PARAMETER SYMMETRIES ARE NOT TASK SYMMETRIES: WHICH INVARIANCES CAN WEIGHT-SPACE LEARNING AFFORD?

Anonymous authors
Paper under double-blind review

ABSTRACT

Weight-space learning predicts properties of neural networks from their parameters, and predictors are commonly made invariant to all parameter symmetries, i.e., to every transformation that preserves the network function. We show that the right invariance depends on the target. Each target has its own task-symmetry group, and quotienting by anything larger incurs an irreducible error, which we compute exactly for compact groups. We then derive task-symmetry groups for broad target classes. In LoRA, a single parameterization exhibits a strict ladder: function-level targets are GL_r-invariant, Euclidean sensitivity is only $\mathrm{O}(r)$-invariant, gradient sparsity is only invariant under signed permutations, and distance from initialization has no symmetry. The two middle levels are selected by the optimizer, since gradient descent and Adam measure progress by the ℓ_2 and ℓ_1 gradient norms, respectively. Finally, neuron-permutation symmetry, which most weight-space architectures quotient, is lost by any target that refers to a reference model, such as drift from a pretrained base. Experiments on synthetic zoos and LoRA-fine-tuned language models confirm the predicted separations.

1 INTRODUCTION

Weight-space learning uses the parameters of trained neural networks as data: given weights $\mathbf{w}$, a meta-model predicts a property of the network, its training data, or its training process (Han et al., 2026). Applications include predicting accuracy and generalization from weights (Unterthiner et al., 2020; Eilertsen et al., 2020; Schürholt et al., 2021), recovering hyperparameters, training-set size, and model lineage (Eilertsen et al., 2020; Salama et al., 2024; Horwitz et al., 2024), and auditing what weights reveal about their training data (Haim et al., 2022). Large model zoos (Schürholt et al., 2022b) and the growing supply of fine-tuned and parameter-efficient checkpoints (Hu et al., 2021) make such problems increasingly relevant.

The basic difficulty is that parameters are not unique. A group G of reparameterizations acts on the weight space $\mathcal{W}$ while leaving the represented function unchanged,

$$f(\cdot; g\mathbf{w}) = f(\cdot; \mathbf{w}), \qquad g \in G. \tag{1}$$

Examples include hidden-unit permutations in multilayer perceptrons (Brea et al., 2019; Simsek et al., 2021), positive rescalings in homogeneous ReLU networks (Neyshabur et al., 2015; Dinh et al., 2017), permutations of attention heads, and changes of basis in low-rank factors (Putterman et al., 2024); see Godfrey et al. (2022) and Ziyin et al. (2025) for broader treatments. A seemingly obvious design principle follows: identify parameters along G-orbits, and require the meta-model to be G-invariant. Much of the weight-space architecture literature is built on this principle, from permutation-equivariant networks for MLP weights (Navon et al., 2023; Zhou et al., 2023a;b; Lim et al., 2023; Kofinas et al., 2024), through networks that also respect scaling and sign symmetries (Kalogeropoulos et al., 2024; Tran et al., 2024), to GL_r-equivariant networks for LoRA factors (Putterman et al., 2024).

The principle is correct only if the prediction target is itself constant on G-orbits. Functional equivalence does not imply equivalence for every property of a *parameterized* model: two weights can compute the same function yet differ in gradients, curvature, sparsity of updates, robustness

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

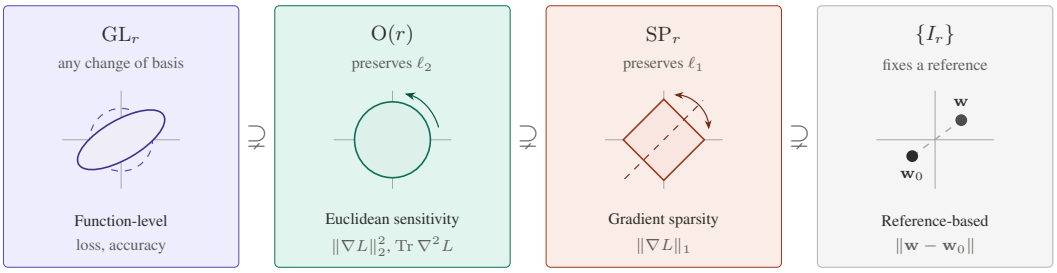


Figure 1: Task symmetries of LoRA targets. Every $C \in \mathrm{GL}_r$ maps LoRA factors (A, B) to $(CA, C^{-\top}B)$ without changing the network, yet each target keeps only the subgroup that preserves the geometry it measures: function-level targets keep all of GL_r, Euclidean sensitivity keeps rotations $\mathrm{O}(r)$, gradient sparsity keeps signed permutations SP_r, and distances to a reference keep nothing (Theorem 4.3). Quotienting by a larger group than the target's discards information that no predictor can recover (Theorem 3.3).

to parameter noise, distance from a reference model, and the trajectory an optimizer takes from them. That parameter-space geometry is not determined by the realized function is well known in the flat-minima literature (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017): Dinh et al. (2017) used the rescaling symmetry of ReLU networks to construct functionally equivalent minima of arbitrary sharpness, which motivated reparameterization-invariant sharpness measures (Kwon et al., 2021; Andriushchenko et al., 2023). This observation has not been carried over to weight-space learning, where architectures are typically designed to be invariant to the full model-symmetry group. Our goal is to make it operational there: to say, for a given target, exactly which symmetries may be removed, what it costs to remove more, and why.

Approach. We separate the *model-symmetry group* G, which preserves $f(\cdot; \mathbf{w})$, from the *task-symmetry group* $G_T \leq G$, which preserves a particular target T. The question "retain or remove?" then has a precise answer. Quotienting by $H \leq G_T$ is lossless; quotienting by any $H \not\leq G_T$ incurs an error that no amount of data or capacity can remove, and which we compute exactly. The remaining work is to compute G_T for the targets people actually care about, which we do for derivative-based, trajectory-based, and reference-based targets.

In short, in this paper we make the following contributions:

- **The exact price of the wrong invariance (Section 3).** For a compact subgroup H, the minimal squared error of any H-invariant predictor equals the expected within-orbit variance $B_H(T) = \mathbb{E}_{\mathbf{w}} \mathrm{Var}_{h\sim H}\, T(h\mathbf{w})$, it is monotone along the subgroup lattice, and it vanishes if and only if $H \leq G_T$. Its normalized form gives an R^2 ceiling for all invariant predictors. Paired with known sample-complexity gains of invariance, this turns "it depends" into a rule: quotient exactly G_T.
- **A criterion for derivative targets and the LoRA ladder (Section 4).** For targets of the form $\int \Phi(\nabla_{\mathbf{w}}^{\otimes k} f)\, \mathrm{d}\mu$, the task-symmetry group is the set of model symmetries whose contragredient action preserves Φ on the attainable derivatives. In LoRA we compute it exactly and obtain the strict ladder $\mathrm{GL}_r \supsetneq \mathrm{O}(r) \supsetneq \mathrm{SP}_r \supsetneq \{I_r\}$ for function-level, Euclidean sensitivity, gradient-sparsity, and distance-from-initialization targets.
- **The optimizer selects the symmetry (Section 5).** Gradient descent is equivariant only under isometries of its metric, and sign-based methods, including the first step of Adam, only under signed permutations. Their first-order progress is the squared ℓ_2 and the ℓ_1 gradient norm respectively, which ties the ladder to training dynamics and explains why gradient sparsity is a natural target.
- **Permutations and reference-based targets (Section 6).** Every target in the ladder above except the last is permutation-invariant, so permutation-invariant architectures are safe for them. What breaks permutation symmetry is an external reference frame: the task-symmetry group of drift from a base model is the stabilizer of the base, which is trivial exactly when its neurons are pairwise distinct. We also show that permutation-invariant linear order-k targets form a $\mathrm{Bell}(k)/h^k$ fraction of all such targets.

- **Experiments (Section 7).** Meta-models that enforce GL_r, $\mathrm{O}(r)$, SP_r or S_h invariance match unconstrained ones exactly on targets their group preserves and stay below the predicted ceilings on the others. Optimizers are equivariant exactly under the predicted groups, and on LoRA-fine-tuned language models the Hessian trace and perturbation robustness separate as predicted.

2 Model Symmetry and Task Symmetry

Consider a model $f(x; \mathbf{w})$ with input $x \in \mathcal{X}$ and parameters $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$. In weight-space learning the examples are parameter–label pairs $\{(\mathbf{w}_i, y_i)\}_{i=1}^n$ with $y_i \approx T(\mathbf{w}_i)$ for a target $T : \mathcal{W} \to \mathbb{R}$, and the learner fits a meta-predictor $h : \mathcal{W} \to \mathbb{R}$. This formulation permits both *function-level targets*, which depend on $\mathbf{w}$ only through $f(\cdot; \mathbf{w})$, and *parameterization-level targets*, which also depend on how the function is parameterized. Treating the data as function–label pairs would exclude the latter by definition.

Let a group G act linearly on $\mathcal{W}$ via a representation $\rho : G \to \mathrm{GL}_d(\mathbb{R})$, $g\mathbf{w} = \rho(g)\mathbf{w}$. We call G a *model-symmetry group* if $f(\cdot; g\mathbf{w}) = f(\cdot; \mathbf{w})$ for all $g \in G$ and $\mathbf{w} \in \mathcal{W}$. Linearity covers all examples in this paper; the results of Section 3 hold for arbitrary measurable actions.

Running example: LoRA. Let $\mathbf{W}_0 \in \mathbb{R}^{p\times p}$ be frozen pretrained weights and $\mathbf{W} = \mathbf{W}_0 + A^\top B$ with $A, B \in \mathbb{R}^{r\times p}$ and $r < p$. Every $C \in \mathrm{GL}_r(\mathbb{R})$ acts by

$$C \cdot (A, B) = (CA,\ C^{-\top}B), \qquad (CA)^\top (C^{-\top}B) = A^\top B, \tag{2}$$

so $\mathrm{GL}_r(\mathbb{R})$ is a model-symmetry group of every LoRA-parameterized model. Its subgroups $\mathrm{O}(r)$ (orthogonal matrices) and SP_r (signed permutation matrices, of order $2^r r!$) will play a central role. We write $\{I_r\}$ for the trivial group.

Definition 2.1 (Task-symmetry group)**.** For a target $T : \mathcal{W} \to \mathbb{R}$, the task-symmetry group is

$$G_T \coloneqq \{g \in G :\ T(g\mathbf{w}) = T(\mathbf{w}) \text{ for every } \mathbf{w} \in \mathcal{W}\}. \tag{3}$$

Targets usually come in families: "the Hessian trace" is a different function of $\mathbf{w}$ for every architecture, loss, and dataset. We therefore also consider a class $\mathcal{M}$ of admissible models, all with model-symmetry group G, and a family $\mathcal{T} = \{T_{f,\mu}\}$ indexed by models $f \in \mathcal{M}$ and data distributions μ on $\mathcal{X}$. The *universal task-symmetry group* of the family is $G_{\mathcal{T}} \coloneqq \bigcap_{f,\mu} G_{T_{f,\mu}}$: the transformations that may be quotiented no matter which model zoo the data come from. For LoRA we take $\mathcal{M}$ to be all models of the form $f(x; A, B) = \phi(x, \mathbf{W}_0 + A^\top B)$ with ϕ smooth, i.e., all models that use the adapted weight matrix in any differentiable way.

3 The Price of the Wrong Invariance

Let $\pi : \mathcal{W} \to \mathcal{W}/H$ denote the orbit map of a subgroup $H \leq G$. A predictor that operates on the quotient, or equivalently an H-invariant predictor, cannot distinguish points on the same orbit.

Theorem 3.1 (Lossless quotient criterion)**.** *For a target $T : \mathcal{W} \to \mathbb{R}$ and a subgroup $H \leq G$, the following are equivalent: (i) T is H-invariant; (ii) $H \leq G_T$; (iii) there is a unique $\overline{T} : \mathcal{W}/H \to \mathbb{R}$ with $T = \overline{T} \circ \pi$. If $\mathcal{W}/H$ carries the quotient topology, a continuous invariant T induces a continuous $\overline{T}$. Consequently G_T is the largest subgroup of G that can be quotiented without loss.*

Theorem 3.2 (Pointwise lower bound)**.** *Let $\mathcal{H}_H$ be any class of H-invariant predictors. For every target T,*

$$\inf_{h\in\mathcal{H}_H} \sup_{\mathbf{w}\in\mathcal{W}} |T(\mathbf{w}) - h(\mathbf{w})| \ \geq\ \frac{1}{2} \sup_{\mathbf{w}\in\mathcal{W},\ g\in H} |T(g\mathbf{w}) - T(\mathbf{w})|. \tag{4}$$

Both statements are elementary (proofs in Appendix D); their role is to fix the logic. The next result is quantitative and is what makes the choice of H a measurable trade-off rather than a binary one.

Theorem 3.3 (Exact cost of quotienting)**.** *Let $H \leq G$ be compact with normalized Haar measure ν_H, let μ be an H-invariant probability measure on $\mathcal{W}$, and let $T \in L^2(\mu)$ be non-constant. Define*

the orbit bias *and its normalized version*

$$B_H(T) \coloneqq \mathbb{E}_{\mathbf{w}\sim\mu}\Big[\operatorname{Var}_{h\sim\nu_H}\, T(h\mathbf{w})\Big], \qquad \rho_H(T) \coloneqq \frac{B_H(T)}{\operatorname{Var}_\mu(T)} \in [0,1]. \tag{5}$$

Then:

1. $\min_{h\in L^2(\mu),\, h \text{ } H\text{-invariant}} \mathbb{E}_\mu\big[(T-h)^2\big] = B_H(T)$, *attained by the orbit average* $\bar{T}_H(\mathbf{w}) = \int_H T(h\mathbf{w})\,\mathrm{d}\nu_H(h)$.
2. *(Monotonicity) If* $H_1 \le H_2$ *are compact, then* $B_{H_1}(T) \le B_{H_2}(T)$.
3. $B_H(T) = 0$ *if and only if* $T(h\mathbf{w}) = T(\mathbf{w})$ *for* $\nu_H \otimes \mu$*-almost every* $(h,\mathbf{w})$. *If* T *is continuous and* μ *has full support, this holds if and only if* $H \le G_T$.
4. *(*R^2 *ceiling) Every* H*-invariant predictor* h *satisfies* $R^2(h) \coloneqq 1 - \mathbb{E}_\mu(T-h)^2/\operatorname{Var}_\mu(T) \le 1 - \rho_H(T)$.

Item 4 is the form we use empirically: $1-\rho_H(T)$ is a horizontal line that no H-invariant architecture, however large and however well trained, can cross. It is estimable on a model zoo by sampling group elements, and unlike the algebraic criterion it degrades gracefully: a target can be *nearly* invariant (ρ_H small), which is exactly the regime where practical decisions are made. For noncompact groups such as GL_r there is no invariant probability measure, B_H depends on a chosen prior over the group, and we state results algebraically through G_T; Theorem 3.2 remains valid.

The other side of the ledger. Invariance also has a benefit. For kernel regression with H-invariant kernels, the excess risk behaves as if the number of samples were multiplied by $|H|$ for finite groups, or the input dimension reduced by the orbit dimension for positive-dimensional groups (Tahmasebi & Jegelka, 2023). For $H \le G_T$ this gain comes for free. For $H \not\le G_T$ the gain is paid for with the bias $B_H(T) > 0$, which does not decrease with the sample size. Combining both sides:

> **Prescription.** Quotient by the task-symmetry group G_T. It is the largest lossless choice (Theorem 3.1) and collects the full sample-complexity gain; every larger compact group pays an irreducible error $B_H(T)$ that bounds the achievable R^2 by $1-\rho_H(T)$ (Theorem 3.3). With finite data, a larger group is justified only when $\rho_H(T)$ is small, and $\rho_H(T)$ can be measured.

The rest of the paper computes G_T for the main families of weight-space targets.

4 Task Symmetries of Derivative Targets

4.1 A General Criterion

Function-level targets, such as test loss, accuracy, or any functional $\phi(\int \psi(f(x;\mathbf{w}))\,\mathrm{d}\mu(x))$ of the realized function, are constant on model-symmetry orbits, so their task-symmetry group is all of G and quotienting is lossless for them. Parameterization-level targets arise as soon as derivatives with respect to $\mathbf{w}$ enter. For a linear action, differentiating $f(x;\rho(g)\mathbf{w}) = f(x;\mathbf{w})$ gives the contragredient transformation rule

$$\nabla_{\mathbf{w}}^{\otimes k} f(x; g\mathbf{w}) = \rho^\star(g)^{\otimes k}\, \nabla_{\mathbf{w}}^{\otimes k} f(x;\mathbf{w}), \qquad \rho^\star(g) \coloneqq \rho(g)^{-\top}. \tag{6}$$

Gradients therefore do not move with the weights: they move with the inverse transpose. Whether a derivative target survives is decided by how the functional applied to the derivatives interacts with $\rho^\star$.

