TOWARDS A COMPLETE THEORY OF NEURAL NETWORKS WITH FEW NEURONS

Anonymous authors
Paper under double-blind review

ABSTRACT

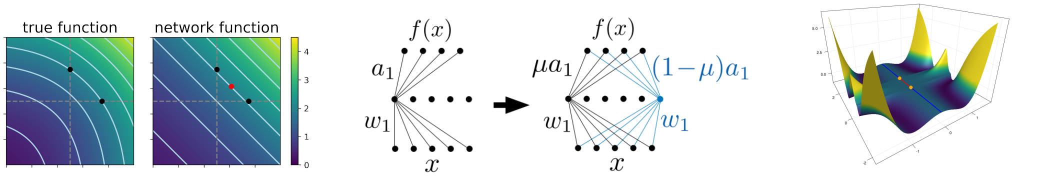
Deep learning has seen unprecedented progress thanks to the deployment of models with millions of parameters. On the theoretical side, an immense amount of effort has gone into understanding the dynamics of overparameterized networks. Although there now is a well-developed theory of networks with infinitely many neurons, the classic problem of understanding how a neural network with a few neurons learns remains unsolved. To attack this problem, we analytically study the landscapes of neural networks with few neurons. For the one-neuron network learning from a teacher network with arbitrarily many orthogonal incoming vectors, we prove that the optimal neuron implements a damped average in its incoming vector and balances the missing neurons in its outgoing weight. In addition, we prove how a neuron splitting mechanism turns a minimum into a line of critical points with transitions from saddles to local minima via non-strict saddles. Finally, we discuss how the insights we get from our novel proof techniques may shed light on the dynamics of neural networks with few neurons.

1 INTRODUCTION

The theory of deep learning has advanced rapidly and we know now that over-parameterized shallow networks provably converge to a global minimizer for typical initializations in the presence of infinitely many neurons (Jacot et al., 2018; Chizat & Bach, 2018; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2020) or a polynomial number of neurons as a function of the number of training data points suffices for the same guarantee (Du et al., 2018; Arora et al., 2019; Oymak & Soltanolkotabi, 2020). However, a good theoretical understanding of underparametrized shallow neural networks is still lacking. The main question we are interested in is: *"Can we characterize the compression strategy of an underparameterized neural network?"*

In this work, we take the first steps towards understanding the structure of the solution found by a neural network with few neurons. We first develop a rigorous analysis for the one-neuron network learning from a teacher network with multiple neurons. Although there has been some work in this direction, it is limited to the case when the teacher network also has one neuron (Tian, 2017; Yehudai & Ohad, 2020; Vardi et al., 2021; Wu, 2022). For a teacher network with multiple orthogonal incoming vectors, we show that the one-neuron network compresses the teacher neurons into a single neuron by implementing a damped average in its incoming vector and by balancing the missing neurons with its outgoing weight.

Going beyond the one-neuron network, we investigate the compression strategy implemented by neural networks with few neurons. Building upon Fukumizu & Amari (2000); Şimşek et al. (2021), we first give a detailed second-order analysis of the neuron splitting mechanism for neural networks with arbitrary width and depth. In particular, we prove under which conditions the splitting of neurons leads to harmful local minima or harmless strict saddles, revealing intriguing transitions from saddles to minima on a line. Then we study the structure of the minima found by underparameterized neural networks by growing them – one neuron at a time – in the teacher-student setting, similar to Wu et al. (2019). All our empirical results are based on integrating the gradient flow of the loss function with a higher order differential equation solver, thereby overcoming some of the numerical difficulties encountered in previous works.



(a) one-neuron network **(b)** neuron splitting **(c)** line of critical points

Figure 1: *Summary of contributions.* **(a)** The optimal solution of a one-neuron network learning from a true function generated by multiple teacher neurons: its incoming vector implements a damped average of the teacher incoming vectors, its outgoing weight is larger than the number of teacher neurons. **(b)** Neuron splitting (as shown in the figure) preserves criticality and the network function. **(c)** In the loss landscape of the two-neurons network, there exists a line of critical points originating from the minimum of the one-neuron network, and identical to it in terms of representing the same function. The same neuron splitting principle applies to neural networks with arbitrary width and depth, resulting in intriguing transitions in the second-order derivatives of the loss at the so-called symmetry-induced critical points. An example is shown in the figure where strict saddles (blue) transit into local minima (red) through non-strict saddles (orange).

1.1 Related Work

The classic line of work in teacher-student setting approximates the gradient flow ODEs when the input dimension grows to infinity (Saad & Solla, 1995), an approach that has been made rigorous by Goldt et al. (2020). The limiting dynamics in finite-size learning rate regime has been studied in recent work (Veiga et al., 2022; Arous et al., 2022). Another line of work focused on studying overparameterization in teacher-student setup in finite input dimension regime. Soltanolkotabi et al. (2018) prove that there is no local minimum if the number of neurons is twice more than the input dimension, but using quadratic activation functions. For differentiable activation functions, they prove convergence to a global minimum only when the initialization is close to a minimum in a teacher student setup, where the student has the same number of neurons as the teacher; and both bigger than the input dimension. When the number of student neurons, teacher neurons, and the input dimension are equal, Safran & Shamir (2018) show that local minima exist for ReLU activation and some of the families of minima are characterized by (Arjevani & Field, 2021). Moreover, despite the vast literature on modeling the true function as a teacher network, there is yet no rigorous result on the solution found by the one-neuron network.

An orthogonal line of work studies a neuron splitting mechanism in neural networks from a small width into a large width. Fukumizu & Amari (2000) study the splitting of one neuron at a critical point and the resulting line of critical points. This is generalized to the splitting of multiple neurons of the same type in Fukumizu et al. (2019). Şimşek et al. (2021) study the combinatorics of the problem by allowing the splitting of all types of neurons and characterizes the scaling of the number of manifolds of critical points as well the number of components of the global minima manifold. Zhang et al. (2021) provide some further discussions and numerics. Jacot et al. (2021); Boursier et al. (2022) analyze a training regime where the gradient flow visits a sequence of saddles.

2 Preliminaries and Overview

Consider a network function $f : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$ of a two layer network with n neurons

$$f(x) = \sum_{j=1}^{n} a_j \sigma\left(w_j \cdot x\right) \tag{1}$$

where $w_j \in \mathbb{R}^d$ represents an incoming vector and $a_j \in \mathbb{R}^{d_{\text{out}}}$ represents an outgoing vector of a neuron. To emphasize the structure of a network function in terms of the summation of its neurons $x \to a_j\sigma(w_j \cdot x)$, we denote the parameter $\theta \in \mathbb{R}^P$ with $P = (d + d_{\text{out}})n$ as follows

$$\theta = (w_1, a_1) \oplus \cdots \oplus (w_n, a_n). \tag{2}$$

Sometimes the parameter θ is made explicit in the network function $f(x|\theta) = f(x)$. The input distribution is arbitrary and denoted by $\mathbb{P}$ (it can be discrete or continuous).

We denote the mapping from inputs to targets by $f^* : \mathbb{R}^d \to \mathbb{R}^{d_{\text{out}}}$. The cost function $c : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}$ is assumed to be twice-differentiable in its first component. The loss is then defined as

$$L(\theta) = \langle c(f(x|\theta), f^*(x)) \rangle_{\mathbb{P}} \,. \tag{3}$$

2.1 Neural Networks with Few Neurons

Specifically, we assume that the input data distribution is d-dimensional standard Gaussian, i.e. $x \sim \mathcal{N}(0, I)$, the cost function is quadratic, i.e. $c(\hat{y}, y) = (\hat{y} - y)^2$, and the targets are generated by a teacher network

$$f^*(x) = \sum_{i=1}^{k} \sigma(v_i \cdot x) \tag{4}$$

where $v_1, \ldots, v_k \in \mathbb{R}^d$ are orthogonal vectors with unit norm. We drop the subscript $\mathbb{P}$ in Eq. 3 when the input distribution is standard Gaussian which is commonly studied Tian (2017); Zhong et al. (2017); Safran & Shamir (2018); Soltanolkotabi et al. (2018). The loss of a neural network with n neurons (w_j, a_j) is then given by

$$L((w_j, a_j)_{j=1}^n) = \left\langle (\sum_{j=1}^{n} a_j \sigma(w_j \cdot x) - \sum_{i=1}^{k} \sigma(v_i \cdot x))^2 \right\rangle . \tag{5}$$

Thanks to the Gaussianity of the input data, the loss can be reparametrized such that it only depends on outgoing weights a_j, incoming vector norms $r_j = \|w_j\|$, and (cosine-)similarity between pairs of normalized incoming vectors $u_{ij} = w_j/r_j \cdot v_i$ (student and teacher) and $\tilde{u}_{jj'} = w_j/r_j \cdot w'_j/r'_j$ (both student). The loss can then be expressed as

$$L((w_j, a_j)_{j=1}^n) = \sum_{j,j'=1}^{n} a_j a_{j'} g(r_j, r_{j'}, \tilde{u}_{jj'}) - 2 \sum_{j=1}^{n} \sum_{i=1}^{k} a_j g(r_j, 1, u_{ji}) + \text{const}, \tag{6}$$

$$\text{subject to} \sum_{i=1}^{k} u_{ji}^2 \leq 1 \; \forall j \in [k], \tag{7}$$

$$\left| \tilde{u}_{jj'} - \sum_{i=1}^{k} u_{ji} u_{j'i} \right| \leq \sqrt{1 - \sum_{i=1}^{k} u_{ji}^2} \sqrt{1 - \sum_{i=1}^{k} u_{j'i}^2} \; \forall j' \neq j; \tag{8}$$

where the potential $g : \mathbb{R}^2_{\geq 0} \times [-1, 1] \to \mathbb{R}$ is given by the following Gaussian average

$$g(r_1, r_2, u) = \langle \sigma(z_1)\sigma(z_2) \rangle = \langle \sigma(w_1 \cdot x)\sigma(w_2 \cdot x) \rangle, \tag{9}$$

where $z_1 = w_1 \cdot x$ and $z_2 = w_2 \cdot x$ are univariate centered Gaussians with standard deviations r_1 and r_2 respectively, and with correlation $u = (w_1/r_1) \cdot (w_2/r_2)$.

2.2 Neuron Splitting Mechanism

We review a general mapping mechanism between the critical points of a narrow network into a wider network which we develop further in Section 4. If an outgoing vector is zero, we can drop the corresponding neuron without changing the network function, similarly, if two incoming vectors are equivalent, we can merge them into one neuron without changing the network function. If a parameter does not allow reducing any of its neurons via one of these two operations, we call it *irreducible* following Şimşek et al. (2021). Assume that

$$\theta = (w_1, a_1) \oplus \cdots \oplus (w_n, a_n)$$

is an irreducible critical point of the network with r neurons. The key observation is that splitting any one of the n neurons into two neurons by copying the incoming vector and splitting the outgoing vector with an arbitrary scalar μ as follows

$$(w_j, a_j) \to (w_j, \mu a_j) \oplus (w_j, (1 - \mu) a_j), \tag{10}$$

not only preserves the network function, but also preserves the gradients with respect to all neurons at zero; thus yielding a so-called *symmetry-induced* critical point of the network with $n+1$ neurons. We denote this critical point by $\oplus^{j,\mu}(\theta)$. Moreover, moving the free parameter μ, we get a line of critical points which are identical to the network function generated by the parameter θ. The neuron splitting principle reviewed here was discovered by Fukumizu & Amari (2000). Following the same procedure, splitting a neuron into h neurons, or equivalently, repeating Eq. 10 for $(h-1)$ times

$$(w_j, a_j) \to (w_j, \mu_1 a_j) \oplus \cdots \oplus (w_j, \mu_h a_j) \tag{11}$$

such that the scalar coefficients add up to one (i.e. $\mu_1 + ... + \mu_h = 1$), we obtain a critical point of the network with $n+h-1$ neurons (Fukumizu et al., 2019; Zhang et al., 2021). Şimşek et al. (2021) work out the combinatorics of this problem: a new critical point can be generated either by splitting one of the n neurons into multiple neurons as discussed above, or by splitting *several* neurons into multiple neurons, for example:

$$(w_i, a_i) \to (w_i, \mu_1^i a_i) \oplus \cdots \oplus (w_i, \mu_h^i a_i), \;\; (w_j, a_j) \to (w_j, \mu_1^j a_j) \oplus \cdots \oplus (w_j, \mu_k^j a_j) \tag{12}$$

such that the groups of free parameters add up to one, i.e. $\mu_1^i + ... + \mu_h^i = 1$. Although there is an enormous amount of manifolds of symmetry-induced critical points in neural networks due to various combinations of neuron groups and permutation-symmetry, we zoom into the scenario of adding a single neuron and study their second-order nature in this paper.

3 The Optimal Solution of the One-Neuron Network

As a first step towards characterizing the optimal solution in the networks with a few neurons, we focus on the case of a single neuron ($n=1$). In this case, the loss simplifies to

$$L(w, a) = a^2 g(r, r, 1) - 2a \sum_{i=1}^{k} g(r, 1, u_i) + \text{const}, \quad \text{subject to} \sum_{i=1}^{k} u_i^2 \leq 1, \tag{13}$$

where r is the incoming vector norm and a is the outgoing weight of the one-neuron network. First let us introduce the properties of the potential g

Assumption 3.1. *We assume that the potential satisfies some or all of the following properties*

i. $g(r_1, r_2, u)$ *is increasing*[1] *as a function of* u *for* $r_1, r_2 > 0$ *and* $u \in (-1, 1)$*,*

ii. a. $\partial_u g(r_1, r_2, u)$ *is increasing as a function of* u *for* $r_1, r_2 > 0$ *and* $u \in [-1, 1]$*,*
b. $\partial_u g(r_1, r_2, u)$ *is increasing as a function of* u *for* $r_1, r_2 > 0$ *and* $u > 0$*, decreasing for* $u < 0$*.*

Thanks to Lemma B.1, any differentiable and increasing activation function satisfies Assumption 3.1. More generally, we show in Lemma B.2 that the common activation functions such as softplus, sigmoid, tanh, and ReLU

$$\sigma(x) = \frac{1}{\beta}\log(e^{\beta x} + 1) \text{ with } \beta > 0, \;\; \sigma(x) = \frac{1}{1 + e^{-x}}, \;\; \sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \;\; \sigma(x) = \max\{0, x\},$$

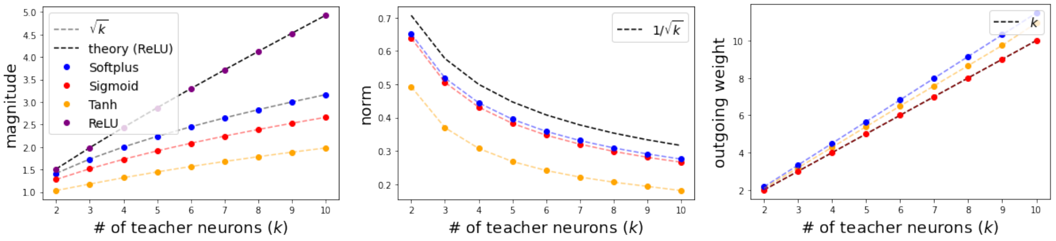
respectively, satisfy Assumptions 3.1. For any activation function satisfying Assumptions 3.1, we show that the incoming vector of any critical point of the one-neuron network must align *equally* with all of the incoming vectors of the teacher network

Theorem 3.2. *Under Assumptions 3.1 (i)-(ii) on the potential, the following loss on the similarities for some fixed* $a, r \neq 0$

$$\tilde{L}((u_i)_{i=1}^k) = -2a \sum_{i=1}^{k} g(r, 1, u_i), \quad \textit{subject to} \;\; \sum_{i=1}^{k} u_i^2 \leq 1 \tag{14}$$

has a unique critical point at $u_i = 1/\sqrt{k}$ *for* $i \in [k]$ *if* $a > 0$*, and* $u_i = -1/\sqrt{k}$ *for* $i \in [k]$ *if* $a < 0$*.*

[1] Increasing means strictly increasing everywhere in this paper.



