Understanding Mode Connectivity via Parameter Space Symmetry

Bo Zhao
University of California San Diego
bozhao@ucsd.edu

Nima Dehmamy
IBM Research
nima.dehmamy@ibm.com

Robin Walters
Northeastern University
r.walters@northeastern.edu

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

Abstract

It has been observed that the global minimum of neural networks is connected by curves on which train and test loss is almost constant. This phenomenon, often referred to as mode connectivity, has inspired various applications such as model ensembling and fine-tuning. Despite empirical evidence, a theoretical explanation is still lacking. We explore the connectedness of minimum through a new approach, parameter space symmetry. By relating topology of symmetry groups to topology of minima, we provide the number of connected components of full-rank linear networks. In particular, we show that skip connections reduce the number of connected components. We then prove mode connectivity up to permutation for linear networks. We also provide explicit expressions for connecting curves in minimum induced by symmetry.

1 Introduction

Among recent studies on the loss landscape, a particularly interesting discovery is mode connectivity [5, 10], which refers to the phenomenon that distinct minima found by stochastic gradient descent (SGD) can be connected by continuous paths through the high-dimensional parameter space of neural networks. Mode connectivity has implications on other phenomena in deep learning such as the lottery ticket hypothesis [8] and loss landscape and training trajectory analysis [11]. Additionally, mode connectivity has inspired applications in diverse fields, including model ensembling [10, 2, 3], model averaging [15, 30], pruning [8], improving adversarial robustness [34], and fine-tuning for altering prediction mechanism [20].

Discrete symmetry, especially permutation, is well-known to be related to mode connectivity. In particular, the neural network output is invariant to permuting the neurons [13]. [6] conjectures that all minima found by SGD are linearly connected up to permutation. Various algorithms have since been developed to find the optimal permutation for linear mode connectivity [1]. However, compared to discrete symmetry, the role of continuous symmetry remains less studied. Continuous symmetry groups with continuous actions define positive dimensional connected spaces in the minimum [32]. We explore the connectedness of minimum through continuous symmetries in the parameter space.

We reveal the role of symmetry in the connectivity of minimum by relating properties of topological groups to their orbits and the minimum. Our results show that both continuous and discrete symmetry are important and useful in understanding the origin and failure cases of mode connectivity. Our work highlights a new approach towards understanding the topology of the minimum and complements previous theories on mode connectivity [31, 9, 23, 24, 18, 27, 25].

Preprint.

2 Connectedness of minima

2.1 Linear network with invertible weights

Let **Param** be the space of parameters. Consider the multi-layer loss function L : **Param** $\to \mathbb{R}$,

$$L : \textbf{Param} \to \mathbb{R}, \qquad (W_1, ..., W_l) \mapsto ||Y - W_l...W_1X||_2^2. \tag{1}$$

where $X, Y \in \mathbb{R}^{h\times h}$ are the input and output of the network. In this subsection, we assume that both X, Y have rank h, and **Param** $= (\mathbb{R}^{h\times h})^l$. Then L has a $GL_h(\mathbb{R})^{l-1}$ symmetry, which acts on **Param** by $g \cdot (W_1, ..., W_l) = (g_1W_1, g_2W_2g_1^{-1}, ..., g_{l-1}W_{l-1}g_{l-2}^{-1}, W_lg_{l-1}^{-1})$, for $(g_1, ..., g_{l-1}) \in GL_h(\mathbb{R})^{l-1}$.

Let $L^{-1}(c) = \{\theta \in \textbf{Param} : L(\theta) = c\}$ be a level set of L. Since $\|\cdot\|_2 \geq 0$ and $L^{-1}(0) \neq \emptyset$, the minimum value of L is 0. By relating the topology of $GL(\mathbb{R})$ and $L^{-1}(0)$, we have the following observations on the structure of the minimum of L.

Proposition 2.1. *There is a homeomorphism between $L^{-1}(0)$ and $(\mathrm{GL}_h)^{l-1}$.*

Since $(\mathrm{GL}_h)^{l-1}$ has 2^{l-1} connected components and homeomorphism preserves topological properties, $L^{-1}(0)$ also has 2^{l-1} connected components.

Corollary 2.2. *The minimum of L has 2^{l-1} connected components.*

2.2 ResNet with 1D weights

The topological properties of the minimum depend on the architecture. As an example of this dependency, we show that adding a skip connection changes the number of connected components of the minimum.

Consider a residual network $W_3(W_2W_1X + \varepsilon X)$ and loss function

$$L(W_3, W_2, W_1) = ||Y - W_3(W_2W_1X + \varepsilon X)||_2, \tag{2}$$

where $(W_1, W_2, W_3) \in \textbf{Param} = \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$, $\varepsilon \in \mathbb{R}$, and data $X \in \mathbb{R}^{n\times n}, Y \in R^{n\times n}$. The following proposition states that for a three-layer residual network with weight matrices of dimension 1×1, the number of components of the minimum is smaller than that of a linear network without the skip connection.

Proposition 2.3. *Let $n = 1$. Assume that $X, Y \neq 0$. When $\varepsilon = 0$, the minimum of L has 4 connected components. When $\varepsilon \neq 0$, the minimum of L has 3 connected components.*

The $\varepsilon = 0$ case follows from Corollary 2.2. For the $\varepsilon \neq 0$ case, the proof decomposes the minimum of L into two sets S_1 and S_0, corresponding to the minima without the skip connection and an extra set of solutions because of the skip connection. S_1 is homeomorphic to $GL_1 \times GL_1$ and has 4 connected components. S_0 is a line and has 1 connected component. Two components of S_1 are connected to S_0, while the other two components of S_1 are not. Therefore, S_0 connects two components of S_1. As a result, the minimum of L has 3 connected components. Full proof can be found in Appendix C.3.