Definition 4.1 (Derivative targets)**.** Fix $k \ge 1$ and a function $\Phi : \mathrm{Sym}^k(\mathbb{R}^d) \to \mathbb{R}$. The order-$k$ Φ-family is $\mathcal{T}_\Phi = \{T_{f,\mu}\}$ with

$$T_{f,\mu}(\mathbf{w}) \coloneqq \int_{\mathcal{X}} \Phi\big(\nabla_{\mathbf{w}}^{\otimes k} f(x;\mathbf{w})\big)\,\mathrm{d}\mu(x), \qquad f \in \mathcal{M}. \tag{7}$$

The *attainable set* is $\mathcal{V}_k \coloneqq \{\nabla_{\mathbf{w}}^{\otimes k} f(x;\mathbf{w}) : f \in \mathcal{M},\ x \in \mathcal{X},\ \mathbf{w} \in \mathcal{W}\}$.

The family includes Euclidean sensitivity ($k = 1$, $\Phi = \|\cdot\|_2^2$), whose average over data is the Hessian trace of the squared loss at interpolation; gradient sparsity ($k = 1$, $\Phi = \|\cdot\|_1$); linear order-k features $\Phi = \langle a, \cdot \rangle$; and, taking f to be a per-example loss, the corresponding quantities for the loss.

Proposition 4.2 (Reduction to the attainable set)**.** *The universal task-symmetry group of $\mathcal{T}_\Phi$ is*

$$G_{\mathcal{T}_\Phi} = \{g \in G : \ \Phi(\rho^\star(g)^{\otimes k} v) = \Phi(v) \ \ \forall v \in \mathcal{V}_k\}. \tag{8}$$

In particular, (a) if $\Phi = \langle a, \cdot \rangle$ is linear and $V_k = \operatorname{span} \mathcal{V}_k$, then $g \in G_{\mathcal{T}_\Phi}$ if and only if $\rho(g)^{\otimes k} a - a \in V_k^\perp$; when $V_k = \mathrm{Sym}^k$ this is the tensor-stabilizer condition $\rho(g)^{\otimes k} a = a$. (b) If Φ is a norm, $G_{\mathcal{T}_\Phi}$ consists of the g for which $\rho^\star(g)$ is an isometry of Φ on $\mathcal{V}_k$.

The proposition is a bookkeeping device: it moves all the difficulty into two concrete questions, namely what the isometries of Φ are and what the attainable set looks like for a given architecture. The attainable set matters. Because the model is G-invariant, its derivatives annihilate orbit directions, so $\mathcal{V}_k$ is typically a proper subset, and stabilizer conditions stated on all of Sym^k would be too strong. For LoRA we compute everything explicitly.

4.2 The LoRA Ladder

Theorem 4.3 (Four task-symmetry groups in one parameterization)**.** *Consider LoRA with $r < p$, the model class $\mathcal{M}$ of Section 2, and the action Equation (2). Let $\|\cdot\|_q$ denote the entrywise ℓ_q norm on $\mathbb{R}^{r\times p} \times \mathbb{R}^{r\times p}$.*

1. Function-level targets *have task-symmetry group* $\mathrm{GL}_r(\mathbb{R})$.
2. Euclidean sensitivity $T(\mathbf{w}) = \int \|\nabla_{\mathbf{w}} f(x; \mathbf{w})\|_2^2 \, \mathrm{d}\mu(x)$ *has universal task-symmetry group* $\mathrm{O}(r)$. *The same holds for the Hessian trace* $\operatorname{Tr} \nabla_{\mathbf{w}}^2 L$ *of the loss, and for isotropic Gaussian perturbation robustness* $T_\sigma(\mathbf{w}) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}[L(\mathbf{w} + \epsilon)] - L(\mathbf{w})$ *required for all σ in an interval.*
3. ℓ_q gradient targets $T(\mathbf{w}) = \int \|\nabla_{\mathbf{w}} f(x; \mathbf{w})\|_q \, \mathrm{d}\mu(x)$ *with $q \in [1, \infty] \setminus \{2\}$, in particular gradient sparsity ($q = 1$), have universal task-symmetry group* SP_r.
4. Distance from initialization $T(\mathbf{w}) = \|A - A_0\|_F^2 + \|B - B_0\|_F^2$ *with* $\operatorname{rank} A_0 = r$ *(e.g., the standard initialization with Gaussian A_0 and $B_0 = 0$) has task-symmetry group $\{I_r\}$.*

For $r \geq 2$ all inclusions in $\mathrm{GL}_r(\mathbb{R}) \supsetneq \mathrm{O}(r) \supsetneq \mathrm{SP}_r \supsetneq \{I_r\}$ are strict.

The proof (Appendix F) rests on an explicit description of attainable LoRA gradients. Writing $\Gamma = \partial f / \partial \mathbf{W}$, the gradients are $\nabla_A f = B\Gamma^\top$ and $\nabla_B f = A\Gamma$, and under Equation (2) they transform as $(C^{-\top} \nabla_A f, \ C \nabla_B f)$. Choosing Γ with columns in $\ker A$ produces gradients of the form $(zn^\top, 0)$ with arbitrary $z \in \mathbb{R}^r$ and a fixed nonzero n, which reduces each target to a norm on $\mathbb{R}^r$; the claims then follow from the isometry groups of ℓ_2 (orthogonal matrices) and of ℓ_q for $q \neq 2$ (signed permutations).

Three remarks place the ladder in context. First, the $\mathrm{O}(r)$ level is the LoRA instance of the observation of Dinh et al. (2017); what is new is that the surviving group is identified exactly, and that it is only one level of a ladder whose position is set by the target functional. Second, gradient sparsity loses the continuous part of the symmetry entirely but keeps the finite group SP_r of permutations and sign flips of the rank-one components $a_i^\top b_i$; this is the natural discrete symmetry of LoRA, and Section 5 shows why ℓ_1 targets are natural. Third, reference targets must be read carefully. Drift from the frozen base, $\|\mathbf{W} - \mathbf{W}_0\| = \|A^\top B\|$, is function-level and keeps the full GL_r, whereas distance of the *factors* from their initialization loses everything. What matters is whether the reference is expressed in the quotiented coordinates. The same analysis applied to curvature gives a parallel ladder (Appendix A.3): the Morse index of the Hessian keeps all of G, its spectrum (e.g., sharpness) keeps O-type symmetries, and entrywise curvature statistics keep signed permutations. Canonicalizations such as balancing are quotients and obey the same rule (Theorem A.2).

5 The Optimizer Selects the Symmetry

A standard family of weight-space targets concerns training itself: loss after further fine-tuning, convergence speed, or the hyperparameters used. Such targets depend on the trajectory an optimizer takes from $\mathbf{w}$, and optimizers are not G-equivariant in general.

Proposition 5.1 (Equivariance of optimizers)**.** *Let L be G-invariant and let $\mathbf{w}_t(\mathbf{w})$ denote the t-th iterate started from $\mathbf{w}$. (a) Preconditioned gradient descent $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta M \nabla L(\mathbf{w}_t)$ with fixed symmetric positive definite M satisfies $\mathbf{w}_t(g\mathbf{w}) = g\,\mathbf{w}_t(\mathbf{w})$ for every g with $\rho(g) M \rho(g)^\top = M$; for $M = I$ these are the g with $\rho(g)$ orthogonal. (b) Sign descent $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \operatorname{sign}(\nabla L(\mathbf{w}_t))$ (Bernstein et al., 2018) and Adam (Kingma & Ba, 2015) (with or without decoupled weight decay, moments initialized at zero) satisfy the same identity for every g with $\rho(g)$ a signed permutation matrix. Consequently, any target $T(\mathbf{w}) = \Psi(f(\cdot; \mathbf{w}_t(\mathbf{w})))$ that is function-level in the t-th iterate is invariant under these subgroups.*

The two optimizers also measure progress with different norms. To first order,

$$L\big(\mathbf{w} - \eta \nabla L(\mathbf{w})\big) = L(\mathbf{w}) - \eta \,\|\nabla L(\mathbf{w})\|_2^2 + o(\eta), \tag{9}$$

$$L\big(\mathbf{w} - \eta \operatorname{sign} \nabla L(\mathbf{w})\big) = L(\mathbf{w}) - \eta \,\|\nabla L(\mathbf{w})\|_1 + o(\eta). \tag{10}$$

The squared ℓ_2 gradient norm is the one-step progress of gradient descent; the ℓ_1 norm is the one-step progress of sign descent, which is also the first step of Adam (with $\epsilon = 0$); more generally, Adam behaves as a variance-adapted sign method (Balles & Hennig, 2018). Gradient sparsity is therefore not an arbitrary target: it is the progress measure of the optimizer used to train nearly every modern checkpoint. Combining both facts with Theorem 4.3:

Theorem 5.2 (Optimizer-dependent task symmetry in LoRA)**.** *In the setting of Theorem 4.3, consider the one-step lookahead targets $T_\eta^{\mathrm{GD}}(\mathbf{w}) = L(\mathbf{w} - \eta \nabla L(\mathbf{w}))$ and $T_\eta^{\mathrm{sign}}(\mathbf{w}) = L(\mathbf{w} - \eta \operatorname{sign} \nabla L(\mathbf{w}))$ over all admissible losses. The group of $C \in \mathrm{GL}_r$ preserving T_η for all η in an interval $(0, \eta_0)$ is $\mathrm{O}(r)$ for gradient descent and SP_r for sign descent.*

The optimizer thus chooses the symmetry of trajectory-dependent targets: zoos trained with SGD retain orthogonal symmetry for these targets, zoos trained with Adam retain only signed permutations. Practice already reflects this: LoRA+ (Hayou et al., 2024) uses different learning rates for A and B precisely because gradient descent is not GL_r-equivariant.

6 Permutation Symmetry and Reference-Based Targets

Most equivariant weight-space architectures quotient neuron permutations. Let $f(x) = V\sigma(Ux)$ with $U \in \mathbb{R}^{h \times p}$ and $V \in \mathbb{R}^{q \times h}$, so $d = h(p+q)$. A permutation $P \in S_h$ acts by $(U, V) \mapsto (PU, VP^\top)$. After grouping the coordinates of each hidden neuron (its row of U and column of V) into a block of size $m = p + q$, this is $\rho(P) = P \otimes I_m$, a permutation of coordinates.

What permutations preserve. A permutation matrix is both orthogonal and a signed permutation matrix. By Theorems 4.2, 5.1 and A.3, every target in the upper three levels of Table 2 is therefore S_h-invariant: function-level targets, Euclidean sensitivity, Hessian spectra, gradient sparsity, and SGD- and Adam-trajectory targets. Permutation-invariant architectures are safe for all of them, and one cannot break permutation symmetry by changing the norm. What breaks it is an external reference frame.

Proposition 6.1 (Reference-based targets)**.** *Let $\mathbf{w}_0 \in \mathcal{W}$ and let N be a norm with $\rho(G) \subseteq \mathrm{Isom}(N)$ (for permutations, any ℓ_q norm). The target $T(\mathbf{w}) = N(\mathbf{w} - \mathbf{w}_0)$ has task-symmetry group $G_T = \mathrm{Stab}_G(\mathbf{w}_0) = \{g \in G : \rho(g)\mathbf{w}_0 = \mathbf{w}_0\}$.*

Theorem 6.2 (Reference targets generically have no symmetry)**.** *If G is finite and ρ is faithful, then $\mathrm{Stab}_G(\mathbf{w}_0) = \{e\}$ for Lebesgue-almost every $\mathbf{w}_0$. For the MLP above, $\mathrm{Stab}_{S_h}(\mathbf{w}_0)$ is trivial if and only if the h hidden neurons of $\mathbf{w}_0$ have pairwise distinct weight blocks.*

The standard fine-tuning zoo is exactly this setting. All checkpoints share a pretrained base $\mathbf{w}_0$, and the quantities one predicts, such as how far fine-tuning moved the model or how large the update was, are displacement targets. Permuting the hidden units of a fine-tuned model leaves its function

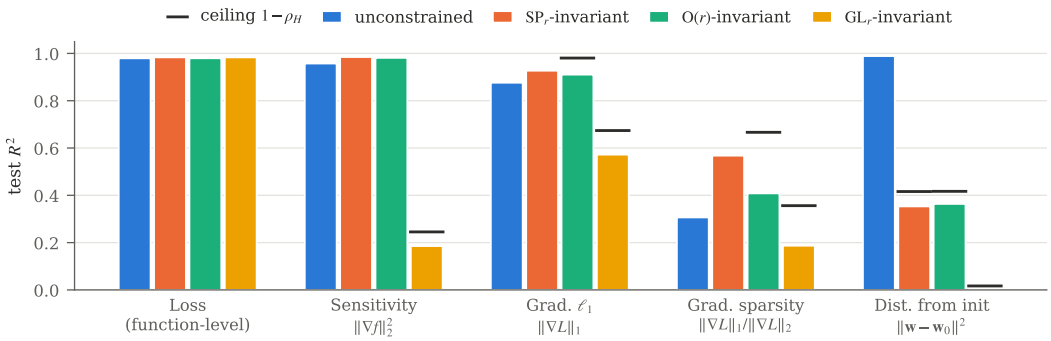


Figure 2: Meta-learning across the LoRA ladder. Test R^2 of four meta-models that differ only in the symmetry they enforce, on five targets ordered by task symmetry. Black ticks mark the ceiling $1-\rho_H(T)$ of Theorem 3.3 for each invariant model (omitted where it equals 1). Every invariant model matches the unconstrained one on targets whose task-symmetry group contains its group, and stays below its ceiling on the others.

unchanged but does not move the base, so a permutation-quotiented meta-model provably cannot represent these targets for any base model with distinct neurons. The same mechanism applies to alignment and merging targets relative to another network, which is why Git Re-Basin must solve for a permutation (Ainsworth et al., 2022), and to lineage and provenance targets (Horwitz et al., 2024). Drift from a base is a zeroth-order target, so the safe class is not "targets without derivatives" but function-level targets: drift is zeroth-order yet not function-level, because $\mathbf{w}_0$ is extra data. Quantitatively, permutation-invariant linear order-k targets form only a $\mathrm{Bell}(k)/h^k$ fraction of all such targets for hidden width h (2×10^{-6} for $h=1000$, $k=2$), and a generic one has no symmetry at all (Theorem A.5 in Appendix A.4).

7 Experiments

We test three predictions of the theory. First, an H-invariant meta-model matches an unconstrained one on targets whose task-symmetry group contains H and is capped by the ceiling $1-\rho_H(T)$ of Theorem 3.3 otherwise. Second, the cost of quotienting is realized only to the extent that the zoo contains several representatives of the same orbit, while invariance helps when the target allows it. Third, optimizers are equivariant exactly under the groups of Theorem 5.1. Finally, we confirm the separations at the scale of LoRA-fine-tuned language models. Details of all setups are in Appendix J.

7.1 Meta-Learning Across the LoRA Ladder

Setup. We generate a zoo of 12,000 rank-3 LoRA adapters of a two-layer network $f(x;\mathbf{W}) = u^\top \tanh(\mathbf{W}x)$ with $p=8$. Adapters are fine-tuned-like (A close to a shared initialization A_0, B small) and each is stored in a random gauge $C = Q_1 D Q_2$ with Haar-random $Q_1, Q_2 \in \mathrm{O}(r)$ and log-normal diagonal D, so the zoo distribution is $\mathrm{O}(r)$-invariant and has a known GL_r gauge prior. We predict five targets, one or more per level: the loss, Euclidean sensitivity, the ℓ_1 gradient norm, the gradient sparsity ratio $\|\nabla L\|_1/\|\nabla L\|_2$, and the distance from initialization (all in log scale). We compare four meta-models with the same capacity that differ only in how they see the factors: an unconstrained MLP on (A,B); an SP_r-invariant set network over the rank-one components (a_i,b_i) with a sign-symmetrized encoder (Zaheer et al., 2017); an $\mathrm{O}(r)$-invariant MLP on the complete invariants $(A^\top A, A^\top B, B^\top B)$; and a GL_r-invariant MLP. All four also receive the same function-level context computed from $A^\top B$, so differences are due to the factorization alone. We estimate the ceilings $1-\rho_H(T)$ by Monte Carlo on an independent sample of 4,000 adapters; for GL_r the ceiling is taken with respect to the zoo's gauge prior.

