Due to the small norm bound constraint, our model is more robust then the standard LoRA across a wide range of hyper-parameters.

	Food-101												CUB-200-2011										
Н. Р.	LoRA						$\text{NB-LoRA}_{p=1}^{\gamma=0.02}$							NB-LoRA $_{p=1}^{\gamma=0.02}$									
Rank	F101	IN1k	11k Max Ratio $S_1/S_2/S_{\infty}$			F101	7101 IN1k Norm Ratio $S_1/S_2/S_{\infty}$				C200	IN1k Max Ratio $S_1/S_2/S_\infty$					IN1k	Max R	tatio S_1	S_2/S_{∞}			
8 16 32 64	84.8 85.2 86.0 failed	72.8 68.7 63.2 failed	0.682 1.060 1.760 failed	2.189 2.599 4.071 failed	6.019 6.490 6.913 failed	84.0 84.0 83.8 83.2	78.2 78.6 79.1 79.5	0.019 0.019 0.019 0.019	0.146 0.108 0.079 0.059	0.778 0.366 0.235 0.160	78.5 78.4 77.9 77.3	77.4 76.1 74.0 68.0	0.258 0.455 0.723 1.304	0.692 1.182 1.444 2.836	3.075 4.062 3.949 6.695	77.9 77.7 78.0 77.6	79.2 79.2 79.3 79.5	0.012 0.012 0.013 0.014	0.087 0.072 0.055 0.048	0.525 0.310 0.199 0.137			
 L. R.	failed F101	failed IN1k	failed Max F	failed Ratio $S_{1/2}$	failed $/S_2/S_{\infty}$	82.3 F101	79.8 IN1k	0.015 Max 1	Ratio S_1	$\frac{0.056}{/S_2/S_{\infty}}$	failed C200	failed IN1k	failed Max R	failed atio S_1 /	failed S_2/S_∞	76.3 C200	79.7 IN1k	0.013 Max R	0.017 Latio S_1	$\frac{0.036}{S_2/S_{\infty}}$			
0.002 0.005 0.010 0.020	85.5 85.2 84.9 failed	73.6 68.7 59.5 failed	0.388 1.060 2.766 failed	1.271 2.599 8.449 failed	3.992 6.491 15.785 failed	83.8 84.0 83.9 83.5	78.8 78.6 77.9 76.7	0.016 0.019 0.020 0.020	0.094 0.108 0.110 0.109	0.322 0.366 0.417 0.484	79.9 78.4 76.2 failed	77.5 76.1 70.1 failed	0.151 0.455 0.972 failed	0.410 1.182 2.088 failed	1.818 4.062 5.959 failed	78.1 77.7 75.8 74.0	79.3 79.2 79.3 79.2	0.010 0.012 0.013 0.015	0.055 0.072 0.077 0.087	0.261 0.310 0.461 0.676			
Epochs	C100	IN1k	Max Ratio $S_1/S_2/S_{\infty}$		S_2/S_{∞}	C100	100 IN1k Max Ratio $S_1/S_2/S_{\infty}$		$/S_2/S_{\infty}$	C100	IN1k	Max Ratio $S_1/S_2/S_\infty$		C100	IN1k	Max Ratio $S_1/S_2/S$		S_2/S_{∞}					
10 20 30 40	85.2 85.2 85.1 84.9	68.7 65.0 62.0 60.0	1.060 1.419 1.624 1.612	2.599 3.531 3.943 3.759	6.491 8.151 8.580 8.213	84.0 83.9 83.6 83.7	78.6 78.1 77.9 77.6	0.019 0.020 0.020 0.020	0.108 0.109 0.109 0.109	0.366 0.401 0.413 0.414	78.4 78.4 76.6 76.9	76.1 74.3 72.7 72.1	0.455 0.545 0.870 0.704	1.182 1.325 2.227 1.691	4.062 4.057 4.657 4.308	77.7 77.0 76.3 75.7	79.2 78.9 78.6 78.5	0.012 0.013 0.014 0.014	0.072 0.078 0.079 0.078	0.310 0.326 0.328 0.322			

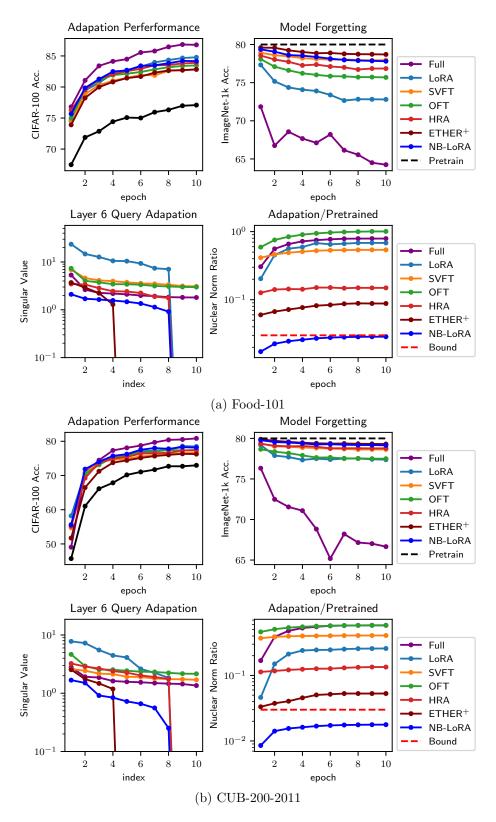


Figure 6: Performance of model adaptation and forgetting on Food-101 and CUB-200-2011 datasets during training.