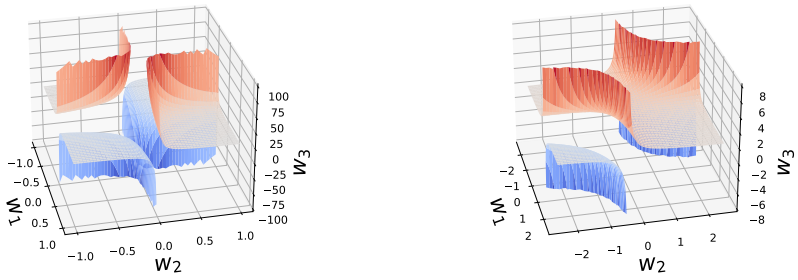
Figure 1 visualizes the minimum without and with the skip connection. This result reveals the effect of skip connection on the connectedness of minimum, which may lead to a new explanation of the effectiveness of ResNets [12] and DenseNets [14]. We leave the connection between the topology of minimum and the optimization and generalization property of neural networks to future work.

3 Mode connectivity

From the examples in the previous section, the connectedness of the minimum is related to the symmetry of the loss function under certain conditions. In this section, we explore applications of this insight in explaining mode connectivity.

3.1 Mode connectivity up to permutation

For the family of linear neural networks defined in Section 2.1, we show that permutation allows us to connect points in the minimum that are not connected without permutation. Our results support the empirical observation that neuron alignment by permutation improves mode connectivity [29].



(a) Linear (b) Resnet

Figure 1: Minimum of (a) 3-layer linear net $||Y - W_3W_2W_1X||_2$ and (b) 3-layer linear net with a residual connection $||Y - W_3(W_2W_1X + X)||_2$, where $X = 1$, $Y = 1$, and $W_1, W_2, W_3 \in \mathbb{R}$.

Consider again the linear network (1) with full rank weights. When $l = 2$, the minimum of L has 2 connected components. Any $g \in GL$ that is not on the identity component can take a point on one connected component of the minimum to the other.

Lemma 3.1. *Consider two points $(W_1, W_2), (W_1', W_2') \in L^{-1}(0)$ that are not connected in $L^{-1}(0)$. For any $g \in GL(h)$ such that $det(g) < 0$, $g \cdot (W_1, W_2)$ and (W_1', W_2') are connected in $L^{-1}(0)$.*

When the hidden dimension $h \geq 2$, there exists a permutation g such that $det(g) > 0$, and a permutation g such that $det(g) < 0$. Therefore, all points on the minimum of L are connected up to permutation.

Proposition 3.2. *Assume that $h \geq 2$. For all $(W_1, ..., W_l), (W_1', ..., W_l') \in L^{-1}(0)$, these exists a list of permutation matrices $P_1, ..., P_{l-1}$ such that $(W_1P_1, P_1^{-1}W_2P_2, ..., P_{l-2}W_{l-1}P_{l-1}, P_{l-1}W_l)$ and $(W_1', ..., W_l')$ are connected in $L^{-1}(0)$.*

3.2 Failure case of linear mode connectivity

In addition to helping show the connectedness of minimum, symmetry relates the set of points in minimum to groups and provides a way to find all points in the minimum. As an application, we show that linear mode connectivity fails to hold in multi-layer regressions. The following proposition says that in two-layer full-rank linear networks, the error barrier in the linear interpolation between two solutions can be arbitrarily large.

Proposition 3.3. *Consider the setting in Section 2.1. For any $k > 0$, there exist $(W_1, W_2), (W_1', W_2') \in L^{-1}(0)$ that belong to the same connected component of $L^{-1}(0)$ and $0 < \alpha < 1$, such that $L((1-\alpha)W_1 + \alpha W_1', (1-\alpha)W_2 + \alpha W_2') > k$.*

The result holds when there is a homogeneous activation ($\sigma(cz) = c^\alpha\sigma(z)$). However, when $\alpha \neq 1$, the proof needs a different choice of m. Figure 2 visualizes the two points on the minimum of a two-layer network with weights of dimension 1×1 and the linear interpolation between them. One possible reason why linear mode connectivity is observed in practice is that only a small part of the minima is reachable by SGD due to implicit bias [21].

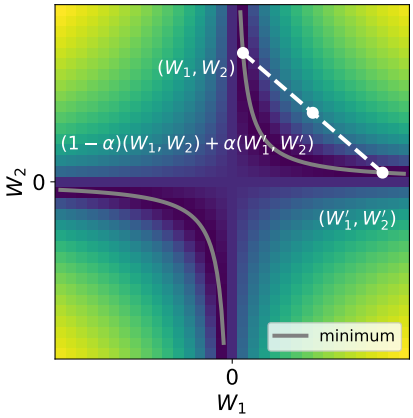


Figure 2: Interpolation between 2 minima of loss function $||Y - W_2W_1X||_2$ with 1 dimensional weights. Loss on the interpolation can be unbounded.

4 Curves on minimum from group actions

The minima of overparametrized ReLU networks consist of affine subspaces [28]. With activations that are not piecewise linear, the minimum may be curved. As a result, the paths connecting two points in the minimum may not be linear. Previously, these paths are discovered empirically by finding parametric curves on which the expected loss is minimized [10]. An alternative and principled way to find curves on the minima is to use parameter space symmetry.