Results. Figure 2 reproduces the ladder. On the loss, all four models reach $R^2 \approx 0.98$. On sensitivity, the GL_r-invariant model drops to 0.19 against a ceiling of 0.25, while the others remain above 0.95. On the two sparsity targets, the $\mathrm{O}(r)$-invariant model is capped (ceilings 0.98 and 0.67)

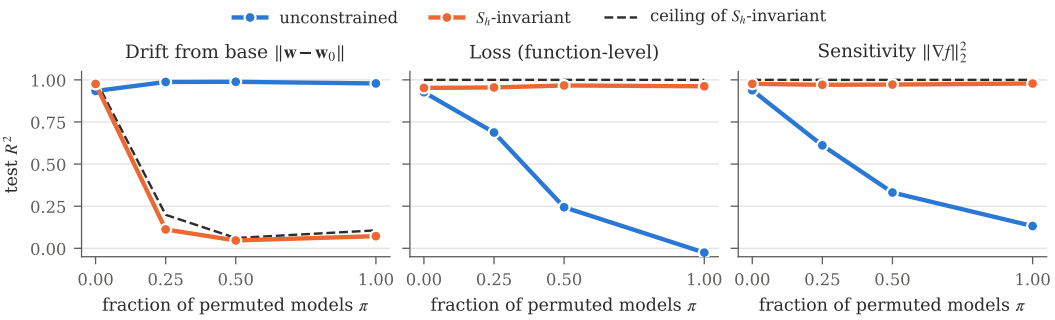


Figure 3: Permutation symmetry cuts both ways. As the fraction π of models stored with permuted neurons grows, the S_h-invariant meta-model collapses to its ceiling on drift from the base (left), a reference-based target, while it keeps near-perfect accuracy on the permutation-invariant targets (center, right), where the unconstrained model degrades.

and the SP_r-invariant model is not. On distance from initialization, every invariant model is capped, at ceilings of 0.42 for SP_r and $\mathrm{O}(r)$ and 0.02 for GL_r, while the unconstrained model reaches 0.99. No invariant model exceeds its ceiling. Two further observations are worth noting. The ℓ_1 gradient norm has an $\mathrm{O}(r)$ ceiling of 0.98: its task symmetry is exactly SP_r, yet quotienting $\mathrm{O}(r)$ costs only 2% of the variance on this zoo, which is precisely the kind of approximate invariance that ρ_H quantifies and an algebraic criterion cannot. And on the sparsity ratio the SP_r-invariant model outperforms the unconstrained one (0.57 versus 0.31), an instance of the sample-complexity benefit of the correct invariance.

7.2 Permutation Symmetry and Drift from a Base Model

Setup. We generate $6{,}000$ two-layer MLPs ($h = 12$ hidden units) as perturbations of a shared base $\mathbf{w}_0$ with distinct neurons, with log-normal perturbation scale. A fraction π of the zoo is stored with randomly permuted hidden units, modeling zoos assembled from different sources or training pipelines. We compare an unconstrained MLP with an S_h-invariant set network over neurons on three targets: drift from the base, the loss, and Euclidean sensitivity. For $\pi < 1$ the zoo distribution is not S_h-invariant, and we report the error of the Bayes-optimal S_h-invariant predictor (the conditional mean given the orbit) as the ceiling.

Results. Figure 3 shows both sides of the prescription. When no model is permuted ($\pi = 0$), every orbit meets the zoo in a single point and the invariant model loses nothing: it predicts drift with $R^2 = 0.98$. As soon as the zoo contains permuted representatives, its drift predictions collapse to the ceiling ($R^2 = 0.11$, 0.05 and 0.07 at $\pi = 0.25$, 0.5 and 1, against ceilings of 0.20, 0.06 and 0.11), while the unconstrained model stays at 0.98 or above. On loss and sensitivity, whose task-symmetry group contains S_h, the picture reverses: the invariant model stays at $R^2 \geq 0.95$ and the unconstrained model degrades to -0.03 and 0.13 at $\pi = 1$, because it must learn the symmetry from data. The cost of the wrong invariance is thus real but conditional on how the zoo populates each orbit, and ρ_H measures it.

7.3 The Optimizer Selects the Symmetry

We fine-tune a rank-4 LoRA adapter on the first layer of a small network trained on the digits dataset, starting from a pretrained base. From a mid-training LoRA state $\mathbf{w}$, we continue training for 200 steps from $\mathbf{w}$ and from $C \cdot \mathbf{w}$ for ten random C drawn from each of $\mathrm{O}(r)$, SP_r, and GL_r, with either SGD or Adam, in double precision. Table 1 reports the difference in final test loss. SGD ends at the same loss, to machine precision, from every orthogonal reparameterization, whereas Adam does so only from signed permutations and deviates by up to 5.7×10^{-2} under rotations. Both optimizers deviate under generic GL_r reparameterizations. This matches Theorems 5.1 and 5.2 exactly.

Table 1: Mean (max) absolute difference in test loss after 200 further training steps from $C \cdot \mathbf{w}$ versus from $\mathbf{w}$, over ten draws of C per group. Reference test losses are 0.345 (SGD) and 0.331 (Adam).

	$C \in \mathrm{O}(r)$	$C \in \mathrm{SP}_r$	$C \in \mathrm{GL}_r$
SGD	3×10^{-16} (1×10^{-15})	3×10^{-17} (2×10^{-16})	1.8×10^{-2} (2.9×10^{-2})
Adam	3.0×10^{-2} (5.7×10^{-2})	3×10^{-17} (1×10^{-16})	2.5×10^{-2} (5.1×10^{-2})

7.4 LoRA-Fine-Tuned Language Models

At the scale of language models (Appendix A.5), the Hessian trace of a LoRA-fine-tuned BERT model is invariant under $\mathrm{O}(r)$ reparameterizations up to estimation noise but changes by more than three orders of magnitude under generic GL_r ones, and the Gaussian perturbation robustness of Llama 3.2 3B on GSM8K varies by more than an order of magnitude in test loss along a single GL_r orbit, with the balanced representative the most robust.

8 Related Work

Reparameterization and sharpness. Flatness of minima has long been linked to generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2020). Dinh et al. (2017) showed that rescaling symmetries of ReLU networks produce functionally equivalent minima of arbitrary sharpness, which motivated scale-invariant and adaptive sharpness measures (Kwon et al., 2021) and a reassessment of the sharpness–generalization link (Andriushchenko et al., 2023); related effects arise from transformer symmetries (da Silva et al., 2025). Our $\mathrm{O}(r)$ level is the LoRA instance of this phenomenon. We differ in identifying the exact surviving group for whole families of targets, in quantifying the cost of quotienting beyond it, and in transferring the question to weight-space learning.

Weight-space learning. Meta-models on weights predict accuracy and generalization (Unterthiner et al., 2020; Eilertsen et al., 2020; Schürholt et al., 2021), training-set size (Salama et al., 2024), and lineage (Horwitz et al., 2024); see Han et al. (2026) for a survey. Architectures are typically equivariant to the model group: to neuron permutations (Navon et al., 2023; Zhou et al., 2023a;b; Lim et al., 2023; Kofinas et al., 2024), additionally to scalings and sign flips (Kalogeropoulos et al., 2024; Tran et al., 2024), or to GL_r for LoRA (Putterman et al., 2024). Our results say when such equivariance is appropriate and what it costs otherwise.

Benefits and costs of invariance. Invariance provably improves sample complexity (Elesedy & Zaidi, 2021; Bietti et al., 2021; Tahmasebi & Jegelka, 2023), and approximation–generalization trade-offs under approximate or misspecified symmetry have been studied (Wang et al., 2022; Petrache & Trivedi, 2023). Invariance can also be obtained by canonicalization or frame averaging (Kaba et al., 2023; Puny et al., 2022; Dym et al., 2024). We complement the gain side with an exact cost for weight-space targets and compute which group is correct for a given target. Further related work is discussed in Appendix I.

9 Conclusion

Parameter symmetries and task symmetries answer different questions: the first identify weights that realize the same function, the second weights that have the same prediction target. We showed that the task-symmetry group is the largest lossless quotient, that quotienting beyond it costs an exactly computable orbit variance with an associated R^2 ceiling, and we computed task-symmetry groups for derivative, curvature, trajectory, and reference-based targets. In LoRA these form a strict ladder $\mathrm{GL}_r \supsetneq \mathrm{O}(r) \supsetneq \mathrm{SP}_r \supsetneq \{I_r\}$, whose middle levels are set by the optimizer's geometry. For permutation symmetry, the dominant quotient in weight-space architectures, the breaking mechanism is a reference frame, and it affects the displacement targets that fine-tuning zoos are built around. The practical message is not to abandon invariant learning but to choose the group from the target. Limitations and open directions are discussed in Appendix B.

AI USE STATEMENT

Large language model tools (ChatGPT 5.6 Sol and Claude Fable 5.1) were used primarily for copyediting, including improvements to grammar, wording, clarity, and LaTeX presentation. The research direction, core ideas, theoretical results, and experimental design are original contributions of the authors. The authors carefully reviewed and verified all technical content and take full responsibility for the final manuscript.

REFERENCES

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022. 7, 25

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning (ICML)*, pp. 840–902, 2023. 2, 9

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016. 26

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. 26

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning (ICML)*, pp. 404–413, 2018. 6, 26

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, pp. 560–569, 2018. 6, 26

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9

Johanni Brea, Berfin Simsek, Bernd Illing, and Wulfram Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *arXiv preprint arXiv:1907.02911*, 2019. 1, 25

Marvin F da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & riemannian geometry finds it. In *International Conference on Machine Learning (ICML)*, 2025. 9

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 26

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, pp. 1019–1028, 2017. 1, 2, 5, 9

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 384–395, 2018. 16, 26

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 26

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *International Conference on Machine Learning (ICML)*, pp. 12228–12267, 2024. 9, 16

Gabriel Eilertsen, Daniel Jönsson, Timo Ropinski, Jonas Unger, and Anders Ynnerman. Classifying the classifier: Dissecting the weight space of neural networks. In *European Conference on Artificial Intelligence (ECAI)*, pp. 1119–1126, 2020. 1, 9

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. 26

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning (ICML)*, pp. 2959–2969, 2021. 9

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-aware dataset selection with datamodels. In *International Conference on Machine Learning (ICML)*, pp. 12491–12526, 2024. 26

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022. 25

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017. 26

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. 26

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 26

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020. 25

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. 26

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018. 25

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 25

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018. 26

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 26

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. 1, 26

Xiaolong Han, Zehong Wang, Bo Zhao, Binchi Zhang, Jundong Li, Damian Borth, Rose Yu, Haggai Maron, Yanfang Ye, Lu Yin, and Ferrante Neri. A survey of weight space learning: Understanding, representation, and generation, 2026. 1, 9

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024. 6, 26

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012. 26

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. 2, 9

Eliahu Horwitz, Bar Cavia, Jonathan Kahana, and Yedid Hoshen. Representing model weights with language using tree experts. *arXiv preprint arXiv:2410.13569*, 2024. 1, 7, 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. 1, 26

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022. 26

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 26

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020. 9

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning (ICML)*, pp. 15546–15566, 2023. 9, 16

Jonathan Kahana, Eliahu Horwitz, Imri Shuval, and Yedid Hoshen. Deep linear probe generators for weight space learning. *arXiv preprint arXiv:2410.10811*, 2024. 26

Ioannis Kalogeropoulos, Giorgos Bouritsas, and Yannis Panagakis. Scale equivariant graph metanetworks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 9

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 9

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6, 26

Boris Knyazev, Abhinav Moudgil, Guillaume Lajoie, Eugene Belilovsky, and Simon Lacoste-Julien. Accelerating training with neuron interaction and nowcasting networks. *arXiv preprint arXiv:2409.04434*, 2024. 26

Miltiadis Kofinas, Boris Knyazev, Yan Zhang, Yunlu Chen, Gertjan J Burghouts, Efstratios Gavves, Cees GM Snoek, and David W Zhang. Graph neural networks for learning equivariant representations of neural networks. *arXiv preprint arXiv:2403.12143*, 2024. 1, 9

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations (ICLR)*, 2021. 26

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 5905–5914, 2021. 2, 9

John W. Lamperti. On the isometries of certain function-spaces. *Pacific Journal of Mathematics*, 8 (3):459–466, 1958. 22

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 598–605, 1989. 16, 26

Bingcong Li, Liang Zhang, and Niao He. Implicit regularization of sharpness-aware minimization for scale-invariant problems. *arXiv preprint arXiv:2410.14802*, 2024. 26

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 25

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. 26

Derek Lim, Haggai Maron, Marc T Law, Jonathan Lorraine, and James Lucas. Graph metanetworks for processing diverse neural architectures. *arXiv preprint arXiv:2312.04501*, 2023. 1, 9

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023. 26

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015. 26

Dhagash Mehta, Tianran Chen, Tingting Tang, and Jonathan D Hauenstein. The loss surface of deep linear networks viewed through the algebraic geometry lens. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5664–5680, 2021. 25

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. 26

Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. g-sgd: Optimizing relu neural networks in its positively scale-invariant space. *arXiv preprint arXiv:1802.03713*, 2018. 25

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. 26

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 26

Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. In *International Conference on Machine Learning*, pp. 25790–25816. PMLR, 2023. 1, 9

Behnam Neyshabur, Ruslan Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2422–2430, 2015. 1, 25

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 26

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 9

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. 26

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations (ICLR)*, 2022. 9

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. 1, 9

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019. 26

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 26

Mohammad Salama, Jonathan Kahana, Eliahu Horwitz, and Yedid Hoshen. Dataset size recovery from lora weights. *arXiv preprint arXiv:2406.19395*, 2024. 1, 9

Konstantin Schürholt, Dimche Kostadinov, and Damian Borth. Self-supervised representation learning on neural network weights for model characteristic prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 9

Konstantin Schürholt, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Hyper-representations as generative models: Sampling unseen neural network weights. *Advances in Neural Information Processing Systems*, 35:27906–27920, 2022a. 26

Konstantin Schürholt, Diyar Taskiran, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Model zoos: A dataset of diverse populations of neural network models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022b. 1

Konstantin Schürholt, Michael W Mahoney, and Damian Borth. Towards scalable and versatile weight space learning. *arXiv preprint arXiv:2406.09997*, 2024. 26

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. 26

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning (ICML)*, pp. 9722–9732, 2021. 1, 25

Bedionita Soro, Bruno Andreis, Hayeon Lee, Wonyong Jeong, Song Chong, Frank Hutter, and Sung Ju Hwang. Diffusion-based neural network weights generation. *arXiv preprint arXiv:2402.18153*, 2024. 26

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4, 9

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. 26

Viet-Hoang Tran, Thieu N. Vo, Tho Tran-Huu, An T. Nguyen, and Tan M. Nguyen. Monomial matrix group equivariant neural functional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 9

Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. *arXiv preprint arXiv:2002.11448*, 2020. 1, 9

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024. 26

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning (ICML)*, pp. 23078–23091, 2022. 9

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *arXiv preprint arXiv:2407.05000*, 2024a. 26

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Lora-pro: Are low-rank adapters properly optimized? *arXiv preprint arXiv:2407.18242*, 2024b. 26

YuQing Xie and Tess Smidt. A tale of two symmetries: Exploring the loss landscape of equivariant models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 25

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*, pp. 9807–9825. PMLR, 2023. 26

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. 26

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3391–3401, 2017. 7, 26

Guy Zamir, Aryan Dokania, Bo Zhao, and Rose Yu. Improving learning to optimize using parameter symmetries. *arXiv preprint arXiv:2504.15399*, 2025. 26

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023. 26

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. 25

Allan Zhou, Kaien Yang, Kaylee Burns, Adriano Cardace, Yiding Jiang, Samuel Sokota, J Zico Kolter, and Chelsea Finn. Permutation equivariant neural functionals. *Advances in neural information processing systems*, 36:24966–24992, 2023a. 1, 9

Allan Zhou, Kaien Yang, Yiding Jiang, Kaylee Burns, Winnie Xu, Samuel Sokota, J. Zico Kolter, and Chelsea Finn. Neural functional transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. 1, 9

Liu Ziyin, Yizhou Xu, Tomaso Poggio, and Isaac Chuang. Parameter symmetry potentially unifies deep learning theory. *arXiv preprint arXiv:2502.05300*, 2025. 1, 25

A Additional Results Deferred from the Main Text

A.1 Summary of the LoRA Ladder

Table 2: The LoRA ladder (Theorems 4.3 and 5.2). Each level is lost by a different mechanism. All groups are exact.

