SYMMETRIES, FLAT MINIMA AND THE CONSERVED QUANTITIES OF GRADIENT FLOW

Bo Zhao[*†]
University of California, San Diego
bozhao@ucsd.edu

Iordan Ganev[*]
Radboud University
iganev@cs.ru.nl

Robin Walters
Northeastern University
r.walters@northeastern.edu

Rose Yu
University of California, San Diego
roseyu@ucsd.edu

Nima Dehmamy
IBM Research
nima.dehmamy@ibm.com

ABSTRACT

Empirical studies of the loss landscape of deep networks have revealed that many local minima are connected through low-loss valleys. Ensemble models sampling different parts of a low-loss valley have reached SOTA performance. Yet, little is known about the theoretical origin of such valleys. We present a general framework for finding continuous symmetries in the parameter space, which carve out low-loss valleys. Our framework uses equivariances of the activation functions and can be applied to different layer architectures. We also introduce a novel set of nonlinear, data-dependent symmetries for neural networks. These symmetries can transform a trained model such that it performs similarly on new samples. We then show that conserved quantities associated with linear symmetries can be used to define coordinates along low-loss valleys. The conserved quantities help reveal that using common initialization methods, gradient flow only explores a small part of the global minimum. By relating conserved quantities to convergence rate and sharpness of the minimum, we provide insights on how initialization impacts convergence and generalizability. We also find the nonlinear action to be viable for ensemble building to improve robustness under certain adversarial attacks.

1 INTRODUCTION

Training deep neural networks (NNs) is generally a highly non-convex problem. The loss landscape of a NN, which is shaped by the model architecture and the dataset, is generally very rugged, with the number of local minima growing rapidly with model size (Bray & Dean, 2007; Şimşek et al., 2021). Despite this complexity, recent work has revealed many interesting structures in the loss landscape. For example, NN loss landscapes often contain approximately flat directions along which the loss does not change significantly (Freeman & Bruna, 2017; Garipov et al., 2018). Flat minima have been used to make ensemble or mixture models by sampling different parameter configurations that yield similar loss values (Garipov et al., 2018; Benton et al., 2021). However, finding such flat directions is mostly done empirically, with few theoretical results.

One source of flat directions is parameter transformations that keep the loss invariant (i.e. symmetries). Specifically, moving in the parameter space from a minimum in the direction of a symmetry takes us to another minimum. Motivated by the fact that continuous symmetries of the loss result in flat directions in local minima, we derive a general class of such symmetries in this paper.

[*]Equal contribution.
[†]Work done during an internship at IBM.

Our key insight is to focus on *equivariances of the nonlinear activation functions*; most known continuous symmetries can be derived using this framework. Models related by exact equivalence cannot behave differently on different inputs. Hence, for ensembling or robustness tasks, we need to find *data-dependent symmetries*. Indeed, aside from the familiar "linear symmetries" of NN, the framework of equivariance allows us to introduce a novel class of symmetries which act *nonlinearly* on the parameters and are data-dependent. These nonlinear symmetries cover a much larger class of continuous symmetries than their linear counterparts, as they apply for almost any activation function. We provide preliminary experimental evidence that ensembles using these nonlinear symmetries are more robust to attacks.

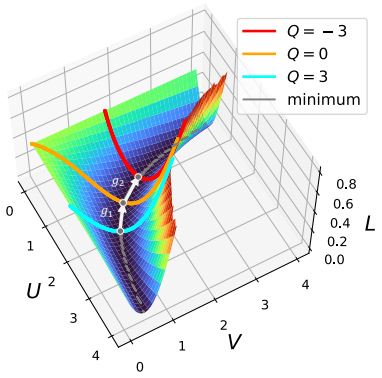

Figure 1: Visualization of the extended minimum in a 2-layer linear network with loss $\mathcal{L} = \|Y - UVX\|^2$. Points along the minima are related to each other by scaling symmetry $U \to Ug^{-1}$ and $V \to gV$. Conserved quantities, Q, associated with scaling symmetry parametrize points along the minimum.

Extended flat minima arise frequently in the loss landscape of NNs; we show that symmetry-induced flat minima can be parametrized using *conserved quantities*. Furthermore, we provide a method of deriving explicit conserved quantities (CQ) for different continuous symmetries of NN parameter spaces. CQ had previously been derived from symmetries for one-parameter groups (Kunin et al., 2021; Tanaka & Kunin, 2021). Using a similar approach we derive the CQ for general continuous symmetries. This approach *fails* to find CQ for rotational symmetries. Nevertheless, we find the conservation law resulting from the symmetry implies a *cancellation of angular momenta* between layers. To summarize, our contributions are:

1. A general framework based on **equivariance** for finding symmetries in NN loss landscapes.
2. A derivation of the **dimensions of minima** induced by symmetries.
3. A new class of **nonlinear, data-dependent symmetries** of NN parameter spaces.
4. An expansion of prior work on **deriving conserved quantities** (CQ) associated with symmetries, and a discussion of its failure for rotation symmetries.
5. A **cancellation of angular momenta** result for between layers for rotation symmetries.
6. A **parameterization** of symmetry-induced flat minima via the associated CQ.

This paper is organized as follows. First, we review existing literature on flat minima, continuous symmetries of parameter space, and conserved quantities. In Section 3, we define continuous symmetries and flat minima, and show how linear symmetries lead to extended minima. We illustrate our constructions through examples of linear symmetries of NN parameter spaces. In Section 4, we define nonlinear, data-dependent symmetries. In Section 5, we use infinitesimal symmetries to derive conserved quantities for parameter space symmetries, extending the results in Kunin et al. (2021) to larger groups and more activation functions. Additionally, we show how CQ can be used to define coordinates along flat minima. We close with experiments involving nonlinear symmetries, conserved quantities and a discussion of potential use cases.

