Annihilation of Families of Spurious Minima in Two-Layer ReLU Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the optimization problem associated with fitting two-layer ReLU neural networks with respect to the squared loss, where labels are generated by a target network. Use is made of the rich symmetry structure to develop a novel set of tools for studying the mechanism by which over-parameterization annihilates spurious minima. Sharp analytic estimates are obtained for the loss and the Hessian spectrum at different minima, and it is proved that adding neurons can turn symmetric spurious minima into saddles; minima of lesser symmetry require more neurons. Using Cauchy's interlacing theorem, we prove the existence of descent directions in certain subspaces arising from the symmetry structure of the loss function. This analytic approach uses techniques, new to the field, from algebraic geometry, representation theory and symmetry breaking, and confirms rigorously the effectiveness of over-parameterization in making the associated loss landscape accessible to gradient-based methods. For a fixed number of neurons and inputs, the spectral results remain true under symmetry breaking perturbation of the target.

1 Introduction

An outstanding question in deep learning (DL) concerns the ability of simple gradient-based methods to successfully train neural networks despite the nonconvexity of the associated optimization problems. Indeed, nonconvex optimization landscapes may have spurious (i.e., non-global) minima with large basins of attraction and this can cause a complete failure of these methods. Evidence suggests that this problem can be circumvented by the use of a large number of parameters in DL models; that is, *over-parameterization*. In view of the complexity exhibited by contemporary neural networks and the absence of suitable analytic tools, much recent research has focused on two-layer ReLU networks as a realistic starting point for a theoretical study [10, 12, 34, 27, 20, 36, 5, 30]. The two-layer networks considered were typically of the form:

$$f(\boldsymbol{x}; W, \boldsymbol{\alpha}) = \sum_{i=1}^{k} \alpha_i \varphi(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle),\ W \in M(k,d),\ \boldsymbol{\alpha} \in \mathbb{R}^k, \tag{1}$$

where $\varphi(z) = \max\{0, z\}$ is the ReLU activation, $M(k,d)$ denotes the space of $k \times d$ matrices and $\boldsymbol{w}_i$ denotes the i'th row of W. In order to isolate the study of optimization-related obstructions due to nonconvexity from issues pertaining to the expressive power of two-layer networks, data has been often assumed to be fully *realizable*. This was further motivated by hardness results which indicated a strict barrier inherent to the explanatory power of all distribution-free approaches operating in complete generality [9, 11, 33]. For the squared loss, the resulting, highly nonconvex, expected loss is

$$\mathcal{L}(W, \boldsymbol{\alpha}) \coloneqq \frac{1}{2} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\Big(f(\boldsymbol{x}; W, \boldsymbol{\alpha}) - f(\boldsymbol{x}; V, \boldsymbol{\beta}) \Big)^2 \right], \tag{2}$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

where $\mathcal{D}$ denotes a probability distribution over the input space, $W \in M(k,d)$, $\alpha_i \in \mathbb{R}^k$ are the optimization variables, and $V \in M(d,d)$, $\beta_i \in \mathbb{R}^d$ are fixed parameters.

The choice of the d-variate normal Gaussian distribution for the input distribution has drawn a considerable amount of interest, e.g., [14, 38, 26, 35, 16, 8, 5, 1]. Empirically, it has been observed that as the number of neurons k increases, the loss of models obtained through the stochastic gradient descent (SGD) algorithm under Xavier initialization [18] decreases, see [29]. The objective of this work is to study the mathematical mechanism behind this fundamental phenomenon. Using ideas based on *symmetry breaking* (see Section 2), we are able to use methods from representation theory and algebraic geometry to study how the loss landscape is transformed due to over-parameterization. A key feature of our approach is the use of fractional power series (FPS) in d and path based techniques. For example, we may define the spectrum of the Hessian as an FPS for *real* values of d and so determine the values of d (typically not integers) where eigenvalues change sign. These methods allow the proof of powerful analytic results: in this paper, we derive sharp analytic estimates of the loss and the Hessian spectrum at local minima allowing us to analyze the mechanisms whereby spurious minima are annihilated when the number of neurons k is increased. Our contributions can be summarized by

Theorem 1 *We describe several infinite families of minima, partly characterized by their symmetry, and prove that:*
• *Over-parameterization transforms symmetric spurious minima into saddles—minima of lesser symmetry require more neurons.*
• *The Hessian spectrum remains extremely skewed with $\Theta(d)$ eigenvalues growing linearly with d, and $\Theta(kd)$ eigenvalues being $\Theta(1)$.*
• *Mild over-parameterization adds $\Theta(d)$ descent directions in a $\Theta(d)$-dimensional subspace dictated by the symmetry structure (i.e., the isotypic decomposition) of the loss function.*
• *The loss remains (essentially) unchanged.*

To our knowledge, prior to the present work, it has not been possible to investigate analytically the mechanism which makes the highly nonconvex population loss accessible to gradient-based methods for mild over-parameterization corresponding to regimes of k and d encountered in practice. To demonstrate the subtly of the results given in Theorem 1, consider a family of spurious minima with symmetry ΔS_{d-1} (defined in Section 2). The minima occur for $d \geq 6$. Adding one neuron results in these minima occurring for $d \geq 8$ — not promising. Adding two neurons annihilates *all* these spurious minima and creates no new spurious minima. The mechanism is finite, cannot be inferred from a limiting case, and *persists under forced symmetry breaking* and so applies beyond symmetric targets [3]. See Table 1 for the effect of adding one neuron to a family of spurious minima with maximal symmetry. Next, we relate our results to the existing literature.

