Symmetries in Weight Space Learning: To Retain or Remove?

Fynn Kiwitt[*] FYNN.KIWITT@TUM.EDU
Technical University of Munich

Behrooz Tahmasebi[*] BZT@MIT.EDU
MIT EECS & MIT CSAIL

Stefanie Jegelka STEFJE@MIT.EDU
Technical University of Munich (CIT, MCML, MDSI)
MIT EECS & MIT CSAIL

Abstract

Weight space learning, an emerging paradigm that seeks to understand neural networks through their space of parameters (weights), has shown promise in a variety of applications, including but not limited to predicting model behavior and addressing privacy concerns. However, weight spaces often exhibit inherent symmetries that impact both theory and practice, such as the scale and rotational invariances found in the Low-Rank Adaptation (LoRA) method, which is the state-of-the-art fine-tuning algorithm for Large Language Models (LLMs). In this work, we investigate a general weight space learning problem under symmetries, focusing on a fundamental question: What is the appropriate formulation for this problem in the presence of symmetries (such as those in LoRA), and should redundant representations that encode the same end-to-end function be removed? We address this question by fully characterizing a new space of symmetric weights, demonstrating that the relevance of redundancy depends on the function being predicted. Specifically, we show that end-to-end symmetries (such as those in LoRA) should not always be removed, as doing so may compromise the universality of the weight space learning problem. To our knowledge, this is the first time this phenomenon has been formally identified and presented, yielding insights into a broad class of weight space learning problems.

1. Introduction

Low-Rank Adaptation (LoRA) is a state-of-the-art fine-tuning method for Large Language Models (LLMs). It aims to reduce the computational cost of full-parameter fine-tuning by learning low-rank updates to the model's weights. The primary objective is to efficiently adapt a pretrained model to new data while ensuring the updates remain meaningful with respect to the fine-tuning dataset.

The LoRA weight space encodes partial information about the fine-tuning data—as expected, since the method is explicitly designed to learn from it. However, this property raises important privacy concerns, as fine-tuning datasets often contain sensitive information. Beyond privacy, various characteristics of the fine-tuned model—such as its sensitivity to weight perturbations, its generalization ability, and its behavior on specific data subsets—are correlated with, and can potentially be inferred from, the information embedded in the weight space.

[*] Equal contribution.

© F. Kiwitt, B. Tahmasebi & S. Jegelka.

Weight space learning refers to the task of using a model's parameters to predict properties that are implicitly encoded in its weight space. This problem has practical applications in areas such as privacy leakage, sensitivity analysis, generalization prediction, and model behavior forecasting. With the rise of LLMs and the abundance of publicly available fine-tuned models, weight space learning has recently garnered significant interest within the deep learning community.

Importantly, many real-world models exhibit symmetries in their weight spaces—that is, different sets of parameters can produce the same end-to-end function. For example, in LoRA, the low-rank factors can be scaled or rotated without altering the resulting function. Similar symmetries appear in other neural architectures, such as neuron permutation invariance in feedforward networks and scaling invariance in ReLU-based models.

In this paper, we study the problem of weight space learning, with a particular focus on understanding the role of symmetries in this setting. A natural question arises: should we remove these symmetries and use invariant neural networks to process the weight space, or should we retain the original weights in their raw form? As a first step toward addressing this important question, we propose a general formulation of weight space learning and demonstrate a perhaps surprising result:

> Removing symmetries can, in some cases, compromise the expressive power (or universality) of the weight space learning problem, even when the model itself exhibits symmetries.

Specifically, we show that the symmetries relevant to weight space learning are a subset of the symmetries of the underlying model. In some cases, the weight space learning problem exhibits no symmetries at all, even when the original model is symmetric. Consequently, the decision to remove or preserve symmetries in weight space learning depends on the structure of the downstream task, as there is no universal rule that guarantees expressivity preservation across all settings.

This result lays a conceptual foundation for handling symmetries in weight space learning. In this extended abstract, we highlight the core ideas and techniques that lead to our main findings.