Target class	Examples	Task-symmetry group	Reason
Function-level	loss, accuracy, $\|A^\top B\|$	GL_r	depends on f only
Euclidean derivative	$\|\nabla L\|_2^2$, $\mathrm{Tr}\,\nabla^2 L$, SGD step, Gaussian noise	$\mathrm{O}(r)$	isometries of ℓ_2
ℓ_q derivative, $q \neq 2$	$\|\nabla L\|_1$ (sparsity), sign/Adam step	SP_r	isometries of ℓ_q
Reference-based	distance from initialization	$\{I_r\}$	stabilizer of reference

A.2 Remarks on the LoRA Ladder

Remark A.1 (Coordinate targets). At the extreme, a single partial derivative $\partial L/\partial w_i$ is preserved only by g with $\rho(g)^\top e_i = e_i$, and the intersection over all coordinates is the kernel of ρ. Such targets are less natural than those in Theorem 4.3, and we mention them only to complete the picture.

Remark A.2 (Balancing canonicalizes GL_r to $\mathrm{O}(r)$). The balanced representative used in Section 7, which minimizes $\|CA\|_F^2 + \|C^{-\top}B\|_F^2$ over the orbit, is unique up to left multiplication by $\mathrm{O}(r)$. Balancing therefore maps each GL_r-orbit to an $\mathrm{O}(r)$-orbit. An $\mathrm{O}(r)$-invariant predictor applied to balanced factors is GL_r-invariant, and by Theorem 4.3 it cannot predict the sensitivity of the factorization actually deployed. Canonicalization (Kaba et al., 2023; Dym et al., 2024) is a quotient and must obey the same prescription. Training itself can also act as an implicit gauge choice: gradient flow conserves $AA^\top - BB^\top$ (Du et al., 2018), so factors trained from a given initialization lie on a slice of each orbit determined by $A_0A_0^\top - B_0B_0^\top$.

A.3 Curvature Targets

Second-order targets give a systematic family rather than isolated examples. For a G-invariant loss, Equation (6) with $k = 2$ reads $\nabla^2 L(g\mathbf{w}) = \rho^\star(g)\,\nabla^2 L(\mathbf{w})\,\rho^\star(g)^\top$, a congruence rather than a similarity.

Proposition A.3 (Curvature ladder). *Let L be G-invariant and twice differentiable. (i) The inertia of $\nabla^2 L$ (numbers of positive, negative and zero eigenvalues, hence the Morse index and the rank) is invariant under all of G. (ii) Every spectral function of $\nabla^2 L$ (trace, top eigenvalue, spectral gap, effective rank, the full eigenvalue distribution) is invariant under $\{g \in G : \rho(g) \in \mathrm{O}(d)\}$. (iii) Entrywise functions such as $\sum_{i,j} |(\nabla^2 L)_{ij}|$ or $\sum_i |(\nabla^2 L)_{ii}|$ are invariant under $\{g \in G : \rho(g)$ is a signed permutation matrix$\}$.*

Part (i) is Sylvester's law of inertia. Thus the qualitative type of a critical point is a function-level property, sharpness is an O-level property, and entrywise curvature statistics used in pruning (LeCun et al., 1989) are signed-permutation-level properties. In LoRA, the sharpness level is exactly $\mathrm{O}(r)$ by Theorem 4.3. Restricting to linear functionals hides this family: in Sym^k, the only $\mathrm{O}(d)$-invariant linear functionals are multiples of the full trace (Theorem E.3 in Appendix E), while nonlinear spectral functionals are unrestricted.

A.4 How Much Does Permutation Quotienting Remove?

For linear order-k targets $\langle a, \nabla^{\otimes k} f\rangle$ with a in the ambient tensor space, let $A_G^{(k)} = \{a : \rho(g)^{\otimes k} a = a\ \forall g \in G\}$ be the targets a G-invariant meta-model can represent.

Proposition A.4 (Character formula). *For finite or compact G with Haar probability ν, $\dim A_G^{(k)} = \int_G (\mathrm{Tr}\,\rho(g))^k \,\mathrm{d}\nu(g)$.*

Theorem A.5 (Permutation-invariant targets are rare). *For the MLP above,*

$$\dim A_{S_h}^{(k)} = m^k \sum_{j=1}^{\min(h,k)} S(k,j), \qquad \textit{so} \qquad \frac{\dim A_{S_h}^{(k)}}{d^k} = \frac{\mathrm{Bell}(k)}{h^k} \quad \textit{for } h \geq k, \tag{11}$$

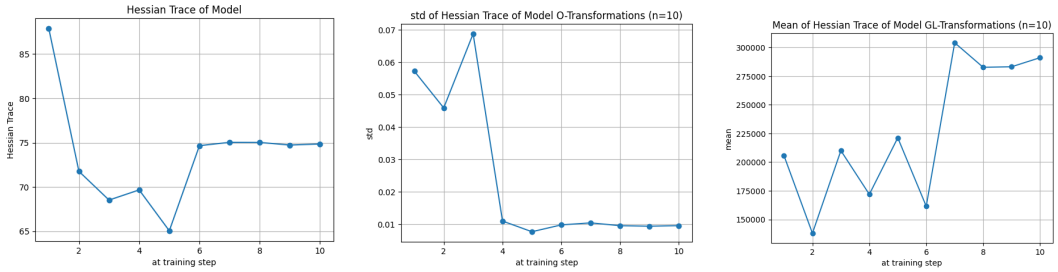


Figure 4: Hessian trace of a LoRA-fine-tuned BERT model across training (left); standard deviation of the trace over ten Haar-random $\mathrm{O}(r)$ reparameterizations (center); mean trace over ten generic GL_r reparameterizations (right). All reparameterized models compute the same function.

where $S(k,j)$ are Stirling numbers of the second kind and $\mathrm{Bell}(k) = \sum_j S(k,j)$. Moreover, for every $k \geq 1$ and $h \geq 2$, a Lebesgue-generic a is fixed by no nontrivial permutation.

For hidden width $h = 1000$ and $k = 2$, permutation-invariant linear second-order targets form a 2×10^{-6} fraction of all linear second-order targets. The invariant tensors are indexed by set partitions of the k derivative slots: they can only contract hidden-unit indices by equality patterns, i.e., by traces and sums over neurons. Everything resolved at the level of individual neurons is removed, which is precisely the information displacement targets need. Restricting to symmetric tensors changes the constant but not the h^{-k} decay (Appendix H).

A.5 LoRA-Fine-Tuned Language Models

Hessian trace: $\mathrm{O}(r)$ but not GL_r. We fine-tune BERT with LoRA on SST-2 under the squared loss and, at each epoch, transform every LoRA layer with the same matrix C, drawn either from the Haar measure on $\mathrm{O}(r)$ or with i.i.d. standard Gaussian entries (a generic element of GL_r); we use ten draws of each. By Theorem A.3 the Hessian trace is exactly invariant under the first family. Figure 4 confirms this: under orthogonal transformations the standard deviation across draws stays below 0.07 on a trace of about 75 (estimation noise), whereas generic invertible transformations change the trace by more than three orders of magnitude, although all transformed models compute the same function.

Perturbation robustness along a GL_r orbit. For each LoRA layer i we compare the trained factors with scalar-rescaled factors $(s^{-1}A_i, sB_i)$, which preserve $A_i^\top B_i$ exactly, and with the balanced representative $(P_i^\star A_i, P_i^{\star -\top} B_i)$ minimizing $\|P_i A_i\|_F^2 + \|P_i^{-\top} B_i\|_F^2$. Writing $S_A = A_i A_i^\top$ and $S_B = B_i B_i^\top$, the optimal metric $X_i^\star = P_i^{\star\top} P_i^\star$ solves $X_i^\star S_A X_i^\star = S_B$, so $X_i^\star = S_A^{-1/2}(S_A^{1/2} S_B S_A^{1/2})^{1/2} S_A^{-1/2}$, and we take the symmetric square root $P_i^\star = (X_i^\star)^{1/2}$; any other choice differs by an orthogonal factor (Theorem A.2). The target is the Gaussian perturbation robustness T_σ of Theorem 4.3, whose task-symmetry group is $\mathrm{O}(r)$: if $\rho(g)$ is orthogonal then $\rho(g)^{-1}\epsilon$ has the same law as ϵ.

We fine-tune Llama 3.2 3B with LoRA on GSM8K, draw five perturbations for each σ, and report the average test loss (Table 3); a GPT-2/Wikitext2 replication is in Appendix K. All unperturbed representatives have identical loss. Rescaled representatives become fragile at progressively smaller σ, reaching losses above 13 at $\sigma = 10^{-5}$ for $s = 10^5$, while the balanced representative is at least as robust as the trained factors at every scale. By Theorem 3.2, every GL_r-invariant predictor, including one applied to balanced factors, incurs error at least half the spread in each column.

B Limitations and Future Work

Our quantitative cost is canonical only for compact groups; for noncompact groups such as GL_r it depends on a prior over the group. The exact counts in Theorem A.5 are for two-layer MLPs; extending them to deep MLPs, where layers carry tensor-product rather than block representations, and to attention-head permutations is direct but not done here. Set-level targets (whether models share a base or can be merged), for which the relevant group is the diagonal of G^n, and equivariant

Table 3: Averaged test loss under perturbations with standard deviation σ for different factorizations of the same LoRA model (Llama 3.2 3B, GSM8K). Lower is better.

Model \ σ	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	$3 \cdot 10^{-1}$	$5 \cdot 10^{-1}$
Baseline	**0.9634**	**0.9634**	**0.9634**	**0.9634**	**0.9634**	0.9636	0.9917	1.9998	4.7824
Balanced $P_i^\star$	**0.9634**	**0.9634**	**0.9634**	**0.9634**	**0.9634**	**0.9635**	**0.9898**	**1.3452**	**1.7136**
Scaled $s = 10^2$	**0.9634**	**0.9634**	**0.9634**	0.9649	1.1836	16.4109	20.4792	20.2002	17.9131
Scaled $s = 10^3$	**0.9634**	**0.9634**	0.9641	1.5015	20.1965	20.2504	23.8895	21.9865	18.0028
Scaled $s = 10^4$	**0.9634**	0.9642	1.0733	14.3455	18.1590	18.7772	24.4439	22.4494	18.1642
Scaled $s = 10^5$	0.9639	1.1025	13.8518	18.9335	22.0456	18.1350	24.3826	22.4639	18.1619
Unperturbed	**0.9634**								

outputs such as per-neuron importance are further mechanisms we leave to future work. Empirically, our meta-learning experiments use synthetic zoos in which the spread of each orbit is controlled; measuring $\rho_H(T)$ on large real zoos, where training partly fixes the gauge (Theorem A.2), is the main empirical question left open.

C Notation and Standing Conventions

Throughout, $\mathcal{W} \subseteq \mathbb{R}^d$ is the weight space, and a group G acts on it by a representation $\rho : G \to \mathrm{GL}_d(\mathbb{R})$, $g\mathbf{w} = \rho(g)\mathbf{w}$. We write $\rho^\star(g) = \rho(g)^{-\top}$ for the contragredient representation; it is a representation because $\rho^\star(gh) = (\rho(g)\rho(h))^{-\top} = \rho^\star(g)\rho^\star(h)$. Inner products on $\mathbb{R}^d$, on matrices, and on tensors are the standard Euclidean (Frobenius) ones. For a k-times differentiable $F : \mathbb{R}^d \to \mathbb{R}$ we identify the k-th derivative $\nabla^{\otimes k} F(\mathbf{w})$ with the symmetric k-linear form $D^k F(\mathbf{w})[v_1, \dots, v_k]$, or equivalently with a tensor in $\mathrm{Sym}^k(\mathbb{R}^d)$ via $D^k F(\mathbf{w})[v_1, \dots, v_k] = \langle \nabla^{\otimes k} F(\mathbf{w}), v_1 \otimes \cdots \otimes v_k \rangle$. For a matrix M we write $M^{\otimes k}$ for its k-fold Kronecker power, which acts on tensors by $M^{\otimes k}(v_1 \otimes \cdots \otimes v_k) = Mv_1 \otimes \cdots \otimes Mv_k$ and satisfies $\langle M^{\otimes k} S, T \rangle = \langle S, (M^\top)^{\otimes k} T \rangle$ and $\mathrm{Tr}(M^{\otimes k}) = (\mathrm{Tr}\, M)^k$. For LoRA we write $\mathbf{w} = (A, B) \in \mathbb{R}^{r \times p} \times \mathbb{R}^{r \times p}$ with $r < p$. A signed permutation matrix is a matrix with exactly one nonzero entry, equal to ± 1, in every row and column; these form the hyperoctahedral group SP_r of order $2^r r!$, and $S^{-\top} = S$ for every $S \in \mathrm{SP}_r$. All models are assumed to be as smooth as the derivative targets under consideration require.

D Proofs for Section 3

D.1 Proof of Theorem 3.1

Proof. We prove (ii)⇔(i)⇔(iii) and then the final claim.

(i)⇔(ii). T is H-invariant iff $T(g\mathbf{w}) = T(\mathbf{w})$ for all $g \in H$ and all $\mathbf{w}$, which by Theorem 2.1 says exactly that every $g \in H$ belongs to G_T, i.e., $H \leq G_T$. Note that G_T is indeed a subgroup: it contains the identity; if $g, g' \in G_T$ then $T(gg'\mathbf{w}) = T(g'\mathbf{w}) = T(\mathbf{w})$; and if $g \in G_T$ then, applying the definition at the point $g^{-1}\mathbf{w}$, $T(\mathbf{w}) = T(gg^{-1}\mathbf{w}) = T(g^{-1}\mathbf{w})$, so $g^{-1} \in G_T$.

(iii)⇒(i). If $T = \overline{T} \circ \pi$, then for $g \in H$ we have $\pi(g\mathbf{w}) = Hg\mathbf{w} = H\mathbf{w} = \pi(\mathbf{w})$, hence $T(g\mathbf{w}) = \overline{T}(\pi(g\mathbf{w})) = \overline{T}(\pi(\mathbf{w})) = T(\mathbf{w})$.

(i)⇒(iii). Suppose T is H-invariant and define $\overline{T}(H\mathbf{w}) := T(\mathbf{w})$. This is well defined: if $H\mathbf{w} = H\mathbf{w}'$, then $\mathbf{w}' = g\mathbf{w}$ for some $g \in H$ and $T(\mathbf{w}') = T(\mathbf{w})$. By construction $T = \overline{T} \circ \pi$. If $\overline{T}'$ is another map with $T = \overline{T}' \circ \pi$, then $\overline{T}$ and $\overline{T}'$ agree on the image of π, which is all of $\mathcal{W}/H$, so $\overline{T}' = \overline{T}$.

Continuity. The quotient topology on $\mathcal{W}/H$ is the finest topology making π continuous, and a map $\overline{T}$ out of $\mathcal{W}/H$ is continuous iff $\overline{T} \circ \pi$ is continuous. Since $\overline{T} \circ \pi = T$ is continuous, so is $\overline{T}$.

Final claim. By (ii)⇒(iii), quotienting by G_T is lossless, and by (iii)⇒(ii), every subgroup that can be quotiented without loss is contained in G_T. Hence G_T is the largest such subgroup. □

D.2 Proof of Theorem 3.2

Proof. Let $h \in \mathcal{H}_H$ and set $\varepsilon(h) \coloneqq \sup_{\mathbf{u}\in\mathcal{W}} |T(\mathbf{u}) - h(\mathbf{u})|$. Fix $\mathbf{w} \in \mathcal{W}$ and $g \in H$. Since h is H-invariant, $h(g\mathbf{w}) = h(\mathbf{w})$, and therefore

$$\begin{aligned}|T(g\mathbf{w}) - T(\mathbf{w})| &= |(T(g\mathbf{w}) - h(g\mathbf{w})) + (h(\mathbf{w}) - T(\mathbf{w}))| \\ &\le |T(g\mathbf{w}) - h(g\mathbf{w})| + |T(\mathbf{w}) - h(\mathbf{w})| \le 2\varepsilon(h).\end{aligned}$$

Taking the supremum over $\mathbf{w}$ and g gives $\sup_{\mathbf{w},g} |T(g\mathbf{w}) - T(\mathbf{w})| \le 2\varepsilon(h)$ for every $h \in \mathcal{H}_H$, and taking the infimum over h yields Equation (4). If T is not H-invariant, the right-hand side is strictly positive, so no H-invariant predictor approximates T uniformly to arbitrary precision. □

D.3 Proof of Theorem 3.3

We assume that the action map $H \times \mathcal{W} \to \mathcal{W}$, $(h, \mathbf{w}) \mapsto h\mathbf{w}$, is continuous (this holds for every linear action, and in particular for all examples in the paper), so that $(h, \mathbf{w}) \mapsto T(h\mathbf{w})$ is jointly measurable whenever T is measurable. Let ν_H denote the normalized Haar measure on the compact group H; it is both left and right invariant and invariant under inversion.