Annihilation of spurious minima under over-parameterization. Existing methods for the analysis of (2) are based on: mean-field [27], optimal control [12], and NTK [21, 13, 25] and the thermodynamic limit [8, 19, 28, 20]. These methods operate by passing to limiting regimes where the number of inputs or neurons is taken to infinity. A growing number of works has severely limited, if not invalidated, the explanatory power of these approaches [37, 17]. Several approaches have been developed to address the loss landscapes in the finite parameter regime. [36] obtains several generalities on critical points; [30] study a technical condition under which one can add a single neuron in a contrived way to turn a given minimum in $M(d,d)$ into a saddle in $M(d+1,d)$. However, as is shown in the present work, families of minima in $M(d+1,d)$ exist, and more importantly, families of minima of lesser symmetry provably require at least two additional neurons before turning into saddles.

Symmetry breaking in nonconvex loss landscapes. It has been recently discovered that the symmetry of spurious minima for optimization problem (2) *break* the symmetry of global minima [4, 2] (see Section 2 for a formal exposition). The present work builds on methods developed in a line of work concerning this phenomenon of symmetry breaking. In [7] infinite families of critical points are constructed and represented by convergent fractional power series (FPS) in d^{-1}. The FPS result is used in [5], together with results from the representation theory of the symmetric group, to obtain precise analytic estimates on the Hessian spectrum. In [6], these results are extended to two-layer ReLU networks where it is shown that the spectra are identical to $O(d^{-\frac{1}{2}})$-order for the global minima and several families of spurious minima.

The results above assume that the number of inputs to the network is greater than or equal to the number of neurons: $d \geq k$. The present work concerns the overparametrized case $k > d$, and introduces three new techniques that facilitate the technically more demanding study of over-parameterization. Here, the initial coefficients of FPS cannot always be given in simple algebraic form. Rather than using the explicit formulae used in [5] for computing the Hessian spectrum, we developed a method for computing the spectrum by expressing eigenvalues in terms of the gradient entries. Secondly, the number of isotropic copies of the standard and the trivial representation grows linearly with k for fixed d. Our use of Cauchy's interlacing theorem [23] yields a tight characterization, not attainable with previous methods, of descent directions afforded by over-parameterization. Lastly, in contrast to previous works which addressed local properties of families of minima, we are able to study global properties and provide an exhaustive list of all critical points in fixed point spaces of interest. This result is needed for showing that *all* spurious minima of a certain isotropy type must turn into saddles (see Theorem 1). We illustrate the methods by focusing on six families of critical points. Orthogonality of the target matrices is not required by the symmetry-breaking framework: other choices of target matrices, distributions, activation functions and architectures have been addressed in previous works. Extensions and generalizations of Theorem 1 are discussed in the concluding comments.

Isotypic component	Degree	Spectrum $k = d$	Spectrum $k = d + 1$
$\mathfrak{x}$-representation (mult. 1)	$\frac{(d-1)(d-2)}{2}$	$\frac{1}{4} - \frac{1}{2\pi}$	$\frac{1}{4} - \frac{1}{2\pi}$
$\mathfrak{y}$-representation (mult. 1)	$\frac{d(d-3)}{2}$	$\frac{1}{4} + \frac{1}{2\pi}$	$\frac{1}{4} + \frac{1}{2\pi}$
trivial representation $\mathfrak{t}$: mult. 2, if $k = d$, mult. 3, if $k = d + 1$	1	$\frac{d}{2\pi} + \frac{-6\pi+4+\pi^2}{4\pi(-2+\pi)}$ $\frac{d}{4} + \frac{-4+2\pi+\pi^2}{4\pi(-2+\pi)}$	$\frac{1}{4} - \frac{1}{2\pi}$ $\frac{d}{2\pi} + \frac{-4+\pi}{4(-2+\pi)}$ $\frac{d}{4} + \frac{-2+\pi^2}{2\pi(-2+\pi)}$
standard representation $\mathfrak{s}_d$: mult. 3, if $k = d$ mult. 4, if $k = d + 1$	$d - 1$	$\frac{1}{4} - \frac{1}{2\pi}$ $\frac{1}{4}$ $\frac{d}{4} + \frac{1}{4}$	$\frac{-1-\sqrt{5}}{4\pi} + \frac{1}{4}$ $\frac{1}{4} - \frac{1}{2\pi}$ $\frac{-1+\sqrt{5}}{4\pi} + \frac{1}{4}$ $\frac{d}{4} + \frac{1}{2}$
Loss at critical point		$\frac{1}{2} - \frac{1}{\pi}$	$\frac{1}{2} - \frac{1}{\pi}$

Table 1: For $k = d, d + 1$, we present the Hessian spectrum, labelled by each of the 4 representations of the symmetric group S_d that occur for the family of type I critical points with ΔS_d symmetry. *Degree* is the dimension of the irreducible representation. Entries in columns 3 and 4 are modulo $O(d^{-\frac{1}{2}})$-terms. The analytic description of the Hessian spectrum (see Section B), reveals that adding one neuron turns the family of minima $\mathfrak{C}^{\mathrm{I}}_{0,0}$ into saddles. We show that minima with the smaller symmetry, ΔS_{d-1}, require two neurons before turning into saddles (see Section 5). The eigenspace of the negative eigenvalue (highlighted in red) of the over-parameterized family is a subspace of the factor of the isotopic decomposition corresponding to the standard representation of S_d. This holds for other families of minima studied in this work and highlights the significance of the *standard* representation of S_d in over-parametrization. Characteristically, the Hessian spectrum is extremely skewed, with $\Theta(d)$ eigenvalues growing linearly with d, and $\Theta(d^2)$ being small constant; a phenomenon which has been observed empirically many times, e.g., [24, 31, 32].

Organization of the paper The proof of Theorem 1 is in four parts: symmetry & the loss function, families of symmetric minima, the isotypic decomposition, and analysis of the loss landscape in the presence of over-parameterization. Proofs and technical details are deferred to the appendix.