2. Related Work

Low-Rank Adaptation (LoRA) was introduced to accelerate the fine-tuning of Large Language Models (LLMs) by representing weight updates as low-rank matrices, thereby significantly reducing the number of trainable parameters [9]. Since its introduction, many extensions have been proposed. These include LoRA concatenation for skill composition [14], QLoRA for quantized fine-tuning [2], alternative initialization strategies such as PiSSA [12], and theoretical investigations into LoRA's expressive power [23] (see also [1]).

To bridge the performance gap between LoRA and full fine-tuning, various techniques have been developed. These include introducing fixed learning rate adjustments to account for the different magnitudes of LoRA factors [7], reformulating the gradient update with a low-rank structure [19], and applying scale-invariant optimization strategies [11, 21].

Another active area of research is learning to optimize, a meta-learning approach that designs optimizers to generalize across tasks [4, 5, 13, 17, 18]. Recent work has investigated how symmetry in parameter space can improve generalization in this setting [20, 22, 24, 25].

Learning on LoRA (LoL) [15] is a recent framework that uses trained LoRA weights as inputs to a meta-network for downstream prediction tasks. This meta-network can infer dataset properties (e.g.,

size) or fine-tuned model characteristics (e.g., accuracy). In such setups, handling the symmetries of LoRA weights is crucial for robust generalization.

From an applications perspective, Salama et al. [16] demonstrated how LoRA weights can be used to estimate the size of the training dataset. More ambitiously, Haim et al. [6] showed that image-level training data could be recovered from the weights of a fully connected network. In contrast, Elbaz et al. [3] found that for group-invariant networks, such reconstructions often converge to different but functionally equivalent samples, and they proposed a workaround using task-specific priors. Finally, Horwitz et al. [8] offer a mixture-of-experts approach that organizes fine-tuned models into a hierarchical structure based on their foundation model lineage, enabling weight-space-based reasoning. A related direction is probing in weight space, where Kahana et al. [10] proposed learning structured probes by factorizing them via latent codes, offering a principled way to extract interpretable signals from weights.

3. Problem Statement and Main Results

Consider a model $f(x; \mathbf{w})$, where $x \in \mathcal{X}$ denotes the input and $\mathbf{w} \in \mathcal{W}$ represents the learnable parameters (weights). Both $\mathcal{X}$ and $\mathcal{W}$ are assumed to be complete metric spaces. The associated function space is defined as

$$\mathcal{F} := \{f(\cdot; \mathbf{w}) \mid \mathbf{w} \in \mathcal{W}\}.$$

The goal of the *weight space learning* problem is the following: given a dataset of function-label pairs $(f(\cdot; \mathbf{w}_i), y_i) \in \mathcal{F} \times \mathbb{R}$ for $i \in [n]$, the task is to learn a meta-regression function $\hat{f}_{\text{meta}} : \mathcal{W} \to \mathbb{R}$ that not only predicts y_i accurately on observed weights $\mathbf{w}_i$, but also generalizes well to unseen weights $\mathbf{w} \in \mathcal{W}$.

In practice, different parameter values can correspond to the same function. That is, the mapping $\mathbf{w} \mapsto f(\cdot; \mathbf{w}) \in \mathcal{F}$ is generally not injective. A notable example arises in the LoRA formulation for fine-tuning neural networks. Let $\mathbf{W} \in \mathbb{R}^{d \times d}$ denote the model weights. In LoRA, the weights are parameterized as

$$\mathbf{W} = \mathbf{W}_0 + A^\top B, \quad \text{with } A, B \in \mathbb{R}^{r \times d},$$

where $\mathbf{W}_0 \in \mathbb{R}^{d \times d}$ denotes frozen pre-trained weights, and A, B are low-rank matrices learned during fine-tuning, with $r \ll d$.

In this formulation, any invertible matrix $C \in \mathrm{GL}_r(\mathbb{R})$ induces an equivalence relation, since

$$\mathbf{W} = \mathbf{W}_0 + (C^{-1}A)^\top C^\top B = \mathbf{W}_0 + A^\top B.$$

Thus, the weight-space representations (A, B) and $(C^{-1}A, C^\top B)$ define the same function.

