Analytic Study of Families of Spurious Minima in Two-Layer ReLU Neural 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. We make use of the rich symmetry structure to develop a novel set of tools for studying families of spurious minima. In contrast to existing approaches which operate in limiting regimes, our technique directly addresses the nonconvex loss landscape for finite number of inputs d and neurons k, and provides analytic, rather than heuristic, information. In particular, we derive analytic estimates for the loss at different minima, and prove that, modulo $O(d^{-1/2})$-terms, the Hessian spectrum concentrates near small positive constants, with the exception of $\Theta(d)$ eigenvalues which grow linearly with d. We further show that the Hessian spectrum at global and spurious minima coincide to $O(d^{-1/2})$-order, thus challenging our ability to argue about statistical generalization through local curvature. Lastly, our technique provides the exact *fractional* dimensionality at which families of critical points turn from saddles into spurious minima. This makes possible the study of the creation and the annihilation of spurious minima using powerful tools from equivariant bifurcation theory.

One of the outstanding conundrums of deep learning concerns the ability of simple gradient-based methods to successfully train neural networks despite the nonconvexity of the associated optimization problems. Indeed, generic nonconvex optimization landscapes can exhibit wide and flat basins of attraction around poor local minima which may lead to a complete failure of such methods. The nature by which nonconvex problems associated with neural networks deviate from generic ones is currently not well-understood. In particular, much of the dynamics of gradient-based methods follows from the curvature of the loss landscape around local minima. It is therefore vital to study the local geometry of minima, and in particular to compare the curvature of spurious (i.e., non-global local) minima with that of global minima in order to understand the mysterious mechanism which drives gradient-based methods to reasonably good minima. However, already establishing the very existence of spurious minima seems to be beyond reach of existing analytic tools; let alone rigorously arguing about their height, curvature and structure—the aim of this work.

In this paper, we focus on two-layer ReLU neural networks of the form

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

where $\varphi(z) = \max\{0, z\}$ is the ReLU activation function acting entry-wise, $M(k,d)$ denotes the space of $k \times d$ matrices and $\boldsymbol{w}_i$ denotes the ith row of W. We are primarily interested in characterizing various optimization-related obstructions for local search method, independently of the expressive power of two-layer ReLU networks. Thus, data is understood to be fully realizable. Concretely, we assume that there are d inputs which are drawn from the standard multivariate Gaussian distribution,

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

and are labeled by a *planted* target network. We consider directly optimizing the expected squared loss, resulting in the following highly nonconvex optimization problem:

$$\mathcal{L}(W, \boldsymbol{\alpha}) \coloneqq \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},I_d)}\left[\Big(\sum_{i=1}^{k}\alpha_i\varphi(\langle\boldsymbol{w}_i,\boldsymbol{x}\rangle) - \sum_{i=1}^{d}\beta_i\varphi(\langle\boldsymbol{v}_i,\boldsymbol{x}\rangle)\Big)^2\right], \tag{2}$$

where k denotes the number of hidden neurons, $W \in M(k,d)$ and $\alpha_i \in \mathbb{R}^k$ are the optimization variables, and $\boldsymbol{v}_i$ and β_i are fixed parameters. This setting, in which the data distribution is regulated rather than being allowed to admit worst-case behavior, has drawn a considerable amount of interest in recent years [1, 2, 3, 4, 5, 6, 7, 8, 9], in part due to the growing number of evidences which indicate that any explanation for the empirical success of deep learning (DL) must take into account the intricate interplay between the network architecture, the input distribution and the label distribution (cf. [6, 10, 11] and references therein for hardness results of optimization and learnability under partial sets of assumptions). Moreover, as demonstrated later in the paper, in spite of its apparent simplicity, the nonconvex problem (2) shares a few important characteristics with full-scale neural networks, such as low-dimensional minima and extremely skewed Hessian spectrum.

Learning problem (2) has also been studied in the statistical physics community, starting from the 80' and the 90' [12, 13, 14, 15, 16, 17, 18, 19, 20] under the student-teacher (ST) framework. In this framework, a *student* network is to be adjusted so as to fit the output of a *teacher* network. The ST framework offers a clean venue for analyzing optimization-related aspects of neural network models in the spirit of physical reductionism. The success of DL models in the past decade has reinitiated a surge of interest in this line of work, e.g., [21, 22, 23, 24]. Despite the long tradition in the statistical physics community and the wide effort put nowadays by the machine learning community, the perplexing geometry of problem (2) still seems to be out of reach of existing analytic tools in regimes encountered in practice. In this paper, we present a novel set of symmetry-based tools which allows us, for the first time, to analytically characterize various important aspects of this highly nonconvex landscape, for finite number of inputs and neurons.

Our contributions, in order of appearance, can be stated as follows:

