Annihilation of Spurious Minima in Two-Layer ReLU Networks

Yossi Arjevani
The Hebrew University
yossi.arjevani@gmail.com

Michael Field
UC Santa Barbara
Mike.field@gmail.com

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 local) 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. 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 [1, 2, 3, 4, 5, 6, 7]. The two-layer networks considered were typically of the form:

$$f(\boldsymbol{x}; W, \boldsymbol{\alpha}) \coloneqq \boldsymbol{\alpha}^\top \varphi(W\boldsymbol{x}), \quad W \in M(k,d),\ \boldsymbol{\alpha} \in \mathbb{R}^k, \tag{1}$$

where $\varphi(z) \coloneqq \max\{0, z\}$ is the ReLU function acting entrywise and $M(k,d)$ denotes the space of $k \times d$ matrices. 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 distribution-free approaches operating in complete generality [8, 9, 10]. For the squared loss, the resulting, highly nonconvex, expected loss is

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

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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The choice of the d-variate normal Gaussian distribution for the input distribution has drawn a considerable interest, e.g., [11, 12, 13, 14, 15, 16, 17]. Empirically, it has been observed that as the number of neurons k increases, the loss of models obtained using stochastic gradient descent (SGD) under Xavier initialization [18] decreases (see [9, 19]). 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 employ techniques from representation theory and algebraic geometry to study how the loss landscape is transformed when the number of neurons is increased. A key feature of our approach is the use of *Puiseux series* in d and path based techniques. For example, we may define the spectrum of the Hessian as Puiseux series 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 mechanism whereby spurious minima are annihilated when the number of neurons k is increased. Our contributions can be summarized by

Theorem 1 *(Informal) We describe several infinite families of spurious minima, partly characterized by their symmetry, and prove that:*
• *Adding rather few neurons can transform symmetric spurious minima into saddles.*
• *The Hessian spectrum remains extremely skewed with $\Theta(d)$ eigenvalues growing linearly with d, and $\Theta(kd)$ eigenvalues being $\Theta(1)$.*
• *Increasing the number of neurons 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.*

The formation of new decent directions allows gradient-based methods to escape spurious minima that exist in the original loss landscape ($k = d$), and so detect models of reduced loss. To indicate the subtly of the results given in Theorem 1, consider a family of spurious minima with symmetry $\Delta(S_{d-1} \times S_1)$ (the type II family described in Section 3). Assume $k = d$. Spurious 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. If instead we consider the unique family of spurious minima with isotropy ΔS_d (see Section 3), we find that adding just one neuron annihilates these spurious minima. This occurs through the appearance of a multiplicity $(d-1)$-eigenvalue associated to the standard representation of S_d on $\mathbb{R}^{d-1}$ (all other eigenvalues remain positive). More precisely, if $k = d$ there are 3 strictly positive Hessian eigenvalues associated to the standard representation of S_d, each with multiplicity $d-1$. Modulo $O(d^{-\frac{1}{2}})$ terms, they are $\frac{1}{4} - \frac{1}{2\pi}$, $\frac{1}{4}$, and $\frac{d}{4} + \frac{1}{4}$. After the addition of one neuron ($k = d+1$), there are 4 Hessian eigenvalues associated to the standard representation of S_d and, modulo $O(d^{-\frac{1}{2}})$ terms, these satisfy

$$\frac{d}{4} + \frac{1}{2}, \frac{1}{4} - \frac{1}{2\pi}, \frac{-1+\sqrt{5}}{4\pi} + \frac{1}{4} > 0 > \frac{-1-\sqrt{5}}{4\pi} + \frac{1}{4}.$$

This example highlights the special role that the standard representation plays in the annihilation of spurious minima (see Section 5 and the concluding remarks). The sharp estimates of the Hessian spectrum further demonstrate how symmetry breaking enables a complete characterization of the dynamics of gradient-based methods, locally, in the vicinity of symmetric critical points. The dependence of such methods on stability of critical points therefore indicates that attempts for a global theory should be preceded by a good description of the mechanism by which spurious minima transform into saddles—the aim of this work.

Next, we relate our results to the existing literature.

Annihilation of spurious minima on account of over-parameterization. Existing methods for the analysis of optimization problem (2) include: mean-field [4], optimal transport [2], NTK [20, 21, 22] and the thermodynamic limit [5, 16, 23, 24]. 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 limited the explanatory power of such approaches [25, 26]. Approaches for addressing the loss landscapes in finite parameter regimes exist and include [6] which obtains several generalities on critical points, and [7] which studies conditions under which a single neuron can be added in contrived way so as to

turn a given minimum in $M(d,d)$ into a saddle in $M(d+1,d)$. However, as shown in the present work, families of minima in $M(d+1,d)$ exist. Indeed, families of minima of lesser symmetry are shown to require at least two additional neurons before turning into saddles.

Symmetry breaking in nonconvex loss landscapes. It has been recently found that the symmetry of spurious minima in optimization problem (2) *break the symmetry* of global minima [27] (see Section 2 for a formal exposition). A similar phenomenon has been observed for tensor decomposition problems in [28] and later studied [29]. The present work builds on methods developed in a line of work concerning this phenomenon of symmetry breaking. In [30], path based techniques are introduced to construct infinite families of critical points represented by Puiseux series in d^{-1}. In [31], results from the representation theory of the symmetric group are used, together with the Puiseux series result, to obtain precise analytic estimates on the Hessian spectrum. In [32], it is shown that certain families of saddles transform into spurious minima at a fractional dimensionality. Moreover, the spectra of these families of spurious minima is shown to be identical to that of global minima to $O(d^{-\frac{1}{2}})$-order. In [33], generic S_d-equivariant steady-state bifurcation is studied, emphasizing the complex geometry of the exterior square and the standard representations along which the minima studied in this work are created and annihilated.

The results above assume that the number of neurons is less or equal the number of inputs: $k \leq d$. The present work concerns the over-parameterized case $k > d$ (additional terms used in related contexts are *over-specified*, e.g., [34] and *over-realized*, e.g., [6]). The study of the technically more demanding case of over-parameterization requires new methods and ideas, which we describe below.

- We develop a method which allows the expression of eigenvalues in terms of the gradient entries. The approach is used to evaluate eigenvalues of $\Theta(d^2)$-multiplicity analytically, and reveals hidden algebraic relations between *criticality* and *curvature*.
- Eigenvalues of $O(d)$-multiplicity are computed numerically using estimates of the Puiseux series coefficients. The estimates are obtained through numerical methods used either directly for a system of equations corresponding to different orders of the Puiseux series terms, or for a reduced system of equations obtained by exploiting the geometric structure of the problem.
- Lastly, Cauchy's interlacing theorem is used to reduce the complexity of the computation of descent directions afforded by over-parameterization, and yields a tight characterization of linear subspaces along which spurious minima transform into saddles.

The methods are illustrated for eight 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 considered in previous works, and are a topic of current research.

Organization of the paper. The proof of Theorem 1 is in four parts: symmetry & the loss function, families of symmetric minima, Hessian spectrum, and a change of stability under 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.

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. Thus S_d acts naturally on $[d]$ (as permutations) and orthogonally on $\mathbb{R}^d$ (as permutation matrices).

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 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 $M(k,d)$ plays a central role in the study of invariance properties of $\mathcal{L}$. 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 be defined in terms of permutation matrices. If $k \geq d$, the *diagonal subgroup* ΔS_d of $S_k \times S_d$ is defined 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 $M(d, d)$. Let $V \in M(k, d)$ denote the matrix determined by the identity matrix $I_d \in M(d, d)$. By the orthogonal symmetry of the Gaussian distribution, our results will hold for any V determined by a matrix in $\mathrm{O}(d)$. We shall not consider training processes but rather concrete instances of families of critical points, and so throughout weights are assumed fixed. The weights of the second layer of critical points studied in the present work consists of positive weights and so, by the positive homogeneity of the ReLU activation, there is no loss of generality in assuming that the second layer of weights is set to ones, i.e., $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathcal{I}_{k,1}$, $\mathcal{I}_{i,j}$ being the $i \times j$-matrix with all entries equal to 1.

Optimization problem (2) has a rich symmetry structure; for our choice of $V, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $\mathcal{L}$ is $S_k \times S_d$-invariant [27]. It is natural to ask how the critical points of $\mathcal{L}$ reflect this symmetry. Given $W \in M(k, d)$, the largest subgroup of $S_k \times S_d$ fixing W is called the *isotropy* subgroup of W. The isotropy group quantifies the *symmetry* of W. If $k = d$, V has isotropy group ΔS_d and every global minimizer of $\mathcal{L}$ lies on the $S_d \times S_d$-orbit of V [7, 30]. Empirically, if $k \geq d$, non-degenerate (i.e., no zero Hessian eigenvalues) spurious minima of $\mathcal{L}$ tend to be highly symmetric in that their isotropy groups are conjugate to large subgroups of ΔS_d. Indeed, we suspect that for our choice of V, the isotropy of non-degenerate spurious minima is always non-trivial and conjugate to a subgroup of ΔS_d.

If $k > d$, the isotropy of V is *not* a subgroup of ΔS_d— it contains the subgroup $I_d \times S_{k-d}$ of row permutations. Perhaps surprisingly, $\mathcal{L}$ is more regular at critical points of spurious minima than at points giving the global minimum: if $k = d$, analyticity of $\mathcal{L}$ at $W = V$ fails. If $k > d$, V has a row of zeros and $\mathcal{L}$ is not differentiable at $W = V$ [1] (see also [7]). It may easily 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 (see Section A.7). Necessarily, Λ contains the $S_k \times S_d$ orbit of V. At boundary points of Λ, $\mathcal{L}$ is not differentiable. The Hessian, defined on the interior of the simplex, is *always* singular. Here our focus will always be on families of spurious minima with non-degenerate critical points.

3 Families of minima: structure and basic properties

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. For a systematic study of the optimization landscape of $\mathcal{L}$, we need to categorize minima and understand their distinctive analytic properties. Both isotropy and the notion of a *regular family* play an important role. Throughout we assume $k = d + m$, for all $d \geq d_0$, where $m \geq 0$ is an integer constant.

If G is a subgroup of ΔS_d, let $M(k, d)^G = \{W \in M(k, d) \mid gW = W, \forall g \in G\}$ denote the *fixed point space* for the action of G on $M(k, d)$. Every $S_d \times S_k$-equivariant vector field on $M(k, d)$ is tangent to $M(k, d)^G$ and so $\nabla\mathcal{L}$ is tangent to $M(k, d)^G$. Hence $\mathfrak{c} \in M(k, d)^G$ is a critical point of $\mathcal{L}|M(k, d)^G$ iff $\mathfrak{c}$ is a critical point of $\mathcal{L}$. The inclusion $i_d : [d] \to [d + 1]$ induces a natural inclusion $i_d : S_d \subset S_{d+1}$, where $i_d(S_d)$ fixes $d + 1$. More generally, given a positive integer p and $d_0 > p$, we have a sequence of inclusions $i_{d,p} : S_{d-p} \times S_p \to S_{d+1-p} \times S_p$, $d \geq d_0$. Identifying S_d with ΔS_d, a sequence $(G_d)_{d \geq d_0}$ of subgroups of ΔS_d is *natural* if for some positive integer $p < d_0$, (a) $i_{d,p}(G_d) \subset G_{d+1}$, and (b) $\dim(M(k, d)^{G_d})$ is independent of $d \geq d_0$ (assume $k \geq d$).

Example 1 *Set $G_d = \Delta(S_{d-p} \times S_p)$ and m be a positive integer. If $p = 0$, then $G_d = \Delta S_d$ and dim$(M(d+m, d)^{G_d}) = 2+m$, $d \geq d_0 = 2$; if $p = 1$, then dim$(M(d+m, d)^{\Delta(S_{d-1} \times S_1)}) = 5+2m$, $d \geq 3$; if $p \geq 2$, then dim$(M(d + m, d)^{\Delta(S_{d-p} \times S_p)}) = 6 + 2m$, $d \geq 4$.*

If $(G_d)_{d \geq d_0}$ is natural, we often identify $M(k, d)^{G_d}$ with $\mathbb{R}^N$, $d \geq d_0$, where $\dim(M(k, d)^{G_d}) = N$. We define linear isomorphisms $\Xi : \mathbb{R}^N \to M(k, d)^{G_d}$ for the families of Example 1. Let $I_i^\star = \mathcal{I}_{i,i} - I_i,\ i \in \mathbb{N}$. The matrix $\Xi(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \mathbb{R}^N$, is expressed as block diagonal matrix $[B_{ij}]$, where each $B_{i,j}$ is a linear combination in the coordinates of $\boldsymbol{\xi}$ of the matrices $\mathcal{I}_{i,j}$, I_i and $I_i^\star$. For example, if $p = 1$, $m \geq 1$, then $N = 5 + 2m$ and $\Xi(\xi_1, \cdots, \xi_N)$ is the $(2 + m) \times 2$-block matrix

$[B_{ij}] \in M(d+m,d)^{\Delta(S_{d-1}\times S_1)}$ defined by

$$B_{11} = \xi_1 I_{d-1} + \xi_2 I^{\star}_{d-1},\ B_{12} = \xi_3 \mathcal{I}_{d-1,1},$$
$$B_{i1} = \xi_{2i}\mathcal{I}_{1,d-1},\ B_{i2} = \xi_{2i+1}\mathcal{I}_{1,1},\ 2 \le i \le 2+m.$$

Similar expressions hold for the other families. In practice, we restrict the vector field $\nabla\mathcal{L}$ to $M(k,d)^{G_d}$ and then pull back this vector field using Ξ to a vector field F_d on $\mathbb{R}^N$. The Jacobian of F_d is then equal to the Hessian of $\nabla(\mathcal{L}|M(k,d)^{G_d})$. Observe that F_d does not depend on a choice of inner product on $\mathbb{R}^N$; indeed, we take the standard Euclidean inner product on $\mathbb{R}^N$ ($\Xi : \mathbb{R}^N \to M(k,d)^G$ is *not* an isometry). Since $F_d(\boldsymbol{\xi}) = \Xi^{-1}\nabla\mathcal{L}(\Xi(\boldsymbol{\xi}))$, we may read off the components of $F_d(\boldsymbol{\xi})$ directly from the corresponding matrix entries of $\nabla\mathcal{L}$. We find that F_d is a continuous family of vector fields on $\mathbb{R}^N$ in the *real* parameter d. Obviously, no such statement can hold on $M(k,d)$ as the dimension of $M(k,d)$ depends on d. Moreover, the vector fields F_d will be real analytic outside of a thin (semianalytic) subset of $\mathbb{R}^N$. Indeed, F_d is subanalytic [35] but the real analyticity statement is easily proved directly and suffices for our applications. All of this allows us to "connect" the critical points $\mathfrak{c}(d) \in M(k,d)^{G_d}$, $d \ge d_0$, by curves in $\mathbb{R}^N$ and develop a path-based approach to our problem.