2 Related Work

Continuous symmetry in parameter space Overparametrization in neural networks leads to symmetries in the parameter space (Głuch & Urbanke, 2021). Continuous symmetry has been identified in fully-connected linear networks (Tarmoun et al., 2021), homogeneous neural networks (Badrinarayanan et al., 2015; Du et al., 2018), radial neural networks (Ganev et al., 2022), and softmax and batchnorm functions (Kunin et al., 2021). We provide a unified framework that generalizes previous findings, and identify nonlinear group actions that have not been studied before.

Conserved quantities The imbalance between layers in linear or homogeneous networks is known to be invariant during gradient flow and related to convergence rate (Saxe et al., 2014; Du et al.,

2018; Arora et al., 2018a;b; Tarmoun et al., 2021; Min et al., 2021). Huh (2020) discovered similar conservation laws in natural gradient descents. Kunin et al. (2021) develop a more general approach for finding conserved quantities for certain one-parameter symmetry groups. Tanaka & Kunin (2021) relate continuous symmetries to dynamics of conserved quantities using an approach similar to Noether's theorem (Noether, 1918). We develop a procedure that determines conserved quantities from infinitesimal symmetries, which is closely related to Noether's theorem.

Topology of minimum The global minimum of overparametrized neural networks are connected spaces instead of isolated points. We show that parameter space symmetries lead to extended flat minima. Previously, Cooper (2018) proved that the global minima is usually a manifold with dimension equal to the number of parameters subtracted by the number of data points. We derive the dimensionality of the symmetry-induced flat minima and show they are related to the number of infinitesimal symmetry generators and dimension of weight matrices. Şimşek et al. (2021) study permutation symmetry and show that in certain overparametrized networks, the minimum related by permutations are connected. Entezari et al. (2022) hypothesize that SGD solutions can be permuted to points on the same connected minima. Ainsworth et al. (2023) develop algorithms that find such permutations. Additional discussion on mode connectivity, sharpness of minima, and the role of symmetry in optimization can be found in Appendix A.

3 Continuous Symmetries in Deep Learning

In this section, we first summarize our notation for basic neural network constructions (see Appendix C for more details). Then we consider transformations on the parameter space that leave the loss invariant and demonstrate how they lead to extended flat minima.

3.1 The Parameter Space and Loss Function

The parameters of a neural network consist of weights[1] $W_i \in \mathbb{R}^{n_i \times m_i}$ for each layer i, where n_i and m_i are the layer output and input dimensions, respectively. For feedforward networks, successive output and input dimensions match: $m_i = n_{i-1}$. We group the widths into a tuple $\mathbf{n} = (n_L, \ldots, n_1, n_0)$, and the parameter space becomes $\mathbf{Param} = \mathbb{R}^{n_L \times n_{i-1}} \times \cdots \times \mathbb{R}^{n_1 \times n_0}$. We denote an element therein as a tuple of matrices $\boldsymbol{\theta} = (W_i \in \mathbb{R}^{n_i \times n_{i-1}})_{i=1}^L$. The activation of the i-th layer is a piecewise differentiable function $\sigma_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$, which may or may not be pointwise. For $\boldsymbol{\theta} \in \mathbf{Param}$ and input $x \in \mathbb{R}^{n_0}$, the feature vector of the ith layer in feedforward network is $Z_{i+1}(x) = W_{i+1}\sigma(Z_i(x))$, where the juxtaposition '$W\sigma(Z)$' denotes an arbitrary linear operation depending on the context; for example, matrix product, convolution, etc. For simplicity, we largely focus on the case of multilayer perceptrons (MLPs). We denote the final output by $F_{\boldsymbol{\theta}} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$, defined as $F_{\boldsymbol{\theta}}(x) = \sigma_L(Z_L(x))$. The "loss function" $\mathcal{L}$ of our model is defined as:

$$\mathcal{L} : \mathbf{Param} \times \mathbf{Data} \to \mathbb{R}, \qquad \mathcal{L}(\boldsymbol{\theta}, (x, y)) = \mathrm{Cost}(y, F_{\boldsymbol{\theta}}(x)). \tag{1}$$

where $\mathbf{Data} = \mathbb{R}^{n_0} \times \mathbb{R}^{n_L}$ is the space of data and $\mathrm{Cost} : \mathbb{R}^{n_L} \times \mathbb{R}^{n_L} \to \mathbb{R}$ is a differentiable cost function, such as mean square error or cross-entropy. In the case of multiple samples, we have matrices $X \in \mathbb{R}^{n_0 \times k}$ and $Y \in \mathbb{R}^{n_L \times k}$ whose columns are the k samples[2], and retain the same notation for the feedforward function, namely, $F_{\boldsymbol{\theta}} : \mathbb{R}^{n_0 \times k} \to \mathbb{R}^{n_L \times k}$. Most of our results concern properties of $\mathcal{L}$ that hold for any training data. Hence, unless specified otherwise, we take a fixed batch of data $\{(x_i, y_i)\}_{i=1}^k \subseteq \mathbf{Data}$, and consider the loss as a function of the parameters only.

Example 3.1. Two-layer network with MSE Consider a network with $\mathbf{n} = (n, h, m)$, the identity output activation ($\sigma_L(x) = (x)$), and no biases. The parameter space is $\mathbf{Param}(\mathbf{n}) = \mathbb{R}^{n \times h} \times \mathbb{R}^{h \times m}$ and we denote an element as $\boldsymbol{\theta} = (U, V)$. Taking the mean square error cost function, the loss function for data $(X, Y) \in \mathbb{R}^{n \times k} \times \mathbb{R}^{m \times k}$ takes the form $\mathcal{L}(\boldsymbol{\theta}, (X, Y)) = \frac{1}{k}\|Y - U\sigma(VX)\|^2$.

