SYMMETRIES, FLAT MINIMA AND THE CONSERVED QUANTITIES OF GRADIENT FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Empirical studies have revealed that many minima are connected and reside in low-loss valleys in the loss landscape of deep networks. Ensemble models sampling different parts of a low-loss valley have reached SOTA performance. Yet, little is known about the theoretical origin of these low-loss valleys. We present a general framework for finding continuous symmetries in the parameter space, which give rise to the low-loss valleys. Importantly, we 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.

Training deep neural networks (NN) is generally a highly non-convex problem. The loss landscape of a NN is generally very rugged, with the number of local minima growing rapidly with model size (Bray & Dean, 2007; Şimşek et al., 2021). Yet, recent work has found many interesting structures in the loss landscape (Kunin et al., 2021a; Garipov et al., 2018). It has been observed that many NN loss landscapes contain approximately flat directions along which the loss does not change significantly (Freeman & Bruna, 2017; Garipov et al., 2018). The loss landscape is shaped by the model architecture and the dataset. One source of flat directions in the loss landscape is parameter transformations that keep the loss invariant (i.e. symmetries). Therefore, moving from a local minimum in the direction of a symmetry takes us to another minimum.

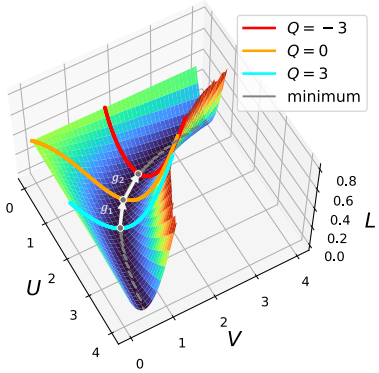


Figure 1: Visualization of the extended minimum in a 2-layer linear network with loss $\mathcal{L} = \|Y - UVX\|^2$. Points along the minum are related to each other by scaling symmetry $U \to Ug^{-1}$ and $V \to gV$. Conserved quantities, Q, associated with scaling symmetry parametrize points along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

We show that continuous symmetries of the parameter space introduce many flat directions in the loss landscape, resulting in extended local minima. It has been hypothesized that gradient descent (GD) prefers flat local minima (He et al., 2019) and that wider local minima may generalize better (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). 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) which improves model performance. However, finding such flat directions is mostly done empirically (Freeman & Bruna, 2017; Garipov et al., 2018; Draxler et al., 2018; Benton et al., 2021; Izmailov et al., 2018), with few theoretical results. One result by Kunin et al. (2021a) relates some flat directions to scaling and translational symmetries in the loss function.

In principle, continuous symmetries of the loss result in flat directions in local minima. Here we derive a general class of continuous symmetries in NN. Our key insight is to focus on *equivariances of*

the nonlinear activation functions. We show that most known continuous symmetries of NN can be derived using this framework. However, models related by such symmetries are *exactly equivalent*, meaning they cannot behave differently on different inputs. Hence, we argue that for ensembling or robustness task, 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 the linear symmetries. They define continuous symmetries for almost any activation function and involve much larger continuous groups than most known linear symmetries. We show theoretically that such nonlinear symmetries may improve model performance on out of distribution (OOD) samples. We also provide preliminary experimental evidence that ensembles using these nonlinear symmetries are more robust to attacks.

Extended flat minima arise quite frequently in the loss of NNs. We show that symmetry-induced flat minima can be parametrized using *conserved quantities*. We discuss a way to derive such conserved quantities and show how it is related to the work on Noether's theorem (Kunin et al., 2021a; Tanaka & Kunin, 2021). We also derive the explicit form of conserved quantities for different continuous symmetries of NN. Our contributions can be summarized as follows:

1. Using equivariance, we present a general framework for finding symmetries in NN.
2. We introduce a new class of nonlinear, data-dependent symmetries in nonlinear NN.
3. We show that continuous symmetries in NN can lead to flat minima in the loss landscape, which can be traversed by applying symmetries.
4. We devise a procedure for deriving conserved quantities (CQ) associated with symmetries.
5. We show that symmetry-induced flat minima can be parametrized using the associated CQ.

This paper is organized as follows. First, we review existing work on flat minima, continuous symmetries of parameter space, and conserved quantities. In sec. 1 we begin by defining continuous symmetries. Next, we define flat minima and show how linear symmetries lead to extended flat minima. Then we work out a few examples of linear symmetries in NN. After that, we introduce nonlinear, data-dependent symmetries. In sec. 2 we introduce a formalism for deriving conserved quantities for parameter space symmetries and show its relation with Noether's theorem used by others. Then, we show how CQ can be used to define coordinates along flat minima. We close with some preliminary experiments involving nonlinear symmetries, conserved quantities and a discussion of potential use cases.

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., 2021a). 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. (2021a) develop a more general approach for finding conserved quantities for certain one-parameter symmetry groups. In physics, Noether's theorem gives a conserved quantity for every continuous symmetry Noether (1918). However, while symmetry of the loss function may be known, the kinetic energy makes it difficult to find symmetry of the Lagrangian that describes the learning dynamics. Tanaka & Kunin (2021) takes a different route by studying the dynamics of the conserved quantity caused by this kinetic symmetry breaking.

Topology of minimum. The global minimum of overparametrized neural networks are connected spaces instead of isolated points. We show that linear symmetries lead to extended flat minima. Cooper (2018) proves that the global minima is usually a manifold with dimension equal to the number of parameters subtracted by the number of data points. Şimşek et al. (2021) study permutation symmetry and show that in certain overparametrized networks, the minimum related by

permutations are connected. Entezari et al. (2021) hypothesize that SGD solutions can be permuted to points on the same connected minima. Ainsworth et al. (2022) develop algorithms that find such permutations. Additional discussion on mode connectivity, ensembles, sharpness of minima and other related topics can be found in Appendix A.

1 CONTINUOUS SYMMETRIES IN DEEP LEARNING

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

1.1 THE PARAMETER SPACE AND LOSS FUNCTION

The parameters of a neural network consist of the 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 input and output 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 $Z_{i+1}(x) = W_{i+1}\sigma(Z_i(x))$ [2]. 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 $F_{\boldsymbol{\theta}}(x) = \sigma_L(Z_{\boldsymbol{\theta},L}(x))$. 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. For multiple samples, where $X \in \mathbb{R}^{n_0 \times k}$ and $Y \in \mathbb{R}^{n_L \times k}$ whose columns are the k samples[3] We retain the same notation with $F_{\boldsymbol{\theta}} : \mathbb{R}^{n_0 \times k} \to \mathbb{R}^{n_L \times k}$.) We first prove properties of $\mathcal{L}$ that hold for any training data. Hence, unless specified otherwise, we take a fixed batch of training data $\{(x_i, y_i)\}_{i=1}^k \subseteq \mathbf{Data}$, and consider the loss to be a function of the parameters only.

Example 1.1. 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 $\mathcal{L}(\boldsymbol{\theta}, (X, Y)) = \frac{1}{k}\|Y - U\sigma(VX)\|^2$.

1.2 ACTION OF CONTINUOUS GROUPS AND INFINITESIMAL SYMMETRIES

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 $I \cdot \boldsymbol{\theta} = \boldsymbol{\theta}$ where I is the identity of G, and $g_1 \cdot (g_2 \cdot \boldsymbol{\theta}) = (g_1 g_2) \cdot \boldsymbol{\theta}$ for all $g_1, g_2 \in G$.

Definition 1.1 (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:*

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

The groups we discuss below are all matrix Lie groups, which are continuous subgroups $G \subseteq \mathrm{GL}_n(\mathbb{R})$ of the general linear group of invertible $n \times n$ real matrices. Any smooth action of such a group 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 thought of as a certain subspace of matrices in $\mathfrak{gl}_n = \mathbb{R}^{n \times n}$, or (equivalently) as the tangent space at the identity I of G. For every matrix $M \in \mathfrak{g}$, we have an exponential map $\exp_M : \mathbb{R} \to G$ defined as $\exp_M(t) = \sum_{k=0}^{\infty} \frac{(tM)^k}{k!}$.

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

[2] Here $W\sigma(Z)$ denotes and arbitrary linear action, including matrix product, convolution, etc.

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

Given an action of G on $\mathbf{Param}$, the *infinitesimal action* of $M \in \mathfrak{g}$ is a vector field $\overline{M}$:

$$\text{Infinitesimal action of } M \text{ vector field:} \quad \overline{M}_{\boldsymbol{\theta}} := \frac{d}{dt}\bigg|_{t\to 0} (\exp_M(t)\cdot\boldsymbol{\theta}), \qquad \forall \boldsymbol{\theta} \in \mathbf{Param}. \tag{3}$$

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. $\langle A, B\rangle = \sum_{ijk\dots} A_{ijk\dots}B_{ijk\dots}$.

Proposition 1.2 (Infinitesimal symmetry)**.** *Let G be a matrix Lie group and a symmetry of $\mathcal{L}$. Then the gradient vector field is point-wise orthogonal to the infinitesimal action of any $M \in \mathfrak{g}$:*

$$\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}\ ,\ \overline{M}_{\boldsymbol{\theta}}\rangle = 0, \qquad \forall \boldsymbol{\theta} \in \mathbf{Param} \tag{4}$$

Sketch of proof. (See Proposition F.5.) Taking $\frac{d}{dt}\big|_0 \mathcal{L}(\exp_M(t)\cdot\boldsymbol{\theta})$ and using chain rule, we arrive at the left hand-side. From invariance $\mathcal{L}(\exp_M(t)\cdot\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})$, yielding the r.h.s. $\frac{d}{dt}\big|_0 \mathcal{L}(\boldsymbol{\theta}) = 0$. □

In the special case where G acts linearly on parameter space, we can identify[4] $\mathbf{Param} \simeq \mathbb{R}^d$. Then, $\mathrm{GL}(\mathbf{Param})$ consists of all invertible linear transformations of $\mathbf{Param}$. Suppose $G \subseteq \mathrm{GL}(\mathbf{Param})$, acting on $\mathbf{Param}$ with matrix multiplication. Then its Lie algebra $\mathfrak{g} \subseteq \mathfrak{gl}_d = \mathbb{R}^{d\times d}$, and $\overline{M}_{\boldsymbol{\theta}} = M\cdot\boldsymbol{\theta}$ acts via matrix multiplication. Hence, a corollary of Proposition 1.2 is:

Corollary 1.3. *Suppose G symmetry of $\mathcal{L}$ acting linearly on $\mathbf{Param}$. Then for $M \in \mathfrak{g}$:*

$$\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}\ ,\ M\cdot\boldsymbol{\theta}\rangle = 0, \qquad \forall \boldsymbol{\theta} \in \mathbf{Param} \tag{5}$$

Next, we discuss the effects of continuous symmetry on the loss landscape.

1.3 Symmetries lead to extended, flat minima

When the group action is linear, applying the action of any $g \in G$ to a local minimum yields another local minimum (see Appendix C for details). Again, we flatten $\boldsymbol{\theta} \in \mathbf{Param} \simeq \mathbb{R}^d$ into a column vector. Let $G \in \mathrm{GL}_d$ act as a linear symmetry. G acts on $\mathbf{Param}$ via a representation $\rho : G \to \mathrm{GL}_d(\mathbb{R})$. Fix a basis $(\theta^1, \dots, \theta^d)$ of the parameter space. Let $\rho(g)^i_j$ denote the element in row i and column j. For $g \in G$ we have $[g\cdot\boldsymbol{\theta}]^i = \sum_j \rho(g)^i_j \boldsymbol{\theta}^j$. Then, the gradient is a row vector with components $[\nabla\mathcal{L}]_i = \partial\mathcal{L}/\partial\theta^i$. We define its inner product with a column vector $\boldsymbol{\theta}' \in \mathbf{Param}$ as $\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}, \boldsymbol{\theta}'\rangle = \nabla_{\boldsymbol{\theta}}\mathcal{L}\boldsymbol{\theta}' = \sum_i [\partial_{\boldsymbol{\theta}}\mathcal{L}]_i \boldsymbol{\theta}'^i$.

Proposition 1.4. *Suppose G is a symmetry of $\mathcal{L}$ acting linearly on $\mathbf{Param}$. Then the gradients of $\mathcal{L}$ at any $\boldsymbol{\theta}$ and $g\cdot\boldsymbol{\theta}$ are related via:*

$$\nabla_{g\cdot\boldsymbol{\theta}}\mathcal{L}\ \rho(g) = \nabla_{\boldsymbol{\theta}}\mathcal{L}, \qquad \forall g \in G,\ \forall \boldsymbol{\theta} \in \mathbf{Param} \tag{6}$$

Sketch of proof. Starting form the invariance $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(g\cdot\boldsymbol{\theta})$ and taking derivatives of both sides

$$[\nabla_{\boldsymbol{\theta}}\mathcal{L}]_i = \frac{\partial\mathcal{L}}{\partial\theta^i}\bigg|_{\boldsymbol{\theta}} = \sum_j \frac{\partial\mathcal{L}}{\partial\theta^j}\bigg|_{g\cdot\boldsymbol{\theta}} \frac{\partial[g\cdot\theta]^j}{\partial\theta^i}\bigg|_{\boldsymbol{\theta}} = \sum_j \frac{\partial\mathcal{L}}{\partial\theta^j}\bigg|_{g\cdot\boldsymbol{\theta}} \rho(g)^j_i = [\nabla_{g\cdot\boldsymbol{\theta}}\mathcal{L}\rho(g)]_i \tag{7}$$

□

Corollary 1.5. *If $\boldsymbol{\theta}^*$ is a critical point (resp. local minimum) of $\mathcal{L}$, then so is $g\cdot\boldsymbol{\theta}^*$.*

Proof. If $\boldsymbol{\theta}^*$ is a critical point of $\mathcal{L}$, then $\nabla_{\boldsymbol{\theta}^*}\mathcal{L} = 0$ and Proposition 1.4 implies that $g\cdot\boldsymbol{\theta}^*$ is another critical point: $\nabla_{g\cdot\boldsymbol{\theta}^*}\mathcal{L} = \nabla_{\boldsymbol{\theta}^*}\mathcal{L}\rho(g^{-1}) = 0$ Moreover, the invariance of the loss under the G-action implies that local minima are sent to other local minimum. □

Hence, if $\boldsymbol{\theta}^*$ is a critical point, then the set $\{g\cdot\boldsymbol{\theta} \mid g \in G\}$ belongs to the critical locus. This set is known as the *orbit* of $\boldsymbol{\theta}$ under the action of G, and is isomorphic to the quotient $G/\mathrm{Stab}_G(\boldsymbol{\theta})$, where $\mathrm{Stab}_G(\boldsymbol{\theta}) = \{g \in G \mid g\cdot\boldsymbol{\theta} = \boldsymbol{\theta}\}$ is the *stabilizer* subgroup of $\boldsymbol{\theta}$ in G. In the case of a linear action, the orbit is a smooth manifold. We provide the dimensionality of the orbit in the examples in the next section (see Appendix C for more details).

[4]Concretely, let $d = \dim(\mathbf{Param})$ be the dimension of the parameter space; in terms of the widths, we have $d = \sum_{i=1}^L n_i n_{i-1}$. Then $\mathbf{Param} \simeq \mathbb{R}^d$ and parameters can be 'flattened' into d-vectors. One can identify $\mathrm{GL}(\mathbf{Param})$ with the group $\mathrm{GL}_d(\mathbb{R})$ of invertible $d\times d$ matrices, and its Lie algebra with $\mathfrak{gl}_d = \mathbb{R}^{d\times d}$.

Need for data-dependent symmetries. A symmetry of $\mathcal{L}$ which is independent of the input data transforms a set of parameters $\boldsymbol{\theta}$ to $\boldsymbol{\theta}' = g \cdot \boldsymbol{\theta}$ such that the loss doesn't change $\mathcal{L}(\boldsymbol{\theta}', X) = \mathcal{L}(\boldsymbol{\theta}, X)$. However, this means that we have $\mathcal{L}(\boldsymbol{\theta}', X') = \mathcal{L}(\boldsymbol{\theta}, X)$ even on new data X', suggesting it is impossible to use these symmetries to improve performance on OOD data. Hence, we conclude that for a symmetry to yield any benefit on unseen data, it should be at least *data-dependent*. Therefore, after reviewing linear symmetries, we introduce a set of data-dependent, nonlinear symmetries.

1.4 Equivariance of the activation function

We now show that a large class of linear symmetries of $\mathcal{L}$ can be described using an *equivariance* property of the activation functions between layers. To illustrate, we use an example of two layers with output $F(x) = U\sigma(Vx)$ for $(U, V) \in \mathbb{R}^{m\times h} \times \mathbb{R}^{h\times n}$ and $x \in \mathbb{R}^n$. All results easily generalize to multiple layers by letting $U = W_i$ and $V = W_{i-1}$ be weights in a deep NN (see appendix D.5 for multi-layer symmetries). Let $G \subseteq \mathrm{GL}_h(\mathbb{R})$ be a subgroup, and let $\pi : G \to \mathrm{GL}_h(\mathbb{R})$ a representation (the simplest example is $\pi(g) = g$). The goal is to find another set of parameters $(U', V') = g \cdot (U, V)$ such that $U\sigma(Vx) = U'\sigma(V'x)$. We propose the action of G on $\mathbf{Param}$ to be

$$g \cdot U = U\pi(g^{-1}), \qquad g \cdot V = gV \tag{8}$$

This action becomes a symmetry of $\mathcal{L}$ if the following condition is satisfied:

Proposition 1.6 (Equivariant activation)**.** *The action in* (8) *becomes a symmetry of $\mathcal{L}$ when*

$$\sigma(gz) = \pi(g)\sigma(z) \tag{9}$$

Sketch of proof. The loss depends on the parameters (U, V) only through $F(x) = UV\sigma(Vx)$, so it suffices to show that (U, V) and $g \cdot (U, V)$ give the same $F(x)$. The key is $(g \cdot U)\sigma((g \cdot V)x) = U\pi(g^{-1})\sigma(gVx) = U\sigma(Vx)$, where $x \in \mathbb{R}^n$ is any input vector. □

To describe the infinitesimal action, note that the representation $\pi : G \to \mathrm{GL}_h$ induces a representation $d\pi : \mathfrak{g} \to \mathfrak{gl}_h$ of the Lie algebra, defined by $d\pi(M) = \frac{d}{dt}\big|_0 \pi(\exp_M(t))$. Using the fact that $\exp_{-M}(t) = \exp_M(t)^{-1}$, the infinitesimal action of the Lie algebra $\mathfrak{g}$ induced by (8) is given by:

$$M \cdot U = -Ud\pi(M), \qquad M \cdot V = MV \tag{10}$$

Denoting the Jacobian of $\sigma : \mathbb{R}^h \to \mathbb{R}^h$ at $z \in \mathbb{R}^h$ by $d\sigma_z \in \mathbb{R}^{h\times h}$, the infinitesimal form of version of $\sigma(gz) = \pi(g)\sigma(z)$ is

$$\langle Mz, d\sigma_z\rangle = d\pi(M)\sigma(z) \tag{11}$$

With indices, the components are $[\langle Mz, d\sigma_z\rangle]^j = \sum_i [Mz]^i \partial_i \sigma(z)^j = [d\pi(M)\sigma(z)]^j$ When the activation is elementwise, we have $Mz \odot \sigma'(z) = d\pi(M)\sigma(z)$ ($\odot$ is elementwise multiplication).

Example 1.2. Linear networks The simplest example of Proposition 1.6 is that of linear networks, where $\sigma(x) = x$. One can take $\pi(g) = g$ and $G = \mathrm{GL}_h(\mathbb{R})$. The GL_h-orbit has dimension $h^2 - \max(0, h - n)\max(0, h - m)$.

Example 1.3. Homogeneous activations Suppose the activation $\sigma : \mathbb{R}^h \to \mathbb{R}^h$ is *homogeneous*, so that σ is applied pointwise in the standard basis. Assume also that there exists $\alpha > 0$ such that $\sigma(cz) = c^\alpha \sigma(z)$ for all $c \in \mathbb{R}_{>0}$ and $z \in \mathbb{R}^h$. These σ are equivariant under the *positive scaling group* $G \subset \mathrm{GL}_h$ consisting of diagonal matrices with positive diagonal entries. For $g \in G$, we have $g = \mathrm{diag}(\mathbf{c})$ is a diagonal matrix with $\mathbf{c} = (c_1, \ldots, c_h) \in \mathbb{R}^h_{>0}$. For $z = (z_1, \ldots, z_h) \in \mathbb{R}^h$, we have $\sigma(gz) = \sum_j \sigma(c_j z_j) = \sum_j c_j^\alpha \sigma(z_j) = g^\alpha \sigma(z)$. Hence, we set $\pi(g) = g^\alpha$. We have $d\pi(M) = \alpha M$, and the infinitesimal version of rescaling invariance of homogeneous σ becomes $Mz \odot \sigma'(z) = \alpha M\sigma(z)$. The orbit of the positive rescaling group has dimension $\min(h, \max(n, m))$.

Example 1.4. LeakyReLU This is a special case of homoegeneous activation, defined as $\sigma(z) = \max(z, 0) - s\min(z, 0)$, with $s \in \mathbb{R}_{>0}$. We have $\alpha = 1$, and $\pi(g) = g$. Since $\sigma(z) = z\sigma'(z)$, infinitesimal equivariance becomes $Mz \odot \sigma'(z) = M\sigma(z)$.

Example 1.5. Radial rescaling activations A less trivial example of continuous symmetries is the case of a radial rescaling activation (Ganev et al., 2022) where for $x \in \mathbb{R}^h \setminus \{0\}$, $\sigma(x) = f(\|x\|)x$ for some function f. Radial rescaling activations are is 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, Proposition 1.6 is satisfied with $\pi(g) = g$. The orbit of $O(h)$ has dimension $\binom{h}{2}$ if $h \leq \max(n, m)$ and $\binom{h}{2} - \binom{h-\max(m,n)}{2}$ otherwise.

1.5 Nonlinear symmetries and general equivariance

Consider the equivariance condition $\pi(g)\sigma(z) = \sigma(gz)$ in Proposition 1.6. For common activation functions such as LeakyReLU we only proved equivariance under positive scaling, which is a small subgroup of the GL_h acting on $z \in \mathbb{R}^h$. In this section we show that, under restricted conditions, it is possible to define a nonlinear, data-dependent action of the *full* GL_h under which many activations, including LeakyReLU, are equivariant.

General equivariance. A map $c : \mathrm{GL}_h \times \mathbb{R}^{h\times k} \to \mathrm{GL}_h$ is said to be an *equivariance* if (1) $c(\mathrm{id}_h, Z) = \mathrm{id}_h$ for all Z, and (2) $c(g_1, g_2 Z)c(g_1, Z) = c(g_1 g_2, Z)$ for all $g_1, g_2 \in \mathrm{GL}_h$ and Z. Every equivariance gives rise an action of GL_h on $\mathbf{Param} \times \mathrm{Data}$:

$$\text{Nonlinear group action:} \qquad g \cdot (U, V, X) = (Uc(g^{-1}, gVX), gV, X) \tag{12}$$

Indeed, the hypotheses (1) and (2) above on c translate directly into the unit and multiplication axioms. This action is a symmetry preserving the loss if and only if $c(g, Z)\sigma(Z) = \sigma(gZ)$ for all $g \in \mathrm{GL}_h$ and $Z \in \mathbb{R}^{h\times k}$. One way to satisfy this equation is to take $h = k$, restricting to a subspace of $\mathbf{Param} \times \mathbb{R}^{h\times h}$ in which $\sigma(gVX)$ is invertible (i.e., the GL_h-non-degenerate locus, see Appendix D), and set $c(g, Z) = \sigma(gZ)\sigma(Z)^{-1}$. In summary:

$$\text{General Equivariance:} \quad c(g, Z) \equiv \sigma(gZ)\sigma(Z)^{-1} \qquad \Rightarrow \qquad c(g, Z)\sigma(Z) = \sigma(gZ) \tag{13}$$

Note that $c : G \times \mathbb{R}^{h\times k} \to \mathrm{GL}_h$ is *not* a representation of G as it will depend on $\mathbf{Param}$ via $Z = VX$. Also, note that $c(g, Z)$ depends on the input data via $Z = VX$. Yet, analogous to (8), using c we can define a nonlinear action $g \cdot (U, V, X) = (Uc(g, VX)^{-1}, gV, X)$, on an appropriate subset of $\mathbf{Param} \times \mathbb{R}^{h\times h}$, as we show next. This means that on a restricted subset of data, we can make a generic σ equivariant to a generic subgroup of GL_h. Next, we show how parameters transformed using one batch X could be used on other batches.

Using nonlinear transformed weights on a different batch. Consider $F(x) = U\sigma(Vx)$. We ask whether the X-dependent transformed U, V in $g \cdot (U, V, X) = (\tilde{U}(g, VX), gV, X)$ can replace the original U, V on data other than X. Specifically, if we use $(U', V') = (Uc(g^{-1}, gVX), gV)$ to define $F'(x') = U'\sigma(V'x')$, when is $F'(x') \approx F(x')$? Using (13), we are asking when is $c(g, Z)\sigma(Z') \approx \sigma(gZ')$ with $Z = VX$ and $Z' = VX'$. First, note that for any permutation $X' = XP$, (12) and (13) yield an exact symmetry because $\sigma(Z') = \sigma(Z)P$ and $c(g, Z)\sigma(Z)P = \sigma(gZ)P$. Thus, on permutations of X even if the batch size k' of XP is larger than h, the nonlinear action is an exact symmetry. Next, assume X' is close to X with $X' = XP + \lambda H$ for $\lambda \ll 1$ and normalized $H \in \mathbb{R}^{n\times k'}$. When σ is pointwise and Lipschitz continuous, meaning $|\sigma(a) = \sigma(b)| \leq \eta|a - b|$ with constant $\eta \in \mathbb{R}_{>0}$, we can find bound on $\|F'(x') - F(x')\| \sim O(\eta\lambda)$ (Appendix E).

Adapting models to OOD using nonlinear symmetries. Consider the same $F(x) = U\sigma(Vx)$ model. (assume $\sigma(X_i) = \sigma(X)_i$ acts on each column X_i). We want to find the parameters $\boldsymbol{\theta}'$ such that for a new data point $X' \in \mathbb{R}^{n\times k'}$, the loss $\mathrm{mean}_i \mathcal{L}(\boldsymbol{\theta}', X'_i) \approx \mathrm{mean}_i \mathcal{L}(\boldsymbol{\theta}, X_i)$. We now show this can be done if X' is related to X and σ satisfies the equivariance (13). Denote $Z = VX$ and $Z' = VX'$. Assume X' is such that $Z' = gZP$, where $P \in \{0, 1\}^{k\times k'}$ is a permutation acting on samples in X (i.e. $\sum_{j=1}^{k} P_i^j = 1$). Note that because P is a permutation and σ acts column-wise, $\sigma(XP) = \sigma(X)P$. We want to find U' such that $U'\sigma(VX') = U\sigma(VX)P$. If $\sigma(gZ) = c(g, Z)\sigma(Z)$ then the answer is $U' = Uc(g^{-1}, gZ)$ (using (12)), as

$$U'\sigma(VX') = U'\sigma(gZ)P = U'c(g, Z)\sigma(Z)P = U\sigma(Z)P \tag{14}$$

which also yields $\mathcal{L}((U', V), X'_i) = \mathcal{L}((U, V), [XP]_i)$. Thus, equivariance can be a powerful tool to adapt models trained on one dataset to slightly different data (e.g. OOD). Note that the g in (14) involves the statistics of X and X' and can be an arbitrary element of GL_h. Specifically,

assuming $k, k' \geq h$ and that Z, Z' are full rank, we can find a $g \in \mathrm{GL}_h$ satisfying $Z' = gZP$ as $g = Z'P^T Z^T (ZPP^T Z^T)^{-1}$. When X' is OOD relative to the batch X, $ZPP^T Z^T$ and $Z'PZ^T$ involve the correlation matrix of X and the cross-correlation $X'P^T X^T$.

Next, we discuss how continuous symmetries give rise to conserved quantities.

2 Conserved quantities of gradient flow

We showed earlier how continuous symmetries may lead to extended flat minima in the loss. Symmetries can move us along flat minima. Now we want to identify quantities that allow us to parametrize these minima. We first show that certain quantities Q, called conserved quantities, remain constant during gradient flow (GF), a continuous version of gradient descent (GD). The value of Q is thus fixed by the initialization. We then show that applying symmetries can change the value of Q and, therefore Q can be used to parametrize flat minima.

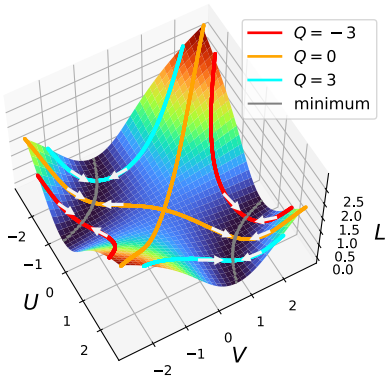


Figure 2: Gradient flow for $L(U, V) = \frac{1}{2}\|Y - UVX\|^2$, where $U, V \in \mathbb{R}$, $Y = 2$, and $X = 1$. Trajectories corresponding to different values of Q intersect the minima at different points.

Gradient flow (GF). The update rule for GD is $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ where ε is the learning rate, which in general can be a symmetric matrix. GD can be regarded as a discrete approximation of a continuous GF process, where $t \in \mathbb{R}_{>0}$. In GF, the parameters have the dynamics

$$\dot{\boldsymbol{\theta}}(t) = d\boldsymbol{\theta}(t)/dt = -\varepsilon \nabla \mathcal{L}(\boldsymbol{\theta}(t)). \tag{15}$$

At time $t = 0$, initializing at $\boldsymbol{\theta}(0) \in \mathbf{Param}$, GF defines a trajectory $\boldsymbol{\theta}(t) \in \mathbf{Param}$ for $t \in \mathbb{R}_{>0}$.