- We show that, empirically, and in a well-defined sense, spurious minima in two-layer ReLU neural networks *breaks* the symmetry of the target weight matrix. Although ReLU networks have been studied for many years, this phenomenon seems to have gone largely unnoticed.
- This phenomenon of symmetry breaking makes possible the derivation of *analytic* expressions for families of spurious minima in the form of (fractional) power series in terms of d and k. Crucially, in contrast to existing approaches which employ various limiting processes, e.g., [21, 22, 23, 24, 25, 26, 27, 28, 29], our method operates in the natural regime where d and k are finite.
- We develop a novel technique which yields an analytic characterization of the Hessian spectrum of spurious minima in terms of d for $d \geq k$. Here, eigenvalues are computed to $O(d^{-1/2})$-order. Our technique further allows us to determine the exact *fractional* value of d at which a family of critical points turns from saddles into spurious minima, and back—a key ingredient in understanding how over-parameterization annihilates spurious minima.
- The access to high-dimensional spectral information allows us to revisit a number of hypotheses in the machine learning literature pertaining to curvature, optimization and generalization, on formal grounds. In particular, we prove that the Hessian spectrum at minima is extremely ill-conditioned and tends to concentrate near small positive constants with the exception of $\Theta(d)$ eigenvalues which grow linearly with d. Although this phenomenon of extremely skewed spectrum has been observed many times [30, 31, 32, 33], to our knowledge, this is the first time it has been established *rigorously* for two-layer ReLU networks. In addition, our analysis shows that quite remarkably the Hessian spectra of spurious and global minima are identical to $O(d^{-1/2})$-terms. Consequently, the inductive bias of stochastic gradient descent (SGD), provably, can not be exclusively explained in terms of local curvature [34, 35, 36, 37, 38, 39, 40].

The results presented in the paper are threefold: identifying a symmetry breaking principle for critical points of two-layer ReLU networks, deriving analytic characterization of families of spurious minima for finite width, and obtaining a detailed spectral analysis of the latter. The next three sections are

organized accordingly, along with a literature survey of related work. The last section is devoted for a high-level description of the symmetry-based technique used in this work. Proofs, formalities and lengthy technical details are deferred to the appendix.

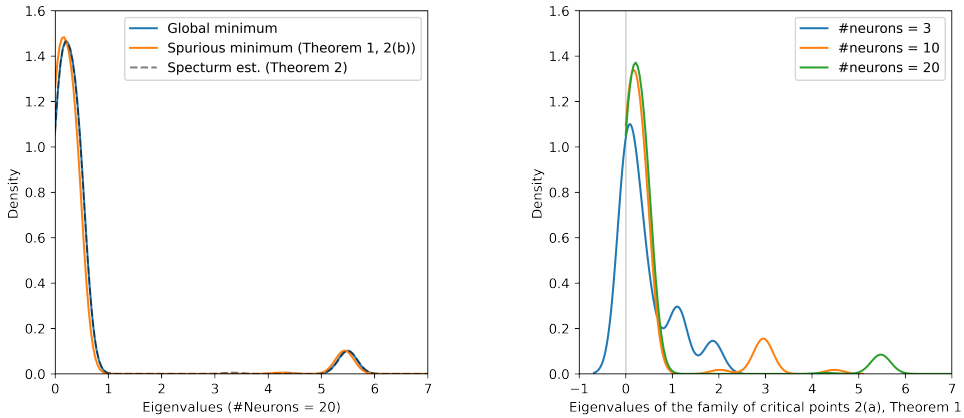


Figure 1: (Left) The symmetry-based technique developed in this work yields an *analytic* characterization of the Hessian spectrum (Theorem 2) to $O(d^{-1/2})$-order (shown in dashed line), which provides a good approximation already for $k = d = 20$ (computed numerically using LinAlg, Python). The analysis further implies that, perhaps surprisingly, the Hessian spectrum of global and spurious minima agree to within $O(d^{-1/2})$-accuracy, as demonstrated above. (Right) the Hessian spectrum is extremely skewed and tends to concentrate near small positive constants with the exception of $\Theta(d)$ eigenvalues which grow linearly with d. Observe that the family of critical points considered here (see Theorem 1, case 2a) turns from saddles at $d = 3$ into spurious minima when $d \geq 10$. Our symmetry-based technique shows that the change of stability occurs at $d \approx 5.71$, and more importantly, indicates that the process can in fact be reversed, namely, spurious minima may turn into saddles, by using over-parameterization.

1 Symmetry breaking in two-layer ReLU neural networks

Optimization problem (2) has a very rich symmetry structure. Indeed, the loss function $\mathcal{L}$ is invariant to left- and right-multiplication of W by permutation matrices (see Section A.1 for a formal proof), i.e., $\mathcal{L}(W, \boldsymbol{\alpha}) = \mathcal{L}(P_\pi W P_\rho^\top, \boldsymbol{\alpha})$ for all $(\pi, \rho) \in S_k \times S_d$, where S_m generally denotes the symmetric group of degree m, and

$$(P_\pi)_{ij} = \begin{cases} 1 & i = \pi(j), \\ 0 & \text{o.w.} \end{cases}.$$

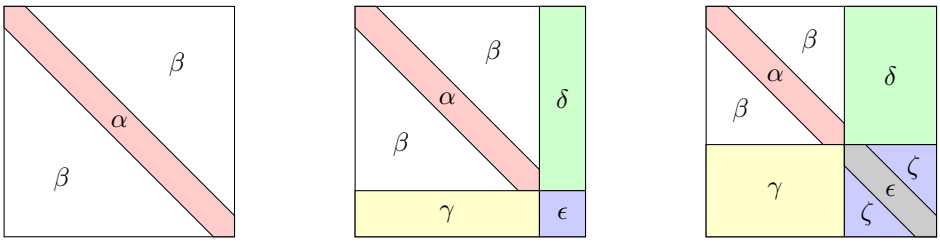
It is therefore natural to ask how the critical points of $\mathcal{L}$ reflect this symmetry. For example, the identity matrix $W = I_d$, one of the global minimizers of $\mathcal{L}$ for the case where $V = I_d$, $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{1}_d$ and $k = d$, is invariant under *simultaneous* left- and right-multiplication by any permutation matrix. Indeed, $P_\pi W P_\pi^\top = P_\pi P_\pi^\top = I_d = W$ for any $\pi \in S_d$. (Modulo group conjugation, this holds for any global minimizer of $\mathcal{L}$. See [41, Proposition 4.14.].) This simple observation is conveniently stated using the concept of the *isotropy* group defined by