Lemma D.1 (Orbit averaging is an orthogonal projection)**.** *For $h \in H$ define $U_h : L^2(\mu) \to L^2(\mu)$ by $(U_hF)(\mathbf{w}) = F(h\mathbf{w})$, and define $\mathcal{R}_H F \coloneqq \int_H U_h F \,\mathrm{d}\nu_H(h)$, i.e., $(\mathcal{R}_H F)(\mathbf{w}) = \int_H F(h\mathbf{w}) \,\mathrm{d}\nu_H(h)$. Then $\mathcal{R}_H$ is the orthogonal projection of $L^2(\mu)$ onto the closed subspace $L^2_H \coloneqq \{F \in L^2(\mu) : F(h\mathbf{w}) = F(\mathbf{w})$ for all $h \in H$ and μ-a.e. $\mathbf{w}\}$.*

Proof. U_h is unitary. By H-invariance of μ, $\|U_hF\|^2 = \int F(h\mathbf{w})^2 \,\mathrm{d}\mu(\mathbf{w}) = \int F(\mathbf{u})^2 \,\mathrm{d}\mu(\mathbf{u}) = \|F\|^2$, and U_h is invertible with inverse $U_{h^{-1}}$. Hence $U_h^* = U_h^{-1} = U_{h^{-1}}$.

$\mathcal{R}_H$ is well defined and bounded. By Jensen's inequality and Tonelli's theorem, $\|\mathcal{R}_H F\|^2 \le \int_H \|U_hF\|^2 \,\mathrm{d}\nu_H(h) = \|F\|^2$.

Range. For $h' \in H$, $(U_{h'}\mathcal{R}_H F)(\mathbf{w}) = \int_H F(hh'\mathbf{w}) \,\mathrm{d}\nu_H(h) = \int_H F(h\mathbf{w}) \,\mathrm{d}\nu_H(h) = (\mathcal{R}_H F)(\mathbf{w})$ by right invariance of ν_H. Thus $\mathcal{R}_H F \in L^2_H$. Conversely, if $F \in L^2_H$, then $U_hF = F$ for all h and hence $\mathcal{R}_H F = F$. Therefore $\mathcal{R}_H^2 = \mathcal{R}_H$ and the range of $\mathcal{R}_H$ is exactly L^2_H, which is closed because it is the range of a bounded idempotent.

Self-adjointness. For $F, F' \in L^2(\mu)$, $\langle \mathcal{R}_H F, F'\rangle = \int_H \langle U_hF, F'\rangle \,\mathrm{d}\nu_H(h) = \int_H \langle F, U_{h^{-1}}F'\rangle \,\mathrm{d}\nu_H(h) = \int_H \langle F, U_hF'\rangle \,\mathrm{d}\nu_H(h) = \langle F, \mathcal{R}_H F'\rangle$, using inversion invariance of ν_H. A self-adjoint idempotent is the orthogonal projection onto its range. □

Proof of Theorem 3.3. Write $\bar{T}_H = \mathcal{R}_H T$.

Item 1. By Theorem D.1 and the Hilbert projection theorem, the minimum of $\mathbb{E}_\mu(T - h)^2 = \|T - h\|^2_{L^2(\mu)}$ over $h \in L^2_H$ is attained at $h = \bar{T}_H$ and equals $\|T - \bar{T}_H\|^2$. It remains to show that this equals $B_H(T)$. For fixed $\mathbf{w}$, the number $\bar{T}_H(\mathbf{w})$ is the mean of the function $h \mapsto T(h\mathbf{w})$ under ν_H, so $\mathrm{Var}_{h\sim\nu_H} T(h\mathbf{w}) = \int_H (T(h\mathbf{w}) - \bar{T}_H(\mathbf{w}))^2 \,\mathrm{d}\nu_H(h)$. Using Tonelli's theorem, the invariance $\bar{T}_H(\mathbf{w}) = \bar{T}_H(h\mathbf{w})$, and unitarity of U_h,

$$\begin{aligned}B_H(T) &= \int_{\mathcal{W}} \int_H \big(T(h\mathbf{w}) - \bar{T}_H(\mathbf{w})\big)^2 \,\mathrm{d}\nu_H(h) \,\mathrm{d}\mu(\mathbf{w}) = \int_H \int_{\mathcal{W}} \big(T(h\mathbf{w}) - \bar{T}_H(h\mathbf{w})\big)^2 \,\mathrm{d}\mu(\mathbf{w}) \,\mathrm{d}\nu_H(h) \\ &= \int_H \big\|U_h(T - \bar{T}_H)\big\|^2_{L^2(\mu)} \,\mathrm{d}\nu_H(h) = \int_H \big\|T - \bar{T}_H\big\|^2_{L^2(\mu)} \,\mathrm{d}\nu_H(h) = \|T - \bar{T}_H\|^2_{L^2(\mu)}.\end{aligned}$$

Item 2. If $H_1 \le H_2$ and F is H_2-invariant, it is in particular H_1-invariant, so $L^2_{H_2} \subseteq L^2_{H_1}$. A minimum of the same functional over a larger set cannot be larger, hence by Item 1 $B_{H_1}(T) = \min_{h\in L^2_{H_1}} \|T - h\|^2 \le \min_{h\in L^2_{H_2}} \|T - h\|^2 = B_{H_2}(T)$.

Item 3. Suppose first that $B_H(T) = 0$. By the computation in Item 1, the set $E \coloneqq \{(h, \mathbf{w}) : T(h\mathbf{w}) \ne \bar{T}_H(h\mathbf{w})\}$ is $\nu_H \otimes \mu$-null. By Tonelli, for ν_H-a.e. h the section $E_h = \{\mathbf{w} : T(h\mathbf{w}) \ne \bar{T}_H(h\mathbf{w})\}$ is μ-null. But $E_h = h^{-1}\{\mathbf{u} : T(\mathbf{u}) \ne \bar{T}_H(\mathbf{u})\}$ and μ is H-invariant, so $\mu(E_h) = \mu\{\mathbf{u} : T(\mathbf{u}) \ne \bar{T}_H(\mathbf{u})\}$ for every h. Hence $T = \bar{T}_H$ μ-a.e. Since $\bar{T}_H$ is invariant, for $\nu_H \otimes \mu$-a.e. $(h, \mathbf{w})$

we get $T(h\mathbf{w}) = \bar{T}_H(h\mathbf{w}) = \bar{T}_H(\mathbf{w}) = T(\mathbf{w})$. Conversely, if $T(h\mathbf{w}) = T(\mathbf{w})$ for a.e. $(h, \mathbf{w})$, then for μ-a.e. $\mathbf{w}$ the function $h \mapsto T(h\mathbf{w})$ is ν_H-a.e. constant, so its variance vanishes and $B_H(T) = 0$.

Now let T be continuous and let μ have full support. The set $O := \{(h, \mathbf{w}) : T(h\mathbf{w}) \neq T(\mathbf{w})\}$ is open in $H \times \mathcal{W}$, as the preimage of the open set $\mathbb{R}\backslash\{0\}$ under the continuous map $(h, \mathbf{w}) \mapsto T(h\mathbf{w}) - T(\mathbf{w})$. Haar measure on a compact group has full support, so $\nu_H \otimes \mu$ assigns positive mass to every nonempty open subset of $H \times \mathcal{W}$ (every such set contains a product of nonempty open sets). Since O is null, it is empty, i.e., $T(h\mathbf{w}) = T(\mathbf{w})$ for all $h \in H$ and all $\mathbf{w}$, which is $H \leq G_T$. The converse implication is immediate from the first part.

Item 4. For every H-invariant $h \in L^2(\mu)$, Item 1 gives $\mathbb{E}_\mu(T - h)^2 \geq B_H(T)$; dividing by $\mathrm{Var}_\mu(T) > 0$ and rearranging gives $R^2(h) \leq 1 - \rho_H(T)$. Finally, $\rho_H(T) \in [0, 1]$: it is nonnegative as a ratio of variances, and since the constant function $\mathbb{E}_\mu T$ is invariant, Item 1 gives $B_H(T) \leq \mathbb{E}_\mu(T - \mathbb{E}_\mu T)^2 = \mathrm{Var}_\mu(T)$. □

Remark D.2 (Noncompact groups). A locally compact group admits a finite invariant measure only if it is compact. For noncompact groups such as $\mathrm{GL}_r(\mathbb{R})$ or the positive rescalings of a ReLU network there is therefore no canonical orbit average. Given any probability measure π on H (a "gauge prior"), the quantity $\mathbb{E}_{\mathbf{w}} \mathrm{Var}_{h\sim\pi} T(h\mathbf{w})$ is still well defined, and when the zoo is generated by drawing an orbit representative and then an independent gauge $h \sim \pi$, it equals the error of the Bayes-optimal H-invariant predictor (the conditional mean given the orbit). This is the "prior-dependent" ceiling reported for GL_r in Section 7; it depends on π and is not intrinsic to T.

E Proofs for Section 4

E.1 The transformation rule Equation (6)

Lemma E.1. *Let $f(x; \cdot)$ be k times differentiable and satisfy $f(x; \rho(g)\mathbf{w}) = f(x; \mathbf{w})$ for all $\mathbf{w}$. Then $\nabla_{\mathbf{w}}^{\otimes k} f(x; \rho(g)\mathbf{w}) = \rho^\star(g)^{\otimes k} \nabla_{\mathbf{w}}^{\otimes k} f(x; \mathbf{w})$.*

Proof. Fix x and write $F = f(x; \cdot)$ and $M = \rho(g)$. Since $\mathbf{w} \mapsto M\mathbf{w}$ is linear, the chain rule gives, for all directions $v_1, \ldots, v_k$, $D^k(F \circ M)(\mathbf{w})[v_1, \ldots, v_k] = D^k F(M\mathbf{w})[Mv_1, \ldots, Mv_k]$ (there are no higher-order terms because M has vanishing second derivative). By invariance, $F \circ M = F$, so $D^k F(M\mathbf{w})[Mv_1, \ldots, Mv_k] = D^k F(\mathbf{w})[v_1, \ldots, v_k]$ for all v_i. Substituting $v_i = M^{-1}u_i$ gives $D^k F(M\mathbf{w})[u_1, \ldots, u_k] = D^k F(\mathbf{w})[M^{-1}u_1, \ldots, M^{-1}u_k]$. In tensor notation the right-hand side is $\langle \nabla^{\otimes k} F(\mathbf{w}), (M^{-1})^{\otimes k}(u_1 \otimes \cdots \otimes u_k) \rangle = \langle (M^{-\top})^{\otimes k} \nabla^{\otimes k} F(\mathbf{w}), u_1 \otimes \cdots \otimes u_k \rangle$. Since the u_i are arbitrary, $\nabla^{\otimes k} F(M\mathbf{w}) = (M^{-\top})^{\otimes k} \nabla^{\otimes k} F(\mathbf{w})$. For $k = 1$ this reads $\nabla F(M\mathbf{w}) = M^{-\top} \nabla F(\mathbf{w})$. □

E.2 Proof of Theorem 4.2

Proof. The main identity. By Theorem E.1, for every $f \in \mathcal{M}$, μ, $g \in G$ and $\mathbf{w}$,

$$T_{f,\mu}(g\mathbf{w}) = \int_{\mathcal{X}} \Phi\big(\rho^\star(g)^{\otimes k} \nabla^{\otimes k} f(x; \mathbf{w})\big) \, \mathrm{d}\mu(x). \tag{12}$$

($\supseteq$) If $\Phi(\rho^\star(g)^{\otimes k} v) = \Phi(v)$ for every $v \in \mathcal{V}_k$, then the integrands in Equation (12) and in the definition of $T_{f,\mu}(\mathbf{w})$ coincide pointwise, because $\nabla^{\otimes k} f(x; \mathbf{w}) \in \mathcal{V}_k$. Hence $g \in G_{T_{f,\mu}}$ for all (f, μ), i.e., $g \in G_{\mathcal{T}_\Phi}$. ($\subseteq$) Let $g \in G_{\mathcal{T}_\Phi}$ and $v \in \mathcal{V}_k$, say $v = \nabla^{\otimes k} f(x_0; \mathbf{w}_0)$. Taking $\mu = \delta_{x_0}$ in Equation (12), the identity $T_{f,\delta_{x_0}}(g\mathbf{w}_0) = T_{f,\delta_{x_0}}(\mathbf{w}_0)$ reads $\Phi(\rho^\star(g)^{\otimes k} v) = \Phi(v)$.

(a) Linear Φ. Let $\Phi = \langle a, \cdot \rangle$. Then $\Phi(\rho^\star(g)^{\otimes k} v) = \langle a, (\rho(g)^{-\top})^{\otimes k} v \rangle = \langle (\rho(g)^{-1})^{\otimes k} a, v \rangle$. The condition $\Phi(\rho^\star(g)^{\otimes k} v) = \Phi(v)$ for all $v \in \mathcal{V}_k$ is equivalent, by linearity, to the same identity for all $v \in V_k = \mathrm{span}\, \mathcal{V}_k$, i.e., to $b_g := (\rho(g)^{-1})^{\otimes k} a - a \in V_k^\perp$. To express this without the inverse, note that $\mathcal{V}_k$ is invariant under $\rho^\star(h)^{\otimes k}$ for every $h \in G$ (by Theorem E.1, the derivative at $h\mathbf{w}$ is again attainable), hence so is V_k. For $u \in V_k^\perp$ and $v \in V_k$ we get $\langle \rho(h)^{\otimes k} u, v \rangle = \langle u, (\rho(h)^\top)^{\otimes k} v \rangle = \langle u, \rho^\star(h^{-1})^{\otimes k} v \rangle = 0$, so $V_k^\perp$ is invariant under $\rho(h)^{\otimes k}$ for all h. Applying $\rho(g)^{\otimes k}$ to b_g gives $\rho(g)^{\otimes k} b_g = -(\rho(g)^{\otimes k} a - a)$, so $b_g \in V_k^\perp$ iff $\rho(g)^{\otimes k} a - a \in V_k^\perp$. When $V_k = \mathrm{Sym}^k(\mathbb{R}^d)$

and $a \in \mathrm{Sym}^k$, the vector $\rho(g)^{\otimes k}a - a$ lies in Sym^k (which is invariant under $\rho(g)^{\otimes k}$), and it is orthogonal to Sym^k only if it vanishes.

(b) Norms. This is the main identity with Φ a norm: the condition says that $\rho^\star(g)^{\otimes k}$ preserves Φ at every attainable point. □

Remark E.2. An earlier version of the order-k criterion assumed that the attainable derivative tensors span all of $\mathrm{Sym}^k(\mathbb{R}^d)$. This fails in general: differentiating $f(x;\rho(\exp(t\xi))\mathbf{w}) = f(x;\mathbf{w})$ at $t=0$ for a Lie algebra element ξ shows that $\nabla f(x;\mathbf{w})$ is orthogonal to the orbit tangent direction $\mathrm{d}\rho(\xi)\mathbf{w}$, so at a fixed $\mathbf{w}$ the attainable gradients lie in a proper subspace. Theorem 4.2(a) states the criterion on the attainable span and requires no spanning assumption.

E.3 Orthogonally invariant coefficient tensors

Proposition E.3. *Let $a \in \mathrm{Sym}^k(\mathbb{R}^d)$ with $Q^{\otimes k}a = a$ for all $Q \in \mathrm{O}(d)$. If k is odd, then $a = 0$. If $k = 2m$ is even, then $a = c\,\mathrm{Sym}(I_d^{\otimes m})$ for some $c \in \mathbb{R}$, and $\langle a, u^{\otimes 2m}\rangle = c\|u\|_2^{2m}$. Thus, among linear functionals of order-k derivatives, only multiples of the full trace are $\mathrm{O}(d)$-invariant.*

Proof. Odd order. $Q = -I_d \in \mathrm{O}(d)$ acts on $(\mathbb{R}^d)^{\otimes k}$ as $(-I_d)^{\otimes k} = (-1)^k I$. Invariance gives $(-1)^k a = a$, so $a = 0$ if k is odd.