Definition 1 *(Notation & Assumptions as above.) A family $\{\mathfrak{C}(d) = \Xi(\mathfrak{c}(d)) \mid d \ge d_0\}$ of critical points of $\mathcal{L}$ with isotropy $G_d \subset \Delta S_d$ is* weakly regular *if $(G_d)_{d\ge d_0}$ is natural and for $d \ge d_0$*

- *(a) There is a continuous curve $\gamma_d : [0,1] \to \mathbb{R}^N$ of critical points of F_d joining $\mathfrak{c}(d)$ to $\mathfrak{c}(d+1)$.*
- *(b) $\mathcal{L}$ is real analytic at $\Xi(\mathfrak{c}(d)) \in M(k,d)$.*
- *(c) F_d is real analytic on a neighbourhood of $\gamma_d([0,1]) \subset \mathbb{R}^N$.*
- *(d) The Jacobian of F_d along γ_d is non-singular.*

If $\lim_{d\to\infty} \mathfrak{c}(d) \doteq \mathfrak{c}_\infty \in \mathbb{R}^N$ exists and is bounded (Euclidean norm on $\mathbb{R}^N$), the family is regular.

Using (c,d), and the real analytic implicit function theorem, γ_d is real analytic. The family defined by $\mathfrak{c}(d) = V$ is not weakly regular as (b) fails even if $k = d$.

It may be shown, using results on subanalytic sets and the Curve Selection Lemma [36], that every regular family has a fractional power series (FPS) representation. That is, with the notation and assumptions of Definition 1, there exist $d_1 \ge d_0$ and a minimal $b \in \mathbb{N}$ such that each component $\mathfrak{c}_i(d)$ of $\mathfrak{c}(d) \in \mathbb{R}^N$ is given by the convergent power series

$$\mathfrak{c}(d)_i = \sum_{j=0}^{\infty} c_{i,j} d^{-\frac{j}{b}},\ i \in [N]. \tag{4}$$

Under the assumption of weak regularity, there may be Puiseux series representations [30, Exam. 4.13] and $\mathfrak{c}_\infty$ lies in the one point compactification of $\mathbb{R}^N$. In practice, rather than use the general result, we prove directly that a family has an FPS representation. Verifying regularity for sufficiently large d_0, is usually straightforward or trivial. We refer to Section A for examples of construction of FPS representations when $k > d$.

The FPS representation for families of critical points is important both theoretically, and computationally and yields Puiseux series representations of the objective value and Hessian spectrum. It was shown in [30, Section 8] that for $k = d$ several (regular) families of critical points with isotropy $G_d = \Delta(S_{d-p} \times S_p)$, $p \in \{0,1\}$, had FPS representations in $d^{-\frac{1}{2}}$ (so $b = 2$). Each coordinate of $\mathfrak{c}_\infty = \lim_{d\to\infty} \mathfrak{c}(d) \in \mathbb{R}^N$ was either ± 1 or zero. For examples of FPS representations with $k = d$ and $G_d = \Delta(S_{k-p} \times S_p)$, $p \in \{2,3\}$, see [32]. Explicit construction of the coefficients in these FPS examples is relatively straightforward and algebraic formulae can be given for low order terms. When $k > d$, analysis is harder. It is not always possible to give low order coefficients in a simple algebraic form. Moreover, there may be multiple regular families with the same limiting value $\mathfrak{c}_\infty \in \mathbb{R}^N$.

Definition 2 *Let $p \ge 0$ and take $G_d = \Delta(S_{d-p} \times S_p)$ as in Example 1. A regular family of critical points with isotropy $(G_d)_{d\ge d_0}$ 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. $+1$).*

In terms of FPS, a family is of type I (resp. type II) if $c_{1,0} = -1$ (resp. $+1$).

Remark 1 *When $k = d$, we suspect that the initial coefficients $c_{i,0}$ of the FPS of a regular family with isotropy $\Delta(S_{d-p} \times S_p)$, $p \geq 0$, satisfy $c_{i,0} \in \{\pm 1, 0\}$. However, this is false if $k > d$—we give an example later. Moreover if $k = d$, and with a slight extension of the notion of regular family, there is a regular family with $G_{2d} = \Delta(S_d \times S_d)$, $d \in \mathbb{N}$, with $c_{1,0} = +1$, $c_{5,0} = -1$ (coefficients corresponding to the diagonal entries of the principal blocks associated to $S_d \times I_d$ and $I_d \times S_d$).*

Henceforth, we emphasize $k > d$, and families with isotropy $\Delta(S_{d-p} \times S_p)$, $p \in \{0, 1\}$. However, the methods are quite general.

Theorem 2 *Suppose that $\mathfrak{C}$ is a regular family of critical points. Assume that initial terms of the asociated FPS do not all vanish and $b \in \{2, 4\}$. If the isotropy $G_d = \Delta S_d$, $k \in \{d, d+1\}$, then $\mathfrak{C}$ is of type I; if the isotropy $G_d = \Delta(S_{d-1} \times S_1)$, $k \in \{d, d+1, d+2\}$, then $\mathfrak{C}$ is either of type I or type II and there exists at least one family of each type. If $k = d$ there is precisely one type I family and, if $p \neq 0$, one type II family.*

Remark 2 *For $k = d$, both type I and type II families are spurious minima [31, 32]. However, empirically, type I minima are not detected by SGD when Xavier initialization is used. 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 may be 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: Assume $k = d$. Under Xavier initialization, $\big(1 - \frac{2}{\pi}\big)d \leq E_W[\mathcal{L}(W)] \leq \big(1 - \frac{1}{\pi}\big)d$ (see Section A.6).*

As we increase $k - d$, the original type I and II critical points of spurious minima persist as degenerate critical sets and new regular families of critical points of the same type are generated. Thus, if $k - d = 1$, and $G_d \subsetneq \Delta S_d$, two regular families of type I points are generated which are swapped by the permutation of rows d and $d+1$. Similarly for families of type II. When $k - d = 2$, $3! = |S_3|$ new regular families of type I critical points appear; similarly for type II. Additional families of critical points, which do not originate from the original families and are not spurious minima, may appear. See Section A.7 for degenerate critical point sets occurring on account of over-parameterization.

Our focus will be on the families of type I and II critical points that arise through the above mechanism. For a given isotropy $G_d = \Delta(S_{d-p} \times S_p)$, $p \in \{0, 1\}$, a type X and $k \geq d$, with $m = k - d$ fixed, let $\mathfrak{C}^{\mathrm{X}}_{p,m}$ denote a choice of regular family of critical points $\{\mathfrak{c}^{\mathrm{X}}_{p,m}(d) \in M(k,d)^{G_p}\}$ that originates from the unique regular family of critical points of type X that exists when $k = d$. It is enough to analyze just one of the type X families when $k > d$ as, by equivariance, the choices lie on the same $S_d \times S_k$-orbit and so have similar Hessians. Once the existence of the FPS representation for the families $\mathfrak{C}^{\mathrm{X}}_{p,m}$ has been proved, the next step is to estimate the Hessian spectrum, the topic of the next section.

We conclude with examples illustrating the quantitative power of our approach.

Example 2 *(1) We investigate how the loss $\mathcal{L}(\mathfrak{c}^{II}_{1,m})$ depends on $k - d$ for the type II families $\mathfrak{C}^{\mathrm{II}}_{1,m}$, $m \in \{0, 1, 2\}$. For $m > 0$, the initial coefficients of the FPS are found using Newton-Raphson method applied either directly for a system of equations corresponding to different orders of the FPS coefficients Section A.5, or for a reduced system of equations obtained through an explicit use of the geometry of the problem Section A.1. We give the asymptotics modulo $O(d^{-\frac{3}{2}})$ and find that if $\mathcal{L}(\mathfrak{c}^{II}_{1,m}) = \alpha_m d^{-1} + O(d^{-\frac{3}{2}})$, then*

$$\alpha_0 = 2.97357632715\ldots (= \frac{1}{2} - \frac{2}{\pi^2}),\ \alpha_1 = 2.67254813889\ldots,\ \alpha_2 = 2.67193392202\ldots$$

(2) Consider $\mathcal{L}(\mathfrak{c}^{I}_{1,m})$ for type I families. We find that $\mathcal{L}(\mathfrak{c}^{I}_{1,m}) = \frac{1}{2} - \frac{1}{\pi} + O(d^{-\frac{1}{2}})$, for $m \in \{0, 1, 2\}$. Higher order terms are m-dependent but can be computed, as they can if $p = 0$. For example, if $p = 0$, $m = 1$ then $\mathcal{L}(\mathfrak{c}^{I}_{0,1}) = \frac{1}{2} - \frac{1}{\pi} - \frac{4}{3\pi} d^{-\frac{1}{2}} + \big(-1 - \frac{2}{\pi^2} + \frac{4}{\pi}\big) d^{-1} + O(d^{-\frac{3}{2}})$.
(3) We conclude with an example where the FPS is in powers of $d^{-\frac{1}{4}}$ and two components of $\mathfrak{c}_\infty \in \mathbb{R}^7$ do not lie in $\{\pm 1, 0\}$. If $k = d + 1$, the initial terms of the FPS for a type I family $\mathfrak{c}(d)$ of critical

points of spurious minima are given by

$$\mathfrak{c}(d)_1 = -1 + 2d^{-1} + \frac{\pi}{2}d^{-\frac{3}{2}} + O(d^{-\frac{7}{4}}), \quad \mathfrak{c}(d)_2 = 2d^{-1} - \sqrt{\pi-2}d^{-\frac{7}{4}} + O(d^{-2}),$$
$$\mathfrak{c}(d)_3 = d^{-1} - \frac{6+3\pi}{4\pi\sqrt{\pi-2}}d^{-\frac{5}{4}} + O(d^{-\frac{3}{2}}), \ \mathfrak{c}(d)_4 = \frac{\sqrt{\pi-2}}{2}d^{-\frac{3}{4}} + O(d^{-1}),$$
$$\mathfrak{c}(d)_5 = \frac{1}{2} + \frac{6+3\pi}{8\pi\sqrt{\pi-2}}d^{-\frac{1}{4}} + O(d^{-\frac{1}{2}}), \quad \mathfrak{c}(d)_6 = \frac{\sqrt{\pi-2}}{2}d^{-\frac{3}{4}} + O(d^{-1}),$$
$$\mathfrak{c}(d)_7 = -\frac{1}{2} + \frac{6+3\pi}{8\pi\sqrt{\pi-2}}d^{-\frac{1}{4}} + O(d^{-\frac{1}{2}})$$

(see Section A.5 for details). The loss $\mathcal{L}(\mathfrak{c}^I_{1,1}) = \frac{1}{2} - \frac{1}{\pi} - \frac{4}{3\pi}d^{-\frac{1}{2}} - \frac{(\pi-2)^{\frac{3}{2}}}{3\pi}d^{-\frac{3}{4}} + O(d^{-1})$.

4 Hessian spectrum

The FPS representation makes possible an analytic characterization of the Hessian spectrum using tools from the representation theory of groups (see Section B for a brief review). The main tool used is the *isotypic decomposition* relating the isotropy of a given point to minimal invariant subspaces of the Hessian. We begin by presenting the isotypic decomposition needed for the over-parameterized case.

Let $k \geq d$. Regard $M(k,d)$ as an S_d-representation (diagonal action on $M(d,d) \subset M(k,d)$). By restriction, $M(k,d)$ is a $S_q \times S_p$-representation, where $q = d-p$, $p < q$, 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, when $d \geq 4$: 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 $(q-2,2)$ [37, 38]). For $p \in \{0,1\}$, the isotypic decomposition is

$$M(k,d) = (m+3p+2)\mathfrak{t} + (pm+2p+3)\mathfrak{s}_q + \mathfrak{x}_q + \mathfrak{y}_q. \tag{5}$$

Since the representations $\mathfrak{x}_q, \mathfrak{x}_q$ contribute 2 eigenvalues, of total multiplicity $q^2 - 2q + 1$, we have

Lemma 1 *If $k - d = m \geq 0$ and $p \in \{0,1\}$, then of the kd eigenvalues of the Hessian at a 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 strictly restricts the number of distinct eigenvalues of the Hessian spectrum. For fixed $k-d$, the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalues account for $kd - \Theta(d)$ of the eigenvalues. We show that for all families of critical points considered here, the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalues are identical to order $O(d^{-\frac{1}{4}})$.