$$\mathrm{Iso}(W) \coloneqq \{(\pi, \rho) \mid (\pi, \rho) \in S_k \times S_d,\ P_\pi W P_\rho^\top = W\}, \tag{3}$$

for a given weight matrix W. Thus, we have $\mathrm{Iso}(I_d) = \Delta S_d$, where Δ maps any subgroup $H \subseteq S_d$ to its diagonal counterpart group $\Delta H \coloneqq \{(h, h) \mid h \in H\} \subseteq S_d \times S_d$. Empirically, and somewhat miraculously, spurious minima and saddles of $\mathcal{L}$ tend to be highly symmetric in the formal sense that their isotropy groups are (conjugated to) large subgroups of ΔS_d, the isotropy of the global minima (see Figure 2). This principle of symmetry breaking may be concisely phrased as follows:

Spurious minima *break the symmetry* of global minima.

The principle extends to more general target networks. When the isotropy of the target weight matrix V changes, so does the symmetry of spurious minima of the respective optimization problem. In Section A, we provide a series of experiments which empirically corroborates this symmetry breaking principle.



Isotropy ΔS_5 Isotropy $\Delta S_4 \times \Delta S_1$ Isotropy $\Delta S_3 \times \Delta S_2$

Figure 2: A schematic description of 5×5 matrices with isotropy ΔS_5, $\Delta S_4 \times \Delta S_1$ and $\Delta S_3 \times S_2$, from left to right (borrowed from [42]). $\alpha, \beta, \gamma, \delta, \epsilon$ and ζ are assumed to be 'sufficiently' different.

The principle of symmetry breaking for ReLU networks was first studied in [42, 41], and was later extended to various tensor decomposition problems in [43]. Here, we analyze two-layer ReLU networks where *both* layers are trainable, an architecture which has received a great amount of attention in recent years [1, 2, 3, 4, 5, 6, 7]. We note in passing that, in its broader sense, the principle of symmetry breaking has been observed many times in various scientific fields, e.g., Higgs-Landau theory, equivariant bifurcation theory and replica symmetry-breaking (see, e.g., [44, 45, 46, 47]). Two-layer ReLU networks seems to form a rather unexpected instance of this principle.

One intriguing quality of nonconvex landscapes in which the symmetry breaking principle applies is that local minima lie in fixed low-dimensional spaces (see Section 4). This phenomenon of a hidden low-dimensional structure has been observed in various DL learning problems with real datasets [48, 49], and is widely believed to be an important factor of learnability in nonconvex settings. In the context of this work, the hidden low-dimensionality of spurious minima turns out to be a key ingredient to our analytic study, as we now present.

2 Power series representation of families of spurious minima

Although the problem of fitting neural networks under the ST framework have been studied for more than 30 years, the symmetry breaking principle exhibited by optimization problem (2) seems to have gone largely unnoticed. Early studies adopted tools (at times, heuristic, e.g., [50, 51]) from statistical physics to analyze phase transitions and generalization errors [12, 13, 14, 15]. Later works focused on the dynamics of SGD and studied the progress of the generalization error along the optimization process through a set of carefully-derived ODEs [16, 17, 18, 19, 20]. Following the empirical success of DL in the past decade, this line of work has recently drawn a renewed interest, e.g, [21, 22, 24, 25], which puts past analyses on rigorous grounds and addresses a broader class of architectures and activation functions. The analytic method used in this long line of works operates in the *thermodynamic limit*—where the number of inputs is taken to infinity. Thus, the formal validity of the results to finite width networks is currently limited.

Additional major approaches for addressing problem (2) are based on: mean-field [26], optimal control [27], NTK [28] and compositional kernels [29]. These approaches offer, in essence, convex surrogates which apply in strict parameter regimes (including algorithmic parameters, such as learning rates). Indeed, similarly to the thermodynamic limit approach, the convex surrogates are obtained by limiting processes—this time by taking the number of hidden neurons to infinite. A growing number of works has severely limited, if not invalidated, the explanatory power of these approaches for network widths encountered in practice [52, 53].

In sharp contrast to the approaches discussed above, our technique directly addresses the associated highly nonconvex optimization landscape in the natural regime where the number of inputs and neurons is finite, and provides analytic, rather than heuristic, information. This is obtained by exploiting the presence of symmetry breaking phenomena whereby (fractional) power series representation for families of spurious minima are derived.

Below, we provide analytic expressions for families of minima of different isotropy (see Definition 3 and Figure 2), along with their respective objective value. For brevity, expressions are given to $O(d^{-3/2})$-order. In the appendix, we list additional $O(d^{-5/2})$-order terms which are required for computing the Hessian spectrum. The invariance properties of optimization problem (2) imply that new families of minima can be obtained by permuting the rows and the columns of a given one. The number of distinct new minima generated using such transformations (i.e., the multiplicity) depends on the very structure of the family of minima under consideration, and is stated in Theorem 1. Lastly, for any $\lambda > 0$ and $i \in [d]$, the value of the ReLU network (1) remains fixed under $(\boldsymbol{w}_i, \alpha_i) \mapsto (\lambda \boldsymbol{w}_i, \alpha_i/\lambda)$. This degree-of-freedom is expressed using the slack variables $\lambda_i > 0$, $i \in [d]$, introduced below.