2 Symmetry and the loss function

The presentation of our results requires some familiarity with group and representation theory. Key ideas and concepts are introduced as needed; elementary definitions are assumed (see [5, 7] for more details). After establishing notational conventions, we review background material.

The *symmetric group* S_d, $d \in \mathbb{N}$, is the group of permutations of $[d] \doteq \{1, \ldots, d\}$; the *orthogonal group* $\mathrm{O}(d)$ is the subgroup of all orthogonal linear maps on $\mathbb{R}^d$. We may identify S_d with the

subgroup of $\mathrm{O}(d)$ consisting of permutation matrices. Both groups define group actions: S_d acts naturally on $[d]$ (as permutations) and on $\mathbb{R}^d$ (as permutation matrices); $\mathrm{O}(d)$ acts on $\mathbb{R}^d$ as orthogonal linear transformations. For $k, d \in \mathbb{N}$, there are natural actions of S_k and S_d on the space $M(k,d)$ of $k \times d$ matrices: S_k permutes rows, S_d permutes columns. The loss function (2) is always invariant under row permutations: $\mathcal{L}(\sigma W) = \mathcal{L}(W)$, for all $\sigma \in S_k$—whatever the choice of $V, \boldsymbol{\beta}$.

The product action of $S_k \times S_d$ on $M(k,d)$ will play an important role in studying invariance properties of $\mathcal{L}$. Formally, if $A = [A_{ij}] \in M(k,d)$, $(\pi, \rho) \in S_k \times S_d$, then

$$(\pi, \rho)[A_{ij}] = [A_{\pi^{-1}(i), \rho^{-1}(j)}],\ \pi \in S_k,\ \rho \in S_d,\ (i,j) \in [k] \times [d]. \tag{3}$$

The action can equivalently be defined in terms of permutation matrices. If $k \geq d$, define the *diagonal subgroup* ΔS_d of $S_k \times S_d$ by $\Delta S_d = \{(g,g) \mid g \in S_d \subseteq S_k\}$. Clearly, $\Delta S_d \approx S_d$.

Henceforth assume $k \geq d$. Regard $M(d,d)$ as the linear subspace of $M(k,d)$ defined by appending $k-d$ zero rows to each matrix in $M(d,d)$. Set $V = I_d \in M(k,d)$. By the orthogonal symmetry of the Gaussian distribution, our results will hold for any matrix in $V \in \mathrm{O}(d)$. For simplicity, we shall assume that the second layer of weights is fixed to ones. Indeed, by [6] our symmetry-based analysis is readily applicable to trainable second-layers, mutatis mutandis. Optimization problem (2) exhibits a rich symmetry structure. For our choice of $V = I_d$, $\mathcal{L}$ is $S_k \times S_d$-invariant [4, Section 4.1]. It is natural to ask how the critical points of $\mathcal{L}$ reflect this symmetry.

Remark 1 *If $k > d$, $V \in M(k,d)$ has a row of zeros and $\mathcal{L}$ is not differentiable at $W = V$ [10]. It may be shown that there is a $k-d$-dimensional compact connected $S_k \times S_d$-invariant simplicial complex $\Lambda \subset M(k,d)$ consisting of all matrices that define the global minimum value of zero. At boundary points of Λ, $\mathcal{L}$ is not differentiable. The Hessian, defined on the interior of the simplex, is* always *singular. None of this matters here: our focus will be on families of spurious minima with non-degenerate critical points (no zero eigenvalues).*

If we take $V = I_d$ and $k = d$ (thereby avoiding the issue raised in Remark 1), V is fixed by the action of ΔS_d. Modulo group conjugation, this holds for any global minimizer of $\mathcal{L}$, see [7, Proposition 4.14.]. It turns out that the subgroup ΔS_d of $S_k \times S_d$ is the group to focus on when $k \geq d$. Given $W \in M(k,d)$, the largest subgroup of $S_k \times S_d$ fixing W is called the *isotropy* subgroup of W. This group measures the *symmetry* of W. Empirically, nondegenerate spurious minima of $\mathcal{L}$ tend to be highly symmetric in the sense that their isotropy groups are (conjugated to) large subgroups of ΔS_d. Indeed, we suspect that for our choice of target V the isotropy of non-degenerate spurious minima is always conjugate to a (large) proper subgroup of the isotropy group of V.

3 Families of minima: structure and basic properties

Recent work shows that families of spurious minima often have characteristic properties. For example, the asymptotics in d of the loss or their chance of being detected by SGD. In order to pursue a systematic study of the optimization landscape of $\mathcal{L}$, we need to categorize minima and understand their distinctive analytic properties. Isotropy is a natural candidate for this task.

Assume $k = d + m \geq 0$, for all $d \geq d_0$, where $m \geq 0$ is an integer constant. Our aim is to define the notion of a *regular family of critical points of* $\mathcal{L}$. Some preliminaries are needed. Let G be a subgroup of ΔS_d and $M(k,d)^G = \{W \in M(k,d) \mid gW = W, g \in G\}$ denote the *fixed point space* for the action of G on $M(k,d)$. It follows by equivariance [7, Section 3.3] that $\mathfrak{c} \in M(k,d)^G$ is a critical point of $\mathcal{L}$ if and only if $\mathfrak{c}$ is a critical point of $\mathcal{L}|M(k,d)^G$.

A sequence $(G_d)_{d \geq d_0}$ of subgroups of ΔS_d is *natural* if (a) there are inclusions $i_d : \Delta S_d \subset \Delta S_{d+1}$ such that $i_d(G_d) \subset G_{d+1}$, (b) $\dim(M(k,d)^{G_d})$ is independent of d.