Suppose the loss function $L : \textbf{Param} \to \mathbb{R}$ admits a G symmetry. Consider the following curve for a point $\boldsymbol{w} \in \textbf{Param}$ and $M \in \text{Lie}(G)$:

$$\gamma_M : \mathbb{R} \times \textbf{Param} \to \textbf{Param},$$
$$\gamma_M(t, \boldsymbol{w}) = \exp(tM) \cdot \boldsymbol{w}. \tag{3}$$

Since $\exp(tM) \in G$ and the action of G preserves the value of L, every point on γ_M is in the same L level set as $\boldsymbol{w}$. This provides a way to find a curve of constant loss between two points that are in the same orbit. Concretely, given two points $\boldsymbol{w}_1$ and $\boldsymbol{w}_2 = g \cdot \boldsymbol{w}_1$, let γ be the following curve:

$$\gamma : [0, 1] \times G \times \textbf{Param} \to \textbf{Param},$$
$$\gamma(t, g, \boldsymbol{w}) = \exp(t \log(g)) \cdot \boldsymbol{w}. \tag{4}$$

Note that $\gamma(0, g, \boldsymbol{w}_1) = \boldsymbol{w}_1$, $\gamma(1, g, \boldsymbol{w}_1) = \boldsymbol{w}_2$, and $L(\gamma(t, g, \boldsymbol{w}_1)) = L(\boldsymbol{w}_1) = L(\boldsymbol{w}_2)$ for all $t \in [0, 1]$. Hence, γ is a curve that connects the points $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, and every point on γ has the same loss value as $L(\boldsymbol{w}_1) = L(\boldsymbol{w}_2)$.

For a group G, the curve γ is defined when the map $\cdot : G \times \textbf{Param} \to \textbf{Param}$ is continuous and $\text{id} \cdot \boldsymbol{w} = \boldsymbol{w}$ for all $\boldsymbol{w} \in \textbf{Param}$, even if it is not a group action or does not preserve loss. However, when $\cdot$ does not preserve loss, the loss can change on γ. Consider our two-layer network and the following map:

$$\cdot : GL(h, \mathbb{R}) \times \textbf{Param} \to \textbf{Param}$$
$$g \cdot (U, V) = (U\sigma(VX)\sigma(gVX)^\dagger, gV). \tag{5}$$

When σ is the identity function, $\cdot$ preserves the loss value, and γ defines a curve on the minimum. In general, the map (5) does not preserve loss when batch size k is larger than hidden dimension h. However, the maximum change of loss on γ can be bounded as follows. Let $U', V' = g \cdot (U, V)$. We have

$$\|U\sigma(VX) - U'\sigma(V'X)\| = \|U\sigma(VX)\left(I - \sigma(gVX)^\dagger\sigma(gVX)\right)\| \leq \|U\sigma(VX)\|. \tag{6}$$

The last steps follows from the fact that $\sigma(gVX)^\dagger\sigma(gVX)$ is a projection.

5 Discussion

In this work, we study topological properties of the loss level sets by relating their topology to the topology of symmetry groups. We derive the number of connected components of full-rank multi-layer networks with and without skip connections, and prove mode connectivity up to permutation for full-rank linear regressions. Using symmetry in the parameter space, we construct an explicit expression for curves that connect two points in the same orbit.

While symmetry appears to be a useful tool for studying the loss landscape, our current results rely on the existence of a homeomorphism between symmetry groups and the minimum. A future direction is to explore the possibility of removing this assumption. Another interesting direction is to investigate additional links between different architecture choices, such as normalization, and connectedness of the minimum. On the application side, the impact of these results can benefit from further study on the connection between the topology of minimum and generalization ability of neural networks.

The connectedness results obtained from symmetry raise a number of interesting questions related to mode connectivity. For example, it would be interesting to understand when and why there is no significant change in loss on the linear interpolation between two minima. One possible explanation is that there always exists a γ defined in the way above that is close to the line formed by the linear interpolation. Another possible reason is that the dimension of minimum is usually high, and a significant part of the linear interpolation is within the minimum with high probability. Moreover, it has been observed that the train and test accuracy are both near constant on the paths that connect different SGD solutions [10]. If these paths correspond to a group action, this implies that the action's dependence on data is weak.

Acknowledgments and Disclosure of Funding

We thank Iordan Ganev for helpful comments on proofs in Appendix B. This work was supported in part by the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, NSF Grants #2205093, #2146343, and #2134274. R. Walters is supported by NSF grants #2107256 and #2134178.

References

[1] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *International Conference on Learning Representations*, 2023.

[2] 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*, pages 769–779. PMLR, 2021.

[3] Frederik Benzing, Simon Schug, Robert Meier, Johannes Von Oswald, Yassir Akram, Nicolas Zucchet, Laurence Aitchison, and Angelika Steger. Random initialisations performing above chance and how to find them. *14th Annual Workshop on Optimization for Machine Learning (OPT2022)*, 2022.

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

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

[6] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *International Conference on Learning Representations*, 2022.

[7] Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, and Aymeric Dieuleveut. Proving linear mode connectivity of neural networks via optimal transport. *arXiv preprint arXiv:2310.19103*, 2023.

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

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

[10] 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.

[11] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Using mode connectivity for loss landscape analysis. *35th International Conference on Machine Learning's Workshop on Modern Trends in Nonconvex Optimization for Machine Learning*, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers*, pages 129–135. Elsevier, 1990.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[15] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *Conference on Uncertainty in Artificial Intelligence*, 2018.

[16] Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. Repair: Renormalizing permuted activations for interpolation repair. *International Conference on Learning Representations*, 2023.

[17] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. *International Conference on Learning Representations*, 2023.