Conserved quantities A *conserved quantity* of GF is a function $Q : \mathbf{Param} \to \mathbb{R}$ such that for any two time points $s, t \in \mathbb{R}_{>0}$ along a GF trajectory $Q(\boldsymbol{\theta}(s)) = Q(\boldsymbol{\theta}(t))$. In other words, we have $dQ(\boldsymbol{\theta}(t))/dt = 0$. A few conserved Q of GF have been discovered, most notably layer imbalance $Q_{\text{imb}} \equiv \|W_i\|^2 - \|W_{i-1}\|^2$ (Du et al., 2018) for each pair of feedforward linear ($\sigma(x) = x$) layers $i-1, i$, and its full matrix version $Q_i = W_i^T W_i - W_{i-1} W_{i-1}^T$. It is known that imbalance and certain other conserved Q can be related to symmetries (Huh, 2020; Kunin et al., 2021a). We generalize these results to the symmetries discussed above. Using chain rule and the GF equation we have

$$\frac{dQ}{dt} = \left\langle \left.\frac{\partial Q}{\partial \theta}\right|_{\boldsymbol{\theta}(t)}, \frac{d\boldsymbol{\theta}(t)}{dt} \right\rangle = -\left\langle \nabla Q, \varepsilon \nabla \mathcal{L} \right\rangle = 0 \tag{16}$$

which also equals $-\left\langle \varepsilon \nabla Q, \nabla \mathcal{L} \right\rangle = 0$ because $\varepsilon = \varepsilon^T$. Note that, if $f : \mathbb{R} \to \mathbb{R}$ is any function, and Q is a conserved quantity, then $f \circ Q$ is also a conserved quantity. Additionally, any linear combination or product of conserved quantities is again a conserved quantity.

Conserved quantities from symmetries. For simplicity, set the learning rate to identity, $\varepsilon = I$. Notice the similarity of $dQ/dt = 0$ (16) with the infinitesimal group action (5) $\langle M \cdot \boldsymbol{\theta}, \nabla \mathcal{L} \rangle = 0$. We can associate a conserved Q to a symmetry by finding solutions to $\nabla Q(\boldsymbol{\theta}) = M \cdot \boldsymbol{\theta}$ (or more generally to $f(\boldsymbol{\theta}) \nabla Q(\boldsymbol{\theta}) = M \cdot \boldsymbol{\theta}$, with $f : \mathbf{Param} \to \mathbb{R}$). Alternatively, we can use the GF equation to write (5) as $\dot{Q}_M \equiv \left\langle \dot{\boldsymbol{\theta}}, M \cdot \boldsymbol{\theta} \right\rangle = 0$. Matrices in the Lie algebra $\mathfrak{g}$ give rise to conserved quantities for the gradient flow, as explained in the following proposition:

Proposition 2.1. *Suppose the action of G on $\mathbf{Param}$ is linear*[5] *and leaves $\mathcal{L}$ invariant. For any $M \in \mathfrak{g}$, the function $Q_M : \mathbf{Param} \to \mathbb{R}$ is a conserved quantity:*

$$Q_M(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, M \cdot \boldsymbol{\theta} \rangle \tag{17}$$

The space of distinct conserved quantities of the form Q_M for $M \in \mathfrak{g}$ is in one-to-one correspondence with the space of symmetric matrices in $\mathfrak{g}$.

[5]For simplicity, we also assume that G is closed under taking transposes, and that G acts *faithfully*. These assumptions generally hold in practice.

Sketch of proof. Flattening to $\boldsymbol{\theta} \in \mathbb{R}^d$, we get $M \in \mathbb{R}^{d\times d}$[6]. Assuming, $M^T \in \mathfrak{g}$, let $M^{\text{sym}} = \frac{1}{2}(M+M^T) \in \mathfrak{g}$. Using GF $\dot{\theta} = -\nabla\mathcal{L}$, we have $Q_M = \boldsymbol{\theta}^T M\boldsymbol{\theta} = Q_{M^{\text{sym}}}$ and $\dot{Q}_M = \left\langle \dot{\boldsymbol{\theta}}, M\cdot\boldsymbol{\theta}\right\rangle + \left\langle \boldsymbol{\theta}, M\cdot\dot{\boldsymbol{\theta}}\right\rangle = -\left\langle \nabla\mathcal{L}, M^{\text{sym}}\cdot\boldsymbol{\theta}\right\rangle = 0.$ □

The conserved $Q_M = Q_{M_S}$ above are only nonzero for the symmetric part of M. If M is an anti-symmetric matrix in $\mathfrak{g}$, then $Q_M \equiv 0$, so we do not obtain meaningful conserved quantities. Instead, we can only conclude that flow lines satisfy the differential equation $\left\langle \dot{\boldsymbol{\theta}}, M\cdot\boldsymbol{\theta}\right\rangle = 0$. 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 $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \in \mathbb{R}^2$ (Appendix G.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. As discussed earlier, applying a symmetry g to a minimum $\boldsymbol{\theta}^*$ of $\mathcal{L}$ yields another minimum $g\cdot\boldsymbol{\theta}^*$. Using flattened $\boldsymbol{\theta} \in \mathbb{R}^d$ it is easy to show that acting with g changes some $Q_M(\boldsymbol{\theta}) = \boldsymbol{\theta}^T M\boldsymbol{\theta}$. Let $g = \exp_{M'}(\eta) \approx I + \eta M$, with $M' \in \mathfrak{g}$ and $0 < \eta \ll 1$ be a $g \in G$ close to identity. We have $Q_M(g\cdot\boldsymbol{\theta}) = Q_M + \eta\boldsymbol{\theta}^T\left({M'}^T M + MM'\right)\boldsymbol{\theta} + O(\eta^2)$. Thus, whenever ${M'}^T M + MM' \neq 0$, applying g changes the value of Q_M. Therefore, Q_M can be used to parameterize the flat minima. However, for anti-symmetric M, we could not find nonzero Q explicitly. So we may lack sufficient number of Q_M to fully parametrize a flat minimum. For example, in the linear network UVx, $G = \text{GL}_h$ and flat minima can have h^2 dimensions, whereas the number of independent nonzero Q_M is $h(h+1)/2$, the number of symmetric basis $M \in \mathfrak{g}$.

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 (15), results from Noether's theorem reduce to our conservation law $\left\langle \overline{M}_{\boldsymbol{\theta}}, \nabla\mathcal{L}\right\rangle = 0$ (5). 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 (15), $J_{0M} = \left\langle \overline{M}_{\boldsymbol{\theta}}, \nabla\mathcal{L}\right\rangle = 0$ and the conservation $dJ_M/dt = 0$ implies $J_{0M} = 0$, meaning we recover (4). See Appendix B.2 for details.

Example 2.1. Equivariant σ Let σ be equivariant under a linear action of $g \in G \subseteq \text{GL}_h(\mathbb{R})$ with $\pi(g)\sigma(z) = \sigma(gz)$. Hence the network $F(z) = U\sigma(Vz)$ in invariant under G, and so is the loss function. For symmetric $M \in \mathfrak{g}$, using the infinitesimal action (10), the following Q_M is conserved

$$Q_M = \text{Tr}\left[Ud\pi(M)U^T\right] - \text{Tr}\left[V^T MV\right]. \tag{18}$$

Example 2.2. Imbalance in linear layers Consider our example $F(z) = U\sigma(Vz)$ with linear activation $\sigma(z) = z$. $\mathcal{L}$ is invariant under $G = \text{GL}_h(\mathbb{R})$. For symmetric $M \in \mathfrak{gl}_h(\mathbb{R})$ we have

$$Q_M = \text{Tr}\left[V^T MV\right] - \text{Tr}\left[UMU^T\right] \tag{19}$$

Using the one-hot matrices E_{kl} (only component k, l is 1, rest are 0) and setting $M = E_{\{kl\}} = E_{kl} + E_{lk}$, we get $Q_{W_{\{kl\}}} = \left[VV^T - U^T U\right]_{kl}$. This shows that each component kl of the whole matrix $\mathbf{Q} = VV^T - U^T U$ is conserved in linear layers.

Example 2.3. Homogeneous activation under scaling For $g \in G = \mathbb{R}^h_{>0}$ (positive scaling group) we have $\sigma(gz) = g^\alpha\sigma(z)$ and $d\pi(M) = \alpha M$, which yields the conserved $Q_M = \alpha\text{Tr}\left[UMU^T\right] - \text{Tr}\left[V^T MV\right]$. Note, for scaling, all M are diagonal and hence symmetric. Using the basis $M = E_{kk}$, we see that the set of all conserved Q_M becomes $\mathbf{Q} = \text{diag}\left[\alpha U^T U - VV^T\right]$ (here, $\text{diag}[A]$ is the leading diagonal). Special cases of this are LeakyReLU and ReLU with $\alpha = 1$.

[6] By abuse of notation, we denote the representation $\rho(g)$ which acts on $\mathbb{R}^d$ simply by $\rho(g) = g$, and similarly $d\rho(M) = M$.

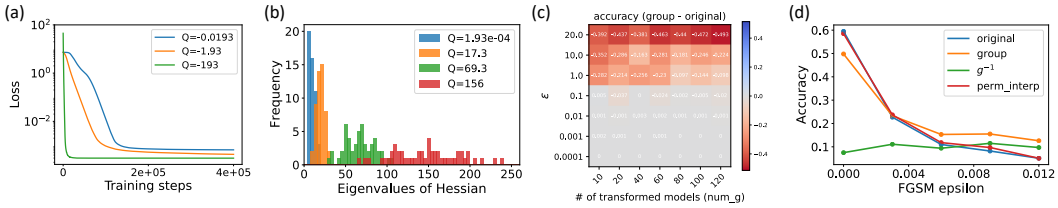


Figure 3: Overview of empirical observations with more details in Appendix M, N, and O. (a) In a two-layer neural network, the convergence rate depends on the conserved quantity Q. (b) The distribution of the eigenvalues of the Hessian at the minimum is related to the value of Q. (c) The ensemble created by group actions has similar loss values when ε is small. (d) The ensemble model improves robustness against fast gradient sign method attacks.

Example 2.4. Radial rescaling activation Consider the network $F(z) = U\sigma(Vz)$ with dimensions (n, h, m) and let $\sigma(z) = f(\|z\|)z$ be the radial rescaling activation. As in sec. 1.4, $G = O(h)$ commutes with σ, making it a symmetry of $\mathcal{L}$. The Lie algebra $\mathfrak{g}$ consists only of anti-symmetric matrices $M = -M^T$, and so (17) yields $Q_M = 0$, and no nontrivial Q_M can be found this way. However, $M \in \mathfrak{g}$ still satisfies

$$\left\langle \boldsymbol{\theta}, M \cdot \dot{\boldsymbol{\theta}} \right\rangle = -\mathrm{Tr}\left[U M \dot{U}^T\right] + \mathrm{Tr}\left[V^T M \dot{V}\right] = 0. \tag{20}$$

Using the canonical basis of $E_{[kl]} \in \mathfrak{g} = o(h)$ given by $E_{[kl]} = E_{kl} - E_{lk}$ ($[kl]$ meaning anti-symmetrized indices), (20) becomes

$$\left\langle \boldsymbol{\theta}, E_{[kl]} \cdot \dot{\boldsymbol{\theta}} \right\rangle = [\dot{U}^T U]_{[kl]} - [\dot{V} V^T]_{[kl]} = \sum_{s=1}^{n} r_{U,s}^2 \dot{\phi}_{U,s} - \sum_{s=1}^{m} r_{V^T,s}^2 \dot{\phi}_{V^T,s} = 0, \tag{21}$$

where $(r_{U,s}, \phi_{U,s})$ and $(r_{V^T,s}, \phi_{V^T,s})$ are the polar coordinates of row U_s and column $V^s = V_s^T$ in the 2D plane with basis (e_k, e_l) [7]. This is analogous to the "angular momentum" in 2D $x \wedge \dot{x} = r^2\dot{\phi}$. Intuitively, (21) states that in every 2D plane (k, l), the angular momenta of rows of U and columns of V are equal. These results also apply to the rotational symmetry $O(h) \subseteq \mathrm{GL}_h(\mathbb{R})$ which is part of the symmetries of the linear network $F(x) = UVx$.

All the result above trivially generalize to multi-layer networks, by setting $U = W_i$ and $V = W_{i-1}$.

3 Discussion of Applications

We ran a limited set of experiments aimed at assessing the utility of the nonlinear group action and conserved quantities. Some of the results are shown in Fig. 3. We find that Q 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.

Convergence rate and generalizability. Conserved Q are fixed at initialization. We find that values of Q can influence the convergence rate or the sharpness of the trained model. In Appendix M, we derive the relation between Q and convergence rate for two additional example optimization problems. We show that initializing parameters at certain Q values accelerates convergence. In Appendix N, we derive the relation between conserved quantities and sharpness of minima in a simple two-layer network, and show empirically that Q values affect sharpness (and possibly generalizability) in larger networks. While Q is often unbounded, common initialization methods such as (Glorot & Bengio, 2010) limit the values of Q to a small range (Appendix L). As a result, only a small part of the minimum is reachable by the models. Symmetries allow us to explore portions of flat minima which gradient descent rarely reaches.

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 O).

[7] The polar coordinate are found by setting $U_{sl} = r_{U,s} \cos \phi_{U,s}$ and $U_{sk} = r_{U,s} \sin \phi_{U,s}$ (same for V^T).

REFERENCES

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

Ferran Alet, Dylan Doblar, Allan Zhou, Josh Tenenbaum, Kenji Kawaguchi, and Chelsea Finn. Noether networks: meta-learning useful conserved quantities. *Advances in Neural Information Processing Systems*, 34, 2021.

Marco Armenta, Thierry Judge, Nathan Painchaud, Youssef Skandarani, Carl Lemaire, Gabriel Gibeau Sanchez, Philippe Spino, and Pierre-Marc Jodoin. Neural teleportation. *arXiv preprint arXiv:2012.01118*, 2020.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations*, 2018a.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018b.

Vijay Badrinarayanan, Bamdev Mishra, and Roberto Cipolla. Symmetry-invariant optimization in deep networks. *arXiv preprint arXiv:1511.01754*, 2015.

Robert Bamler and Stephan Mandt. Improving optimization for models with continuous symmetry breaking. In *International Conference on Machine Learning*, pp. 423–432. PMLR, 2018.

Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In *International Conference on Machine Learning*, pp. 769–779. PMLR, 2021.

Johannes Brandstetter, Max Welling, and Daniel E Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2241–2256, 2022.

Alan J Bray and David S Dean. Statistics of critical points of gaussian fields on large-dimensional spaces. *Physical review letters*, 98(15):150201, 2007.

Shih-Kang Chao, Zhanyu Wang, Yue Xing, and Guang Cheng. Directional pruning of deep neural networks. *Advances in Neural Information Processing Systems*, 33:13986–13998, 2020.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *International Conference on Learning Representations*, 2017.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Yaim Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Neural Information Processing Systems*, 2018.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

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.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Iordan Ganev, Twan van Laarhoven, and Robin Walters. Universal approximation and model compression for radial neural networks. *arXiv preprint arXiv:2107.02550v2*, 2022.

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.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Grzegorz Głuch and Rüdiger Urbanke. Noether: The more things change, the more stay the same. *arXiv preprint arXiv:2104.05508*, 2021.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Using mode connectivity for loss landscape analysis. *arXiv preprint arXiv:1806.06977*, 2018.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

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

Dongsung Huh. Curvature-corrected learning dynamics in deep neural networks. In *International Conference on Machine Learning*, pp. 4552–4560. PMLR, 2020.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

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. *International Conference on Learning Representations*, 2017.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755. PMLR, 2018.

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

Daniel Kunin, Javier Sagastuy-Brena, Lauren Gillespie, Eshed Margalit, Hidenori Tanaka, Surya Ganguli, and Daniel LK Yamins. Rethinking the limiting dynamics of sgd: modified loss, phase space oscillations, and anomalous diffusion. *arXiv preprint arXiv:2107.09133*, 2021b.

Jonathan Leake and Nisheeth K Vishnoi. Optimization and sampling under continuous symmetry: Examples and lie theory. *arXiv preprint arXiv:2109.01080*, 2021.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. $\mathcal{G}$-SGD: Optimizing relu neural networks in its positively scale-invariant space. *International Conference on Learning Representations*, 2019.

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*. PMLR, 2021.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Emmy Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pp. 235–257, 1918.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *35th Conference on Neural Information Processing Systems*, 2021.

Fabrizio Pittorino, Antonio Ferraro, Gabriele Perugini, Christoph Feinauer, Carlo Baldassi, and Riccardo Zecchina. Deep networks on toroids: Removing symmetries reveals the structure of flat regions in the landscape geometry. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17759–17781, 2022.

Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International conference on machine learning*, pp. 2892–2901. PMLR, 2017.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120v3*, 2014.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Berfin Şimşek, 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*, pp. 9722–9732. PMLR, 2021.

Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pp. 10153–10161. PMLR, 2021.

Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *Advances in Neural Information Processing Systems*, 2017.

Andre Wibisono and Ashia C Wilson. On accelerated methods in optimization. *arXiv preprint arXiv:1509.03616*, 2015.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Lechao Xiao and Jeffrey Pennington. Synergy and symmetry in deep learning: Interactions between the data, model, and inference algorithm. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 24347–24369, 2022.

Yuqian Zhang, Qing Qu, and John Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. *arXiv preprint arXiv:2205.10637*, 2022.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

CONTENTS

1 Continuous symmetries in deep learning **3**
1.1 The parameter space and loss function . . . 3
1.2 Action of continuous groups and infinitesimal symmetries . . . 3
1.3 Symmetries lead to extended, flat minima . . . 4
1.4 Equivariance of the activation function . . . 5
1.5 Nonlinear symmetries and general equivariance . . . 6

2 Conserved quantities of gradient flow **7**

3 Discussion of applications **9**

A Additional related works **16**

B Extended derivations **17**
B.1 Absorbing bias into weights . . . 17
B.2 Relation to Noether's theorem. . . . 17

C Linear symmetries lead to extended, flat minima **18**
C.1 Running example . . . 18

D Nonlinear group action **20**
D.1 Running example . . . 20
D.2 Set-up . . . 22
D.3 The action . . . 22
D.4 Infinitesimal action . . . 24
D.5 Linear symmetries: multilayer case . . . 25

E Using nonlinearly transformed weights on different inputs **25**

F Jacobians and differentials **27**
F.1 Conserved quantities . . . 28
F.2 Linear group action . . . 28
F.3 Anti-symmetric case . . . 30
F.4 Non-linear actions . . . 30

G Neural networks: linear actions **31**
G.1 General Set-up . . . 31
G.2 The linear hidden symmetry group action . . . 32
G.3 Conserved quantities for linear networks . . . 32
G.4 Conserved quantities for ReLU networks . . . 33

G.5 Conserved angular momentum for radial rescaling networks 33
G.6 Two-layer networks 34
G.6.1 ReLU networks 34
G.6.2 Radial rescaling networks 34
G.6.3 Linear networks 35

H Neural networks: non-linear actions 35
H.1 Set-up 35
H.2 The action 35
H.3 Failure of an action in the non-square case 36

I The Lagrangian approach 37
I.1 Lagrangian 37
I.2 Second-order gradient flow conserved quantities 38
I.3 Addendum: more on the Lagrangian 39
I.3.1 Constrained uniqueness 39
I.3.2 Symmetries 39
I.3.3 Limiting procedures 39

J Background: Lagrangians and Noether's theorem 40
J.1 Euler-Lagrange equation 40
J.2 First-order gradient flow cannot be derived from variational principle 40
J.3 Bregman Lagrangian 40
J.4 Relating γ to δt 41
J.5 Noether's theorem 41
J.6 Relating conserved quantity to convergence rate 42
J.7 Conserved quantities in neural networks 42
J.7.1 Two-layer linear network 42
J.7.2 Two-layer nonlinear neural networks 43
J.7.3 Two-layer nonlinear neural networks - alternative transformations 45

K Gradient descent and drifting conserved quantities 47
K.1 Change in Q in gradient descent (linear layers) 47
K.2 Empirical observations 48

L Distribution of Q under Xavier initialization 49

M Conserved quantity and convergence rate 50
M.1 Example 1: ellipse 50
M.2 Example 2: radial activation functions 51
M.3 Experiments 54

N Conserved quantity and generalization ability **54**

N.1 Example: two-layer linear network with 1D parameters 55

N.2 Experiments: two-layer networks 56

O Ensemble models **56**

A Additional related works

Conserved quantities in gradient descent While gradient flows are well approximated by gradient descent Elkabetz & Cohen (2021), the conserved quantities of gradient descent are no longer conserved in gradient flow because of the time discretization. Kunin et al. (2021a) study the dynamics of the conserved quantities in gradient descent. Kunin et al. (2021b) study continuous models that better represent stochastic gradient descent.

Mode connectivity / flat regions / ensembles In neural networks, the optima of the loss functions are connected by curves or volumes, on which the loss is almost constant (Freeman & Bruna, 2017; Garipov et al., 2018; Draxler et al., 2018; Benton et al., 2021; Izmailov et al., 2018). In these works, various algorithms have been proposed to find these low-cost curves, which provides a low-cost way to create an ensemble of models from a single trained model. Other related works include linear mode connectivity Frankle et al. (2020), using mode connectivity for loss landscape analysis Gotmare et al. (2018), and studying flat region by removing symmetry Pittorino et al. (2022).

Mixture of Experts Mixtures of Experts consist of several experts functions and a gating function that decides which experts to use (Jacobs et al., 1991; Jordan & Jacobs, 1994). The experts and gating functions can both be neural networks (Eigen et al., 2013), and the gating network can have sparse outputs so that only a small number of experts is used for each input (Shazeer et al., 2017; Fedus et al., 2022).

Sharpness of minima and generalization Recent theory and empirical studies suggest that sharp minimum do not generalize well (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Petzka et al., 2021). Explicitly searching for flat minimum has been shown to improve generalization bounds and model performance (Chaudhari et al., 2017; Foret et al., 2020; Kim et al., 2022). The sharpness of minimum can be defined using the largest loss value in the neighborhood of a minima (Keskar et al., 2017; Foret et al., 2020; Kim et al., 2022), visualization of the change in loss under perturbation with various magnitudes on weights (Izmailov et al., 2018), singularity of Hessian (Sagun et al., 2017), or the volume of the basin that contains the minimum (approximated by Radon measure (Zhou et al., 2020) or product of the eigenvalues of the Hessian (Wu et al., 2017)). Under most of these metrics, however, equivalent models can be built to have minimum with different sharpness but same generalization ability (Dinh et al., 2017). Applications include explaining the good generalization of SGD by examining asymmetric minimum (He et al., 2019), and new pruning algorithms that search for minimizers close to flat regions (Chao et al., 2020).

Parameter space symmetry and optimization While sets of parameter values related by the symmetries produce the same output, the gradients at these points are different, resulting in different learning dynamics (Kunin et al., 2021a; Van Laarhoven, 2017). This insight leads to a number of new advances in optimization. Neyshabur et al. (2015) and Meng et al. (2019) propose optimization algorithms that are invariant to symmetry transformations on parameters. Armenta et al. (2020) and Zhao et al. (2022) apply loss-invariant transformations on parameters to improve the magnitude of gradients, and consequently the convergence speed.

The structures encoded in known symmetries have also led to new optimization methods and insights of the loss landscape. Bamler & Mandt (2018) improves the convergence speed when optimizing in the direction of weakly broken symmetry. Zhang et al. (2020) discusses how symmetry helps in obtaining global minimizers for a class of nonconvex problems. The potential relevance of continuous symmetries in optimization problems was also discussed in Leake & Vishnoi (2021), which also provides an overview of Lie groups.

Symmetry in data While we focus on symmetry in the parameter space, a closely related topic is symmetry in the data (Xiao & Pennington, 2022). Equivariant neural networks incorporate data symmetry into models (Cohen & Welling, 2016; Ravanbakhsh et al., 2017; Kondor & Trivedi, 2018; Maron et al., 2018). Other related works include symmetry discovery (Alet et al., 2021; Dehmamy et al., 2021) and using symmetry for data augmentation for PDE solvers (Brandstetter et al., 2022).

B EXTENDED DERIVATIONS

B.1 ABSORBING BIAS INTO WEIGHTS

Consider the linear function $F(X) = WX + b$, where $W \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$ and $X \in \mathbb{R}^{m\times k}$ We can always absorb b into a weight matrix by appending a row of ones $\mathbf{1} \in \mathbb{R}^k$ to x as its last row, defining $\tilde{X} = (X, \mathbf{1}) \in \mathbb{R}^{(m+1)\times k}$, and appending b as the last column of the weight matrix $\tilde{W} = (W|b) \in \mathbb{R}^{n\times(m+1)}$. We have

$$\tilde{F} = \tilde{W}\tilde{X} = \left(\begin{array}{c|c} W & b \end{array}\right)\left(\frac{X}{\mathbf{1}}\right) = WX + b = F \tag{22}$$

For the group actions on $\tilde{W} \in \mathbb{R}^{n\times(m+1)}$, we need to make sure b is not mixed with W, meaning we want $g \cdot \tilde{W} = (g \cdot W|g \cdot b)$. For the action of GL_n from the left, this does not pose a restriction:

$$g \in \mathrm{GL}_n(\mathbb{R}) \qquad\qquad g \cdot \tilde{W} = g\tilde{W} = (gW|gb) \tag{23}$$

For the action of GL_m from the right we have

$$g \in \mathrm{GL}_m(\mathbb{R}) \qquad\qquad g \cdot \tilde{W} = \tilde{W}\left(\begin{array}{c|c} g & 0 \\ \hline 0 & 1 \end{array}\right) = (Wg|b) \tag{24}$$

B.2 RELATION TO NOETHER'S THEOREM.

We will now show how the approach in Tanaka & Kunin (2021) relates to our conservation law $dQ/dt = \left\langle \dot{\boldsymbol{\theta}}, M\boldsymbol{\theta} \right\rangle = 0$. Assuming a small time-step $\tau \ll 1$, we can write GD as $\boldsymbol{\theta}(t+\tau) - \boldsymbol{\theta}(t) = -\tilde{\varepsilon}\nabla\mathcal{L}(\boldsymbol{\theta}(t))$. Expanding the l.h.s to second order in τ and discarding $O(\tau^3)$ terms defines the 2nd order GF equation

$$\text{2nd order GF:} \qquad \frac{d\boldsymbol{\theta}}{dt} + \frac{\tau}{2}\frac{d^2\boldsymbol{\theta}}{dt^2} = -\varepsilon\nabla\mathcal{L}. \tag{25}$$

Here $\varepsilon = \tilde{\varepsilon}/\tau$. To use Noether's theorem, the dynamics (i.e. GF) must be a variational (Euler-Lagrange (EL)) equation derived from an "action" $S(\boldsymbol{\theta})$ (objective function), which for (25) is the time integral of Bregman Lagrangian Wibisono & Wilson (2015) L

$$S(\boldsymbol{\theta}) = \int dt L(\boldsymbol{\theta}(t), \dot{\boldsymbol{\theta}}(t); t) = \int_\gamma dt e^{t/\tau}\left[\frac{\tau}{2}\left\langle \dot{\boldsymbol{\theta}}, \varepsilon^{-1}\dot{\boldsymbol{\theta}}\right\rangle - \mathcal{L}(\boldsymbol{\theta})\right] \tag{26}$$

where $\boldsymbol{\theta} : \mathbb{R} \to \mathbf{Param}$ is a trajectory (flow path) in $\mathbf{Param}$, parametrized by t. The variational EL equations find the paths γ^* which minimize the action, meaning $\partial S_\gamma/\partial\gamma|_{\gamma^*} = 0$.

Noether's theorem states that if $M \in \mathfrak{g}$ is a symmetry of the *action* $S(\boldsymbol{\theta})$ (26) (not just the loss $\mathcal{L}(\boldsymbol{\theta})$), then the Noether current J_M is conserved

$$\begin{aligned} \text{Noether current:} \quad & J_M = \left\langle \overline{M}_{\boldsymbol{\theta}}, \frac{\partial L}{\partial\dot{\boldsymbol{\theta}}}\right\rangle = e^{t/\tau}\left\langle \overline{M}_{\boldsymbol{\theta}}, \varepsilon^{-1}\dot{\boldsymbol{\theta}}\right\rangle = e^{t/\tau}J_{0M}, \\ \text{Conservation:} \quad & \frac{dJ_M}{dt} = e^{t/\tau}\left[\frac{1}{\tau}J_{0M} + \frac{dJ_{0M}}{dt}\right] = 0, \quad \Rightarrow \quad J_{0M}(t) = J_{0M}e^{-t/\tau} \end{aligned} \tag{27}$$

Tanaka & Kunin (2021) also derived the Noether current (27), but concludes that because $L(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) \neq \mathcal{L}(\boldsymbol{\theta})$, the symmetries are "broken" and therefore doesn't derive conserved charges for the types of symmetries we discussed above. However, while Tanaka & Kunin (2021) focuses on 2nd order GF,

we note that our conserved Q were derived for *first order GF*, which is found from the $\tau \to 0$ limit of 2nd oder GF. In this limit $L \to e^{t/\tau}\mathcal{L}$ and thus symmetries of L also becomes symmetries of $\mathcal{L}$. When $\tau \to 0$, 2nd order GF reduces to $\varepsilon^{-1}\dot{\boldsymbol{\theta}} = -\nabla\mathcal{L}$ the conserved charge goes to

$$\lim_{\tau\to 0} J_{0M} = \left\langle \overline{M}_{\boldsymbol{\theta}}, \nabla\mathcal{L}\right\rangle = J_{0M}(0) \lim_{\tau\to 0} e^{-t/\tau} = 0, \tag{28}$$

which means that we recover the invariance under infinitesimal action (5). In fact, for linear symmetries and symmetric $M \in \mathfrak{g}$, $J_{0M} = dQ_M/dt = 0$.