3.2 Action of Continuous Groups and Flat Minima

Let G be a group. An action of G on the parameter space $\mathbf{Param}$ is a function $\cdot : G \times \mathbf{Param} \to \mathbf{Param}$, written as $g \cdot \boldsymbol{\theta}$, that satisfies the unit and multiplication axioms of the group, meaning $\mathrm{id} \cdot \boldsymbol{\theta} = \boldsymbol{\theta}$ where id is the identity of G, and $g_1 \cdot (g_2 \cdot \boldsymbol{\theta}) = (g_1 g_2) \cdot \boldsymbol{\theta}$ for all $g_1, g_2 \in G$.

[1] For clarity, we suppress the bias vectors; all results can be extended to include bias; see appendix C.

[2] We use capital letters for matrix data and small letters for individual samples.

Definition 3.1 (Parameter space symmetry)**.** *The action $G \times \mathbf{Param} \to \mathbf{Param}$ is a symmetry of $\mathcal{L}$ if it leaves the loss function invariant, that is:*

$$\mathcal{L}(g \cdot \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}), \qquad \forall \boldsymbol{\theta} \in \mathbf{Param}, \quad g \in G. \tag{2}$$

We describe examples of parameter space symmetries in the next section. Before doing so, we show how a parameter space symmetry leads to flat minima (see Appendix C.6):

Proposition 3.2. *Suppose $G \times \mathbf{Param} \to \mathbf{Param}$ is a symmetry of $\mathcal{L}$. If $\boldsymbol{\theta}^*$ is a critical point (resp. local minimum) of $\mathcal{L}$, then so is $g \cdot \boldsymbol{\theta}^*$ for any $g \in G$.*

The proof of this result relies on using the differential of the action of g to relate the gradient of $\mathcal{L}$ at θ^* with the gradient at $g \cdot \theta^*$. We see that, if $\boldsymbol{\theta}^*$ is a local minimum, then so is every element of the set $\{g \cdot \boldsymbol{\theta}^* \mid g \in G\}$. This set is known as the *orbit* of $\boldsymbol{\theta}^*$ under the action of G. The orbits of different parameter values may be of different dimensions. However, in many cases, there is a "generic" or most common dimension, which is the orbit dimension of any randomly chosen $\boldsymbol{\theta}$.

3.3 Equivariance of the activation function

In this section, we describe a large class of linear symmetries of $\mathcal{L}$ using an *equivariance* property of the activations between layers. For accessibility, we focus on the example of two layers with output $F(x) = U\sigma(Vx)$ for $(U, V) \in \mathbf{Param} = \mathbb{R}^{m\times h} \times \mathbb{R}^{h\times n}$ and $x \in \mathbb{R}^n$. All results generalize to multiple layers by letting $U = W_i$ and $V = W_{i-1}$ be weights of two successive layers in a deep neural network (see Appendix C.5). Let $G \subseteq \mathrm{GL}_h(\mathbb{R})$ be a subgroup of the general linear group, and let $\pi : G \to \mathrm{GL}_h(\mathbb{R})$ a representation (the simplest example is $\pi(g) = g$). We consider the following action of the group G on the parameter space $\mathbf{Param}$:

$$g \cdot U = U\pi(g^{-1}), \qquad g \cdot V = gV \tag{3}$$

This action becomes a symmetry of $\mathcal{L}$ if and only if the following identity holds:

$$\sigma(gz) = \pi(g)\sigma(z) \qquad \forall g \in G, \quad \forall z \in \mathbb{R}^h \tag{4}$$

We now turn our attention to examples. To ease notation, we write GL_h instead of $\mathrm{GL}_h(\mathbb{R})$.

Example 3.2. Linear networks A simple example of (4) is that of linear networks, where σ is the identity function: $\sigma(x) = x$. One can take $\pi(g) = g$ and $G = \mathrm{GL}_h$.

Example 3.3. Homogeneous activations Suppose the activation $\sigma : \mathbb{R}^h \to \mathbb{R}^h$ is *homogeneous*, meaning that (1) σ is applied pointwise in the standard basis and (2) there exists $\alpha > 0$ such that $\sigma(cz) = c^\alpha \sigma(z)$ for all $c \in \mathbb{R}_{>0}$ and $z \in \mathbb{R}^h$. Such an activation is equivariant under the *positive scaling group* $G \subset \mathrm{GL}_h$ consisting of diagonal matrices with positive diagonal entries. Explicitly, the group G consists of diagonal matrices $g = \mathrm{diag}(\mathbf{c})$ with $\mathbf{c} = (c_1, \ldots, c_h) \in \mathbb{R}^h_{>0}$. For $z = (z_1, \ldots, z_h) \in \mathbb{R}^h$ and $g \in G$, we have $\sigma(gz) = \sum_j \sigma(c_j z_j) = \sum_j c_j^\alpha \sigma(z_j) = g^\alpha \sigma(z)$. Hence, the equivariance equation is satisfied with $\pi(g) = g^\alpha$.

Example 3.4. LeakyReLU This is a special case of homogeneous activation, defined as $\sigma(z) = \max(z, 0) + s\min(z, 0)$, with $s \in \mathbb{R}_{\geq 0}$. We have $\alpha = 1$, and $\pi(g) = g$.