[18] Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019.

[19] John Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.

[20] Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR, 2023.

[21] 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*, pages 7760–7768. PMLR, 2021.

[22] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[23] Quynh Nguyen. On connected sublevel sets in deep learning. In *International conference on machine learning*, pages 4790–4799. PMLR, 2019.

[24] Quynh Nguyen. A note on connectivity of sublevel sets in deep learning. *arXiv preprint arXiv:2101.08576*, 2021.

[25] Quynh N Nguyen, Pierre Bréchet, and Marco Mondelli. When are solutions connected in deep networks? *Advances in Neural Information Processing Systems*, 34:20956–20969, 2021.

[26] 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*, pages 17759–17781, 2022.

[27] Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks. In *International Conference on Machine Learning*, pages 8773–8784. PMLR, 2020.

[28] 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*, pages 9722–9732. PMLR, 2021.

[29] Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing mode connectivity via neuron alignment. *Advances in Neural Information Processing Systems*, 33:15300–15311, 2020.

[30] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.

[31] David Yunis, Kumar Kshitij Patel, Pedro Henrique Pamplona Savarese, Gal Vardi, Jonathan Frankle, Matthew Walter, Karen Livescu, and Michael Maire. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[32] Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. *International Conference on Learning Representations*, 2023.

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

[34] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *International Conference on Learning Representations*, 2020.

[35] Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *arXiv preprint arXiv:2307.08286*, 2023.

Appendix

A Related Work

Mode connectivity [10] and [5] discover empirically that the global minimum of neural networks are connected by curves on which train and test loss is almost constant. It is then observed that SGD solutions are linearly connected if they are trained from pre-trained weights [22] or share a short period of training at the beginning [8]. Additionally, neuron alignment by permutation improves mode connectivity [29]. Then, [6] conjecture that all minima found by SGD are linearly connected up to permutation. Following the conjecture, [1] develop algorithms that finds the optimal alignment for linear mode connectivity, and [16] further reduce the barrier by rescaling the preactivations of interpolated networks. A few papers propose theoretical explanation of linear mode connectivity using different tools. [35] shows that the feature maps of each layer are also linearly connected and identify conditions that guarantees linear connectivity. [31] seeks to explain linear mode connectivity through finding a convex hull defined by SGD trajectory endpoints. [7] uses optimal transport theory to prove that wide two-layer neural networks trained using SGD are linearly connected with high probability. It is worth noting that linear mode connectivity does not always hold outside of computer vision. Language models that are not linearly connected have different generalization strategies [17]. [20] further show that the lack of linear connectivity indicates that the two models rely on different attributes to make predictions.

Theory on connectedness of minimum Several work explores the theoretical explanation of mode connectivity by studying the connectedness of sub-level sets. [9] shows that the minimum is connected for 2-layer linear network without regularization, and for deeper linear networks with $L2$ regularization. Futhermore, they show that the minimum of a two-layer ReLU network is asymptotically connected, that is, there exists a path connecting any two solutions with bounded error. [23] proves that the sublevel sets are connected in pyramidal networks with piecewise linear activation functions and first hidden layer wider than $2N$, where N is the number of training data). The width requirement is later improved to $N + 1$ [24]. [18] prove the existence of a piece-wise linear path between two solutions for ReLU networks, if they are both dropout stable, or both noise stable and sufficiently overparametrized. [27] generalizes this proof to show that wider neural networks are more connected, following the observation that SGD solutions for wider neural network are more dropout stable. [25] gives a new upper bound of the loss barrier between solutions using the loss of sparse subnetworks that are optimized, which is a milder condition than dropout stability.

Symmetry in the loss landscape Discrete symmetries have inspired a line of work on loss landscape topology. [4] shows that permutations of a given layer are connected within a loss level set. Through examining the permutation symmetries, [28] characterize the geometry of the global minima manifold for networks without other symmetries and show that adding one neuron to each layer in a minimal

network connects the permutation equivalent global minima. By removing permutation and rescaling symmetries, [26] study the geometry of minima in the functional space. [32] finds a set of nonlinear continuous symmetries that partially parametrizes the minimum. [33] uses symmetry induced curves to approximate the curvature of the minimum.

B Background

In this section, we review mathematical concepts used in the paper and list some useful results on the number of connected components of topological spaces. We refer readers to [19] for a more detailed introduction to this topic.

B.1 Connected components

Consider two topological spaces X and Y. A map $f : X \to Y$ is *continuous* if for every open subset $U \subseteq Y$, its preimage $f^{-1}(U)$ is open in X. If X and Y are metric spaces with metrics d_X and d_Y respectively, this is equivalent to the delta-epsilon definition. That is, f is continuous if at every $x \in X$, for any $\epsilon > 0$ there exists $\delta > 0$ such that $d_X(x, y) < \delta$ implies $d_Y(f(x), f(y)) < \epsilon$ for all $y \in Y$.

A topological space is *connected* if it cannot be expressed as the union of two disjoint, nonempty, open subsets. A topological space X is *path connected* if for every $p, q \in X$, there is a continuous map $f : [0, 1] \to X$ such that $f(0) = p$ and $f(1) = q$. Path connectedness implies connectedness, but the converse is not true [19]. [23] studies the path connectedness of sublevel sets of loss functions.

The following theorem is the main intuition of this paper and will appear frequently in proofs.