Theorem 1. *Optimization problem (2) with $k = d$, $V = I_d$ and $\boldsymbol{\beta} = \mathbf{1}_d$ possesses the following families of minima for $d \geq 9$:*

1. *Two families of ΔS_d-minima of multiplicity $d!$ with $W(d) = \mathit{Diag}(\lambda_1, \dots, \lambda_d) A_d(a_1, a_2)$ and $\alpha(d) = (\lambda_1^{-1}, \dots, \lambda_d^{-1})$, where $\lambda_i > 0$,*

$$A_d(a_1, a_2) := \begin{pmatrix} a_1 & a_2 & \dots & & a_2 \\ a_2 & a_1 & a_2 & \dots & a_2 \\ & \vdots & & \vdots & \\ a_2 & \dots & & a_2 & a_1 \end{pmatrix} \in M(d, d),$$

   $\mathrm{Diag}(\cdot)$ *maps a given vector to the diagonal entries of a diagonal matrix, and*

   (a) $a_1 = 1$ *and* $a_2 = 0$ *(a global minimizer), in which case*

$$\mathcal{L}(W(d), \alpha(d)) = 0.$$

   (b) $a_1 = -1 + \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = -\frac{1}{\pi} + \frac{1}{2} - \frac{4}{3\pi\sqrt{d}} + \frac{-\frac{1}{2} - \frac{2}{\pi^2} + \frac{3}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right). \tag{4}$$

2. *Two families of $\Delta(S_{d-1} \times S_1)$-minima of multiplicity $d \cdot d!$ with*

$$W(d) = \mathit{Diag}(\lambda_1, \dots, \lambda_d) \left(\begin{array}{c|c} A_{d-1}(a_1, a_2) & a_3 \mathcal{I}_{d-1,1} \\ \hline a_4 \mathcal{I}_{1,d-1} & a_5 \end{array} \right),$$

   and $\alpha(d) = (\lambda_1^{-1}, \dots, \lambda_d^{-1})$, *where* $\lambda_i > 0$ *and*

   (a) $a_1 = 1 + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = O\left(d^{\frac{-3}{2}}\right)$, $a_3 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_4 = \frac{4}{\pi d} + O\left(d^{\frac{-3}{2}}\right)$ *and* $a_5 = -1 + \frac{\frac{8}{\pi^2} + 2 + \frac{8}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = \frac{\frac{1}{2} - \frac{2}{\pi^2}}{d} + O\left(d^{\frac{-3}{2}}\right). \tag{5}$$

   (b) $a_1 = -1 + \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_3 = O\left(d^{\frac{-3}{2}}\right)$, $a_4 = \frac{2 - \frac{4}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right)$ *and* $a_5 = 1 + \frac{8(-1+\pi)}{\pi^2 d} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = -\frac{1}{\pi} + \frac{1}{2} - \frac{4}{3\pi\sqrt{d}} + \frac{-1 - \frac{4}{\pi^2} + \frac{5}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right). \tag{6}$$

3. *Two families of $\Delta(S_{d-2} \times S_2)$-minima of multiplicity $d! \binom{d}{2}$ with*

$$W(d) = \mathit{Diag}(\lambda_1, \dots, \lambda_d) \left(\begin{array}{c|c} A_{d-2}(a_1, a_2) & a_3 \mathcal{I}_{d-2,2} \\ \hline a_4 \mathcal{I}_{2,d-2} & A_2(a_5, a_6) \end{array} \right),$$

   and $\alpha(d) = (\lambda_1^{-1}, \dots, \lambda_d^{-1})$, *where* $\lambda_i > 0$ *and*

(a) $a_1 = 1 + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = 0 + O\left(d^{\frac{-3}{2}}\right)$, $a_3 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_4 = \frac{4}{\pi d} + O\left(d^{\frac{-3}{2}}\right)$, $a_5 = -1 + \frac{\frac{8}{\pi^2}+2+\frac{8}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right)$ *and* $a_6 = \frac{2\left(-12\pi+16+\pi^3+4\pi^2\right)}{\pi^2 d(2+\pi)} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = \frac{-4+\pi^2}{\pi^2 d} + O\left(d^{\frac{-3}{2}}\right). \tag{7}$$

(b) $a_1 = -1 + \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_3 = 0 + O\left(d^{\frac{-3}{2}}\right)$, $a_4 = \frac{2-\frac{4}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_5 = 1 + \frac{8(-1+\pi)}{\pi^2 d} + O\left(d^{\frac{-3}{2}}\right)$ *and* $a_6 = \frac{4\left(-\pi^2-8+6\pi\right)}{\pi^2 d(2+\pi)} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = -\frac{1}{\pi} + \frac{1}{2} - \frac{4}{3\pi\sqrt{d}} + \frac{-\frac{3}{2} - \frac{6}{\pi^2} + \frac{7}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right). \tag{8}$$