(a) **(b)** **(c)**

Figure 2: *Structure of the optimal solution of the one-neuron network for various activation functions.* We train one-neuron networks learning from teacher networks with number of neurons from 2 to 10. For ReLU, we use the analytic formula of the potential giving an exact approach to study the loss averaged over Gaussian data, whereas for softplus, sigmoid, and tanh, we approximate the population loss with 10^5 input points. We run simulations for 10 different seeds, all of which converge to the same solution where the similarities are equal at $1/\sqrt{k}$ except for tanh for which there is also a sign-symmetric solution[3]. **(a)** For ReLU, the magnitude $\|w^*\|a^*$ matches exactly between simulations and Theorem 3.3. For softplus, the magnitude matches exactly with $\sqrt{k}$; for softplus and tanh, it is just below $\sqrt{k}$. **(b)** The norm of the incoming vector is smaller than $1/\sqrt{k}$ not only for softplus as shown in Theorem 3.4, but also for sigmoid and tanh. **(c)** The outgoing weight is larger than k for softplus and tanh; and it is exactly k for sigmoid.

Intuitively, the incoming vector of the one-neuron network is pulled towards the teacher incoming vectors for positive outgoing weights, hence any critical point has to be in the span of the teacher incoming vectors. Since the teacher incoming vectors are orthogonal to each other and they are equal in strength – with unit norm and unit outgoing weight – the student incoming vector should align with each of them equally to be stationary.

For ReLU activation function, since the potential has an analytical expression (Cho & Saul, 2009; Safran & Shamir, 2018), we are able to fully characterize the landscape of the one-neuron network.

Theorem 3.3. *Assume that the activation function is $\sigma(x) = \max\{0, x\}$. Any global minima (w^*, a^*) of the loss in Eq. 13 satisfies*

$$\|w^*\|a^* = k\frac{h(u)}{h(1)}, \tag{15}$$

forming an equal-loss hyperbola with the loss given by

$$L^* = k^2\left(h(0) - \frac{h(u)^2}{h(1)}\right) + k(h(1) - h(0)), \tag{16}$$

where $u = 1/\sqrt{k}$. For a teacher network with $k > 1$ neurons, the only other critical point of the loss satisfies $w^* = 0$ and $a^* = 0$. For the teacher network with one neuron, i.e. $k = 1$, there is an additional ray of critical points with $a^* = 0$, $w^* = \alpha v_1$ for $\alpha < 0$.

For softplus activation function, there is no known analytical expression for the potential. We develop a novel proof technique to describe the global landscape of the one-neuron network that is sketched in Appendix C.3. Although the proof relies on some specific properties of softplus (Lemma C.3), we numerically observe that the minimum reached by gradient flow has a similar characteristic for sigmoid and tanh (see Figure 2).

Theorem 3.4. *Assume that the activation function is $\sigma(x) = (1/\beta)\cdot\log(e^{\beta x} + 1)$ with $\beta > 0$. Any critical point (w^*, a^*) of the loss in Eq. 13 satisfies*

$$\|w^*\| \leq \frac{1}{\sqrt{k}}, \quad \textit{and} \quad k \leq a^*. \tag{17}$$

For a teacher network with one neuron, there is a unique critical point with $\|w^*\| = 1$ and $\|a^*\| = 1$.

[3] For tanh, we find two solutions that are sign-symmetric: the usual one where the similarities are all $1/\sqrt{k}$, with a norm and outgoing weight denoted by (r, a) with $a > 0$, and its symmetric solution where the similarities are all $-1/\sqrt{k}$, with a norm and outgoing weight $(r, -a)$. Only the first solution is plotted in the figure for finer comparison in the positive scale.

For softplus activation function, combining Theorem 3.2 and Theorem 3.4, the incoming vector of the optimal solution of the one-neuron network can be expressed as

$$w^* = \frac{r}{\sqrt{k}} \sum v_i,$$

with $r \leq 1/\sqrt{k}$. Hence, the incoming vector implements a damped average of the incoming vectors of the teacher with a damping factor of $r/\sqrt{k} \leq 1$. The outgoing weight of the minimum is at least k since the one neuron should compensate for approximating k teacher neurons. Although the proof does not directly apply to other activation functions such as sigmoid, [4] we numerically observe a very similar structure for the solution of the one-neuron network (see Fig. 2). Generalizing the proof to non-orthogonal incoming teacher vectors should be possible since the objective in Eq. 13 does not change, however, the constraint on the similarities needs to be adapted.

What about the global landscape of the two-neurons network? We discuss in the next section 4 that there is a line of symmetry-induced critical points representing the same network function as the solution of the one-neuron network. In addition, there is an optimal solution exploiting the full network capacity where the incoming vectors are not identical. Whether there are other non-trivial critical points remains an open question that we believe can be answered within our framework.

4 Neuron Splitting Creates Lines of Critical Points

We reviewed the atypical structure of critical points of the neural networks loss landscapes in Section 2: symmetry-induced critical points form manifolds thus they are not isolated. For example in the landscape of a neural network with at least two neurons, there exists a line of symmetry-induced critical points induced by a critical point of the one-neuron network. Hence, the celebrated Kac-Rice formula describing the scaling of the number of critical points of complex landscapes (Auffinger et al., 2013; Ros et al., 2019; Maillard et al., 2020) does not apply to neural network landscapes with more than one neuron.

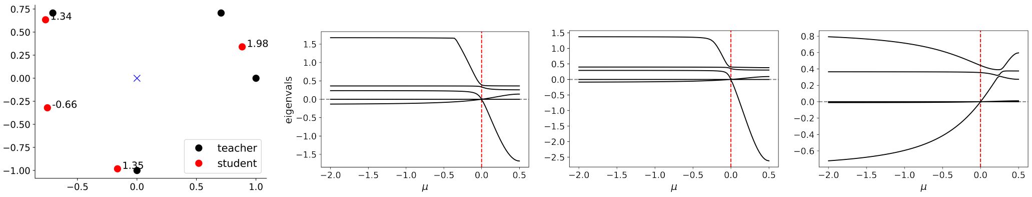
In this section, we assume that the activation function σ is twice-differentiable. We denote the Hessian of the loss in Eq. 3 at a parameter θ by $HL(\theta)$ which is matrix of size $P \times P$. Then we zoom into one such line of symmetry-induced critical points and work out their second-order nature that enables us to identify whether these are local minima, strict saddles, or non-strict saddles. Below, we recall the definitions:

- *Strict saddle:* A critical point with at least one negative eigenvalue in the Hessian, i.e. there is an escape direction towards decreasing loss in its neighborhood.
- *Non-strict saddle:* A critical point where the Hessian has no negative eigenvalues, yet there is an escape direction towards decreasing loss in its neighborhood.
- *Local minimum:* A critical point where the Hessian has no negative eigenvalues and the loss increases in all directions of perturbation.

If the Hessian does not have a negative eigenvalue, the critical point can be either a non-strict saddle or a local minimum and a higher-order analysis is needed in this case (Anandkumar & Ge, 2016). For one output neuron (i.e. $d_{\text{out}} = 1$), Fukumizu & Amari (2000) proved that the second-order nature of the symmetry-induced critical points is determined by the mixing ratio μ and a second-order derivative matrix Y of size $d \times d$ which we explicitly write in Theorem 4.1. More precisely, using a linear transformation of the weight vectors of the splitted neuron, Fukumizu & Amari (2000) show that depending on the eigenvalue signs of the matrix $\mu(1-\mu)Y$, a symmetry-induced critical point $\oplus^{j,\mu}(\theta)$ is either a local minimum or a strict saddle. However, the linear transformation of Fukumizu & Amari (2000) at $\mu \in \{0, 1\}$ is not invertible, thus the corresponding symmetry-induced critical points are unclassified. Using a novel decomposition of the Hessian, we generalize their result in two ways (Theorem 4.1):

i. For arbitrary d_{out}, we explicitly give the number of positive, negative, and zero eigenvalues of the Hessian at a symmetry-induced critical point; which depends on two matrices of second-

[4] The sigmoid activation function does not satisfy the property that $\sigma'(x)x/\sigma(x)$ is increasing – which is needed in Lemma C.1.



(a) local min. θ **(b)** $a_j = 1.35$ **(c)** $a_j = 1.98$ **(d)** $a_j = -0.66$

Figure 3: *The Hessian spectrum of the symmetry-induced critical points as a function of the mixing ratio μ.* We plot the 5 smallest eigenvalues in (b-c-d). **(a)** A local minimum θ with the largest empirical basin of attraction (see the simulation details in Appendix A.1). The dataset is composed of 2-dim. input points sampled from standard gaussian ($N = 10^5$ data points) and targets that are generated by a teacher network (its incoming vectors are shown as black dots, and the outgoing weights are all one). The incoming vectors of the networks at convergence are shown in red and the outgoing weights are written on the plot. We split the neurons of the local minimum θ. **(b-c)** *Neuron splitting at $a_j = 1.35$ and $a_j = 1.98$:* Y has one positive and one negative eigenvalue, hence for any $\mu \neq 0$, the Hessian has a negative eigenvalue. **(d)** *Neuron splitting at $a_j = -0.66$:* Y is positive definite, hence we find symmetry-induced local minima on $(0, 0.5]$ and strict saddles of index-2 on $(-\infty, 0)$. In all cases, the zero eigenvalue is degenerate at $\mu = 0$, and it appears exactly thrice in the spectrum since we have two input neurons and one output neuron.

order derivatives, namely Y and V (see Eq. 19). This is the relevant case for deep networks which was not treated in (Fukumizu & Amari, 2000).

ii. In particular, for $d_{\text{out}} = 1$ and $\mu \in \{0, 1\}$, we show that the symmetry-induced critical points are non-strict saddles for any distribution of the eigenvalue signs of the matrix Y; which may cause the optimization algorithm to get stuck.

Theorem 4.1. *Let $\theta \in \mathbb{R}^P$ be an irreducible critical point of a neural network with n neurons. Let $\oplus^{j,\mu}(\theta)$ be a symmetry-induced critical point. The spectrum of $HL(\oplus^{j,\mu}(\theta))$ is composed of two parts. The bulk of $(d+d_{out})n$ eigenvalues have the same signs as the eigenvalue signs of $HL(\theta)$. The remaining $(d+d_{out})$ eigenvalues have the same signs as the eigenvalue signs of the following matrix*

$$\begin{bmatrix} \mu(1-\mu)Y(w_j, a_j) & -V(w_j, a_j) \\ -V(w_j, a_j)^T & 0 \end{bmatrix} \tag{18}$$

where

$$\begin{aligned} Y(w_j, a_j) &= \left\langle \sigma''(w_j \cdot x) x x^T a_j \cdot c'(f(x), f^*(x)) \right\rangle_{\mathbb{P}} \in \mathbb{R}^{d \times d}, \\ V(w_j, a_j)_{k\ell} &= \langle \sigma'(w_j \cdot x) x_k c'(f(x), f^*(x))_\ell \rangle_{\mathbb{P}} \text{ with } k \in [d], \ell \in [d_{out}]. \end{aligned} \tag{19}$$

This theorem can be generalized to multi-layers networks by considering any of its hidden layers. Only the expressions of Y and V need to be updated by (i) substituting x with the post-activation vector coming before the hidden layer and (ii) the derivative of the cost needs to be calculated w.r.t. the pre-activation vector coming after the hidden layer, before applying the activation on the next layer.

We need to unpack the submatrix in Eq. 18 to understand the second-order nature of the symmetry-induced critical points. Note that

$$V(w_j, a_j) a_j = \frac{\partial L}{\partial w_j}(\theta) = 0. \tag{20}$$

Since the original critical point θ is irreducible, the outgoing vector a_j has to be non-zero, which implies that dim(Null(V)) ≥ 1. In the case of one output neuron, V vanishes thus the submatrix in Eq. 18 reduces to

$$\begin{bmatrix} \mu(1-\mu)Y(w_j, a_j) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{21}$$

Therefore, the eigenvalue signs of the Hessian in the new directions are determined by the spectrum of Y and the sign of $\mu(1-\mu)$ up to the trivial zero-eigenvalue coming from the vanishing V. Note that the two symmetry-induced critical points $\oplus^{j,\mu}(\theta)$ and $\oplus^{j,1-\mu}(\theta)$ have identical Hessian spectra since they have the same set of neurons. Hence, it is enough to discuss the ray $(-\infty, 0.5]$.

An interesting case is when θ is a local minimum, with a Hessian $HL(\theta)$ without any negative eigenvalues. What is the effect of neuron splitting in this case? Does a minimum turn into a strict saddle? Depending on the spectrum of Y, there are three main scenarios (see Figure 3)

i. *positive definite Y:* strict saddles on $\mu < 0$ transit into local minima on $\mu \in (0, 0.5]$ via a non-strict saddle at $\mu = 0$ with a one-sided escape route,

ii. *negative definite Y:* local minima on $\mu < 0$ transit into strict saddles on $\mu \in (0, 0.5]$ via a non-strict saddle at $\mu = 0$ with a one-sided escape route,

iii. *Y has at least one positive and one negative eigenvalue:* strict saddles on $\mu < 0$ transit into other strict saddles on $\mu \in (0, 0.5]$ via a non-strict saddle at $\mu = 0$ with a two-sided escape route.