Even order. Let $k = 2m$ and define the homogeneous polynomial of degree $2m$ $p_a(u) := \langle a, u^{\otimes 2m}\rangle$. For $Q \in \mathrm{O}(d)$, $p_a(Qu) = \langle a, Q^{\otimes 2m}u^{\otimes 2m}\rangle = \langle (Q^\top)^{\otimes 2m}a, u^{\otimes 2m}\rangle = p_a(u)$, since $Q^\top \in \mathrm{O}(d)$ fixes a. The group $\mathrm{O}(d)$ acts transitively on each sphere $\{\|u\|_2 = s\}$ (any unit vector can be mapped to any other by an orthogonal matrix, e.g., a Householder reflection), so p_a is constant on spheres. Homogeneity then gives $p_a(u) = \|u\|_2^{2m}p_a(e_1) = c\|u\|_2^{2m}$ with $c = p_a(e_1)$.

Identification. Let $J = \mathrm{Sym}[(\sum_i e_i \otimes e_i)^{\otimes m}] \in \mathrm{Sym}^{2m}(\mathbb{R}^d)$, where Sym averages over permutations of the $2m$ tensor slots. Since $u^{\otimes 2m}$ is symmetric, $\langle J, u^{\otimes 2m}\rangle = \langle (\sum_i e_i \otimes e_i)^{\otimes m}, u^{\otimes 2m}\rangle = (\sum_i u_i^2)^m = \|u\|_2^{2m}$. Thus $p_a = c\,p_J = p_{cJ}$.

Injectivity of $a \mapsto p_a$. A symmetric tensor is determined by its polynomial through the polarization identity $\langle a, v_1 \otimes \cdots \otimes v_k\rangle = \frac{1}{k!}\frac{\partial^k}{\partial t_1 \cdots \partial t_k} p_a(t_1v_1 + \cdots + t_kv_k)$, which follows by expanding $(t_1v_1 + \cdots + t_kv_k)^{\otimes k}$ and using symmetry of a. Hence $p_a = p_{cJ}$ implies $a = cJ = c\,\mathrm{Sym}(I_d^{\otimes m})$. □

E.4 Proof of Theorem A.3

Proof. Let $M = \rho(g)$. By Theorem E.1 with $k = 2$, in matrix form, $\nabla^2 L(M\mathbf{w}) = M^{-\top}\nabla^2 L(\mathbf{w})M^{-1}$, a congruence of the symmetric matrix $\nabla^2 L(\mathbf{w})$ by the invertible matrix $M^{-\top}$.

(i) Sylvester's law of inertia states that congruent real symmetric matrices have the same numbers of positive, negative, and zero eigenvalues. Hence the inertia, the Morse index (number of negative eigenvalues), and the rank are G-invariant.

(ii) If M is orthogonal, then $M^{-\top} = M$ and $M^{-1} = M^\top$, so $\nabla^2 L(M\mathbf{w}) = M\nabla^2 L(\mathbf{w})M^\top$ is similar to $\nabla^2 L(\mathbf{w})$ and has the same eigenvalues with multiplicities. Every function of the spectrum is therefore preserved.

(iii) Let $M = S$ be a signed permutation matrix, $Se_j = s_j e_{\pi(j)}$ with $s_j \in \{\pm 1\}$. Then $S^{-\top} = S$ and, writing $H = \nabla^2 L(\mathbf{w})$, $(SHS^\top)_{\pi(i)\pi(j)} = s_i s_j H_{ij}$. The entries of $SHS^\top$ are thus a permutation of the entries of H up to signs, and diagonal entries are mapped to diagonal entries. Any function of the multiset of absolute values of entries, or of diagonal entries, is preserved. □

F Proof of the LoRA Ladder

Recall $\rho(C)(A,B) = (CA, C^{-\top}B)$ for $C \in \mathrm{GL}_r(\mathbb{R})$.

Lemma F.1 (Contragredient action on LoRA gradients)**.** $\rho^\star(C)(X,Y) = (C^{-\top}X,\ CY)$. *In particular, $\rho(C)$ is orthogonal iff $C \in \mathrm{O}(r)$, and $\rho(C)$ is a signed permutation matrix of $\mathbb{R}^{2rp}$ iff $C \in \mathrm{SP}_r$.*

Proof. The candidate map satisfies $\langle (C^{-\top}X, CY), \rho(C)(A,B)\rangle = \langle C^{-\top}X, CA\rangle + \langle CY, C^{-\top}B\rangle = \mathrm{tr}(X^\top C^{-1}CA) + \mathrm{tr}(Y^\top C^\top C^{-\top}B) = \langle (X,Y),(A,B)\rangle$ for all arguments, i.e., it is the inverse transpose of $\rho(C)$. In the standard basis, $\rho(C)$ is block diagonal with blocks $I_p \otimes C$ (acting on the columns of A) and $I_p \otimes C^{-\top}$. It is orthogonal iff C and $C^{-\top}$ are, i.e., iff $C \in \mathrm{O}(r)$; it is a signed permutation matrix iff both blocks are, i.e., iff C and $C^{-\top}$ are signed permutation matrices, which holds iff $C \in \mathrm{SP}_r$ because $S^{-\top} = S$ on SP_r. □

Lemma F.2 (Attainable LoRA gradients)**.** *For $f(x; A, B) = \phi(x, \mathbf{W}_0 + A^\top B)$ write $\Gamma = \partial\phi/\partial\mathbf{W} \in \mathbb{R}^{p\times p}$, evaluated at $\mathbf{W}_0 + A^\top B$. Then $\nabla_A f = B\Gamma^\top$ and $\nabla_B f = A\Gamma$. If A and B have rank $r < p$, there is a nonzero $n \in \mathbb{R}^p$ such that for every $z \in \mathbb{R}^r$ some model in $\mathcal{M}$ has gradient $(zn^\top, 0)$ at (A, B).*

Proof. Gradients. For a perturbation $(\mathrm{d}A, \mathrm{d}B)$, $\mathrm{d}(A^\top B) = \mathrm{d}A^\top B + A^\top \mathrm{d}B$, so $\mathrm{d}f = \langle \Gamma, \mathrm{d}A^\top B + A^\top \mathrm{d}B\rangle = \mathrm{tr}(\Gamma^\top \mathrm{d}A^\top B) + \mathrm{tr}(\Gamma^\top A^\top \mathrm{d}B)$. Using cyclicity, $\mathrm{tr}(\Gamma^\top \mathrm{d}A^\top B) = \mathrm{tr}(\mathrm{d}A^\top B\Gamma^\top) = \langle \mathrm{d}A, B\Gamma^\top\rangle$ and $\mathrm{tr}(\Gamma^\top A^\top \mathrm{d}B) = \langle A\Gamma, \mathrm{d}B\rangle$.

Construction. Since $\mathrm{rank}\, A = r < p$, $\ker A \subseteq \mathbb{R}^p$ has dimension $p - r \geq 1$. Let $N \in \mathbb{R}^{p\times(p-r)}$ have columns forming a basis of $\ker A$ and put $n := Ne_1 \neq 0$. Since B has full row rank, its Moore–Penrose pseudoinverse satisfies $BB^+ = I_r$. Given z, set $M := (B^+ze_1^\top)^\top \in \mathbb{R}^{(p-r)\times p}$ and $\Gamma := NM \in \mathbb{R}^{p\times p}$, and take the linear model $\phi(x, \mathbf{W}) = \langle \Gamma, \mathbf{W}\rangle \in \mathcal{M}$, whose derivative is Γ everywhere. Then $A\Gamma = (AN)M = 0$ and $B\Gamma^\top = BM^\top N^\top = BB^+ze_1^\top N^\top = z(Ne_1)^\top = zn^\top$. □

Lemma F.3 (Linear isometries of ℓ_q)**.** *Let $M \in \mathbb{R}^{r\times r}$ satisfy $\|Mz\|_q = \|z\|_q$ for all z. If $q = 2$, M is orthogonal. If $q \in [1,\infty] \setminus \{2\}$, M is a signed permutation matrix. Conversely, orthogonal matrices preserve ℓ_2, and signed permutation matrices preserve every ℓ_q.*

Proof. The converse statements are immediate: $\|Qz\|_2^2 = z^\top Q^\top Qz = \|z\|_2^2$, and a signed permutation matrix permutes the absolute values of the coordinates. In all cases M is injective (it preserves a norm), hence invertible.

$q = 2$. By polarization, $\langle Mz, Mz'\rangle = \frac{1}{4}(\|M(z+z')\|^2 - \|M(z-z')\|^2) = \langle z, z'\rangle$, so $M^\top M = I$.

$q = 1$. M maps the unit ball $\mathbb{B}_1$ bijectively onto itself and, being linear, maps extreme points of $\mathbb{B}_1$ to extreme points. The extreme points of $\mathbb{B}_1$ are $\pm e_1, \ldots, \pm e_r$. Hence $Me_i = s_i e_{\pi(i)}$ with $s_i \in \{\pm 1\}$, and π is injective because M is. Thus M is a signed permutation matrix.

$q = \infty$. The dual norm of ℓ_∞ is ℓ_1, and M is an ℓ_∞ isometry iff $M^{-\top}$ is an ℓ_1 isometry: $\|M^{-\top}y\|_1 = \sup_{\|z\|_\infty\leq 1}\langle M^{-\top}y, z\rangle = \sup_{\|z\|_\infty\leq 1}\langle y, M^{-1}z\rangle = \sup_{\|u\|_\infty\leq 1}\langle y, u\rangle = \|y\|_1$, using that M^{-1} maps the ℓ_∞ unit ball onto itself. By the case $q = 1$, $M^{-\top}$ is a signed permutation matrix, hence so is $M = (M^{-\top})^{-\top}$.

$q \in (1,\infty) \setminus \{2\}$. We show that $u = Me_i$ and $v = Me_j$ have disjoint supports for $i \neq j$. For all $t \in \mathbb{R}$,

$$h(t) := \sum_{k=1}^{r} |u_k + tv_k|^q = \|M(e_i + te_j)\|_q^q = \|e_i + te_j\|_q^q = 1 + |t|^q.$$

Let $K = \{k : u_k \neq 0\}$. For $k \in K$ the map $t \mapsto |u_k + tv_k|^q$ is smooth near $t = 0$, with Taylor expansion $|u_k|^q + \alpha_k t + \beta_k t^2 + O(t^3)$, where $\alpha_k = q|u_k|^{q-1}\,\mathrm{sign}(u_k)v_k$ and $\beta_k = \frac{1}{2}q(q-1)|u_k|^{q-2}v_k^2 \geq 0$. For $k \notin K$ the term equals $|t|^q|v_k|^q$. Hence, as $t \to 0$,

$$h(t) = \textstyle\sum_{k\in K}|u_k|^q + \alpha t + \beta t^2 + \sigma|t|^q + O(t^3), \qquad \alpha = \sum_{k\in K}\alpha_k,\ \beta = \sum_{k\in K}\beta_k,\ \sigma = \sum_{k\notin K}|v_k|^q.$$

Comparing with $1 + |t|^q$: the constant terms agree, and $\alpha = 0$ because all remaining terms are $o(|t|)$. If $q > 2$, then $|t|^q = o(t^2)$, and dividing $\beta t^2 + (\sigma - 1)|t|^q + O(t^3) = 0$ by t^2 and letting $t \to 0$ gives $\beta = 0$, hence $v_k = 0$ for all $k \in K$. If $1 < q < 2$, then $t^2 = o(|t|^q)$, and dividing by $|t|^q$ gives $\sigma = 1 = \|v\|_q^q = \sigma + \sum_{k\in K}|v_k|^q$, hence again $v_k = 0$ for all $k \in K$. In both cases u and v have disjoint supports. The r nonzero vectors $Me_1, \ldots, Me_r$ therefore have pairwise disjoint supports in a space of dimension r, so each is supported on a single coordinate and distinct vectors on distinct coordinates: $Me_i = c_i e_{\pi(i)}$ with π a permutation. Finally $|c_i| = \|Me_i\|_q = \|e_i\|_q = 1$. (For the infinite-dimensional version of this result, see Lamperti (1958).) □

Proof of Theorem 4.3. Throughout, fix (A, B) of full row rank r, which exists since $r < p$, and let n and the models f_z be as in Theorem F.2. By Theorems E.1 and F.1, the gradient at $C \cdot \mathbf{w}$ is $(C^{-\top}\nabla_A f,\ C\nabla_B f)$, where the gradients on the right are evaluated at $\mathbf{w}$.

Item 1. $A^\top B$ is invariant under Equation (2), so $f(\cdot; C \cdot \mathbf{w}) = f(\cdot; \mathbf{w})$ for every $f \in \mathcal{M}$, and every function-level target is invariant under all of GL_r.

Item 2, sensitivity. ($\supseteq$) If $C \in \mathrm{O}(r)$, then $C^{-\top} = C$ and $\|CX\|_F = \|X\|_F$ for all X, so $\|\nabla f(x; C \cdot \mathbf{w})\|_2^2 = \|\nabla_A f\|_F^2 + \|\nabla_B f\|_F^2$ pointwise in x, and the target is invariant for every model and data distribution. ($\subseteq$) Let C preserve every Euclidean sensitivity target. Take $f = f_z$ and $\mu = \delta_x$. At $\mathbf{w}$ the target equals $\|zn^\top\|_F^2 = \|z\|_2^2\|n\|_2^2$; at $C \cdot \mathbf{w}$ it equals $\|C^{-\top}zn^\top\|_F^2 = \|C^{-\top}z\|_2^2\|n\|_2^2$. Since $n \neq 0$, $\|C^{-\top}z\|_2 = \|z\|_2$ for all z, so $C^{-\top}$ is orthogonal by Theorem F.3, and so is $C = (C^{-\top})^{-\top}$.

Item 2, Hessian trace. ($\supseteq$) If $C \in \mathrm{O}(r)$, $\rho(C)$ is orthogonal (Theorem F.1) and Theorem A.3(ii) applies to every admissible loss. ($\subseteq$) Let Γ be the matrix constructed in Theorem F.2 for a given z, and consider the squared loss $L(\mathbf{w}) = \frac{1}{2}(\langle \Gamma, \mathbf{W}_0 + A^\top B\rangle - y)^2$ with $y := \langle \Gamma, \mathbf{W}_0 + A^\top B\rangle$ evaluated at the fixed $\mathbf{w}$. The residual $\varrho(\mathbf{w}') = \langle \Gamma, \mathbf{W}_0 + A'^\top B'\rangle - y$ is function-level, so it vanishes at both $\mathbf{w}$ and $C \cdot \mathbf{w}$. Since $\nabla^2 L = \nabla\varrho\nabla\varrho^\top + \varrho\nabla^2\varrho$, at both points $\operatorname{Tr}\nabla^2 L = \|\nabla\varrho\|_2^2$, which is the Euclidean sensitivity of the model $\langle \Gamma, \mathbf{W}\rangle$. The previous argument now forces $C \in \mathrm{O}(r)$.

Item 2, Gaussian robustness. ($\supseteq$) If $C \in \mathrm{O}(r)$, then $\rho(C)$ is orthogonal, $\rho(C)^{-1}\epsilon$ has the same law as ϵ, and, using invariance of L, $T_\sigma(C \cdot \mathbf{w}) = \mathbb{E}\, L(\rho(C)\mathbf{w} + \epsilon) - L(\rho(C)\mathbf{w}) = \mathbb{E}\, L(\rho(C)(\mathbf{w} + \rho(C)^{-1}\epsilon)) - L(\mathbf{w}) = \mathbb{E}\, L(\mathbf{w} + \epsilon) - L(\mathbf{w}) = T_\sigma(\mathbf{w})$. ($\subseteq$) For the quadratic loss above, L is a polynomial of degree four in $\mathbf{w}$. Expanding $L(\mathbf{w} + \epsilon)$ in ϵ and using that odd Gaussian moments vanish and $\mathbb{E}[\epsilon^\top H\epsilon] = \sigma^2 \operatorname{Tr} H$, $T_\sigma(\mathbf{w}) = \frac{\sigma^2}{2}\operatorname{Tr}\nabla^2 L(\mathbf{w}) + \sigma^4 c(\mathbf{w})$ for some polynomial c. If $T_\sigma(C \cdot \mathbf{w}) = T_\sigma(\mathbf{w})$ for all σ in an interval, the coefficients of this polynomial in σ^2 agree, in particular the Hessian traces agree, and we conclude as before.