C Linear symmetries lead to extended, flat minima

In this section, we show that, in the case of a linear group action, applying the action of any element of the group to a local minimum yields another local minimum. This fact is a corollary of a more general result; in order to describe it and remove ambiguity, we include the following clarifications. Let G be a matrix Lie group acting as a linear symmetry. Fix a basis $(\theta_1, \dots, \theta_d)$ of the parameter space. The gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ of the loss $\mathcal{L}$ at a point $\boldsymbol{\theta} \in \mathbf{Param}$ in the parameter space as another vector in $\mathbf{Param} \simeq \mathbb{R}^d$, whose i-th coordinate is the partial derivative $\frac{\partial\mathcal{L}}{\partial\theta_i}\big|_{\boldsymbol{\theta}}$. Hence, it makes sense to apply the group action to the gradient: $g \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}$. We regard vectors in $\mathbf{Param} \simeq \mathbb{R}^d$ as column vectors with d rows. Thus, the transpose of any vector is a row vector with d columns. In the case of the gradient, its transpose at $\boldsymbol{\theta}$ matches the *Jacobian* $d\mathcal{L}_{\boldsymbol{\theta}} \in \mathbb{R}^{1\times d}$ of $\mathcal{L}$ (see Appendix F), that is: $d\mathcal{L}_{\boldsymbol{\theta}} = (\nabla_{\boldsymbol{\theta}}\mathcal{L})^T$. Alternative notation for the Jacobian is $d\mathcal{L}_{\boldsymbol{\theta}_0} = \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\big|_{\boldsymbol{\theta}_0}$, where we now use $\boldsymbol{\theta}$ as a dummy variable and $\boldsymbol{\theta}_0 \in \mathbf{Param}$ as a specific value. As noted above, we are interested in matrix Lie groups $G \subseteq \mathrm{GL}_d(\mathbb{R}) = \mathrm{GL}(\mathbf{Param})$, and assume that the matrix transpose g^T belongs to G for any $g \in G$. These assumptions hold in all examples of interest.

Proposition C.1. *Suppose the action of G on the parameter space is linear and leaves the loss invariant. Then the gradients of $\mathcal{L}$ at any $\boldsymbol{\theta}_0$ and $g \cdot \boldsymbol{\theta}_0$ are related as follows:*

$$g^T \cdot \nabla_{g\cdot\boldsymbol{\theta}_0}\mathcal{L} = \nabla_{\boldsymbol{\theta}_0}\mathcal{L} \qquad \forall g \in G,\ \forall \boldsymbol{\theta}_0 \in \mathbf{Param} \tag{29}$$

Sketch of proof. Let $T_g : \mathbf{Param} \to \mathbf{Param}$ be the transformation corresponding to $g \in G$. The the Jacobian $d\mathcal{L}_{\boldsymbol{\theta}_0}$ is given by:

$$d\mathcal{L}_{\boldsymbol{\theta}_0} = \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_0} = \frac{\partial(\mathcal{L}\circ T_g)}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_0} = \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\bigg|_{g\cdot\boldsymbol{\theta}_0}\frac{\partial T_g}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_0} = d\mathcal{L}_{g\cdot\boldsymbol{\theta}_0}\circ T_g$$

where we have used the definition of the Jacobian, the invariance of the loss ($\mathcal{L}\circ T_g = \mathcal{L}$), the chain rule, the linearity of the action. The result follows from applying $T_{g^{-1}}$ on the right to both sides, and using the fact that the gradient is the transpose of the Jacobian (see Appendix **??**). □

The following corollary follows from Proposition C.1, the invariance of $\mathcal{L}$ under the action of G, and the fact that $\nabla_{\boldsymbol{\theta}^*}\mathcal{L} = 0$ at a critical point $\boldsymbol{\theta}^*$ of $\mathcal{L}$.

Corollary C.2. *If $\boldsymbol{\theta}^*$ is a critical point (resp. local minimum) of $\mathcal{L}$, then so is $g \cdot \boldsymbol{\theta}^*$.*

Hence, if $\boldsymbol{\theta}^*$ is a critical point, then the set $\{g \cdot \boldsymbol{\theta} \mid g \in G\}$ belongs to the critical locus. This set is known as the *orbit* of $\boldsymbol{\theta}$ under the action of G, and is isomorphic to the quotient $G/\mathrm{Stab}_G(\boldsymbol{\theta})$, where $\mathrm{Stab}_G(\boldsymbol{\theta}) = \{g \in G \mid g \cdot \boldsymbol{\theta} = \boldsymbol{\theta}\}$ is the *stabilizer* subgroup of $\boldsymbol{\theta}$ in G. In the case of a linear action, the orbit is a smooth manifold. While the corollary above implies that the critical locus is a union of G-orbits, it does not imply, in general, that the critical locus is a single G-orbit. It also does not rule out the case that the stabilizer is a somewhat 'large' subgroup of G, in which case the orbit would have low dimension.

C.1 Running example

Recall that the parameter space of a two-layer network is $\mathbf{Param} = \mathbb{R}^{m\times h} \times \mathbb{R}^{h\times n}$, where the dimension vector is (m, h, n), and we write elements as (U, V). The action of GL_h is $g \cdot (U, V) = (Ug^{-1}, gV)$. Let $\mathbf{Param}^\circ \subseteq \mathbf{Param}$ be the subset of pairs (U, V) where each of U and V have full rank. This is an open dense subset of $\mathbf{Param}$, and is preserved by the GL_h-action.

Proposition C.3. *The* GL_h*-orbit of each element of* $\mathbf{Param}^\circ$ *has dimension*

$$\dim(\mathrm{Orbit}) = h^2 - \max(0, h-n)\max(0, h-m).$$

Proof. Fix $(U, V) \in \mathbf{Param}^\circ$. Suppose $h \leq n$, so that $h^2 - \max(0, h-n)\max(0, h-m) = h^2$. Then $V \in \mathbb{R}^{h\times n}$ defines a surjective linear map, so has a right inverse $V^\dagger \in \mathbb{R}^{n\times h}$. If $g \in \mathrm{GL}_h$ belongs to the stabilizer of (U, V), then we have $gV = V$. Applying $V^\dagger$ on the right to both sides, we obtain $g = \mathrm{id}_h$. Thus, the stabilizer of (U, V) is trivial, and the orbit has dimension equal to the dimension of the group, namely h^2. The case $h \leq m$ is similar.

Now suppose $h > \max(n, m)$. In this case, $h^2 - \max(0, h-n)\max(0, h-m) = h(n+m) - nm$. Set $U_0 = [\mathrm{id}_m \quad 0] \in \mathbb{R}^{m\times h}$ and $V_0 = \begin{bmatrix}\mathrm{id}_n \\ 0\end{bmatrix} \in \mathbb{R}^{h\times n}$, so that the last $h-m$ rows of U_0 are zero, and the last $h-n$ columns of V_0 are zero. Then, by the rank assumption, there exists $g_1 \in \mathrm{GL}_h$ such that $Ug_1 = U_0$, and there exists $g_2 \in \mathrm{GL}_h$ such that $g_2^{-1}V = V_0$. Without loss of generality, we can take g_1 and g_2 such that $\det(g_1) > 0$ and $\det(g_2) > 0$. Thus, both g_1 and g_2 belong to the component of the identity in GL_h.

Next, consider the action of $G = \mathrm{GL}_h$ on full rank matrices in $\mathbb{R}^{m\times h}$ and $\mathbb{R}^{h\times n}$ individually. We have that $\mathrm{Stab}_G(U) = g_1\mathrm{Stab}_G(U_0)g_1^{-1}$ and $\mathrm{Stab}_G(V) = g_2\mathrm{Stab}_G(V_0)g_2^{-1}$. The stabilizer in GL_h of the pair $(U, V) \in \mathbf{Param}^\circ$ can be written as:

$$\mathrm{Stab}_G(U, V) = \{g \in G \mid Ug^{-1} = U \text{ and } gV = V\} \tag{30}$$
$$= \mathrm{Stab}_G(U) \cap \mathrm{Stab}_G(V) \tag{31}$$
$$= (g_1\mathrm{Stab}_G(U_0)g_1^{-1}) \cap (g_2\mathrm{Stab}_G(V_0)g_2^{-1}) \tag{32}$$

Since g_1 and g_2 belong to the connected component of the identity, the dimension of $(g_1\mathrm{Stab}_G(U_0)g_1^{-1}) \cap (g_2\mathrm{Stab}_G(V_0)g_2^{-1})$ is equal to the dimension[8] of $\mathrm{Stab}_G(U_0) \cap \mathrm{Stab}_G(V_0)$. Hence we reduce the problem to computing the dimension of $\mathrm{Stab}_G(U_0) \cap \mathrm{Stab}_G(V_0)$.

To this end, observe that a matrix g belongs to $\mathrm{Stab}_G(U_0)$ if and only if is of the form:

$$g = \begin{bmatrix}\mathrm{id}_m & 0 \\ * & *\end{bmatrix}$$

where the lower left $* \in \mathbb{R}^{(h-m)\times m}$ and the lower right $* \in \mathrm{GL}_{h-m}$ are arbitrary. Similarly, a matrix g belongs to $\mathrm{Stab}_G(V_0)$ if and only if is of the form:

$$g = \begin{bmatrix}\mathrm{id}_n & * \\ 0 & *\end{bmatrix}$$

where the upper right $* \in \mathbb{R}^{n\times(h-n)}$ and lower right $* \in \mathrm{GL}_{h-n}$ are arbitrary. If $m \geq n$, taking the intersection amounts to considering matrices of the form:

$$g = \begin{bmatrix}\mathrm{id}_n & 0 & 0 \\ 0 & \mathrm{id}_{m-n} & 0 \\ 0 & * & *\end{bmatrix}$$

where the rows and columns are divided according to the partition $h = n + (m-n) + (h-m)$. If $m \geq n$, taking the intersection amounts to considering matrices of the form:

$$g = \begin{bmatrix}\mathrm{id}_m & 0 & 0 \\ 0 & \mathrm{id}_{n-m} & * \\ 0 & 0 & *\end{bmatrix}$$

where the rows and columns are divided according to $h = m + (n-m) + (h-n)$. In both cases, the dimension of the intersection is $(h-n)(h-m) = h^2 - hn - hm + nm$. We obtain the dimension of the orbit as:

$$h^2 - (h-n)(h-m) = h(n+m) - nm.$$

□

[8]Explicitly, fix a continuous path $\gamma_i : [0, 1]$ such that $\gamma_i(0)$ is the identity in G and $\gamma_i(1) = g_i$, for $i = 1, 2$. The dimension of $(\gamma_1(t)\mathrm{Stab}_G(U_0)\gamma_1(t)^{-1}) \cap (\gamma_2(t)\mathrm{Stab}_G(V_0)\gamma_2(t)^{-1})$ is constant along this path.

Corollary C.4. *The orbit of a point in* $\mathbf{Param}^\circ$ *under the appropriate symmetry group is given by:*

Type of activation	Symmetry group	Dimension of random orbit
Linear	GL_h	$h^2 - \max(0, h-n)\max(0, h-m)$
Homogeneous	$T_+(h)$	$\min(h, \max(n,m))$
Radial rescaling	$O(h)$	$\begin{cases} \binom{h}{2} & \text{if } h \leq \max(n,m) \\ \binom{h}{2} - \binom{h-\max(m,n)}{2} & \text{otherwise} \end{cases}$

where $T_+(h)$ *is the coordinate-wise positive rescaling subgroup of* GL_h*, consisting of diagonal matrices with positive entries along the diagonal.*

Proof. Adopt the notation of the proof of the above Proposition. The stabilizer in $T_+(h)$ of (U_0, V_0) is the intersection of the stabilizer in GL_h of (U_0, V_0) and $T_+(h)$. This intersection is easily seen to have dimension $\max(0, h - \max(n,m))$. Subtracting this from $\dim(T_+(h)) = h$, we obtain the result for the homogeneous case. For the orthogonal case, the stabilizer in $O(h)$ of (U_0, V_0) is the intersection of the stabilizer in GL_h of (U_0, V_0) and $O(h)$. This intersection has dimension 0 if $h \leq \max(n,m)$ and $\binom{h-\max n,m}{2}$ otherwise. Subtracting from $\dim(O(h)) = \binom{h}{2}$, we obtain the result for the radial rescaling case. □

$$\text{Linear case} \qquad \dim = h^2 - \max(0, h-n)\max(0, h-m) \tag{33}$$

$$\text{Homogeneous case} \qquad \dim = \min(h, \max(n,m)) \tag{34}$$

$$\text{Radial rescaling case} \qquad \dim = \begin{cases} \binom{h}{2} & \text{if } h \leq \max(n,m) \\ \binom{h}{2} - \binom{h-\max(m,n)}{2} & \text{otherwise} \end{cases} \tag{35}$$

D NONLINEAR GROUP ACTION

D.1 RUNNING EXAMPLE

Consider the case $L = 2$ with no bias vectors. The dimension vector is $\mathbf{n} = (n, h, m)$, so the parameter space is:

$$\mathbf{Param} = \mathbb{R}^{m \times h} \times \mathbb{R}^{h \times n}$$

We write and element as (U, V) instead of (W_2, W_1). Let $\sigma = \sigma_1 : \mathbb{R}^h \to \mathbb{R}^h$ be the activation. By abuse of notation, we also write $\sigma : \mathbb{R}^{h \times k} \to \mathbb{R}^{h \times k}$ for the map defined by applying σ to each column separately. As usual, k denotes the number of samples. For the remainder of this section, we will consider the case where the number of samples to be equal to the hidden dimension, that is: $k = h$.

Definition D.1. *For any subgroup G of the hidden symmetry group $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} = \mathrm{GL}_h$, the G-non-degenerate locus is defined as the set of all $(U, V, X) \in \mathbf{Param} \times \mathbb{R}^{n \times h}$ such that the $h \times h$ matrix $\sigma(gVX)$ is invertible for all $g \in G$, that is:*

$$\left(\mathbf{Param} \times \mathbb{R}^{n \times h}\right)^\circ_G = \{(U, V, X) \in \mathbb{R}^{m \times h} \times \mathbb{R}^{h \times n} \times \mathbb{R}^{n \times h} \mid \sigma(gVX) \in \mathrm{GL}_h \ \ \forall g \in G\}$$

Lemma D.2. *There is non-linear action of G on the non-degenerate locus given by:*

$$G \circlearrowright \left(\mathbf{Param} \times \mathbb{R}^{n \times h}\right)^\circ_G \qquad g \cdot (U, V, X) = \left(U\sigma(VX)\left(\sigma(gVX)\right)^{-1}, gV, X\right)$$

Moreover, this action leaves the feedforward function unchanged.

Proof. The unit axiom follows easily:

$$\begin{aligned} \mathrm{id}_h \cdot (U, V, X) &= \left(U\sigma(VX)\left(\sigma(\mathrm{id}_h VX)\right)^{-1}, \mathrm{id}_h V, X\right) = \left(U\sigma(VX)\left(\sigma(VX)\right)^{-1}, V, X\right) \\ &= (U, V, X) \end{aligned}$$

So does the multiplication axiom:

$$\begin{aligned} g_2 \cdot (g_1 \cdot (U, V, X)) &= g_2 \cdot \left(U\sigma(VX) \left(\sigma(g_1 VX)\right)^{-1}, g_1 V, X \right) \\ &= \left(U\sigma(VX) \left(\sigma(g_1 VX)\right)^{-1} \sigma(g_1 VX) \left(\sigma(g_2 g_1 VX)\right)^{-1}, g_2 g_1 V, X \right) \\ &= \left(U\sigma(VX) \left(\sigma(g_2 g_1 VX)\right)^{-1}, g_2 g_1 V, X \right) = g_2 g_1 \cdot (U, V, X) \end{aligned}$$

The last claim follows from the observation that the product $U\sigma(VX)$ is preserved under this group action. □

Finally, we discuss the infinitesimal action. Since we have a non-linear action, the infinitesimal action is a vector field on the non-degenerate locus. We identify the tangent space of the non-degenerate locus at (U, V) with the ambient vector space $\mathbf{Param} \times \mathbb{R}^{n \times h}$.

Lemma D.3. *The value of the infinitesimal action vector field of $M \in \mathfrak{gl}_h$ at a point (U, V, X) in the non-degenerate locus is given by:*

$$\overline{M}_{(U,V,X)} = \left(U \circ d\sigma_{VX}(MVX) \circ \left(\sigma(VX)\right)^{-1}, MV, 0 \right)$$

Before giving the proof we clarify the notation and give two remarks. The first factor is the product of the matrix $U \in \mathbb{R}^{m \times h}$; the matrix output of $d\sigma_{VX} : \mathbb{R}^{h \times h} \to \mathbb{R}^{h \times h}$ applied to MVX, where, as noted above, $\sigma : \mathbb{R}^{h \times h} \to \mathbb{R}^{h \times h}$ denotes the map applying the activation column-wise; and the matrix $\sigma(VX)^{-1} \in \mathbb{R}^{h \times h}$.

The first remark is: the tuple $\overline{M}_{U,V,X}$ never belongs to the non-degenerate locus if $h > 1$. Indeed, since its last coordinate is the zero $n \times h$ matrix, we compute the first partial feedforward function as $\sigma(MV0) = \sigma(0)$. This is a matrix whose columns are all equal, and hence is non-invertible. In particular, the infinitesimal action vector field does not reduce, in general, to an action of the Lie algebra $\mathfrak{gl}_h$ on $\mathbf{Param} \times \mathbb{R}^{n \times h}$.

The second remark is: suppose $\pi : G \to \mathrm{GL}_h$ is a representation of G, and suppose σ satisfies $\sigma(gz) = \pi(z)\sigma(z)$ for all $z \in \mathbb{R}^h$. Then one verifies that $d\sigma_{VX}(MVX) = d\pi(M) \circ \sigma(VX)$. Hence, in this case, the infinitesimal action vector field reduces to a linear action on the parameter space given by: $M \cdot (U, V) = (-U \circ d\pi(M), MV)$.

Sketch of proof of Lemma D.3. The key identity is:

$$\left.\frac{d}{dt}\right|_0 \left(\sigma(\exp_M(t)VX)\right)^{-1} = \sigma(VX)^{-1} \circ d\sigma_{VX}(MVX) \circ \sigma(VX)^{-1} \tag{36}$$

for any $M \in \mathfrak{gl}_h$ and any (U, V, X) in the non-degenerate locus. One way to verify this identity is to use the fact that, for all t, we have:

$$\sigma(\exp_M(t)VX) \left(\sigma(\exp_M(t)VX)\right)^{-1} = \mathrm{id}_h$$

Taking the derivative of both sides with respect to t and rearranging terms gives identity. We now compute the value of the infinitesimal action vector field of M at $(U, V, X) \in \left(\mathbf{Param} \times \mathbb{R}^{n \times h}\right)^{\circ}_G$:

$$\begin{aligned} \overline{M}_{(U,V,X)} &= \left.\frac{d}{dt}\right|_0 \left(U \circ \sigma(VX) \circ \left(\sigma(\exp_M(t)VX)\right)^{-1}, \exp_M(t)V, X \right) \\ &= \left(U \circ \sigma(VX) \circ \left.\frac{d}{dt}\right|_0 \left(\sigma(\exp_M(t)VX)\right)^{-1}, MV, 0 \right) \\ &= \left(U \circ \sigma(VX) \circ \sigma(VX)^{-1} \circ d\sigma_{VX}(MVX) \circ \sigma(VX)^{-1}, MV, 0 \right) \\ &= \left(U \circ d\sigma_{VX}(MVX) \circ \left(\sigma(VX)\right)^{-1}, MV, 0 \right) \end{aligned}$$

where the first equality is the definition of $\overline{M}_{(U,V,X)}$, the second is elementary differentiation, and the third is the application of identity (36) from above. □

D.2 SET-UP

For the general case, we require some additional notation. Taking the parameters into account, we form the following extension of the feedforward function F:

$$\tilde{F} : \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^m, \qquad (\boldsymbol{\theta}, x) \mapsto F_{\mathbf{W}}(x)$$

One can also define the extension of the partial feedforward function $\widetilde{F}_i : \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ as $(\boldsymbol{\theta}, x) \mapsto F_{\boldsymbol{\theta},i}(x)$. Grouping the parameters, input data, and output value together, we obtain a map:

$$\widetilde{\mathcal{L}} : \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

taking $(\boldsymbol{\theta}, x, y)$ to $\mathcal{L}_{x,y}(\boldsymbol{\theta}) = \|y - F_{\boldsymbol{\theta}}(x)\|^2$. We summarize these maps in the following diagram, which holds for any choice of parameters $\boldsymbol{\theta}$, input data x, and output value y:

$$\begin{tikzcd}
\mathbb{R}^n \arrow[r, "F_{\boldsymbol{\theta}}=\text{Feedforward}"] \arrow[d, hook] & \mathbb{R}^m \arrow[dr, bend left, "\|y-(-)\|^2"] & \\
\mathbf{Param} \times \mathbb{R}^n \arrow[ur, "\widetilde{F}"] \arrow[r, hook] & \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \arrow[r, "\widetilde{\mathcal{L}}"] & \mathbb{R} \\
\mathbf{Param} \arrow[u, hook] \arrow[urr, bend right, "\mathcal{L}_{x,y}=\text{Loss}"'] & &
\end{tikzcd}$$

The inclusions $\mathbb{R}^n \hookrightarrow \mathbf{Param} \times \mathbb{R}^n$, $\mathbf{Param} \hookrightarrow \mathbf{Param} \times \mathbb{R}^n$, and $\mathbf{Param} \times \mathbb{R}^n \hookrightarrow \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m$ are given by $x \mapsto (\boldsymbol{\theta} x)$, $\boldsymbol{\theta} \mapsto (\boldsymbol{\theta}, x)$, and $(\boldsymbol{\theta}, x) \mapsto (\boldsymbol{\theta}, x, y)$. The top right map takes $\hat{y} \in \mathbb{R}^m$ to $\|y - \hat{y}\|^2$.

Example D.1. Consider the case $L = 2$ with no bias. The dimension vector is $(n_0, n_1, n_L) = (n, h, m)$, so the parameter space is:

$$\mathbf{Param}(n, n, m) = \mathbb{R}^{h \times n} \times \mathbb{R}^{m \times h}$$

Corresponding to parameters $\boldsymbol{\theta} = (V, U) \in \mathbf{Param}(n, h, m)$ and a data point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, we have the functions:

$$\begin{aligned}
F_{V,U} &: \mathbb{R}^n \to \mathbb{R}^m & x &\mapsto U\sigma(Vx) \\
\widetilde{F} &: \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^m, & (V, U, x) &\mapsto U\sigma(Vx) \\
\mathcal{L}_{x,y} &: \mathbf{Param} \to \mathbb{R}, & (V, U) &\mapsto \|y - U\sigma(Vx)\|^2 \\
\widetilde{\mathcal{L}} &: \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, & (V, U, x, y) &\mapsto \|y - U\sigma(Vx)\|^2.
\end{aligned}$$

Remark We note alternative notation that can be useful for certain reformulations. For $i = 1, \ldots, L$, define a map $\Phi_i : \mathbf{Param} \times \mathbb{R}^{n_0 \times k} \to \mathbb{R}^{n_i \times k}$ by setting:

$$\Phi_i(\boldsymbol{\theta}, X) = W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X) + b_i.$$

Hence Φ_i satisfies the recursive relation $\Phi_i(\boldsymbol{\theta}, X) = W_i \sigma_{i-1}(\Phi_{i-1}(\boldsymbol{\theta}, X)) + b_i$, with $\Phi_1(\boldsymbol{\theta}, X) = W_1 X + b_1$. Additionally, we have $\tilde{F}_i(\boldsymbol{\theta}, X) = \sigma_i(\Phi_i(\boldsymbol{\theta}, X))$ for $i = 1, \ldots, L$.

D.3 THE ACTION

In this section, we define a certain non-linear action on the space of parameters and data. This action is only well-defined with certain width vectors. Specifically, suppose all hidden dimensions are equal, so that $n_1 = n_2 = \cdots = n_{L-1}$, and call this common value h. Suppose further that the number of samples is also equal to h, that is: $k = h$. In this case, for $i = 1, \ldots, L-1$, the i-th partial feedforward function is a function $F_i : \mathbb{R}^{h \times h} \to \mathbb{R}^{h \times h}$. The network input dimension $n = n_0$ and network output dimension $m = n_L$ are not required to be equal to h. In this case, the hidden symmetry group is the product of $L-1$ copies of the GL_h, that is: $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} = (\mathrm{GL}_h)^{L-1} = \mathrm{GL}_h \times \cdots \times \mathrm{GL}_h$.

Definition D.4. *Let G be a subgroup of $(\mathrm{GL}_h)^{L-1}$. The G-non-degenerate locus of the product $\big(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n \times h}\big)$ is defined as the set of pairs $(\boldsymbol{\theta}, X)$ such that*

$$\sigma_i(g_i W_i F_{\boldsymbol{\theta},i-1}(X) + g_i b_i) \in \mathrm{GL}_h$$

for $i = 1, \ldots, L-1$. We denote this subset as $\big(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n \times h}\big)^{\circ}_G$.

(Recall that $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b}) = (W_i, b_i)_{i=1}^L$ encodes the weights and the biases.) We write an element of the hidden symmetry group as $g = (g_i)_{i=1}^{L-1}$, with each $g_i \in \mathrm{GL}_h$, and denote the identity element of the group as id. Consider the map:

$$\alpha : G \times \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n \times n}\right)_G^\circ \to \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n \times n}\right)_G^\circ$$

given by:

$$(g, \boldsymbol{\theta}, X) \mapsto \left(\left(g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X) \sigma_{i-1}(g_{i-1} W_{i-1} \tilde{F}_{i-2}(\boldsymbol{\theta}, X) + g_{i-1} b_{i-1})^{-1} \ , \ g_i b_i \right)_{i=1}^L \ , \ X \right)$$

where, for convenience, we set $g_0 = \mathrm{id}_n$ and $g_L = \mathrm{id}_{n_L}$, and the case $i = 1$ is actually $(g_1 W_1, g_1 b_1)$. In other words, for $i = 2, \ldots, L-1$, to update the weight matrix, we first compute the non-linear composition $\sigma_i(g_{i-1} W_{i-1} \widetilde{F}_{i-2}(\boldsymbol{\theta} X) + g_{i-1 b_{i-1}})$, which is an $h \times h$ matrix. Then, we multiply the inverse of this matrix on the right to the product of the $h \times h$ matrices g_i, W_i and $F_{i-1}(X)$. The bias transforms more easily as $g_i b_i$. In the case $i = L$, we have that W_L is $n_L \times h$ rather than $h \times h$, and $g_L = 1$, but otherwise the procedure is the same. The input data X is unchanged.

Remark In the notation with Φ_i, the map α can be stated as:

$$(g, \boldsymbol{\theta}, X) \mapsto \left(\left(g_i W_i \sigma_{i-1}(\Phi_{i-1}(\boldsymbol{\theta}, X)) \left(\sigma_{i-1}(g_{i-1} \Phi_{i-1}(\boldsymbol{\theta}, X))\right)^{-1} \ , \ g_i b_i \right)_{i=1}^L \ , \ X \right)$$

or more simply as: $W_i \mapsto g_i W_i \sigma_{i-1}(\Phi_{i-1}) \left(\sigma_{i-1}(g_{i-1} \Phi_{i-1})\right)^{-1}$.

Lemma D.5. *The map α has the following properties:*

1. *For $g \in G$ and $(\boldsymbol{\theta}, X)$ in the non-degenerate locus, we have the following relation between the extensions of the partial feedforward functions, for $i = 1, \ldots, L$:*

$$\widetilde{F}_i(\alpha(g, \boldsymbol{\theta}, X)) = \sigma_i \left(g_i W_i \widetilde{F}_{i-1}(\boldsymbol{\theta}, X) + g_i b_i \right)$$

   and the following relation between the extensions of the feedforward functions:

$$\widetilde{F}(\alpha(g, \boldsymbol{\theta}, X)) = \widetilde{F}(\boldsymbol{\theta}, X).$$

2. *The map α defines a group action. That is:*

$$\alpha(\mathrm{id}, \boldsymbol{\theta}, X) = (\mathbf{W}, X) \qquad \alpha(g, \alpha(h, \boldsymbol{\theta}, X) = \alpha(gh, \mathbf{W}, X)$$

   for any $g, h \in G$ and any $(\boldsymbol{\theta}, X)$ in the non-degenerate locus.

Remark In the notation with Φ_i, the claim regarding the partial feedforward functions becomes: $\Phi_i(\alpha(g, \boldsymbol{\theta}, X)) = g_i \Phi_i(\boldsymbol{\theta}, X)$.