Theorem B.1 (Theorem 4.7 in [19])**.** *Let X, Y be topological spaces and let $f : X \to Y$ be a continuous map. If X is connected, then $f(X)$ is connected.*

A map f is a *homeomorphism* from X to Y if f is bijective and both f and f^{-1} are continuous. X and Y are *homeomorphic* if such a map exists. A *(connected) component* of a topological space X is a maximal nonempty connected subset of X. The components of X form a partition of X. The next two corollaries of Theorem B.1 show that connectedness and the number of connected components are topological properties. That is, they are preserved under homeomorphisms.

Corollary B.2. *Let $f : X \to Y$ be a homeomorphism from X to Y, and let $U \subseteq X$ be a subset of X with the subspace topology. Then U is connected if and only if $f(U) \subseteq Y$ is connected.*

Proof. By the definition of homeomorphism, f and f^{-1} are continuous. From Theorem B.1, if $U \in X$ is connected, then $f(U) \in Y$ is connected. Similarly, if $f(U)$ is connected, then $f^{-1}(f(U)) = U$ is connected. □

Corollary B.3. *Let X be a topological space that has N components. Let Y be a topological space homeomorphic to X. Then Y has N components.*

Proof. Let $C_1, ..., C_N$ be the components of X. Let f be a homeomorphism from X to Y. Since f is bijective and $C_1, ..., C_N$ is a partition of X, $f(C_1), ..., f(C_N)$ is a partition of Y. From Theorem B.1, since $C_1, ..., C_N$ are all connected, so are $f(C_1), ..., f(C_N)$.

Lastly, we need to show that $f(C_1), ..., f(C_N)$ are maximally connected. Suppose there exists a set $U \subseteq Y$, such that $U \nsubseteq f(C_i)$ and $f(C_i) \cup U$ is connected for some i. Then by Theorem B.1, $f^{-1}(f(C_i) \cup U) \supset C_i$ is connected in X. This contradicts the fact that C_i is a maximal component in X. Therefore, $f(C_1), ..., f(C_N)$ are maximally connected.

Since $f(C_1), ..., f(C_N)$ partitions Y and are maximally connected, Y has N components. □

Another consequence of Theorem B.1 is the following upper bound on the number of components of the image of a continuous map.

Proposition B.4. *Let $f : X \to Y$ be a continuous map. The number of components of the image $f(X) \subseteq Y$ is at most the number of components of X.*

Proof. Let $C_1, ..., C_N$ be the components of X. Since C_i is continuous and the action is continuous, according to Theorem B.1, $f(C_i)$ is continuous for all $i \in \{1, ..., N\}$. Additionally, since $\bigcup_{i=1}^{N} C_i = X$, we have $\bigcup_{i=1}^{N} f(C_i) = f(X)$. Therefore, there is a surjective map from $\{f(C_1), ..., f(C_N)\}$ to the set of components of $f(X)$, which implies that $f(X)$ has at most N components. □

Let $X_1, ..., X_n$ be topological spaces. The *product space* is their Cartesian product $X_1 \times ... \times X_n$ endowed with the product topology. Denote $\pi_0(X)$ as the set of connected components of a space X. The following proposition provides a way to count the components of a product space.

Proposition B.5. *Consider n topological spaces $X_1, ..., X_n$. Then $|\pi_0(X_1 \times ... \times X_n)| = \prod_{i=0}^{n} |\pi_0(X_i)|$.*

Proof. When $n = 1$, the number of components of the product space is $|\pi_0(X_1)|$.

For the $n > 1$ case, since $X_1 \times ... \times X_n = (X_1 \times ... \times X_{n-1}) \times X_n$, it suffices to show that $|\pi_0(A \times B)| = |\pi_0(A)||\pi_0(B)|$ for any topological spaces A and B. Let $f : \pi_0(A) \times \pi_0(B) \to \pi_0(A \times B)$ be the map that assigns $C \in \pi_0(A) \times \pi_0(B)$ to the element in $\pi_0(A \times B)$ that contains C. Then f is surjective because $\pi_0(A) \times \pi_0(B)$ forms a partition of $A \times B$. To prove that f is injective, suppose that $f(C_1) = f(C_2)$ for $C_1, C_2 \in \pi_0(A) \times \pi_0(B)$. Consider the projection $\pi_A : A \times B \to A$. Since π_A is continuous and C_1, C_2 belong to the same component of $A \times B$, $\pi_A(C_1)$ and $\pi_A(C_2)$ belong to the same component of A. Similarly, $\pi_B(C_1)$ and $\pi_B(C_2)$ belong to the same component of B under the projection $\pi_B : A \times B \to B$. Since all components of A and B are maximally connected, we have $C_1 = C_2$, which implies that f is injective. Since f is a bijection from $\pi_0(A) \times \pi_0(B)$ to $\pi_0(A \times B)$, $|\pi_0(A \times B)| = |\pi_0(A)||\pi_0(B)|$. □

B.2 Groups

A *group* is a set G together with a law of composition, that satisfies $(ab)c = a(bc)$ $\forall a, b, c \in G$, $\exists 1$ such that $1a = a1 = a$ $\forall a \in G$, and $\forall a \in G$, $\exists b$ such that $ab = ba = 1$. An *action* of a group G on a set S is a map $\cdot : G \times S \to S$, that satisfies $\mathrm{id} \cdot s = s$ for all $s \in S$ and $(gg') \cdot s = g \cdot (g' \cdot s)$ for all g, g' in G and all s in S. The *orbit* of $s \in S$ is the set $O(s) = \{s' \in S \mid s' = gs \text{ for some } g \text{ in } G\}$.

A *topological group* is a group G endowed with a topology such that multiplication and inverse are both continuous. A recurring example is the general linear group $GL_n(\mathbb{R})$, with the subspace topology obtained from $\mathbb{R}^{n^2}$. $GL_n(\mathbb{R})$ has two connected components, which correspond to the preimages of the positive and negative reals under the determinant map.