4. *One family of* $\Delta(S_{d-3} \times S_3)$*-minima of multiplicity* $d!\binom{d}{3}$ *with*

$$W(d) = \mathit{Diag}(\lambda_1, \ldots, \lambda_d)\left(\begin{array}{c|c} A_{d-3}(a_1, a_2) & a_3\mathcal{I}_{d-3,3} \\ \hline a_4\mathcal{I}_{3,d-3} & A_2(a_5, a_6) \end{array}\right),$$

and $\alpha(d) = (\lambda_1^{-1}, \ldots, \lambda_d^{-1})$, *where* $\lambda_i > 0$ *and* $a_1 = 1 + O\left(d^{\frac{-3}{2}}\right)$, $a_2 = O\left(d^{\frac{-3}{2}}\right)$, $a_3 = \frac{2}{d} + O\left(d^{\frac{-3}{2}}\right)$, $a_4 = \frac{4}{\pi d} + O\left(d^{\frac{-3}{2}}\right)$, $a_5 = -1 + \frac{\frac{8}{\pi^2}+2+\frac{8}{\pi}}{d} + O\left(d^{\frac{-3}{2}}\right)$ *and* $a_6 = \frac{2\left(-12\pi+16+\pi^3+4\pi^2\right)}{\pi^2 d(2+\pi)} + O\left(d^{\frac{-3}{2}}\right)$, *in which case*

$$\mathcal{L}(W(d), \alpha(d)) = \frac{\frac{3}{2} - \frac{6}{\pi^2}}{d} + O\left(d^{\frac{-3}{2}}\right). \tag{9}$$

The idea of the proof of Theorem 1 is given in Section 4.1, and the multiplicity computation, a simple application of the orbit-stabilizer theorem, in Section B. We note that the technique also applies to other choices of target networks, mutatis mutandis. The power series stated in Theorem 1 also represent spurious minima for the under-parameterized $d > k$-case. This is obtained by appropriately padding the entries of the weight matrices by zeros (see [41, section 4.3] for details). The derivation of the respective Hessian spectrum is then an immediate application of [9, section E.1] to the spectral analysis given in Theorem 2 below (for the case where $d = k$). By contrast, studying over-parameterized regime where $d < k$ requires qualitatively different tools and is outside the scope of this work.

The analysis above reveals that not all local minima are alike: while the objective value of some minima decays like $\Theta(1/d)$, in other cases the objective value converges to the positive constant $\frac{1}{2} - \frac{1}{\pi}$. (In particular, except for the global minima case 1a, all minima described above are spurious.) The difference between the two types of behavior seems to lie in the limiting values of the entries of $W(d)$; in the former, the diagonal mainly consists of ones, whereas in the latter mainly minus ones. This is consistent with the fact that all the teacher's diagonal entries are ones.

For reasons which are yet to be understood, empirically, the bias induced by Xavier initialization [54] seems to favor the class of minima for which the objective value decays as $\Theta(1/d)$. This is not to say that under Xavier initialization the expected value upon convergence decreases with d. Indeed, the objective value decays to zero at different rates (see cases 2a, 3a, 4 above). Moreover, the expected objective value also depends on the probability to converge to a given type of minima. (A possible proxy to the latter is the minima multiplicity stated in Theorem 1.)

3 Analytic study of the Hessian spectrum

Once a power series representation has been obtained, it is possible to analytically characterize various important properties of families of minima. Here, we use this representation to compute yet another (fractional) power series, this time, of the Hessian spectrum.

Theorem 2. *Assuming all the weights in the second layer of (1) are normalized to one, the (nonnegative) Hessian spectrum of the families of minima considered in Theorem 1 is*

Eigenvalue	*Multiplicity*
$O(d^{-1/2})$	d
$\frac{1}{4} - \frac{1}{2\pi} + O(d^{-1/2})$	$\frac{(d-1)(d-2)}{2}$
$\frac{1}{2} - \frac{1}{\pi} + O(d^{-1/2})$	$d-1$
$\frac{1}{4} + O(d^{-1/2})$	$d-1$
$\frac{1}{4} + \frac{1}{2\pi} + O(d^{-1/2})$	$\frac{d(d-3)}{2}$
$\frac{d}{4} + \frac{1}{2} + O(d^{-1/2})$	$d-1$
$\frac{d}{4} + \frac{-4+\pi+\pi^2}{2\pi(-4+\pi)} + O(d^{-1/2})$	1
$\frac{d}{\pi} + \frac{-10\pi+8+\pi^2}{2\pi(-4+\pi)} + O(d^{-1/2})$	1

The method we developed for computing the Hessian spectrum builds on [9], but differs in three crucial aspects. First, in our setting the second layer is trainable. The associated so-called *isotypic decomposition* (a set of 'simple' subspaces compatible with the action of row- and column-permutations, which we formally introduced in Section D) must therefore cover the space of $d \times d$ weight matrices, as well as the d weights of the second layer. Secondly, and more significantly, rather then expressing the Hessian entries in terms of $d, W(d)$ and $\boldsymbol{\alpha}(d)$, we directly express the eigenvalues in these terms. This considerably simplifies computations and, in principle, allows one to compute the eigenvalues to any order. (In Theorem 2, eigenvalues are given to $O(d^{-1/2})$-order). Thirdly, and perhaps most importantly, computing the Hessian to high accuracy reveals that for small values of d the families of minima presented in Theorem 1 are in fact saddles. The method we develop in this paper allows us to regard d as a real number and to pinpoint the exact *fractional* dimension at which a given family of critical points turns from being a saddle into a spurious minimum. Observing the process by which spurious minima are created (see also Figure 1), one is led into the dual question: is there a mechanism by which spurious minima may be annihilated (that is, turn into saddles)? We have found that the process of the creation of minima can be reversed in the over-parameterized regime. This provides a strong evidence for the benefits of increasing the number of student neurons from an optimization point of view. As mentioned earlier, a rigorous study of the mechanism by which families of saddles turn into spurious minima and back requires tools from the equivariant bifurcation theory, and is deferred to future work.