In particular, in any case, we have a non-strict saddle at $\mu = 0$ since the submatrix in Eq. 18 completely vanishes. Although the Hessian spectrum $HL(\oplus^{j,0}(\theta))$ does not suffice to classify this critical point, since there are neighboring strict saddles on one side or on both sides, we conclude that it is a non-strict saddle with a one or two-sided escape route towards lower loss [5]. The example of Fig. 3 with the small network of 5 neurons (where one neuron is duplicated) shows that we can always find a mixing ratio μ that turns the local minimum of the 4-neuron network into a saddle for the 5-neuron network. Indeed, we have an 'exclusive-or' situation: either for all $\mu \in (0, 0.5]$ or for all $\mu > 0$, the Hessian of the critical points has a negative eigenvalue.

We discuss the case of multiple output neurons in detail in Appendix Section D.1.

4.1 The Minimal Hessian Eigenvalue as a Function of the Mixing Ratio

The escape speed from a strict saddle point depends on the magnitude of the minimum eigenvalue of the Hessian (Jin et al., 2017). For the local minima, the smallest eigenvalue of the Hessian gives a measure of flatness. In this section, we dig deeper into the second-order nature of the symmetry-induced critical points by studying the smallest non-trivial eigenvalue of the Hessian, focusing on the case of one-output neuron. If the Hessian has a negative eigenvalue, $\lambda^\dagger$ denotes the smallest eigenvalue; if not, it denotes the second smallest eigenvalue excluding the trivial zero.

Lemma 4.2. *The smallest non-trivial eigenvalue of the Hessian of a symmetry-induced critical point can be bounded as follows*

$$\lambda^\dagger(HL(\oplus^{j,\mu}(\theta))) \leq \begin{cases} u(\mu)\lambda_{\min}(Y) & \text{for } \mu \in (0,1), \\ u(\mu)\lambda_{\max}(Y) & \text{for } \mu \in \mathbb{R}/[0,1], \end{cases}$$

where

$$u(\mu) = \frac{\mu(1-\mu)}{(1-\mu)^2 + \mu^2}.$$

As a sanity check: if Y is positive definite, $\lambda^\dagger$ is non-negative for $\mu \in (0, 1)$ and so is the upper bound (similarly for Y negative definite); if Y has at least one negative and one positive eigenvalue, $\lambda^\dagger$ is negative and so is the upper bound. We also provide a lower bound in the Appendix D.2.

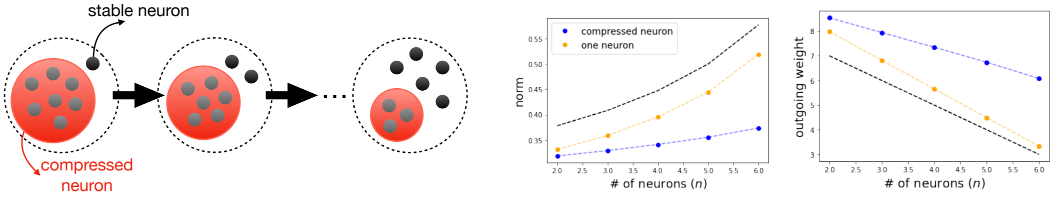
4.2 A Curious Case: The Gradient Flow Diverges Towards Infinity Following the Ray of Symmetry-Induced Local Minima

If Y is negative definite, the symmetry-induced critical points on $\mu \in \mathbb{R}/[0, 1]$ are local minima (Theorem 4.1). In the limit $\mu \to \infty$ (equivalently, $\mu \to -\infty$ due to symmetry), we get

$$-\frac{1}{2}\lambda_{\max}(Y) \geq \lim_{\mu \to \infty} \lambda^\dagger(HL(\oplus^{j,\mu}(\theta))).$$

The upper bound on the smallest non-trivial eigenvalue grows as μ extends towards ∞ and eventually saturates. This suggests that the symmetry-induced local minimum gets steeper as $|\mu|$ grows

[5] Except for the case of vanishing Y with no positive and no negative eigenvalues, for ex. this is the case for the linear activation function.



(a) (compressed) neuron splitting

(b) minima structure in few-neuron networks

Figure 4: *From one-neuron to six-neurons network by splitting one neuron at a time. The targets are generated by a teacher network with eight-neurons as in Eq. 4, the input data is* 9*-dimensional standard Gaussian, and the activation function is softplus.* To get a tight approximation of the gradient flow in high-dimensions, we use numerical integration to approximate the potentials. **(a)** A schematic for the neuron splitting process. In two-neurons network, the gradient flow converges the following solution: one *compressed* neuron averaging seven teacher neurons, and one *stable* neuron matching one teacher neuron. We find a tiny but non-zero correlation between the two neurons. Splitting the compressed neuron with a mixing ratio of 0.5, we get a three-neuron solution with one compressed neuron averaging six teacher neurons, and two stable neurons. The same trend continues until we reach six-neurons network. The norm and the outgoing weight of the stable neuron is approximately 1.70 and 0.42 for all few-neurons networks. **(b)** We plot the norm and the outgoing weight of the compressed neuron as a function of the number of neurons. The compressed neuron attains a smaller norm compare to the one-neuron solution averaging the same number of teacher neurons, hence attains a bigger outgoing weight to balance. We also find that the correlation between the compressed and a stable neuron decreases as the number of neurons increase.

larger, although they represent the same network function (in particular, same generalization behavior). In general, we observed that the gradient flow trajectories initialized close to the origin (standard training) does not get stuck at symmetry-induced local minima, except for a tiny fraction of initializations (see Appendix A.1).

We wondered where gradient flow trajectories end up when they are initialized close to the symmetry-induced saddles, in particular whether they converge to other symmetry-induced local minima at a lower loss. We found that the destination is determined by the sign of the dot product between the smallest eigenvector of the Hessian and the perturbation vector (see Fig. 7 of Appendix). The destination also depends on the splitted neuron creating the symmetry-induced saddles. We also found a curious case for some perturbation scenarios: the gradient flow travels near a line of symmetry-induced local minima at a lower loss towards infinity, without converging to any local minima on the line (see Fig. 9 in Appendix). We propose a low dimensional toy model to explain the apperance of a local minimum at infinity in Appendix D.2.3 near a line of critical points, inspired by the upper bound on the minimal eigenvalue of the Hessian.

5 Conclusion & Future Directions

In this paper, we proposed a novel proof to study the global landscape of the one-neuron network when it learns from a teacher network with arbitrary number of neurons. Going beyond the one-neuron case, we analyzed the second-order nature of the symmetry-induced critical points ubiquitous in the loss landscapes of neural networks Şimşek et al. (2021). Numerically analyzing the neural networks with few neurons trained through a neuron splitting procedure, we identified that the compression strategy of averaging teacher neurons is prevalent in underparameterized networks. Generalizing the proofs for the one-neuron network to more general activation functions is an intriguing direction to pursue. An another interesting direction is generalizing the proofs for the case of non-orthogonal incoming vectors, and also for the case of arbitrary outgoing weights for the teacher, which requires a finer analysis. A more challenging case is the rigorous analysis of the global landscape of the two-neurons networks, which is an open question. We believe our paper establishes the first steps towards understanding the compression strategy of underparameterized networks, hence opening a way to understand and rigorously study the solutions found by neural networks with finite width.

REFERENCES

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on learning theory*, pp. 81–102. PMLR, 2016.

Yossi Arjevani and Michael Field. Analytic study of families of spurious minima in two-layer relu neural networks: a tale of symmetry ii. *Advances in Neural Information Processing Systems*, 34: 15162–15174, 2021.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *arXiv preprint arXiv:2206.04030*, 2022.

Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:2206.00939*, 2022.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

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.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-ichi Mototake, and Mirai Tanaka. Semi-flat minima and saddle points by embedding neural networks to overparameterization. *Advances in Neural Information Processing Systems*, 32:13868–13876, 2019.

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.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, 2020.

Reiner Horst and Panos M Pardalos. *Handbook of global optimization*, volume 2. Springer Science & Business Media, 2013.

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

Arthur Jacot, François Ged, Franck Gabriel, Berfin Şimşek, and Clément Hongler. Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization. *arXiv preprint arXiv:2106.15933*, 2021.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732. PMLR, 2017.

Christopher A. Kennedy and Mark H. Carpenter. Additive runge–kutta schemes for convection–diffusion–reaction equations. *Applied Numerical Mathematics*, 44(1–2):139–181, Jan 2003. ISSN 0168-9274. doi: 10.1016/s0168-9274(02)00138-1. URL `http://dx.doi.org/10.1016/S0168-9274(02)00138-1`.

Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning*, pp. 287–327. PMLR, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018. doi: 10.21105/joss.00615.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Christopher Rackauckas and Qing Nie. Differentialequations.jl – a performant and feature-rich ecosystem for solving differential equations in julia. *The Journal of Open Research Software*, 5 (1), 2017. doi: 10.5334/jors.151. URL `https://app.dimensions.ai/details/publication/pub.1085583166andhttp://openresearchsoftware.metajnl.com/articles/10.5334/jors.151/galley/245/download/`. Exported from https://app.dimensions.ai on 2019/05/05.

Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52 (4):4225, 1995.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International conference on machine learning*, pp. 3404–3413. PMLR, 2017.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, 34:28690–28700, 2021.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *arXiv preprint arXiv:2202.00293*, 2022.

Lei Wu. Learning a single neuron for non-monotonic activation functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 4178–4197. PMLR, 2022.

Lemeng Wu, Dilin Wang, and Qiang Liu. Splitting steepest descent for growing neural architectures. *Advances in neural information processing systems*, 32, 2019.

Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pp. 3756–3786. PMLR, 2020.

Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhiqin J Xu. Embedding principle of loss landscape of deep neural networks. *Advances in Neural Information Processing Systems*, 34:14848–14859, 2021.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pp. 4140–4149. PMLR, 2017.

A Simulations 13

A.1 Minima Found by a Four-Neuron Network (Fig 3) 13

B General Properties of the Potentials 14

C Proofs for the One-Neuron Network 16

C.1 Proof of Theorem 3.2 for General Activation Functions 16

C.2 Proof of Theorem 3.3 for ReLU 17

C.3 Proof of Theorem 3.4 for Softplus 18

D Second-Order Nature of Symmetry-Induced Critical Points 24

D.1 Multiple Output Neurons 25

D.2 Bounding the Minimal Hessian Eigenvalue 27

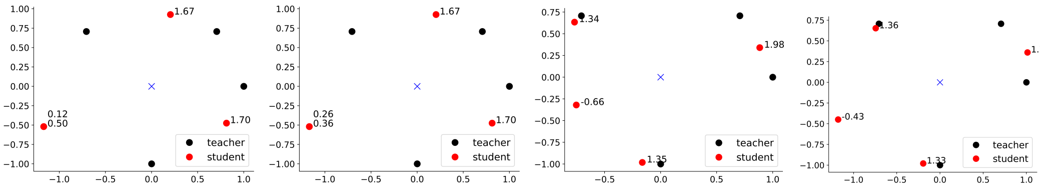
A Simulations

We integrate the gradient flow with an ESDIRK integration method (KenCarp58) Kennedy & Carpenter (2003); Rackauckas & Nie (2017) and chain the result into a trust-region Newton optimizer Mogensen & Riseth (2018) to obtain highly accurate estimates of fixed points of the gradient flow, i.e. critical points of the loss function.

A.1 Minima Found by a Four-Neuron Network (Fig 3)

We create a two-layer teacher network with $r = 4$ neurons: unit 2-dimensional incoming vectors $w_j = [\cos(\beta_j), \sin(\beta_j)]$ with $\beta_j \in \{0, \pi/4, 3\pi/4, 3\pi/2\}$ and with outgoing weights $a_j = 1$. We sample $N = 10^5$ input data points x_i from a standard Gaussian. The teacher network then creates the targets $y_i = \sum_{j=1}^{4} \sigma(\cos(\beta_j)x_i^1 + \sin(\beta_j)x_i^2)$. We trained 1000 seeds of students with the same number of neurons using the normal Glorot initialization Glorot & Bengio (2010).

In this subsection, we present the empirical local minimum that we found in 1000 simulations above in terms of the distribution of its neurons (Fig. 5) and some statistics such as the size of the empirical basin of attraction and the Hessian spectrum.



(a) sym. ind. local min.-I **(b)** sym. ind. local min.-II **(c)** local min.-I **(d)** local min.-II

Figure 5: *Three different local minima.* Out of 1000 random initializations, gradient flow converged to a either one of the three local minima up to permutation shown in the figure or to a global minimum. For the symmetry-induced local minimum (a-b), the mixing ratio of the outgoing weights differs depending on the initialization, however it is always on the line segment $\mu \in (0, 1)$ corresponding to the splitting of the neuron in the lower left corner. Although the other two local minima are very similar to each other (c-d), we believe they are different, i.e. the separation is not an artifact due to numerical issues. In (c-d), we note the extra symmetry between the two neurons with very similar outgoing weights (i.e. $a_j \sim 1.35$). This extra symmetry is due to the angular internal symmetry between the teacher neurons. Further statistics on these three local minima are given in the Table 1.

We observe that the global minimum is the flattest in terms of the minimum non-trivial eigenvalue of the Hessian (i.e. excluding the trivial zero of the symmetry-induced local minimum). It also has the largest empirical basin of attraction with $696/1000$ initializations converging to it. However, although the local minimum-II is flatter than local minimum-I, its empirical basin of attraction is much smaller – which contradicts with the hypothesis that the gradient flow converges to a flatter minimum more often.

—	sym. ind. local min	local min.-I	local min.-II	global min.
fractions	13/1000	227/1000	64/1000	696/1000
loss	2.02e-5	1.63e-6	1.61e-6	1.17e-29
grad	3.97e-11	1.41e-11	1.57e-11	1.90e-11
min. eig.	-2.3e-12	0.335	0.190	0.085
2-nd min. eig.	0.815	0.387	0.260	0.270
max. eig.	3.3e6	3.0e6	3.3e6	3.1e6

Table 1: Statistics of all three minima of the student network. The average loss, gradient norm; the minimum, the second minimum, and the maximum eigenvalue of the Hessian are reported for each one of the minima.

B General Properties of the Potentials

In this section, we introduce some general properties of the potentials. At the moment, we use these only for the one-neuron networks (see Section C). We expect these properties to play a crucial role in studying the networks with two or more neurons.

We first prove a simple rule for the partial derivative of the potential with respect to the similarity.