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

$$\frac{1}{4} - \frac{1}{2\pi} + O(d^{-\frac{1}{4}}) \quad \textit{and} \quad \frac{1}{4} + \frac{1}{2\pi} + O(d^{-\frac{1}{4}})$$

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

The derivation of Theorem 3 builds on a technique used in [31, 32] and is directed towards the case where $k > d$ and the coefficients of FPS may not be given in a simple algebraic form. Specifically, we rewrite the expression for the Hessian eigenvalues in terms of the gradient entries. Since gradient entries vanish at critical points, this allow us to evaluate the eigenvalue expressions. For example, the Puiseux series of the $\mathfrak{x}$-eigenvalue of Type I $\Delta(S_{d-1} \times S_1)$-critical points is

$$\lambda^d_{\mathfrak{x}} = \frac{1}{4} - \frac{1}{2\pi} + [d^0]F_{d,1} - c_{1,2}[d^{\frac{1}{2}}]F_{d,1} - [d^0]F_{d,2} + [d^{\frac{1}{4}}]F_{d,1}d^{\frac{1}{4}} + [d^{\frac{1}{2}}]F_{d,1}d^{\frac{1}{2}} + O(d^{-\frac{1}{4}}), \tag{6}$$

with $[d^\alpha]F_{d,i}$ indicating the coefficient of d^α in $F_{d,i}$. Since $F_{d,i}$ vanish at critical points, $\lambda^d_{\mathfrak{x}} = \frac{1}{4} - \frac{1}{2\pi} + O(d^{-\frac{1}{4}})$. Equation 6 further demonstrates the sensitivity of the $\mathfrak{x}$-eigenvalue to variations in the FPS coefficients and in different orders of the gradient terms, see Section B.1. Algebraic relations between criticality and curvature indicate therefore a certain rigidity of the loss landscape. Relations of similar nature exist between *criticality* and the *loss* at a point.

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 computed the $\mathfrak{x}$- and the $\mathfrak{y}$-eigenvalues, we now turn to describe how the eigenvalues associated to the trivial and the standard representations vary when the number of neurons is increased. We find that while eigenvalues associated to the trivial representation remain strictly positive for all sufficiently large d—some eigenvalues, associated to the standard representation, become negative, indicating a transition from minima to saddles along the isotypic component of the standard representation. We start with points of isotropy ΔS_d.

5.1 Critical points of isotropy ΔS_d

By Theorem 2, if $k \in \{d, d+1\}$, there is one regular family of critical points with isotropy ΔS_d: the type I family $\mathfrak{C}^{\mathrm{I}}_{0,i}$, $i = k - d$. The representation-theoretic tools used in Section 4, yield a complete characterization of the Hessian spectrum of both families of critical points (see the discussion following the statement of Theorem 1 in the introduction for more details).

The 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 $\mathfrak{C}^{\mathrm{I}}_{0,1}$ of non-degenerate saddles where the negative eigenvalue of the Hessian at $\mathfrak{C}^{\mathrm{I}}_{0,1}$ is associated to the standard representation $\mathfrak{s}_d$.

Since the negative eigenvalue of the Hessian at $\mathfrak{C}^{\mathrm{I}}_{0,1}$ is associated to $\mathfrak{s}_d$, there are exactly $d-1$ descent directions, out of $d(d+1)$ possible directions in $M(d+1, d)$, lying in the $4d-4$-dimensional isotypic component $4\mathfrak{s}_d$ spanned by (only nonzero elements are described):

1. The $(d-1)$-dimensional space of $(d+1) \times d$-matrices $[y_{ij}]$ where for $i, j \in [d]$, $y_{ij} = z_i - z_j$, for some $(z_1, \cdots, z_d) \in \mathbb{R}^d$ with $\sum_{i\in[d]} z_i = 0$.
2. The $(d-1)$-dimensional space of $(d+1)\times d$-matrices $[y_{ij}]$ where for $i, j \in [d]$, $i \neq j$, $y_{ij} = z_i + z_j$, where $(z_1, \cdots, z_d) \in \mathbb{R}^d$ with $\sum_{i\in[d]} z_i = 0$.
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.

No regular families with isotropy ΔS_d exists if two neurons or more are added, i.e., $k - d \geq 2$ (see Section A.7).

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

Consider the type II regular families $\mathfrak{C}^{\mathrm{II}}_{1,0}$, $\mathfrak{C}^{\mathrm{II}}_{1,1}$ and $\mathfrak{C}^{\mathrm{II}}_{1,2}$, corresponding to $k = d, d+1$ and $d+2$ respectively. We show that negative eigenvalues of the Hessian appear when $k - d = 2$ but not $k - d = 1$. The same result (and proof) hold for the type I family.

For $k = d, k = d+1$, the eigenvalues associated to the trivial and standard representation $\mathfrak{s}_{d-1}$ are strictly positive—see Table 5.2. By Theorem 3, the $\mathfrak{x}$- and the $\mathfrak{y}$-representation eigenvalues are also

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 $\mathfrak{s}_{d-1}$	$d - 2$	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		$\Theta(1/d)$	$\Theta(1/d)$

Table 1: Type II critical points with symmetry ΔS_{d-1}. The Hessian eigenvalues associated to the trivial and the standard representation are given to four decimal places, modulo $O(d^{-\frac{1}{2}})$-terms.

strictly positive. Therefore, we have families of spurious minima for $k = d, d + 1$. For $k = d + 2$, critical points in the family $\mathfrak{C}^{\mathrm{II}}_{1,2}$ are saddles, with strictly negative eigenvalues associated to $\mathfrak{s}_{d-1}$. To show this, we take a different route so as to reduce the complexity of the computation. Consider the $2d \times 2d$-submatrix $\widehat{H}$ of the Hessian corresponding to the last two rows of the weight matrix. By Cauchy's interlacing theorem [39], 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 $\mathfrak{s}_{d-1}$, the problem is reduced to computing the spectrum of a 2×2 matrix. Using Puiseux series representation, we show that modulo $O(d^{-\frac{1}{2}})$-terms the two eigenvalues of $\hat{H}$ are $\lambda_1 = 0.8060\ldots$ and $\lambda_2 = -0.1198\ldots$. Hence there exists a $(d - 2)$-dimensional eigenspace of descent directions, projecting onto the associated eigenspace for $\widehat{H}$. Applying Cauchy's interlacing theorem again, there must also exist positive eigenvalues associated to $\mathfrak{s}_{d-1}$.

Theorem 4 *(Notation & Assumptions as above.)*

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

Empirically, when $k = d + 1$, minima of isotropy $\Delta(S_{d-1} \times S_1)$ are not seen for $d < 8$ (they are if $k = d$ and $d \geq 6$) and the probability to detect them using gradient descent is much lower for small values of $d \geq 8$ [19] (type I minima are not detected under Xavier initialization). Theorem 4 implies that when $k = d + 2$ there are no spurious minima of type II of symmetry $\Delta(S_{d-1} \times S_1)$ and the descent directions are tangent to a copy of $\mathfrak{s}_{d-1}$. The last point is crucial for understanding the empirical results. The small eigenvalue 0.0230 associated to $\mathfrak{s}_{d-1}$ 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 $\mathfrak{s}_{d-1}$ is special and quite exceptional. In our case, the trivial solution will be a sink for gradient descent (i.e., a strict local minimum of the loss function), when $k = d + 1$, and a source (i.e., a strict local maximum of the loss function) 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. 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 [33, Sections 1.1, 4] for more on this phenomenon. As we increase d, families $\mathfrak{C}^{\mathrm{II}}_{p,}$, $p > 1$, of spurious minima appear which may not be annihilated by adding two neurons. However, no such minima were found in [19] when $k = d + 2, d \leq 20$ (they were for $k = d + 1$). The empirical results provide strong support for a change of stability associated to $\mathfrak{s}_{d-1}$ and suggest the unique role this representation may play in understanding over-parametrization.

Concluding Comments

The rich symmetry structure exhibited by the loss function (2) makes possible an analytic study of the associated nonconvex loss landscape. The approach is twofold. First, the presence of symmetry breaking allows an efficient organization of an otherwise highly complex set of critical points, offering new ways of recognizing, differentiating and understanding local minima (Section 2 and Section 3). Second, for a given family of critical points, symmetry grants a parameterization of fixed dimensionality, independent of the ambient space, which permits a new array of analytic and algebraic tools (Section 4 and Section 5).

In this work, the symmetry breaking framework is used for investigating the nature by which over-parameterization contributes to making the loss landscape of (2) accessible for gradient-based methods. We find that increasing the number of neurons transforms spurious minima into saddles: decent directions are formed along linear subspaces corresponding the standard representation $\mathfrak{s}_d$ of S_d, ascent directions along other representations of S_d persist, and the loss remains essentially the same. The process by which spurious minima turn into saddles suggests a powerful mechanism enabling minimization of the *loss* (rather than the gradient norm [40, 41, 42]) via computationally efficient local search methods, and further highlights the importance of the intricate interplay between symmetries inherent to data distributions and those displayed by neural network models.

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. Hence, at integer values of d, critical points are non-degenerate. It follows that for the families $\mathfrak{C}^{\mathrm{X}}_{p,m}$ we consider, there is an open neighborhood $\mathcal{U}$ of $V = I_d \in M(d,d)$, such that for all $V' \in \mathcal{U}$, the loss function $\mathcal{L}'$ for V' has a non-degenerate critical point close to each point of the $\Delta(S_{d-p} \times S_p)$-orbit of $\mathfrak{C}^{\mathrm{X}}_{p,m}$ with the same number of negative eigenvalues (counting multiplicities). Critical points and eigenvalues depend continuously on $V' \in \mathcal{U}$ and the Hessian spectrum remains extremely skewed.

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 d is even [33]). 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-degenerate 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 $2^{d-1} - 1$ non-degenerate saddles of index $\geq \ell$ passing simultaneously through the origin and emerging as $2^{d-1} - 1$ non-degenerate saddles of index $\leq \ell$. *No new minima or maxima are created.* Forced symmetry breaking leads to great complexity near the transition but minimal models of complexity can be given (*op. cit.*, Section 4).

Rather than striving for generalization, our approach in this work has been to focus on an analytically tractable case, one already acknowledged as difficult [2, 4, 5, 9, 19, 20], that helps elucidate some of the key foundational questions. The phenomena described are robust and so already have the power to disprove or support general conjectures in DL [30, 31, 32]. The symmetry breaking framework used to study these phenomena generalizes beyond the families of minima considered in the present work [43], and applies to other choices of activation functions and distributions [27, 28]. In addition, numerical work indicates that minima of the empirical loss are also symmetry breaking, and so allow theoretical investigations of the empirical loss landscape as well as algorithmic biases (see, e.g., [44]) within the new analytic framework. The full scope and power of symmetry breaking in DL, and more generally stochastic nonconvex optimization, remain to be discovered.

Acknowledgments and Disclosure of Funding

We thank Noa Aharon, Daniel Soudry and Avi Wigderson for valuable discussions and constructive suggestions. The research was supported by the Israel Science Foundation (grant No. 724/22).

References

[1] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *International conference on machine learning*. PMLR, 2017.

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

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

[4] 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.

[5] 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.

[6] Yuandong Tian. Student specialization in deep rectified networks with finite width and input dimension. In *International Conference on Machine Learning*. PMLR, 2020.

[7] Itay M Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In *Conference on Learning Theory*. PMLR, 2021.

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

[9] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *6th International Conference on Learning Representations, ICLR 2018*.

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

[11] Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*. PMLR, 2018.

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

[13] 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.

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

[15] 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*.

[16] 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.

[17] Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, pages 152–162. PMLR, 2021.

[18] 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.

[19] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*. PMLR, 2018.

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

[21] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 2016.

[22] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*. PMLR, 2020.

[23] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[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] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33, 2020.

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

[28] 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.

[29] Yossi Arjevani, Michael Field, and Gal Vinograd. Symmetry & critical points for symmetric tensor decomposition problems. *In preparation*.

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

[31] Yossi Arjevani and Michael Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[33] Yossi Arjevani and Michael Field. Equivariant bifurcation, quadratic equivariants, and symmetry breaking for the standard representation of s k. *Nonlinearity*, 35(6):2809, 2022.

[34] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. *Advances in neural information processing systems*, 27, 2014.

[35] Edward Bierstone and Pierre D Milman. Subanalytic geometry. *Model theory, algebra, and geometry (ed. by Haskell, D. et al.), MSRI*, 39:151–172, 1998.

[36] John Milnor. Singular points of complex hypersurfaces. *Ann. Math. St*, 61:591–648, 1968.

[37] GD James. The representation theory of the symmetric groups. Springer, 1978.

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

[39] Charles R Johnson and Roger A Horn. *Matrix analysis*. Cambridge university press Cambridge, 1985.

[40] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[41] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[42] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

[43] Yossi Arjevani and Michael Field. Bifurcation, spurious minima and over-paramerization. *In preparation*.