Proof. We prove the claim regarding the partial feedforward functions by induction on i. The base step is $i = 1$, and we have:

$$\widetilde{F}_1(\alpha(g, \boldsymbol{\theta}, X)) = \sigma_1 \left(g_1 W_1 X + g_1 b_1\right) = \sigma_1 \left(g_1 W_1 \widetilde{F}_0(\boldsymbol{\theta}, X) + g_1 b_1 \right)$$

For the induction step, we compute:

$$\begin{aligned} &\widetilde{F}_i(\alpha(g, \boldsymbol{\theta}, X)) = \\ &\quad = \sigma_i \left(g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X) \sigma_{i-1}(g_{i-1} W_{i-1} \tilde{F}_{i-2}(\boldsymbol{\theta}, X) + g_{i-1} b_{i-1})^{-1} \widetilde{F}_{i-1}(\alpha(g, \boldsymbol{\theta}, X)) + g_i b_i \right) \\ &\quad = \sigma_i \left(g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X) + g_i b_i \right) \end{aligned}$$

where the cancellation occurs due to the induction hypothesis for $i - 1$. For the claim regarding the feedforward functions, we use the $i = L$ case and compute:

$$\begin{aligned} \widetilde{F}(\alpha(g, \boldsymbol{\theta}, X)) = \widetilde{F}_L(\alpha(g, \boldsymbol{\theta}, X)) &= \sigma_L \left(g_L W_L \widetilde{F}_{L-1}(\boldsymbol{\theta}, X) + g_L b_L \right) \\ &= \sigma_L \left(W_L \widetilde{F}_{L-1}(\boldsymbol{\theta}, X) + b_L \right) = \widetilde{F}_L(g, \boldsymbol{\theta}, X) = \widetilde{F}(g, \boldsymbol{\theta}, X) \end{aligned}$$

where we invoke the fact that g_L is the identity. For the third claim, we fist show that the unit axiom holds. Take $d = \mathrm{id}$ be the identity of G, which has g_i equal to the identity matrix $\mathrm{id}_h \in \mathrm{GL}_h$ for all $i = 1, \ldots, L$. Then $g_1 W_1 = W_1$, and, for $i = 2, \ldots, L$, we have:

$$\begin{aligned}
&g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\sigma_{i-1}(g_{i-1}W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + g_{i-1}b_{i-1})^{-1} \\
&= W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\sigma_{i-1}(W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + b_{i-1})^{-1} \\
&= W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)(\tilde{F}_{i-1}(\boldsymbol{\theta}, X))^{-1} = W_i
\end{aligned}$$

The biases are all unchanged: $g_i b_i = b_i$ for all i. The input data is also unchanged. Thus, $g = \mathrm{id}$ acts as the identity and the unit axiom is satisfied. For the multiplication axiom, let $g, h \in (\mathrm{GL}_h)^{L-1}$. We compute the i-th weight matrix of $\alpha(g, \alpha(h, \boldsymbol{\theta}, X))$. Fixing X, set:

$$W_i^{(h)} = h_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\sigma_{i-1}(h_{i-1}W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + h_{i-1}b_{i-1})^{-1}$$

and $b_{i-1}^{(h)} = h_{i-1}b_{i-1}$. We have:

$$\begin{aligned}
(W_i^{(h)})^{(g)} &= g_i W_i^{(h)} \tilde{F}_{i-1}(\alpha(h, \boldsymbol{\theta}, X))\sigma_{i-1}(g_{i-1}W_{i-1}^{(h)}\tilde{F}_{i-2}(\alpha(h, \boldsymbol{\theta}, X)) + g_{i-1}b_{i-1}^{(h)})^{-1} \\
&= g_i W_i^{(h)} \tilde{F}_{i-1}(\alpha(h, \boldsymbol{\theta}, X))\sigma_{i-1}(g_{i-1}W_{i-1}^{(h)}\tilde{F}_{i-2}(\alpha(h, \boldsymbol{\theta}, X)) + g_{i-1}h_{i-1}b_{i-1})^{-1}
\end{aligned}$$

The last term of $W_i^{(h)}$ cancels with $\tilde{F}_{i-1}(\alpha(h, \boldsymbol{\theta}, X))$ by the first claim of the lemma, and similarly for $i - 1$. Hence the last expression reduces to:

$$g_i h_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\sigma_{i-1}(g_{i-1}h_{i-1}W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + g_{i-1}h_{i-1}b_{i-1})^{-1}$$

This is precisely $W_i^{(gh)}$, so the weights of $\alpha(g, \alpha(h, \boldsymbol{\theta}, X))$ and $\alpha(gh, \boldsymbol{\theta}, X)$ match. The biases also match, as they are $g_i h_i b_i$ in both cases. Hence α defines a group action. □

We note the this non-linear action reduces to a linear action in special cases. Specifically, for each $i = 1, \ldots, L-1$, choose a subgroup $G_i \subseteq \mathrm{GL}_h$ and a representation $\pi_i : G_i \to \mathrm{GL}_h$. Recall that if $\sigma_i(g_i z) = \pi_i(g_i)\sigma(z)$ for all $z \in \mathbb{R}^h$ and $i = 1, \ldots, L$, then we have a linear action of $G = G_1 \times \cdots \times G_{L_1}$ on the parameter space $\mathbf{Param}$ given by:

$$g \cdot W_i = g_i W_i \pi_{i-1}(g_{i-1})^{-1} \qquad g \cdot b_i = g_i b_i \tag{37}$$

where, as usual, g_0 and g_L are the identity.

Lemma D.6. *Suppose G_i and π_i are as above. If $\sigma_i(g_i z) = \pi_i(g_i)\sigma(z)$ for all $z \in \mathbb{R}^h$ and $i = 1, \ldots, L$, then the action given by α is given by the above formula, and hence is linear.*

Proof. For each $i = 2, \ldots, L$, we have:

$$\begin{aligned}
&g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\sigma_{i-1}(g_{i-1}W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + g_{i-1}b_{i-1})^{-1} \\
&= g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\left(\pi_{i-1}(g_{i-1})\sigma_{i-1}(W_{i-1}\tilde{F}_{i-2}(\boldsymbol{\theta}, X) + b_{i-1})\right)^{-1} \\
&= g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\left(\pi_{i-1}(g_{i-1})\tilde{F}_{i-1}(\boldsymbol{\theta}, X)\right)^{-1} \\
&= g_i W_i \tilde{F}_{i-1}(\boldsymbol{\theta}, X)\left(\tilde{F}_{i-1}(\boldsymbol{\theta}, X)\right)^{-1}\pi_i(g_{i-1}^{-1}) \\
&= g_i W_i \pi_i(g_{i-1}^{-1})
\end{aligned}$$

The case $i = 1$ is easy, and the biases also transform appropriately. Note that the action occurs solely on the parameter space without any dependence on the data. □

D.4 Infinitesimal action

Lemma D.7. *The value of the infinitesimal action vector field of $M \in (\mathfrak{gl}_h)^{L-1}$ at a point $(\boldsymbol{\theta}, X)$ in the non-degenerate locus is given by:*

$$\overline{M}_{(\boldsymbol{\theta}, X)} = \left((*, M_i b_i)_{i=1}^{L}, 0\right)$$

where $*$ is given by:

$$M_i \circ U - U \circ d\sigma_{\Phi_{i-1}(\boldsymbol{\theta}, X)}(M_{i-1}\Phi_{i-1}(\boldsymbol{\theta}, X)) \circ (\sigma(\Phi_{i-1}(\boldsymbol{\theta}, X)))^{-1}$$

D.5 LINEAR SYMMETRIES: MULTILAYER CASE

We now state a more general version of the results of the previous section by combining equivariances across all layers of a multilayer network. Specifically, consider a feedforward fully-connected neural network with widths $\mathbf{n} = (n_0, \dots, n_L)$, so that the parameters space consists of tuples of matrices $\boldsymbol{\theta} = (W_i \in \mathbb{R}^{n_i \times n_{i-1}})_{i=1}^L$. For each layer $1 < i < L$, let G_i be a subgroup of GL_{n_i}, and let $\pi_i : G_i \to \mathrm{GL}_{n_i}(\mathbb{R})$ be a representation (in many cases, we take $\pi_i(g) = g$). Define an action of $G = G_1 \times \cdots \times G_L$ on $\mathbf{Param}$ via

$$\forall g = (g_1, \dots, g_L) \in G, \qquad g \cdot W_i = g_{i+1} W_i \pi_i(g_i^{-1}) \tag{38}$$

For each i, the representation π_i induces a Lie algebra representation $d(\pi_i) : \mathfrak{g}_i \to \mathfrak{gl}_{n_i}$, as described above. The infinitesimal action of the Lie algebra $\mathfrak{g} = \mathfrak{g}_1 \times \cdots \times \mathfrak{g}_L$ induced by 38 is given by:

$$\forall M = (M_1, \dots, M_L) \in \mathfrak{g}, \qquad g \cdot M_i = M_{i+1} W_i - W_i d(\pi_i)(M_i) \tag{39}$$

The proof of the first part of the following Proposition proceeds by induction, where the key computation is that of from Proposition 1.6. The second part relies on Corollary 1.3.

Proposition D.8. *Suppose that, for each $i = 1, \dots, L$, the activation σ_i intertwines the two actions of G_i, that is, $\sigma_i(g_i z_i) = \pi_i(g_i)\sigma(z_i)$ for all $g_i \in G_i$, $z_i \in \mathbb{R}^{n_i}$. Then:*

1. *(Combined equivariance of activations) The action of $G = G_1 \times \cdots \times G_L$ defined in 38 is a symmetry of the parameter space.*
2. *(Infinitesimal equivariant action) The action of $\mathfrak{g} = \mathfrak{g}_1 \times \cdots \times \mathfrak{g}_L$ defined in 39 satisfies $\langle \nabla_\theta \mathcal{L}, M \cdot \theta \rangle$ for all $\theta \in \mathbf{Param}$ and all $M \in \mathfrak{g}$.*

The above proposition is stated for feedforward networks, but can easily be adopted to more general settings, such as networks with skip connections and quiver neural networks.

E USING NONLINEARLY TRANSFORMED WEIGHTS ON DIFFERENT INPUTS

The transformation $U' = c(g^{-1}, gVX')$ depends on the input data batch X via $Z = VX$ Ideally, we would like to find transformed weights U', V' which could be used in place of the original U, V for any input data. Therefore, we want to understand under what circumstances a $U'(Z')$ computed using a given batch of input data $Z' = VX'$ would also work on a different batch $Z = VX$, meaning

$$\begin{aligned} g \cdot (U, Z') &= U'(Z'), gZ' \\ U\sigma(Z) &\stackrel{?}{=} U'(Z')\sigma(gZ) \\ &= U\sigma(Z')\sigma(gZ')^{-1}\sigma(gZ) \\ \sigma(Z) &\stackrel{?}{=} \sigma(Z')\sigma(gZ')^{-1}\sigma(gZ) \end{aligned} \tag{40}$$

In general, (40) is not satisfied by arbitrary Z, Z' and σ. But batches of input can be related to each other, like being samples from an underlying distribution. We will now explore whether, under reasonable assumptions, (40) can be approximately satisfied. We will use three ingredients for this. First, we will assume σ is Lipschitz continuous, meaning $|\sigma(a) - \sigma(b)| \leq \eta |a - b|$, allowing us to relate the differences in the output of σ to differences of inputs. Second ingredient is an assumption about the statistics of Z' being representative of the underlying distribution. Lastly, we will use the Cauchy-Schwartz inequality (CSI) in multiple occasions. Let us first review the matrix for of the CSI.

Cauchy-Schwartz inequality (CSI) for matrices. For real vectors $u, v \in \mathbb{R}^h$ with the Euclidean inner product $\langle u, v \rangle = u^T v \in \mathbb{R}$, the CSI states that

$$(u^T v)^2 \leq \|u\|^2 \|v\|^2 \tag{41}$$

u and v can also be rows and column of two matrices. Let U_i^j denote row i and column j of U. We have

$$\left(\sum_j U_i^j V_j^k \right)^2 = \left([UV]_i^k \right)^2 \leq \|U_i\|^2 \|V^k\|^2 \tag{42}$$

where U_i means row i and V^k means column k. Summing (42) over i and k we get

$$\sum_{i,k} \left([UV]_i^k\right)^2 = \|UV\|^2 \leq \|U\|^2\|V\|^2 \tag{43}$$

which is the matrix form of CSI.

Lipschitz continuous nonlinearity. Assume σ is Lipschitz continuous meaning for $a, b \in \mathbb{R}$ we have

$$|\sigma(a) - \sigma(b)| \leq \eta|a - b| \tag{44}$$

Define the action of σ on matrices to be elementwise, meaning $\sigma(A)_i^j = \sigma(A_i^j)$. Then the Lipschitz condition can be extended to the Frobenius norms as follows. For $A, B \in \mathbb{R}^{m\times n}$ we have

$$\begin{aligned} &\sum_{ij} |\sigma(A_i^j) - \sigma(B_i^j)|^2 \leq \eta^2 \sum_{ij} |A_i^j - B_i^j|^2 \\ \Rightarrow \quad &\|\sigma(A) - \sigma(B)\|^2 \leq \eta^2\|A - B\|^2 \end{aligned} \tag{45}$$

For parts of the derivation we also need the second assumption that $\sigma(0) = 0$. This results in

$$\sigma(0) = 0 \qquad \Rightarrow \qquad \|\sigma(A)\| \leq \eta\|A\|^2 \tag{46}$$

Representative batches We will now show that if if the samples (columns) in the new input Z are close to samples in Z' used to compute $U'(Z')$, then the equivariance condition is satisfied and $U'(Z')$ can be used as weights for batch Z as well. Assume that the input batch $Z' = VX' \in \mathbb{R}^{h\times b'}$ has samples which are close to the samples in the new input $Z \in \mathbb{R}^{h\times b}$. This means that for each column $Z^k \in \mathbb{R}^h$, there exists a column $Z'^{\mu(k)} \in \mathbb{R}^h$ close to it, meaning ($\mathbb{Z}_b = \{1, \ldots, b\}$)

$$\forall k\mathbb{Z}_b \in, \exists\mu(k) \in \mathbb{Z}_{b'} : \qquad \left\|Z^k - Z'^{\mu(k)}\right\| \leq \lambda \tag{47}$$

for some constant $\lambda > 0$ Here $\mu : \mathbb{Z}_b \to \mathbb{Z}_{b'}$ is a map. We can capture the mapping μ into a matrix $P \in \{0, 1\}^{b\times b'}$ as

$$\begin{aligned} P_j^k &\equiv \delta_j^{\mu(k)} & Z'^{\mu(k)} &= [Z'P]^k \\ \|Z - Z'P\|^2 &\leq \lambda^2 b, & \sum_j P_j^k &= 1 \quad \text{(permutation).} \end{aligned} \tag{48}$$

Note that because σ is element-wise and $\sum_k P_j^k = 1$ (i.e. each Z^k is mapped to a single $Z'^{\mu(k)}$) we have

$$\sigma(gZ'P) = \sigma(gZ')P. \tag{49}$$

Using (48) in the Lipschitz bound (45) we get

$$\begin{aligned} &\left\|\sigma\left(gZ^k\right) - \sigma\left(gZ'^{\mu(k)}\right)\right\| \leq \eta\left\|gZ^k - gZ'^{\mu(k)}\right\| \\ &\Rightarrow \|\sigma(gZ) - \sigma\left(gZ'\right)P\| \leq \eta\|gZ - gZ'P\| \end{aligned} \tag{50}$$

Using CSI (43) and (48) yields

$$\begin{aligned} \|gZ - gZ'P\| &\leq \|g\|\|Z - Z'P\| \leq \lambda\sqrt{b}\|g\| \\ \Rightarrow \|\sigma(gZ) - \sigma\left(gZ'\right)P\| &\leq \eta\lambda\sqrt{b}\|g\| \end{aligned} \tag{51}$$

Finally, we can use this result to address (40). We want to check if $U'(Z')\sigma(gZ)$ is approximately the same as $U\sigma(Z)$. We will now show that it is approximately $U\sigma(Z')P = U\sigma(Z) + O(\eta\lambda)$ (from (51)). To do so, we look at the difference between these two term

$$\begin{aligned} U'(Z') &= U\sigma(Z')\sigma(gZ')^{-1} \\ U'(Z')\sigma(gZ) &= U\sigma(Z')\sigma(gZ')^{-1}\sigma(gZ) \\ U'(Z')\sigma(gZ) - U\sigma(Z')P &= U\left[\sigma(Z')\sigma(gZ')^{-1}\sigma(gZ) - \sigma(Z')P\right] \\ &= U\sigma(Z')\left[\sigma(gZ')^{-1}\sigma(gZ) - P\right] \\ &= U\sigma(Z')\sigma(gZ')^{-1}\left[\sigma(gZ) - \sigma(gZ')P\right] \end{aligned} \tag{52}$$

where we used $\sigma(gZ')^{-1}\sigma(gZ') = I_b$ (assuming $b \leq h$). (51) and CSI we have

$$\left\|\sigma(Z')\sigma(gZ')^{-1}\left[\sigma(gZ) - \sigma(gZ')P\right]\right\| \leq \eta\lambda\sqrt{b}\|g\| \left\|\sigma(Z')\sigma(gZ')^{-1}\right\| \tag{53}$$

Therefore we have

$$\begin{aligned} U'(Z')\sigma(gZ) &= U\sigma(Z')\sigma(gZ')^{-1}\sigma(gZ) \\ &= U\sigma(Z')P + O(\lambda\eta) \\ &= U\sigma(Z) + O(\lambda\eta) \end{aligned} \tag{54}$$

where $\|\sigma(Z')P - \sigma(Z)\| \leq \eta\lambda\sqrt{b}$ from (51).

F JACOBIANS AND DIFFERENTIALS

Let $U \subset \mathbb{R}^n$ be an open subset of Euclidean space $\mathbb{R}^n$, and let $F : U \to \mathbb{R}^m$ be a differentiable function. Let $F_1, \ldots, F_m : U \to \mathbb{R}$ be the components of F, so that $F(u) = (F_1(u), \ldots, F_m(u))$. The Jacobian of F, also know as differential of F, at $u \in U$ is the following matrix of partial derivatives evaluated at u:

$$dF_u = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}\Big|_u & \frac{\partial F_1}{\partial x_2}\Big|_u & \cdots & \frac{\partial F_1}{\partial x_n}\Big|_u \\ \frac{\partial F_2}{\partial x_1}\Big|_u & \frac{\partial F_2}{\partial x_2}\Big|_u & \cdots & \frac{\partial F_2}{\partial x_n}\Big|_u \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}\Big|_u & \frac{\partial F_m}{\partial x_2}\Big|_u & \cdots & \frac{\partial F_m}{\partial x_n}\Big|_u \end{bmatrix}$$

The differential dF_u defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, that is, an element of $\mathbb{R}^{m\times n}$. Observe that if F itself is linear, then, as matrices, $dF_u = F$ for all points $u \in U$. If $G : \mathbb{R}^m \to \mathbb{R}^p$ is another differentiable map, then the chain rule implies that, for all $u \in \mathbb{R}^n$, we have:

$$d(G \circ F)_u = dG_{F(u)} \circ dF_u.$$

Case of $m = 1$ In the special case the $m = 1$, the differential is a $1 \times n$ row vector, and the *gradient* $\nabla_u F$ of F at $u \in \mathbb{R}^n$ is defined as the transpose of the Jacobian dF_u:

$$\nabla_v F = (dF_u)^T = \left[\tfrac{\partial F}{\partial x_1}\Big|_u \quad \cdots \quad \tfrac{\partial F}{\partial x_n}\Big|_u\right]^T = \left(\frac{\partial F}{\partial x_i}\Big|_u\right)_{i=1}^n \in \mathbb{R}^n$$

Manifolds More generally, let M and N be smooth manifolds, and suppose $F : M \to N$ is a smooth map. The differential of F at $m \in M$ is a linear map between the tangent spaces:

$$dF_m : T_m M \to T_{F(m)} N$$

The map dF_m is computed in local coordinate charts as the Jacobian of partial derivatives. If $G : N \to L$ is another smooth map, then the chain rule becomes $d(G \circ F)_m = dG_{F(m)} \circ dF_m$, for any $m \in M$.

Matrix case Suppose $L : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a differentiable function. In this case, we regard the Jacobian at $W \in \mathbb{R}^{m\times n}$ as an $n \times m$ matrix:

$$dL_W = \begin{bmatrix} \frac{\partial L}{\partial w_{11}}\Big|_W & \frac{\partial L}{\partial w_{21}}\Big|_W & \cdots & \frac{\partial L}{\partial w_{m1}}\Big|_W \\ \frac{\partial L}{\partial w_{12}}\Big|_W & \frac{\partial L}{\partial w_{22}}\Big|_W & \cdots & \frac{\partial L}{\partial w_{m2}}\Big|_W \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial w_{1n}}\Big|_W & \frac{\partial L}{\partial w_{2n}}\Big|_W & \cdots & \frac{\partial L}{\partial w_{mn}}\Big|_W \end{bmatrix} \in \mathbb{R}^{m\times n}$$

where w_{ij} are the matrix coordinates. If $F : \mathbb{R} \to \mathbb{R}^{m\times n}$ is a differentiable function, we regard its Jacobian at $s \in \mathbb{R}$ as a $m \times n$ matrix:

$$dF_t = \begin{bmatrix} \left.\frac{dF_{11}}{dt}\right|_s & \left.\frac{dF_{12}}{dt}\right|_s & \cdots & \left.\frac{dF_{1n}}{dt}\right|_s \\ \left.\frac{dF_{21}}{dt}\right|_s & \left.\frac{dF_{22}}{dt}\right|_s & \cdots & \left.\frac{dF_{2n}}{dt}\right|_s \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{dF_{m1}}{dt}\right|_s & \left.\frac{dF_{m2}}{dt}\right|_s & \cdots & \left.\frac{dF_{mn}}{dt}\right|_s \end{bmatrix} \in \mathbb{R}^{n\times m}$$

where $F_{ij} : \mathbb{R} \to \mathbb{R}$ are the coordinates of F. Then the chain rule becomes:

$$\left.\frac{d}{dt}\right|_s (L \circ F) = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\left.\frac{\partial L}{\partial w_{ij}}\right|_{F(s)} \left.\frac{dF_{ij}}{dt}\right|_s \right) = \mathrm{Tr}(dL_{F(s)} \cdot dF_s).$$

In other words, the derivative of the composition $L \circ F$ at $s \in \mathbb{R}$ is the trace of the product of the matrices $dL_{F(s)} \in \mathbb{R}^{n\times m}$ and $dF_s \in \mathbb{R}^{m\times n}$.

F.1 Conserved quantities

Let $\mathcal{V} = \mathbb{R}^d$ be the standard vector space of dimension d. Suppose $L : \mathcal{V} \to \mathbb{R}$ is a differentiable function.

Lemma F.1. *Let $Q : \mathcal{V} \to \mathbb{R}$ be a differentiable function. The following are equivalent:*

1. *The function Q is constant on any gradient flow curve for L.*
2. *We have $\frac{d}{dt}(Q \circ \gamma) = 0$ for any gradient flow curve γ.*
3. *For any $v \in \mathcal{V}$, the gradients of Q and L at v are orthogonal: $\langle \nabla_v L, \nabla_v Q\rangle$.*

Proof. The implications $(1) \Leftrightarrow (2)$ are immediate. To show the equivalence of the third and second statements, let $v \in \mathcal{V}$ and let $\gamma : (a, b) \to \mathcal{V}$ be a gradient flow curve such that $0 \in (a, b)$ and $\gamma(0) = v$ (such a curve always exists). We compute:

$$\langle \nabla_v Q, \nabla_v \mathcal{L}\rangle = dQ_{\gamma(0)} \circ \nabla_v \mathcal{L} = dQ_{\gamma(0)} \circ d\gamma_0 = d(Q \circ \gamma)_0 = \frac{d}{dt}(Q \circ \gamma),$$

where $d\gamma_0 = \dot\gamma(0) = \nabla_v L$ since γ is a gradient flow curve, and we use the chain rule in the third equality. □

We note that, if $f : \mathbb{R} \to \mathbb{R}$ is any function, and Q is a conserved quantity, the $f \circ Q$ is also a conserved quantity. Additionally, any linear combination of conserved quantities is again a conserved quantity. Let $\mathrm{Conserv}(\mathcal{V}, L)$ denote the vector space of conserved quantities for the gradient flow of $L : \mathcal{V} \to \mathbb{R}$. For any $v \in \mathcal{V}$, there a map:

$$\nabla_v : \mathrm{Conserv}(\mathcal{V}, L) \to T_v\mathcal{V} = \mathbb{R}^d, \qquad Q \mapsto \nabla_v Q$$

taking a conserved quantity to the value of its gradient at v. By the third part of Lemma F.1, the map ∇_v is valued in the kernel of the differential dL_v.

F.2 Linear group action

Let G be a subgroup of the general linear group $\mathrm{GL}_d(\mathbb{R})$. Thus, there is a linear action of G on $\mathcal{V} = \mathbb{R}^d$. Suppose the function L is invariant for the action of G, that is,

$$L(g \cdot v) = L(v) \qquad \forall v \in \mathcal{V} \quad \forall g \in G.$$

Let $\mathfrak{g} = \mathrm{Lie}(G)$ be the Lie algebra of G, which is a Lie subalgebra of $\mathfrak{gl}_d = \mathbb{R}^{d\times d}$. The *infinitesimal action* of $\mathfrak{g}$ on $\mathcal{V}$ is given by $\mathfrak{g} \times V \to V$, taking (M, v) to Mv.

Lemma F.2. *Let $L : \mathcal{V} \to \mathbb{R}$ be a G-invariant function, and let $M \in \mathfrak{g}$.*

1. *For any $v \in \mathcal{V}$, the gradient of L and the infinitesimal action of M are orthogonal:*
$$\langle \nabla_v L, Mv \rangle = 0.$$
2. *Suppose $\gamma : (a, b) \to \mathcal{V}$ is a gradient flow curve for L. Then:*
$$(\dot{\gamma}(t))^T M\gamma(t) = 0 \qquad \forall t \in (a, b)$$
3. *Suppose M^T belongs to $\mathfrak{g}$. Then the function*
$$Q_M : \mathcal{V} \to \mathbb{R}, \qquad v \mapsto v^T M v$$
is a conserved quantity for the gradient flow of L.

Proof. For the first claim, observe that the invariance of L implies that the following diagram commutes:

$$\begin{tikzcd} G \arrow[r, "\mathrm{inc}"] \arrow[d] & \mathbb{R}^{d\times d} \arrow[r, "\mathrm{ev}_v"] & \mathbb{R}^d \arrow[d, "L"] \\ \{v\} \arrow[rr, "L"'] & & \mathbb{R} \end{tikzcd}$$

where $\mathrm{inc} : G \to \mathbb{R}^{d\times d}$ is the inclusion (which passes through the inclusion of G in to $\mathrm{GL}_d(\mathbb{R})$), $\mathrm{ev}_v : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ be the evaluation map at v, and the left vertical map is the constant map at $\{v\}$. Indeed, the clockwise composition is $g \mapsto L(gv)$, which is equal to the constant map at $g \mapsto L(v)$. The chain rule implies that taking Jacobians at the identity 1 of G results in the following commutative diagram:

$$\begin{tikzcd} \mathfrak{g} \arrow[r, "d\mathrm{inc}_1"] \arrow[d] & \mathbb{R}^{d\times d} \arrow[r, "d(\mathrm{ev}_x)_{\mathrm{id}_d}"] & \mathbb{R}^d \arrow[d, "dL_v"] \\ 0 \arrow[rr] & & \mathbb{R} \end{tikzcd}$$

where $\mathfrak{g}$ is the Lie algebra of G, identified with the tangent space of G at the identity; the tangent space of the vector space $\mathbb{R}^d$ at v is canonically identified with $\mathbb{R}^d$; and the the zero appears because the tangent space of a single point is zero. The derivative of the inclusion map is the inclusion $\mathfrak{g} \hookrightarrow \mathfrak{gl}_d = \mathbb{R}^{d\times d}$, while the derivative the evaluation map is itself as it is a linear map. Hence, for $M \in \mathfrak{g}$, we have:

$$0 = dL_v \circ d(\mathrm{ev}_v)_{\mathrm{id}_d} \circ d\mathrm{inc}_1(M) = dL_v \circ \mathrm{ev}_v(M) = dL_v(Mv) = \langle \nabla_v L, Mv \rangle.$$

The first claim follows. The second claim is consequence of the first claim, together with the definition of a gradient flow curve. For the third claim, we take the derivative of the composition of Q_M with a gradient flow curve γ:

$$\begin{aligned} \frac{d}{dt}(Q_M \circ \gamma) &= (\dot{\gamma}(t))^T (M + M^T)\gamma(t) \\ &= \langle \dot{\gamma}(t), (M + M^T)\gamma(t) \rangle \\ &= -\langle \nabla_{\gamma(t)} L, (M + M^T)\gamma(t) \rangle \\ &= -\langle \nabla_{\gamma(t)} L, M\gamma(t) \rangle - \langle \nabla_{\gamma(t)} L, M^T\gamma(t) \rangle \end{aligned}$$

Both terms in the last expression are constantly zero by the second claim. Hence Q_M is constant on any gradient flow curve, and so it is a conserved quantity. □

We summarize some of the results and constructions of this section diagrammatically. Let $\mathfrak{g}^{\mathrm{sym}}$ denote the vector space of symmetric matrices in $\mathfrak{g}$ (this is not a Lie subalgebra in general). Observe that $\mathfrak{g} \cap \mathfrak{g}^T$ is the set of all $M \in \mathfrak{g}$ such that $M^T \in \mathfrak{g}$. Let $\mathrm{Infin}_G(\mathcal{V}, L)$ denote the vector space of

infinitesimal-action conserved quantities for the gradient flow of the G-invariant function $L : \mathcal{V} \to \mathbb{R}$. We have:

$$\begin{tikzcd}
\mathfrak{g} \cap \mathfrak{g}^T \arrow[dr, "M \mapsto Q_M"] \arrow[d, bend right] & & \\
& \mathrm{Infin}_G(\mathcal{V}, L) \arrow[r, hook] & \mathrm{Conserv}(\mathcal{V}, L) \\
\mathfrak{g}^{\mathrm{sym}} \arrow[u, hook, bend right] \arrow[ur, "\simeq"'] & &
\end{tikzcd}$$

where the map $\mathfrak{g} \cap \mathfrak{g}^T \to \mathfrak{g}^{\mathrm{sym}}$ takes M to its symmetric part $\frac{1}{2}\left(M + M^T\right)$, while the map $\mathfrak{g}^{\mathrm{sym}} \hookrightarrow \mathfrak{g} \cap \mathfrak{g}^T$ is the natural inclusion. We note that $\mathfrak{g} \cap \mathfrak{g}^T$ is the Lie algebra of the group $G \cap G^T$, while $\mathfrak{g}^{\mathrm{sym}}$ is in general not a Lie algebra. By definition, the vector space $\mathrm{Infin}_G(\mathcal{V}, L)$ is the image of the map $M \to Q_M$ defined on $\mathfrak{g} \cap \mathfrak{g}^T$. It is straightforward to verify the following result:

Corollary F.3. *The map $M \to Q_M$ establishes an isomorphism of vector spaces: $\mathfrak{g}^{\mathrm{sym}} \xrightarrow{\sim} \mathrm{Infin}_G(\mathcal{V}, L)$.*

F.3 Anti-symmetric case

Suppose $M \in \mathfrak{g}$ is anti-symmetric, so $M = -M^T$. Let $\gamma : (a, b) \to \mathbb{R}^d$ be a gradient flow curve. Write γ in coordinates as $\gamma = (\gamma_1, \ldots, \gamma_d)$. Lemma F.2 implies that $(\dot{\gamma}(t))^T M\gamma = 0$. Hence we have:

$$0 = \sum_{i<j} m_{ij}\left(\dot{\gamma}_i\gamma_j - \gamma_i\dot{\gamma}_j\right) = \sum_{i<j} m_{ij}\gamma_i^2\left(\frac{\gamma_j}{\gamma_i}\right)' = \sum_{i<j} m_{ij} r_{ij}^2 \dot{\theta}_{ij}$$

where $r_{ij} = r_{ij}(t)$ is equal to $\sqrt{\gamma_i(t)^2 - \gamma_j(t)^2}$ and $\theta_{ij} = \theta_{ij}(t)$ is the angle between the i-th coordinate axis and the ray from the origin to the projection of $\gamma(t)$ to the (i, j)-plane. One verifies the last equality using the definition of θ_{ij} in terms of the arctangent of the quotient γ_j/γ_i. We see that $(r_{ij}(t), \theta_{ij}(t))$ are the polar coordinates for the point $(\gamma_i(t), \gamma_j(t)) \in \mathbb{R}^2$.