Lemma B.1. *If the activation function σ is differentiable, then the partial derivative of the potential $g(r_1, r_2, u) = \langle\sigma(r_1x)\sigma(r_2y)\rangle$ with respect to the correlation $\langle xy\rangle = u$ can be expressed as*

$$\partial_u g(r_1, r_2, u) = r_1 r_2 \langle\sigma'(r_1x)\sigma'(r_2y)\rangle. \tag{22}$$

Applying the partial derivative rule once more, we get

$$\partial^2_{u^2} g(r_1, r_2, u) = r_1^2 r_2^2 \langle\sigma''(r_1x)\sigma''(r_2y)\rangle. \tag{23}$$

Proof. We compute the derivative of $g(r_1, r_2, u)$ by making the correlation u explicit. We denote $u' = \sqrt{1-u^2}$. After the computation, we use the Stein's lemma to reach the desired formula.

$$\partial_u g(r_1, r_2, u) = r_2\langle\sigma(r_1x)\sigma'(r_2(ux+u'z))x\rangle - \frac{r_2 u}{u'}\langle\sigma(r_1x)\sigma'(r_2(ux+u'z))z\rangle \tag{24}$$

where x and z are independent standard Gaussians. Here is a reminder for the Stein's Lemma for a standard Gaussian z

$$\langle v(z)z\rangle = \langle v'(z)\rangle. \tag{25}$$

To remove x in the integrand of the first term, we apply the Stein's formula for $v(x) = \sigma(r_1x)\sigma'(r_2(ux+u'z))$ and get

$$r_1r_2\langle\sigma'(r_1x)\sigma'(r_2(ux+u'z))\rangle + r_2^2u\langle\sigma(r_1x)\sigma''(r_2(ux+u'z))\rangle. \tag{26}$$

To remove z in the integrand of the second term, we apply the Stein's formula for $v(z) = \sigma'(r_2(ux+u'z))$ and we get

$$-r_2^2u\langle\sigma(r_1x)\sigma''(ux+u'z)\rangle. \tag{27}$$

Summing up the two terms, we complete the proof. □

We next prove that the potentials induced by softplus, ReLU, sigmoid, and tanh satisfy Assumptions 3.1 (i) and (ii) (a. or b.). The core of the proof comes from Lemma B.1.

Lemma B.2. *The potentials of increasing and differentiable activation functions such as softplus, sigmoid, and tanh*

$$\sigma(x)=\frac{1}{\beta}\log(e^{\beta x}+1) \textit{ with } \beta>0,\ \ \sigma(x)=\frac{1}{1+e^{-x}},\ \ \sigma(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}},$$

respectively satisfy Assumption 3.1 (i). The potential of the ReLU activation function, i.e. $\sigma(x)=\max\{0,x\}$*, also satisfies it. The potentials of convex and twice differentiable activation functions such as softplus, and ReLU satisfy Assumption 3.1 (ii)-a. Moreover, the potentials of sigmoid and tanh activation functions satisfy Assumption 3.1 (ii)-b.*

Proof. The potentials of increasing and diffentiable activation functions satisfy Assumption (i) since $\sigma'(\cdot)>0$ due to Lemma B.1. Moreover, if the activation function is convex and twice differentiable, its potential also satisfies Assumption 3.1 (ii) since $\sigma''(\cdot)>0$ and from applying Lemma B.1 twice.

Note that the softplus is smooth, increasing and convex since the first two derivatives are positive

$$\sigma'(x)=\frac{e^{\beta x}}{e^{\beta x}+1},\ \ \sigma''(x)=\frac{\beta e^{\beta x}}{(e^{\beta x}+1)^2}. \tag{28}$$

Therefore the softplus and any other twice differentiable, increasing, and convex activation function satisfies Assumptions 3.1 (i) and (ii)-a.

For the ReLU activation function, the following analytic expressions are derived for the potential and its partial derivative with respect to the similarity (Cho & Saul, 2009; Safran & Shamir, 2018)

$$g(r_1,r_2,u)=\frac{r_1r_2}{2\pi}\left(\sqrt{1-u^2}+(\pi-\arccos(u))u\right), \tag{29}$$

$$\partial_u g(r_1,r_2,u)=\frac{r_1r_2}{2\pi}(\pi-\arccos(u)). \tag{30}$$

Since $\partial_u g$ is positive in $(-1,1)$ Assumption 3.1 (i) is satisfied. Since $\arccos(u)$ is decreasing in $[-1,1]$, $\partial_u g$ is increasing, hence Assumption 3.1 (ii)-a is satisfied too.

Finally, we show that potentials of sigmoid and tanh satisfy Assumption 3.1 (ii)-b. The second-order derivatives of these two activation functions are the same since they differ only by a constant

$$\sigma''(x)=\sigma(x)(1-\sigma(x))(1-2\sigma(x))=\sigma'(x)(1-2\sigma(x)) \tag{31}$$

where $\sigma(x)=1/(1+e^{-x})$. Since $\sigma'(\cdot)$ is an even and $(1-2\sigma(\cdot))$ is an odd function, their multiplication $\sigma''(\cdot)$ is an odd function. Hence, it respects the following identity

$$\langle\sigma''(x)\sigma''(y)\rangle=-\langle\sigma''(x)\sigma''(-y)\rangle. \tag{32}$$

Moreover, we have that $\sigma''(x)<0$ for $x>0$ and $\sigma''(0)=0$. At $u=0$, the above expectation factors out and we get thanks to the sign symmetry

$$\langle\sigma''(x)\sigma''(y)\rangle=\langle\sigma''(x)\rangle\langle\sigma''(y)\rangle=0. \tag{33}$$

We next use conditional expectation and the oddness property to rewrite the expression of the potential

$$\begin{aligned}\langle\sigma''(x)\sigma''(y)\rangle&=\langle\sigma''(x)\sigma''(y)\mid xy>0\rangle\mathbb{P}\{xy>0\}+\langle\sigma''(x)\sigma''(y)\mid xy\le 0\rangle\mathbb{P}\{xy\le 0\},\\&=\langle|\sigma''(x)||\sigma''(y)|\mid xy>0\rangle\left(\mathbb{P}\{xy>0\}-\mathbb{P}\{xy\le 0\}\right).\end{aligned} \tag{34}$$

Note that the first term is positive since the integrand is positive. As for the second term, we get

$$\mathbb{P}\{xy>0\}>1/2 \tag{35}$$

since x and y are positively correlated. Hence, the probability of $xy>0$ is bigger than the probability of the complement. Therefore from Lemma B.1 and Eq. 34, we get that $\partial^2_{u^2}g$ is positive for $u>0$ and negative for $u<0$ which suffices to satisfy Assumption 3.1 (ii)-b. □

For the linear activation function, we have an explicit formula for the potential

$$g(r_1,r_2,u)=\langle r_1x\cdot r_2y\rangle=r_1r_2u \tag{36}$$

which satisfies Assumption 3.1 (i) but not (ii) as $\partial_u g$ is constant as a function of u.

C Proofs for the One-Neuron Network

For the one-neuron network, we can expand the loss as follows

$$L = a^2 g(r,r,1) - 2a \sum_{i=1}^{k} g(r,1,u_i) + \text{const}, \tag{37}$$

where the constant represents the teacher-teacher interactions. In this section, we first prove the uniqueness of similarities in Subsection C.1 for a general family of activation functions using symmetry arguments and properties that we developed in Section B. Next, applying the result of Theorem 3.2, the above loss in Eq. 37 at a critical point reduces to

$$L = a^2 g(r,r,1) - 2kag(r,1,u) + \text{const}. \tag{38}$$

where $u = 1/\sqrt{k}$. Moreover, the partial derivatives with respect to the norm and the outgoing weight should also be zero which gives the following two constraints

$$\partial_a L = 2ag(r,r,1) - 2kg(r,1,u) = 0, \tag{39}$$

$$\partial_r L = a^2 \partial_r g(r,r,1) - 2ka\partial_r g(r,1,u) = 0. \tag{40}$$

These constraints are equivalent to the following

$$\frac{a}{k} = \frac{g(r,1,u)}{g(r,r,1)} = \frac{2\partial_r g(r,1,u)}{\partial_r g(r,r,1)}. \tag{41}$$

Note that the second equality between the two ratios of Gaussian averages gives a fixed point equation on the norm. To find the solutions of the fixed point equation, we control the Gaussian averages with the help of the FKG inequality and Stein's lemma in Subsection C.3.

C.1 Proof of Theorem 3.2 for General Activation Functions

In this subsection, we prove that for the general activation functions, any critical point of the one-neuron network should have equal similarities with the teacher incoming vectors, excluding the origin. First, we take a detour to check the applicability of the convex optimization framework.

If Assumption 3.1 (ii)-a is strengthened to the convexity of $g(r_1, r_2, u)$ in the similarity for $r_1, r_2 > 0$ and $u \in [-1, 1]$, we get the uniqueness of the critical point for the case $a < 0$ using the convex optimization framework as follows (see Boyd et al. (2004) Section 4.2). In this case, the loss $\tilde{L}$ is convex since its Hessian is diagonal with entries

$$\partial^2_{u_i^2} \tilde{L} = -2a\partial^2_{u_i^2} g(r,1,u_i) > 0. \tag{42}$$

Note that the constraint on the similarities (Eq. 14) is also convex, thus we get a convex optimization problem which has a unique minimizer. We can argue for the non-existence of critical points by contradiction. Assume that there exists a critical point other than the minimizer. Then the loss on the line segment between the minimizer and the critical point crosses the loss corresponding to the linear interpolation between these two points. This breaks the convexity.

Swapping any pair of u_i does not change the loss since all teacher neurons have equal strength (unit norm incoming vector and unit outgoing weight), therefore the loss is permutation symmetric in $(u_i)_{i=1}^k$. If any two u_i were distinct from each other at the minimizer, then its permutation would also be a minimizer which would violate the uniticity. In this case, we conclude that there is a unique critical point where the similarities are equal to each other.

However for the case $a > 0$, $\tilde{L}$ is concave and the constraint is convex. This falls in the realm of concave minimization Horst & Pardalos (2013). In this case, in general, there can be arbitrarily many local minimizers. Hence, for the proof of the Theorem 3.2, we use a different strategy

Proof. We first show that the similarities should satisfy the constraint in Eq. 14 at a critical point. If the norm of similarities do not achieve the maximum which is unity, we pick a similarity, say u_k. The potential $g(r,1,\cdot)$ is increasing in the similarity due to Assumption 3.1 (i). If the outgoing weight is positive, increasing any of u_k's decreases $\tilde{L}$, resulting in a non-zero derivative. The only exception is the case when all u_k's are positive and the constraint in Eq. 14 is satisfied since then it is not possible to increase any of the similarities. If the outgoing weight is negative, decreasing any of u_k's decreases $\tilde{L}$, resulting in a non-zero derivative. The only exception is the case when all u_k's are negative and the constraint in Eq. 14 is satisfied since then it is not possible to decrease any of the similarities. Hence, there is no critical point in the loss landscape except in the two cases mentioned above, i.e. $\sum_{i=1}^{k} u_i^2 = 1$ with either all $u_i > 0$ or all $u_i < 0$.

Next, we prove that all similarities at a critical point must be equal. Let's assume the contrary, wlog $u_i > u_j$. Since the disk constraint is convex, the linear interpolation between (u_j, u_i) and (u_i, u_j)

$$(u_i(t), u_j(t)) = (u_j + t(u_i - u_j), u_i + t(u_j - u_i)) \tag{43}$$

for $t \in [0, 1]$ remains in the domain. The derivative of the loss along this path is given by

$$\partial_t \tilde{L} = -2a \left(\partial_u g(r, 1, u_i(t))(u_i - u_j) - \partial_u g(r, 1, u_j(t))(u_i - u_j)\right). \tag{44}$$

Evaluated at $(u_i(t), u_j(t))\big|_{t=1} = (u_i, u_j)$, we get

$$\partial_t \tilde{L}\big|_{t=1} = -2a(u_i - u_j) \left(\partial_u g(r, 1, u_i) - \partial_u g(r, 1, u_j)\right). \tag{45}$$

If all $u_k > 0$ and either Assumption 3.1 (ii)-a or Assumption 3.1 (ii)-b is satisfied, we get that $\partial_u g$ is increasing, hence the above derivative cannot be zero. If all $u_k < 0$ and if Assumption 3.1 (ii)-a is satisfied, we get that $\partial_u g$ is increasing; if Assumption 3.1 (ii)-b is satisfied, we get that $\partial_u g$ is decreasing. In any case, the above derivative cannot be zero. Therefore, at a critical point, we have either $u_i = 1/\sqrt{k}$ for $a > 0$, or $u_i = -1/\sqrt{k}$ for $a < 0$ for all $i \in [k]$. □

C.2 Proof of Theorem 3.3 for ReLU

In the case of ReLU, potentials have an analytic expression Cho & Saul (2009); Safran & Shamir (2018)

$$g(r_1, r_2, u) = \frac{r_1 r_2}{2\pi} h(u) \;\text{ with }\; h(u) = \sqrt{1 - u^2} + (\pi - \arccos(u))u. \tag{46}$$

We showed in Theorem 3.2 that at a critical point with $a \neq 0$ and $r \neq 0$, all similarities to the teacher incoming vectors must be equal, and we denote them by u. Plugging in the similarities in the loss and using the factorization of the potential in Eq. 46, we get

$$L = a^2 r^2 \cdot h(1) - 2kar \cdot h(u) + \text{const}. \tag{47}$$

Reparameterizing with $\tilde{a} = ar$, let us observe that the loss is a second-order polynomial

$$L = h(1) \left(\tilde{a}^2 - \tilde{a} 2k \frac{h(u)}{h(1)} + k + k(k-1)\frac{h(0)}{h(1)}\right). \tag{48}$$

where we also made the constant explicit. Since the coefficient of the square term is positive, there is a minimizer (which is the only critical point). Taking the derivative, the minimum is attained at

$$\tilde{a} = k \frac{h(u)}{h(1)}. \tag{49}$$

Note that $h(u) > 0$ for $k \geq 2$, whether we have $u = -1/\sqrt{k}$ or $u = 1/\sqrt{k}$. This implies that the outgoing weight is positive, therefore the only option for $u = 1/\sqrt{k}$ according to Theorem 3.2. If $k = 1$, then we may have $h(-1) = 0$ implying $a = 0$, which is a case we analyze separately in the second part of the proof.

Plugging $u = 1/\sqrt{k}$ in, we get the loss

$$L = h(1) \left(-(k\frac{h(u)}{h(1)})^2 + k + k(k-1)\frac{h(0)}{h(1)}\right), \tag{50}$$

$$= k^2 \left(h(0) - \frac{h(u)^2}{h(1)}\right) + k(h(1) - h(0)). \tag{51}$$

Let us check the cases $a = 0$ and $r = 0$.