Quite remarkably, although the families of minima presented in Theorem 1 differ significantly, their Hessian spectra coincide to $O(d^{-1/2})$-order. Another noticeable feature of the analysis in Theorem 2 is that the Hessian spectrum tends to be extremely skewed and to concentrate near small positive constants with the exception of $\Theta(d)$ eigenvalues which grow linearly with d (see Figure 1). This skewness of the Hessian spectrum has been observed, empirically, many times. Both phenomena above, as well as their consequences for optimization-related aspects, are discussed in detail below.

Positively-skewed Hessian spectral density. Although first reported nearly 30 years ago [30], to our knowledge, this is the first time that this phenomenon of extremely skewed spectral density has been established *rigorously* for two trainable ReLU layers of arbitrarily large dimensionality. Early empirical studies of the Hessian spectrum [30] revealed that local minima tend to be extremely ill-conditioned. This intriguing observation was corroborated and further refined in a series of works [31, 32, 33] which studied how the spectrum evolves along the training process. It was noticed that, upon convergence, the spectral density decomposes into two parts: a bulk of eigenvalues concentrated near small positive constants, and a small set of positive outliers located away from zero.

Due to the high computational cost of an exact computation of the Hessian spectrum (cubic in the problem parameters k and d), this phenomenon of extremely skewed spectral densities has only been confirmed for small-scale networks. Other methods for extracting second-order information in large-scale problems roughly fall into two general categories. The first class of methods approximate the Hessian spectral density by employing various numerical estimation techniques, most notably stochastic Lanczos method (e.g., [55, 56]). These methods have provided various numerical evidences which indicate that a similar skewed spectrum phenomenon also occurs in full-scale modern neural networks. The second class of techniques builds on tools from random matrix theory. The latter approach yields an exact computation of the limiting spectral distribution (where the number of

neurons is taken to infinity), assuming the inputs, as well as the model weights are drawn at random [57, 58, 59]. In contrast, our method gives an analytic description of the spectral density for any number of neurons (granted $d \geq 9$), and at critical points rather than randomly drawn weight matrices.

The flat minima conjecture and implicit bias. It has long been debated whether some notion of local curvature can be used to explain the remarkable generalization capabilities of modern neural networks [34, 35, 36, 37, 38, 39, 40]. One intriguing hypothesis suggests that minima with wider basins of attraction tend to generalize better. The suggested intuitive explanation is that flat minima promote statistical and numerical stability; together with low empirical loss, these ingredients are widely used to achieve good generalization, cf. [60]. Perhaps surprisingly, the analysis in Theorem 2 shows that the spectra of global minima and spurious minima agree to $O(d^{-1/2})$-order. Consequently, in our setting, local second-order curvature *cannot* be used to separate global from spurious minima. This rules out any notions of 'flatness' which rely exclusively on the Hessian spectrum. Of course, other metrics of a 'wideness of basins' may well be applied.

4 A symmetry-based analytic framework

In the sequel, we present the main ingredients of the symmetry-based technique developed in this work. The technique builds on, and significantly extends various components from [9, 42, 41]. Elementary concepts from group theory are assumed known. To ease exposition, we illustrate with reference to the case where $k = d$ and the second layer is kept fixed with entries equal ones. We denote the resulting nonconvex function by $f(W) := \mathcal{L}(W, \mathbf{1})$.

4.1 Power series representation of minima

The first step in deriving a power series representation is to restrict the nonconvex function under consideration to *fixed point subspaces* (see [61] for a more complete account) that consists of weight matrices which remain fixed under row- and column-permutations in a certain subgroup of $S_d \times S_d$. Consider, for concreteness, the space of all matrices which are invariant to permutations in ΔS_d, i.e.,

$$\mathcal{W}_d := \{W \in M(d,d) \mid W = P_\pi W P_\pi^\top \text{ for all } (\pi,\pi) \in \Delta S_d\}. \tag{10}$$

It is easy to verify that $\mathcal{W}_d$ is in fact a 2-dimensional subspace of the form $\mathcal{W}_d = \{(a_1, a_2) \mid a_1 I_d + a_2(\mathbf{1}_d \mathbf{1}_d^\top - I)\}$, and that if $W \in \mathcal{W}_d$, then $\nabla f(W) \in \mathcal{W}_d$ (see [42, Proposition 3]). Thus, one may regard f as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$. Next, we make the dependence of f on d explicit by defining $F : \mathbb{R}^3 \to \mathbb{R}^2 : \ (a_1, a_2, d) \mapsto \nabla f(a_1 I_d + a_2(\mathbf{1}_d \mathbf{1}_d^\top - I_d))$. Note that although d is a natural number, the following explicit expressions of the entries of $F(a_1, a_2, d)$ allows us to regard d as a real variable:

$$\begin{pmatrix} \frac{a_1 d}{2} - \frac{a_1 d \sin\left(\beta_{(1)}^{(1)}\right)}{2\pi\nu_{(1)}} + \frac{a_1\left(d^2-d\right)\sin\left(\alpha_{(1)}^{(2)}\right)}{2\pi} - \frac{a_1\left(d^2-d\right)\sin\left(\beta_{(1)}^{(2)}\right)}{2\pi\nu_{(1)}} - \frac{a_2\alpha_{(1)}^{(2)}\left(d^2-d\right)}{2\pi} + \frac{a_2\left(d^2-d\right)}{2} + \frac{\beta_{(1)}^{(1)} d}{2\pi} - \frac{d}{2} \\ -\frac{a_1\alpha_{(1)}^{(2)}\left(d^2-d\right)}{2\pi} + \frac{a_1\left(d^2-d\right)}{2} - \frac{a_2\alpha_{(1)}^{(2)}\left(d^3-3d^2+2d\right)}{2\pi} + \frac{a_2\left(d^3-2d^2+d\right)\sin\left(\alpha_{(1)}^{(2)}\right)}{2\pi} + \frac{a_2\left(d^3-2d^2+d\right)}{2} - \frac{a_2\left(d^2-d\right)\sin\left(\beta_{(1)}^{(1)}\right)}{2\pi\nu_{(1)}} - \frac{a_2\left(d^3-2d^2+d\right)\sin\left(\beta_{(1)}^{(2)}\right)}{2\pi\nu_{(1)}} + \frac{\beta_{(1)}^{(2)}\left(d^2-d\right)}{2\pi} - \frac{d^2}{2} + \frac{d}{2} \end{pmatrix}$$