The *product* of groups $G_1, ..., G_n$ is a group denoted by $G_1 \times ... \times G_n$. The elements in $G_1 \times ... \times G_n$ is the product set of $G_1, ..., G_n$. The group structure is defined by identity $(1, ..., 1)$, inverse $(g_1, ..., g_n)^{-1} = (g_1^{-1}, ..., g_n^{-1})$, and multiplication rule $(g_1, ..., g_n)(g_1', ..., g_n') = (g_1 g_1', ..., g_n g_n')$.

B.3 Relating connectedness of groups, orbits, and level sets

From Theorem B.1, continuous maps preserve connectedness. Through continuous actions, we study the connectedness of orbits and level sets by relating them to the connectedness of more familiar objects such as the general linear group.

In the main text, we focus on the case of bijective actions. Here we only assume the action to be continuous and try to bound the number of components of the orbits. As an immediate consequence of Proposition B.4, an orbit cannot have more components than the group.

Corollary B.6. *Assume that the action of a group G on S is continuous. Then the number of connected components of orbit $O(s)$ is smaller than or equal to the number of connected components of G, for all s in S.*

Proof. An orbit $O(s)$ is the image of the group action, which we assume to be continuous. The result follows from Proposition B.4. □

Let X be a topological space and $L : X \to \mathbb{R}$ a continuous function on X. A topological group G is said to be a *symmetry group* of L if $L(g \cdot x) = L(x)$ for all $g \in G$ and $x \in X$. In this case, the action can be defined on a level set of L, $L^{-1}(c)$ with a $c \in \mathbb{R}$, as $G \times L^{-1}(c) \to L^{-1}(c)$. If the minimum

of L consists of a single orbit, Corollary B.6 extends immediately to the number of components of the minimum.

Corollary B.7. *Let L be a function with a symmetry group G. If the minimum of L consists of a single G-orbit, then the number of connected components of the minimum is smaller or equal to the number of connected components of G.*

Generally, symmetry groups do not act transitively on a level set $L^{-1}(c) \in$ **Param**. In this case, the connectedness of the orbits does not directly inform the connectedness of the level set.

Proposition B.8.

(a) There exists a space X, a group G, and an action of G, such that each orbit for the action of G is connected and X is not connected.

(b) There exists a space X, a group G, and an action of G, such that each orbit for the action of G is disconnected and X is connected.

Proof. For part (a), consider a subspace of $\mathbb{R}^2$, $X = X_1 \cup X_2$ where $X_1 = \{(x, y) : x = 0, y > 0\}$ and $X_2 = \{(x, y) : x = 1, y > 0\}$. The space X is not connected. Let G be the multiplicative group of positive real numbers and act on X by multiplication on the second coordinate. Then there are two orbits, X_1 and X_2, which are both connected.

For part (b), consider the space $X = \mathbb{R}^2 \setminus \{0\}$. Then X is connected. Let G be the multiplicative group of real numbers, which acts on X by multiplication on both coordinates. That is, $g \cdot (x_1, x_2) = (gx, gx_2), \forall (x_1, x_2) \in X, \forall g \in G$. The orbit of any point $(x_1, x_2) \in X$ is not connected. □

Nevertheless, since the set of orbits partitions the space, we can use the following bound on the number of components of the space.

Proposition B.9. *Let X be a topological space and let $X = \coprod_i X_i$ be a partition of X into disjoint subspaces. Then $|\pi_0(X)| \leq \sum_i |\pi_0(X_i)|$.*

Proof. Let $S = \{A \subseteq X : \exists i, A \text{ is a component of } X_i\}$ be the union of the components of the subspaces. Then S is a partition of X, and every element in S is connected. Therefore, there is a surjective map from S to $\pi_0(X)$, defined by mapping each $s \in S$ to the element of $\pi_0(X)$ that includes s. This implies that $|\pi_0(X)| \leq |S| = \sum_{i=1}^{n} |\pi_0(X_i)|$. □

Consider a topological space X and a group G that acts on X. Let $O = \{O_1, ..., O_n\}$ be the set of orbits. By Proposition B.9, the number of components of the orbits give the following upper bound on the number of components of the space: $|\pi_0(X)| \leq \sum_{i=1}^{n} |\pi_0(O_i)|$.

C Missing Proofs

C.1 Proof of Proposition 2.1

Proof. Recall that $W_1, ..., W_n, X, Y$ are matrices in $\mathbb{R}^{h \times h}$, and X, Y are both full rank. Consider the map

$$f : (\mathrm{GL}_h)^{l-1} \to L^{-1}(0), \quad (g_1, ..., g_{l-1}) \mapsto (g_1 X^{-1}, g_2, ..., g_{l-1}, Y \prod_{i}^{l-1} g_i^{-1}). \tag{7}$$

The inverse $f^{-1} : (W_1, ..., W_l) \mapsto (W_1 X, W_2, W_3, ..., W_{l-1})$ is well defined, because X, $W_1, W_2, W_3, ..., W_{l-1}$ are all full-rank. Since both f and f^{-1} are continuous, f is a homeomorphism between $(\mathrm{GL}_h)^{l-1}$ and $L^{-1}(0)$. □

C.2 Proof of Corollary 2.2

Proof. From Proposition 2.1, $L^{-1}(0)$ is homeomorphic to $(\mathrm{GL}_h)^{l-1}$. According to Corollary B.3, this implies that $L^{-1}(0)$ has the same number of connected components as $(\mathrm{GL}_h)^{l-1}$. From Proposition B.5, $GL_h(\mathbb{R})^{l-1}$ has 2^{l-1} connected components. Therefore, $L^{-1}(0)$ has 2^{l-1} connected components. □

C.3 Proof of Proposition 2.3

Proof. When $\varepsilon = 0$, the skip connection is effectively removed, and the loss function (2) reduces to (1). By Corollary 2.2, the minimum of L has 4 connected components. In the rest of the proof, we consider the case where $\varepsilon \neq 0$.