[44] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

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]

A Supplementary material for Section 3

We begin by outlining the proof of existence for type II families with $k = d+2$ and isotropy $\Delta(S_{d-1} \times S_1)$ following the strategy in [30, Section 8], and briefly discuss other cases given in Theorem 2. Estimates required for a valid use of the real analytic version of the implicit function theorem are obtained using two different methods presented in Section A.1 and in Section A.5.

A.1 Fractional power series representation for type II minima: $k = d+2$

Following Section 3, we restrict the loss $\mathcal{L}$ to the fixed point space $M(d+2,d)^{\Delta S_{d-1}}$, which is of dimension 9 for $d \geq 3$. We recall the linear isomorphism $\Xi : \mathbb{R}^9 \to M(d+2,d)^{\Delta S_{d-1}}$ defined by

$$\Xi(\boldsymbol{\xi}) = \begin{bmatrix} \xi_1 \mathcal{I}_{d-1,d-1} + \xi_2 I^{\star}_{d-1} & \xi_3 \mathcal{I}_{d-1,1} \\ \xi_4 \mathcal{I}_{1,d-1} & \xi_5 \mathcal{I}_{1,1} \\ \xi_6 \mathcal{I}_{1,d-1} & \xi_7 \mathcal{I}_{1,1} \\ \xi_8 \mathcal{I}_{1,d-1} & \xi_9 \mathcal{I}_{1,1} \end{bmatrix}, \ \boldsymbol{\xi} = (\xi_1, \cdots, \xi_9) \in \mathbb{R}^9$$

Regard Ξ as an identification of $M(d+2,d)^{\Delta S_{d-1}}$ with $\mathbb{R}^9$ and recall from Section 3 that F_d is the vector field on $\mathbb{R}^9$ defined as the pullback of $\nabla\mathcal{L}|M(k,d)^{\Delta S_{d-1}}$ by Ξ.

The map Ξ naturally determines a 4×2-block decomposition of matrices in $M(d+2,d)^{\Delta S_{d-1}}$. If a row $\boldsymbol{w}$ of $W \in M(d+2,d)^{\Delta S_{d-1}}$ lies in row ℓ of the block decomposition, the row is said to be of *row-type* ℓ. Necessarily $\ell \in [4]$. Clearly there are $d-1$-rows of row-type 1 and exactly one row of row-type ℓ, $2 \leq \ell \leq 4$.

Remark 3 *The vector fields $(\nabla\mathcal{L})|M(k,d)^G$, $\nabla(\mathcal{L}|M(k,d)^G)$ on $M(k,d)^G$ are equal,* provided that we use the inner product on $M(k,d)^G$ induced from $M(k,d)$ to define the gradient of $\mathcal{L}|M(k,d)^G$*, and so the eigenvalues of the Hessian of $\mathcal{L}$ at $\mathfrak{c}$ corresponding to directions tangent to $M(k,d)^G$ will equal the eigenvalues of the Hessian of $\mathcal{L}|M(k,d)^G$ at $\mathfrak{c}$. Indeed, the Jacobian of $(\nabla\mathcal{L})|M(k,d)^G$ (or F_d) is equal to the Hessian of $\mathcal{L}|M(k,d)^G$.*

Let S_3 denote the subgroup of S_k permuting the last three rows of $M(d+2,d)$. Suppose $\mathfrak{c}$ is a type II critical point of $\mathcal{L}$. Since $M(d+2,d)^{\Delta S_{d-1}}$ is invariant, but not fixed, by the S_3-action, it follows by the $S_k \times S_d$-invariance of $\mathcal{L}$ that the S_3-orbit of $\mathfrak{c}$ is a subset of $M(d+2,d)^{\Delta S_{d-1}}$ containing at most six points ($|S_3| = 6$).

We claim that if $\mathfrak{c}$ is fixed by a non-identity element of S_3, then the Hessian of $\mathfrak{c}$ has to be singular. This follows since otherwise at least two of the final three rows of $\mathfrak{c}$ must be parallel and so $\mathfrak{c}$ lies in a set of critical points of $\mathcal{L}|M(k,d)^{\Delta S_{d-1}}$ of dimension at least one (see Section A.7 below). In particular, for a regular family, (a) critical points cannot be fixed by a non-identity element of S_3, and (b) if $\mathfrak{C}$ is a regular family in $M(d+2,d)^{\Delta S_{d-1}}$, so is $\sigma\mathfrak{C}$, $\sigma \in S_3$, and their paths do not cross. Generally, we regard the six type II families that result from this observation as being essentially the same and focus on just one of them. That choice comes naturally from an order on the critical points that we discuss shortly. Unlike type I families, the type II family has a rich geometric structure that plays a significant role in the analysis.

The next step is to give an explicit expression for the gradient vector field of $\mathcal{L}$ restricted to $M(d+2,d)^{\Delta S_{d-1}} \approx \mathbb{R}^9$. Given $W \in M(d+2,d)^{\Delta S_{d-1}}$, denote the rows of W by $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{d+2}$. After defining norms, dependent only the row-type, we define angles between different rows of W (capital Greek) and angles between rows of W and rows of V (lower case Greek). Two subscripts are needed for angles between rows of different row-type, one subscript suffices for the same row type. For $p \in [3]$, it is often convenient to set $d_p = d - p$.

For $a \in [4]$, let τ_a denote the norm of $\boldsymbol{w}_i \in M(d+2,d)^{\Delta S_{d-1}}$, where $\boldsymbol{w}_i$ is of row-type a and the norm induced from standard Euclidean inner product on $M(d+2,d)$. In $\mathbb{R}^N$ coordinates, $\tau_1 = \sqrt{\xi_1^2 + d_2\xi_2^2 + \xi_3^2}$, $\tau_2 = \sqrt{d_1\xi_4^2 + \xi_5^2}$, etc.

For $i,j \in [d-1]$, $i \neq j$, let Θ_1 denote the angle between rows $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ and θ_1 denote the angle between row $\boldsymbol{w}_i$ and $\boldsymbol{v}_j$ (here and below, angles are well-defined independently of i,j using symmetry). For $a,b \in [4]$, $a \neq b$, denote the angle between $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$ by Λ_{ab} (that is, the angle between rows of row-types a and b, $a \neq b$). For $a \in [4]$, $b \in [2]$, let λ_{ab} denote the angle between a W-row of type a and a V-row of type b, where if $a = b \leq 2$, we assume W, V are different rows.

Note that Λ_{ab} is symmetric in a, b but λ_{ab} is not; indeed, λ_{ba} is not even defined if $a > 2$. Finally, let β_1 denote the angle between $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$, where $i \in [d-1]$, and β_2 denote the angle between $\boldsymbol{w}_d$ and $\boldsymbol{v}_d$. All these angles may be expressed as inverse cosines of expressions in the variable $\xi_1, \cdots, \xi_9$. For example

$$\Lambda_{24} \doteq \cos^{-1}\left(\frac{\langle w^d, w^{d+2}\rangle}{\tau_2\tau_4}\right) = \cos^{-1}\left(\frac{d_1\xi_4\xi_8 + \xi_5\xi_9}{\tau_2\tau_4}\right)$$

The norms and angles are well defined—depend only row-type—on account of symmetry.

Next we give expressions for the vector field induced on $\mathbb{R}^9$ by $\nabla\mathcal{L}|M(d+2,d)^{\Delta S_{d-1}}$. Define

$$\begin{aligned}
\Gamma_1(\boldsymbol{\xi}) &= d_2\left(\frac{\tau_1\sin(\Theta_1) - \sin(\theta_1)}{\tau_1}\right) + \frac{[\sum_{j\neq 4}\tau_j\sin(\Lambda_{1j})] - \sin(\lambda_{12}) - \sin(\beta_1)}{\tau_1} \\
\Gamma_2(\boldsymbol{\xi}) &= d_1\left(\frac{\tau_1\sin(\Lambda_{21}) - \sin(\lambda_{21})}{\tau_2}\right) + \frac{\tau_3\sin(\Lambda_{23}) + \tau_4\sin(\Lambda_{24}) - \sin(\beta_2)}{\tau_2} \\
\Gamma_3(\boldsymbol{\xi}) &= d_1\left(\frac{\tau_1\sin(\Lambda_{31}) - \sin(\lambda_{31})}{\tau_3}\right) + \frac{\tau_2\sin(\Lambda_{32}) + \tau_4\sin(\Lambda_{34}) - \sin(\lambda_{32})}{\tau_3} \\
\Gamma_4(\boldsymbol{\xi}) &= d_1\left(\frac{\tau_1\sin(\Lambda_{41}) - \sin(\lambda_{41})}{\tau_4}\right) + \frac{\tau_2\sin(\Lambda_{42}) + \tau_3\sin(\Lambda_{43}) - \sin(\lambda_{42})}{\tau_4}
\end{aligned}$$

Define nine "angle" terms.

$A_1^1 = d_2\Theta_1\xi_2 + \Lambda_{i2}\xi_4 + \Lambda_{13}\xi_6 + \Lambda_{14}\xi_8 - \beta_1,\ A_\star^1 = (\xi_1 + d_3\xi_2)\Theta_1 + \Lambda_{12}\xi_4 + \Lambda_{13}\xi_6 + \Lambda_{14}\xi_8 - \theta_1$

$A_2^1 = d_2\Theta_1\xi_3 + \Lambda_{12}\xi_5 + \Lambda_{13}\xi_7 + \Lambda_{14}\xi_9 - \lambda_{12}$

$A_1^2 = (\xi_1 + d_2\xi_2)\Lambda_{12} + \Lambda_{23}\xi_6 + \Lambda_{24}\xi_8 - \lambda_{21},\ A_2^2 = d_1\Lambda_{12}\xi_3 + \Lambda_{23}\xi_7 + \Lambda_{24}\xi_9 - \beta_2$

$A_1^3 = (\xi_1 + d_2\xi_2)\Lambda_{13} + \Lambda_{23}\xi_4 + \Lambda_{34}\xi_8 - \lambda_{31},\ A_2^3 = d_1\Lambda_{13}\xi_3 + \Lambda_{23}\xi_5 + \Lambda_{34}\xi_9 - \lambda_{32}$

$A_1^4 = (\xi_1 + d_2\xi_2)\Lambda_{14} + \Lambda_{24}\xi_4 + \Lambda_{34}\xi_6 - \lambda_{41},\ A_2^4 = d_1\Lambda_{14}\xi_3 + \Lambda_{24}\xi_5 + \Lambda_{34}\xi_7 - \lambda_{42}$

Finally, define

$$\begin{aligned}
\Omega_1 &= \pi\left[\xi_1 + d_2\xi_2 + \xi_4 + \xi_6 + \xi_8 - 1\right] \\
\Omega_2 &= \pi\left[d_1\xi_3 + \xi_5 + \xi_7 + \xi_9 - 1\right]
\end{aligned}$$

Note that Ω_1 is the column sum of any one of the first $d-1$ columns of the matrix $\pi(\Xi(\boldsymbol{\xi}) - V)$ and Ω_2 is the sum of column d of $\pi(\Xi(\boldsymbol{\xi}) - V)$.

The components $(F_{d,1}, \cdots, F_{d,9})$ of F_d are given by

$$\begin{aligned}
F_{d,1}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_1\xi_1 - A_1^1 + \Omega_1) & F_{d,2}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_1\xi_2 - A_\star^1 + \Omega_1) \\
F_{d,3}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_1\xi_3 - A_2^1 + \Omega_2) \\
F_{d,4}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_2\xi_4 - A_2^1 + \Omega_1) & F_{d,5}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_2\xi_5 - A_2^2 + \Omega_2) \\
F_{d,6}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_3\xi_6 - A_1^3 + \Omega_1) & F_{d,7}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_3\xi_7 - A_2^3 + \Omega_2) \\
F_{d,8}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_4\xi_8 - A_1^4 + \Omega_1) & F_{d,9}(\boldsymbol{\xi}) &= \frac{1}{2\pi}(\Gamma_4\xi_9 - A_2^4 + \Omega_2)
\end{aligned}$$

and so the critical point equations on $\mathbb{R}^9$ are $F_{d,i}(\boldsymbol{\xi}) = 0$, $i \in [9]$. That is,

$$\begin{aligned}
&\frac{1}{2\pi}(\Gamma_1\xi_1 - A_1^1 + \Omega_1) = 0 && \frac{1}{2\pi}(\Gamma_1\xi_2 - A_\star^1 + \Omega_1) = 0 \\
&\frac{1}{2\pi}(\Gamma_1\xi_3 - A_2^1 + \Omega_2) = 0 \\
&\frac{1}{2\pi}(\Gamma_2\xi_4 - A_2^1 + \Omega_1) = 0 && \frac{1}{2\pi}(\Gamma_2\xi_5 - A_2^2 + \Omega_2) = 0 \\
&\frac{1}{2\pi}(\Gamma_3\xi_6 - A_1^3 + \Omega_1) = 0 && \frac{1}{2\pi}(\Gamma_3\xi_7 - A_2^3 + \Omega_2) = 0 \\
&\frac{1}{2\pi}(\Gamma_4\xi_8 - A_1^4 + \Omega_1) = 0 && \frac{1}{2\pi}(\Gamma_4\xi_9 - A_2^4 + \Omega_2) = 0
\end{aligned}$$

If $\boldsymbol{\xi} \in \mathbb{R}^9$ determines a type II critical point, then $\xi_1 \to +1$ as $d \to \infty$. Numerics also indicate that for type II one of ξ_5, ξ_7, ξ_9 converges to -1 as $d \to \infty$. Permuting with an element of S_3, we may and shall hypothesize that $\xi_9 \to -1$ as $d \to \infty$. It helps to use some geometry concerning the rows $\boldsymbol{w}_d, \boldsymbol{w}_{d+1}$ of $W = \Xi(\boldsymbol{\xi})$. Let $\boldsymbol{u}_d = \sum_{i \in [d-1]} \boldsymbol{v}_i \in \mathbb{R}^d$ and $\mathbb{F} \subset \mathbb{R}^d$ be the 2-dimensional subspace spanned by $\boldsymbol{v}_d, \boldsymbol{u}_d$. Observe that $\boldsymbol{w}_d, \boldsymbol{w}_{d+1}, \boldsymbol{w}_k, \boldsymbol{v}_d \in \mathbb{F}$ and so are coplanar. By the analyticity properties of regular families, $\boldsymbol{w}_d$ and $\boldsymbol{w}_{d+1}$ cannot be parallel. Hence either $\beta_2 < \lambda_{32}$ or $\beta_2 > \lambda_{32}$: *curves of regular families do not cross*. In the first case $\Lambda_{23} = \lambda_{32} - \beta_2$; in the second $\Lambda_{23} = \beta_2 - \lambda_{32}$. Composing with a unique $\sigma \in S_3$, fixing the last row, we may always assume $\Lambda_{23} = \lambda_{32} - \beta_2$. Numerics indicate that $\lambda_{32} > \beta_2 > \pi/2 > \Lambda_{23} > 0$—but this is not assumed in what follows.