Case $d = 2$. Then $M = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}$ for some nonzero $a \in \mathbb{R}$, and so:

$$0 = \dot{\gamma}^T M\gamma = a\dot{\gamma}_1(t)\gamma_2(t) - a\gamma_1(t)\dot{\gamma}_2(t) = ar(t)^2\dot{\theta}(t)$$

where $r(t)$ and $\theta(t)$ are polar coordinates. Setting the final expression equal to zero, we obtain that $\theta(t)$ is constant along any flow line $\gamma(t)$ that begins away from the origin.

Case $d = 3$. Then $M = \begin{bmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{bmatrix}$ for some $a, b, c \in \mathbb{R}$, and so:

$$0 = \dot{\gamma}^T M\gamma = ar_{12}^2\dot{\theta}_{12} + br_{13}^2\dot{\theta}_{13} + cr_{23}^2\dot{\theta}_{23}$$

F.4 Non-linear actions

Let G be a Lie group with Lie algebra $\mathfrak{g} = \mathrm{Lie}(G)$, identified with the tangent space at the identity 1 of G. Suppose $G \times \mathcal{V} \to \mathcal{V}$ is an action of G on $\mathcal{V} = \mathbb{R}^d$. For now, this action is not necessarily linear. For $v \in \mathcal{V}$, set:

$$a_v : G \to \mathcal{V}; \qquad g \mapsto g \cdot v$$

Recall that vector fields on the vector space $\mathcal{V}$ are equivalent to sections of the tangent bundle $T\mathcal{V}$ of $\mathcal{V}$. We write $\Gamma(\mathcal{V}, T\mathcal{V})$ for this space of sections.

Definition F.4. *The infinitesimal action of G on $\mathcal{V}$ is the following map of Lie algebras:*

$$\mathfrak{g} \to \Gamma(\mathcal{V}, T\mathcal{V}); \qquad \xi \mapsto \overline{\xi} = [v \mapsto d(a_v)_1(\xi)]$$

Hence the value $\overline{\xi}_v$ of the vector field $\overline{\xi}$ at $v \in \mathcal{V}$ is given by first computing the differential of the map $g \mapsto g^{-1}v$ at the identity of G, and then evaluating at $U \in \mathfrak{g}$.

Suppose $L : \mathcal{V} = \mathbb{R}^d \to \mathbb{R}$ is a differentiable, G-invariant function, so that $\mathcal{L}(g \cdot v) = L(v)$ for all $v \in \mathcal{V}$ and $g \in G$.

Proposition F.5. *Fix $v \in \mathcal{V}$. The infinitesimal action of any $\xi \in \mathfrak{g}$ at v is orthogonal to the gradient of L at v:*

$$\langle \overline{\xi}_v, \nabla_v L \rangle = 0$$

Proof. We compute:

$$\langle \overline{\xi}_v, \nabla_v L \rangle = dL_v\left(\overline{\xi}_v\right) = dL_v \circ d(a_v)_1(\xi) = d(L \circ a_v)_1(U)$$

The fact that L is G-invariant implies that the composition $L \circ a_v : G \to \mathbb{R}$ factors through the constant map $g \mapsto v$, so its differential is constantly zero. □

Two relevant commutative diagrams are displayed below. The left encodes the fact that $\mathcal{L}$ is G-invariant. The right is obtained from the left when restricting to $G \times \{v\}$ (and possibly adding an inverse).

$$\begin{tikzcd} G \times \mathcal{V} \arrow[r, "\text{action}"] \arrow[d, "\text{project}"'] & \mathcal{V} \arrow[d, "L"] \\ V \arrow[r, "L"'] & \mathbb{R} \end{tikzcd} \qquad \begin{tikzcd} G \arrow[r, "a_v"] \arrow[d, "\text{const}"'] & \mathcal{V} \arrow[d, "L"] \\ \{v\} \arrow[r, "L"'] & \mathbb{R} \end{tikzcd}$$

We briefly explain the connection to the linear case. If the action of G on $\mathcal{V}$ is linear, then it is given by a group homomorphism $\rho : G \to \mathrm{GL}_d(\mathbb{R}) = \mathrm{GL}(\mathcal{V})$. The differential of ρ at the identity is a map of Lie algebras $d\rho_1 : \mathfrak{g} \to \mathfrak{gl}_d$. The differential of a_v at the identity is the map $\mathfrak{g} \to T_v\mathcal{V} = \mathbb{R}^d$ taking ξ to $d\rho_1(\xi)$ evaluated at v. Hence we recover Lemma F.2 from Proposition F.5.

Note that Lemma **??** generalizes to Riemannian manifolds. Namely, if M is a Riemannian manifold with an action of G, then the gradient vector field of an invariant function is pointwise orthogonal to the infinitesimal action of $\mathfrak{g}$.

G Neural networks: linear actions

G.1 General Set-up

Consider a neural network with L layers, input dimension $n_0 = n$, output dimension $n_L = m$, no bias vectors, and hidden dimensions given by $n_1, \ldots, n_{L-1}$. For convenience, we group the dimensions into a tuple $\mathbf{n} = (n_0, n_1, \ldots, n_L)$. The parameter space is given by:

$$\mathbf{Param}(\mathbf{n}) = \mathbb{R}^{n_1 \times n} \times \mathbb{R}^{n_2 \times n_1} \times \cdots \times \mathbb{R}^{n_{L-1} \times n_{L-2}} \times \mathbb{R}^{m \times n_{L-1}} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_L}.$$

We write an element therein as a pair $(\mathbf{W}, \mathbf{b})$ of tuples $\mathbf{W} = (W_1, \ldots, W_L)$ and $\mathbf{b} = (b_1, \ldots, b_L)$, so that W_i is a $n_i \times n_{i-1}$ matrix and b_i is a vector in $\mathbb{R}^{n_i}$ for $i = 1, \ldots, L$. When $\mathbf{n}$ is clear from context, we write simply $\mathbf{Param}$ for the parameter space. Fix a piecewise differentiable function $\sigma_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ for each $i = 1, \ldots, L$. The activations can be pointwise (as is conventionally the case), but are not necessarily so. The feedforward function

$$F = F_{(\mathbf{W},\mathbf{b})} : \mathbb{R}^n \to \mathbb{R}^m$$

corresponding to parameters $(\mathbf{W}, \mathbf{b}) \in \mathbf{Param}$ with activations σ_i is defined in the usual recursive way. To be explicit, we define the partial feedforward function $F_{(\mathbf{W},\mathbf{b}),i} = F_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$ to be the map taking $x \in \mathbb{R}^n$ to $\sigma_i(W_i F_{i-1}(x) + b_i)$, for $i = 1, \ldots, L$, with $F_0 = \mathrm{id}_{\mathbb{R}^{n_0}}$. Then the feedforward function is $F = F_L$. If $y \in \mathbb{R}^m$ is a training output value corresponding to $x \in \mathbb{R}^n$, we obtain a loss function

$$\mathcal{L}_{x,y} : \mathbf{Param} \to \mathbb{R}$$

by taking $(\mathbf{W}, \mathbf{b})$ to $\|y - F_{(\mathbf{W},\mathbf{b})}(x)\|^2$.

The above constructions generalize to multiple samples. Specifically, instead of $x \in \mathbb{R}^{n_0}$ and $y \in \mathbb{R}^{n_L}$, one has matrices $X \in \mathbb{R}^{n_o \times k}$ and $Y \in \mathbb{R}^{n_L \times k}$ whose columns are the k samples. Additionally, one uses the Frobenius norm of $n_L \times k$ matrices to compute the loss function. The i-th partial feedforward function is $F_{\mathbf{W},i} : \mathbb{R}^{n_0 \times k} \to \mathbb{R}^{n_i \times k}$, and we have the feedforward function $F_{\mathbf{W}} : \mathbb{R}^{n_0 \times k} \to \mathbb{R}^{n_L \times k}$. (We use the same notation as in the case $k = 1$; the number of samples will be understood from context.)

Example G.1. Consider the case $L = 2$ with no biases and no output activation. The dimension vector is $(n_0, n_1, n_2) = (n, h, m)$, so the parameter space is $\mathbf{Param}(n, h, m) = \mathbb{R}^{h \times n} \times \mathbb{R}^{m \times h}$. Corresponding to parameters $\mathbf{W}$ and a data point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, we have the functions:

$$F_{V,U} : \mathbb{R}^n \to \mathbb{R}^m, \quad x \mapsto U\sigma(Vx); \qquad \mathcal{L}_{x,y} : \mathbf{Param} \to \mathbb{R}, \quad (V, U) \mapsto \|y - U\sigma(Vx)\|^2$$

G.2 The linear hidden symmetry group action

Consider a neural network with L layers and dimensions $\mathbf{n}$. The *hidden symmetry group* corresponding to dimensions $\mathbf{n}$ is defined as the following product of general linear groups:

$$\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} = \mathrm{GL}_{n_1} \times \cdots \times \mathrm{GL}_{n_{L-1}}$$

An element is a tuple of invertible matrices $\mathbf{g} = (g_1, \ldots, g_{L-1})$, where $g_i \in \mathrm{GL}_{n_i}$. Consider the action of the hidden symmetry group on the parameter space given by

$$\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} \circlearrowright \mathbf{Param}(\mathbf{n}) \qquad \mathbf{g} \cdot (\mathbf{W}, \mathbf{b}) = \left(g_i W_i g_{i-1}^{-1} \ , \ g_i b_i\right)_{i=1}^{L}. \tag{55}$$

where $g_0 = \mathrm{id}_{n_0}$ and $g_L = \mathrm{id}_{n_L}$. The Lie algebra of $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}}$ is

$$\mathfrak{gl}_{\mathbf{n}^{\text{hidden}}} = \mathfrak{gl}_{n_1} \times \cdots \times \mathfrak{gl}_{n_{L_1}},$$

and the infinitesimal action of the tuple $M \in \mathfrak{gl}_{\mathbf{n}^{\text{hidden}}}$ is given by:

$$\mathfrak{gl}_{\mathbf{n}^{\text{hidden}}} \circlearrowright \mathbf{Param}(\mathbf{n}) \qquad \mathbf{M} \cdot (\mathbf{W}, \mathbf{b}) \mapsto (M_i W_i - W_i M_{i-1} \ , \ M_i b_i)_{i=1}^{L}$$

where we set M_0 and M_L to be the zero matrices.

Example G.2. In the case $L = 2$, with dimension vector $\mathbf{n} = (n, h, m)$. The hidden symmetry group is GL_h with Lie algebra $\mathfrak{gl}_h$. The action of the group and the infinitesimal action of the Lie algebra are given by:

$$\mathrm{GL}_h \circlearrowright \mathbf{Param}(n, h, m) = \mathbb{R}^{h \times n} \times \mathbb{R}^{m \times h} \qquad g \cdot (V, U) = \left(gV, Ug^{-1}\right)$$

$$\mathfrak{gl}_h \circlearrowright \mathbf{Param}(n, h, m) = \mathbb{R}^{h \times n} \times \mathbb{R}^{m \times h} \qquad M \cdot (V, U) = (MV, -UM)$$

G.3 Conserved quantities for linear networks

Suppose a neural network with L layers has the identity activation $\sigma_i = \mathrm{id}_{n_i}$ in each layer, so that the resulting network is linear. Then it is straightforward to verify that the networks with parameters $\mathbf{g} \cdot (\mathbf{W}, \mathbf{b})$ and $(\mathbf{W}, \mathbf{b})$ have the same feedforward function. Consequently, the loss is invariant for the group action: its value the original and transformed parameters is the same for any choice of training data. (As we will see below, for more sophisticated activations, one needs to restrict to a subgroup of the hidden symmetry group to achieve such invariance.)

In what follows, we set $\odot$ to the Hadamard product of matrices, defined by entrywise multiplication. We also set $\tau(M)$ to be the sum of all entries in a matrix M.

Suppose $\mathbf{M} \in \mathfrak{gl}_{\mathbf{n}^{\text{hidden}}}$ is such that $M_i \in \mathfrak{gl}_{n_i}$ is symmetric for each i. The the corresponding conserved quantity is given by:

$$Q_{\mathbf{M}}(\mathbf{W}, \mathbf{b}) = \sum_{i=1}^{L-1} \left(\tau(W_i \odot M_i W_i) + \tau(b_i \odot M_i b_i) - \tau(W_{i+1} \odot W_{i+1} M_i)\right)$$

We examine these conserved quantities in the following convenient basis for the space of symmetric matrices in $\mathfrak{gl}_{\mathbf{n}^{\text{hidden}}}$. For $j = 1, \ldots, L$ and $\{k, \ell\} \subseteq \{1, \ldots, n_j\}$, set:

$$E^{(j)}_{\{k,\ell\}} := \begin{cases} E^{(n_j)}_{kk} & \text{if } k = \ell \\ \frac{1}{2}\left(E^{(n_j)}_{k\ell} + E^{(n_j)}_{\ell k}\right) & \text{if } k \neq \ell \end{cases}$$

where $E^{(n_j)}_{k\ell}$ is the elementary $n_j \times n_j$ matrix with the entry in the k-th row and ℓ-th column equal to one, and all other entries equal to zero. Then one computes:

$$Q_{E^{(j)}_{\{k,\ell\}}}(\mathbf{W}, \mathbf{b}) = b^{(j)}_k b^{(j)}_\ell + \sum_{t=1}^{n_{j-1}} w^{(j)}_{kt} w^{(j)}_{\ell t} - \sum_{r=1}^{n_{j+1}} w^{(j+1)}_{rk} w^{(j+1)}_{r\ell}$$

In other words, we take the sum of the following three terms: the product of the k-th and ℓ-th entries of the bias vector b_j, the dot product of the k-th and ℓ-th rows of W_j, and the dot product of the k-th and ℓ-th columns of W_{j+1}.

G.4 Conserved quantities for ReLU networks

The pointwise ReLU activation commutes with positive rescaling, so we consider the subgroup of the hidden symmetry group consisting of tuples of diagonal matrices with positive diagonal entries, that is:

$$G = \{\mathbf{g} \in \mathrm{GL}_{\mathbf{n}^{\text{hidden}}} \mid g_i = \mathrm{Diag}(s_1, \ldots, s_{n_i}) \ , \ s_j > 0\}$$

This subgroup is isomorphic to the product $(\mathbb{R}_{>0})^{\sum_{i=1}^{L-1} n_i}$. Its Lie algebra is spanned by the elements $E^{(j)}_{kk}$ defined above, for $j = 1, \ldots, L-1$ and $k = 1, \ldots, n_j$. The corresponding conserved quantity is:

$$Q_{E^{(j)}_{kk}}(\mathbf{W}, \mathbf{b}) = \left(b^{(j)}_k\right)^2 + \sum_{t=1}^{n_{j-1}} \left(w^{(j)}_{kt}\right)^2 - \sum_{r=1}^{n_{j+1}} \left(w^{(j+1)}_{rk}\right)^2$$

In other words, we take the sum of the following three terms: the square of the k-th entry of the bias vector b_j, the norm of the k-th row of W_j, and the norm of the k-th column of W_{j+1}.

G.5 Conserved angular momentum for radial rescaling networks

Suppose each σ_i is a radial rescaling activation $\sigma_i(z) = \lambda_i(|z|)z$, where $\lambda_i : \mathbb{R} \to \mathbb{R}$ is the rescaling factor. Each such activation commutes with orthogonal transformations, so we consider the subgroup of the hidden symmetry group consisting of tuples of orthogonal matrices:

$$G = \{\mathbf{g} \in \mathrm{GL}_{\mathbf{n}^{\text{hidden}}} \mid g_i g_i^T = \mathrm{id}_{n_i} \text{ for all } i\}$$

The Lie algebra of this subgroup consists only of anti-symmetric matrices, and so there are no infinitesimal-action conserved quantities. However, given an anti-symmetric matrix $M_i \in \mathfrak{gl}_{n_i}$ for each i, any gradient flow curve satisfies the following differential equation (encoding conservation of angular momentum):

$$\sum_{i=1}^{L-1} \left(\tau(\dot{W}_i \odot M_i W_i) + \tau(\dot{b}_i \odot M_i b_i) - \tau(\dot{W}_{i+1} \odot W_{i+1} M_i)\right) = 0$$

A basis for the space of anti-symmetric matrices in $\mathfrak{gl}_{\mathbf{n}^{\text{hidden}}}$ is given by:

$$E^{(j)}_{k<\ell} := E^{(n_j)}_{k\ell} - E^{(n_j)}_{\ell k}$$

where $j = 1, \ldots, L-1$, and $k, \ell \in \{1, \ldots, n_j\}$ satisfy $k < \ell$. The differential equation corresponding to $E^{(j)}_{k\leq\ell}$ is given by:

$$r^2_{b_j;k,\ell} \dot{\theta}_{b_j;k,\ell} + \sum_{t=1}^{n_{j-1}} r^2_{W_j;ks,\ell s} \dot{\theta}_{W_j;ks,\ell s} + \sum_{r=1}^{n_{j+1}} r^2_{W_{j+1};rk,r\ell} \dot{\theta}_{W_{j+1};rk,r\ell} = 0$$

where $(r_{b_j;k\ell}, \theta_{b_j;k,\ell})$ are the polar coordinates of the image of b_j under the projection $\mathbb{R}^{n_j} \to \mathbb{R}^2$ which selects only the k-th and ℓ-th coordinates. Similarly, for any pair matrix entries we have a projection $\mathbb{R}^{n_j \times n_{j-1}} \to \mathbb{R}^2$ and can take the polar coordinates of the image of W_j under this projection.

G.6 TWO-LAYER NETWORKS

To illustrate the main ideas of the previous sections, we now turn our attention to examples of two-layer neural network with input and output dimensions both equal to one, hidden dimension equal to two, and no bias vectors. The parameter space is $\mathbf{Param} = \mathbb{R}^{2\times 1} \times \mathbb{R}^{1\times 2}$, with elements written as a pair of matrices: $(U, V) = \left(\begin{bmatrix} u_1 & u_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right)$. For an activation $\sigma : \mathbb{R}^2 \to \mathbb{R}^2$ (not necessarily pointwise), the loss function corresponding to a data point $(x, y) \in \mathbb{R} \times \mathbb{R}$ is given by:

$$\mathcal{L} = \mathcal{L}_{x,y} : \mathbf{Param} \to \mathbb{R}, \qquad (U, V) \mapsto \|y - U\sigma(Vx)\|^2$$

Consider the action of $\mathrm{GL}_2(\mathbb{R})$ on $\mathcal{V}$ given by:

$$\mathrm{GL}_2(\mathbb{R}) \times \mathbf{Param} \to \mathbf{Param}, \qquad (g, U, V) \mapsto g \cdot (U, V) = \left(Ug^{-1}, gV\right).$$

The Lie algebra $\mathfrak{gl}_2$ of $\mathrm{GL}_2(\mathbb{R})$ consists of all two-by-two matrices. In what follows, we consider several different choices of activations. For each, there is a corresponding subgroup G of $\mathrm{GL}_2(\mathbb{R})$ that leaves the loss function $\mathcal{L}$ invariant for any choice of training data. One obtains conserved quantities corresponding to symmetric elements of the Lie algebra of G, if any such exist.

G.6.1 RELU NETWORKS

The loss function with the ReLU activation $\sigma = \sigma_{\mathrm{ReLU}}$ applied pointwise is given by: $\mathcal{L}(U, V) = \|y - u_1\sigma(v_1 x) - u_2\sigma(v_2 x)\|^2$. Recall that the ReLU activation commutes with positive rescaling, so $\sigma(sx) = s\sigma(x)$ for any positive $s \in \mathbb{R}_{>0}$ and any $x \in \mathbb{R}$. This implies that $\mathcal{L}$ is invariant for the action of the subgroup G of $\mathrm{GL}_2(\mathbb{R})$ consisting of diagonal matrices with positive diagonal entries:

$$G = \left\{ \begin{bmatrix} g_1 & 0 \\ 0 & g_2 \end{bmatrix} \in \mathrm{GL}_2(\mathbb{R}) \mid g_1 \text{ and } g_2 \text{ are positive.} \right\}$$

That is: $\mathcal{L}(g \cdot (U, V)) = \mathcal{L}(U, V)$ for any $g \in G$. The Lie algebra of G is the two-dimensional space of diagonal matrices in $\mathfrak{gl}_2$ (with not necessarily positive diagonal entries). In other words, $\mathfrak{g}$ is spanned by the matrices $E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. Using Lemma **??**, one computes the conserved quantities corresponding to these elements as:

$$Q_{E_{11}}(U, V) = v_1^2 - u_1^2 \qquad Q_{E_{22}}(U, V) = v_2^2 - u_2^2$$

Hence there is a two-dimensional space of conserved quantities coming from the infinitesimal action.

G.6.2 RADIAL RESCALING NETWORKS

Consider the case when $\sigma : \mathbb{R}^2 \to \mathbb{R}^2$ is a *radial rescaling* activation, that is, one of the form $\sigma(a) = \lambda(|z|)z$ for $z \in \mathbb{R}^2$, where $\lambda : \mathbb{R} \to \mathbb{R}$ is a function applied to the norm of v. Hence, every vector z is rescaled by a scalar factor that depends only on the norm of z. Radial rescaling activations have been studied in Ganev et al. (2022), and it is well-known that they commute with orthogonal transformations. Consequently, $\mathcal{L}$ is invariant for the action of the subgroup $G = O(2)$ of $\mathrm{GL}_2(\mathbb{R})$ consisting of orthogonal matrices:

$$G = O(2) = \left\{ g \in \mathrm{GL}_2(\mathbb{R}) \mid g^T g = \mathrm{id} \right\}$$

The Lie algebra of G consists of anti-symmetric matrices in $\mathfrak{gl}_2$, and contains no non-zero symmetric matrices. Hence, we do not obtain any conserved quantities from the infinitesimal action in this case. However, using the element $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \in \mathfrak{g}$, Lemma **??** does imply that the certain differential equation holds along any gradient flow curve:

$$r_U^2 \dot{\theta}_U + r_V^2 \dot{\theta}_V = 0 \tag{56}$$

where (r_U, θ_U) are the polar coordinates of $(u_1, u_2) \in \mathbb{R}^2$, and similarly for (r_V, θ_V). Note that the left-hand side of Equation 56 is a function of t; so if $\gamma : (a, b) \to \mathcal{V}$ is a gradient flow curve, then a more precise version of the equation is $\left(r_U^2 \circ \gamma\right)(t) \cdot (\theta_U \circ \gamma)'(t) + \left(r_V^2 \circ \gamma\right)(t) \cdot (\theta_V \circ \gamma)'(t) = 0$ for all t.

G.6.3 LINEAR NETWORKS

Finally, we consider the simpler case of a linear network, where σ is the identity activation. In this case, the action of every element of $\mathrm{GL}_2(\mathbb{R})$ leaves the loss function invariant; hence the symmetry group is the full general linear group, and the Lie algebra is all of $\mathfrak{gl}_2$. The space of symmetric matrices in $\mathfrak{gl}_2$ is spanned by the matrices E_{11} and E_{22} from above, in addition to the matrix $E_{(1,2)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The corresponding conserved quantities are:

$$Q_{E_{11}}(U,V) = v_1^2 - u_1^2 \qquad Q_{E_{22}}(U,V) = v_2^2 - u_2^2 \qquad Q_{E_{(1,2)}}(U,V) = v_1 v_2 - u_1 u_2$$

Thus, we obtain a three-dimensional space of conserved quantities. Since $\mathrm{GL}_2(\mathbb{R})$ also contains the orthogonal group $O(2)$, Equation 56 holds along any gradient flow curve.

H NEURAL NETWORKS: NON-LINEAR ACTIONS

H.1 SET-UP

We require some additional notation. Taking the parameters into account, we form the following extension of the feedforward function F:

$$\tilde{F} : \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^m, \qquad (\mathbf{W}, x) \mapsto F_{\mathbf{W}}(x)$$

One can also define the extension of the partial feedforward function $\widetilde{F}_i : \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^{n_i}$ as $(\mathbf{W}, x) \mapsto F_{\mathbf{W},i}(x)$. Grouping the parameters, input data, and output value together, we obtain a map:

$$\widetilde{\mathcal{L}} : \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

taking $(\mathbf{W}, x, y)$ to $\mathcal{L}_{x,y}(\mathbf{W}) = \|y - F_{\mathbf{W}}(x)\|^2$. We summarize these maps in the following diagram, which holds for any choice of parameters $\mathbf{W}$, input data x, and output value y:

$$\begin{tikzcd}
\mathbb{R}^n \arrow[r, "F_{\mathbf{W}}=\text{Feedforward}"] \arrow[d, hook] & \mathbb{R}^m \arrow[dr, bend left, "\|y-(-)\|^2"] & \\
\mathbf{Param} \times \mathbb{R}^n \arrow[ur, "\widetilde{F}"] \arrow[r, hook] & \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \arrow[r, "\widetilde{\mathcal{L}}"] & \mathbb{R} \\
\mathbf{Param} \arrow[u, hook] \arrow[urr, bend right, "\mathcal{L}_{x,y}=\text{Loss}"'] & &
\end{tikzcd}$$

The inclusions $\mathbb{R}^n \hookrightarrow \mathbf{Param} \times \mathbb{R}^n$, $\mathbf{Param} \hookrightarrow \mathbf{Param} \times \mathbb{R}^n$, and $\mathbf{Param} \times \mathbb{R}^n \hookrightarrow \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m$ are given by $x \mapsto (\mathbf{W}, x)$, $\mathbf{W} \mapsto (\mathbf{W}, x)$, and $(\mathbf{W}, x) \mapsto (\mathbf{W}, x, y)$. The top right map takes $\hat{y} \in \mathbb{R}^m$ to $\|y - \hat{y}\|^2$.

Example H.1. Consider the case $L = 2$. The dimension vector is $(n_0, n_1, n_L) = (n, h, m)$, so the parameter space is:

$$\mathbf{Param}(n, n, m) = \mathbb{R}^{h \times n} \times \mathbb{R}^{m \times h}$$

Corresponding to parameters $\mathbf{W} = (V, U) \in \mathbf{Param}(n, h, m)$ and a data point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, we have the functions:

$$\begin{aligned}
F_{V,U} &: \mathbb{R}^n \to \mathbb{R}^m & x &\mapsto U\sigma(Vx) \\
\widetilde{F} &: \mathbf{Param} \times \mathbb{R}^n \to \mathbb{R}^m, & (V, U, x) &\mapsto U\sigma(Vx) \\
\mathcal{L}_{x,y} &: \mathbf{Param} \to \mathbb{R}, & (V, U) &\mapsto \|y - U\sigma(Vx)\|^2 \\
\widetilde{\mathcal{L}} &: \mathbf{Param} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, & (V, U, x, y) &\mapsto \|y - U\sigma(Vx)\|^2.
\end{aligned}$$

H.2 THE ACTION

Suppose $n_i = n$ for $i = 0, \ldots, L-2$ and $k = n$. In other words, all hidden dimensions are equal to the input dimension, which is also equal to the number of samples. The output dimension $m = n_L$

and the last hidden dimension n_{L-1} are not required to be equal to n. In this special case, we define the *non-degenerate locus* by requiring the partial feedforward matrices to be invertible:

$$\left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n\times n}\right)^\circ = \{(\mathbf{W}, X) \in \mathbf{Param}\times\mathbb{R}^{n\times n} \mid F_{\mathbf{W},i}(X) \in \mathrm{GL}_n \text{ for } i = 0, 1, \ldots, L-2\}$$

Furthermore, the hidden symmetry group is the product of $L-2$ copies of the general linear group GL_n and one copy of $\mathrm{GL}_{n_{L-1}}$, that is, $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} = (\mathrm{GL}_n)^{L-2} \times \mathrm{GL}_{n_{L-1}}$.. We write an element as $g = (g_i)_{i=1}^{L-1}$, and denote the identity element as id. Consider the following map:

$$\alpha : \mathrm{GL}_{\mathbf{n}^{\text{hidden}}} \times \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n\times n}\right)^\circ \to \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n\times n}\right)^\circ \tag{57}$$

$$(g, \mathbf{W}, X) \mapsto \left(\left(\sigma^{-1} g_i \sigma(W_i F_{i-1}(X)) \left(g_{i-1} F_{i-1}(X)\right)^{-1}\right)_{i=1}^{L}, X\right) \tag{58}$$

where, for convenience, we set $g_0 = \mathrm{id}_n$ and $g_L = \mathrm{id}_{n_L}$. In words, for $i = 1, \ldots, L-1$, we first take the product of the $n \times n$ matrices W_i and $F_{i-1}(X)$, and then apply the non-linear composition $\sigma^{-1} \circ g_i \circ \sigma$ to this product. Note that, as usual, σ and σ^{-1} are applied elementwise. This produces an $n \times n$ matrix, which we then multiply on the right with the inverse of the $n \times n$ matrix $g_{i-1} F_{i-1}(X)$. In the case $i = L$, we have simply $W_L g_{L-1}^{-1}$. The input data X is unchanged.

Lemma H.1. *The map α has the following properties:*

1. *For $g \in \mathrm{GL}_{\mathbf{n}^{\text{hidden}}}$ and $(\mathbf{W}, X)$ in the non-degenerate locus, we have the following relation between the extensions of the partial feedforward functions, for $i = 1, \ldots, L-1$:*

$$\widetilde{F}_i(\alpha(g, \mathbf{W}, X)) = g_i \widetilde{F}_i(\mathbf{W}, X),$$

and the following relation between the extensions of the feedforward functions:

$$\widetilde{F}(\alpha(g, \mathbf{W}, X)) = \widetilde{F}(\mathbf{W}, X).$$

2. *The map α defines a group action. That is:*

$$\alpha(\mathrm{id}, \mathbf{W}, X) = (\mathbf{W}, X) \qquad \alpha(g, \alpha(h, \mathbf{W}, X) = \alpha(gh, \mathbf{W}, X)$$

for any $g, h \in \mathrm{GL}_{\mathbf{n}^{\text{hidden}}}$ and any $(\mathbf{W}, X)$ in the non-degenerate locus.