Example 1 *Let $d/2 \geq p \geq 0$, and set $G_d = \Delta(S_{d-p} \times S_p)$. If $p = 0$, then $G_d = \Delta S_d$ and $\dim(M(d,d)^{G_d}) = 2$, $d \geq d_0 = 2$. If $p = 1$, then $\dim(M(d+1,d)^{\Delta(S_{d-1} \times S_1)}) = 7$, for all $d \geq 3$. If $p \geq 2$, then $\dim(M(d+2,d)^{\Delta(S_{d-p} \times S_p)}) = 10$, all $d \geq d_0 = 5$. $2 \leq p$. These families are natural—the inclusion i_d is the restriction of the product of the natural inclusion of ΔS_{d-p} in ΔS_{d+1-p} with the identity on ΔS_p.*

If $(G_d)_{d \geq d_0}$ is natural and $n = \dim(M(k,d)^{G_d})$, we may identify each $M(k,d)^{G_d}$ with $\mathbb{R}^n$. For ReLU activation, $\mathcal{L}|\mathbb{R}^n$ will be a continuous family in the real variable d ($k = d + m$), which is real

analytic outside of a thin subset of $\mathbb{R}^n$ ([7, Section 4]). This allows us to "connect" the critical points $\mathfrak{c}(d), \mathfrak{c}(d+1)$, $d \geq d_0$.

Definition 1 *A family $\mathfrak{C} = \{\mathfrak{c}(d) \mid d \geq d_0\}$ of critical points of $\mathcal{L}$ is* regular *if $(G_d = G_{\mathfrak{c}(d)})_{d \geq d_0}$ is natural and for $d \geq d_0$*

(a) There is a continuous curve $\gamma_d : [0,1] \to \mathbb{R}^n$ of critical points of $\mathcal{L}|\mathbb{R}^n$ joining $\mathfrak{c}(d)$ to $\mathfrak{c}(d+1)$.

(b) For all real $d \geq d_0$, $\mathcal{L}$ is real analytic on a neighbourhood of $\gamma_d([0,1]) \subset M(k,d)$.

(c) The Hessian of $\mathcal{L}|\mathbb{R}^n$ at $\mathfrak{c}(d)$ is non-singular, all real $d \geq d_0$.

Using (b,c), the implicit function theorem for real analytic maps implies that γ_d is real analytic.

The loss $\mathcal{L}$ is real analytic on a neighbourhood of $\mathfrak{c}$ if no two rows of $\mathfrak{c}$ are parallel and no row of $\mathfrak{c}$ is parallel to a non-zero row of the target matrix V. Note that $V \in M(d,d)$ does not satisfy the second condition (and so (b) of the definition fails, even if $k = d$).

It was shown in [7] that for $k \leq d$ several families of critical points had fractional power series (FPS) representations in $1/\sqrt{d}$ (and so the families are regular). More generally, if we are given a regular family of critical points $\mathfrak{C} = \{\mathfrak{c}(d) \mid d \geq d_0\}$, there is an FPS representation of the critical points if there exist $d_1 \geq d_0$ and minimal $q \in \mathbb{N}$ such that for $1 \leq i \leq d, 1 \leq j \leq k$, $d \geq d_1$, there are unique real sequences $(C^n_{i,j})_{n \geq 0}$, such that the ij-component of $\mathfrak{c}(d)$ is given by the convergent series

$$\mathfrak{c}(d)_{ij} = \sum_{n=0}^{\infty} C^n_{i,j} d^{-\frac{n}{q}}. \tag{4}$$

For the results in [7], $q = 2$. However, if $k > d$, there are examples where $q = 4$, described below.

The FPS representation is an important tool both theoretically and computationally since it implies that the objective value and the Hessian spectrum have an FPS representation and so we can use asymptotic arguments to establish properties of the Hessian spectrum for large d. We refer to Section A for the existence of FPS for $k > d$.

In addition to the isotropy type, an additional property required for our classification of non-degenerate families of critical points concerns their FPS representation.

Definition 2 *Let $d/2 \geq p \geq 0$ and take $G_d = S_{d-p} \times S_p$ as in Example 1. A regular family $\mathfrak{C}$ of critical points with isotropy $(G_d)_{d \geq d_0}$ is of type I (resp. type II) if as $d \to \infty$, the diagonal elements of the $(d-p) \times (d-p)$-block corresponding to the action of ΔS_{d-p} converge to -1 (resp. $+1$).*

The definition may be equivalently stated in terms of FPS series: for type I (resp. type II), $C^0_{ii} = -1$ (resp. $+1$) for all $i \leq d-p$. Note that condition (a) of definition 1 implies that if $p = 0, k = d$, the type II family cannot be the family $(V = I_d)$ of critical points giving the global mimimum.

Theorem 2 *For isotropy ΔS_d with $k \in \{d, d+1\}$, there is one family of critical points of type I; for isotropy $\Delta(S_{d-p} \times S_p)$, $p \in \{1,2\}$ with $k \in \{d, d+1, d+2\}$, there exist two families of nondegenerate critical points: one of type I, one of type II.*

For a given specification covered by Theorem 2 of an isotropy group $\Delta(S_{d-p} \times S_p)$, a type X and a pair of k and d with $m = k - d$ fixed, let $\mathfrak{C}^{\mathrm{X}}_{p,m}$ denote the family of critical points corresponding to these parameters. Thus, $\mathfrak{c}^{\mathrm{X}}_{p,m}(d) \in M(k,d)$. For $k = d$, type I and type II are families of spurious minima (see Section B below). For reasons which are not yet understood, the bias induced by Xavier initialization seems to favor type II minima. Since the loss at type II minima decays as $\Theta(1/d)$ and the loss at type I is $\Theta(1)$ (independently of the isotropy), it is tempting to argue that the expected initial loss under Xavier initialization is smaller than the loss at type I minima. However, this turns out to be false.