Let $(W_{1_0}, W_{2_0}, W_{3_0}) = (I, (\alpha - \varepsilon)I, \alpha^{-1}YX^{-1})$, where $\alpha \in \mathbb{R}$ is an arbitrary number such that $\alpha \neq \varepsilon$ and $\alpha \neq 0$. Then $(W_{1_0}, W_{2_0}, W_{3_0})$ is a point in $L^{-1}(0)$. Define set $G_1 = \{g \in R^{h\times h} : \det(gW_{2_0}W_{1_0}X + \varepsilon X) \neq 0\}$. Let $a : GL_1 \times G_1 \to$ **Param** be the following map:

$$\begin{aligned} g_1, g_2 \mapsto (&g_1 W_{1_0}, \\ &g_2 W_{2_0} g_1^{-1}, \\ &W_{3_0}(W_{2_0}W_{1_0}X + \varepsilon X)(g_2 W_{2_0}W_{1_0}X + \varepsilon X)^{-1}). \end{aligned} \tag{8}$$

From the definition of G_1, $(g_2W_{2_0}W_{1_0}X + \varepsilon X)$ is invertible, so a is well defined. Additionally, we have $L(a(g_1, g_2)) = L(W_{1_0}, W_{2_0}, W_{3_0}) = 0, \forall g_1, g_2 \in GL_1 \times G_1$. Therefore, denoting the image of a as S_1, we have $S_1 \subseteq L^{-1}(0)$.

Let $S_0 = \{(W_1, W_2, W_3) : W_3 = Y(\varepsilon X)^{-1}$ and $W_1 = 0\}$ if $\varepsilon \neq 0$, or $\emptyset$ otherwise. For $(W_1, W_2, W_3) \in S_0$, we have $L(W_1, W_2, W_3) = ||Y - Y(\varepsilon X)^{-1}(0 + \varepsilon X)||_2 = 0$. Therefore, $S_0 \subseteq L^{-1}(0)$.

We then show that the minimum of L is the union of S_1 and S_0. Consider a point $(W_1, W_2, W_3) \in L^{-1}(0)$. If $W_1 = 0$, then $\varepsilon \neq 0$, otherwise (W_1, W_2, W_3) cannot be in $L^{-1}(0)$. In this case, W_3 must equal to $Y(\varepsilon X)^{-1}$, and $(W_1, W_2, W_3) \in S_0$. If $W_1 \neq 0$, then $W_1W_{1_0}^{-1} \in GL_1$ and $W_2W_1W_{1_0}^{-1}W_{2_0}^{-1} \in G_1$. The second part is due to $W_2W_1W_{1_0}^{-1}W_{2_0}^{-1}W_{2_0}W_{1_0}X + \varepsilon X = W_2W_1X + \varepsilon X \neq 0$ since $(W_1, W_2, W_3) \in L^{-1}(0)$. In this case we have $(W_1, W_2, W_3) = a(W_1W_{1_0}^{-1}, W_2W_1W_{1_0}^{-1}W_{2_0}^{-1})$, which means that $(W_1, W_2, W_3) \in S_1$.

The number of connected components of S_1 and S_0 can be obtained from their structures. Since $W_{2_0}W_{1_0}X \neq 0$, there is a homeomorphism between G_1 and GL_1 defined by the map

$$f : G_1 \to GL_1, g \mapsto gW_{2_0}W_{1_0}X + \varepsilon X \tag{9}$$

with inverse $f^{-1} : GL_1 \to G_1, g \mapsto \varepsilon(g - \varepsilon X)(W_{2_0}W_{1_0}X)^{-1}$. Since a is also a homeomorphism, its image S_1 is homeomorphic to $GL_1 \times GL_1$ and has 4 connected components. When $\varepsilon \neq 0$, S_0 is a line and thus has 1 connected component.

The last part of the proof shows the connectedness of the connected components of S_1 and S_0. Let $G_1^+ = \{g_2 \in G_1 : f(g_2) \in GL^{sign(\varepsilon X)}\}$ be the connected component in G_1 that correspond to $GL^{sign(\varepsilon X)}$, and $G_1^- = \{g_2 \in G_1 : f(g_2) \in GL^{-sign(\varepsilon X)}\}$ be the component that correspond to $GL^{-sign(\varepsilon X)}$. For convenience, we name the connected components of $Im(a)$ as follows:

$$\begin{aligned} C_1 &= \{(W_1, W_2, W_3) \in \textbf{Param} : (W_1, W_2, W_3) = a(g_1, g_2), g_1 \in GL^+, g_2 \in G_1^+\} \\ C_2 &= \{(W_1, W_2, W_3) \in \textbf{Param} : (W_1, W_2, W_3) = a(g_1, g_2), g_1 \in GL^-, g_2 \in G_1^+\} \\ C_3 &= \{(W_1, W_2, W_3) \in \textbf{Param} : (W_1, W_2, W_3) = a(g_1, g_2), g_1 \in GL^+, g_2 \in G_1^-\} \\ C_4 &= \{(W_1, W_2, W_3) \in \textbf{Param} : (W_1, W_2, W_3) = a(g_1, g_2), g_1 \in GL^-, g_2 \in G_1^-\} \end{aligned}$$