Sketch of proof. One proves the first claim with a straightforward inductive argument. The unit axiom of the third claim is immediate. Verification of the multiplication axiom relies on the first claim. □

Observe that the action above reduces to the linear action when σ is the identity (or any multiple of the identity).

Example H.2. Consider the case $L = 2$. The dimension vector is (n, h, m), so the parameter space is:

$$\mathbf{Param}(n, h, m) = \mathbb{R}^{h\times n} \times \mathbb{R}^{m\times h}$$

We write and element as (V, U) instead of (W_1, W_2). Set $k = n$. The non-degenerate locus is:

$$\left(\mathbf{Param}(n, h, m) \times \mathbb{R}^{n\times n}\right)^\circ = \{(V, U, X) \in \mathbb{R}^{h\times n} \times \mathbb{R}^{m\times h} \times \mathbb{R}^{n\times n} \mid X \in \mathrm{GL}_n\}$$

The hidden symmetry group is $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}} = \mathrm{GL}_h$ and the action is given by:

$$\mathrm{GL}_h \circlearrowright \left(\mathbf{Param}(n, h, m) \times \mathbb{R}^{n\times n}\right)^\circ \qquad g \cdot (V, U, X) = \left((\sigma^{-1} g \sigma(VX))X^{-1}, Ug^{-1}, X\right)$$

H.3 Failure of an action in the non-square case

The version of the non-degenerate locus for a general dimension vector is defined as follows:

Definition H.2. *Let $\mathbf{n}$ be a dimension vector for a neural network with L layers and let k be a positive integer. The corresponding* non-degenerate locus *is defined as:*

$$\left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n_0\times k}\right)^\circ = \{(\mathbf{W}, X) \in \mathbf{Param}\times\mathbb{R}^{n_0\times k} \mid \mathrm{rank}(F_{\mathbf{W},i}(X)) = \min(n_i, k) \text{ for } i = 0, 1, \ldots, L-2\}$$

If $(\mathbf{W}, X)$ belongs to the non-degenerate locus, then, for $i = 0, \dots L-2$, the matrix $F_{\mathbf{W},i}(X) \in \mathbb{R}^{n_i \times k}$ is full rank, so has a left inverse if and only if $k \leq n_i$ and right inverse if and only if $k \geq n_i$. These inverses are unique if and only if $n_i = k$, in which case the right and left inverses coincide. Consider the map:

$$\beta : \mathrm{GL}_{\mathbf{n}^{\text{hidden}}} \times \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n_0 \times k}\right)^\circ \to \left(\mathbf{Param}(\mathbf{n}) \times \mathbb{R}^{n_0 \times k}\right)^\circ \tag{59}$$

$$(g, \mathbf{W}, X) \mapsto \left(\left(\sigma^{-1} g_i \sigma(W_i F_{i-1}(X)) F_{i-1}(X)^\dagger g_{i-1}^{-1}\right)_{i=1}^{L} \ , \ X\right) \tag{60}$$

where $F_{i-1}(X)^\dagger$ denotes a left inverse of $F_{i=1}(X)$ if $k \leq n_i$ and a right inverse if $k \geq n_i$.

Proposition H.3. *With no assumptions on the bijective map σ, the map β defines an action of $\mathrm{GL}_{\mathbf{n}^{\text{hidden}}}$ on the non-degenerate locus preserving the map $\widetilde{F}$ if an only if $n_i = k$ for $i = 0, 1, \dots, L-2$.*

Sketch of proof. Let $i \in \{0, \dots, L-2\}$ The unit axiom requires that $F_i(X)F_i(X)^\dagger$ is the identity $n_i \times n_i$ matrix; hence $F_i(X)$ has a right inverse, so $k \geq n_i$. On the other hand, both the preservation of $\widetilde{F}$ and the multiplication axiom require that $F_i(X)^\dagger F_i(X)$ be the identity $k \times k$ matrix, so $k \leq n_i$. It follows that $k = n_i$. □

Example H.3. Consider a two-layer neural network with dimension vector $(1, 1, 1)$ and $k = 2$ samples. The non-degenerate locus is:

$$\left(\mathbf{Param}(1,1,1) \times \mathbb{R}^{1\times 2}\right)^\circ = \{(v, u, X = [x_1 \ x_2]) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{1\times 2} \mid x_1 \text{ and } x_2 \text{ are not both zero}\}$$

A right inverse for $X = [x_1 \ \ x_2]$ is given by $X^\dagger = \frac{1}{x_1^2 + x_2^2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. We show that the transformation under β of any $g \in \mathbb{R}^\times$ does not preserve the feedforward function. Indeed, given (v, u, X) in the non-degenerate locus, we have:

$$\widetilde{F}(v, u, X) = [u\sigma(vx_1) \ \ u\sigma(vx_2)]$$

For $u = v = 1 = x_1 = 1$ and $x_2 = 2$, we obtain $[\sigma(1) \ \sigma(2)]$. On the other hand, one computes that $\widetilde{F}(\beta(g, v, u, X))$ is equal to:

$$\left[ug^{-1}\sigma\left(\tfrac{x_1^2}{x1^2+x_2^2}\left(\sigma^{-1}g\sigma(vx_1)\right) + \tfrac{x_1x_2}{x1^2+x_2^2}\left(\sigma^{-1}g\sigma(vx_2)\right)\right) \quad ug^{-1}\sigma\left(\tfrac{x_1x_2}{x1^2+x_2^2}\left(\sigma^{-1}g\sigma(vx_1)\right) + \tfrac{x_2^2}{x1^2+x_2^2}\left(\sigma^{-1}g\sigma(vx_2)\right)\right)\right]$$

For $u = v = 1 = x_1 = 1$ and $x_2 = 2$, we obtain

$$\left[g^{-1}\sigma\left(\tfrac{1}{5}\left(\sigma^{-1}g\sigma(1)\right) + \tfrac{2}{5}\left(\sigma^{-1}g\sigma(2)\right)\right) \quad g^{-1}\sigma\left(\tfrac{1}{5}\left(\sigma^{-1}g\sigma(1)\right) + \tfrac{4}{5}\left(\sigma^{-1}g\sigma(2)\right)\right)\right]$$

I THE LAGRANGIAN APPROACH

I.1 LAGRANGIAN

We identify the tangent bundle TM with $\mathbb{R}^d \times \mathbb{R}^d$, writing elements as pairs (x, y). Given V as a above and $\lambda \in \mathbb{R}^\times$, define a Lagrangian[9]:

$$L_\lambda : \mathbb{R} \times TM \to \mathbb{R} \qquad (t, x, y) \mapsto e^{t/\lambda}\left(V(x) - \frac{\lambda}{2}|y|^2\right) \tag{61}$$

Observe that at $t = 0$ and $y = 0$, we recover V, that is, $L_\lambda(0, x, 0) = V(x)$. It is straightforward to show that a curve $f : (a, b) \to M$ satisfies the Euler–Lagrange equation for L_λ if and only if:

$$\lambda \ddot{f}(t) + \dot{f}(t) + \nabla_{f(t)} V = 0 \qquad \forall t \in ()a, b) \tag{62}$$

While Equation 62 reduces to Equation **??** when $\lambda = 0$, note that it does not make sense to set $\lambda = 0$ directly in the expression for the Lagrangian L_λ. We discuss some constrained uniqueness properties of this Lagrangian in an addendum to these notes. Orthogonal symmetries of V are also symmetries of the Lagrangian L_λ. More precisely, let $\rho^{-1}(O(d))$ be the subgroup of G consisting of all $g \in G$ such that the image $\rho(g)$ defines an orthogonal transformation of $\mathbb{R}^d$.

[9] An equivalent version replaces λ with $\gamma = 1/\lambda \in \mathbb{R}^\times$. A more general version replaces $|y|^2$ by a positive semi-definite symmetric bilinear form.

Lemma I.1. *Suppose $g \in \rho^{-1}(O(d))$. Then, for all $\lambda \neq 0$, and for all $(x, y) \in TM \simeq \mathbb{R}^d \times \mathbb{R}^d$, we have:*

$$L_\lambda(t\ ,\ \rho(g)x\ ,\ \rho(g)y) = L_\lambda(t\ ,\ x\ ,\ y).$$

I.2 Second-order gradient flow conserved quantities

For $U \in \mathfrak{g}$ and $\lambda \in \mathbb{R}^\times$, set:

$$\overline{Q}_{U,\lambda} : \mathbb{R} \times TM \to \mathbb{R} \qquad (t, x, y) \mapsto e^{t/\lambda} x^T \overline{U} y \tag{63}$$

Given $f : (a, b) \to \mathbb{R}^d$, we write $\overline{Q}_{U,\lambda} \circ \dot{f}$ for the map $t \mapsto \overline{Q}_{U,\lambda}(t, f(t), \dot{f}(t))$. The following lemma shows two facts. First, the function $\overline{Q}_{U,\lambda}$ vanishes on the path in TM defined by differentiating an integral curve for the gradient flow. Second, when $\overline{U}$ is anti-symmetric, the function $\overline{Q}_{U,\lambda}$ is a conserved charge for the Lagrangian L_λ.

Proposition I.2. *Let $U \in \mathfrak{g}$ and $\lambda \in \mathbb{R}^\times$. Let $f : (a, b) \to M$ be a smooth curve.*

1. *If f satisfies the gradient flow equation for V (Equation **??**), then:*

$$\overline{Q}_{U,\lambda} \circ \dot{f} = 0$$

2. *If f satisfies the Euler–Lagrange equation for L_λ (Equation 62), then:*

$$\frac{d}{dt}\left(\overline{Q}_{U,\lambda} \circ \dot{f}\right) = e^{t/\lambda} \dot{f}(t) \overline{U}^{\text{sym}} \dot{f}(t).$$

where $\overline{U}^{\text{sym}}$ is the symmetric part of the matrix $\overline{U} \in \mathfrak{gl}_d$.

Proof. The first claim is a direct consequence of Corollary **??** and the definition of $\overline{Q}_{U,\lambda}$. For the second claim, we have:

$$\frac{d}{dt}\left(\overline{Q}_{U,\lambda} \circ \dot{f}\right) = \frac{e^{t/\lambda}}{\lambda}\left(f(t)^T \overline{U} \dot{f}(t) + \lambda \dot{f}(t)^T \overline{U} \dot{f}(t) + \lambda f(t)^T \overline{U} \ddot{f}(t)\right) \tag{64}$$

An standard (elementary) calculation shows that $\dot{f}(t)^T \overline{U} \dot{f}(t) = \dot{f}(t) \overline{U}^{\text{sym}} \dot{f}(t)$. If f satisfies Equation 62, then we can write $\ddot{f}$ in terms of $\dot{f}$ and $\nabla_{f(t)} V$ to obtain:

$$\frac{d}{dt}\left(\overline{Q}_{U,\lambda} \circ \dot{f}\right) = \frac{e^{t/\lambda}}{\lambda}\left(f(t)^T \overline{U} \dot{f}(t) + \lambda \dot{f}(t) \overline{U}^{\text{sym}} \dot{f}(t) - f(t)^T \overline{U} \dot{f}(t) - f(t)^T \overline{U} \nabla_{f(t)} V\right)$$

The term $f(t)^T \overline{U} \dot{f}(t)$ cancels, while Proposition **??** implies that $f(t)^T \overline{U}\left(\nabla_{f(t)} V\right) = 0$. The second claim follows. □

Corollary I.3. *Suppose $\overline{U} \in \mathfrak{gl}_d$ is anti-symmetric. If f satisfies the Euler–Lagrange equation for L_λ (Equation 62), then:*

$$\frac{d}{dt}\left(\overline{Q}_{U,\lambda} \circ \dot{f}\right) = 0$$

Corollary I.3 is a special case of Noether's theorem, since, as noted above, orthogonal symmetries of V are also symmetries of L_λ. We also invoke the fact that the Lie algebra of the orthogonal group consists of anti-symmetric matrices, so that the Lie algebra of the subgroup $\rho^{-1}(O(d))$ of G can be identified the Lie subalgebra of $\mathfrak{g}$ consisting of those $U \in \mathfrak{g}$ such that $\overline{U}$ is skew-symmetric.

For clarity, we perform the computation of the proof of Proposition I.2 in alternative notation:

$$\begin{aligned}
\lambda \dot{Q} &= e^{t/\lambda}\left(x^T \overline{U} \dot{x} + \lambda \dot{x}^T \overline{U} \dot{x} + \lambda x^T \overline{U} \ddot{x}\right) \\
&= e^{t/\lambda}\left(x^T \overline{U} \dot{x} + \lambda \dot{x}^T \overline{U}^{\text{sym}} \dot{x} + \lambda x^T \overline{U}\left(-\frac{1}{\lambda}\dot{x} - \frac{1}{\lambda}\nabla_x V\right)\right) \\
&= e^{t/\lambda}\left(x^T \overline{U} \dot{x} + \lambda \dot{x}^T \overline{U}^{\text{sym}} \dot{x} - x^T \overline{U} \dot{x} - x^T \overline{U} \nabla_x V\right) \\
&= \lambda \dot{x}^T \overline{U}^{\text{sym}} \dot{x}
\end{aligned}$$

I.3 ADDENDUM: MORE ON THE LAGRANGIAN

I.3.1 CONSTRAINED UNIQUENESS

Conjecture I.4. *Consider functions of the form*

$$L : \mathbb{R} \times TM \to \mathbb{R} \qquad (t, x, y) \mapsto h(t)V(x) - g(t)|y|^2$$

for non-vanishing coefficient functions $g, h : \mathbb{R} \to \mathbb{R}^\times$. The Lagrangian L_λ is the unique such function that satisfies the following two conditions: (1) For all $x \in M$, we have $L(0, x, 0) = V(x)$. (2) The Euler–Lagrange equation for L is equivalent to Equation 62.

Proof. The two conditions imply the following identities:

$$h(0) = 1 \qquad 2g(t) = \lambda h(t) \qquad 2\dot{g}(t) = h(t)$$

From these, one concludes that $g(t) = \frac{\lambda}{2}e^{t/\lambda}$ and $h(t) = e^{t/\lambda}$. □

Remark Note that L_λ is not the unique Lagrangian that satisfies the following two conditions: (1) For all $x \in M$, we have $L_\lambda(0, x, 0) = V(x)$. (2) The Euler–Lagrange equation for L_λ limits to the gradient flow equation for V as $\lambda \to 0$.

I.3.2 SYMMETRIES

Any symmetry of V (not necessarily orthogonal) can be used to obtain an asymptotic symmetry of L_λ in the sense of first setting $t = 0$ and then taking $\lambda \to 0$. Specifically, for $g \in G$, we have:

$$L_\lambda(0\ ,\ \rho(g)x\ ,\ \rho(g)y) = V(\rho(g)x) + \frac{\lambda}{2}|\rho(g)y|^2 = V(x) - \frac{\lambda}{2}|\rho(g)y|^2$$

which limits to $V(x)$ as $\lambda \to 0$, just as $L_\lambda(0, x, y) = V(x) + \frac{\lambda}{2}|y|^2$ does.

I.3.3 LIMITING PROCEDURES

Let $F : \mathbb{R}^\times \times (a, b) \to \mathbb{R}^d$ be a smooth map. Writing $f_\lambda(t)$ for $F(\lambda, t)$, we see that F encodes a smooth family of curves f_λ indexed by $\lambda \in \mathbb{R}^\times$. We say that a curve $f : (a, b) \to \mathbb{R}^d$ is a limit of the family f_λ if $\lim_{\lambda\to 0} f_\lambda(t) = f(t)$ for every $t \in (a, b)$. For a given family, a limit may or may not exist; if it exists, it is unique.

Lemma I.5. *Suppose f_λ satisfies the Euler–Lagrange equation for L_λ (Equation 62), and suppose the limit $f = \lim_{\lambda\to 0} f_\lambda$ exists. Then f satisfies the gradient flow equation for V (Equation* **??***).*

Proof. If the limit exists, then $\dot{f} = \lim_{\lambda\to 0} \dot{f}_\lambda$, and similarly for the second derivative. The result now follows from the fact that the Euler–Lagrange equation for L_λ limits to the gradient flow equation for V as $\lambda \to 0$. □

Example I.1. Suppose $V : \mathbb{R}^d \to \mathbb{R}$ is given by $V(x) = \frac{1}{2}|x|^2$, so that $\nabla_x V = x$ for every $x \in \mathbb{R}^d$. Given $x \in \mathbb{R}^d$, set:

$$f_\lambda^{(x)} : \mathbb{R} \to \mathbb{R}^d \qquad t \mapsto \exp\left(-t\frac{1 - \sqrt{1 - 4\lambda}}{2\lambda}\right) x$$

Then $f_\lambda^{(x)}(0) = x$ and $f_\lambda^{(x)}$ satisfies the Euler–Lagrange equation for L_λ. The limit of $f_\lambda^{(x)}$ as $\lambda \to 0$ is given by:

$$f^{(x)} : \mathbb{R} \to \mathbb{R}^d \qquad t \mapsto e^{-t}x$$

i.e., the unique flow line for the reverse gradient vector field of V starting at x.

Remark Suppose each f_λ satisfies the Euler–Lagrange equation for L_λ, and that the limit $f = \lim_{\lambda\to 0} f_\lambda$ exists. By Corollary I.3, we have that

$$\frac{d}{dt}\left(\overline{Q}_{U,\lambda} \circ \dot{f}\right) = 0$$

for $\overline{U}$ antisymmetric. This implies that $\overline{Q}_{U,\lambda} \circ \dot{f} = K$ for some constant K. Using the definition of $\overline{Q}_{U,\lambda}$, we have $f_\lambda \bar{U} \dot{f}_\lambda = e^{-t/\lambda} K$. Taking the limit as $\lambda \to 0$, we obtain $f \bar{U} \dot{f} = 0$, and so

$$\overline{Q}_{U,\lambda} \circ \dot{f} = 0$$

for any λ. Thus we recover Part 1 of Proposition I.2 in the anti-symmetric case. We emphasize that we have proven this result by more elementary means already. Anyway, this proof only works for anti-symmetric $\overline{U}$, and for flows that are limits of solutions to the Euler–Lagrange equation.

J BACKGROUND: LAGRANGIANS AND NOETHER'S THEOREM

J.1 EULER-LAGRANGE EQUATION

The action $\mathcal{L}$ is defined as the integral of a Lagrangian L:

$$\mathcal{L} = \int dt L(x, \dot{x}, t) \tag{65}$$

From the stationary action principle, $\mathcal{L}$ is minimized. In other words, we want $\frac{\partial \mathcal{L}}{\partial x} = 0$. Consider a small change (δx) in x. The first-order Taylor expansion of $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} + \delta\mathcal{L} &= \int dt L\left(x + \delta x, \frac{d(x+\delta x)}{dt}\right) \\ &= \int dt \left[L(x,\dot{x}) + \delta x^\mu \frac{\partial L(x,\dot{x})}{\partial x^\mu} + \frac{d\delta x^\mu}{dt}\frac{\partial L(x,\dot{x})}{\partial \dot{x}^\mu} + O((\delta x)^2)\right] \end{aligned} \tag{66}$$

We can remove $\frac{d\delta x^\mu}{dt}$ by integration by parts:

$$\int dt \frac{d\delta x^\mu}{dt}\frac{\partial L(x,\dot{x})}{\partial \dot{x}^\mu} = \delta x^\mu \left.\frac{\partial L(x,\dot{x})}{\partial \dot{x}^\mu}\right|_i^f - \int dt \delta x^\mu \frac{d}{dt}\frac{\partial L}{\partial \dot{x}^\mu} \tag{67}$$

Then,

$$\delta\mathcal{L} = \int dt \delta x \delta L = \int dt \delta x^\mu \left[\frac{\partial L}{\partial x^\mu} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}^\mu}\right] + \delta x^\mu \left.\frac{\partial L(x,\dot{x})}{\partial \dot{x}^\mu}\right|_i^f. \tag{68}$$

Recall that we want $\delta\mathcal{L} = 0$. If we assume $\delta x = 0$ on the boundary, the second integral evaluates to 0 and we get the Euler-Lagrange equation:

$$\frac{\partial L(x,\dot{x})}{\partial x^\mu} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}^\mu} = 0. \tag{69}$$

J.2 FIRST-ORDER GRADIENT FLOW CANNOT BE DERIVED FROM VARIATIONAL PRINCIPLE

Consider the Lagrangian

$$L = \frac{1}{2}\dot{x}^i \varepsilon_{ij}^{-1} \dot{x}^j - \gamma V \tag{70}$$

To minimize this Lagrangian, we use the Euler-Lagrange equation,

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = -\nabla V - \varepsilon \ddot{x} = 0 \tag{71}$$

This equation does not contain first-order terms of x. Therefore, we cannot get the original gradient flow dynamics from the variational principle.

J.3 BREGMAN LAGRANGIAN

The Bregman Lagrangian (Wibisono & Wilson, 2015) provides a framework for modeling dynamics in neural networks with different learning rules. The form of this Lagrangian that we will be using is

$$L_{\alpha,\beta,\gamma}(x,\dot{x},t) = e^{\gamma(t)}\left(\eta(\dot{x},\dot{x}) - e^{\beta(t)} f(x)\right), \tag{72}$$

where $\eta : \mathbb{R}^d \otimes \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a degenerate Riemannian metric (positive semi-definite). In this paper, we consider the following version of the Bregman Lagrangian:

$$L = e^{\frac{t}{\gamma}} \left(\frac{1}{2} \dot{x}^i \varepsilon_{ij}^{-1} \dot{x}^j - \frac{1}{\gamma} V(x) \right) \tag{73}$$

where $\frac{1}{\gamma}$ is the friction constant, x is the trainable parameters in our optimization problem, and $V(x)$ is the loss function. For simplicity, we consider $\epsilon = I$ in the main text.

From the Euler-Lagrange equation (71), the equation of motion of this Lagrangian is

$$0 = \frac{\partial L}{\partial x} - \frac{d}{dt} \frac{\partial L}{\partial \dot{x}} = -e^{\frac{t}{\gamma}} \frac{1}{\gamma} \nabla V - \left(e^{\frac{t}{\gamma}} \varepsilon^{-1} \ddot{x} + \frac{1}{\gamma} e^{\frac{t}{\gamma}} \varepsilon^{-1} \dot{x} \right) = -e^{\frac{t}{\gamma}} \left(\frac{1}{\gamma} \nabla V + \varepsilon^{-1} \ddot{x} + \frac{1}{\gamma} \varepsilon^{-1} \dot{x} \right). \tag{74}$$

Simplifying the above equation, we have

$$\dot{x} = -\varepsilon \nabla V - \gamma \ddot{x}. \tag{75}$$

J.4 Relating γ to δt

We show that the coefficient γ in the Lagrangian defined in (26) is proportional to the time step δt in gradient descent.

The second order expansion of x is

$$x(t + \delta t) = x + \delta t \dot{x} + \frac{\delta t^2}{2} \ddot{x} \tag{76}$$

From this expansion, $\delta x = \delta t \dot{x} + \frac{\delta t^2}{2} \ddot{x}$. On the other hand,

$$\frac{\delta x}{\delta t} = -\varepsilon \frac{dV}{dx} \tag{77}$$

so $\delta x = -\delta t \varepsilon \frac{dV}{dx}$. Equating the δx obtained in the two methods, we have

$$\dot{x} = -\varepsilon \frac{dV}{dx} - \frac{\delta t}{2} \ddot{x} \tag{78}$$

Comparing this with the equation of motion derived from (26), we have

$$\gamma = \frac{\delta t}{2} \tag{79}$$

J.5 Noether's theorem

If we consider $x, \dot{x}$ that is a solution of the system, the first integral in (68) becomes 0, and the remaining term in $\delta \mathcal{L}$ is

$$\delta \mathcal{L} = \delta x^\mu \left. \frac{\partial L(x, \dot{x})}{\partial \dot{x}^\mu} \right|_i^f \tag{80}$$

From Stoke's theorem,

$$\delta \mathcal{L} = \int_i^f dt \partial_t \left(\delta x^\mu \frac{\partial L}{\partial \dot{x}^\mu} \right) \tag{81}$$

Setting $\delta \mathcal{L} = 0$ gives us the Noether's theorem:

$$\frac{d}{dt} \delta x^\mu \frac{\partial L}{\partial \dot{x}^\mu} = 0 \tag{82}$$

The conserved charge is $J = \delta x^\mu \frac{\partial L}{\partial \dot{x}^\mu}$. The term $\frac{\partial L}{\partial \dot{x}^\mu} = p_\mu$ is the conjugated momentum.

J.6 RELATING CONSERVED QUANTITY TO CONVERGENCE RATE

The Euler-Lagrange equation for our Lagrangian (26) gives the following dynamics:

$$\frac{dV}{dx} = -\varepsilon^{-1}\dot{x} - \frac{1}{\gamma}\varepsilon^{-1}\ddot{x} \tag{83}$$

When $\gamma \to \infty$,

$$\frac{dV}{dt} = (-\varepsilon^{-1}\dot{x})^T\dot{x} = -\dot{x}^T\varepsilon^{-1}\dot{x} \tag{84}$$

Then, we write the change in x in terms of the change in the coefficient of the basis:

$$x(t+\delta t) = gx = (I + \delta\theta^i T_i)x \tag{85}$$

$$\dot{x} = \frac{\delta\theta^i T_i x}{\delta t} = \dot{\theta}^i T_i x \tag{86}$$

Assume that $\varepsilon = I$. Substitute $\dot{x}$ back, we have

$$\begin{aligned}\frac{dV}{dt} &= -x^T T_i^T \dot{\theta}^i \varepsilon^{-1}\dot{\theta}^j T_j x \\ &= -\dot{\theta}^i\dot{\theta}^j x^T T_i^T T_j x\end{aligned} \tag{87}$$

Recall that the conserved quantity for the second-order gradient flow is

$$Q_i = e^{\gamma t}(\dot{Q}_i^{GF} + J_i) \tag{88}$$

where $\dot{Q}_i^{GF} = x^T S_i \dot{x} = x^T S_i \dot{\theta}^j T_j x$ is the time derivative of the conserved quantity for the first-order gradient flow, and $J_i = x^T A_i \dot{x} = x^T A_i \dot{\theta}^j T_j x$ is the angular momentum. Then, the components of Q_i appears in $\frac{dV}{dt}$ in the following way:

$$-\frac{dV}{dt} = [\text{nonsym}] + \dot{\theta}^i \dot{Q}_i^{GF} + \dot{\theta}^i J_i \tag{89}$$

J.7 CONSERVED QUANTITIES IN NEURAL NETWORKS

J.7.1 TWO-LAYER LINEAR NETWORK

In this section, we find the conserved quantity corresponding to the following Lagrangian:

$$\mathcal{L} = e^{\gamma t}\left(\frac{|\dot{U}|^2 + |\dot{V}|^2}{2} - \gamma L(UV)\right) \tag{90}$$

where the parameters are $U \in \mathbb{R}^{m\times h}$ and $V \in \mathbb{R}^{h\times n}$. For simplicity, we assumed that $\varepsilon = 1$.

Let T_i be an element of the Lie algebra of GL_h. Assume that for $g \in GL_h$, the transformation $U \to Ug^{-1}$, $V \to gV$ leaves L unchanged. When g is close to identity, we can use the first-order expansion of the group action $U(I - \eta T_i)$ and $(I + \eta T_i)V$, where $\eta \in \mathbb{R}$ is a small number. Putting U and V together in a matrix, we have

$$X = \begin{bmatrix} U \\ V^T \end{bmatrix}, \qquad g\cdot X = \begin{bmatrix} Ug^{-1} \\ (gV)^T \end{bmatrix} = \begin{bmatrix} Ug^{-1} \\ V^T g^T \end{bmatrix} = \begin{bmatrix} U(I - \delta\theta^i T_i) \\ V^T(I + \delta\theta^i T_i^T) \end{bmatrix} \tag{91}$$

Let S_i, A_i be the symmetric and antisymmetric parts of T_i, that is $T_i = S_i + A_i$. Then,

$$\begin{aligned}\delta X &= \begin{bmatrix} -U\delta\theta^i T_i \\ V^T\delta\theta^i T_i^T \end{bmatrix} = \begin{bmatrix} -U\delta\theta^i A_i \\ V^T\delta\theta^i A_i^T \end{bmatrix} + \begin{bmatrix} -U\delta\theta^i S_i \\ V^T\delta\theta^i S_i^T \end{bmatrix} \\ &= -\begin{bmatrix} U\delta\theta^i A_i \\ V^T\delta\theta^i A_i \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} U\delta\theta^i S_i \\ V^T\delta\theta^i S_i \end{bmatrix} = -\delta\theta^i X A_i - \delta\theta^i \sigma_z X S\end{aligned} \tag{92}$$

where $\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ is the Pauli matrix.