The idea now is to take formal FPS expansions for the components of a type II critical point, substitute in the critical point equations described above, equate like coefficients and thereby obtain FPS solutions. Guided by the numerics, we seek a power series in $d^{-\frac{1}{2}}$. Granted our hypothesis on the constant terms in the FPS, knowledge of vanishing coefficients (see Section A.5), and an easy computation comparing like terms giving $c_{3,2}$, we have

$$\begin{aligned}
&c_{1,0} = 1,\ c_{1,i} = 0,\ i \in [2],\ \ c_{2,i} = 0,\ i \leq 3,\ \ c_{3,i} = 0,\ i < 2,\ c_{3,2} = 2,\\
&c_{4,i} = 0,\ i < 2,\ \ c_{5,0} = 0,\ \ c_{6,i} = 0, i < 2,\ \ c_{7,0} = 0,\ \ c_{8,i} = 0,\ i < 2,\\
&c_{9,0} = -1.
\end{aligned}$$

For notational clarity, we relabel the 9 unknown coefficients $c_{i,j}$ giving the next terms in the FPS so that we aim to find $c_3, e_4, f_3, g_2, h_1, p_2, q_1, a_2, b_1 \in \mathbb{R}$ such that

$$\begin{aligned}
&\xi_1 = 1 + c_3 d^{-\frac{3}{2}} + \cdots \quad \xi_2 = e_4 d^{-2} + \cdots \quad \xi_3 = 2d^{-1} + f_3 d^{-\frac{3}{2}} \cdots\\
&\xi_4 = g_2 d^{-1} + \cdots \qquad \xi_5 = h_1 d^{-\frac{1}{2}} + \cdots\\
&\xi_6 = p_2 d^{-1} + \cdots \qquad \xi_7 = q_1 d^{-\frac{1}{2}} + \cdots\\
&\xi_8 = a_2 d^{-1} + \cdots \qquad \xi_9 = -1 + b_1 d^{-\frac{1}{2}} + \cdots
\end{aligned}$$

The condition we gave on angles holds if and only if $h_1 p_2 > g_2 q_1$ (both sides are negative). Set $R_2 = \sqrt{g_2^2 + h_1^2}$, $R_3 = \sqrt{p_2^2 + q_1^2}$. We derive expressions for the angles Λ_{ab}, $a, b \in \{2, 3, 4\}$ and find that

$$\Lambda_{34} = \Lambda_{34}^0 + O(d^{-\frac{1}{2}}),\ \ \Lambda_{23} = \Lambda_{23}^0 + O(d^{-\frac{1}{2}}),\ \ \Lambda_{24} = \Lambda_{24}^0 + O(d^{-\frac{1}{2}}),$$

where

$$\begin{aligned}
&\Lambda_{24}^0 = \sin^{-1}(g_2/R_2) \in (0, \pi/2)\\
&\Lambda_{23}^0 = \sin^{-1}((h_1 p_2 - g_2 q_1)/(R_2 R_3)) \in (0, \pi/2) \text{ (since } h_1 p_2 > g_2 q_1)\\
&\Lambda_{34}^0 = \sin^{-1}(p_2/R_3) \in (0, \pi/2)
\end{aligned}$$

Using standard trigonometric formulas, we deduce the relationship $\Lambda_{23} + \Lambda_{34} = \Lambda_{24}$ and, letting $d \to \infty$, $\Lambda_{23}^0 + \Lambda_{34}^0 = \Lambda_{24}^0$. We have similar expressions for β_2 and λ_{a2}, $a \in \{3, 4\}$:

$$\beta_2 = \beta_2^0 + O(d^{-\frac{1}{2}})\ \ \lambda_{32} = \lambda_{32}^0 + O(d^{-\frac{1}{2}})\ \ \lambda_{42} = \lambda_{42}^0 + O(d^{-\frac{1}{2}})$$

$$\beta_2^0 = \cos^{-1}(h_1/R_2) \in (\pi/2, \pi)\ \ \lambda_{32}^0 = \cos^{-1}(q_1/R_3) \in (\pi/2, \pi)\ \ \lambda_{42}^0 = \pi$$

It follows that $\Lambda_{24} + \beta_2 = \lambda_{42}$ and $\Lambda_{34} + \lambda_{32} = \lambda_{42}$, with the same identities holding between the constant terms by letting $d \to \infty$. In particular, all the constant terms for the FPS expansions of these angles can be expressed in terms of R_2, Λ_{24}^0 and R_3, Λ_{34}^0 (polar coordinates on (g_2, h_1)- and (p_2, q_1)-space).

Equating like coefficients of terms in the equations and noting in particular that the coefficients $\pi(e_4 + g_2 + p_2 + a_2)$ of d^{-1} in Ω_1, and $\pi(f_3 + h_1 + p_1 + b_1)$ of $d^{-\frac{1}{2}}$ in Ω_2 are zero, we derive a

system of nine nonlinear equations that determine $c_3, e_4, f_3, g_2, h_1, p_2, q_1, a_2, b_1$:

$$
\begin{aligned}
c_3 - b_1 + R_2 + R_3 &= 0 \\
e_4 + g_2 + p_2 + a_2 &= 0 \\
f_3 + h_1 + q_1 + b_1 &= 0 \\
g_2\left(\frac{c_3R_2^2 - 2g_2h_1}{R_2^3} + \frac{-b_1g_2 + h_1a_2 - g_2q_1 + h_1p_2}{R_2^2}\right) &= \Lambda_{23}^0 p_2 + \Lambda_{24}^0 a_2 + \frac{\pi}{2}e_4 - \frac{2h_1}{R_2} \\
h_1\left(\frac{c_3R_2^2 - 2g_2h_1}{R_2^3} + \frac{-b_1g_2 + h_1a_2 - g_2q_1 + h_1p_2}{R_2^2}\right) &= \Lambda_{23}^0 q_1 + \Lambda_{24}^0 b_1 + \frac{\pi}{2}f_3 - \frac{2g_2}{R_2} + a_2 \\
p_2\left(\frac{c_3R_3^2 - 2p_2q_1}{R_3^3} + \frac{-b_1p_2 + q_1a_2 + p_2h_1 - q_1g_2}{R_3^2}\right) &= \Lambda_{23}^0 g_2 + \Lambda_{24}^0 a_2 + \frac{\pi}{2}e_4 - \frac{2q_1}{R_3} \\
q_1\left(\frac{c_3R_3^2 - 2p_2q_1}{R_3^3} + \frac{-b_1p_2 + q_1a_2 + p_2h_1 - q_1g_2}{R_3^2}\right) &= \Lambda_{23}^0 h_1 + \Lambda_{24}^0 b_1 + \frac{\pi}{2}f_3 - \frac{2p_2}{R_3} + a_2 \\
2 - \frac{\pi}{2}e_4 &= (\pi - \Lambda_{24}^0)g_2 + (\pi - \Lambda_{34}^0)p_2 + \pi a_2 \\
c_3 - \frac{\pi}{2}f_3 + g_2 + p_2 &= (\pi - \Lambda_{24}^0)h_1 + (\pi - \Lambda_{34}^0)q_1 + \pi b_1
\end{aligned}
$$

It is possible to reduce to a system of four equations in $R_2, \Lambda_{24}^0, R_3, \Lambda_{34}^0$ (or two in $\Lambda_{24}^0, \Lambda_{34}^0$), solve and then express the coefficients $c_3, e_4, \cdots, b_1$ in terms of $R_2, \Lambda_{\ell k}^0, R_3, \Lambda_{dk}^0$. In practice, we either use Newton-Raphson method on the original system, using an initialization based on numerics, or eliminate c_3, e_4, f_3 from the 9 equations and solve the resulting nonlinear system of 6 equations using Newton-Raphson. We used the reduction to a system of six equations, and found that

$$
\begin{array}{rclrcl}
c_3 & = & -0.5748287640041448964\ldots & e_4 & = & -1.6165352425422284608\ldots \\
f_3 & = & 0.2969965493462016520\ldots & & & \\
g_2 & = & 0.7877659431796313120\ldots & h_1 & = & -1.1161365378487412475\ldots \\
p_2 & = & 0.1562694812799615923\ldots & q_1 & = & -0.4248280138040598900\ldots \\
a_2 & = & 0.6724998180826355564\ldots & b_1 & = & 1.2439680023065994855\ldots
\end{array}
$$

Newton's method was initialized using numerical data for the critical points when $d = 10^4$ to get rough estimates for the coefficients; the original computation was done in long double precision. The values were compared with those obtained by directly solving FPS equations corresponding to different orders of the FPS terms (see Section A.5), and through a high precision computation of the critical point for $d = 10^{512}$—the values of the components of the gradient at the critical point were all less than 10^{-4000} and matched with those obtained by solutions of the equations above. Note that for $d = 10^{512}$, one expects to be able to read off the required coefficients from the computed critical points to 250 or more decimal places of accuracy (the series is in integer powers of $d^{-\frac{1}{2}}$).

Remark 4 *It is straightforward to check that the Jacobian of the 9 variable system in $(c_3, \cdots, b_1)$ with respect to $(c_3, \cdots, b_1)$ is non-singular (for this, we need expressions for the angles Λ_{ab}^0, $a, b \in \{2, 3, 4\}$, $a \neq b$, in terms of $(c_3, \cdots, b_1)$). This is significant for the next and final step.*

Having determined $c_3, e_4, \cdots, b_1$, we set $t^{-\frac{1}{2}} = s$ and define $\widetilde{\xi}_i(s)$, $i \in [9]$ by

$$
\begin{aligned}
&\xi_1(s) = -1 + s^3\widetilde{\xi}_1(s),\ \xi_2(s) = s^4\widetilde{\xi}_2(s),\ \xi_1(s) = 2s^{-1} + s^3 \\
&\xi_4(s) = s^2\widetilde{\xi}_4(s),\ \xi_5(s) = s\widetilde{\xi}_5(s) \\
&\xi_6(s) = s^2\widetilde{\xi}_6(s),\ \xi_7(s) = s\widetilde{\xi}_5(s) \\
&\xi_8(s) = s^2\widetilde{\xi}_8(s),\ \xi_9(s) = -1 + s\widetilde{\xi}_9(s)
\end{aligned}
$$

where the values of $\widetilde{\xi}_1(0), \ldots, \widetilde{\xi}_9(0)$ are given by $(c_3, \cdots, b_1)$ respectively. Substitute in the critical point equations—with d everywhere replaced by s^{-2}—and cancel the powers of s that occur in each equation. The Jacobian of this system with respect to $(\widetilde{\xi}_1, \ldots, \widetilde{\xi}_9)$ at $s = 0$ is non-singular (this uses Remark 4). Finally, use the real analytic version of the implicit function theorem to obtain the FPS.

Remark 5 *Although the argument only gives the convergence of the FPS for sufficiently large d (that is, sufficiently small $d^{-\frac{1}{2}}$), it appears that the series converges for small d, possibly all d for which the problem is defined. This is similar to what happens when $k = d$ [30].*

A.2 Type II, $k = d + 1$

If $k = d + 1$, we can reduce the analysis to solving a single equation $p(\vartheta) = 0$, where p is a polynomial in $\vartheta, \sin(\vartheta)$ and $\cos(\vartheta)$. We solve $p = 0$ directly using Newton's method, initializing with a value of ϑ suggested by numerics. For completeness, we give p explicitly as well as the coefficients of interest in the associated FPS.