Proposition 1 *Under Xavier initialization, $\left(1 - \frac{2}{\pi}\right)d \leq E_W[\mathcal{L}(W)] \leq \left(1 - \frac{1}{\pi}\right)d$.*

The explanation for the bias of gradient-based methods towards type II should therefore be addressed based on different analytic properties.

Once the existence (and the uniqueness) of an FPS representation has been established, the next goal is to extract sufficient information about the initial coefficients. A description of this process is given in Section A. To demonstrate the quantitative power of the symmetry-based approach, we present here the FPS representation of $\mathfrak{C}^{\mathrm{I}}_{0,0}$ and $\mathfrak{C}^{\mathrm{I}}_{0,1}$ Let $\mathcal{I}_{k,d}$ be the all-ones $(k \times d)$-matrix, and $A_d(a_1, a_2) := a_1 I_d + (\mathcal{I}_{d,d} - I_d)a_2$, then

1. $\mathfrak{c}^{\mathrm{I}}_{0,0}(d) = A(a_1(d), a_2(d))$ where

$$a_1 = -1 + 2d^{-1} + O\big(d^{-2}\big), \text{ and } a_2 = 2d^{-1} + O\big(d^{-2}\big). \tag{5}$$

$$\mathcal{L}(\mathfrak{c}^{\mathrm{I}}_{0,0}(d)) = \frac{1}{2} - \frac{1}{\pi} - \frac{4}{3\pi\sqrt{d}} + \left(-\frac{1}{2} - \frac{2}{\pi^2} + \frac{3}{\pi}\right)d^{-1} + \left(-\frac{8}{\pi^2} + \frac{58}{15\pi}\right)d^{-\frac{3}{2}} + O(d^{-2}).$$

2. $\mathfrak{c}^{\mathrm{I}}_{0,1}(d) = \left(\dfrac{A_d(a_1, a_2)}{a_3 \mathcal{I}_{1,d}}\right)$, where

$$a_1 = -1 + 2d^{-1} + 2d^{-\frac{3}{2}} + O\big(d^{-2}\big),\ a_2 = 2d^{-1} + O\big(d^{-2}\big),\ \text{ and } a_3 = \frac{2}{d} + \frac{4}{\pi}d^{-\frac{3}{3}} + O\big(d^{-2}\big), \tag{6}$$

$$\mathcal{L}(\mathfrak{c}^{\mathrm{I}}_{0,1}(d)) = \frac{1}{2} - \frac{1}{\pi} - \frac{4}{3\pi}d^{-\frac{1}{2}} + \left(-1 - \frac{2}{\pi^2} + \frac{4}{\pi}\right)d^{-1} + \left(-\frac{4}{\pi^2} + \frac{58}{15\pi}\right)d^{-\frac{3}{2}} + O\big(d^{-2}\big).$$

Despite the sharp analytic estimates given above, it may not be possible to obtain simple algebraic expressions for the coefficients of low-order terms of an FPS representation; they may arise as roots of nonlinear equations. Although this is the situation for type II families if $k > d$, coefficients of low order terms may be computed to arbitrarily high precision. For type I families, $k = d + 1, d + 2$, the FPS are power series in $d^{-\frac{1}{4}}$. For example, if $k = d + 1$, $i \in \{d, d + 1\}$, then $(\mathfrak{c}^{\mathrm{I}}_{1,1}(d))_{i,d} = \pm\frac{1}{2} + \alpha d^{-\frac{1}{4}} + O(d^{-\frac{1}{2}})$, where $\alpha \approx 0.57441\ldots$. However, as discussed in the next section, we may avoid these computations but still characterize the Hessian spectrum to the desired order.

4 Hessian spectrum

In [5], it is shown how the FPS representation makes possible an analytic characterization of the Hessian spectrum using tools from the representation theory of groups. For completeness, a description of the method is given in Section B. The main tool used is the *isotypic decomposition* which relates the isotropy of a given point to invariant minimal subspaces of the Hessian. We start by presenting the isotypic decomposition needed for the over-parameterized cases we consider.

Let $k \geq d$. Regard $M(k, d)$ as an S_d-representation (diagonal action). By restriction, $M(k, d)$ is a $S_q \times S_p$-representation, where $0 \leq p \leq d$, $q = d - p$, and $S_q \times S_p \subset S_d$. If $p \in \{0, 1\}$ (the case of interest here) the isotypic decomposition uses 4 irreducible representations of S_q, $d \geq 4$: the trivial representation $\mathfrak{t}$ of degree 1, the standard representation $\mathfrak{s}_q$ of S_q of degree $q - 1$, the exterior square representation $\mathfrak{x}_q = \wedge^2\mathfrak{s}_q$ of degree $\frac{(q-1)(q-2)}{2}$ and a representation $\mathfrak{y}_q$ of degree $\frac{(q)(q-3)}{2}$ (associated to the partition $(q-2, 2)$ [22, 15]). If $p \in \{0, 1\}$, the isotypic decomposition of $(M(k, d), S_{d-p} \times S_p)$ is

$$M(k, d) = (k - d + a + 2)\mathfrak{t} + (k - d + b + 3)\mathfrak{s}_{d-p} + \mathfrak{x}_{d-p} + \mathfrak{y}_{d-p}, \tag{7}$$

where $a = b = 0$ if $p = 0$, and $a = 3, b = 2$ if $p = 1$. By the isotypic decomposition,

Lemma 1 *Fix $p < d/2$ and assume $k \geq d$. Of the kd eigenvalues of the Hessian at any point of isotropy $\Delta(S_{d-p} \times S_p)$:*

1. *$\Theta(d^2)$ are populated by two eigenvalues: The $\mathfrak{x}$- and $\mathfrak{y}$-representation eigenvalues.*
2. *At most $O(k - d)$ eigenvalues are distinct.*

Lemma 1 implies that the isotropy type of a point restricts the number of distinct eigenvalues of the Hessian spectrum. For fixed small values of $k - d$, the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalues account for $kd - \Theta(d)$ of the kd eigenvalues. Using the FPS representation, we show that, perhaps surprisingly, for all families of critical points considered in the paper the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalue are identical to $O(1/\sqrt{d})$-order.