Example 3.5. Radial rescaling activations A less trivial example of continuous symmetries is the case of a radial rescaling activation (Ganev et al., 2022) where for $z \in \mathbb{R}^h$, we have $\sigma(z) = f(\|z\|)z$ for some function $f : \mathbb{R} \to \mathbb{R}$. Radial rescaling activations are equivariant under rotations of the input: for any orthogonal transformation $g \in O(h)$ (that is, $g^T g = I$) we have $\sigma(gz) = g\sigma(z)$ for all $z \in \mathbb{R}^h$. Indeed, $\sigma(gz) = f(\|gz\|)(gz) = g(f(\|z\|)z) = g\sigma(z)$, where we use the fact that $\|gz\| = z^T g^T gz = z^T z = \|z\|$ for $g \in O(h)$. Hence, (4) is satisfied with $\pi(g) = g$.

We arrive at our first novel result, whose proof appears in Appendix C.6.

Theorem 3.3. *The dimension of a generic orbit in $\mathbf{Param}$ under the appropriate symmetry group is given as follows. The cases are divided based on whether $h \leq \max(n, m)$ or not.*

		Orbit Dimension	
Activation	Symmetry Group	$h \leq \max(n,m)$	$h \geq \max(n,m)$
Identity	$\mathrm{GL}_h(\mathbb{R})$	h^2	$h(n+m) - nm$
Homogeneous	Positive rescaling	h	$\max(n,m)$
Radial rescaling	$O(h)$	$\binom{h}{2}$	$\binom{h}{2} - \binom{h-\max(m,n)}{2}$

As an aside, we note that a familiar example where (4) is satisfied involves the permutation of neurons. More precisely, suppose σ is pointwise and let G be the finite group of $h \times h$ permutation matrices. Then (4) holds with $\pi(g) = g$. However, the permutation group is finite (0-dimensional), and so does not imply the presence of flat minima.

3.4 INFINITESIMAL SYMMETRIES

Deriving conserved quantities from symmetries requires the infinitesimal versions of parameter space symmetries. Recall that any smooth action of a matrix Lie group $G \subseteq \mathrm{GL}_h$ induces an action of the infinitesimal generators of the group, i.e., elements of its Lie algebra. Concretely, let $\mathfrak{g} = \mathrm{Lie}(G) = T_I G$ be the Lie algebra, which can be identified with a certain subspace of matrices in $\mathfrak{gl}_h = \mathbb{R}^{h \times h}$. For every $M \in \mathfrak{g}$, we have an exponential map $\exp_M : \mathbb{R} \to G$ defined as $\exp_M(t) = \sum_{k=0}^{\infty} \frac{(tM)^k}{k!}$. If $\rho : G \to \mathrm{GL}_h$ is a (linear) representation, then the *infinitesimal action* is given by $d\rho : \mathfrak{g} \to \mathfrak{gl}_h$ by $d\rho(M) = \frac{d}{dt}\big|_0 \rho(\exp_M(t))$. In the case of the action appearing in (3), the corresponding infinitesimal action of the Lie algebra $\mathfrak{g}$ induced by (3) is given by:

$$M \cdot U = -U d\pi(M), \qquad M \cdot V = MV \tag{5}$$

More generally, suppose G acts linearly on parameter space (see Appendix C for non-linear versions). Set d to be the dimension of the parameter space[3], and make the identification $\mathbf{Param} \simeq \mathbb{R}^d$ by flattening matrices into column vectors. The general linear group $\mathrm{GL}(\mathbf{Param}) \simeq \mathrm{GL}_d(\mathbb{R})$ consists of all invertible linear transformations of $\mathbf{Param}$. Suppose G is a subgroup of $\mathrm{GL}(\mathbf{Param})$, so its Lie algebra $\mathfrak{g}$ is a Lie subalgebra of $\mathfrak{gl}_d = \mathbb{R}^{d \times d}$. For $M \in \mathfrak{g}$ and $\boldsymbol{\theta} \in \mathbf{Param}$, the *infinitesimal action* is given simply by matrix multiplication: $M \cdot \boldsymbol{\theta}$.

In the case of a parameter space symmetry, the invariance of $\mathcal{L}$ translates into the following orthogonality condition, where the inner product $\langle , \rangle : \mathbf{Param} \times \mathbf{Param} \to \mathbb{R}$ is calculated by contracting all indices, e.g. $\langle A, B \rangle = \sum_{ijk\ldots} A_{ijk\ldots} B_{ijk\ldots}$.

Proposition 3.4. *Let G be a matrix Lie group and a symmetry of $\mathcal{L}$. Then the gradient vector field is point-wise orthogonal to the action of any $M \in \mathfrak{g}$:*

$$\langle \nabla_{\boldsymbol{\theta}} \mathcal{L} \, , \, M \cdot \boldsymbol{\theta} \rangle = 0, \qquad \forall \boldsymbol{\theta} \in \mathbf{Param} \tag{6}$$

4 NONLINEAR DATA-DEPENDENT SYMMETRIES

For common activation functions, the equivariance $\sigma(gz) = \pi(g)\sigma(z)$ of (4) holds only for g belonging to a relatively small subgroup of GL_h. For ReLU, g must be in the positive scaling group, while for the usual sigmoid activation, the equation only holds for trivial $g = \mathrm{id}$. However, under certain conditions, it is possible to define a nonlinear action of the *full* GL_h which applies to many different activations. The subtlety of such an action is that it is data-dependent, which means that, for any $g \in \mathrm{GL}_h$, the transformation of the parameter space depends on the input data[4] x.