where, here and below, $\alpha_{(i)}^{(j)}$ (resp. $\beta_{(i)}^{(j)}$) denotes the angles between the ith row of W and the jth of W (resp. V). Using the implicit function theorem one can establish the existence (and the uniqueness) of a path of critical points $(a_1(d), a_2(d))$. (These formalities are covered in [41].) This gives a formal meaning for examining a given family of minima at a *fractional* dimensionality.

We are left with the following two main issues. First, the implicit function theorem does not yield an explicit form of $a_1(d)$ and $a_2(d)$. This issue is addressed using a *real analytic* version of the implicit function theorem (and computing the coefficients of the power series of $a_1(d)$ and $a_2(d)$ to the desired order). Secondly, the implicit function theorem requires initial values for a_1 and a_2 which solve $F(a_1, a_2, d) = 0$. To handle this issue, we form power series in $1/d$ and, loosely speaking, develop them at $d = \infty$. The limiting entries of $W(d)$ take a simple form which can be used as initial values to the implicit function theorem. In fact, due to the dependence of the ∇f on the angels between rows of W and V, the power series of $a_1(d)$ and $a_2(d)$ are expressed in terms of $1/\sqrt{d}$. Theorem 1 is established by following the recipe described here for various families of critical points of different isotropy. The gradient expressions involved in the process are lengthy and are therefore relegated to Section E.

4.2 Computing the Hessian spectrum

Our next goal is to derive an analytic characterization of the Hessian spectrum. The invariance properties of $\mathcal{L}$ imply that although the dimension of the Hessian, $k(d+1) \times k(d+1)$, depends on k and d, the number of distinct eigenvalues remains fixed. This is an immediate consequence of the isotypic decomposition (see Section D). A formal introduction of the isotypic decomposition requires some familiarity with group actions, and is therefore deferred to Section D.2. Here, we shall focus on one particularly simple case.

The isotypic decomposition associated with ΔS_d, the isotropy of I_d, implies that for any d,

$$\overline{\mathfrak{Y}_d} := \mathrm{vec}\left(\begin{bmatrix} 0 & d-3 & 3-d & \cdots & 0 & 0 & 0 \\ d-3 & 0 & 0 & \cdots & -1 & -1 & -1 \\ 3-d & 0 & 0 & \cdots & 1 & 1 & 1 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \end{bmatrix}\right),$$

is an eigenvector of $\nabla^2 f(I_d)$, where vec denotes the linear transformation which stacks the rows of a given matrix on top of one another as a column vector. Thus, the eigenvalue corresponding to $\overline{\mathfrak{Y}_d}$ equals $(\nabla^2 f(I_d)\overline{\mathfrak{Y}_d})_2/(d-3)$, which takes the following explicit form

$$\frac{a_1^2}{2\pi\nu_{(1)}^2\nu_{(1)}^{(2)}} + \frac{a_1^2 \sin\left(\alpha_{(1)}^{(2)}\right)}{2\pi\nu_{(1)}^2\left(\nu_{(1)}^{(2)}\right)^2} - \frac{a_1 a_2}{\pi\nu_{(1)}^2\nu_{(1)}^{(2)}} - \frac{a_1 a_2 \sin\left(\alpha_{(1)}^{(2)}\right)}{\pi\nu_{(1)}^2\left(\nu_{(1)}^{(2)}\right)^2} + \frac{a_2^2}{2\pi\nu_{(1)}^2\nu_{(1)}^{(2)}} + \frac{a_2^2 \sin\left(\alpha_{(1)}^{(2)}\right)}{2\pi\nu_{(1)}^2\left(\nu_{(1)}^{(2)}\right)^2}$$
$$+ \frac{\alpha_{(1)}^{(2)}}{2\pi} + \frac{(d-1)\sin\left(\alpha_{(1)}^{(2)}\right)}{2\pi} - \frac{(d-1)\sin\left(\beta_{(1)}^{(2)}\right)}{2\pi\nu_{(1)}} - \frac{\sin\left(\beta_{(1)}^{(1)}\right)}{2\pi\nu_{(1)}} - \frac{\sin\left(\beta_{(1)}^{(2)}\right)}{2\pi\left(\mu_{(1)}^{(2)}\right)^2\nu_{(1)}},$$

where $\nu_{(i)}$ denotes the norm of the ith row of W, and $\nu_{(i)}^{(j)}$ (resp. $\mu_{(i)}^{(j)}$) denotes $\sin\arccos(\alpha_{(i)}^{(j)})$ (resp. $\sin\arccos(\beta_{(i)}^{(j)})$). To complete the derivation of the eigenvalue $\overline{\mathfrak{Y}_d}$ for families of isotropy ΔS_d, e.g., cases 1(a) and 1(b) in Theorem 1, one plugs-in the power series representation of $a_1(d)$ and $a_2(d)$ into the expression above, and forms a new power series to get an estimate for $\overline{\mathfrak{Y}_d}$. The same procedure is used for all families of minima presented in Theorem 1, and is described in detail in Section 4.2.