We have $\dim(M(d+1,d)^{\Delta S_{d-1}}) = 7$. The critical point equations are read off easily from those we gave for $k = d + 2$: drop the last two equations and all terms that involve the variables ξ_8, ξ_9 or an angle indexed with a '4'. Just as in the case when when $k = d + 2$, we reduce to finding the initial coefficients $c_3, e_4, f_3, g_2, h_1, p_2, q_1$ in the FPS expansion

$$\xi_1 = 1 + c_3 d^{-\frac{3}{2}} + \cdots,\ \xi_2 = e_4 d^{-2} + \cdots,\ \xi_3 = 2d^{-1} + f_3 d^{-\frac{3}{2}} + \cdots,$$
$$\xi_4 = g_2 d^{-1} + \cdots, \qquad \xi_5 = h_1 d^{-\frac{1}{2}} + \cdots$$
$$\xi_6 = p_2 d^{-1} + \cdots, \qquad \xi_7 = -1 + q_1 d^{-\frac{1}{2}} + \cdots$$

As in the case $k = d + 2$, we obtain a system of seven nonlinear equations.

$$c_3 - b_1 + R_2 = 0$$
$$e_4 + g_2 + a_2 = 0$$
$$f_3 + h_1 + b_1 = 0$$
$$g_2 \left(\frac{c_3 R_2^2 - 2g_2 h_1}{R_2^{\frac{3}{2}}} + \frac{-b_1 g_2 + h_1 a_2}{R_2^2} \right) = \vartheta a_2 + \frac{\pi}{2} e_4 - \frac{2h_1}{R_2}$$
$$h_1 \left(\frac{c_3 R_2^2 - 2g_2 h_1}{R_2^{\frac{3}{2}}} + \frac{-b_1 g_2 + h_1 a_2}{R_2^2} \right) = \frac{\pi}{2} f_3 - \frac{2g_2}{R_2} + \vartheta b_1 + a_2$$
$$\frac{\pi}{2} e_4 + 2 + \vartheta g_2 = 0$$
$$c_3 + \frac{\pi}{2} f_3 + g_2 + \vartheta h_1 = 0,$$

where $R_2 = \sqrt{g_2^2 + h_1^2}$ and $\vartheta = \Lambda_{23}^0 = \sin^{-1}\left(\frac{g_2}{R_2}\right) \in (0, \frac{\pi}{2})$.

We may reduce to a single equation $p(\vartheta) = 0$, where $p = AQ - BP$ and

$$\begin{aligned}
A(\vartheta) &= \frac{2}{2-\pi}\left[\frac{\sin(2\vartheta)}{2}(1-\sin(\vartheta))(\vartheta - \frac{\pi}{2}) + \sin(\vartheta)(1-\sin(\vartheta))^2\right] \\
&\quad + \sin(\theta)\left[2\theta - \frac{2\theta^2}{\pi} - 1 - \frac{\sin(2\theta)}{2}(\frac{2\theta}{\pi} - 1)\right] \\
B(\vartheta) &= \frac{2}{2-\pi}\left[-\cos(\vartheta)(\frac{\pi}{2} - \vartheta)^2 + (\frac{\pi}{2} - \vartheta)(1-\sin(\vartheta))(2 - \sin^2\vartheta) - \cos(\vartheta)(1-\sin(\vartheta))^2\right] \\
&\quad + \left[\cos(\vartheta)(1 - \frac{\pi}{2}) - \sin^3\vartheta(\frac{2\vartheta}{\pi} - 1)\right] \\
P(\vartheta) &= 2 - \frac{4\vartheta}{\pi} - \frac{2\sin(2\vartheta)}{\pi} - 2\cos^3\vartheta \\
Q(\vartheta) &= 2\sin^3\vartheta - \frac{4}{\pi}\sin^2\vartheta
\end{aligned}$$

Using Newton-Raphson, the required solution to $p(\vartheta) = 0$ is given by
$\vartheta = 0.58416413506022510436594641534260755532740719514252671834097577387592202\ldots$
with $|p'(\vartheta)| > 0.4$ and $|p(\vartheta)| < 10^{-2400}$

We have $g_2 = R_2 \sin(\vartheta)$, $h_1 = -R_2 \cos(\vartheta)$ and $R_2 = -P(\vartheta)/A(\vartheta)$. It follows easily that the remaining coefficients $c_3, e_2, \cdots, b_1$ are uniquely determined by ϑ. We find that

$$\begin{array}{llll}
c_3 = & -0.57228787893585490607\ldots & e_4 = & -1.61458989052095508224\ldots \\
f_3 = & 0.29629854644604431015\ldots & & \\
g_2 = & 0.91787878976036618322\ldots & h_1 = & -1.38833511087258399162\ldots \\
a_2 = & 0.69671110076058889902\ldots & b_1 = & 1.09203656442653968 15\ldots
\end{array}$$

The existence of the FPS now follows the method for $k = d + 2$.

A.3 Type I, $k = d+1, d+2$

We conclude with a brief discussion of the type I family when $k = d+1, d+2$. This behaves rather differently from type II. For example, no row of a critical point converges to $\boldsymbol{v}_d$ as $d \to \infty$. Instead, one row converges to $\boldsymbol{v}_d/2$, another to $-\boldsymbol{v}_d/2$. Moreover, the fractional power series are in $d^{-\frac{1}{4}}$ rather than $d^{-\frac{1}{2}}$. As a result, the convergence of critical points as $d \to \infty$ is slow. If $k = d+1$, explicit expressions for the critical coefficients are given in Example 1 of Section 3 (see Section A.5 for a detailed derivation). The analysis then proceeds as in the type II case.

If $k = d+2$, we again have the constant coefficients $1/2, -1/2$ for row-types $2, 3$ respectively (if necessary after a permutation by an element of S_3. Here is probably easiest (certainly faster), to obtain the system of nine equations, as we did for type II, and then solve using Newton-Raphson. In brief, after some work the critical coefficients for the FPS in $d^{-\frac{1}{4}}$ are given by

$$\begin{array}{rclrcl}
c_6 & = & 2.6472714633048307498\ldots & e_7 & = & -1.0684533932698202809\ldots \\
f_5 & = & -0.8644915139550179823\ldots & & & \\
g_3 & = & 0.5342266966349101404\ldots & h_1 & = & 0.4322457569775089911\ldots \\
p_3 & = & 0.5342266966349101404\ldots & q_1 & = & 0.4322457569775088806\ldots \\
a_4 & = & 1.2534701553854549462\ldots & b_1 & = & -0.8753051450722888701\ldots
\end{array}$$

Note that $c_4 = e_4 = 2$, $f_4 = 1$, $g_3 = p_3$ and $h_1 = q_1$, just as for $k = d+1$. The existence of the FPS follows as above.

A.4 Hessian spectrum for type I

In Table 2 we give the Hessian spectrum associated to the standard and trivial representations for type I points when $k = d+1, d+2$ for $d = 10, 100$.

Isotypic comp.	$k = d+1$	$k = d+2$	$k = d+1$	$k = d+2$
d	10	10	100	100
$\mathfrak{t}$	$0.01859, 0.03574$ $0.08472, 0.25925$ $1.5634, 2.90219$ 3.3678	$0.00613, 0.01715$ $0.04434, 0.05308$ $0.2309, 0.3080$ $1.6453, 3.2022$ 3.7551	$0.006647, 0.05914$ $0.20056, 0.27534$ $15.746, 25.4661$ 26.055	$0.006467, 0.01348$ $0.06335, 0.09132$ $0.2697, 0.4127$ $15.846, 25.74$ 26.395
$\mathfrak{s}_{d-1}$	$-0.00230, 0.04343$ $0.1135, 0.2324$ $0.3915, 2.9398$	$-0.03903, 0.00423$ $0.04824, 0.1206$ $0.2630, 0.4559$ 3.2410	$0.03178, 0.0680$ $0.09210, 0.24363$ $0.48426, 25.478$	$-0.03432, 0.03707$ $0.07250, 0.09303$ $0.3103, 0.5329$ 25.751

Table 2: Type I critical points with isotropy ΔS_{d-1} & the Hessian eigenvalues associated to the trivial and the standard representation to four decimal places when $k = d+1, d+2$ and $d = 10$ and 100. The spectrum associated to the $\mathfrak{x}$- and $\mathfrak{y}$-representations is strictly positive and not shown.

Referring to the table, the addition of one extra neuron results in the type I critical point becoming a saddle when $d = 10$ (it defines a spurious minimum if $k = d = 10$) but is a spurious minimum for $d = 100$. If we add a second neuron, the type I critical point becomes a saddle for $d = 100$; most likely a saddle for all $d \geq 4$ (certainly sufficiently large d by our results).

A.5 Type of regular families and derivation of case (3) in Example 2

Evaluating formal Puiseux series at critical points gives rise to algebraic relations between the Puiseux series coefficients, allowing one to argue about the structure of regular families (a-priori, independently of their existence). We show how these relations can be used to deduce that for $k = d+1$, any regular family of $\Delta(S_{d-1} \times S_1)$-critical points with $b = 4$ must be either type I, type II or have all its initial terms vanish. Other isotropy types and pairs of d and k are addressed similarly. Although only the diagonal of the main $(d-1) \times (d-1)$-block is required for determining the type of a family (see Definition 2), we evaluate all the entries which belong to the $(d-1) \times d$-upper block to low-order terms. This makes the derivation of the eigenvalue expressions given in (Section B.1)

more transparent, and further serves as a preparation for the detailed derivation below for the type I $\Delta(S_{d-1} \times S_1)$-critical points given in Example 2.

As in the previous sections, denote the FPS coefficients corresponding to $\xi_1, \xi_2, \ldots, \xi_7$ by $c_i, e_i, f_i, g_i, h_i, p_i, q_i$, and let $F_{d,i}$ denote the i'th component of the vector field F. Note that regularity assumptions imply that all coefficients with (strictly) negative index must vanish (e.g., necessarily, $c_{-1} = 0$). We show that for any regular family of critical points in $M(d+1,d)^{\Delta(S_{d-1}\times S_1)}$ with $b = 4$, it follows that $c_0 = \pm 1$ and $c_1 = e_0 = e_1 = e_2 = e_3 = f_0 = f_1 = f_2 = f_3 = 0$. The following notation comes handy when handling expressions involving Puiseux series coefficients. Given a Puiseux series $E = \sum_{j=j_0}^{\infty} \eta_j d^{-\frac{j}{4}}$ where $j_0 \in \mathbb{Z}$, let $[d^{-\frac{j}{4}}]E := \eta_j$.

Observe that

$$[d]F_{d,1} = \frac{e_0}{2} \quad \text{and} \quad [d]F_{d,3} = \frac{f_0}{2}.$$

Since the gradient entries vanish at critical points, necessarily, $e_0 = f_0 = 0$. Similarly,

$$[d^{\frac{3}{4}}]F_{d,1} = \frac{e_1}{2} \quad \text{and} \quad [d^{\frac{3}{4}}]F_{d,3} = \frac{f_1}{2},$$

imply $e_1 = f_1 = 0$. Assuming momentarily $c_0 \neq 0$, we have

$$[d]F_{d,1} = \frac{c_0\sqrt{-\frac{e_2^4}{c_0^4+2c_0^2e_2^2+e_2^4}+1}}{2\pi} - \frac{c_0}{2\pi\sqrt{c_0^2+e_2^2}},$$
$$[d^{\frac{1}{2}}]F_{d,2} = \frac{e_2\sqrt{-\frac{e_2^4}{c_0^4+2c_0^2e_2^2+e_2^4}+1}}{2\pi} - \frac{e_2 \operatorname{acos}\left(\frac{e_2^2}{c_0^2+e_2^2}\right)}{2\pi} + \frac{e_2}{2} - \frac{e_2}{2\pi\sqrt{c_0^2+e_2^2}}.$$

Dividing the first expression by c_0 and using the resulting expression to simplify the second one gives

$$\frac{e_2}{2} - \frac{e_2 \operatorname{acos}\left(\frac{e_2^2}{c_0^2+e_2^2}\right)}{2\pi} = 0.$$

If, by way of contradiction, e_2 is assumed to be non-zero then the preceding equation gives

$$\operatorname{acos}\left(\frac{e_2^2}{c_0^2+e_2^2}\right) = \pi,$$

a contradiction, as the argument of $\arccos$, $\frac{e_2^2}{c_0^2+e_2^2}$, is nonnegative. Hence, $e_2 = 0$ and

$$[d]F_{d,1} = c_0\left(\frac{1}{2\pi} - \frac{1}{2\pi|c_0|}\right),$$

whence $c_0 = \pm 1$. Evaluating $[d^{\frac{3}{4}}]F_{d,1}$, $[d^{\frac{1}{2}}]F_{d,3}$, $[d^{\frac{1}{4}}]F_{d,2}$ and $[d^{\frac{1}{4}}]F_{d,3}$ then yields $c_1 = e_3 = f_2 = f_3 = 0$, as required.

The case $c_0 = 0$ is addressed similarly. We shall only point out a possible course of derivation rather than provide the full expressions. Recall that $e_0 = e_1 = f_0 = f_1 = 0$ holds regardless of the value assigned to c_0. Evaluating $[d^{\frac{1}{2}}]F_{d,2}$ and $[d^{\frac{1}{2}}]F_{d,3}$ gives $e_2 = f_2 = 0$. Evaluating $[d^{\frac{1}{4}}]F_{d,2}$ gives $e_3 = 0$. By $[d^{\frac{1}{2}}]F_{d,4}$ and $[d^{\frac{1}{2}}]F_{d,6}$, $g_0 = p_0 = 0$. Lastly, evaluating $[d]F_{d,5}$ and $[d]F_{d,7}$ gives $h_0 = q_0 = 0$, concluding the derivation.

The procedure just presented is based on the direct approach described in [30, Section 8] by which one extracts coefficients by directly solving the FPS equations, exactly or numerically, to an increasing order. One proceeds until sufficient information has been obtained so as to establish the existence of an FPS and estimate the Hessian spectrum to a desired order. Below we give a detailed derivation of the type I $\Delta(S_{d-1} \times S_1)$-minima in Example 2 to demonstrate how the approach may be used in practice.