The nonlinear action. For any nonzero vector $z \in \mathbb{R}^h$, let $(r, \alpha_1, \ldots, \alpha_{h-1})$ be the spherical coordinates[5] of z, and define the following h by h matrix:

$$(R_z)_{ij} = \begin{cases} z_i \cos(\alpha_{j-1}) \left(\prod_{k=1}^{j-1} \sin(\alpha_k) \right)^{-1} & \text{if } j \leq i \text{ and } \prod_{k=1}^{i-1} \sin(\alpha_k) \neq 0 \\ -r \sin(\alpha_i) & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha_0 = 0$ by convention. We observe that R_z is the product of a rotation matrix and rescaling by $|z|$. Moreover, since $z \neq 0$, the first column of R_z is the unit vector $z/|z|$ and R_z has inverse given by $R_z^{-1} = \frac{1}{|z|^2} R_z^T$. Using these facts, one arrives at the following result, stated in the case of a two-layer neural network with notation from Section 3.3, and proven in Appendix D:

[3] In terms of the widths, we have $d = \sum_{i=1}^{L} n_i n_{i-1}$.

[4] That is, rather than being a map $\mathrm{GL}_h \times \mathbf{Param} \to \mathbf{Param}$ satisfying the group action axioms, a data-dependent action is a map $\mathrm{GL}_h \times (\mathbf{Param} \times \mathbb{R}^n) \to \mathbf{Param} \times \mathbb{R}^n$ satisfying the same axioms.

[5] Hence, $r = |z|$ is the norm, and the i-th coordinate of z is $z_i = r \cos(\alpha_i) \prod_{k=1}^{i-1} \sin(\alpha_k)$, where $\alpha_h = 0$.

Theorem 4.1. *Suppose $\sigma(z)$ is nonzero for any $z \in \mathbb{R}^h$. Then there is an action $\mathrm{GL}_h \times (\mathbf{Param} \times \mathbb{R}^n) \to \mathbf{Param} \times \mathbb{R}^n$ given by*

$$g \cdot (U, V, x) = (UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}\ ,\ gV\ ,\ x). \tag{7}$$

The evaluation of the feedforward function at x unchanged: $F_{(U,V)}(x) = F_{(UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}, gV)}(x)$.

We emphasize that a necessary and sufficient condition for the particular action of Theorem 4.1 to be well-defined is that $\sigma(z)$ be nonzero for any $z \in \mathbb{R}^h$; this is the case for usual sigmoid. Moreover, in Appendix D.2, we provide a generalization to the case where $\sigma(z)$ is only required to be nonzero for any *nonzero* $z \in \mathbb{R}^h$, a condition satisfied by hyperbolic tangent, leaky ReLU, and many other activations. The cost of such a generalization is a restriction to a 'non-degenerate locus' of $\mathbf{Param} \times \mathbb{R}^n$ where $Vx \neq 0$. Theorem 4.1 also generalizes to mutli-layer networks, as explained in Appendix D.3. We have the following explicit algorithm to compute the action of Theorem 4.1:

0. Input: weight matrices (U, V), input vector $x \in \mathbb{R}^n$, matrix $g \in \mathrm{GL}_h$.
1. Determine the spherical coordinates of $\sigma(Vx)$ and $\sigma(gVx)$, and construct the matrices $R_{\sigma(Vx)}$ and $R_{\sigma(gVx)}$.
2. Compute the inverse $R^{-1}_{\sigma(gVx)} = \frac{1}{|\sigma(gVx)|^2} R^T_{\sigma(gVx)}$.
3. Set $U' = UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}$ and $V' = gV$.
4. Output: the transformed weights (U', V'). The data $x \in \mathbb{R}^n$ remains unchanged.

Lipschitz bounds. Unlike the exact symmetries of Section 3, a data-dependent action may alter the loss in the *function space*. This is evident from (7): while the transformed and original feedforward functions have the same value at x, they will differ at other points. That is, if $\tilde{x} \in \mathbb{R}^h$ is an input value different from x, then $F_{(U,V)}(\tilde{x}) \neq F_{(UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}, gV)}(\tilde{x})$ in general.

However, the transformed feedforward function will differ from the original one in a controlled way. More precisely, when σ is Lipschitz continuous, we show that there is a bound on how much the Lipschitz bound of the feedforward changes after the nonlinear action. The relevance of such a bound originates in the fact that we expect the distance between data points to encode important information about shared features. To be more specific, fix weight matrices (U, V), which provide the feedforward function $F(\tilde{x}) = U\sigma(V\tilde{x})$. For any input vector $x \in \mathbb{R}^n$ and matrix $g \in \mathrm{GL}_h$, the transformed weight matrices $(UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}, gV)$ provide a new feedforward function given by:

$$F^{(g,x)}_{(U,V)} : \mathbb{R}^n \to \mathbb{R}^m \qquad F^{(g,x)}_{(U,V)}(\tilde{x}) = UR_{\sigma(Vx)}R^{-1}_{\sigma(gVx)}\sigma(gV\tilde{x}) \tag{8}$$

Proposition 4.2 (Lipschitz bounds from equivariance)**.** *Let σ be Lipschitz continuous with Lipschitz constant η. Then $F^{(g,x)}_{(U,V)}$ is Lipschitz continuous with bound $\eta\|U\|\|V\|\frac{|\sigma(Vx)|\|g\|}{|\sigma(gVx)|}$.*

In particular, the Lipschitz bound of the original feedforward function is $\eta\|U\|\|V\|$. Thus, if it happens that $|\sigma(Vx)|\|g\| < |\sigma(gVx)|$, then the Lipschitz bound decreases when transforming the parameters. Additionally, we observe that the nonlinear action does not disrupt latent distribution of data significantly. See Appendix D.5 for proof of Proposition 4.2, which relies on iterative applications of the Cauchy-Schwarz inequality, as well as the fact that $\|R_z^{\pm 1}\| = |z|^{\pm 1}$.