For the case $a = 0$, first note that the derivatives of the loss w.r.t r and u_i's are zero. The derivative of the loss w.r.t a is

$$\partial_a L\big|_{a=0} = -2 \sum_{i=1}^{k} r h(u_i), \tag{52}$$

which is zero if $r = 0$. If $r > 0$, the sum of potentials is positive if any of the similarities satisfy $u_i > -1$ which yields a negative derivative. Therefore, $u_1 = -1$ with arbitrary $r > 0$ is also a critical point for the case with one teacher neuron.

For the case $a \neq 0$ and $r = 0$, first note that the derivatives w.r.t a and u_i's are zero thanks to the factorization property of ReLU (in Eq. 46). The derivative of the loss w.r.t. r is

$$\partial_r L\big|_{r=0} = a^2 \partial_r r^2 h(1)\big|_{r=0} - 2a \sum_{i=1}^{k} \partial_r r h(u_i)\big|_{r=0} = -2a \sum_{i=1}^{k} h(u_i). \tag{53}$$

Since $r = 0$ enforces $u_i = 0$, we get that the sum of potentials is positive which yields a non-zero derivative. This completes the proof of Theorem 3.3.

C.3 Proof of Theorem 3.4 for Softplus

In this section, we prove Theorem 3.4 where the activation function is assumed to be a softplus

$$\sigma(x) = \frac{1}{\beta} \log(e^{\beta x} + 1)$$

with $\beta > 0$. The potential does not have a known analytic expression in this case unlike ReLU, hence the proof involves some techniques to compare some ratios of averages, in particular the FKG inequality. We also rely on some specific properties of the softplus family that are developed in Section C.3.4. Unfortunately, some of these properties do not apply to other activation functions such as sigmoid and tanh. However, as a first example of managing potentials without an analytic expression, we hope that the proof inspires generalizations to other activation functions.

Below we present the proof skeleton. In the following Subsections C.3.1, C.3.2, C.3.3, and C.3.4, the components of the proof are presented in detail.

Proof Sketch. Rearranging the terms in the Eq. 41 and writing the potentials explicitly, we get

$$f(u, r) = \frac{\langle \sigma'(rx)\sigma(rx)x \rangle}{\langle \sigma(rx)^2 \rangle} - \frac{\langle \sigma'(rx)\sigma(y)x \rangle}{\langle \sigma(rx)\sigma(y) \rangle}. \tag{54}$$

Defining the helper functions

$$G(r) = \frac{\langle \sigma'(rx)\sigma(rx)x \rangle}{\langle \sigma(rx)^2 \rangle} = \frac{1}{2}\frac{d}{dr} \log(\langle \sigma(rx)^2 \rangle), \tag{55}$$

$$\tilde{G}(u, r) = \frac{\langle \sigma'(rx)\sigma(y)x \rangle}{\langle \sigma(rx)\sigma(y) \rangle} = \frac{d}{dr} \log(\langle \sigma(rx)\sigma(y) \rangle), \tag{56}$$

we will use the following expression

$$f(u, r) = G(r) - \tilde{G}(u, r). \tag{57}$$

We first characterize the zeros of f in Section C.3.1. In particular, we show that for $r \in [0, 1]$, there is a unique correlation $u \in [0, 1]$ such that $f(u, r) = 0$. We denote this correlation by $h(r)$ such that $h : [0, 1] \to [0, 1]$. Thanks to Lemma C.1, we have that $\partial_u f \neq 0$. Therefore, we can invoke the implicit function theorem and obtain that h is smooth since f is smooth. We also show that $h(0) = 0$ and $h(1) = 1$. Finally for $r > 1$, we show that there is no solution of f.

In Section C.3.2, we give a lower bound on the correlation $h(r)$. Using this lower bound and with the help of Stein's Lemma, we give a lower bound on the outgoing weight in Section C.3.3.

Combining all, we conclude that any critical point of the one neuron incoming vector which has similarity/correlation $u = 1/\sqrt{k}$ with the teacher incoming vectors must have a norm r that is smaller than or equal to u so as to satisfy the derivative constraints in the Eq. 41, and a lower bound on the outgoing weight follows. This completes the proof of Theorem 3.4. *End of Proof Sketch.*

C.3.1 Characterizing the Zeros of f

In this subsection, we will find all zero-crossings of $f : [-1, 1] \times [0, \infty) \to \mathbb{R}$.

Let us check the corner cases of $r = 0$ and $r = 1$ first.

For $r = 0$: Note that $G(0) = \tilde{G}(0, 0)$. Since $\tilde{G}$ is increasing in the correlation (Lemma C.1), the only solution in this case is $u = 0$.

For $r = 1$: Note that $G(1) = \tilde{G}(1, 1)$ since $y = x$ due to correlation 1. Since $\tilde{G}$ is increasing in the correlation (Lemma C.1), the only solution in this case is $u = 1$.

We next show that there is no solution for $r > 1$.

For $r > 1$: We first show that $G(r) > \tilde{G}(1, r)$ for $r > 1$, which is equivalent to

$$\langle \sigma'(rx)\sigma(rx)x\rangle\langle\sigma(rx)\sigma(x)\rangle > \langle\sigma'(rx)\sigma(x)x\rangle\langle\sigma(rx)^2\rangle. \tag{58}$$

Changing the measure of x from the standard Gaussian $p(x)$ to $\tilde{p}(x) = p(x)\sigma(rx)^2/\langle\sigma(rx)^2\rangle$, we get the following equivalent inequality

$$\left\langle \frac{\sigma'(rx)x}{\sigma(rx)} \right\rangle \left\langle \frac{\sigma(x)}{\sigma(rx)} \right\rangle > \left\langle \frac{\sigma'(rx)x}{\sigma(rx)} \frac{\sigma(x)}{\sigma(rx)} \right\rangle. \tag{59}$$

From the property (iv) of Lemma C.3, we have that $\sigma'(rx)x/\sigma(rx)$ is increasing after a substitution $x \to rx$. We need to show $\sigma(x)/\sigma(rx)$ is decreasing in x for $r > 1$, thus we take the derivative

$$\frac{d}{dx}\frac{\sigma(x)}{\sigma(rx)} = \frac{\sigma'(x)\sigma(rx) - \sigma(x)\sigma'(rx)r}{\sigma(rx)^2}. \tag{60}$$

Since $\sigma'(x)x/\sigma(x)$ is increasing, we also have

$$\frac{\sigma'(x)x}{\sigma(x)} < \frac{\sigma'(rx)rx}{\sigma(rx)} \text{ for } x > 0, \text{ and } \frac{\sigma'(x)x}{\sigma(x)} > \frac{\sigma'(rx)rx}{\sigma(rx)} \text{ for } x < 0. \tag{61}$$

This yields $\sigma'(x)\sigma(rx) < \sigma(x)\sigma'(rx)r$, hence that $\sigma(x)/\sigma(rx)$ is decreasing from Eq. 60. Thanks to the FKG inequality, we conclude that $\sigma'(rx)x/\sigma(rx)$ and $\sigma(x)/\sigma(rx)$ are negatively correlated.

Since From Lemma C.1, $\tilde{G}$ is increasing in correlation, we have $\tilde{G}(1, r) > \tilde{G}(u, r)$ for any $u \leq 1$, therefore there is no solution of f for the case of $r > 1$.

For $r \in (0, 1)$: Now we can reduce the search domain to $r \in (0, 1)$ to find the solutions of f. We next show that for any $r \in (0, 1)$, there is a unique $u \in (0, 1)$ such that $f(u, r) = 0$. It suffices to show

$$\tilde{G}(0, r) < G(r) < \tilde{G}(1, r),$$

since $\tilde{G}$ is continuous and increasing in correlation (Lemma C.1), it crosses $G(r)$ at a unique u.

First inequality: $\tilde{G}(0, r) < G(r)$. In this case, x and y are Gaussians with zero correlation, therefore they are independent. We can expand $\tilde{G}(0, r)$ by splitting the average

$$\tilde{G}(0, r) = \frac{\langle\sigma'(rx)x\rangle\langle\sigma(y)\rangle}{\langle\sigma(rx)\rangle\langle\sigma(y)\rangle} = \frac{\langle\sigma'(rx)x\rangle}{\langle\sigma(rx)\rangle}. \tag{62}$$

We want to show

$$\langle\sigma'(rx)x\rangle\langle\sigma(rx)^2\rangle < \langle\sigma'(rx)\sigma(rx)x\rangle\langle\sigma(rx)\rangle. \tag{63}$$

which is equivalent to the following inequality after changing the measure from standard Gaussian $p(x)$ to $\tilde{p}(x) = p(x)\sigma(rx)/\langle\sigma(rx)\rangle$

$$\left\langle \frac{\sigma'(rx)x}{\sigma(rx)} \right\rangle_{\tilde{p}} \langle\sigma(rx)\rangle_{\tilde{p}} < \left\langle \frac{\sigma'(rx)x}{\sigma(rx)}\sigma(rx) \right\rangle_{\tilde{p}}. \tag{64}$$

This is again a consequence of the FKG inequality since we have that both $\sigma'(x)x/\sigma(x)$ and $\sigma(x)$ are increasing from the properties (i) and (iv) of the softplus (Lemma C.3).

Second inequality: $G(r) < \tilde{G}(1, r)$ for $r < 1$. This is equivalent to the Ineq. 58, but the direction of the inequality is reversed. We showed that $\sigma(x)/\sigma(r'x)$ is decreasing in x for all $r' > 1$, therefore its reciprocal $\sigma(r'x)/\sigma(x)$ is increasing in x. Now substituting $x \to rx$ where $r = 1/r' < 1$, we get that $\sigma(x)/\sigma(rx)$ is increasing in x for all $r > 1$. This yields positive correlation between $\sigma'(rx)x/\sigma(rx)$ and $\sigma(x)/\sigma(rx)$ from the FKG inequality, hence the proof is complete.

In summary, we showed that there is zero-crossing of f for $r > 1$ and for $r \in [0, 1]$, there is a unique correlation u, denoted by $h(r)$ which satisfies the following

i. $h(0) = 0$ and $h(1) = 1$,

ii. for $r \in (0, 1)$, we have $h(r) \in (0, 1)$.

Therefore, there is a unique map h such that $f(h(r), r) = 0$ for all $r \in [0, 1]$. f is smooth since σ is smooth. Moreover, since $\tilde{G}$ is increasing (Lemma C.1), its derivative is not zero, which yields $\partial_u f = -\partial_u \tilde{G} \neq 0$. Therefore, the implicit function theorem gives that for every $r \in (0, 1)$, there is a neighborhood of r and $\bar{h} \in \mathcal{C}_\infty$ such that $f(\bar{h}(r), r) = 0$. Since h is unique, $h(r) = \bar{h}(r)$ in the neighborhood of r for every $r \in (0, 1)$, hence h is also $\mathcal{C}_\infty$.

C.3.2 Bound on the Norm

In this subsection, we will show that $h(r) \geq r$ for all $r \in (0, 1)$. Let us assume the contrary, which implies

$$\tilde{G}(h(r), r) < \tilde{G}(r, r)$$

due to Lemma C.1. It suffices to show that for all $r \in (0, 1)$, we have

$$\tilde{G}(r, r) \leq G(r), \tag{65}$$

since this yield a contradiction with $G(r) = \tilde{G}(h(r), r)$. This piece is equivalent to

$$\langle \sigma'(rx)\sigma(rx + r'z)x \rangle \langle \sigma(rx)^2 \rangle \leq \langle \sigma'(rx)\sigma(rx)x \rangle \langle \sigma(rx)\sigma(rx + r'z) \rangle \tag{66}$$

where $r' = \sqrt{1 - r^2}$. After a change of measure from standard Guassian $p(x)$ to

$$\tilde{p}(x) = p(x) \frac{\langle \sigma(rx + r'z) \rangle_z}{\sigma(rx)} \Bigg/ \left\langle \frac{\sigma(rx + r'z)}{\sigma(rx)} \right\rangle, \tag{67}$$

this is equivalent to the following inequality

$$\left\langle \frac{\sigma'(rx)x}{\sigma(rx)} \right\rangle_{\tilde{p}} \left\langle \frac{\sigma(rx)}{\langle \sigma(rx + r'z) \rangle} \right\rangle_{\tilde{p}} \leq \left\langle \frac{\sigma'(rx)x}{\sigma(rx)} \frac{\sigma(rx)}{\langle \sigma(rx + r'z) \rangle} \right\rangle_{\tilde{p}}. \tag{68}$$

What remains to show is that

$$\frac{\langle \sigma(rx + r'z) \rangle_z}{\sigma(rx)}$$

is non-increasing in x since then we can conclude by the FKG inequality. Since $r > 0$ we can drop it up to a change in the standard deviation of x. We want to show that its derivative is non-positive

$$\sigma(x)\langle \sigma'(x + r'z) \rangle \leq \sigma'(x)\langle \sigma(x + r'z) \rangle \;\Leftrightarrow\; \frac{\sigma(x)}{\sigma'(x)} \leq \frac{\langle \sigma(x + r'z) \rangle}{\langle \sigma'(x + r'z) \rangle}. \tag{69}$$

From the property (iii) of softplus (Lemma C.3), we have that $R(x) = \sigma(x)/\sigma'(x)$ is convex. Applying Jensen, we get

$$\frac{\sigma(x)}{\sigma'(x)} \leq \left\langle \frac{\sigma(x + r'z)}{\sigma'(x + r'z)} \right\rangle. \tag{70}$$

What remains to show is that

$$\left\langle \frac{\sigma(x + r'z)}{\sigma'(x + r'z)} \right\rangle \langle \sigma'(x + r'z) \rangle \leq \langle \sigma(x + r'z) \rangle. \tag{71}$$

Note that $\langle \sigma'(x + r'z) \rangle_z$ is increasing in x since σ'' is positive everywhere. Moreover, the function

$$\left\langle \frac{\sigma(x + r'z)}{\sigma'(x + r'z)} \right\rangle_z$$

is increasing in x since its integrand R is increasing from the property (iii) of softplus (Lemma C.3). Finally, the FKG inequality completes the proof.

In summary, we showed that for a given $u \in (0, 1)$, if r solves the fixed point Eq. 41, it satisfies $r \geq u$.