A detailed derivation of case (3) in Example 2. (Notation and assumption as above.) For any family of $\Delta(S_{d-1} \times S_1)$-critical points, $c_0 = \pm 1$ and $c_1 = e_0 = e_1 = e_2 = e_3 = f_0 = f_1 = f_2 = f_3 = 0$. The family of critical points given in Example 2 is type I, and so $c_0 = -1$. The derivation proceeds as follows.

1. Observe that

$$[d^0]F_{d,2} = \frac{e_4}{4} + \frac{g_0}{4} + \frac{p_0}{4} - \frac{1}{2},$$
$$[d^0]F_{d,3} = \frac{f_4}{4} + \frac{h_0}{4} + \frac{q_0}{4} - \frac{1}{4}.$$

   We use these relations to substitute e_4 and f_4 for lower-order terms. That is,

$$e_4 = -g_0 - p_0 + 2,$$
$$f_4 = -h_0 - q_0 + 1.$$

2. By $[d^{\frac{1}{2}}]F_{d,1} = \frac{c_2}{2\pi} - \frac{|g_0|}{2\pi} - \frac{|p_0|}{2\pi}$, $c_2 = |g_0| + |p_0|$.

3. We now show that $g_0 = p_0 = 0$. If $g_0 = 0$ (resp. $p_0 = 0$) then by $[d^0]F_{d,6} = p_0\big(\frac{1}{4} - \frac{1}{2\pi}\big)$ (resp. $[d^0]F_{d,4} = q_0\big(\frac{1}{4} - \frac{1}{2\pi}\big)$), $p_0 = 0$ (resp. $g_0 = 0$). Assume then, by way of contradiction, that both g_0 and p_0 are non-zero, i.e., $g_0 \neq 0$ and $p_0 \neq 0$. Then g_0 and p_0 must satisfy the following four equations (effectively two, by symmetry) obtained by evaluating $[d^0]F_{d,4}$, $[d^0]F_{d,5}$, $[d^0]F_{d,6}$ and $[d^0]F_{d,7}$:

$$-\frac{g_0}{2\pi} + \frac{g_0}{4} - \frac{g_0|p_0|}{2\pi|g_0|} - \frac{p_0 \operatorname{acos}\left(\frac{g_0 p_0}{|g_0||p_0|}\right)}{2\pi} + \frac{p_0}{4} = 0, \tag{7}$$

$$-\frac{h_0}{2\pi} + \frac{h_0}{4} - \frac{h_0|p_0|}{2\pi|g_0|} - \frac{q_0 \operatorname{acos}\left(\frac{g_0 p_0}{|g_0||p_0|}\right)}{2\pi} + \frac{q_0}{4} = 0, \tag{8}$$

$$-\frac{g_0 \operatorname{acos}\left(\frac{g_0 p_0}{|g_0||p_0|}\right)}{2\pi} + \frac{g_0}{4} - \frac{p_0|g_0|}{2\pi|p_0|} - \frac{p_0}{2\pi} + \frac{p_0}{4} = 0, \tag{9}$$

$$-\frac{h_0 \operatorname{acos}\left(\frac{g_0 p_0}{|g_0||p_0|}\right)}{2\pi} + \frac{h_0}{4} - \frac{q_0|g_0|}{2\pi|p_0|} - \frac{q_0}{2\pi} + \frac{q_0}{4} = 0. \tag{10}$$

   If $g_0, p_0 > 0$ or $g_0, p_0 < 0$, then by

$$[d^0]F_{d,4} = -\frac{g_0}{2\pi} + \frac{g_0}{4} - \frac{p_0}{2\pi} + \frac{p_0}{4},$$

   it follows that $g_0 = -p_0$, a contradiction. If $p_0 < 0 < g_0$ then by $[d^0]F_{d,4}$ again, $p_0 = g_0$, still a contradiction. The case where $g_0 < 0 < p_0$ is treated similarly (and in fact follows by symmetry). Thus, necessarily, $g_0 = p_0 = 0$, and so $c_2 = 0$ and $e_4 = 2$.

4. Next, we have $[d^{\frac{1}{4}}]F_{d,1} = \frac{c_3}{2\pi} - \frac{|g_1|}{2\pi} - \frac{|p_1|}{2\pi}$ and $[d^{-\frac{1}{4}}]F_{d,2} = \frac{e_5}{4} + \frac{g_1}{4} + \frac{p_1}{4}$. Solving for c_3 and e_5 shows, by the same argument used in (3), that any of the four cases concerning the (strict) signs of g_1 and p_1 yields a contradiction. Thus, $g_1 = p_1 = 0$, and so $c_3 = e_5 = 0$.

5. We have $[d^0]F_{d,1} = \frac{c_4}{2\pi} - \frac{\sqrt{g_2^2+h_0^2}}{2\pi} - \frac{\sqrt{p_2^2+q_0^2}}{2\pi} - \frac{1}{2\pi}$. In addition, by $[d^{-\frac{1}{4}}]F_{d,3}$ we have $f_5 = -h_1 - q_1$, and by $[d^{-\frac{1}{2}}]F_{d,2}$, $e_6 = -g_2 - p_2$.

6. Using $c_4 = \sqrt{g_2^2+h_0^2}+\sqrt{p_2^2+q_0^2}+1$, we obtain two (effectively one by symmetry) equations corresponding respectively to $[d^0]F_{d,5}$ and $[d^0]F_{d,7}$:

$$\text{(I)}\quad 0 = -\frac{h_0}{2\pi}+\frac{h_0}{4}-\frac{h_0|g_2|}{2\pi\left(g_2^2+h_0^2\right)}+\frac{h_0|g_2p_2+h_0q_0|}{2\pi\left(g_2^2+h_0^2\right)}-\frac{h_0\sqrt{p_2^2+q_0^2}}{2\pi\sqrt{g_2^2+h_0^2}}+\frac{h_0}{2\pi\sqrt{g_2^2+h_0^2}}$$
$$-\frac{q_0\operatorname{acos}\left(\frac{g_2p_2}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}+\frac{h_0q_0}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}\right)}{2\pi}+\frac{q_0}{4}+\frac{\operatorname{acos}\left(\frac{h_0}{\sqrt{g_2^2+h_0^2}}\right)}{2\pi}-\frac{1}{4},$$

$$\text{(II)}\quad 0 = -\frac{h_0\operatorname{acos}\left(\frac{g_2p_2}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}+\frac{h_0q_0}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}\right)}{2\pi}+\frac{h_0}{4}-\frac{q_0\sqrt{g_2^2+h_0^2}}{2\pi\sqrt{p_2^2+q_0^2}}-\frac{q_0}{2\pi}+\frac{q_0}{4}$$
$$-\frac{q_0|p_2|}{2\pi\left(p_2^2+q_0^2\right)}+\frac{q_0|g_2p_2+h_0q_0|}{2\pi\left(p_2^2+q_0^2\right)}+\frac{q_0}{2\pi\sqrt{p_2^2+q_0^2}}+\frac{\operatorname{acos}\left(\frac{q_0}{\sqrt{p_2^2+q_0^2}}\right)}{2\pi}-\frac{1}{4}.$$

In addition, by $[d^{-\frac{1}{2}}]F_{d,4}$,

$$\text{(III)}\quad 0 = -\frac{g_2}{2\pi}+\frac{g_2}{4}-\frac{g_2|g_2|}{2\pi\left(g_2^2+h_0^2\right)}+\frac{g_2|g_2p_2+h_0q_0|}{2\pi\left(g_2^2+h_0^2\right)}-\frac{g_2\sqrt{p_2^2+q_0^2}}{2\pi\sqrt{g_2^2+h_0^2}}+\frac{g_2}{2\pi\sqrt{g_2^2+h_0^2}}$$
$$-\frac{p_2\operatorname{acos}\left(\frac{g_2p_2}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}+\frac{h_0q_0}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}\right)}{2\pi}+\frac{p_2}{4}.$$

The system of the FPS equations is symmetric under $(g_2,h_0)\leftrightarrow(p_2,q_0)$, and so we get the following symmetrized version of (III),

$$\text{(IV)}\quad 0 = -\frac{g_2\operatorname{acos}\left(\frac{g_2p_2}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}+\frac{h_0q_0}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}\right)}{2\pi}+\frac{g_2}{4}-\frac{p_2\sqrt{g_2^2+h_0^2}}{2\pi\sqrt{p_2^2+q_0^2}}-\frac{p_2}{2\pi}$$
$$+\frac{p_2}{4}-\frac{p_2|p_2|}{2\pi\left(p_2^2+q_0^2\right)}+\frac{p_2|g_2p_2+h_0q_0|}{2\pi\left(p_2^2+q_0^2\right)}+\frac{p_2}{2\pi\sqrt{p_2^2+q_0^2}}.$$

7. Our next goal is prove $p_2=g_2=0$. This step is somewhat more involved. Recall that for the family described in Example 2 $h_0=1/2=-q_0$, and so the following expressions are well-defined,

$$x = -\frac{1}{2\pi}+\frac{1}{4}-\frac{|g_2|}{2\pi\left(g_2^2+h_0^2\right)}+\frac{|g_2p_2+h_0q_0|}{2\pi\left(g_2^2+h_0^2\right)}-\frac{\sqrt{p_2^2+q_0^2}}{2\pi\sqrt{g_2^2+h_0^2}}+\frac{1}{2\pi\sqrt{g_2^2+h_0^2}},$$
$$y = -\frac{1}{2\pi}+\frac{1}{4}-\frac{|p_2|}{2\pi\left(p_2^2+q_0^2\right)}+\frac{|g_2p_2+h_0q_0|}{2\pi\left(p_2^2+q_0^2\right)}-\frac{\sqrt{g_2^2+h_0^2}}{2\pi\sqrt{p_2^2+q_0^2}}+\frac{1}{2\pi\sqrt{p_2^2+q_0^2}},$$
$$z = -\frac{\operatorname{acos}\left(\frac{g_2p_2}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}+\frac{h_0q_0}{\sqrt{g_2^2+h_0^2}\sqrt{p_2^2+q_0^2}}\right)}{2\pi}+\frac{1}{4}.$$

Equations (I-IV) now read:

$$\text{(I)}\quad 0 = h_0x+q_0z+\frac{\operatorname{acos}\left(\frac{h_0}{\sqrt{g_2^2+h_0^2}}\right)}{2\pi}-\frac{1}{4},$$
$$\text{(II)}\quad 0 = h_0z+q_0y+\frac{\operatorname{acos}\left(\frac{q_0}{\sqrt{p_2^2+q_0^2}}\right)}{2\pi}-\frac{1}{4},$$
$$\text{(III)}\quad 0 = g_2x+p_2z,$$
$$\text{(IV)}\quad 0 = g_2z+p_2y.$$

Combining (I) and (III) (resp. (II) and (IV)) by solving (I) for x (resp. solving (II) for y) and substituting yields,

$$\text{(A)} \quad 0 = \frac{-g_2}{h_0}\left(q_0 z + \frac{\operatorname{acos}\left(\frac{h_0}{\sqrt{g_2^2+h_0^2}}\right)}{2\pi} - \frac{1}{4}\right) + p_2 z,$$

$$\text{(B)} \quad 0 = g_2 z - \frac{p_2}{q_0}\left(h_0 z + \frac{\operatorname{acos}\left(\frac{q_0}{\sqrt{p_2^2+q_0^2}}\right)}{2\pi} - \frac{1}{4}\right).$$

Summing (A) and (B) we have,

$$0 = \left(p_2 + g_2 - \frac{g_2 q_0}{h_0} - \frac{p_2 h_0}{q_0}\right) z - \frac{g_2}{h_0}\left(\frac{\operatorname{acos}\left(\frac{h_0}{\sqrt{g_2^2+h_0^2}}\right)}{2\pi} - \frac{1}{4}\right) - \frac{p_2}{q_0}\left(\frac{\operatorname{acos}\left(\frac{q_0}{\sqrt{p_2^2+q_0^2}}\right)}{2\pi} - \frac{1}{4}\right).$$

For the derivation so far to be valid only $h_0, q_0 \neq 0$ is required. Plugging-in $h_0 = 1/2 = -q_0$, the above becomes

$$0 = \frac{g_2 \operatorname{acos}\left(\frac{1}{2\sqrt{g_2^2+\frac{1}{4}}}\right)}{\pi} - \frac{g_2}{2} - \frac{p_2 \operatorname{acos}\left(-\frac{1}{2\sqrt{p_2^2+\frac{1}{4}}}\right)}{\pi} + \frac{p_2}{2}. \tag{11}$$

The function

$$f(x) = \frac{x \operatorname{acos}\left(\frac{1}{2\sqrt{x^2+\frac{1}{4}}}\right)}{\pi} - \frac{x}{2}$$

is injective, and so $g_2 = p_2$ by Equation 11. Plugging in this into (A) yields $f(g_2) = 0$. Since $f(0) = 0$ we have, by the injectivity of f again, that $g_2 = 0$, hence $h_2 = 0$ as well. Backward substitution then gives $c_4 = 2,\ f_4 = 1, e_6 = 0$. The FPS equations encountered in the reminder of the derivation are simpler.