Theorem 3 *For a family of critical points of isotropy $\Delta(S_{d-p} \times S_p),\ p \in \{0,1,2\}$ and k as in Theorem 2, $kd - \Theta(d)$ of the Hessian eigenvalues are populated by the two eigenvalues:*

$$\frac{1}{4} - \frac{1}{2\pi} + O\left(\frac{1}{\sqrt{d}}\right), \quad \textit{and} \quad \frac{1}{4} + \frac{1}{2\pi} + O\left(\frac{1}{\sqrt{d}}\right),$$

associated with the $\mathfrak{x}$- and the $\mathfrak{y}$-representation, respectively.

The derivation of Theorem 3 builds on a technique used in [5], but differs in that it applies to values of k which are strictly greater than d. Crucially, in this regime, the coefficients of FPS may not be given in a simple algebraic form and this makes it hard to use the methods of [5] which depend on explicit formulas for the lower order terms of the FPS. We develop a novel technique which circumvents this issue. The technique proceeds by re-writing the expression of the Hessian eigenvalues in terms of the gradient entries. Since gradient entries vanish at critical points, this allow us to evaluate the eigenvalue expressions with little knowledge of the coefficients. A description of this technique is given in Section A.7.

It follows from Theorem 3 that all Hessian eigenvalues not associated to the trivial or standard representations are strictly positive for sufficiently large d. Consequently, annihilation of spurious minima in a family must be tangent to an invariant subspace of the sum of the isotypic components for the trivial and standard representations. Generically, it is to be expected that the subspace will be isomorphic to either the trivial representation or the standard representation.

5 Over-parameterization

Having proven that the Hessian eigenvalues associated to the $\mathfrak{x}$- and the $\mathfrak{y}$-representation are strictly positive for all large enough d, it remains to analyze the eigenvalues associated with the trivial and the standard representations. We show that when the number of neurons is increased, some eigenvalues associated to the standard representation become negative. Eigenvalues associated to the trivial representation remain strictly positive. We start with points of isotropy ΔS_d.

5.1 Type I critical points on $M(d,d)$ and $M(d+1,d)$

By Theorem 2, there is precisely one family of ΔS_d-critical points: the type I family $\mathfrak{C}^{\mathrm{I}}_{0,i}$ defined on $M(d+i,d)$ for $i = 0,1$ (and given in (5 and (6)). Using the representation-theoretic tools presented in Section B, we derive the complete characterization of the Hessian spectrum of the family of critical points given in Table 1. Spectral analysis of the Hessian reveals that: (a) $\mathfrak{C}^{\mathrm{I}}_{0,0}$ is a family of minima, (b) adding one neuron turns it into the family $\mathfrak{C}^{\mathrm{I}}_{1,0}$ of non-degenerate saddles. More can be said. Since the negative eigenvalue of the Hessian at $\mathfrak{C}^{\mathrm{I}}_{0,1}$ is associated with the standard representation, it follows that there exist exactly $d-1$ descent directions in the $4d-4$-dimensional isotypic component $4\mathfrak{s}_d$ of $(M(d+1,d), S_d)$ (see Section A.7). This isotypic component may be written as the sum of 4 standard representations (only non-zero elements are listed):

1. The $(d-1)$-dimensional space of $(d+1)\times d$-matrices $[a_{ij}]$ where for $i,j \in [d]$, $a_{ij} = x_i - x_j$, for some $(x_1, \cdots, x_d) \in \mathbb{R}^d$ with $\sum_{i\in[d]} x_i = 0$.
2. The $(d-1)$-dimensional space of $(d+1) \times d$-matrices $[a_{ij}]$ where for $i,j \in [d]$, $i \neq j$, $a_{ij} = x_i + x_j$, where $(x_1, \cdots, x_d) \in \mathbb{R}^d$ with $\sum_{i\in[d]} x_i = 0$.
3. The $(d-1)$-dimensional space of $(d+1) \times d$-matrices whose diagonal elements sum up to zero.
4. The $(d-1)$-dimensional space of $(d+1) \times d$-matrices whose $(d+1)$'th row elements sum up to zero.

Given that ΔS_d-minima of type I turn into saddles when one neuron is added, should we expect all families of minima to be annihilated at the same amount of over-parameterization? Empirically, we know that although for $k = d+1$ the chance to get stuck at a spurious minimum appears lower than for $k = d$, it is still quite large (except for small d, see [29]). In the next section, we study $\Delta(S_{d-1} \times S_1)$ families of minima and show that in this case two neurons are required.

Isotypic component	Degree	$k = d$	$k = d + 1$
		Symmetry ΔS_{d-1}	Symmetry ΔS_{d-1}
Trivial representation	1	0.0908	0.0044, 0.0843
		0.25	0.2632, 0.3121
mult. 5, if $k = d$		0.1591d - 0.3471	0.1591d + 0.7546
mult. 7, if $k = d + 1$		0.25d + 0.25	0.25d + 0.5
		0.25d + 0.8471	0.25d + 1.0979
Standard representation	$d - 2$	0.0908	0.0230, 0.0908
		0.0908	0.0936
mult. 5, if $k = d$		0.25	0.2693
mult. 6, if $k = d + 1$		0.4091	0.5340
		0.25d + 0.25	0.25d + 0.5
Loss		$\Theta(1/d)$	$\Theta(1/d)$

Figure 1: Type II critical points with symmetry ΔS_{d-1}. The Hessian eigenvalues associated to the trivial and the standard representation to four decimal places for, modulo $O(1/\sqrt{d})$-terms.