C.3.3 Bounding the Outgoing Weight

To get a bound on a, let us analyze the ratio of potentials in Eq. 41

$$\frac{g(r,1,u)}{g(r,r,1)} = \frac{a}{k}. \tag{72}$$

Using the convexity of softplus (property (i) of Lemma C.3), we get

$$\frac{\langle \sigma(rx)\sigma(ux+u'z)\rangle}{\langle \sigma(rx)^2\rangle} \geq \frac{\langle \sigma(rx)\sigma(rx)\rangle + \langle \sigma(rx)((u-r)x+u'z)\sigma'(rx)\rangle}{\langle \sigma(rx)^2\rangle} \tag{73}$$

$$= 1 + (u-r)\frac{\langle \sigma'(rx)\sigma(rx)x\rangle}{\langle \sigma(rx)^2\rangle}. \tag{74}$$

We can transform the numerator using the Stein's lemma with $v(x) = \sigma(rx)\sigma'(rx)$

$$\langle \sigma(rx)\sigma'(rx)x\rangle = r\langle \sigma'(rx)\sigma'(rx) + \sigma(rx)\sigma''(rx)\rangle \tag{75}$$

which is positive since softplus is positive, increasing, and convex. Combining it with $u \geq r$, we get that the ratio is bounded below by 1 which yields the lower bound k on the outgoing weight a.

C.3.4 Helper Lemmas

In this subsection, we provide helper lemmas used in the proof of Theorem 3.4. We first present Lemma C.1 and its proof where we use Lemma C.2, which shows that $\tilde{G}$ is increasing in correlation for all $u \in [0,1]$ for any given $r \geq 0$. Finally, we present several properties of the softplus family in Lemma C.3.

Lemma C.1. *The following function is increasing in* $u \in [0,1]$

$$\tilde{G}(u,r) = \frac{\langle \sigma'(rx)\sigma(y)x\rangle}{\langle \sigma(rx)\sigma(y)\rangle} \tag{76}$$

for any $r \geq 0$*, where* x *and* y *are standard Gaussians with correlation* $\langle xy\rangle = u$.

Proof. Let us first reparameterize $y = ux + u'z$ where $u' = \sqrt{1-u^2}$ to make the correlation u explicit. Derivating it w.r.t u, we get

$$\partial_u \tilde{G}(u,r) = \frac{\langle \sigma'(rx)\sigma'(y)x^2\rangle}{\langle \sigma(rx)\sigma(y)\rangle} - \frac{\langle \sigma'(rx)\sigma(y)x\rangle\langle \sigma(rx)\sigma'(y)x\rangle}{\langle \sigma(rx)\sigma(y)\rangle^2}.$$

Showing $\partial_u \tilde{G}(u,r) > 0$ is equivalent to the following

$$\langle \sigma'(rx)\sigma'(y)x^2\rangle\langle \sigma(rx)\sigma(y)\rangle > \langle \sigma'(rx)\sigma(y)x\rangle\langle \sigma(rx)\sigma'(y)x\rangle.$$

Changing the measure from the standard Gaussian $p(x)$ to

$$\tilde{p}(x) = p(x)\frac{\sigma(rx)\langle \sigma(y)\rangle_z}{\langle \sigma(rx)\sigma(y)\rangle} \tag{77}$$

where $\langle \sigma(y)\rangle_z$ denotes the conditional average over z with fixed x, we get

$$\left\langle \frac{\sigma'(rx)}{\sigma(rx)}x\frac{\langle \sigma'(y)\rangle_z}{\langle \sigma(y)\rangle_z}x\right\rangle_{\tilde{p}} > \left\langle \frac{\sigma'(rx)}{\sigma(rx)}x\right\rangle_{\tilde{p}}\left\langle \frac{\langle \sigma'(ux+u'z)\rangle}{\langle \sigma(ux+u'z)\rangle}x\right\rangle_{\tilde{p}}. \tag{78}$$

Thanks to the property (iv) of softplus (Lemma C.3), we have that $\sigma'(rx)x/\sigma(rx)$ is increasing in x after a substitution $x \to rx$. For $r=0$, the function reduces to αx with some $\alpha > 0$, thus it is increasing also in this case. In Lemma C.2, we show also that $\langle \sigma'(y)\rangle_z x/\langle \sigma(y)\rangle_z$ is increasing in x considering x with standard deviation u and z with standard deviation u'. For $u=0$, the function reduces to αx with some $\alpha > 0$, thus it is increasing also in this case.

Therefore, we conclude with the FKG inequality that $\sigma'(rx)x/\sigma(rx)$ and $\langle \sigma'(y)\rangle_z x/\langle \sigma(y)\rangle_z$ are positively correlated. □

Lemma C.2. *The following function*

$$\frac{\langle \sigma'(x+z)\rangle_z x}{\langle \sigma(x+z)\rangle_z}$$

is increasing in x where z is a centered Gaussian with arbitrary standard deviations.

Proof. In the proof, all averages are w.r.t. the standard Gaussian z, hence we drop the subindex. Taking the derivative, we want to show

$$\left(\frac{\langle \sigma''(x+z)\rangle x}{\langle \sigma'(x+z)\rangle} + 1\right)\langle \sigma(x+z)\rangle > \langle \sigma'(x+z)\rangle x \tag{79}$$

which is equivalent to the following due to the property $\sigma''(z) = \beta\sigma'(z)(1-\sigma'(z))$

$$\left(\beta x\left(1 - \frac{\langle \sigma'(x+z)^2\rangle}{\langle \sigma'(x+z)\rangle}\right) + 1\right)\langle \sigma(x+z)\rangle > \langle \sigma'(x+z)\rangle x. \tag{80}$$

In the case $x \geq 0$, the LHS is bigger than $\langle \sigma(x+z)\rangle$ since $\sigma'(\cdot)$ is upper bounded by 1. Moreover, since $\sigma(x) > x$ and from the convexity of softplus, we get $\langle \sigma'(x+z)\rangle > x$. This yields the above inequality by again noting that $\langle \sigma'(x+z)\rangle$ is upper bounded by 1.

In the case $x < 0$, we need another strategy. We have thanks to Cauchy-Schwartz

$$\frac{\langle \sigma'(x+z)^2\rangle}{\langle \sigma'(x+z)\rangle} \geq \langle \sigma'(x+z)\rangle, \tag{81}$$

thus it suffices to show

$$(\beta x - \beta x\langle \sigma'(x+z)\rangle + 1)\langle \sigma(x+z)\rangle > \langle \sigma'(x+z)\rangle x. \tag{82}$$

We will show the following

$$\langle \sigma'(x+z)\rangle \geq \sigma'(x), \tag{83}$$

which does not trivially come from Jensen's inequality since the sigmoid is convex only for $x < 0$. Let us explicitly write out the sigmoid

$$\left\langle \frac{e^{\beta(x+z)}}{e^{\beta(x+z)}+1}\right\rangle - \frac{e^{\beta x}}{e^{\beta x}+1} \geq 0 \;\Leftrightarrow\; \left\langle \frac{e^{\beta z}-1}{e^{\beta(x+z)}+1}\right\rangle \geq 0. \tag{84}$$

Since we have $e^{\beta(x+z)} \leq e^{\beta z}$ for all $x < 0$, it suffices to show

$$\left\langle \frac{e^{\beta z}-1}{e^{\beta z}+1}\right\rangle \geq 0 \;\Leftrightarrow\; \langle \sigma'(z)\rangle \geq \frac{1}{2} \tag{85}$$

which comes from noting that $\sigma'(z) + \sigma'(-z) = 1$, and since the distribution of z is symmetric around the origin. Hence, it remains to show

$$\frac{\beta x + e^{\beta x} + 1}{e^{\beta x}+1}\langle \sigma(x+z)\rangle > \langle \sigma'(x+z)\rangle x. \tag{86}$$

From the proof of Lemma C.3, we have that $\sigma'(x) < \beta\sigma(x)$, which in combination with the following observation for all $x < 0$

$$\frac{\beta x + e^{\beta x} + 1}{e^{\beta x}+1} = \frac{\beta x}{e^{\beta x}+1} + 1 > \beta x + 1 \tag{87}$$

completes the proof of Eq. 86, hence the proof of the lemma. □

Lemma C.3. *The softplus family has the following properties*

i. *$\sigma(x)$ is increasing and convex,*

ii. *$\sigma(x)$ is log concave (equivalently, $\sigma(x)/\sigma'(x)$ is increasing),*

iii. *$\sigma(x)/\sigma'(x)$ is convex,* *iv.* $\sigma'(x)x/\sigma(x)$ *is increasing.*

Proof. For the property (i), see the proof of Lemma B.2. We next prove each one of the properties one after the other. Let us start with the property (ii). First note that $\sigma(x)$ is log concave if and only if $\sigma(x)/\sigma'(x)$ is increasing since

$$\frac{d}{dx}\frac{\sigma(x)}{\sigma'(x)} = 1 - \frac{\sigma(x)\sigma''(x)}{\sigma'(x)^2} > 0 \Leftrightarrow \sigma'(x)^2 > \sigma(x)\sigma''(x). \tag{88}$$

We will prove that $R(x) := \sigma(x)/\sigma'(x)$ is increasing. Let us write out the ratio explicitly

$$R(x) = \frac{1}{\beta}\left(\log(e^{\beta x}+1) + \frac{\log(e^{\beta x}+1)}{e^{\beta x}}\right). \tag{89}$$

Let us compute the first derivative of R (dropping the factor $1/\beta$ since it does not change the sign)

$$R'(x) = \sigma'(x) + \frac{\sigma'(x) - \beta\sigma(x)}{e^{\beta x}} = \frac{e^{\beta x} - \beta\sigma(x)}{e^{\beta x}}. \tag{90}$$

Since $\log$ is concave, expanding it around 1 we get $\log(y+1) < y$ for all $y > 0$. Substituting $y = e^{\beta x}$, we get that the numerator of R' is positive, thus R is increasing. This completes the proof of property (ii). Computing the second derivative of R, we get

$$R''(x) = \frac{\sigma''(x)(e^{\beta x}+1) - 2\beta\sigma'(x) + \beta^2\sigma(x)}{e^{\beta x}} = \beta\left(\frac{-\sigma'(x) + \beta\sigma(x)}{e^{\beta x}}\right). \tag{91}$$

What remains to show is that $\beta\sigma(x) > \sigma'(x)$. We have the following identity due to the fundamental theorem of calculus

$$\frac{\log(y+1)}{y} = \frac{1}{y}\int_0^y \frac{1}{t+1}dt > \frac{1}{y+1} \tag{92}$$

since $1/(y+1)$ is a lower bound of the integrand. Substituting $y = e^{\beta x}$, we get

$$\log(e^{\beta x}+1) > \frac{e^{\beta x}}{e^{\beta x}+1} \tag{93}$$

which completes the proof of the property (iii). Let us prove the property (iv) by taking the derivative of the function of interest

$$\frac{d}{dx}\frac{\sigma'(x)x}{\sigma(x)} = \frac{(\sigma''(x)x + \sigma'(x))\sigma(x) - \sigma'(x)^2 x}{\sigma(x)^2} \tag{94}$$

Using $\sigma''(x) = \beta\sigma'(x)(1-\sigma'(x))$ and after dropping the positive term $\sigma'(x)$, the numerator of the derivative is

$$((1-\sigma'(x))\beta x + 1)\sigma(x) - \sigma'(x)x = \left(\frac{\beta x}{1+e^{\beta x}} + 1\right)\sigma(x) - \frac{e^{\beta x}}{1+e^{\beta x}}x \tag{95}$$

$$= \frac{e^{\beta x}}{e^{\beta x}+1}\left(\frac{1}{e^{\beta x}}(e^{\beta x} + \beta x + 1)\sigma(x) - x\right) \tag{96}$$

For the case $x \geq 0$, we have $\sigma(x) > x$ and $(\beta x + 1)/e^{\beta x} > 0$, hence the derivative is positive.

For the case $x < 0$, we want to show

$$\left(e^{\beta x} + \beta x + 1\right)\frac{\log(e^{\beta x}+1)}{e^{\beta x}} > \beta x. \tag{97}$$

If $e^{\beta x} + \beta x + 1 > 0$, it is done since the LHS is positive. If $e^{\beta x} + \beta x + 1 \leq 0$, we have

$$\left(e^{\beta x} + \beta x + 1\right)\frac{\log(e^{\beta x}+1)}{e^{\beta x}} \geq \left(e^{\beta x} + \beta x + 1\right)\sup\frac{\log(e^{\beta x}+1)}{e^{\beta x}} \tag{98}$$

since $\log(e^{\beta x}+1)/e^{\beta x}$ is positive. We will next show that $\log(e^{\beta x}+1)/e^{\beta x}$ is a decreasing function therefore its supremum is achieved at $x \to -\infty$. From the integral expression in Eq. 92, we derive that $\log(y+1)/y$ is a decreasing function since adding smaller terms in the average decreases it. Thus the following limit gives us the supremum

$$\lim_{y\to 0}\frac{\log(y+1)}{y} = \lim_{y\to 0}\frac{1}{y+1} = 1. \tag{99}$$

Combining it with the Eq. 98 after the substitution $y = e^{\beta x}$, we get the desired Ineq. 97 which implies that the derivative is positive in this case too. This completes the proof of property (iv). □

D Second-Order Nature of Symmetry-Induced Critical Points

In this section, we prove Theorem 4.1. First we present a proof sketch.

Proof Sketch. We decompose the Hessian of a symmetry-induced critical point $\oplus^{j,\mu}(\theta)$ using a specific linear transformation $A(\mu)$ as follows

$$HL(\oplus^{j,\mu}(\theta)) = A(\mu)^T \begin{bmatrix} \mu(1-\mu)Y & -V & 0 \\ -V & 0 & 0 \\ 0 & 0 & HL(\theta) \end{bmatrix} A(\mu)$$

$$H(\mu) = A(\mu)^T \cdot \widetilde{H}L(\oplus^{j,\mu}(\theta)) \cdot A(\mu).$$

In our decomposition, $A(\mu)$ is invertible for all μ. Thanks to the Sylvester's law of inertia, the number of positive, negative, and zero eigenvalues of $H(\mu) = HL(\oplus^{j,\mu}(\theta))$ are the same as those of $\widetilde{H}(\mu) = \widetilde{H}L(\oplus^{j,\mu}(\theta))$ which is a congruent matrix.