Item 3. ($\supseteq$) If $C \in \mathrm{SP}_r$, then $C^{-\top} = C$, and left multiplication by C permutes the rows of a matrix and flips some of their signs. This preserves every entrywise ℓ_q norm of both gradient blocks, so the target is invariant for every model and data distribution. ($\subseteq$) With $f = f_z$ and $\mu = \delta_x$, and using that entrywise norms factor on rank-one matrices, $\|zn^\top\|_q = \|z\|_q\|n\|_q$, the target equals $\|z\|_q\|n\|_q$ at $\mathbf{w}$ and $\|C^{-\top}z\|_q\|n\|_q$ at $C \cdot \mathbf{w}$. Hence $C^{-\top}$ is an ℓ_q isometry of $\mathbb{R}^r$, thus a signed permutation matrix by Theorem F.3, and so is C.

Item 4. Write $\mathbf{w}_0 = (A_0, B_0)$, $\rho = \rho(C)$ and $T(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}_0\|_2^2$. ($\subseteq$) Suppose $T(\rho\mathbf{w}) = T(\mathbf{w})$ for all $\mathbf{w}$. Replacing $\mathbf{w}$ by $t\mathbf{w}$ with $t \in \mathbb{R}$ gives $t^2\|\rho\mathbf{w}\|^2 - 2t\langle\rho\mathbf{w}, \mathbf{w}_0\rangle = t^2\|\mathbf{w}\|^2 - 2t\langle\mathbf{w}, \mathbf{w}_0\rangle$ for all t, so $\|\rho\mathbf{w}\| = \|\mathbf{w}\|$ and $\langle\mathbf{w}, \rho^\top\mathbf{w}_0\rangle = \langle\mathbf{w}, \mathbf{w}_0\rangle$ for all $\mathbf{w}$. The first identity makes ρ orthogonal (case $q = 2$ of Theorem F.3), the second gives $\rho^\top\mathbf{w}_0 = \mathbf{w}_0$, and together $\rho\mathbf{w}_0 = \rho\rho^\top\mathbf{w}_0 = \mathbf{w}_0$. By Theorem F.1, $C \in \mathrm{O}(r)$, and the A-block of $\rho\mathbf{w}_0 = \mathbf{w}_0$ reads $CA_0 = A_0$. Since A_0 has full row rank, $A_0A_0^+ = I_r$, and $C = CA_0A_0^+ = A_0A_0^+ = I_r$. ($\supseteq$) The identity preserves every target.

Strictness. For $r \geq 2$: $\mathrm{diag}(2, 1, \ldots, 1) \in \mathrm{GL}_r \setminus \mathrm{O}(r)$; a rotation by $\pi/4$ in the first two coordinates lies in $\mathrm{O}(r) \setminus \mathrm{SP}_r$; and $-I_r \in \mathrm{SP}_r \setminus \{I_r\}$. (For $r = 1$, $\mathrm{O}(1) = \mathrm{SP}_1 = \{\pm 1\}$.) □

G Proofs for Section 5

Proof of Theorem 5.1. Write $M_g = \rho(g)$. Invariance of L and Theorem E.1 give $\nabla L(M_g\mathbf{w}) = M_g^{-\top}\nabla L(\mathbf{w})$ for all $\mathbf{w}$.

(a) Let $\mathbf{w}_t$ be the iterates from $\mathbf{w}_0 = \mathbf{w}$ and $\mathbf{u}_t$ those from $\mathbf{u}_0 = M_g\mathbf{w}$. Assume $\rho(g)M\rho(g)^\top = M$, which is equivalent to $MM_g^{-\top} = M_gM$. We show $\mathbf{u}_t = M_g\mathbf{w}_t$ by induction. The case $t = 0$ holds by definition. If $\mathbf{u}_t = M_g\mathbf{w}_t$, then

$$\mathbf{u}_{t+1} = M_g\mathbf{w}_t - \eta M\nabla L(M_g\mathbf{w}_t) = M_g\mathbf{w}_t - \eta MM_g^{-\top}\nabla L(\mathbf{w}_t) = M_g\mathbf{w}_t - \eta M_gM\nabla L(\mathbf{w}_t) = M_g\mathbf{w}_{t+1}.$$

For $M = I$ the condition reads $M_gM_g^\top = I$.

(b) Let $M_g = S$ be a signed permutation matrix, $Se_j = s_je_{\pi(j)}$. Then $S^{-\top} = S$, so $\nabla L(S\mathbf{w}) = S\nabla L(\mathbf{w})$. The coordinatewise sign function satisfies $\mathrm{sign}(Sv) = S\,\mathrm{sign}(v)$, since $(Sv)_{\pi(j)} = s_jv_j$ and $\mathrm{sign}(s_jv_j) = s_j\,\mathrm{sign}(v_j)$ (with $\mathrm{sign}(0) = 0$). The induction of part (a) therefore applies to sign descent.

For Adam, write the iteration from $\mathbf{w}_0$ as $m_0 = v_0 = 0$ and, with $\mathrm{g}_t = \nabla L(\mathbf{w}_{t-1})$, $m_t = \beta_1 m_{t-1} + (1-\beta_1)\mathrm{g}_t$, $v_t = \beta_2 v_{t-1} + (1-\beta_2)\mathrm{g}_t \odot \mathrm{g}_t$, $\hat{m}_t = m_t/(1-\beta_1^t)$, $\hat{v}_t = v_t/(1-\beta_2^t)$, and $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta\big(\hat{m}_t/(\sqrt{\hat{v}_t}+\epsilon) + \lambda \mathbf{w}_{t-1}\big)$, where $\lambda \geq 0$ is decoupled weight decay and all operations on vectors are coordinatewise. Let $|S|$ be the permutation matrix with $|S|e_j = e_{\pi(j)}$. We claim that the iteration started at $S\mathbf{w}_0$ has states $(S\mathbf{w}_t, Sm_t, |S|v_t)$. This holds at $t=0$. If it holds at $t-1$, the new gradient is $S\mathrm{g}_t$; hence the first moment is Sm_t by linearity, and the second moment is $|S|v_t$ because $(S\mathrm{g}) \odot (S\mathrm{g}) = |S|(\mathrm{g} \odot \mathrm{g})$ (signs square away) and $|S|$ is linear. The update direction is $S\hat{m}_t/(\sqrt{|S|\hat{v}_t}+\epsilon)$, whose $\pi(j)$-th coordinate is $s_j \hat{m}_{t,j}/(\sqrt{\hat{v}_{t,j}}+\epsilon)$, i.e., it equals $S(\hat{m}_t/(\sqrt{\hat{v}_t}+\epsilon))$. The weight decay term is $\lambda S\mathbf{w}_{t-1}$. Hence the new iterate is $S\mathbf{w}_t$.

Finally, if $\mathbf{w}_t(g\mathbf{w}) = g\mathbf{w}_t(\mathbf{w})$, then $f(\cdot; \mathbf{w}_t(g\mathbf{w})) = f(\cdot; g\mathbf{w}_t(\mathbf{w})) = f(\cdot; \mathbf{w}_t(\mathbf{w}))$, so any $\Psi(f(\cdot; \mathbf{w}_t(\cdot)))$ is invariant. □

First-order expansions. For fixed $\mathbf{w}$ and a fixed direction s, the function $\eta \mapsto L(\mathbf{w} - \eta s)$ is differentiable with derivative $-\nabla L(\mathbf{w})^\top s$ at $\eta = 0$. For $s = M\nabla L(\mathbf{w})$ this is $-\nabla L^\top M \nabla L = -\|\nabla L\|_M^2$, and for $s = \operatorname{sign} \nabla L(\mathbf{w})$ it is $-\sum_i \partial_i L \operatorname{sign}(\partial_i L) = -\sum_i |\partial_i L| = -\|\nabla L\|_1$. For Adam's first step with $\epsilon = 0$, $\hat{m}_1 = \mathrm{g}_1$ and $\hat{v}_1 = \mathrm{g}_1 \odot \mathrm{g}_1$, so the direction is $\mathrm{g}_1/|\mathrm{g}_1| = \operatorname{sign} \mathrm{g}_1$ on coordinates with $\mathrm{g}_{1,i} \neq 0$.

Proof of Theorem 5.2. ($\supseteq$) For $C \in \mathrm{O}(r)$, $\rho(C)$ is orthogonal (Theorem F.1), so one step of gradient descent is equivariant by Theorem 5.1(a): $\mathbf{w}_1(C \cdot \mathbf{w}) = C \cdot \mathbf{w}_1(\mathbf{w})$, and therefore $T_\eta^{\mathrm{GD}}(C \cdot \mathbf{w}) = L(C \cdot \mathbf{w}_1(\mathbf{w})) = L(\mathbf{w}_1(\mathbf{w})) = T_\eta^{\mathrm{GD}}(\mathbf{w})$ for every η. For $C \in \mathrm{SP}_r$, $\rho(C)$ is a signed permutation matrix (Theorem F.1) and the same argument applies to sign descent via Theorem 5.1(b).

($\subseteq$) Suppose $T_\eta(C \cdot \mathbf{w}) = T_\eta(\mathbf{w})$ for all $\eta \in (0, \eta_0)$, all $\mathbf{w}$, and all admissible losses. Since $T_0 = L$ is invariant, $(T_\eta(\cdot) - L(\cdot))/\eta$ is also invariant, and letting $\eta \downarrow 0$ and using the first-order expansions gives $\|\nabla L(C \cdot \mathbf{w})\|_2^2 = \|\nabla L(\mathbf{w})\|_2^2$ for gradient descent and $\|\nabla L(C \cdot \mathbf{w})\|_1 = \|\nabla L(\mathbf{w})\|_1$ for sign descent, for all admissible L. Admissible losses $L(\mathbf{w}) = \ell(\mathbf{W}_0 + A^\top B)$ include the linear losses $\ell(\mathbf{W}) = \langle \Gamma, \mathbf{W} \rangle$ of Theorem F.2, so these are exactly the Euclidean-sensitivity and ℓ_1-gradient targets of Items 2 and 3 of Theorem 4.3 with $\mu = \delta_x$. The necessity parts of those items give $C \in \mathrm{O}(r)$ and $C \in \mathrm{SP}_r$, respectively. □

H Proofs for Section 6

Proof of Theorem 6.1. ($\supseteq$) If $\rho(g)\mathbf{w}_0 = \mathbf{w}_0$, then for every $\mathbf{w}$, $T(g\mathbf{w}) = N(\rho(g)\mathbf{w} - \mathbf{w}_0) = N(\rho(g)(\mathbf{w} - \mathbf{w}_0)) = N(\mathbf{w} - \mathbf{w}_0) = T(\mathbf{w})$, because $\rho(g)$ is an isometry of N. ($\subseteq$) If $g \in G_T$, evaluate the invariance at $\mathbf{w} = \mathbf{w}_0$: $N(\rho(g)\mathbf{w}_0 - \mathbf{w}_0) = T(g\mathbf{w}_0) = T(\mathbf{w}_0) = N(0) = 0$. Since N is a norm, $\rho(g)\mathbf{w}_0 = \mathbf{w}_0$. For neuron permutations $\rho(P)$ is a permutation matrix, which preserves every ℓ_q norm. □

Proof of Theorem 6.2. *Generic statement.* For $g \neq e$, faithfulness gives $\rho(g) \neq I$, so the fixed-point set $\mathrm{Fix}(g) = \ker(\rho(g) - I)$ is a proper linear subspace of $\mathbb{R}^d$ and has Lebesgue measure zero. The set $\mathcal{S} = \bigcup_{g \neq e} \mathrm{Fix}(g)$ is a finite union of null sets, hence null. For $\mathbf{w}_0 \notin \mathcal{S}$, no $g \neq e$ fixes $\mathbf{w}_0$, i.e., $\mathrm{Stab}_G(\mathbf{w}_0) = \{e\}$.

MLP characterization. Write $\mathbf{w}_0 = (\beta_1, \ldots, \beta_h)$ with $\beta_i \in \mathbb{R}^m$ the block of neuron i (its incoming weights and outgoing weight). Then $\rho(P)\mathbf{w}_0$ has block $\beta_{P^{-1}(i)}$ in position i, so $\rho(P)\mathbf{w}_0 = \mathbf{w}_0$ iff $\beta_{P^{-1}(i)} = \beta_i$ for all i. If the blocks are pairwise distinct, this forces $P^{-1}(i) = i$ for all i, i.e., $P = I$. If $\beta_i = \beta_j$ for some $i \neq j$, the transposition $(i\,j)$ fixes $\mathbf{w}_0$ and the stabilizer is nontrivial. □

Proof of Theorem A.4. Let $\Pi := \int_G \rho(g)^{\otimes k} \, \mathrm{d}\nu(g)$, an operator on $(\mathbb{R}^d)^{\otimes k}$. Since $\rho^{\otimes k}$ is a representation and ν is invariant, $\rho(h)^{\otimes k}\Pi = \int_G \rho(hg)^{\otimes k} \, \mathrm{d}\nu(g) = \Pi$ for every h; hence every vector in the range of Π is invariant, and $\Pi^2 = \int_G \rho(h)^{\otimes k}\Pi \, \mathrm{d}\nu(h) = \Pi$. If $a \in A_G^{(k)}$, then $\Pi a = \int_G a \, \mathrm{d}\nu = a$. Thus Π is an idempotent with range exactly $A_G^{(k)}$. The trace of an idempotent equals its rank, so $\dim A_G^{(k)} = \operatorname{Tr} \Pi = \int_G \operatorname{Tr}(\rho(g)^{\otimes k}) \, \mathrm{d}\nu(g) = \int_G (\operatorname{Tr} \rho(g))^k \, \mathrm{d}\nu(g)$, using linearity of the trace and $\operatorname{Tr}(M^{\otimes k}) = (\operatorname{Tr} M)^k$. □

This exact identity replaces the trace-ratio bound of an earlier version of this work, which assumed $|\operatorname{Tr}\rho(g)| \le cd$ for all nontrivial g with a constant $c < 1$. That assumption fails for every positive-dimensional group, since $\operatorname{Tr}\rho(g) \to d$ as $g \to e$, and the bound was loose even for finite groups.

Proof of Theorem A.5. Reduction. Order the coordinates of $\mathbf{w}$ by neuron blocks, so that $\mathbb{R}^d = \mathbb{R}^h \otimes \mathbb{R}^m$ and $\rho(P) = P \otimes I_m$. Reordering tensor factors gives an isomorphism $(\mathbb{R}^h \otimes \mathbb{R}^m)^{\otimes k} \cong (\mathbb{R}^h)^{\otimes k} \otimes (\mathbb{R}^m)^{\otimes k}$ under which $\rho(P)^{\otimes k}$ becomes $P^{\otimes k} \otimes I$. A vector is fixed by all $P^{\otimes k} \otimes I$ iff it lies in $((\mathbb{R}^h)^{\otimes k})^{S_h} \otimes (\mathbb{R}^m)^{\otimes k}$ (expand in a basis of the second factor), so $\dim A^{(k)}_{S_h} = m^k \dim((\mathbb{R}^h)^{\otimes k})^{S_h}$.

Orbit counting. $P^{\otimes k}$ maps the basis vector $e_{i_1} \otimes \cdots \otimes e_{i_k}$ to $e_{P(i_1)} \otimes \cdots \otimes e_{P(i_k)}$, so it permutes the basis indexed by $[h]^k$. For a group permuting a basis, the invariant subspace is spanned by the orbit sums, which are linearly independent (they have disjoint supports); its dimension is the number of orbits. To a multi-index $(i_1, \ldots, i_k)$ associate the set partition of $[k]$ with $s \sim t \Leftrightarrow i_s = i_t$. Permuting values does not change this partition. Conversely, two multi-indices with the same partition, say with j blocks, take j distinct values each, and any bijection between these value sets extends to a permutation of $[h]$ mapping one multi-index to the other. Hence orbits correspond bijectively to set partitions of $[k]$ with at most h blocks (a partition with j blocks is realized iff $j \le h$), and there are $\sum_{j=1}^{\min(h,k)} S(k,j)$ of them. For $h \ge k$ this is $\operatorname{Bell}(k)$, and $\dim A^{(k)}_{S_h}/d^k = m^k \operatorname{Bell}(k)/(hm)^k = \operatorname{Bell}(k)/h^k$. As a check, Theorem A.4 with $\operatorname{Tr}\rho(P) = m \cdot \#\mathrm{fix}(P)$ gives the same number, since $\mathbb{E}_P[\#\mathrm{fix}(P)^k] = \sum_{j \le \min(h,k)} S(k,j)$ for a uniform permutation.