Weight space symmetries are not unique to LoRA. Other notable examples include permutation symmetries among neurons in feedforward networks, and scaling symmetries in ReLU networks, where appropriate rescaling of adjacent layers can leave the output function unchanged.

In this paper, we formalize a general framework to study weight space learning under such symmetries. Specifically, we consider a Lie group G that acts smoothly on the weight space $\mathcal{W}$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is a full-dimensional submanifold (possibly with boundary). We assume that G fully captures the symmetry structure of the model in the sense that

$$f(\cdot, \mathbf{w}) \equiv f(\cdot, g\mathbf{w}) \quad \forall g \in G,\ \mathbf{w} \in \mathcal{W},$$

where $g\mathbf{w}$ denotes the group action of $g \in G$ on the parameter $\mathbf{w} \in \mathcal{W}$.

3.1. Zeroth-Order Weight Space Learning

Consider a weight space learning problem with data $(f(\cdot;\mathbf{w}_i), y_i) \in \mathcal{F} \times \mathbb{R}$ for $i \in [n]$, where the goal is to learn a meta-regression function of the form:

$$f_{\text{meta}}(\mathbf{w};\phi,\psi) \coloneqq \phi\left(\int_{\mathcal{X}} \psi\left(f(x;\mathbf{w})\right) dx\right), \tag{1}$$

where ϕ and ψ are parametrized functions (to be learned from the dataset), and $\mathcal{X}$ is a measurable subset of $\mathbb{R}^d$. In practice, this corresponds to learning zeroth-order features from the model $f(x;\mathbf{w})$, using samples from the input domain $\mathcal{X}$ to approximate the integral and train ϕ and ψ.

In this setting, we can directly observe that:

$$f_{\text{meta}}(g\mathbf{w};\phi,\psi) = \phi\left(\int_{\mathcal{X}} \psi\left(f(x;g\mathbf{w})\right) dx\right) = \phi\left(\int_{\mathcal{X}} \psi\left(f(x;\mathbf{w})\right) dx\right), \tag{2}$$

since $f(x;g\mathbf{w}) = f(x;\mathbf{w})$ for all $g \in G$. This implies that zeroth-order features in weight space learning are invariant under weight space symmetries. Consequently, standard methods for learning under symmetries can be effectively applied in such cases.

In particular, we obtain the following result for LoRA weights:

Corollary 1 *Weight space learning with LoRA is* $\mathrm{GL}_r(\mathbb{R})$*-invariant for zeroth-order meta-regression functions (Equation* (2)*). This symmetry matches that of the original LoRA formulation.*

3.2. Higher-Order Weight Space Learning

In this section, we consider meta-regression functions that require features beyond zeroth order. As a concrete example, consider weight space learning on the zero-loss manifold of a pre-trained neural network, with training data $(x_j, z_j) \in \mathcal{X} \times \mathbb{R}$, $j \in [J]$. Under the square loss, we have

$$L(\mathbf{w}) \coloneqq \frac{1}{2J}\sum_{j=1}^{J}(f(x_j;\mathbf{w}) - z_j)^2 \quad \Longrightarrow \quad \nabla^2_{\mathbf{w}} L(\mathbf{w}) = \frac{1}{J}\sum_{j=1}^{J} \nabla_{\mathbf{w}} f(x_j;\mathbf{w}) \left(\nabla_{\mathbf{w}} f(x_j;\mathbf{w})\right)^\top.$$

In particular, consider the function

$$f_{\text{sens}}(\mathbf{w}) \coloneqq \operatorname{tr}\left(\nabla^2_{\mathbf{w}} L(\mathbf{w})\right) = \frac{1}{J}\sum_{j=1}^{J} \|\nabla_{\mathbf{w}} f(x_j;\mathbf{w})\|_2^2, \tag{3}$$

which captures the average sensitivity of the model $f(\cdot;\mathbf{w})$ with respect to its parameters.