5.2 Critical points of isotropy $\Delta(S_{d-1} \times S_1)$

Consider the type II $\Delta(S_{d-1} \times S_1)$-critical points $\mathfrak{C}^{\mathrm{II}}_{1,0}$, $\mathfrak{C}^{\mathrm{II}}_{1,1}$ and $\mathfrak{C}^{\mathrm{II}}_{1,2}$, corresponding to $k = d, d + 1$ and $d + 2$ respectively. As for the case of ΔS_d-isotropy, negative eigenvalues are introduced by over-parametrization in the isotypic component associated to the standard representation. However, now two additional neurons are required for this to happen and so turn minima into saddles.

For $k = d$ and $k = d + 1$, the eigenvalues associated to the trivial and the standard representation are strictly positive, see Table 5.2 below. By Theorem 3, the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalues are also strictly positive. Therefore, $\mathfrak{C}^{\mathrm{II}}_{1,i}$ are families of minima for $i = 0, 1$. The two families differ in the magnitude of the eigenvalues (see Table 5.2), and in the dimension d above which any critical points is a minimum, $d = 6$ for $\mathfrak{C}^{\mathrm{II}}_{1,0}$ and $d = 7$ for $\mathfrak{C}^{\mathrm{II}}_{1,1}$. For $k = d + 2$, critical points in the family $\mathfrak{C}^{\mathrm{II}}_{1,2}$ are saddles, with strictly negative eigenvalues associated to the standard representation. To show this, we take a different route to reduce the complexity of the computation. Consider the $2d \times 2d$-submatrix $\widehat{H}$ of the Hessian corresponding to rows $(d + 1)$ and $(d + 2)$ of the weight matrix. By Cauchy's interlacing theorem, the smallest eigenvalue of the Hessian is bounded from above by the smallest eigenvalue of $\widehat{H}$. Therefore, it suffices to prove that $\widehat{H}$ has a negative eigenvalue. Since the isotypic decomposition corresponding to $\widehat{H}$ consists of exactly two of the subspaces associated to the standard representation, we have reduced the problem to that of computing the spectrum of a 2×2 matrix. Using the FPS representation, we show that, modulo $O(1/\sqrt{d})$-terms, the two eigenvalues of $\hat{H}$ are $\lambda_1 = 0.8060$ and $\lambda_2 = -0.1198$ (four decimal places are shown). This implies that there exists a $(d - 2)$-dimensional eigensubspace of descent directions, projecting onto the associated eigenspace for $\widehat{H}$. By Cauchy's interlacing theorem it further follows that there must exist positive eigenvalues associated to the standard representation. A similar analysis shows that the extremal properties of the $\Delta(S_{d-1} \times S_1)$-family $\mathfrak{C}^{\mathrm{I}}_{1,i}$ of type I critical points minima, are identical to type II, $i = 1, 2$. In summary,

Theorem 4
- *ΔS_d-symmetric critical points of type I are minima for $k = d$ and non-singular saddles for $k = d + 1$ with a $(d - 1)$-dimensional eigenspace of descent directions.*
- *Critical points of symmetry $\Delta(S_{d-1} \times S_1)$, type I or II, define families of spurious minima for $k = d, d+1$, and non-singular saddles for $k = d+2$, with at least a $(d-2)$-dimensional space of descent directions.*

At first sight, it seems unlikely that one can infer from Theorem 4 and Table 5.2 that there is a lower probability of convergence to minima of isotropy $\Delta(S_{d-1} \times S_1)$ when $k = d + 1$: although some eigenvalues are smaller, some are bigger, and it is not clear why size, as opposed to sign, of the eigenvalue should matter. Empirically, these minima are not seen for $d < 8$ (they are if $k = d$ and $d \geq 6$) and the probability is definitely much lower for small values of $d \geq 8$ [29] (note that type I minima are not detected by SGD under Xavier initialization, see Section 3). Theorem 4 implies that when $k = d + 2$ there are no spurious minima of type II Theorem 4 of symmetry $\Delta(S_{d-1} \times S_1)$

and that the descent directions are tangent to a copy of the standard representation of S_{d-1}. The last point is crucial for understanding the empirical results. The small eigenvalue 0.0230 associated to the standard representation of S_{d-1} when $k = d + 1$ indicates that we are close to a change of stability (bifurcation) of the critical point for gradient descent. Bifurcation of the trivial solution on the standard representation is special and quite exceptional. In our case, the trivial solution will be a sink for gradient descent, when $k = d + 1$, and a source when $k = d + 2$. The change of stability results from the collision of a large number of saddles of *high index* with the sink, followed by the emergence of a source and a large number of saddles with low index (see [3, Section 4] and concluding comments). The high index of the saddles converging to the sink, implies that the basin of attraction for the sink shrinks rapidly as the saddles approach the sink. We refer to [3, Sections 1.1, 4] for more on this phenomenon. As we increase d, families $\mathfrak{C}^{\mathrm{II}}_{p,}$, $p > 1$, of spurious mimima appear which are not annihilated by adding two neurons ($k = d + 2$). This is reflected in the rapid increase in the probability of convergence to a spurious minimum when $k = d + 1$ seen in [29, Tables 1 & 2]. In summary, the empirical results provide strong support for a change of stability associated to the *standard representation* of S_{d-1}. This interpretation of the empirical results in [29] is not obvious and suggests the unique and crucial role of the standard representation of the symmetric group.