Note that for $(W_1, W_2, W_3) \in S_1$, there exists a (unique) $g_2 \in G_1$ such that we can write W_3 as

$$W_3 = W_{3_0}[W_{2_0}W_{1_0}X + \varepsilon X][g_2W_{2_0}W_{1_0}X + \varepsilon X]^{-1}) = Yf(g_2)^{-1}.$$

Following from the definition of G_1^+, for a point (W_1, W_2, W_3) in C_1 or C_2, $sign(W_3) = sign(Y(\varepsilon X)^{-1})$. Additionally, when g_2 is close to 0, g_2 belongs to G_1^+. The boundary of both C_1 and C_2 contain a point in S_0:

$$\lim_{g_1 \to 0^+} a(g_1, g_1) = \lim_{g_1 \to 0^-} a(g_1, g_1) = (0, \alpha - \varepsilon, Y(\varepsilon X)^{-1}) \in S_0.$$

Therefore, both C_1 and C_2 are connected to S_0.

For points in C_3 and C_4, $sign(W_3) \neq sign(Y(\varepsilon X)^{-1})$. Therefore, no point in C_3 or C_4 can be sufficiently close to S_0. As a result, these components are not connected to S_0. In summary, when $\varepsilon \neq 0$, S_0 connects 2 components of S_1, and the minimum of L has 3 connected components. □

C.4 Proof of Lemma 3.1

Proof. Consider the map f and its inverse f^{-1} defined in (7) in the proof of Proposition 2.1. Let $g = f^{-1}(W_1, W_2)$ and $g' = f^{-1}(W_1', W_2')$. By Corollary B.2, since (W_1, W_2) and (W_1', W_2') are not in the same connected component of $L^{-1}(0)$, g and g' are not in the same connected component of GL_h. Equivalently, $det(gg') < 0$. Consider a $g_1 \in GL_h$ such that $det(g) < 0$. Then $det(g_1 gg') > 0$, which means that $g_1 g$ and g' belong to the same connected component of GL_h. Therefore, according to Corollary B.2, $g_1 \cdot (W_1, W_2) = f(g_1 g)$ and $(W_1', W_2') = f(g')$ belong to the same connected component of $L^{-1}(0)$. □

C.5 Proof of Proposition 3.2

Proof. Let $(g_1, ..., g_{l-1}), (g_1', ..., g_{l-1}') \in (GL_h)^{n-1}$ such that $f(g_1, ..., g_{l-1}) = (W_1, ..., W_l)$ and $f(g_1', ..., g_{l-1}') = (W_1', ..., W_l')$. Let $P_0 = I$. For $i = 1, ..., l-1$, if $det(g_i g_i' P_{i-1}^{-1}) > 0$, set P_i to I. Otherwise, we set P_i to an arbitrary element in $P \in S_h \setminus A_h$, which is not empty when $h \geq 2$.

Let $(g_1'', ..., g_{l-1}'') \in (GL_h)^{n-1}$ such that $f(g_1'', ..., g_{l-1}'') = (W_1 P_1, P_1^{-1} W_2 P_2, ..., P_{l-2} W_{l-1} P_{l-1}, P_{l-1} W_l)$. By the way we construct P_i's, we have $g_i'' = P_{i-1}^{-1} g_i' P_i$ and $det(g_i g_i'') > 0$. Therefore, g_i and g_i'' belong to the same connected component of $(GL_h)^{l-1}$ for all i. Since f is a homeomorphism between $(\mathrm{GL}_h)^{l-1}$ and $L^{-1}(0)$, $(W_1 P_1, P^{-1} W_2 P_2, ..., P_{l-2} W_{l-1} P_{l-1}, P_{l-1} W_l)$ and $(W_1', ..., W_l')$ are connected in $L^{-1}(0)$. □

C.6 Proof of Proposition 3.3

Proof. Let $(W_2, W_1) \in L^{-1}(0)$ be an arbitrary point on the minimum of L. Let $\alpha = 0.5$, $m = \frac{4\sqrt{k}}{\|Y\|_2} + 2$, and $(W_2', W_1') = (W_2 m^{-1}, m W_1)$. Then

$$
\begin{aligned}
& L((1-\alpha)W_1 + \alpha W_1', (1-\alpha)W_2 + \alpha W_2') \\
=& ||Y - ((1-\alpha)W_2 + \alpha W_2')((1-\alpha)W_1 + \alpha W_1')X||_2^2 \\
=& ||Y - (1-\alpha)^2 W_2 W_1 X - \alpha^2 W_2' W_1' X - (1-\alpha)\alpha(m + m^{-1}) W_2 W_1 X||_2^2 \\
=& ||(1 - (1-\alpha)^2 - \alpha^2 - (1-\alpha)\alpha(m + m^{-1}))Y||_2^2 \\
=& ||(2\alpha(1-\alpha) - \alpha(1-\alpha)(m + m^{-1}))Y||_2^2 \\
=& ||(\alpha(1-\alpha)(2 - m - m^{-1})Y||_2^2 \\
=& (2 - m - m^{-1})^2 0.25^2 ||Y||_2^2
\end{aligned}
\tag{10}
$$

Note that $\alpha(1-\alpha) = 0.25$ and $m + m^{-1} > 2$. Substitute in m, we have

$$
L((1-\alpha)W_1 + \alpha W_1', (1-\alpha)W_2 + \alpha W_2') > ||(2-m)^2 0.25^2 ||Y||_2^2 = k. \tag{11}
$$

□