Can a zeroth-order meta-regression function learn such sensitivity features? Let us examine the symmetries of the sensitivity function in the LoRA weight space, where $\mathbf{w} = (A, B)$. In this setting, for any $g \in G$ corresponding to $C \in \mathrm{GL}_r(\mathbb{R})$, the two configurations $(C^{-1}A, C^\top B)$ and (A, B) are functionally equivalent. However, the sensitivity transforms as

$$f_{\text{sens}}(g\mathbf{w}) = \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{d}\left\{\|C^\top \nabla_{\mathbf{A}_k} f(x_j;\mathbf{w})\|_2^2 + \|C^{-1}\nabla_{\mathbf{B}_k} f(x_j;\mathbf{w})\|_2^2\right\}, \tag{4}$$

where $\mathbf{A}_k$ denotes the k-th column of matrix A (similarly for B).

This expression is equal to $f_{\text{sens}}(\mathbf{w})$ if and only if $C \in \mathrm{O}(r)$. In other words, while all invertible matrices $C \in \mathrm{GL}_r(\mathbb{R})$ correspond to weight space symmetries of LoRA, only orthogonal matrices preserve the sensitivity function. Thus, learning sensitivity-based features requires going beyond zeroth-order meta-regressors.

Corollary 2 *Weight-space learning with LoRA is only $\mathrm{O}(r)$-invariant when learning sensitivity-dependent features (Equation* (3)*). This invariance group is a strict subset of the full symmetry group of the original LoRA formulation, which is $\mathrm{GL}_r(\mathbb{R})$.*

Remark 3 *The above illustrates that the symmetries relevant for weight-space learning can be a strict subset of the symmetries of the underlying model. Furthermore, compressing the weight space based solely on the model's full symmetry group can compromise universality (expressive power), as certain weight-space learning tasks (such as those involving sensitivity) require reduced symmetry.*

One can show that the explanation presented here extends beyond the square loss and holds for a broader class of functionals beyond the trace of the Hessian. We will investigate this class of functionals and their associated symmetries in future work.

3.3. Weight-Space Learning with No Symmetry

Consider the following meta-regressor:

$$f_{\text{meta}}(\mathbf{w}) \coloneqq \partial_1[L(\mathbf{w})], \tag{5}$$

where $L(\mathbf{w})$ is as previously defined, and $\partial_1[\cdot]$ denotes the partial derivative with respect to the first coordinate of the vector $\mathbf{w} \in \mathbb{R}^d$. In this case, what symmetries, if any, does f_{meta} exhibit?

Assume that $D(g) \in \mathrm{GL}_d(\mathbb{R})$ denotes the matrix representation of a group element $g \in G$, such that the group action is given by $g\mathbf{w} \coloneqq D(g)\mathbf{w}$. Then:

$$f_{\text{meta}}(g\mathbf{w}) = f_{\text{meta}}(\mathbf{w}) \quad \text{if and only if} \quad D(g) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix}, \tag{6}$$

and analogously for other indices $i = 2, 3, \ldots, d$.

Therefore, even when using first-order derivatives of the loss function as features, the only allowable transformation is $D(g) = I_d$, implying that no nontrivial symmetries are preserved—despite the underlying model potentially being symmetric. This means that:

> The general problem of weight-space learning (beyond the zeroth-order case) requires considering the full weight space under its inherent symmetries. In other words, any compression of the weight space that eliminates these symmetries compromises the universality (i.e., expressive power) of the weight-space learning framework.

However, the above result holds only in the context of the general formulation. For restricted weight-space learning meta-regressors (such as the zeroth-order function class), it may be possible to

remove certain symmetries from the model's weight space while still maintaining full expressivity. Identifying an appropriate space for weight-space learning thus heavily depends on the downstream task, and no universal solution exists. We defer a detailed investigation of this problem to future work.