By extension, if ΔS_d- and $\Delta(S_{d-1} \times S_1)$-symmetric critical points are assumed representative of other families of symmetric spurious minima, then the chance of convergence to symmetric spurious minima gets smaller when k is increased. We suspect that for this problem, all spurious local minima detected by SGD have symmetry a (large) subgroup of ΔS_d.

Concluding Comments

Much existing literature is concerned with convergence analyses over nonconvex loss landscapes. However, it seems that *static* properties of landscapes, such as the arrangement of spurious minima relative to global minima, have received far less attention, mainly due to lack of suitable analytic tools. In this paper, we demonstrate how the symmetry structure makes possible an analytic study of fundamental phenomena of DL in shallow ReLU networks.

Our spectral results assume the target V has high symmetry but apply also to asymmetric problems. In the cases we discuss, the transition from saddle to minimum, or minimum to saddle, occurs at a non-integer value of d. In particular, at integer values of d critical points are non-singular. It follows, by standard stability results in dynamical systems, that if we fix $k - d = m$, p, and the family $\mathfrak{C}^{\mathrm{X}}_{p,}$, $\mathrm{X} \in \{\mathrm{I}, \mathrm{II}\}$, there is an open neighbourhood $\mathcal{U}$ of V in $M(d, d)$, such that for all $V' \in \mathcal{U}$, the loss function $\mathcal{L}'$ for V' has a non-singular critical point close to each point of the $\Delta S_{k-p} \times S_p$-orbit of $\mathfrak{c}^{\mathrm{X}}_{p,m}$ with the same index as $\mathfrak{c}^{\mathrm{X}}_{p,m}$ (sum of the multiplicities of negative eigenvalues). Critical points and eigenvalues depend continuously on $V' \in \mathcal{U}$: the Hessian spectrum remains extremely skewed when we forcibly break the model symmetry. Of course, we no longer have FPS or isotypic decompositions when V' is asymmetric.

There is the problem of understanding the geometric mechanisms underlying the transition from minimum to saddle. As already indicated, this is closely related to the geometry of the standard representation $\mathfrak{s}_d$ of S_d. For simplicity, assume $d = 2\ell + 1$ is odd (similar results hold if n is even [3]). Gradient vector fields on $\mathfrak{s}_d$ always have a critical point at the origin which is never a non-degenerate saddle but is (generically) either a non-singular minimum or maximum. The transition between minima (sink for the gradient descent, index 2ℓ) and maxima (source, index 0) can be achieved *locally* (that is, as a local deformation of the landscape geometry) by $2^{n-1} - 1$ non-singular saddles of index $\geq \ell$ passing simultaneously through the origin and emerging as $2^{n-1} - 1$ non-singular saddles of index $\leq \ell$ (*op. cit.*, Section 4). *No new minima or maxima are created*—this is rather atypical in bifurcation theory. Analysis of dynamics on fixed point spaces strongly indicates that at non-integer values of d, where the annihilation of minimum occurs, mechanisms of this type apply (*op. cit.*, Section 5). Forced symmetry breaking leads to great complexity near the transition but minimal models of complexity can be given (*op. cit.*, Section 4).

References

[1] S. Akiyama and T. Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. *arXiv preprint arXiv:2106.06251*, 2021.

[2] Y. Arjevani, J. Bruna, M. Field, J. Kileel, M. Trager, and F. Williams. Symmetry breaking in symmetric tensor decomposition. *arXiv preprint arXiv:2103.06234*, 2021.

[3] Y. Arjevani and M. Field. Equivariant bifurcation, quadratic equivariants, and symmetry breaking for the standard representation of s_n. *Nonlinearity, to appear*.

[4] Y. Arjevani and M. Field. On the principle of least symmetry breaking in shallow relu models. *arXiv preprint arXiv:1912.11939*, 2019.

[5] Y. Arjevani and M. Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[6] Y. Arjevani and M. 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, 2021.

[7] Y. Arjevani and M. Field. Symmetry & critical points for a model shallow neural network. *Physica D: Nonlinear Phenomena*, 427:133014, 2021.

[8] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, 2019.

[9] A. Blum and R. L. Rivest. Training a 3-node neural network is np-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[10] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614. JMLR. org, 2017.

[11] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[12] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[13] A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *arXiv preprint arXiv:1602.05897*, 2016.

[14] S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Póczos. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1338–1347, 2018.

[15] W. Fulton and J. Harris. Representation theory, volume 129 of. *Graduate Texts in Mathematics*, 1991.

[16] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[17] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

[18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[19] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *arXiv preprint arXiv:1906.08632*, 2019.

[20] S. Goldt, M. S. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. Generalisation dynamics of online learning in over-parameterised neural networks. *arXiv preprint arXiv:1901.09085*, 2019.

[21] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[22] G. James. The representation theory of the symmetric groups. Springer, 1978.

[23] C. R. Johnson and R. A. Horn. *Matrix analysis*. Cambridge university press Cambridge, 1985.

[24] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[25] Y. Li, T. Ma, and H. R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

[26] Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[27] S. Mei, A. Montanari, and P.-M. 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.

[28] E. Oostwal, M. Straat, and M. Biehl. Hidden unit specialization in layered neural networks: Relu vs. sigmoidal activation. *Physica A: Statistical Mechanics and its Applications*, 564:125517, 2021.

[29] I. Safran and O. Shamir. Spurious local minima are common in two-layer relu neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4430–4438, 2018.

[30] I. M. Safran, G. Yehudai, and O. Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In *Conference on Learning Theory*, pages 3889–3934. PMLR, 2021.

[31] L. Sagun, L. Bottou, and Y. LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[32] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[33] O. Shamir. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 2018.

[34] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[35] Y. Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

[36] Y. Tian. Student specialization in deep rectified networks with finite width and input dimension. In *International Conference on Machine Learning*, pages 9470–9480. PMLR, 2020.

[37] G. Yehudai and O. Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

[38] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]