The conserved quantity is

$$\begin{aligned} J &= \delta X \cdot \frac{\partial \mathcal{L}}{\partial \dot{X}} \\ &= (-\delta\theta^i X A_i - \delta\theta^i \sigma_z X S) \cdot (e^{\gamma t} \dot{X}) \\ &= -\delta\theta^i e^{\gamma t} \text{tr}[X A_i \dot{X}^T + \sigma_z X S_i \dot{X}^T] \end{aligned} \tag{93}$$

Let $\nabla L = \frac{\partial L}{\partial W}$, where $W_a{}^b = U_a{}^\mu V_\mu{}^b$. Assume that $\gamma \to \infty$. Then $\dot{U} = -\frac{\partial L}{\partial U}$ and $\dot{V} = -\frac{\partial L}{\partial V}$. Below is the explicit form of $\dot{X}$ in this limit. Note that $\frac{\partial L}{\partial W_c^b} = \left[\frac{\partial L}{\partial W}\right]_b^c = [\nabla L]_b^c$.

$$\begin{aligned} \dot{U}_\mu^a &= -\frac{\partial L}{\partial U_a^\mu} = -\frac{\partial L}{\partial W_c^b}\frac{\partial W_c^b}{\partial U_a^\mu} = -[\nabla L]_b^c V_\mu^b \delta_{ca} = -[\nabla L]_b^a V_\mu^b = -V\nabla L \\ \dot{V}_b^\mu &= -\frac{\partial L}{\partial V_\mu^b} = -\frac{\partial L}{\partial W_a^c}\frac{\partial W_a^c}{\partial V_\mu^b} = -[\nabla L]_c^a U_a^\mu \delta_{cb} = -[\nabla L]_b^a U_a^\mu = -\nabla L U \\ \dot{X} &= \begin{bmatrix} \dot{U} \\ \dot{V}^T \end{bmatrix} = \begin{bmatrix} \dot{U}_a{}^\mu \\ \dot{V}^\mu{}_b \end{bmatrix} = -\begin{bmatrix} \nabla L^T V^T \\ \nabla L U \end{bmatrix} \end{aligned} \tag{94}$$

Substituting in $\dot{X}$, the first term of J becomes

$$\begin{aligned} \text{Tr}[X A_i \dot{X}^T] &= \text{Tr}[\dot{X}^T X A_i] \qquad (95) \\ &= \text{Tr}\left[\begin{bmatrix} V\nabla L & U^T \nabla L^T \end{bmatrix} \begin{bmatrix} U \\ V^T \end{bmatrix} A_i\right] \\ &= \text{Tr}\left[V\nabla L U A_i + U^T \nabla L^T V^T A_i\right] \\ &= \text{Tr}\left[V\nabla L U A_i - A_i^T U^T \nabla L^T V^T\right] \\ &= 0 \qquad (96) \end{aligned}$$

The second term of J is

$$\begin{aligned} \text{Tr}[\sigma_z X S_i \dot{X}^T] &= \text{Tr}\left[\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} U \\ V^T \end{bmatrix} S_i \begin{bmatrix} \dot{U}^T & \dot{V} \end{bmatrix}\right] \\ &= \text{Tr}[U S_i \dot{U}^T - V^T S_i \dot{V}] \qquad (= \dot{Q}_{GF}) \\ &= \text{Tr}[U S_i V \nabla L - V^T S_i U^T \nabla L^T] \\ &= \text{Tr}[U S_i V \nabla L - \nabla L^T V^T S_i U^T] \\ &= 0 \end{aligned} \tag{97}$$

Therefore, in the limit of first-order gradient flow, $\text{tr}[\sigma_z X S_i \dot{X}^T] = \dot{Q}_{GF} = 0$. In summary, we derived the explicit form of angular momentum and conserved quantity of the two-layer linear reparametrization, verified that angular momentum equals 0, and rediscovered the conserved quantities we had before.

J.7.2 Two-layer nonlinear neural networks

In this section, we find the conserved quantity corresponding to the antisymmetric basis for the following Lagrangian:

$$\mathcal{L} = e^{\gamma t}\left(\frac{|\dot{U}|^2 + |\dot{V}|^2}{2} - \gamma L(U\sigma(VX))\right) \tag{98}$$

where $U \in \mathbb{R}^{m\times h}$ and $V \in \mathbb{R}^{h\times n}$ are parameters, X is data, and σ is an element-wise function. For simplicity, we assumed that $\varepsilon = 1$.

Let T_i be an element of the Lie algebra of GL_h. One action that leaves f unchanged is:

$$g \cdot (U, V, X) = (g \cdot U, g \cdot V, g \cdot X) = (U g^{-1}, \sigma^{-1}(g\sigma(VX))X^{-1}, X). \tag{99}$$

We first derive δU and δV.

The action of infinitesimal $g_\varepsilon = I + \varepsilon \cdot T$ yields

$$\begin{aligned} g_\varepsilon \cdot (U, V, X) &= (U(I - \varepsilon^i T_i), \sigma^{-1}((I + \varepsilon^i T_i)\sigma(VX))X^{-1}, X) \\ &= (U - \varepsilon^i U T_i, \sigma^{-1}(\sigma(VX) + \varepsilon^i T_i \sigma(VX))X^{-1}, X) \end{aligned} \tag{100}$$

For small ϵ,

$$\begin{aligned} &\lim_{\varepsilon \to 0} \sigma^{-1}(\sigma(VX) + \varepsilon^i T_i \sigma(VX))X^{-1} \\ &= \left(\sigma^{-1}(\sigma(VX)) + \varepsilon^i T_i \sigma(VX) \odot \frac{\partial \sigma^{-1}(\sigma(VX))}{\partial \sigma(VX)}\right) X^{-1} \\ &= V + \varepsilon^i \left((T_i \sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1} \end{aligned} \tag{101}$$

$\odot$ here denotes element-wise product. Then, the action of g_ε yields

$$g_\varepsilon \cdot (U, V, X) = (U, V, X) + \varepsilon^i \left(-UT_i, \left((T_i \sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1}, 0\right) \tag{102}$$

For a set of parameters $\theta = (U, V, X) \in \Theta$,

$$T_i \cdot \theta = \left(-UT_i, \left((T_i \sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1}, 0\right) \tag{103}$$

Combining U, V into a single matrix,

$$X = \begin{bmatrix} U \\ V^T \end{bmatrix}, \qquad \delta X = \begin{bmatrix} -U \delta\theta^i T_i \\ \left[\left((\delta\theta^i T_i \sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1}\right]^T \end{bmatrix} \tag{104}$$

Let $\nabla L = \frac{\partial L}{\partial W}$, where $W_a^b = U_a^\mu \sigma(V_\mu^\nu X_\nu^b)$. Assume that $\gamma \to \infty$. Then $\dot{U} = -\frac{\partial L}{\partial U}$ and $\dot{V} = -\frac{\partial L}{\partial V}$. Below is the explicit form of $\dot{X}$ in this limit.

$$\begin{aligned} \frac{\partial L}{\partial U} &= \nabla L^T \sigma(VX)^T \\ \frac{\partial L}{\partial V} &= \left[(U^T \nabla L^T) \odot \sigma'(VX)\right] X^T \\ \dot{X} &= \begin{bmatrix} \dot{U} \\ \dot{V}^T \end{bmatrix} = -\begin{bmatrix} \frac{\partial L}{\partial U} \\ \frac{\partial L}{\partial V}^T \end{bmatrix} = -\begin{bmatrix} \nabla L^T \sigma(VX)^T \\ X\left[(\nabla L U) \odot \sigma'(VX)^T\right] \end{bmatrix} \end{aligned} \tag{105}$$

The conserved quantity is

$$\begin{aligned} J &= \delta X \cdot \frac{\partial \mathcal{L}}{\partial \dot{X}} \\ &= e^{\gamma t} \delta X \cdot \dot{X} \\ &= e^{\gamma t} \mathrm{Tr}[\delta X^T \dot{X}] \end{aligned} \tag{106}$$

Substitute in δX and $\dot{X}$,

$$
\begin{aligned}
J &= -\delta\theta^i e^{\gamma t} Tr\left[\begin{bmatrix} -T_i^T U^T & \left((T_i\sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1} \end{bmatrix} \begin{bmatrix} \nabla L^T \sigma(VX)^T \\ X\left[(\nabla LU) \odot \sigma'(VX)^T\right] \end{bmatrix}\right] \\
&= \delta\theta^i e^{\gamma t} Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T - \left((T_i\sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1} X \left((\nabla LU) \odot \sigma'(VX)^T\right)\right] \\
&= \delta\theta^i e^{\gamma t} Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T - \left((T_i\sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) \left((\nabla LU) \odot \sigma'(VX)^T\right)\right] \\
&= \delta\theta^i e^{\gamma t} \left(Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T\right] - \left((\sigma(VX)^T T_i^T) \odot \frac{1}{\sigma'(VX)^T}\right) \cdot \left((\nabla LU) \odot \sigma'(VX)^T\right)\right) \\
&= \delta\theta^i e^{\gamma t} \left(Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T\right] - (\sigma(VX)^T T_i^T) \cdot (\nabla LU)\right) \\
&= \delta\theta^i e^{\gamma t} Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T - T_i\sigma(VX)\nabla LU\right] \\
&= \delta\theta^i e^{\gamma t} Tr\left[T_i^T U^T \nabla L^T \sigma(VX)^T - \sigma(VX)\nabla LUT_i\right] \\
&= 0
\end{aligned}
\tag{107}
$$

Note that the second term in the 4th line can be evaluated by first taking the element-wise product of 4 matrices with same dimensions, then summing over all indices. The 5th line uses that fact that matrix element-wise product is associative and commutative. Again, we obtained that in the limit of first-order gradient flow ($\gamma \to \infty$), the conserved quantity for the second-order gradient flow $J = 0$ for any T_i.

Conserved quantities For conserved quantity Q and infinitesimal generators T, we have

$$0 = \frac{dQ}{dt} = \frac{\partial Q}{\partial U} \cdot \frac{dU}{dt} + \frac{\partial Q}{\partial V} \cdot \frac{dV}{dt} \tag{108}$$

and

$$0 = (T \cdot \boldsymbol{w}) \cdot \nabla L = -(\delta U \cdot \dot{U} + \delta V \cdot \dot{V}). \tag{109}$$

By matching terms, one way to find a conserved quantity is to find a Q that satisfies $\frac{\partial Q}{\partial U} = \delta U$ and $\frac{\partial Q}{\partial V} = \delta V$. Therefore, one set of conserved quantities is

$$Q = \frac{1}{2}\mathrm{Tr}[UT_iU^T] + Q_V, \tag{110}$$

where the second term Q_V does not depend on U and is the solution to $\frac{\partial Q_V}{\partial V_{ij}} = \left[\left((\delta\theta^i T_i\sigma(VX)) \odot \frac{1}{\sigma'(VX)}\right) X^{-1}\right]_{ij}$.

J.7.3 Two-layer nonlinear neural networks - alternative transformations

Another transformation that leaves f unchanged is:

$$g \cdot (U, V, X) = (g \cdot U, g \cdot V, g \cdot X) = (U\sigma(VX)\sigma(gVX)^{-1}, gV, X). \tag{111}$$

We first derive δU and δV.

The action of infinitesimal $g_\varepsilon = I + \varepsilon \cdot T$ yields

$$
\begin{aligned}
g_\varepsilon \cdot V &= (I + \varepsilon \cdot T)V = V + \varepsilon TV \\
g_\varepsilon \cdot U &= U\sigma(VX)\sigma((I + \varepsilon T)VX)^{-1} \\
&= U\sigma(VX)\sigma(VX + \varepsilon TVX)^{-1}
\end{aligned}
\tag{112}
$$

To simplify notation, define $f : \mathbb{R}^{h\times k} \to \mathbb{R}^{k\times h}, x \mapsto [\sigma(x)]^{-1}$. Then

$$g_\varepsilon \cdot U = U\sigma(VX)f(VX + \varepsilon TVX) \tag{113}$$

For small ε,

$$\begin{aligned}
&\lim_{\varepsilon\longrightarrow 0} U\sigma(VX)f(VX+\varepsilon TVX) \\
&= U\sigma(VX)\left[f(VX)+\varepsilon(TVX)_i^j\left.\frac{\partial f(x)}{\partial x_i^j}\right|_{x=VX}\right] \\
&= U\sigma(VX)\left[\sigma(VX)^{-1}+\varepsilon(TVX)_i^j\left.\frac{\partial f(x)}{\partial x_i^j}\right|_{x=VX}\right] \\
&= U+\varepsilon U\sigma(VX)(TVX)_i^j\left.\frac{\partial f(x)}{\partial x_i^j}\right|_{x=VX}
\end{aligned} \tag{114}$$

The last term can be evaluated to

$$\begin{aligned}
(TVX)_i^j\frac{\partial f(x)_a^b}{\partial x_i^j} &= (TVX)_i^j\frac{\partial[\sigma(x)^{-1}]_a^b}{\partial[\sigma(x)]_c^d}\frac{\partial\sigma(x)_c^d}{\partial x_i^j} \\
&= (TVX)_i^j\frac{\partial[\sigma(x)^{-1}]_a^b}{\partial[\sigma(x)]_c^d}\frac{\partial\sigma(x)_c^d}{\partial x_i^j}\delta_i^c\delta_d^j \\
&= (TVX)_i^j\frac{\partial[\sigma(x)^{-1}]_a^b}{\partial[\sigma(x)]_i^j}\frac{\partial\sigma(x)_i^j}{\partial x_i^j} \\
&= -(TVX)_i^j\left[[\sigma(x)^{-1}]\frac{\partial\sigma(x)}{\partial\sigma(x)_i^j}[\sigma(x)^{-1}]\right]_a^b\sigma'(x)_i^j \\
&= -(TVX)_i^j[\sigma(x)^{-1}]_a^n\left[\frac{\partial\sigma(x)}{\partial\sigma(x)_i^j}\right]_n^k[\sigma(x)^{-1}]_k^b\sigma'(x)_i^j \\
&= -(TVX)_i^j[\sigma(x)^{-1}]_a^n\delta_n^i\delta_j^k[\sigma(x)^{-1}]_k^b\sigma'(x)_i^j \\
&= -(TVX)_i^j[\sigma(x)^{-1}]_a^i[\sigma(x)^{-1}]_j^b\sigma'(x)_i^j \\
&= -\left(\sigma(x)^{-1}\left[TVX\odot\sigma'(x)\right]\sigma(x)^{-1}\right)_a^b
\end{aligned} \tag{115}$$

The fourth line uses the derivative of matrix inverse.

Substitute back and let $H=VX$, we have

$$\begin{aligned}
&\lim_{\varepsilon\longrightarrow 0} U\sigma(H)f(H+\varepsilon TH) \\
&= U-\varepsilon U\sigma(H)\sigma(H)^{-1}\left[TH\odot\sigma'(H)\right]\sigma(H)^{-1} \\
&= U-\varepsilon U\left[TH\odot\sigma'(H)\right]\sigma(H)^{-1}
\end{aligned} \tag{116}$$

$\odot$ here denotes element-wise product. Then, the action of g_ε yields

$$g_\varepsilon\cdot(U,V,X)=(U,V,X)+\varepsilon\left(-U\left[TH\odot\sigma'(H)\right]\sigma(H)^{-1},TV,0\right) \tag{117}$$

Combining U,V into a single matrix,

$$X=\begin{bmatrix}U\\V^T\end{bmatrix},\qquad \delta X=\begin{bmatrix}-\delta\theta^i U\left[TH\odot\sigma'(H)\right]\sigma(H)^{-1}\\ \delta\theta^i V^T T^T\end{bmatrix} \tag{118}$$

Let $\nabla L=\frac{\partial L}{\partial W}$, where $W_a^b=U_a^\mu\sigma(V_\mu^\nu X_\nu^b)$. Assume that $\gamma\to\infty$. Then $\dot U=-\frac{\partial L}{\partial U}$ and $\dot V=-\frac{\partial L}{\partial V}$. Below is the explicit form of $\dot X$ in this limit.

$$\begin{aligned}
\frac{\partial L}{\partial U} &= \nabla L^T\sigma(VX)^T \\
\frac{\partial L}{\partial V} &= \left[(U^T\nabla L^T)\odot\sigma'(VX)\right]X^T \\
\dot X &= \begin{bmatrix}\dot U\\ \dot V^T\end{bmatrix}=-\begin{bmatrix}\frac{\partial L}{\partial U}\\ \frac{\partial L}{\partial V}^T\end{bmatrix}=-\begin{bmatrix}\nabla L^T\sigma(VX)^T\\ X\left[(\nabla LU)\odot\sigma'(VX)^T\right]\end{bmatrix}
\end{aligned} \tag{119}$$

The conserved quantity is

$$\begin{aligned} J &= \delta X \cdot \frac{\partial \mathcal{L}}{\partial \dot{X}} \\ &= e^{\gamma t} \delta X \cdot \dot{X} \\ &= e^{\gamma t} \mathrm{Tr}[\delta X^T \dot{X}] \end{aligned} \tag{120}$$

Substitute in δX and $\dot{X}$,

$$\begin{aligned} J &= -\delta\theta^i e^{\gamma t} Tr\left[\delta U \dot{U}^T + \delta V \cdot \dot{V}^T\right] \\ &= -\delta\theta^i e^{\gamma t} Tr\left[-U\left[TH \odot \sigma'(H)\right]\sigma(H)^{-1}\sigma(H)\nabla L + TVX\left[(\nabla LU) \odot \sigma'(VX)^T\right]\right] \\ &= -\delta\theta^i e^{\gamma t} Tr\left[-U\left[TH \odot \sigma'(H)\right]\nabla L + TH\left[(\nabla LU) \odot \sigma'(H)^T\right]\right] \\ &= -\delta\theta^i e^{\gamma t} Tr\left[-\left[TH \odot \sigma'(H)\right]\nabla LU + TH\left[(\nabla LU) \odot \sigma'(H)^T\right]\right] \\ &= -\delta\theta^i e^{\gamma t}\left[-\left[TH \odot \sigma'(H)\right] \cdot (\nabla LU)^T + TH \cdot \left[(\nabla LU)^T \odot \sigma'(H)\right]\right] \\ &= 0 \end{aligned} \tag{121}$$

Conserved quantities The conserved quantities corresponding to the new group action can be found in a similar way. Note that δU now depends on both U and V. We seek to simplify the differential equations by removing δU's dependence on V. Let

$$\frac{\partial Q}{\partial U} = \delta U F(U, V, T), \qquad \frac{\partial Q}{\partial V} = F(U, V, T)^T \delta V \tag{122}$$

where $F = \left(\left[TH \odot \sigma'(H)\right]\sigma(H)^{-1}\right)^{-1}$. Then dQ/dt is still 0. One set of conserved quantities is

$$Q = \frac{1}{2}\mathrm{Tr}[UU^T] + Q_V, \tag{123}$$

where the second term Q_V does not depend on U and is the solution to $\frac{\partial Q_V}{\partial V_{ij}} = \left[F(U, V, T)^T TV\right]_{ij}$.

K GRADIENT DESCENT AND DRIFTING CONSERVED QUANTITIES

Due to the non-infinitesimal time steps, conserved quantities in gradient flow are no longer conserved in gradient descent. However, with small learning rate, we expect the change in the conserved quantities to be small. In this section, we first prove that the change of Q is bounded by the square of learning rate for two layer linear networks, then show empirically that the change Q is small for nonlinear networks.

K.1 CHANGE IN Q IN GRADIENT DESCENT (LINEAR LAYERS)

Proposition K.1. *Consider the two layer linear network, where $U \in \mathbb{R}^{m\times h}, V \in \mathbb{R}^{h\times n}$ are the only parameters, and the loss function $\mathcal{L}$ is a function of UV. In gradient descent with learning rate η, the change in the conserved quantity $Q = \mathrm{Tr}\left[U^T U - VV^T\right]$ at step t is bounded by*

$$|Q_{t+1} - Q_t| \leq \eta^2 \left|\frac{d\mathcal{L}(t)}{dt}\right|. \tag{124}$$

Proof. Let U_t and V_t be the value of U and V at time t in a gradient descent. The update rule is

$$U_{t+1} = U_t - \eta\frac{\partial \mathcal{L}}{\partial U}, \qquad V_{t+1} = V_t - \eta\frac{\partial \mathcal{L}}{\partial V} \tag{125}$$

Consider the two layer linear reparametrization $W = UV$.

$$\begin{aligned} Q_t &= \mathrm{Tr}\left[U_t^T U_t - V_t V_t^T\right] \\ Q_{t+1} &= \mathrm{Tr}\left[U_{t+1}^T U_{t+1} - V_{t+1} V_{t+1}^T\right] \\ &= \mathrm{Tr}\left[\left(U_t - \eta\frac{\partial \mathcal{L}(U_t)}{\partial U_t}\right)^T\left(U_t - \eta\frac{\partial \mathcal{L}(U_t)}{\partial U_t}\right) - \left(V_t - \eta\frac{\partial \mathcal{L}(V_t)}{\partial V_t}\right)\left(V_t - \eta\frac{\partial \mathcal{L}(V_t)}{\partial V_t}\right)^T\right] \end{aligned} \tag{126}$$

Expanding Q_{t+1} and subtracting by Q_t, we have

$$Q_{t+1} - Q_t = \mathrm{Tr}\left[\eta^2 \left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T \frac{\partial \mathcal{L}}{\partial U_t} - \eta \left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T U_t - \eta U_t^T \frac{\partial \mathcal{L}}{\partial U_t} - \eta^2 \frac{\partial \mathcal{L}}{\partial V_t}\left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T + \eta \frac{\partial \mathcal{L}}{\partial V_t} V_t^T + \eta V_t \left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T\right] \tag{127}$$

Note that

$$\left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T U_t = (\nabla \mathcal{L} V_t^T)^T U_t = V_t \nabla \mathcal{L}^T U_t = V_t \left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T, \tag{128}$$

and similarly

$$U_t^T \frac{\partial \mathcal{L}}{\partial U_t} = \frac{\partial \mathcal{L}}{\partial V_t} V_t^T. \tag{129}$$

Therefore, 127 simplifies to

$$\begin{aligned} Q_{t+1} - Q_t &= \eta^2 \mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T \frac{\partial \mathcal{L}}{\partial U_t} - \frac{\partial \mathcal{L}}{\partial V_t}\left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T\right] \\ &= \eta^2 \left(\mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T \frac{\partial \mathcal{L}}{\partial U_t}\right] - \mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T \frac{\partial \mathcal{L}}{\partial V_t}\right]\right), \end{aligned} \tag{130}$$

and the variation of Q in each step is bounded by the convergence rate:

$$\begin{aligned} |Q_{t+1} - Q_t| &= \eta^2 \left|\mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T \frac{\partial \mathcal{L}}{\partial U_t}\right] - \mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T \frac{\partial \mathcal{L}}{\partial V_t}\right]\right| \\ &\leq \eta^2 \left|\mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial U_t}\right)^T \frac{\partial \mathcal{L}}{\partial U_t}\right] + \mathrm{Tr}\left[\left(\frac{\partial \mathcal{L}}{\partial V_t}\right)^T \frac{\partial \mathcal{L}}{\partial V_t}\right]\right| \\ &= \eta^2 \left|\frac{d\mathcal{L}}{dt}\right| \end{aligned} \tag{131}$$

□

K.2 Empirical observations

In gradient flow, the conserved quantity Q is constant by definition. In gradient descent, Q varies with time. In order to see how applicable our theoretical results are in gradient descent, we investigate the amount of variation in Q in gradient descent using two-layer neural networks.

Since Q is the difference between the two terms $f_1(U) = \frac{1}{2}\mathrm{Tr}[U^T U]$ and $f_2(V) = \sum_{a,j} \int_{x_0}^{V_{aj}} dx \frac{\sigma(x)}{\sigma'(x)}$, we normalize Q by the initial value of $f_1(U)$ and $f_2(V)$, i.e.,

$$\tilde{Q} = \frac{\left|\frac{1}{2}\mathrm{Tr}[U^T U] - \sum_{a,j} \int_{x_0}^{V_{aj}} dx \frac{\sigma(x)}{\sigma'(x)}\right|}{\left|\frac{1}{2}\mathrm{Tr}[U_0^T U_0]\right| + \left|\sum_{a,j} \int_{x_0}^{V_{0aj}} dx \frac{\sigma(x)}{\sigma'(x)}\right|}$$

and denote the amount of change in $\tilde{Q}$ as

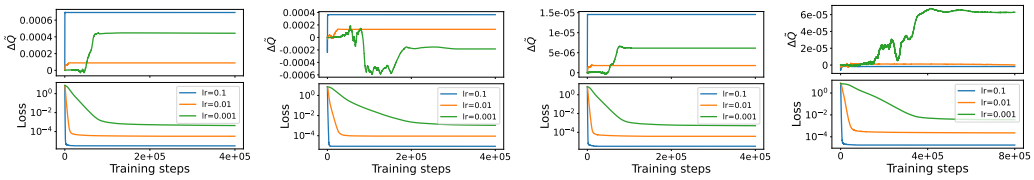
$$\Delta \tilde{Q}(t) = \tilde{Q}(t) - \tilde{Q}(0) \tag{132}$$

We run gradient descent on two-layer networks with whitened input with the following objective

$$\mathrm{argmin}_{U,V}\{\mathcal{L}(U,V) = \|Y - U\sigma(V^T)\|_F^2\} \tag{133}$$

where σ is the identity function, ReLU, sigmoid, or tanh. $Y \in \mathbb{R}^{5\times 10}$, $U \in \mathbb{R}^{5\times 50}$ and $V \in \mathbb{R}^{10\times 50}$ have random Gaussian initialization with zero mean. We repeat the gradient descent with learning rate 0.1, 0.01, and 0.001.

The variation $\Delta\tilde{Q}(t)$ and loss is shown in Fig.4. The amount of change in Q is small relative to the magnitude of $f_1(U)$ and $f_2(V)$, indicating that conserved quantities in gradient flow are approximately conserved in gradient descent. The error in Q grows with step size, as $\Delta\tilde{Q}(t)$ is larger with the largest learning rate we used, although it has the same magnitude as those of smaller learning rates. We also observe that Q stays constant after loss converges.



(a) linear (b) ReLU (c) tanh (d) sigmoid

Figure 4: Dynamics of conserved quantities in GD. The amount of change in Q is small relative to its magnitude, and Q converges when loss converges.

L DISTRIBUTION OF Q UNDER XAVIER INITIALIZATION

We first consider a linear two-layer neural network UVX, where $U \in \mathbb{R}^{m\times h}, V \in \mathbb{R}^{h\times n}$, and $X \in \mathbb{R}^{n\times k}$. We choose the following form of the conserved quantity:

$$Q = \frac{1}{2}\mathrm{Tr}[U^T U - VV^T]. \tag{134}$$

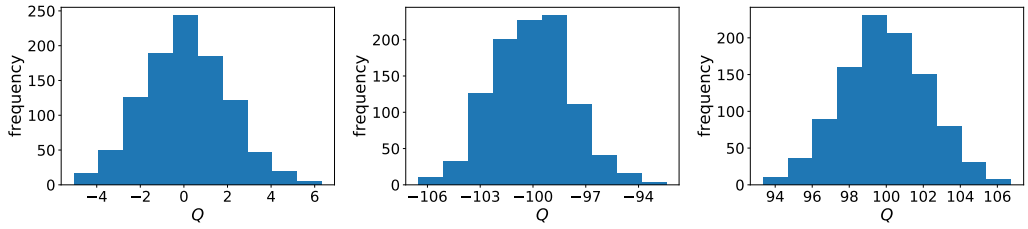
Xavier initialization keeps the variance of each layer's output the same as the variance of the input. Under Xavier initialization (Glorot & Bengio, 2010), each element in a given layer is initialized independently, with mean 0 and variance equal to the inverse of the layer's input dimension:

$$U_{ij} = \mathcal{N}\left(0, \frac{1}{h}\right) \qquad V_{ij} = \mathcal{N}\left(0, \frac{1}{n}\right) \tag{135}$$

The expected value of Q is

$$\mathbb{E}[Q] = Var(U_{ij}) \times m \times h + Var(V_{ij}) \times h \times n = m - h. \tag{136}$$

Figure 5 shows the distribution of Q for 2-layer linear NN with different layer dimensions. For each dimension tuples (m, h, n), we constructed 1000 sets of parameters using Xavier initialization. The centers of the distributions of Q match Eq. (136).



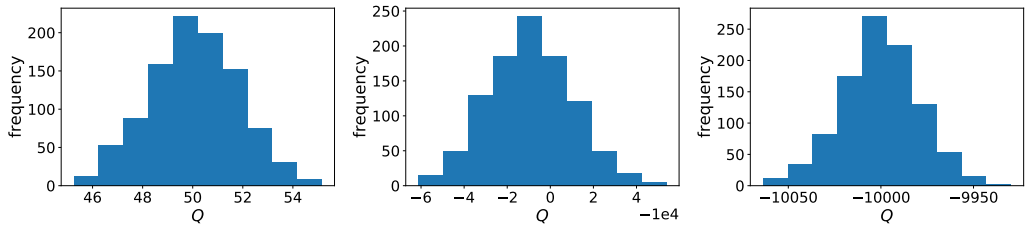
(a) $m, h, n = 100, 100, 100$ (b) $m, h, n = 100, 200, 100$ (c) $m, h, n = 200, 100, 100$

Figure 5: Distribution of Q for 2-layer linear NN with different layer dimensions.

Next, we consider the nonlinear two-layer neural network $U\sigma(VX)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is an element-wise activation function. For simplicity, we assume whitened input ($X = I$). We choose the following form of the conserved quantity:

$$Q = \frac{1}{2}\mathrm{Tr}[U^T U] - \sum_{a,j} \int_0^{V_{aj}} dx \frac{\sigma(x)}{\sigma'(x)} \tag{137}$$

Figure 6 shows the distribution of Q for 2-layer linear NN with different nonlinearities, each with 1000 sets of parameters created under Xavier initialization. The shapes of the distributions are similar to that of linear networks. The value of Q is usually concentrated around a small range of values. Since the range of Q is unbounded, the Xavier initialization limits the model to a small part of the global minimum.



(a) ReLU (b) tanh (c) sigmoid

Figure 6: Distribution of Q for 2-layer linear NN with different nonlinearities, with parameter dimensions $m = h = n = 100$.

M CONSERVED QUANTITY AND CONVERGENCE RATE

The values of conserved quantities are unchanged throughout the gradient flow. Since the conserved quantities parameterize trajectories, initializing parameters with certain conserved quantity values accelerates convergence. For two-layer linear reparametrization, Tarmoun et al. (2021) derived the explicit relation between layer imbalance and convergence rate. We derive the relation between conserved quantities and convergence rate for two example optimization problems and provide numerical evidence that initializing parameters with optimal conserved quantity values accelerates convergence.