Therefore it suffices to study the eigenvalue signs of $\widetilde{H}(\mu)$ which is composed of two block matrices on the diagonal –the matrix in Eq. 18 and the Hessian of the original local minimum $HL(\theta)$– with off-diagonal blocks being all-zero. Thus the eigenvalues of $\widetilde{H}(\mu)$ are identical to the union of eigenvalues of its block-diagonal matrices. Since θ is a local minimum, $HL(\theta)$ is a positive definite matrix which completes the proof of the first part of the statement. Finally, the eigenvalue signs in the new $d + d_{\text{out}}$ directions are determined by the matrix in Eq. 18 which is the second part of the statement. *End of Proof Sketch.*

Let us now present the full-Hessian of a symmetry-induced critical point

$$HL(\oplus^{j,\mu}(\theta)) = \begin{bmatrix} \mu^2 X + \mu Y & \mu(1-\mu)X & \mu U + V & \mu U & 0 \\ \mu(1-\mu)X & (1-\mu)^2 X + (1-\mu)Y & (1-\mu)U & (1-\mu)U + V & 0 \\ \mu U^T + V^T & (1-\mu)U^T & Z & Z & 0 \\ \mu U^T & (1-\mu)U^T + V^T & Z & Z & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j(\theta)) \end{bmatrix} \tag{100}$$

where X and Y are $d \times d$; U and V are $d \times d_{\text{out}}$; and Z is $d_{\text{out}} \times d_{\text{out}}$ and $HL(\ominus^j(\theta))$ is the Hessian corresponding to the parameter θ except for the j-th neuron, which we denote by $\ominus^j(\theta)$. We need to compute the second order derivatives to write out the submatrices explicitly. First let us compute the first order derivatives

$$\partial_{a_i} L(\theta) = \langle \sigma(w_i x) c'(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$\partial_{w_i} L(\theta) = \langle \sigma'(w_i x) x a_i^T \rangle_{\mathbb{P}}.$$

Then the second-order derivatives follow

$$\partial^2_{w_i w_j} L(\theta) = \langle \sigma'(w_i x) \sigma'(w_j x) x x^T a_i^T c''(f(x), f^*(x)) a_j \rangle_{\mathbb{P}}$$
$$\partial^2_{w_i^2} L(\theta) = \langle \sigma'(w_i x)^2 x x^T a_i^T c''(f(x), f^*(x)) a_i + \sigma''(w_i x) x x^T a_i^T c'(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$\partial^2_{w_j a_i} L(\theta) = \langle \sigma(w_i x) \sigma'(w_j x) x a_j^T c''(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$\partial^2_{w_i a_i} L(\theta) = \langle \sigma(w_i x) \sigma'(w_i x) x a_i^T c''(f(x), f^*(x)) + \sigma'(w_i x) x c'(f(x), f^*(x))^T \rangle_{\mathbb{P}}$$
$$\partial^2_{a_i a_j} L(\theta) = \langle \sigma(w_i x) \sigma(w_j x) c''(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$\partial^2_{a_i^2} L(\theta) = \langle \sigma(w_i x)^2 c''(f(x), f^*(x)) \rangle_{\mathbb{P}}.$$

We introduce the following submatrices to ease the notation

$$X = \langle \sigma'(w_i x)^2 x x^T a_i^T c''(f(x), f^*(x)) a_i \rangle_{\mathbb{P}}$$
$$Y = \langle \sigma''(w_i x) x x^T a_i^T c'(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$U = \langle \sigma(w_i x) \sigma'(w_i x) x a_i^T c''(f(x), f^*(x)) \rangle_{\mathbb{P}}$$
$$V = \langle \sigma'(w_i x) x c'(f(x), f^*(x))^T \rangle_{\mathbb{P}}$$
$$Z = \langle \sigma(w_i x)^2 c''(f(x), f^*(x)) \rangle_{\mathbb{P}}$$

which reduces the second-order derivatives into

$$
\begin{aligned}
&\partial^2_{w_i w_j} L(\theta) = X((w_i, a_i), (w_j, a_j)) \\
&\partial^2_{w_i^2} L(\theta) = X(w_i, a_i) + Y(w_i, a_i) \\
&\partial^2_{w_j a_i} L(\theta) = U((w_i, a_i), (w_j, a_j)) \\
&\partial^2_{w_i a_i} L(\theta) = U(w_i, a_i) + V(w_i, a_i) \\
&\partial^2_{a_i a_j} L(\theta) = Z((w_i, a_i), (w_j, a_j)) \\
&\partial^2_{a_i^2} L(\theta) = Z(w_i, a_i).
\end{aligned}
$$

Note that $X(w_i, a_i)$ and $Z(w_i, a_i)$ are positive definite if the cost c is convex. Moreover $Y(w_i, a_i)$ is a symmetric matrix thus it has real eigenvalues.

Next we change the basis via an invertible matrix A. We obtain the following transformed Hessian denoted by $\tilde{H}$ which has an approximate block-diagonal structure after the change of basis, where $P^- = P - (d + d_{\text{out}})$

$$
\begin{aligned}
\tilde{H}(\mu) &= \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & -V & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ -V & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & HL(\ominus^j(\theta)) \end{bmatrix} \\
\tilde{H}(\mu) &= \begin{bmatrix} (1-\mu)I_d & -\mu I_d & 0 & 0 & 0 \\ I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & (1-\mu)I_{d_{\text{out}}} & 0 \\ 0 & 0 & -I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P-} \end{bmatrix} H(\mu) \begin{bmatrix} (1-\mu)I_d & I_d & 0 & 0 & 0 \\ -\mu I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & -I_{d_{\text{out}}} & 0 \\ 0 & 0 & (1-\mu)I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P-} \end{bmatrix} \\
\tilde{H}(\mu) &= (A(\mu)^{-1})^T H(\mu) A(\mu)^{-1}; \qquad (101)
\end{aligned}
$$

where $A(\mu)$ is given by

$$
A(\mu) = \begin{bmatrix} I_d & -I_d & 0 & 0 & 0 \\ \mu I_d & (1-\mu)I_d & 0 & 0 & 0 \\ 0 & 0 & I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & -(1-\mu)I_{d_{\text{out}}} & \mu I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & I_{P-} \end{bmatrix}.
$$

Finally, after a change of block-rows and block columns, we recover the statement of the Thm. 4.1 due to the following observation

$$
HL(\theta) = \begin{bmatrix} X+Y & U+V & 0 \\ U+V & Z & 0 \\ 0 & 0 & HL(\ominus^j(\theta)) \end{bmatrix}. \qquad (102)
$$

In the case of biases, the decomposition applies in the same way. The only important thing to take into account is the update in the submatrices of Y and V

$$
\begin{aligned}
&Y(w_j, b_j, a_j) = \langle \sigma''([w_j, b_j] \cdot [x, 1])[x, 1][x, 1]^T a_j \cdot c'(f(x), f^*(x)) \rangle_{\mathbb{P}} \in \mathbb{R}^{(d+1)\times(d+1)} \\
&V(w_j, b_j, a_j)_{k\ell} = \langle \sigma'([w_j^\ell, b_j^\ell] \cdot [x, 1])[x, 1]_k c'(f(x), f^*(x))_\ell \rangle_{\mathbb{P}} \text{ with } k \in [d+1], \ell \in [d_{\text{out}}].
\end{aligned}
$$

D.1 Multiple Output Neurons

Lemma D.1. *For multiple output neurons with $d_{out} \geq 2$, if the matrix V is non-vanishing, the submatrix in Eq. 18 has at least one negative eigenvalue.*

Proof. We will show that $\tilde{H}$ has a negative eigenvalue as long as at least one entry of V is non-vanishing, i.e. $V_{k\ell} \neq 0$. It suffices to show that a submatrix of the submatrix, say 2×2 in Eq. 18

has one negative eigenvalue since we can that construct an vector such as $[a_1, a_2, 0]$ which returns a negative direction by picking out 2×2 submatrix. We pick the following 2×2 submatrix

$$\begin{bmatrix} Y_{kk} & -V_{k\ell} \\ -V_{k\ell} & 0 \end{bmatrix}. \tag{103}$$

Note that the determinant of the above matrix is $-V_{k\ell}^2 < 0$ since $V_{k\ell} \neq 0$. This completes the proof. □

Lemma D.1 implies that for multiple number of output neurons, if $V \neq 0$, then all symmetry-induced critical points on the line are strict saddles!

For the mixing ratio $\mu = 0$, changing the corresponding incoming vector does not change the network function. Therefore we obtain a d-dimensional subspace that goes through the SI critical point $\oplus^{j,0}(\theta)$ where the loss remains constant. We also have an additional direction of constant loss in the span of the outgoing vectors which is the one pointing towards the line of symmetry-induced critical points. However, this does not guarantee $d + 1$ zero eigenvalues in its Hessian since this subspace may correspond to the directions that are not eigenvectors, nevertheless the second-order derivatives vanish. This happens for the Hessians that have positive and negative eigenvalues and where the second-order derivative vanish on the directions between the eigenvectors. A simple example is $L(w_1, w_2) = w_1^2 - w_2^2$ where the Hessian is

$$HL = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}.$$

The Hessian HL has no zero eigenvalues, however in the direction $w_1 = w_2$, the loss remains constant, which lies in between the two eigenvectors $[1, 0]$ and $[0, 1]$.

Next we investigate the eigenvalues signs of

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix}. \tag{104}$$

to determine the eigenvalue signs of the Hessian of SI critical points at $\mu \in \{0, 1\}$ in the new directions. Note that the dimensionality of the null space of V is at least one due to the constraint in Eq. 20.

Lemma D.2. *Let V be a matrix of size $d \times d_{out}$ such that dim(Null(V)) $= n \geq 1$. Then the number of zero-eigenvalues of the following matrix*

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix} \tag{105}$$

of size $(d+d_{out}) \times (d+d_{out})$ is at least $|d - d_{out}|$. If $d > d_{out}$, then at least $d - d_{out} + n$ zero-eigenvalues are guaranteed in particular for $n = 1$, the exact number of zero-eigenvalues is $d - d_{out} + 2$.

Proof. Non-zero eigenvalues. First, observe that for every non-zero eigenvalue λ with the eigenvector $[a, b]$

$$\begin{bmatrix} 0 & V \\ V^T & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} Vb \\ V^T a \end{bmatrix} = \begin{bmatrix} \lambda a \\ \lambda b \end{bmatrix},$$

$-\lambda$ is an eigenvalue corresponding to the eigenvector $[-a, b]$ due to the following

$$\begin{bmatrix} Vb \\ -V^T a \end{bmatrix} = \begin{bmatrix} -\lambda(-a) \\ -\lambda b \end{bmatrix}.$$

In short, the non-zero eigenvalues of the matrix in Eq. 104 come in pairs $(\lambda, -\lambda)$.

Zero eigenvalues. We search for the number of different solutions (up to sign and scaling) of the following equation

$$\begin{bmatrix} Vb \\ V^T a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For V, recall that we have $Va_j = 0$.

Case1: $d_{out} \leq d$. In this case V^T is $d_{\text{out}} \times d$ so it has a null-space of the dimension at least $d - d_{\text{out}}$. We choose $d - d_{\text{out}}$ orthogonal vectors spanning it, say $v_1, ..., v_{d-d_{\text{out}}}$. Concatenating each one of $d - d_{\text{out}}$ vectors with 0 gives orthogonal eigenvectors, i.e. $[v_1, 0], ..., [v_{d-d_{\text{out}}}, 0]$ of the matrix in Eq. 104 with zero eigenvalues. In addition, we can concatenate the 0 vector with a_j which is in the null space of V, i.e. $[0, a_j]$, which is orthogonal to the others. In general, if $\dim(N(V)) = n$, by concatenating all with zero vectors, we get $[0, v]$ eigenvectors that are orthogonal to each other and others of the form $[v, 0]$. Therefore, we constructed $d - d_{\text{out}} + n$ orthogonal eigenvectors with zero eigenvalues.

Finally we know that the number of non-zero eigenvalues should be even. If n is odd, so is $(d + d_{\text{out}}) - (d - d_{\text{out}} + n) = 2d_{\text{out}} - n$, therefore there has to be at least one more zero eigenvalue.

In this case, there are at least $d - d_{\text{out}} + n + 1$ zero eigenvalues.

On the other hand, the rank of $V^T V$ is $d_{\text{out}} - n$. Let v be an eigenvector with a non-zero eigenvalue λ^2. Therefore, $[\frac{1}{\lambda} Vv, v]$ is an eigenvector of the matrix in Eq. 104 with the eigenvalue λ. Following this construction, overall we get $2(d_{\text{out}} - n)$ non-zero eigenvalues.

If $n = 1$, there are exactly $2(d_{\text{out}} - 1)$ non-zero and $d - d_{\text{out}} + 2$ zero eigenvalues.

Case2: $d < d_{out}$. In this case V is $d \times d_{\text{out}}$, therefore it has a null-space of the dimension at least $d_{\text{out}} - d$. Concatenating each one of them with 0 vectors, we find at least $d_{\text{out}} - d$ zero eigenvalues.

Note the asymmetry between the two cases: a_j is a non-zero vector in the null space of V, but we do not have such a knowledge for V^T thus its null space may be 0-dimensional. □

D.2 Bounding the Minimal Hessian Eigenvalue

Now using the decomposition, we will provide an upper bound for the minimum eigenvalue of the Hessian. In this section we denote the Hessian by H to ease notation (i.e. dropping the loss L) or by $H(\mu)$ where it makes sense to emphasize μ.

D.2.1 Negative Minimum Eigenvalue

The Rayleigh quotient for any $u \neq 0$ upper bounds the minimum eigenvalue

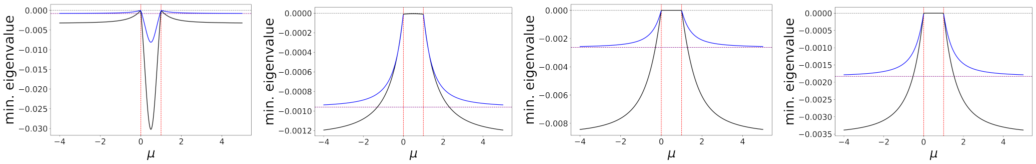
$$\frac{u^T H u}{u^T u} \geq \lambda_{\min}(H).$$

Plugging in the decomposition (Eq. 102), we get

$$\frac{u^T A^T \tilde{H} A u}{u^T u} = \frac{v^T H v}{(A^{-1}v)^T A^{-1} v} \geq \lambda_{\min}(H).$$

for any vector v. Wlog we can assume that it is a unit vector. Let us choose v such that $[v_0, 0, 0, 0, 0]$ where v_0 is a unit eigenvector of Y with an eigenvalue λ_0. Thus we have

$$\begin{aligned}
v^T \tilde{H} v &= \begin{bmatrix} v_0^T & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & -V & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ -V & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & H(\ominus^j(\theta)) \end{bmatrix} \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} v_0^T & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(1-\mu)\lambda_0 v_0 \\ 0 \\ 0 \\ -Vv_0 \\ 0 \end{bmatrix} \\
&= \mu(1-\mu)\lambda_0 \qquad (106)
\end{aligned}$$



(a) neuron-I **(b)** neuron-II **(c)** neuron-III **(d)** neuron-IV

Figure 6: *The minimal Hessian eigenvalue of the sym. ind. strict saddles as a function of μ (black) and the upper bound (blue).* We observe that the upper bound on the most negative eigenvalue of the Hessian qualitatively captures the behavior of the most negative eigenvalue. In the cases (b-c-d), the matrix Y is positive-definite, the upper bound for the line segment $\mu \in (0, 1)$ is positive. Since we already know that the min. eigenvalue for this line segment is zero, the upper bound is not plotted.