5 Conclusion

When present, the symmetry breaking principle (spurious minima break the symmetry of global minima) is a powerful tool of reducing the complexity of nonconvex problems in a nonlinear way. In this work, this principle allows us to characterize important properties of families of spurious to a precision which has not been possible before. Our analysis applies directly to ReLU networks with finite inputs and neurons rather than limiting proxies of varying qualities. The extent to which the symmetry breaking principle applies, at this stage of the theory, is not clear. However, it does seem to hold in a broader class of interesting nonconvex problems, e.g., [43].

The families of spurious minima studied in this paper are illustrative examples in which an exceptional proximity to the objective value and the hessian spectrum is allowed. One intriguing outcome is that spurious and global minima are locally identical (to $O(d^{-1/2})$-order). Considering the fact that the multiplicity local minima for $k = d$ is essentially exponential in d, it is rather unlikely for SGD to be able to escape spurious minima when d grows (see [8, Table 1]). Why is it then that SGD is generally able to find good minima for learning tasks of fitting ReLU neural networks? Preliminary experiments we conducted indicate that *continuously* increasing k and leaving d fixed turns families of spurious minima turn into saddles (the objective value seems to remain constant). We believe that this is observation is instrumental in explaining the empirical success of SGD for DL problem. As mentioned earlier, a thorough study of this phenomenon requires tools from equivariant bifurcation theory and is postponed to future work.

References

[1] Qiuyi Zhang, Rina Panigrahy, Sushant Sachdeva, and Ali Rahimi. Electron-proton dynamics in deep learning. *arXiv preprint arXiv:1702.00458*, pages 1–31, 2017.

[2] Simon S. Du, Jason D. Lee, Yuandong Tian, Aarti Singh, and Barnabás 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.

[3] Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks:(almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017.

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

[5] Yuandong 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.

[6] Alon Brutzkus and Amir 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.

[7] Rong Ge, Jason D. Lee, and Tengyu 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.

[8] Itay Safran and Ohad 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.

[9] Yossi Arjevani and Michael Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. *arXiv preprint arXiv:2008.01805*, 2020.

[10] Avrim L Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992.

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

[12] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

[13] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[14] W Kinzel and Pal Rujan. Improving a network generalization ability by selecting examples. *EPL (Europhysics Letters)*, 13(5):473, 1990.

[15] Timothy LH Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[16] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[17] Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and general*, 28(3):643, 1995.

[18] David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

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

[20] Peter Riegler and Michael Biehl. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20):L507, 1995.

[21] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka 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.

[22] 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. *arXiv preprint arXiv:1906.08632*, 2019.

[23] Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. *arXiv preprint arXiv:2006.15459*, 2020.

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

[25] Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Generalisation dynamics of online learning in over-parameterised neural networks. *arXiv preprint arXiv:1901.09085*, 2019.

[26] 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.

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

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[29] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *arXiv preprint arXiv:1602.05897*, 2016.

[30] Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nimes*, 91 (8):12, 1991.

[31] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[32] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

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

[34] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[35] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[36] Stanis*l*aw Jastrz*k*ebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

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

[38] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. In *Advances in Neural Information Processing Systems*, pages 4949–4959, 2018.

[39] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

[40] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.

[41] Yossi Arjevani and Michael Field. Symmetry & critical points for a model shallow neural network. *CoRR*, abs/2003.10576, 2020.

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

[43] Yossi Arjevani, Joan Bruna, Michael Field, Joe Kileel, Matthew Trager, and Francis Williams. Symmetry breaking in symmetric tensor decomposition. *arXiv preprint arXiv:2103.06234*, 2021.

[44] L Michel. Minima of higgs-landau polynomials. Technical report, 1979.

[45] Martin Golubitsky. The benard problem, symmetry and the lattice of isotropy subgroups. *Bifurcation Theory, Mechanics and Physics. CP Boner et al., eds.(Reidel, Dordrecht, 1983)*, pages 225–256, 1983.

[46] M. J. Field and R. W. Richardson. Symmetry breaking and the maximal isotropy subgroup conjecture for reflection groups. *Archive for Rational Mechanics and Analysis*, 105(1):61–94, Mar 1989. ISSN 1432-0673. doi: 10.1007/BF00251598.

[47] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 59–68, 2015. doi: 10.1145/2746539.2746619.

[48] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[49] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[50] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[51] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

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

[53] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

[54] 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*, pages 249–256, 2010.

[55] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 2019.

[56] Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

[57] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[58] Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, pages 5410–5419, 2018.

[59] Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[60] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[61] Michael J. Field. *Dynamics and symmetry*, volume 3 of *ICP Advanced Texts in Mathematics*. Imperial College Press, London, 2007. ISBN 978-1-86094-828-2; 1-86094-828-6. doi: 10.1142/9781860948541.

[62] Charles Benedict Thomas. *Representations of finite and Lie groups*. World Scientific, 2004.

[63] William Fulton and Joe Harris. Representation theory, volume 129 of. *Graduate Texts in Mathematics*, 1991.

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)? [Yes] , The Hessian spectra shown in Figure 1 are computed using LinAlg, a linear algebra package of Python. The method by which numerical results displayed in Section A.2 are obtained is described in the beginning of the section.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (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]