M.1 EXAMPLE 1: ELLIPSE

We first show that the convergence rate is related to the conserved quantity in a toy optimization problem. Consider the following loss function with $a \in \mathbb{R}$:

$$\begin{aligned} \mathcal{L}(w_1, w_2) &= w_1^2 + a w_2^2 \\ \nabla \mathcal{L} &= (2w_1, 2aw_2) \end{aligned} \tag{138}$$

Assuming gradient flow,

$$\frac{dw_1}{dt} = -\nabla_{w_1}\mathcal{L} = -2w_1 \qquad \frac{dw_2}{dt} = -\nabla_{w_2}\mathcal{L} = -2aw_2 \tag{139}$$

Then w_1, w_2 are governed by the following differential equations:

$$w_1(t) = w_{1_0} e^{-2t} \qquad w_2(t) = w_{2_0} e^{-2at} \tag{140}$$

where w_{1_0}, w_{2_0} are initial values of w_1 and w_2. We can find conserved quantities by using an ansatz $Q = f(w_1^i w_2^k)$ and solving $\nabla Q \cdot \nabla L = 0$ for i, k. Below we use the following form of conserved quantity:

$$Q = \frac{w_1^{2a}}{w_2^2} = \frac{w_{1_0}^{2a}}{w_{2_0}^2} \tag{141}$$

To show the effect of Q on the convergence rate, we fix $L(0)$ and derive how Q affects $L(t)$. Let $L(0) = w_{1_0}^2 + a w_{2_0}^2 = L_0$. Let w_{2_0} continue to be an independent variable. Then $w_{1_0}^2 = L_0 - a w_{2_0}^2$. Substitute in $w_{1_0}^2$, the loss at time t is

$$L(t) = w_1(t)^2 + a w_2(t)^2 = (L_0 - a w_{2_0}^2) e^{-4t} + a w_{2_0}^2 e^{-4at} \tag{142}$$

and Q becomes

$$Q = \frac{w_{1_0}^{2a}}{w_{2_0}^2} = \frac{(L_0 - a w_{2_0}^2)^a}{w_{2_0}^2} \tag{143}$$

The derivative of L in the direction of Q is

$$\begin{aligned}
\partial_Q L(t) &= \frac{dL(t)}{dw_{2_0}} \frac{dw_{2_0}}{dQ} = \frac{dL(t)}{dw_{2_0}} \left(\frac{dQ}{dw_{2_0}} \right)^{-1} \\
&= \left(-2aw_{2_0} e^{-4t} + 2aw_{2_0} e^{-4at} \right) \left(\frac{a(L_0 - aw_{2_0}^2)^{a-1}(-2aw_{2_0})w_{2_0}^2 - 2w_{2_0}(L_0 - aw_{2_0}^2)^a}{w_{2_0}^4} \right)^{-1} \\
&= \frac{\left(-2aw_{2_0} e^{-4t} + 2aw_{2_0} e^{-4at} \right) w_{2_0}^4}{a(L_0 - aw_{2_0}^2)^{a-1}(-2aw_{2_0})w_{2_0}^2 - 2w_{2_0}(L_0 - aw_{2_0}^2)^a} \\
&= \frac{2aw_{2_0}^5 \left(e^{-4at} - e^{-4t} \right)}{2w_{2_0}(L_0 - aw_{2_0}^2)^{a-1} \left(-a^2 w_{2_0}^2 - (L_0 - aw_{2_0}^2) \right)}
\end{aligned} \tag{144}$$

In general, $\partial_Q L(t) \neq 0$, meaning that the loss at time t depends on Q. Since we have fixed the initial loss, the convergence rate $L(t) - L(0)$ also depends on Q. Special cases where $\partial_Q L(t) = 0$ include $a = 1$ (circle), $a = 0$ (collapsed dimension), and certain initializations such as $w_{2_0} = 0$ (local maximum of gradient magnitude).

M.2 Example 2: radial activation functions

In this example, we find the conserved quantities and their relation with convergence rate for two-layer reparametrization with radial activation functions under spectral initialization.

Define radial function $g : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ as

$$g(W)_{ij} = h\left(|W_i|\right) W_{ij}, \tag{145}$$

where $|W_i| = \left(\sum_k W_{ik}^2 \right)^{\frac{1}{2}}$ is the norm of the ith row of W, and $h : \mathbb{R} \to \mathbb{R}$ outputs a scalar.

Consider the following objective:

$$\text{argmin}_{U,V} \{ \mathcal{L}(U, V) = \frac{1}{2} \| Y - U g(V^T) \|_F^2 \} \tag{146}$$

with spectral initializations

$$U_0 = \Phi \overline{U}_0, V_0 = \Psi \overline{V}_0,$$

where Φ, Ψ come from the singular value decomposition $Y = \Phi \Sigma_Y \Psi^T$, and $\overline{U}_0, \overline{V}_0$ are random diagonal matrices.

Proposition M.1. *Under the gradient flow $U = -\nabla_U \mathcal{L}$ and $V = -\nabla_V \mathcal{L}$, the following quantity is an invariant:*

$$Q = \frac{1}{2} \text{Tr}[U^T U] - \sum_i \int_{x_0}^{\bar{V}_{ii}} dx \frac{g(x)}{g'(x)} \tag{147}$$

Proof. Since g is a radial function on rows and Ψ^T is an orthogonal matrix, $g(\overline{V}^T \Psi^T) = g(\overline{V}^T) \Psi^T$. With spectral initialization, the loss function can be reduced to only involving diagonal matrices:

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2} \| Y - U g(V^T) \|_F^2 \\
&= \frac{1}{2} \| \Phi \Sigma \Psi^T - \Phi \overline{U} g[(\Psi \overline{V})^T] \|_F^2 \\
&= \frac{1}{2} \| \Phi \Sigma \Psi^T - \Phi \overline{U} g(\overline{V}^T) \Psi^T \|_F^2 \\
&= \frac{1}{2} \| \Phi \left(\Sigma - \overline{U} g(\overline{V}^T) \right) \Psi^T \|_F^2 \\
&= \frac{1}{2} \| \Sigma - \overline{U} g(\overline{V}^T) \|_F^2
\end{aligned} \tag{148}$$

Since $\overline{V}$ is a diagonal matrix, g is now an element wise function on $\overline{V}$. Let $\overline{W} = \overline{U}g(\overline{V}^T)$. The gradients for $\overline{U}$ and $\overline{V}$ are

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \overline{U}} &= \nabla_{\overline{W}}\mathcal{L}g(\overline{V})^T \\ \frac{\partial \mathcal{L}}{\partial \overline{V}} &= \nabla_{\overline{W}}\mathcal{L}^T\overline{U} \odot g'(\overline{V})\end{aligned} \tag{149}$$

where $g'(x) = dg(x)/dx$ is the derivative of the nonlinearity. Additionally, since $\mathcal{L}$ does not depend on Φ and Ψ,

$$\frac{\partial \mathcal{L}}{\partial \Phi} = \frac{\partial \mathcal{L}}{\partial \Psi} = 0 \tag{150}$$

Since the rows of Φ, Ψ are orthogonal,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial U} &= \frac{\partial \mathcal{L}}{\partial \overline{U}}\Phi^T = \nabla_{\overline{W}}\mathcal{L}g(\overline{V})^T\Phi^T \\ \frac{\partial \mathcal{L}}{\partial V} &= \frac{\partial \mathcal{L}}{\partial \overline{V}}\Psi^T = \left(\nabla_{\overline{W}}\mathcal{L}^T\overline{U} \odot g'(\overline{V})\right)\Psi^T\end{aligned} \tag{151}$$

Φ and Ψ are not changed in gradient flow, so $\frac{\partial Q}{\partial U} = \frac{\partial Q}{\partial \overline{U}}\Phi^T$ and $\frac{\partial Q}{\partial V} = \frac{\partial Q}{\partial \overline{V}}\Psi^T$. Define inner product on matrices as $\langle X, Y\rangle = \mathrm{Tr}[X^TY]$. For Q to be a conserved quantity, we need $\langle \nabla\mathcal{L}, \nabla Q\rangle = 0$:

$$\begin{aligned}\langle \nabla\mathcal{L}, \nabla Q\rangle &= \langle \frac{\partial \mathcal{L}}{\partial U}, \frac{\partial Q}{\partial U}\rangle + \langle \frac{\partial \mathcal{L}}{\partial V}, \frac{\partial Q}{\partial V}\rangle \\ &= \langle \nabla_{\overline{W}}\mathcal{L}g(\overline{V})^T\Phi^T, \frac{\partial Q}{\partial \overline{U}}\Phi^T\rangle + \langle \left(\nabla_{\overline{W}}\mathcal{L}^T\overline{U} \odot g'(\overline{V})\right)\Psi^T, \frac{\partial Q}{\partial \overline{V}}\Psi^T\rangle \\ &= \langle \nabla_{\overline{W}}\mathcal{L}g(\overline{V})^T, \frac{\partial Q}{\partial \overline{U}}\rangle + \langle \left(\nabla_{\overline{W}}\mathcal{L}^T\overline{U} \odot g'(\overline{V})\right), \frac{\partial Q}{\partial \overline{V}}\rangle \\ &= \mathrm{Tr}\left[\partial_{\overline{U}^T}Q\nabla_{\overline{W}}\mathcal{L}g(\overline{V})^T + \overline{U}^T\nabla_{\overline{W}}\mathcal{L}(\partial_V Q \odot g'(\overline{V}))\right] = 0\end{aligned} \tag{152}$$

Following the same procedure as for elementwise functions, to have a Q which satisfies (152) it is sufficient to have

$$\frac{\partial Q}{\partial \overline{U}_{ia}} = f(\overline{U}, \overline{V})\overline{U}_{ia} \qquad \frac{\partial Q}{\partial \overline{V}_{aj}}g'(\overline{V})_{aj} = -f(\overline{U}, \overline{V})g(\overline{V})_{aj} \qquad f(\overline{U}, \overline{V}) \in \mathbb{R} \tag{153}$$

For simplicity, let $f(\overline{U}, \overline{V}) = 1$. Then, (153) is satisfied by

$$Q = \frac{1}{2}\mathrm{Tr}[\bar{U}^T\bar{U}] - \sum_i \int_{x_0}^{\bar{V}_{ii}} dx \frac{g(x)}{g'(x)} \tag{154}$$

□

Tarmoun et al. (2021) shows that the conserved quantity Q appears as a term in the convergence rate of the matrix factorization gradient flow. We observe a similar relationship between Q and convergence rate when the loss function is augmented with a radial activation function, as shown in the following proposition.

Proposition M.2. *Consider the objective function and spectral initialization defined in Proposition M.1. Let $h(|W_i|) = |W_i|^{-2}$, and $X = Ug(V^T) = \Phi\Sigma_X\Psi^T$. Then, the eigencomponent of X approaches the corresponding eigencomponent of Y at a rate of*

$$\dot{\sigma}_i^X = \frac{1}{\lambda_i}(\sigma_i^Y - \sigma_i^X)(\sigma_i^{X\,2} + 1)^2, \tag{155}$$

where $\sigma_i^X = diag(\Sigma_X)_i$, $\sigma_i^Y = diag(\Sigma_Y)_i$, and $\lambda_i = \bar{U}_{ii}^2 + \bar{V}_{ii}^2$ are conserved quantities.

Proof. Similar to Tarmoun et al. (2021), components can be decoupled, and we have a set of differential equations on scalars:

$$\begin{aligned}\dot{\overline{u}}_i &= [\sigma_i^Y - \overline{u}_i g(\overline{v}_i)]g(\overline{v}_i)\\ \dot{\overline{v}}_i &= [\sigma_i^Y - \overline{u}_i g(\overline{v}_i)]\overline{u}_i \frac{dg(\overline{v}_i)}{d\overline{v}_i}\end{aligned} \tag{156}$$

We also have

$$\dot{g}(\overline{v}_i) = \frac{dg}{d\overline{v}_i}\frac{d\overline{v}_i}{dt} = [\sigma_i^Y - \overline{u}_i g(\overline{v}_i)]\overline{u}_i \left(\frac{dg(\overline{v}_i)}{d\overline{v}_i}\right)^2. \tag{157}$$

Let $\sigma_i^X = \overline{u}_i g(\overline{v}_i)$. Then

$$\begin{aligned}\dot{\sigma}_i^X &= \dot{\overline{u}}_i g(\overline{v}_i) + \overline{u}_i \dot{g}(\overline{v}_i)\\ &= \left[\sigma_i^Y - \overline{u}_i g(\overline{v}_i)\right]\left[g(\overline{v}_i)^2 + \overline{u}_i^2\left(\frac{dg(\overline{v}_i)}{d\overline{v}_i}\right)^2\right].\end{aligned} \tag{158}$$

Since $\overline{V}$ is a diagonal matrix, g is now an element wise function on $\overline{V}$. Specifically, $g(\overline{v}_i) = \frac{1}{\overline{v}_i}$. According to Proposition M.1, the following quantity is invariant:

$$\frac{1}{2}\overline{u}_i^2 - \int dx \frac{g(x)}{g'(x)} = \frac{1}{2}\overline{u}_i^2 - \int dx \frac{\overline{v}_i^{-1}}{-\overline{v}_i^{-2}} = \frac{1}{2}\overline{u}_i^2 + \frac{1}{2}\overline{v}_i^2 \tag{159}$$

Since any function of the invariant is also invariant, we will use the following form:

$$Q = \overline{U}^T\overline{U} + \overline{V}^T\overline{V}, \tag{160}$$

and define

$$\lambda_i = Q_{ii} = \overline{u}_i^2 + \overline{v}_i^2 \tag{161}$$

Using the g that we defined,

$$\sigma_i^X = \overline{u}_i g(\overline{v}_i) = \overline{u}_i \overline{v}_i^{-1}. \tag{162}$$

In order to relate σ^X and Q, we first write $\overline{u}_i$ and $\overline{v}_i$ as functions of σ_i^X ad Q using (161) and (162):

$$\overline{u}_i^2 = \frac{\lambda_i {\sigma_i^X}^2}{{\sigma_i^X}^2 + 1}, \qquad \overline{v}_i^2 = \frac{\lambda_i}{{\sigma_i^X}^2 + 1}. \tag{163}$$

Then, substitute $\overline{u}_i$, $\overline{v}_i$, $g(\overline{v}_i)$, and $\frac{dg(\overline{v}_i)}{d\overline{v}_i}$ into (158), and we have

$$\begin{aligned}\dot{\sigma}_i^X &= [\sigma_i - \overline{u}_i g(\overline{v}_i)]\left[g(\overline{v}_i)^2 + \overline{u}_i^2\left(\frac{dg(\overline{v}_i)}{d\overline{v}_i}\right)^2\right]\\ &= \left[\sigma_i^Y - \overline{u}_i g(\overline{v}_i)\right]\left[\left(\frac{1}{\overline{v}_i}\right)^2 + \overline{u}_i^2\left(-\overline{v}_i^{-2}\right)^2\right]\\ &= \left[\sigma_i^Y - \overline{u}_i g(\overline{v}_i)\right]\left[(\overline{v}_i^2)^{-1} + \overline{u}_i^2(\overline{v}_i^2)^{-2}\right]\\ &= \left[\sigma_i^Y - \sigma_i^X\right]\left[\left(\frac{\lambda_i}{{\sigma_i^X}^2+1}\right)^{-1} + \frac{\lambda_i {\sigma_i^X}^2}{{\sigma_i^X}^2+1}\left(\frac{\lambda_i}{{\sigma_i^X}^2+1}\right)^{-2}\right]\\ &= \left[\sigma_i^Y - \sigma_i^X\right]\left[\frac{{\sigma_i^X}^2+1}{\lambda_i} + \frac{{\sigma_i^X}^2({\sigma_i^X}^2+1)}{\lambda_i}\right]\\ &= \left[\sigma_i^Y - \sigma_i^X\right]\left[\frac{{\sigma_i^X}^4 + 2{\sigma_i^X}^2 + 1}{\lambda_i}\right]\\ &= \frac{1}{\lambda_i}(\sigma_i^Y - \sigma_i^X)({\sigma_i^X}^2+1)^2\end{aligned} \tag{164}$$

□

Proposition M.2 relates the rate of change in parameters $\dot{\sigma}_i^X$ and the conserved quantity λ_i. To get a more explicit expression of how λ_i affects convergence rate, we will derive a bound for $|\sigma_i^Y - \sigma_i^X|$, which describes the distance between trainable parameters to their desired value.

Proposition M.3. *The difference between the singular values of $Ug(V^T)$ and Y is bounded by*

$$|\sigma_i^X - \sigma_i^Y| \leq |\sigma_i^X(0) - \sigma_i^Y| e^{-\frac{t}{\lambda_i}}. \tag{165}$$

Proof. Note that

$$\dot{\sigma_i^X} = \frac{1}{\lambda}(\sigma_i^Y - \sigma_i^X)({\sigma_i^X}^2 + 1)^2 \geq \frac{1}{\lambda_i}(\sigma_i^Y - \sigma_i^X) \tag{166}$$

Consider the following two differential equations, with same initialization $a(0) = b(0)$:

$$\begin{aligned} \dot{a} &= \frac{1}{\lambda}(\sigma - a)(a^2 + 1)^2 \\ \dot{b} &= \frac{1}{\lambda}(\sigma - b) \end{aligned} \tag{167}$$

In these equations, both a and b moves from $a(0) = b(0)$ to σ monotonically. Since $\dot{a} \geq \dot{b}$ at every $a = b$, a will always be closer to σ than b does. We can explicitly solve for b, which yields $b(t) = \sigma + (b(0) - \sigma)e^{-\frac{t}{\lambda}}$. Then the distance between b and σ is $|b - \sigma| = |b(0) - \sigma|e^{-\frac{t}{\lambda}}$. Using $|b - \sigma|$, we can bound $|a - \sigma|$:

$$|a - \sigma| \leq |b - \sigma| = |b(0) - \sigma|e^{-\frac{t}{\lambda}} \tag{168}$$

Therefore,

$$|\sigma_i^X - \sigma_i^Y| \leq |\sigma_i^X(0) - \sigma_i^Y| e^{-\frac{t}{\lambda_i}} \tag{169}$$

□

Since λ is a conserved quantity, its value set at initialization remains unchanged throughout the gradient flow. Therefore, we are able to optimize the convergence rate by choosing a favorable value for λ at initialization. In this example, smaller λ_i's lead to faster convergence.

M.3 Experiments

We compare the convergence rate of two-layer networks initialized with different Q values. We run gradient descent on two-layer networks with whitened input with the following objective

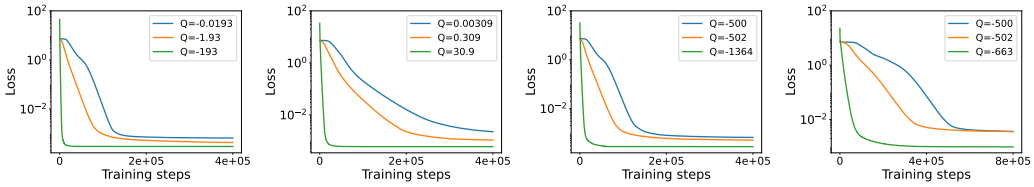
$$\text{argmin}_{U,V}\{\mathcal{L}(U, V) = \|Y - U\sigma(V^T)\|_F^2\} \tag{170}$$

where σ is the identity function, ReLU, sigmoid, or tanh. Matrices $Y \in \mathbb{R}^{5\times 10}$, $U \in \mathbb{R}^{5\times 50}$ and $V \in \mathbb{R}^{10\times 50}$ have random Gaussian initialization with zero mean. We repeat the gradient descent with learning rate 0.1, 0.01, and 0.001. The learning rate is set to 10^{-3}, as we do not observe significant changes in the shape of learning curves at smaller learning rates. U and V are initialized with different variance, which leads to different initial values of Q. As shown in Fig.7, the number of steps required for the loss curves to drop to near convergence level is correlated with Q in both linear and element-wise nonlinear networks. This result provides empirical evidence that initializing parameters with optimal values for Q accelerates convergence.

We then demonstrate the effect of conserved quantity values on the convergence rate of radial neural networks. Fig.8 shows the training curve for loss function defined in Proposition M.2. We initialize parameters $U \in \mathbb{R}^{5\times 5}$ and $V \in \mathbb{R}^{10\times 5}$ with 4 different values of Q and the learning rate is set to 10^{-5}. As predicted in Eq. 165, convergence is faster when $Q = \text{Tr}[U^T U + V^T V]$ is small.

N Conserved quantity and generalization ability

Conserved quantities parameterize the minimum of neural networks and are related to the eigenvalues of the Hessian at minimum. Recent theory and empirical studies suggest that sharp minimum



(a) linear (b) ReLU (c) tanh (d) sigmoid

Figure 7: Training curves of two-layer networks initialized with different Q. The value of Q affects convergence rate.

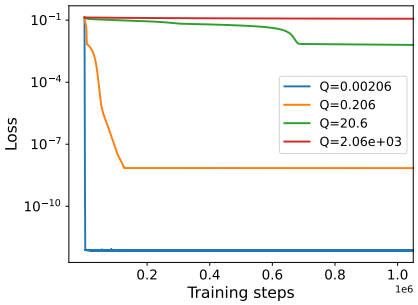


Figure 8: Training curve for the loss function defined in Proposition M.2. Smaller value of $Q = \mathrm{Tr}[U^T U + V^T V]$ at initialization leads to faster convergence.

do not generalize well (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Petzka et al., 2021). Explicitly searching for flat minimum has been shown to improve generalization bounds and model performance (Chaudhari et al., 2017; Foret et al., 2020; Kim et al., 2022). We derive their relationship for the simplest two-layer network, and show empircally that conserved quantity values affect sharpness. Like convergence rate, a systematic study of the relationship between conserved quantity and generalization ability of the solution is an interesting future direction.

N.1 Example: two-layer linear network with 1D parameters

We again consider the two-layer linear network with loss $\mathcal{L} = \frac{1}{2}\|Y - UVX\|^2$. For simplicity, we work with one dimensional parameters $U, V \in \mathbb{R}$ and assume $X = Y = 1$ in this example. We show that at the point to which the gradient flow converges, the eigenvalues of the Hessian are related to the value of the conserved quantity.

The gradients and Hessian of $\mathcal{L}$ are

$$\nabla\mathcal{L} = \begin{bmatrix} -(Y - UVX)VX \\ -(Y - UVX)UX \end{bmatrix} \qquad \mathcal{H} = \begin{bmatrix} V^2X^2 & -YX + 2UVX^2 \\ -YX + 2UVX^2 & U^2X^2 \end{bmatrix} \tag{171}$$

At the minima, U, V are related by $UVX = Y$. Recall that $Q = U^2 - V^2$ is a conserved quantity. From the above two equations, we can write U, V as functions of Q. Taking the solution $U = \sqrt{\frac{1}{2}(Q + \sqrt{Q^2+4})}, V = \sqrt{\frac{1}{2}(-Q + \sqrt{Q^2+4})}$ and substitute in $X = Y = 1$, we have

$$\mathcal{H} = \begin{bmatrix} \frac{1}{2}(-Q + \sqrt{Q^2+4}) & 1 \\ 1 & \frac{1}{2}(Q + \sqrt{Q^2+4}) \end{bmatrix}, \tag{172}$$

and the eigenvalues of $\mathcal{H}$ are

$$\lambda_1 = 0, \qquad \lambda_2 = 2\sqrt{Q^2+4}. \tag{173}$$

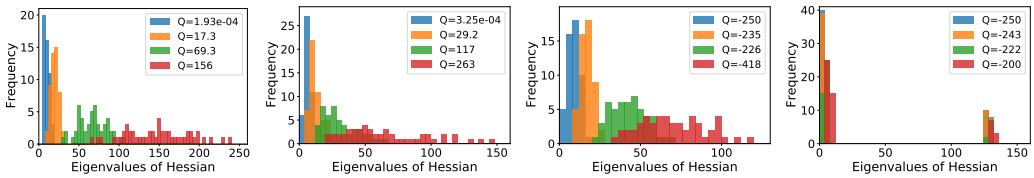
We have shown that Q is related the eigenvalues of the Hessian at the minimum. Since the eigenvalues determines the curvature, Q also determines the sharpness of the minimum, which is believed to

related to model's generalization ability. The result in this example can also be observed in Figure 1, where the minimum of the $Q = 0$ trajectory lies at the least sharp point of the loss valley.

N.2 EXPERIMENTS: TWO-LAYER NETWORKS

The goal of this section is to explore the relation between Q and the sharpness of the trained model. We measure sharpness by the magnitude of the eigenvalues of the Hessian, which are related to the curvature at the minima. We use the same loss function (170) in Section M.3. The parameters are $U \in \mathbb{R}^{10\times 50}$ and $V \in \mathbb{R}^{5\times 50}$, each initialized with zero mean and various standard deviations that lead to different Q's. We first train the models using gradient descent. We then use the vectorized parameters in the trained model to compute the eigenvalues of the Hessian.

The linear model extends the example in Section N.1 to higher dimension parameter spaces. 700 out of the 750 eigenvalues are around 0 (with magnitude $\leq 10^{-3}$), which verifies the dimension of the minima in Proposition C.3. After removing the small eigenvalues, the center of the eigenvalue distribution correlates positively with the value of Q (Figure 9(a)). In models with nonlinear activations, Q is still related to eigenvalue distributions, although the relations seem to be more complicated.



(a) linear (b) ReLU (c) tanh (d) sigmoid

Figure 9: Eigenvalues of the Hessian from trained models initialized with different conserved quantity values (Q). The distribution of the eigenvalues and the value of Q appear to be related.

O ENSEMBLE MODELS

In neural networks, the optima of the loss functions are connected by curves or volumes, on which the loss is almost constant (Freeman & Bruna, 2017; Garipov et al., 2018; Draxler et al., 2018; Benton et al., 2021; Izmailov et al., 2018). Various algorithms have been proposed to find these low-cost curves, which provides a low-cost way to create an ensemble of models from a single trained model. Using our group actions, we propose a new way of constructing models with similar loss values. We show that even with stochasticity in the data, the loss is approximately unchanged under the group action (Appendix O). This provides an efficient alternative to build ensemble models, since the transformation only requires random elements in the symmetry group, without any searching or additional optimization.

We implement our group actions by modifying the activation function between two consecutive layers. Let $H = VX$ be the output of the previous layer. The group action on the weights U, V is

$$g \cdot (U, V) = (U\pi(g, H), gV) \tag{174}$$

where $\pi(g, H) = \sigma(H)\sigma(gH)^{\dagger}$. The new activation implements the symmetry group action

$$U\sigma(H) \to U\pi(g, H)\sigma(gH) \tag{175}$$

by wrapping the transformations around an activation function $\sigma'(x) = \pi(g, x)\sigma(gx)$, so that $U\sigma'(H) = U\pi(g, H)\sigma(gH)$.

We test the group action on CIFAR-10. The model contains a convolution layer with kernel size 3, followed by a max pooling, a fully connected layer, a leaky ReLU activation, and another fully connected layer. The group action is on the last two fully connected layers. After training a single model, we create transformed models using $g = I + \varepsilon M$, where $M \in \mathbb{R}^{32\times 32}$ is a random matrix and ε controls the magnitude of movement in the parameter space. We then use the mode of the transformed models' prediction as the final output.

We compare the ensemble formed by group actions to four ensembles formed by various random transformation. Let $g = I + \varepsilon M$. The random baselines are:

- 'group': $(U, V) \mapsto (U\pi(g, H), gV)$. This is the model created by group actions.
- 'g^{-1}': $(U, V) \mapsto (Ug^{-1}, gV)$.
- 'random': $(U, V) \mapsto (Ug', gV)$, where $g' = I + \varepsilon D$ and D is a random diagonal matrix.
- 'shuffle': $(U, V) \mapsto (U\pi'(g, H), gV)$, where $\pi'(g, H)$ is constructed by randomly shuffling $\pi(g, H)$.
- 'interpolated permute' or 'perm_interp': $(U, V) \mapsto (U\left(\frac{I+\frac{\varepsilon}{2}(I+S)}{I+\varepsilon}\right)^{-1}, \frac{I+\frac{\varepsilon}{2}(I+S)}{I+\varepsilon}V)$, where $S \in \mathbb{R}^{32\times 32}$ is a random permutation matrix.

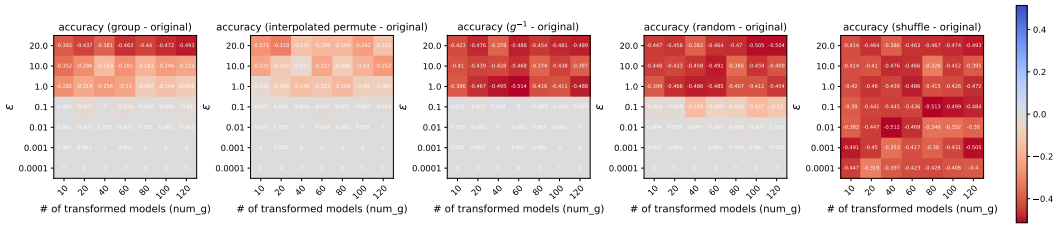
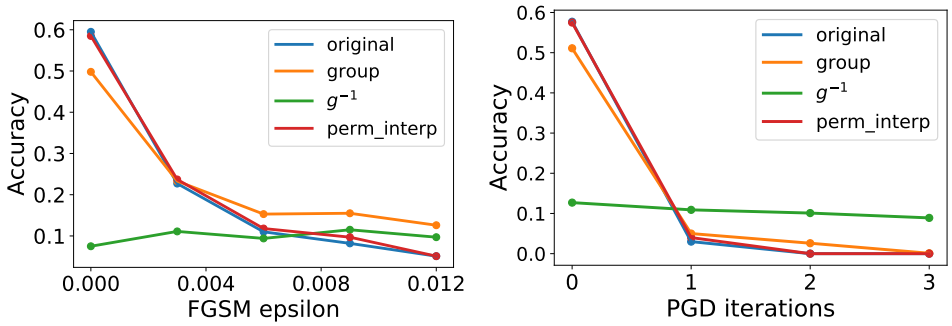


Figure 10: Change in accuracy compared to the original single model when using the ensemble model and 4 baselines. The red color indicates degradation in model performance. The ensemble created by group actions has similar loss values when ε is small.



(a) FGSM (b) PGD

Figure 11: Adversarial attacks on the original model and the ensemble models with various strengths. In FGSM, the group ensemble model improves robustness. In PGD, the ensemble has negligible effects on robustness.

Figure 10 shows the accuracy of the ensembles compared to single models. The ensemble formed by group actions preserves the model accuracy for small ε and has smaller accuracy drop at larger ε. The ensemble model also improves robustness against Fast Gradient Signed Method (FGSM) attacks (Figure 11). Under FGSM attacks with various strength, the ensemble model created using group actions consistently performs better than the baselines with random transformations. However, the same improvement is not observed under Projected Gradient Descent (PGD) attacks.