We need to check

$$A^{-1}v = \begin{bmatrix} (1-\mu)I_d & I_d & 0 & 0 & \\ -\mu I_d & I_d & 0 & 0 & 0 \\ 0 & 0 & \mu I_{d_{\text{out}}} & -I_{d_{\text{out}}} & 0 \\ 0 & 0 & (1-\mu)I_{d_{\text{out}}} & I_{d_{\text{out}}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (1-\mu)v_0 \\ -\mu v_0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

so the norm of $A^{-1}v$ is

$$\|A^{-1}v\|^2 = (1-\mu)^2 + \mu^2.$$

Therefore by choosing a specific unit eigenvector v, we obtained the following upper bound on the minimum eigenvalue

$$\frac{\mu(1-\mu)}{(1-\mu)^2 + \mu^2}\lambda_0 \geq \lambda_{\min}(H)$$

which is valid for every μ and every eigenvalue of Y. To make the bound tightest using this form of v, we need to choose v_0 as the extreme eigenvectors of Y. We obtain (see Fig. 6):

- *for* $\mu \in (0, 1)$*:*

$$\frac{\mu(1-\mu)}{(1-\mu)^2 + \mu^2}\lambda_{\min}(Y) \geq \lambda_{\min}(H(\mu)) \text{ for } \mu \in (0, 1),$$

- *for* $\mu \in \mathbb{R}/[0, 1]$*:*

$$\frac{\mu(1-\mu)}{(1-\mu)^2 + \mu^2}\lambda_{\max}(Y) \geq \lambda_{\min}(H(\mu)) \text{ for } \mu \in \mathbb{R}/[0, 1].$$

In particular, in the limits as $\mu \to \pm\infty$, we get

$$-\frac{1}{2}\lambda_{\max}(Y) \geq \lim_{\mu \to \pm\infty} \lambda_{\min}(H(\mu)).$$

Next, using an additive decomposition, we will give a lower bound on the minimum eigenvalue. Note that the entries of the Hessian are at most quadratic in μ, which can be written out as

$$\underbrace{\mu^2 \begin{bmatrix} X & -X & 0 & 0 & 0 \\ -X & X & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{M_2} + \mu \underbrace{\begin{bmatrix} Y & X & U & U & 0 \\ X & -2X - Y & -U & -U & 0 \\ U^T & -U^T & 0 & 0 & 0 \\ U^T & -U^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{M_1} + \underbrace{\begin{bmatrix} 0 & 0 & V & 0 & 0 \\ 0 & X + Y & U & U + V & 0 \\ V^T & U^T & Z & Z & 0 \\ 0 & U^T + V^T & Z & Z & 0 \\ 0 & 0 & 0 & 0 & H(\ominus^j(\theta)) \end{bmatrix}}_{M_0}$$

since X is positive definite. Moreover, $\lambda_{\min}(M_2) = 0$ since we can choose an eigenvector $[v, v, 0, 0]$ where v is an eigenvector of X.

Using the following general inequality twice

$$\lambda_{\min}(A+B) = \min_{\|u\|=1} u^T(A+B)u \geq \min_{\|u\|=1} u^T Au + \min_{\|u\|=1} u^T Bu = \lambda_{\min}(A)+\lambda_{\min}(B),$$

we obtain the following bound on the min. eigenvalue for $\mu \geq 0$

$$\lambda_{\min}(H(\mu)) \geq \mu^2\lambda_{\min}(M_2) + \mu\lambda_{\min}(M_1) + \lambda_{\min}(M_0) \geq \mu\lambda_{\min}(M_1) + \lambda_{\min}(M_0),$$

and for $\mu < 0$

$$\lambda_{\min}(H(\mu)) \geq \mu\lambda_{\max}(M_1) + \lambda_{\min}(M_0).$$

In particular, as $\mu \to \infty$, we show that $\lambda_{\min}(H(\mu))$ is at most linearly decreasing.

D.2.2 Minimum Positive Eigenvalue & One Output Neuron

If the minimum eigenvalue is not negative, we know that it is zero since all symmetry-induced critical points have a zero-eigenvalue in the Hessian. In this case, we want to bound the minimum positive eigenvalue to get a measure of sharpness for the symmetry-induced local minimum. Recalling the decomposition in the case of one output neuron, we have

$$H = A^T \begin{bmatrix} \mu(1-\mu)Y & 0 & 0 & 0 & 0 \\ 0 & X+Y & U+V & 0 & 0 \\ 0 & U+V & Z & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & H(\ominus^j(\theta)) \end{bmatrix} A. \tag{107}$$

We know the eigenvector of $\tilde{H}$ corresponding to the trivial-zero eigenvalue, let's denote it by $e = [0, 0, 0, 1, 0]$. First note that

$$HA^{-1}e = A^T\tilde{H}AA^{-1}e = 0. \tag{108}$$

Let's denote the minimum non-negative eigenvalue of H by $\lambda^+_{\min}$ excluding the trivial zero corresponding to the eigenvector $A^{-1}e$. We have the following upper bound for all $u \perp A^{-1}e$

$$\frac{u^T Hu}{u^T u} \geq \lambda^+_{\min}(H). \tag{109}$$

Plugging in the decomposition we get

$$\frac{u^T A^T\tilde{H}Au}{u^T u} = \frac{v^T\tilde{H}v}{(A^{-1}v)^T A^{-1}v} \geq \lambda^+_{\min}(H) \tag{110}$$

where for any $v \perp e$. We can choose $v = [v_0, 0, 0, 0, 0]$ which is orthogonal to e where v_0 is an eigenvector of Y as in the previous case (Sec. D.2.1) which gives us the following upper bound

$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_0 \geq \lambda^+_{\min}(H).$$

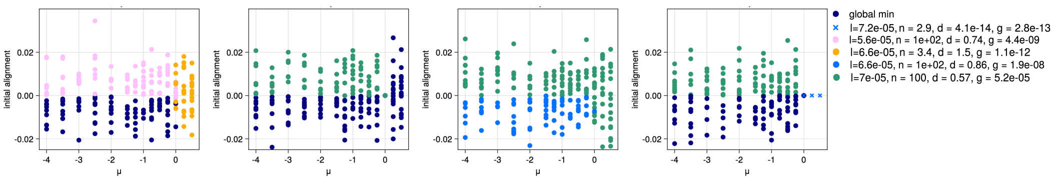
Therefore the tightest bounds are as follows:

- *positive definite Y, for $\mu\in(0,1)$:*

$$\frac{\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_{\min}(Y) \geq \lambda^+_{\min}(H);$$

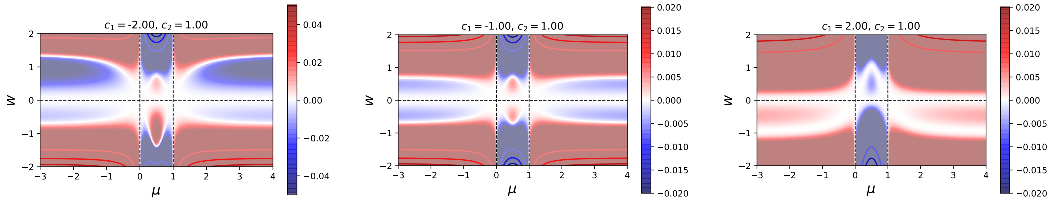
- *negative definite Y, for $\mu\in\mathbb{R}/[0,1]$:*

$$\frac{-\mu(1-\mu)}{(1-\mu)^2+\mu^2}\lambda_{\min}(|Y|) \geq \lambda^+_{\min}(H).$$



(a) $\oplus^{1,\mu}(\theta)$ **(b)** $\oplus^{2,\mu}(\theta)$ **(c)** $\oplus^{3,\mu}(\theta)$ **(d)** $\oplus^{4,\mu}(\theta)$

Figure 7: *The sign of the initial alignment of the perturbation with the minimum eigenvector determines where the gradient flow trajectories end up.* The gradient trajectories can converge to (i) a global minimum point (dark blue), (ii) a local minimum at infinity (pink-green), (iii) a local minimum (orange), or (iv) an SI-minima (blue cross). Each panel represents splitting one of the four neurons. In the legend, n represents the norm of the final parameter and d is the distance to the initial line of symmetry-induced critical points (and l is the loss, g is the gradient norm). **(a-b-c)** All SI-critical points are strict saddles except for $\mu = 0$. In **(a)** for $\mu \in (0, 0.5)$, both positive and negative perturbations converge to the same local minimum (orange), for $\mu < 0$, positive perturbations converge to a local min. at infinity (pink), and the negative ones converge to a global min.; in **(b)** for $\mu < 0$, positive perturbations converge to a local min. at infinity (green), and all others converge to a global min.; in **(c)** for $\mu < 0$, negative perturbations converge to another local min. at infinity (blue), positive ones converge to the green local min. at infinity, also for $\mu \in (0, 0.5)$, all perturbations converge to the same local min. In **(d)** SI-critical points are local minima for $\mu \in (0, 0.5)$, so all perturbations converge to them; for $\mu < 0$, positive perturbations converge to the green local min. and the negative perturbations converge to a global min. At the non-strict saddles at $\mu = 0$, since the multiplicity of the zero eigenvalue is three; we cannot identify a unique eigenvector corresponding to the minimum eigenvalue.



(a) **(b)** **(c)**

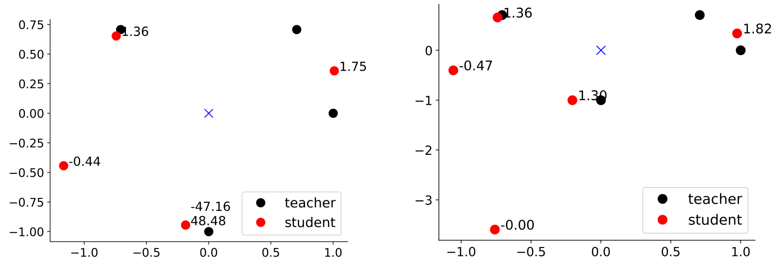
Figure 8: *Loss surfaces corresponding to the two-dimensional model with constants* $c_2 = 1$, $c_1 \in \{-2, -1, 2\}$ *in (a)-(b)-(c), respectively.* **(a)-(b)** $c_1 c_2 < 0$: perturbations from the strict saddles at $(0, \mu)$ with $\mu \in \mathbb{R}/[0, 1]$ move towards a local minimum at infinity, the horizontal alignment of the local minima is $c_1/2$ or $c_2/2$. **(c)** $c_1, c_2 > 0$: perturbations from the strict saddles at $(0, \mu)$ with $\mu \in (0, 1)$ converge towards either a local minimum at $(\min\{c_1, c_2\}, 0.5)$ (for positive perturbations, i.e. $w > 0$), or diverge away to $-\infty$ (for negative perturbations, i.e. $w < 0$).

D.2.3 Perturbations & a model of local minimum at infinity near the line of critical points

We propose the following model for understanding the geometry of the loss landscape near the line of symmetry-induced saddles. Our focus is to understand the mechanism of how a local minimum emerges at infinity near the line of critical points and how the perturbation from the line of critical points either converge to one of the two types of minima.

$$\mathcal{L}(w, \mu) = -w^2 \left(\frac{w^2}{4} - \frac{w}{3} u(\mu)(c_1 + c_2) + \frac{1}{2} u(\mu)^2 c_1 c_2 \right) u(\mu)$$

$$\text{where } u(\mu) = \frac{\mu(1-\mu)}{\mu^2 + (1-\mu)^2}$$



(a) symmetry-induced local minimum **(b)** new local minimum

Figure 9: *Perturbations from the symmetry-induced saddles corresponding to local minimum-I shown in Fig. 3 either converges to (a) symmetry-induced local minimum at infinity or to a (b) new local minimum. We plot the network that we obtained at a late snapshot in training in Fig. 9-a. Note that the parameter vector effectively corresponds to a symmetry-induced local minimum corresponding to local minimum-II of the network with* 4 *neurons. The gradient flow dynamics pushes the parameter vector towards infinity such that the outgoing weights grow in opposite signs and cancel out each other.*

Note that the loss is zero on the line of $\mu = \{0, 1\}$ and $w = 0$. The partial derivatives are given by

$$\partial_w \mathcal{L}(w, \mu) = -w(w - u(\mu)c_1)(w - u(\mu)c_2)u(\mu),$$

$$\partial_\mu \mathcal{L}(w, \mu) = -w^2 \left(-\frac{w}{3}u'(\mu)(c_1 + c_2) + u(\mu)u'(\mu)c_1c_2\right) u(\mu) - w^2 \left(\frac{w^2}{4} - \frac{w}{3}u(\mu)(c_1 + c_2) + \frac{1}{2}u(\mu)^2 c_1c_2\right) u'(\mu),$$

where both are zero at $w = 0$. Therefore, we have a line of critical points at zero-loss on the line $w = 0$. On the line of critical points, due to the w^2 term in the loss, the mixed second-derivatives $\partial^2_{w\mu}$ and $\partial^2_{\mu\mu}$ vanish. The only non-vanishing second derivative is then

$$\partial^2_{ww}\mathcal{L}(w = 0, \mu) = -2c_1c_2u(\mu)^3.$$

If the constants c_1, c_2 have opposite signs, say $c_1 > 0$ and $c_2 < 0$ wlog, we have that for $\mu \in \mathbb{R}/[0, 1]$ the critical points are strict saddles, and for $\mu \in (0, 1)$ they are local minima.

In this model, we observe

1. for $\mu \in (0, 1)$, all perturbations converge back to the line segment of local min.
2. for $\mu > 1$, for a positive perturbation such that $w > 0$, the trajectories converge to the local minimum at $(c_1/2, +\infty)$
3. for $\mu > 1$, for a negative perturbation such that $w < 0$, the trajectories converge to the local minimum at $(c_2/2, +\infty)$
4. for $\mu < 0$, due to permutation-symmetry, positive perturbations go to the local minimum at $(c_2/2, -\infty)$ and negative ones go to the local minimum at $(c_1/2, -\infty)$.

If the constants have the same signs, then for $\mu \in \mathbb{R}/[0, 1]$, the critical points are local minima; for $\mu \in (0, 1)$, they are strict saddles. If both $c_1, c_2 > 0$ wlog, perturbing from a strict saddle at some $\mu \in (0, 1)$, for positive $w > 0$, the trajectories get stuck at the local min. of $(\min\{c_1, c_2\}, 0.5)$, and for negative $w < 0$, there is no local min. and the trajectories diverge away.