Acknowledgements

This project was supported by the Office of Naval Research under Grant N00014-20-1-2023 (MURI ML-SCOPE), the National Science Foundation under Award CCF-2112665 (TILOS AI Institute), and an Alexander von Humboldt Professorship.

References

[1] Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. Lora-xs: Low-rank adaptation with extremely small number of parameters. *arXiv preprint arXiv:2405.17604*, 2024.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Ran Elbaz, Gilad Yehudai, Meirav Galun, and Haggai Maron. On the reconstruction of training data from group invariant networks. *arXiv preprint arXiv:2411.16458*, 2024.

[4] Sebastian Flennerhag, Tom Zahavy, Brendan O'Donoghue, Hado P van Hasselt, András György, and Satinder Singh. Optimistic meta-gradients. *Advances in Neural Information Processing Systems*, 36:57852–57862, 2023.

[5] Boyan Gao, Henry Gouk, Hae Beom Lee, and Timothy M Hospedales. Meta mirror descent: Optimiser learning for fast convergence. *arXiv preprint arXiv:2203.02711*, 2022.

[6] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35: 22911–22924, 2022.

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

[8] Eliahu Horwitz, Bar Cavia, Jonathan Kahana, and Yedid Hoshen. Representing model weights with language using tree experts. *arXiv preprint arXiv:2410.13569*, 2024.

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

[10] Jonathan Kahana, Eliahu Horwitz, Imri Shuval, and Yedid Hoshen. Deep linear probe generators for weight space learning. *arXiv preprint arXiv:2410.10811*, 2024.

[11] Bingcong Li, Liang Zhang, and Niao He. Implicit regularization of sharpness-aware minimization for scale-invariant problems. *arXiv preprint arXiv:2410.14802*, 2024.

[12] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*, 2024.

[13] Luke Metz, C Daniel Freeman, Niru Maheswaranathan, and Jascha Sohl-Dickstein. Training learned optimizers with randomly initialized learned optimizers. *arXiv preprint arXiv:2101.07367*, 2021.

[14] Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. Lora soups: Merging loras for practical skill composition tasks. *arXiv preprint arXiv:2410.13025*, 2024.

[15] Theo Putterman, Derek Lim, Yoav Gelberg, Stefanie Jegelka, and Haggai Maron. Learning on loras: Gl-equivariant processing of low-rank weight spaces for large finetuned models. *arXiv preprint arXiv:2410.04207*, 2024.

[16] Mohammad Salama, Jonathan Kahana, Eliahu Horwitz, and Yedid Hoshen. Dataset size recovery from lora weights. *arXiv preprint arXiv:2406.19395*, 2024.

[17] Jiayi Shen, Xiaohan Chen, Howard Heaton, Tianlong Chen, Jialin Liu, Wotao Yin, and Zhangyang Wang. Learning a minimax optimizer: A pilot study. In *International Conference on Learning Representations*, 2020.

[18] Benjamin Thérien, Charles-Étienne Joseph, Boris Knyazev, Edouard Oyallon, Irina Rish, and Eugene Belilovsky. μ lo: Compute-efficient meta-generalization of learned optimizers. *arXiv preprint arXiv:2406.00153*, 2024.

[19] Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *arXiv preprint arXiv:2407.05000*, 2024.

[20] Junjie Yang, Tianlong Chen, Mingkang Zhu, Fengxiang He, Dacheng Tao, Yingbin Liang, and Zhangyang Wang. Learning to generalize provably in learning to optimize. In *International conference on artificial intelligence and statistics*, pages 9807–9825. PMLR, 2023.

[21] Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. Lora done rite: Robust invariant transformation equilibration for lora optimization. *arXiv preprint arXiv:2410.20625*, 2024.

[22] Guy Zamir, Aryan Dokania, Bo Zhao, and Rose Yu. Improving learning to optimize using parameter symmetries. *arXiv preprint arXiv:2504.15399*, 2025.

[23] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

[24] Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. *Advances in neural information processing systems*, 35:16679–16690, 2022.

[25] Bo Zhao, Robert M Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. *arXiv preprint arXiv:2305.13404*, 2023.