General equivariance. To provide perspective, we note that the action described above fits into a more general framework, which we now summarize. A map $c : \mathrm{GL}_h \times \mathbb{R}^h \to \mathrm{GL}_h$ is said to be an *equivariance* if it satisfies (1) $c(\mathrm{id}_h, z) = \mathrm{id}_h$ for all z, and (2) $c(g_1, g_2 z)c(g_2, z) = c(g_1 g_2, z)$ for all $g_1, g_2 \in \mathrm{GL}_h$ and z. These two conditions on c translate directly into the unit and multiplication axioms of a group[6], generalizing $\pi(g_1 g_2) = \pi(g_1)\pi(g_2)$, and $\pi(\mathrm{id}_h) = \mathrm{id}_h$. Every equivariance gives rise to a nonlinear action of GL_h on $\mathbf{Param} \times \mathbb{R}^h$ given by $g \cdot (U, V, x) = (Uc(g, Vx)^{-1}, gV, x)$. This action is a symmetry preserving the loss if and only if the following generalization of (4) holds:

$$\text{General Equivariance:} \qquad \sigma(gz) = c(g, z)\sigma(z) \qquad \forall g \in \mathrm{GL}_h\ \forall z \in \mathbb{R}^h \tag{9}$$

An explicit example of such an equivariance is $c(g, z) = R_{\sigma(gz)}R^{-1}_{\sigma(z)}$, and Proposition 4.2 generalizes to any general equivariance by replacing $\frac{|\sigma(Vx)|\|g\|}{|\sigma(gVx)|}$ with $\|c(g, Vx)^{-1}\|$.

[6] In fact, c defines a GL_h-equivariant structure on the tangent bundle of $\mathbb{R}^h$.

5 CONSERVED QUANTITIES OF GRADIENT FLOW

We have shown that continuous symmetries lead us along extended flat minima in the loss landscape. In this section, we identify quantities that (partially) parameterize these minima. We first show that certain real-valued functions on the parameter space remain constant during gradient flow. We refer to such functions as *conserved quantities*. Applying symmetries changes the value of the conserved quantity. Therefore, conserved quantities can be used to parameterize flat minima.

Gradient flow (GF). Recall that GD proceeds in discrete steps with the update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ where ε is the learning rate (which in general can be a symmetric matrix), and $t = 0, 1, 2, \dots$ are the time steps. In gradient flow, rather than updating the parameter values in steps, one can instead define a smooth curve in the parameter space from a choice of initial values to the limiting local minimum. The curve is a function of a continuous time variable $t \in \mathbb{R}$, and velocity of this curve at any point is equal to the gradient of the loss function, scaled by the negative of the learning rate. In other words, the dynamics of the parameters under GF are given by:

$$\dot{\boldsymbol{\theta}}(t) = d\boldsymbol{\theta}(t)/dt = -\varepsilon \nabla_{\boldsymbol{\theta}(t)} \mathcal{L}. \tag{10}$$

From an initialization $\boldsymbol{\theta}(0)$ at $t = 0$, GF defines a trajectory $\boldsymbol{\theta}(t) \in \mathbf{Param}$ for $t \in \mathbb{R}_{>0}$, which limits to a critical point. In this way, GF is a continuous version of GD.

Conserved quantities. A *conserved quantity* of GF is a function $Q : \mathbf{Param} \to \mathbb{R}$ such that the value of Q at any two time points $s, t \in \mathbb{R}_{>0}$ along a GF trajectory is the same: $Q(\boldsymbol{\theta}(s)) = Q(\boldsymbol{\theta}(t))$. In other words, we have $dQ(\boldsymbol{\theta}(t))/dt = 0$. Note that, if $f : \mathbb{R} \to \mathbb{R}$ is any function, and Q is a conserved quantity, then the composition $f \circ Q$ is also a conserved quantity. Several conserved quantities of GF have appeared in the literature, most notably layer imbalance $Q_{\text{imb}} \equiv \|W_i\|^2 - \|W_{i-1}\|^2$ (Du et al., 2018) for each pair of successive feedforward linear layers ($\sigma(x) = x$), and its full matrix version $Q_i = W_i^T W_i - W_{i-1} W_{i-1}^T$.

We now propose a generalization of the layer imbalance results mentioned above by associating a conserved quantity to any infinitesimal symmetry. As in Section 3.2, suppose a matrix Lie group G acts linearly on the parameter space. Then, from (6), we have the identity $\langle \nabla_\theta \mathcal{L}, M \cdot \boldsymbol{\theta} \rangle = 0$ for any element M in the Lie algebra $\mathfrak{g}$. Using the gradient flow dynamics (10), this identity becomes:

$$\left\langle \varepsilon^{-1} \dot{\boldsymbol{\theta}} \ , \ M \cdot \boldsymbol{\theta} \right\rangle = 0 \tag{11}$$

In other words, the velocity at any point of a gradient flow curve is orthogonal to the infinitesimal action. For simplicity, we set the learning rate to the identity: $\varepsilon = I$ (all results generalize to symmetric ε.) The following proposition (whose proof is elementary and well-known) provides a way of 'integrating' equation (11), in the appropriate sense, in order to obtain conserved quantities:

Proposition 5.1. *Suppose the action of G on $\mathbf{Param}$ is linear*[7] *and leaves $\mathcal{L}$ invariant. For any $M \in \mathfrak{g}$, there is a conserved quantity $Q_M : \mathbf{Param} \to \mathbb{R}$ given by $Q_M(\boldsymbol{\theta}) = \langle \boldsymbol{\theta} \ , \ M \cdot \boldsymbol{\theta} \rangle$.*