Generic linear targets. By the argument in the proof of Theorem 6.2, applied to the finite group S_h acting on $(\mathbb{R}^d)^{\otimes k}$ (or on $\operatorname{Sym}^k(\mathbb{R}^d)$, which is invariant) through $\rho^{\otimes k}$, it suffices that this action is faithful. If $P \ne I$, pick i with $P(i) \ne i$ and any $c \in [m]$; the symmetric tensor $(e_i \otimes e_c)^{\otimes k}$ is mapped to $(e_{P(i)} \otimes e_c)^{\otimes k} \ne (e_i \otimes e_c)^{\otimes k}$, so $\rho(P)^{\otimes k}$ is not the identity on Sym^k (and hence not on the ambient space). □

Symmetric tensors. Since derivative tensors are symmetric, one may prefer to count invariant targets inside $\operatorname{Sym}^k(\mathbb{R}^d)$. The same decay holds. Because $\operatorname{Sym}^k(\mathbb{R}^d)$ is a $\rho^{\otimes k}$-invariant subspace, its invariant vectors are among those of the ambient space, so $\dim \operatorname{Sym}^k(\mathbb{R}^d)^{S_h} \le m^k \operatorname{Bell}(k)$, while $\dim \operatorname{Sym}^k(\mathbb{R}^d) = \binom{d+k-1}{k} \ge d^k/k!$. Therefore

$$\frac{\dim \operatorname{Sym}^k(\mathbb{R}^d)^{S_h}}{\dim \operatorname{Sym}^k(\mathbb{R}^d)} \le \frac{k!\,\operatorname{Bell}(k)}{h^k} \qquad (h \ge k).$$

For $k = 2$ the count is exact: the S_h-invariants of $(\mathbb{R}^h)^{\otimes 2}$ are spanned by $\delta = \sum_i e_i \otimes e_i$ and $\mathbf{1} \otimes \mathbf{1} - \delta$, both symmetric, and the swap of the two tensor slots acts on $(\mathbb{R}^h)^{\otimes 2} \otimes (\mathbb{R}^m)^{\otimes 2}$ as the swap on each factor. Hence the invariant symmetric tensors are $\operatorname{span}\{\delta, \mathbf{1} \otimes \mathbf{1} - \delta\} \otimes \operatorname{Sym}^2(\mathbb{R}^m)$, of dimension $m(m+1)$, out of $\dim \operatorname{Sym}^2(\mathbb{R}^d) = hm(hm+1)/2$, a fraction of about $2/h^2$. Hidden biases can be included by taking $m = p + q + 1$.

I Extended Related Work

Parameter symmetries and loss landscapes. Parameter spaces of neural networks carry permutation, scaling, and other symmetries (Brea et al., 2019; Simsek et al., 2021; Godfrey et al., 2022; Ziyin et al., 2025), which shape the geometry of the loss landscape, including permutation saddles, mode connectivity, and visualizations of loss surfaces (Garipov et al., 2018; Li et al., 2018; Mehta et al., 2021; Frankle et al., 2020). Symmetries of equivariant models themselves have also been analyzed (Xie & Smidt, 2025). Permutation symmetry explains much of the barrier between independently trained networks (Entezari et al., 2022), and aligning permutations enables model merging (Ainsworth et al., 2022); these alignment targets are set-level targets in our terminology, whose relevant group is the diagonal of G^n.

Symmetry in optimization and training dynamics. Optimizers that respect or exploit symmetries include path-normalized SGD (Neyshabur et al., 2015), optimization in the positively scale-invariant space of ReLU networks (Meng et al., 2018), symmetry teleportation (Zhao et al., 2022; 2023),

and scale-invariant LoRA optimizers (Yen et al., 2024; Li et al., 2024). Continuous symmetries induce conservation laws of gradient flow that are broken by discretization and regularization (Kunin et al., 2021), and homogeneous models are automatically balanced (Du et al., 2018). Adaptive and sign-based methods (Duchi et al., 2011; Hinton et al., 2012; Kingma & Ba, 2015; Bernstein et al., 2018; Balles & Hennig, 2018) and curvature-based preconditioners (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2024) change the metric and hence, by Theorem 5.1, the symmetry group of trajectory-dependent targets. Sharpness-aware minimization (Foret et al., 2020) and the NTK (Jacot et al., 2018) are further examples of parameter-space quantities that depend on the parameterization.

Learning to optimize and meta-learning. Learned optimizers and meta-learning methods operate on weights or gradients (Andrychowicz et al., 2016; Finn et al., 2017; Li et al., 2017; Nichol et al., 2018; Rusu et al., 2018; Rajeswaran et al., 2019; Shen et al., 2020; Metz et al., 2021; Gao et al., 2022; Flennerhag et al., 2023; Thérien et al., 2024), as do hypernetworks (Ha et al., 2016). Parameter symmetries have been used to improve their generalization (Zamir et al., 2025; Yang et al., 2023).

Weight generation and representation. Beyond prediction, weight-space models are used to generate weights (Schürholt et al., 2022a; Soro et al., 2024), to learn scalable weight representations (Schürholt et al., 2024), to accelerate training (Knyazev et al., 2024), and to probe networks (Kahana et al., 2024). Set-structured encoders (Zaheer et al., 2017) underlie many of these architectures and our signed-permutation-invariant meta-models.

LoRA. LoRA (Hu et al., 2021) has many extensions, including merging for skill composition (Prabhakar et al., 2024), quantized fine-tuning (Dettmers et al., 2024), principal-subspace initialization (Meng et al., 2024), extremely small adapters (Bałazy et al., 2024), analyses of expressive power (Zeng & Lee, 2023), kernel views of fine-tuning (Malladi et al., 2023), and methods that close the gap to full fine-tuning (Hayou et al., 2024; Wang et al., 2024a;b). Several of these address the fact that gradient descent is not equivariant under the GL_r gauge, which Theorem 5.2 formalizes.

Quantization, pruning, and privacy. Rescaling symmetries are exploited to equalize weight ranges before quantization (Nagel et al., 2019), a choice of representative within an orbit much like our balanced factorization, and second-order pruning uses entrywise curvature (LeCun et al., 1989). On the privacy side, training data can be reconstructed from weights (Haim et al., 2022), and for group-invariant networks reconstructions converge to functionally equivalent samples (Elbaz et al., 2024); data attribution methods (Ilyas et al., 2022; Engstrom et al., 2024) ask related questions about what weights encode about training data.

J Experimental Details

All synthetic experiments run on a CPU in a few minutes; code will be released. Meta-models are trained with AdamW (learning rate 10^{-3}, weight decay 10^{-5}, batch size 256, cosine schedule, at most 300 epochs) with early stopping on a validation split, using a $70/15/15$ train/validation/test split and standardized inputs and targets. MLPs have two hidden layers of width 256 with GELU activations; set networks use a two-layer encoder of width 128, sum pooling, and a two-layer decoder of width 256. We report a single training run per configuration.

LoRA ladder (Section 7.1). The base network is $f(x;\mathbf{W}) = u^\top \tanh(\mathbf{W}x)$ with $p = 8$, $\mathbf{W}_0, u$ Gaussian with variance $1/p$, and $J = 16$ fixed probe inputs with labels from a fixed teacher. Adapters have rank $r = 3$: $A' = A_0 + 0.3\,E_A/\sqrt{p}$ with a shared Gaussian A_0, and $B' = s\,E_B/\sqrt{p}$ with $\log s \sim \mathcal{N}(-0.5, 0.5^2)$ and Gaussian E_A, E_B. Each adapter is stored as $(CA', C^{-\top}B')$ with $C = Q_1 D Q_2$, Q_1, Q_2 Haar on $\mathrm{O}(3)$ and $D = \mathrm{diag}(e^{0.5z})$, $z \sim \mathcal{N}(0, I)$. The targets are the logarithms of the probe loss, of the Euclidean sensitivity $\frac{1}{J}\sum_j \|\nabla_{\mathbf{w}} f(x_j;\mathbf{w})\|_2^2$, of $\|\nabla_{\mathbf{w}} L\|_1$, of $\|\nabla_{\mathbf{w}} L\|_1/\|\nabla_{\mathbf{w}} L\|_2$, and of $\|A - A_0\|_F^2 + \|B\|_F^2$. The function-level context given to all meta-models consists of $\mathbf{W}_0 + A^\top B$, $\nabla_{\mathbf{W}} L$, and the two matrices $\frac{1}{J}\sum_j \Gamma_j^\top \Gamma_j$ and $\frac{1}{J}\sum_j \Gamma_j \Gamma_j^\top$ with $\Gamma_j = \partial f(x_j)/\partial \mathbf{W}$; all are functions of $A^\top B$. The SP_r-invariant network computes $\sum_i [\phi(a_i, b_i, c) + \phi(-a_i, -b_i, c)]$ for the context c, followed by a decoder, which is exactly invariant under permutations and sign flips of the rank-one components. Ceilings are estimated on 4,000 fresh adapters with 128 group samples each (exact enumeration of the 48 elements is replaced by uniform sampling for SP_3);

for GL_r we resample the gauge from its prior, which gives the Bayes-optimal GL_r-invariant error for this zoo. Table 4 lists all numbers.

Table 4: Test R^2 for Figure 2; ceilings $1 - \rho_H(T)$ in parentheses.

Meta-model	Loss	Sensitivity	Grad. ℓ_1	Grad. sparsity	Dist. from init
Unconstrained	0.979	0.957	0.876	0.306	0.989
SP_r-invariant	0.983 (1.000)	0.985 (1.000)	0.927 (1.000)	0.568 (1.000)	0.353 (0.416)
$\mathrm{O}(r)$-invariant	0.979 (1.000)	0.981 (1.000)	0.910 (0.980)	0.408 (0.666)	0.364 (0.417)
GL_r-invariant	0.983 (1.000)	0.185 (0.245)	0.572 (0.674)	0.187 (0.356)	-0.014 (0.017)

Permutations (Section 7.2). The MLP is $f(x) = h^{-1/2} \sum_i v_i \tanh(u_i^\top x/\sqrt{p})$ with $h = 12$ hidden units and $p = 6$, so each neuron has a block $(u_i, v_i) \in \mathbb{R}^7$. The base $\mathbf{w}_0$ has standard Gaussian blocks. Models are $\mathbf{w} = \mathbf{w}_0 + s\,E/\sqrt{7}$ with $\log s \sim \mathcal{N}(-1, 0.6^2)$, and a fraction π are stored with a uniformly random permutation of their neurons. Targets are the logarithms of $\|\mathbf{w} - \mathbf{w}_0\|_2^2$, of the loss on 32 probe inputs with teacher labels, and of the Euclidean sensitivity. The ceiling uses the fact that, given the orbit, the stored representative is the unpermuted one with probability $1 - \pi$ and a uniformly permuted one with probability π; we estimate the conditional variance with 256 random permutations on 4,000 fresh models.

Table 5: Test R^2 for Figure 3: unconstrained (U) and S_h-invariant (I) meta-models, with the ceiling of the invariant model in parentheses.

	Drift from base		Loss		Sensitivity	
π	U	I	U	I	U	I
0	0.934	0.976 (1.000)	0.927	0.952 (1.000)	0.938	0.976 (1.000)
0.25	0.987	0.112 (0.199)	0.687	0.955 (1.000)	0.611	0.971 (1.000)
0.5	0.988	0.047 (0.061)	0.244	0.967 (1.000)	0.331	0.972 (1.000)
1	0.979	0.072 (0.107)	-0.025	0.962 (1.000)	0.132	0.978 (1.000)

Optimizers (Section 7.3). A network $x \mapsto W_2 \tanh(W_1 x)$ with $W_1 \in \mathbb{R}^{64\times 64}$ is trained with Adam on the digits 0–4 of the scikit-learn digits dataset (1,300 training and 497 test images, 8×8 pixels) and frozen. A rank-4 LoRA adapter on W_1 is then trained on all ten classes with cross-entropy loss, from Gaussian A_0 and $B_0 = 0$, for 100 Adam steps (learning rate 3×10^{-3}) to obtain $\mathbf{w}$. From $\mathbf{w}$ and from $C \cdot \mathbf{w}$ we run 200 full-batch steps of SGD (learning rate 0.5) or Adam (learning rate 3×10^{-3}, default β and ϵ) in float64. Orthogonal matrices are Haar-distributed, signed permutations uniform, and generic GL_r elements are $Q_1 \operatorname{diag}(e^{0.5z}) Q_2$.

K Additional Experiments

K.1 GPT-2 fine-tuned on Wikitext2

We repeat the perturbation experiment of Section 7 for GPT-2 fine-tuned with LoRA on Wikitext2, with the same protocol.

K.2 Hessian trace under reparameterization

K.3 Loss landscape along a rescaling orbit

Model \ σ	Unpert.	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	$3 \cdot 10^{-1}$	$5 \cdot 10^{-1}$
Baseline	**3.2777**	**3.2777**	**3.2777**	**3.2777**	**3.2777**	**3.2777**	**3.2787**	3.4709	8.3331	9.8147
Balanced $P_i^\star$		**3.2777**	**3.2777**	**3.2777**	**3.2777**	**3.2777**	**3.2787**	**3.3746**	**7.0786**	**8.6564**
Scaled $s = 10^2$		**3.2777**	**3.2777**	**3.2777**	3.2794	3.4713	15.6909	21.5531	24.7776	21.1475
Scaled $s = 10^3$		**3.2777**	3.2778	3.2793	3.4894	11.4180	21.1974	23.1106	24.4113	21.1934
Scaled $s = 10^4$		**3.2777**	3.2803	3.4918	12.8537	18.3714	21.7953	23.3156	24.3576	21.1724
Scaled $s = 10^5$		3.2787	3.5227	10.6500	18.5895	19.4165	21.8154	23.3321	24.3471	21.1932

Table 6: Averaged test loss under perturbations with standard deviation σ for different LoRA factorizations of the same GPT-2 model. Lower is better.

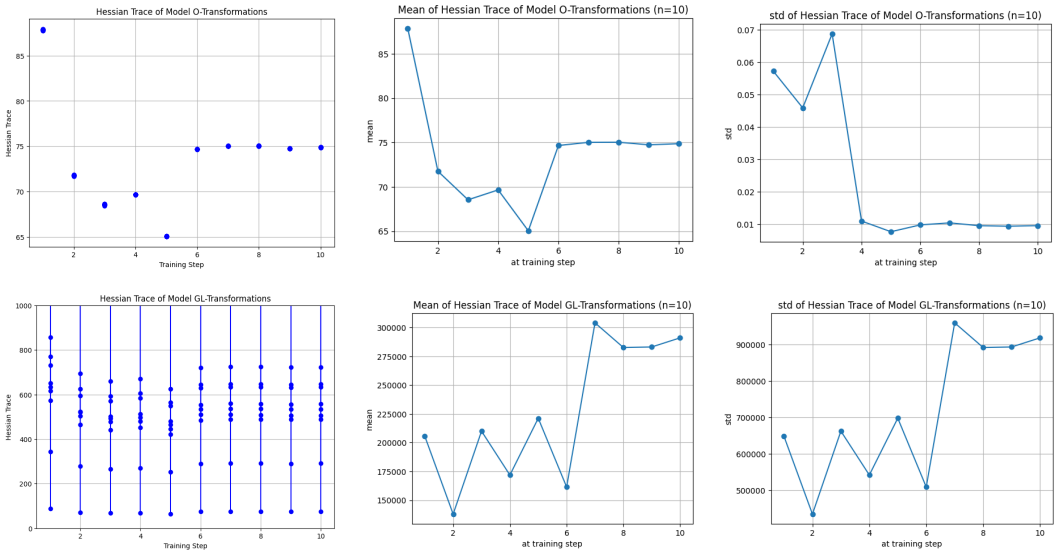


Figure 5: Hessian trace under ten Haar-random $\mathrm{O}(r)$ (top) and ten generic GL_r (bottom) reparameterizations: all draws, mean, and standard deviation per epoch. Some GL_r values fall outside the plotted range.



(a) Unperturbed. (b) Scaled with $c = 5$. (c) Scaled with $c = 10$.

Figure 6: Loss landscape of the same LoRA model in three factorizations related by scalar rescaling. The function is identical; the landscape in parameter coordinates is not.