While Proposition 5.1 directly links the infinitesimal action to conserved quantities, it has the limitation that the conserved quantity corresponding to an anti-symmetric matrix $M = -M^T$ in $\mathfrak{g}$ is constantly zero, and we do not obtain meaningful conserved quantities. Instead, we can only conclude that flow curves satisfy the differential equation (11). Fixing a basis $(\theta^1, \cdots \theta^d)$ for $\mathbf{Param} \simeq \mathbb{R}^d$, this equation becomes $\sum_{i<j} M_{ij} r_{ij}^2 \dot{\phi}_{ij} \equiv 0$ where (r_{ij}, ϕ_{ij}) are the polar coordinates for the point $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \in \mathbb{R}^2$ (see Appendix C.9.5). In summary, we find:

$M \in \mathfrak{g}$	M symmetric	M anti-symmetric
differential equation	conserved quantity	differential equation
$\dot{\theta}^T M \theta = 0$	$Q_M(\theta) = \theta^T M \theta$	$\sum_{i<j} m_{ij} r_{ij}^2 \dot{\phi}_{ij} \equiv 0$

Conserved quantities parametrize symmetry flat directions. We now observe that applying a symmetry changes the values of the conserved quantities Q_M (Figure 1). Indeed, for $M \in \mathfrak{g}$ and

[7] For simplicity, we also assume that G is closed under taking transposes, and acts faithfully on the parameter space. These assumptions generally hold in practice; see Appendix C for a version with fewer assumptions.

$g \in G$, we have $Q_M(g \cdot \boldsymbol{\theta}) = Q_{g^T M g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbf{Param}$, so applying the group action[8] transforms the conserved quantity Q_M to $Q_{g^T M g}$. As discussed in Section 3, applying a symmetry g to a minimum $\boldsymbol{\theta}^*$ of $\mathcal{L}$ yields another minimum $g \cdot \boldsymbol{\theta}^*$; hence applying symmetries leads to a partial parameterization of flat minima. Note that, in general, we may lack sufficient number of Q_M to fully parameterize a flat minimum. For example, in the linear network UVx, $G = \mathrm{GL}_h$ and flat minima generically have h^2 dimensions, whereas the number of independent nonzero Q_M is $h(h+1)/2$, which is the dimension of the space of symmetric matrices $M = M^T$ in $\mathfrak{gl}_h$.

In gradient descent, the values of these conserved quantities may change due to the time discretization. However, the change in Q is expected to be small. For example, in two-layer linear networks, the change of Q is bounded by the square of learning rate. Appendix E contains derivations and empirical observations of the magnitude of change in Q.

Relation to Noether's theorem. In physics, Noether's theorem (Noether, 1918) states that continuous symmetries give rise to conserved quantities. Recently, Tanaka & Kunin (2021) showed that Noether's theorem can also be applied to GD by approximating it as a *second order* GF. We show that in the limit where the second order GF reduces to first order GF (10), results from Noether's theorem reduce to our conservation law $\left\langle \overline{M}_{\boldsymbol{\theta}}, \nabla \mathcal{L} \right\rangle = 0$ (6). In short, using Noether's theorem, the conserved Noether current is $J_M = e^{t/\tau} J_{0M}$ with $J_{0M} = \left\langle \overline{M}_{\boldsymbol{\theta}}, \varepsilon^{-1} \dot{\boldsymbol{\theta}} \right\rangle$. In the limit $\tau \to 0$, using (10), $J_{0M} = \left\langle \overline{M}_{\boldsymbol{\theta}}, \nabla \mathcal{L} \right\rangle = 0$ and the conservation $dJ_M/dt = 0$ implies $J_{0M} = 0$, meaning we recover (6). Details appear in Appendix B.

Examples. We present examples of conserved quantities for two-layer neural networks. all of which generalize in a straightforward way to the multi-layer case. See Appendix C.9 for full derivations (which heavily rely on properties of the trace). We adopt the notation of Section 3.3.

Example 5.1. General equivariant activation Suppose σ is equivariant under a linear action of a subgroup $G \subseteq \mathrm{GL}_h(\mathbb{R})$, so that $\pi(g)\sigma(z) = \sigma(gz)$. Then the two-layer network $F(z) = U\sigma(Vz)$ is invariant under G, as is the loss function. For symmetric $M \in \mathfrak{g}$, Proposition 5.1 yields the following conserved quantity:

$$Q_M : \mathbf{Param} \to \mathbb{R}, \qquad Q_M(U, V) = \mathrm{Tr}\left[V^T M V\right] - \mathrm{Tr}\left[U^T U d\pi(M)\right] \tag{12}$$

Indeed, this follows from the fact that $M \cdot (U, V) = (-U d\pi(M), MV)$, as in (5).

Example 5.2. Imbalance in linear layers Suppose the network is linear. Then $\sigma(z) = z$ and the loss is invariant under $\mathrm{GL}_h(\mathbb{R})$. For symmetric M we have the conserved quantity $Q_M(U, V) = \mathrm{Tr}\left[(VV^T - U^T U)M\right]$. Moreover, each component of the matrix $VV^T - U^T U$ is conserved.

Example 5.3. Homogeneous activation under scaling Suppose σ is a homogeneous activation of degree α. Let $G = (\mathbb{R}_{>0})^h$ be the positive rescaling group, so that $\sigma(gz) = g^\alpha \sigma(z)$ for any $g \in G$ and $z \in \mathbb{R}^h$. Note that the Lie algebra of G consists of all diagonal matrices in $\mathfrak{gl}_h$, so that, in particular, each $M \in \mathfrak{g}$ is symmetric. Since $d\pi(M) = \alpha M$ for any $M \in \mathfrak{g}$, we obtain the conserved quantity $Q_M(U, V) = \mathrm{Tr}\left[(VV^T - \alpha U^T U)M\right]$. Using the basis $M = E_{kk}$, we see that $\mathbf{Q} = \mathrm{diag}\left[VV^T - \alpha U^T U\right]$ is conserved (here, $\mathrm{diag}[A]$ is the leading diagonal). Special cases of this are LeakyReLU and ReLU with $\alpha = 1$.

Example 5.4. Radial rescaling activations Let σ be such a radial rescaling activation. As in Section 3.3, the orthogonal group $G = O(h)$ is a symmetry of $\mathcal{L}$. The Lie algebra $\mathfrak{g} = \mathfrak{so}_h$ comprises anti-symmetric matrices, and so Proposition 5.1 yields no non-trivial conserved quantities. However, using the canonical basis of $\mathfrak{g} = \mathfrak{so}_h$ given by $E_{[kl]} = E_{kl} - E_{lk}$ (so $[kl]$ indicates anti-symmetrized indices), one uses equation (11) to deduce the following novel result (see Appendix C.9):

Theorem 5.2. *When σ is a radial rescaling activation, we have:*

$$V\dot{V}^T - \dot{V}V^T + U^T\dot{U} - \dot{U}^T U = 0 \tag{13}$$

for any $(U, V) \in \mathbf{Param}$, where the dots indicate derivatives with respect to gradient flow.

Expanding the (k, l) entry of the matrix on the left-hand-side of (13), we obtain: $\sum_{s=1}^n r_{U,s;kl}^2 \dot{\phi}_{U,s;kl} + \sum_{s=1}^m r_{V^T,s;kl}^2 \dot{\phi}_{V^T,s;kl} = 0$, where $(r_{U,s;kl}, \phi_{U,s;kl})$ and $(r_{V^T,s;kl}, \phi_{V^T,s;kl})$

[8] Note that this procedure only works if $g^T M g$ belongs to $\mathfrak{g}$, which is the case the examples we consider.

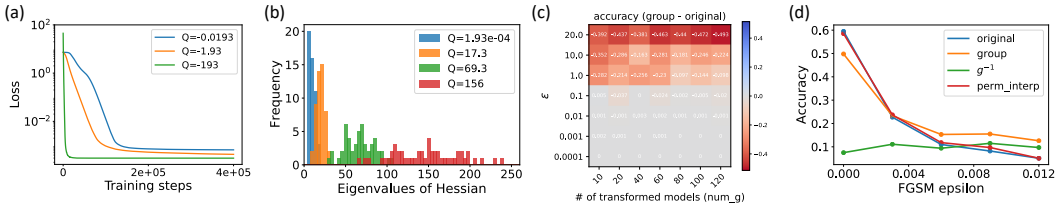

Figure 2: Overview of empirical observations with more details in Appendix G, H, and I. (a) In a two-layer neural network, the convergence rate depends on the conserved quantity Q. (b) The distribution of the eigenvalues of the Hessian at the minimum is related to the value of Q. (c) The ensemble created by group actions has similar loss values when ε is small. (d) The ensemble model improves robustness against fast gradient sign method attacks.

are the 2D polar coordinates of the points (U_{sk}, U_{sl}) and (V_{sk}^T, V_{sl}^T). This is analogous to the "angular momentum" in 2D, that is: $x \wedge \dot{x} = r^2\dot{\phi}$. Intuitively, Theorem 5.2 implies that in every 2D plane (k, l), the angular momenta of the rows of U and the columns of V sum to zero. These results also apply to linear networks $F(x) = UVx$, since rotational symmetries are linear.

6 Applications

We present a set of experiments aimed at assessing the utility of the nonlinear group action and conserved quantities. A summary of the results are shown in Figure 2. We show that the value of conserved quantities can impact convergence rate and generalizability. We also find the nonlinear action to be viable for ensemble building to improve robustness under some adversarial attacks.

Exploration of the minimum. While Q is often unbounded, common initialization methods such as (Glorot & Bengio, 2010) limit the values of Q to a small range (Appendix F). As a result, only a small part of the minimum is reachable by the models. Symmetries allow us to explore portions of flat minima that gradient descent rarely reaches.

Convergence rate and generalizability. Conserved quantities are by definition unchanged during gradient flows. By relating the values of conserved quantities to convergence rate and model generalizability, we have access to properties of the trajectory and the final model before the gradient flow starts. This knowledge allows us to choose good conserved quantity values at initialization, in order to train faster and find better solutions. In Appendix G, we derive the relation between Q and convergence rate for two example optimization problems, and provide numerical evidence that initializing parameters with optimal conserved quantity values accelerates convergence. In Appendix H, we derive the relation between conserved quantities and sharpness of minima in a simple two-layer network, and show empirically that Q values affect the eigenvalues of the Hessian (and possibly generalizability) in larger networks.

Ensemble models. Applying the nonlinear group action allows us to obtain an ensemble without any retraining or searching. We show that even with stochasticity in the data, the loss is approximately unchanged under the group action. The ensemble has the potential to improve robustness under adversarial attacks (Appendix I).

7 Discussion

While our work sheds new light onto the link between symmetries and group, it contains several limitations, which merit further investigation. First, we have not been able to determine conserved quantities in the radial rescaling case, only a differential equation that gradient flow curves must satisfy. In future work, we plan to address further the question of whether conserved quantities exist for rotation symmetries (e.g. in the radial rescaling case). These will likely not be defined on the full parameter space, but only certain subsets. Second, one major contribution of this paper is the non-linear group action of Section 4. However, our formulation only gurantees full GL_h equivariance for batch size $k = 1$. In future work, we plan to explore more consequences and variations of this non-linear group action, with the hope of generalizing to greater batch size. Finally, in many cases, parameter space symmetries lead to model compression: i.e., finding a lower-dimension space of parameters with the same expressivity of the original space of original space. Such questions are beyond the scope of the current paper; this topic would be an area of future work.