8. We have $[d^{-\frac{1}{4}}]F_{d,5} = -\frac{h_1}{2\pi} + \frac{h_1}{4} - \frac{q_1}{4} + \frac{q_1}{2\pi}$, hence $h_1 = q_1$. Therefore, $[d^{-\frac{1}{4}}]F_{d,1} = \frac{c_5}{2\pi} - \frac{h_1}{2\pi} + \frac{q_1}{2\pi}$ implies $c_5 = 0$.
9. By $[d^{-\frac{1}{2}}]F_{d,3} = \frac{f_6}{4} + \frac{h_2}{4} + \frac{q_2}{4}$ and $[d^{-\frac{1}{2}}]F_{d,5} = -\frac{c_6}{2\pi} + \frac{f_6}{4} + \frac{h_2}{2}$, hence $f_6 = -h_2 - q_2$ and $c_6 = \frac{\pi h_2}{2} - \frac{\pi q_2}{2}$.
10. By $[d^{-\frac{3}{4}}]F_{d,2} = \frac{e_7}{4} + \frac{g_3}{4} + \frac{p_3}{4}$, $[d^{-\frac{3}{4}}]F_{d,3} = \frac{f_7}{4} + \frac{g_3}{2\pi} + \frac{h_3}{4} - \frac{p_3}{2\pi} + \frac{q_3}{4}$, $[d^{-\frac{3}{4}}]F_{d,4} = \frac{e_7}{4} + \frac{g_3}{2}$ and $[d^{-\frac{3}{4}}]F_{d,6} = \frac{e_7}{4} + \frac{p_3}{2}$, $e_7 = -2p_3,\ f_7 = -h_3 - j_3$ and $g_3 = p_3$.
11. By $[d^{-\frac{3}{4}}]F_{d,5}$, $c_7 = \frac{\pi h_3}{2} - 4p_3 - \frac{\pi q_3}{2}$, and by $[d^{-1}]F_{d,2}$, $e_8 = -g_4 - p_4 + \frac{4}{\pi}$.
12. We now have

$$[d^{-1}]F_{d,4} = \frac{g_4}{4} + \frac{2p_3|p_3|}{\pi} - \frac{p_4}{4} - \frac{1}{2} + \frac{3}{2\pi},$$
$$[d^{-1}]F_{d,6} = -\frac{g_4}{4} + \frac{2p_3|p_3|}{\pi} + \frac{p_4}{4} - \frac{1}{2} + \frac{1}{2\pi}.$$

Summing the two equations gives

$$\frac{4p_3|p_3|}{\pi} - 1 + \frac{2}{\pi} = 0.$$

The equation has a single root at $p_3 = g_3 = \frac{\sqrt{-2+\pi}}{2}$. Backward substitution then gives $e_7 = -\sqrt{-2+\pi}$.

13. The expression $[d^{-\frac{1}{2}}]F_{d,1}$ depends on g_3 which has just been determined. Re-evaluating, we have $h_2 = q_2 + 1$, and so $c_6 = \frac{\pi}{2}$.
14. By $[d^{-1}]F_{d,4}$, $g_4 = p_4 - \frac{2}{\pi}$, by $[d^{-1}]F_{d,3}$, $f_8 = -h_4 - q_4 + \frac{4}{\pi^2}$, and by $[d^{-\frac{3}{4}}]F_{d,1}$,

$$h_3 = \frac{-4\pi p_4\sqrt{-2+\pi} - q_3(\pi^2 + 2\pi) + 4\sqrt{-2+\pi}}{\pi\,(2-\pi)}.$$

15. Now,

$$[d^{-1}]F_{d,5} = -\frac{c_8}{2\pi} + \frac{h_4}{4} - \frac{2p_4}{\pi} - \frac{4q_1\sqrt{-2+\pi}}{3\pi} + \frac{2q_1\sqrt{-2+\pi}}{3} - \frac{q_4}{4} - \frac{1}{2} - \frac{1}{2\pi} + \frac{\pi}{8} + \frac{7}{\pi^2},$$
$$[d^{-1}]F_{d,7} = \frac{c_8}{2\pi} - \frac{h_4}{4} + \frac{2p_4}{\pi} - \frac{4q_1\sqrt{-2+\pi}}{3\pi} + \frac{2q_1\sqrt{-2+\pi}}{3} + \frac{q_4}{4} - \frac{5}{\pi^2} - \frac{\pi}{8} + \frac{1}{2\pi}.$$

Summing the two equations above, we get

$$4q_1\sqrt{-2+\pi}\left(\frac{1}{3} - \frac{2}{3\pi}\right) - \frac{1}{2} + \frac{2}{\pi^2} = 0,$$

hence $q_1 = h_1 = \frac{6+3\pi}{8\pi\sqrt{-2+\pi}}$ (recall that $q_1 = h_1$). Consequently, $f_5 = -\frac{6+3\pi}{4\pi\sqrt{-2+\pi}}$.

16. The procedure may be further iterated by observing, e.g., that $[d^{-1}]F_{d,5}$ implies $c_8 = \frac{\pi h_4}{2} - 4p_4 - \frac{\pi q_4}{2} - \frac{\pi}{2} - 1 + \frac{\pi^2}{4} + \frac{12}{\pi}$ and so on, if additional coefficients are needed.

For type II critical points, the equations corresponding to Equations (I-IV) above are different, and we have not been able to solve them exactly. Rather, Newton-Raphson method was used to obtain numerical estimates. The same procedure was then applied iteratively, with the aid of numerical methods, giving estimates for higher-order terms. The estimates obtained through this process match with those obtained by the method described in Section A.1. Other choices of types, k, d and isotropy were addressed similarly.

A.6 Expected Initial Value

Bounding $\mathbb{E}_W[\mathcal{L}(W)]$ follows by a straightforward computation of the expected loss. We first derive explicit expressions for the terms used in computation.

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{w})\sim\mathcal{N}(0,I_d)^{\otimes 2}}[\sigma^2(\langle \boldsymbol{w},\boldsymbol{x}\rangle)] = \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}\frac{\|\boldsymbol{w}\|^2}{2} = \frac{d}{2},$$

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{v}_2,\boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{v}_1,\boldsymbol{x}\rangle)] = \begin{cases} 1/2 & \boldsymbol{v}_1 = \boldsymbol{v}_2, \\ 1/(2\pi) & \boldsymbol{v}_1 \neq \boldsymbol{v}_2, \end{cases}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle)] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle) \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle \geq 0] + \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle) \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle < 0] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\langle \boldsymbol{w},\boldsymbol{x}\rangle \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle \geq 0] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\boldsymbol{w} \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle \geq 0]^\top \boldsymbol{x} \\
&= \frac{\|\boldsymbol{x}\|}{\sqrt{2\pi}},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle)] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle) \mid \boldsymbol{w}^\top\boldsymbol{x} \geq 0] + \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\sigma(\langle \boldsymbol{w},\boldsymbol{x}\rangle) \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle < 0] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\langle \boldsymbol{w},\boldsymbol{x}\rangle \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle \geq 0] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,I_d)}[\boldsymbol{w} \mid \langle \boldsymbol{w},\boldsymbol{x}\rangle \geq 0]^\top \boldsymbol{x} \\
&= \frac{\|\boldsymbol{x}\|}{\sqrt{2\pi}},
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_{\|\boldsymbol{x}\|=1}[\sigma(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)] &= \frac{\|\boldsymbol{v}\|}{2}\mathbb{E}_{\theta}[\langle \boldsymbol{v}/\|\boldsymbol{v}\|, \boldsymbol{x}\rangle \mid \langle \boldsymbol{v}/\|\boldsymbol{v}\|, \boldsymbol{x}\rangle \geq 0] \\
&= \frac{\|\boldsymbol{v}\|}{2}\mathbb{E}_{\|\boldsymbol{x}\|=1}[x_1 \mid x_1 \geq 0] \\
&= \frac{\|\boldsymbol{v}\|}{2}\frac{2}{d\text{Beta}((n+1)/2, 1/2)} \\
&= \frac{\|\boldsymbol{v}\|}{d\text{Beta}((n+1)/2, 1/2)},
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{\substack{\boldsymbol{w}\sim\mathcal{N}(0,I_d)\\ \boldsymbol{x}\sim\mathcal{N}(0,I_d)}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)] \\
&\quad= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(0,I_d)}\left[\frac{1}{\sqrt{2\pi}}\|\boldsymbol{x}\|\sigma(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)\right] \\
&\quad= \mathbb{E}_{r}\mathbb{E}_{\|\boldsymbol{\theta}\|=1}\left[\frac{1}{\sqrt{2\pi}}r\|\boldsymbol{\theta}\|\sigma(\langle \boldsymbol{v}, r\boldsymbol{\theta}\rangle)\right] \\
&\quad= \mathbb{E}_{r}\frac{v}{\sqrt{2\pi}}r^2\mathbb{E}_{\|\boldsymbol{\theta}\|=1}[\sigma(\langle \boldsymbol{v}, \boldsymbol{\theta}\rangle)] \\
&\quad= \frac{1}{\sqrt{2\pi}}\frac{\|\boldsymbol{v}\|}{d\text{Beta}((d+1)/2, 1/2)}\mathbb{E}_{r}r^2 \\
&\quad= \frac{\|\boldsymbol{v}\|}{\sqrt{2\pi}\text{Beta}((d+1)/2, 1/2)},
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x},\boldsymbol{v},\boldsymbol{w})\sim\mathcal{N}(0,I_d)^{\otimes 3}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{v}, \boldsymbol{x}\rangle)] \\
&\quad= \frac{\mathbb{E}_{\boldsymbol{v}}\|\boldsymbol{v}\|}{\sqrt{2\pi}\text{Beta}((d+1)/2, 1/2)} \\
&\quad= \frac{1}{\sqrt{\pi}\text{Beta}((d+1)/2, 1/2)}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \\
&\quad= \frac{1}{\pi\text{Beta}(d/2, 1)} \\
&\quad= \frac{d}{2\pi}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\substack{\boldsymbol{x}\sim\mathcal{N}(0,I_d)\\ W\sim\mathcal{N}\left(0_{d\times d}, I_{d^2}\right)}}[(\mathbf{1}^\top\sigma\left(\frac{1}{\sqrt{d}}W\boldsymbol{x}\right) - \mathbf{1}^\top\sigma(V\boldsymbol{x}))^2] \\
&\quad= \mathbb{E}[d\frac{1}{d}\sigma^2(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) + \frac{1}{d}d(d-1)\sigma(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}_2, \boldsymbol{x}\rangle) - \frac{1}{\sqrt{d}}2d^2\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{v}, \boldsymbol{x}\rangle) + d\sigma^2(\langle \boldsymbol{v}, \boldsymbol{x}\rangle) \\
&\quad+ d(d-1)\sigma(\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{v}_2, \boldsymbol{x}\rangle)] \\
&\quad= \frac{d}{2} + (d-1)\frac{d}{2\pi} - \frac{2d^{3/2}}{\sqrt{2\pi}\text{Beta}((d+1)/2, 1/2)} + \frac{d}{2} + d(d-1)\frac{1}{2\pi} \\
&\quad= d + \frac{d(d-1)}{\pi} - \frac{\sqrt{2}d^{3/2}\Gamma(d/2+1)}{\sqrt{\pi}\Gamma((d+1)/2)\Gamma(1/2)} \\
&\quad\leq d + \frac{d(d-1)}{\pi} - \frac{\sqrt{2}d^{3/2}}{\pi}\left(\frac{d}{2}\right)^{1/2} \\
&\quad= \left(1 - \frac{1}{\pi}\right)d,
\end{aligned}
$$

where the penultimate transition uses Gautschi's inequality. The same inequality also gives the following lower bound,

$$\begin{aligned}\mathbb{E}[(\mathbf{1}^\top\sigma\left(\frac{1}{\sqrt{d}}W\boldsymbol{x}\right)-\mathbf{1}^\top\sigma(V\boldsymbol{x}))^2] &\geq \left(1-\frac{1}{\pi}\right)d+\frac{1}{\pi}(d^2-\sqrt{2}d^{3/2}(d/2+1)^{1/2})\\ &= \left(1-\frac{1}{\pi}\right)d+\frac{d^2\left(1-\sqrt{1+\frac{2}{d}}\right)}{\pi}\\ &\geq \left(1-\frac{1}{\pi}\right)d+\frac{d^2(1-(1+1/d))}{\pi}\\ &= \left(1-\frac{2}{\pi}\right)d,\end{aligned}$$

with the penultimate transition following by the first-order Taylor expansion of $\sqrt{1+x}$.

A.7 Adding more than two neurons: beyond non-degenerate critical points

When we add neurons to a shallow network, new critical points appear and old critical points become simplices with singular Hessian spectrum (at points where the Hessian is defined). This phenomenon is well-known and not restricted to ReLU networks. We shall refer to this process here as "fossilization", as the set of (connected) fossils together, with the discrete critical points, encodes information about the number of additional neurons involved, and how they were added. Thus the fossil record generated when $p > 1$ neurons are added simultaneously may be less informative than that generated when p neurons are added one-by-one. Moreover, as we show, symmetry plays a significant role in the description of the fossilized sets; even if the target V is asymmetric.

The critical points giving the global minimum fossilize when neurons are added. Recall that $\mathcal{L}$ is always S_k-invariant, where S_k is the group of row permutations of $M(k,d)$. If $d = k$, we add the superscript r (resp. c) to emphasize row (resp. column) permutations. In our setting, if $k = d$, $V = I_d$, the $d!$ points in the S_d^r-orbit of V will be non-degenerate critical points of $\mathcal{L}$ giving the global minimum zero: these will be the only points in $M(d,d)$ that give the global minimum. If we add $p \geq 1$ neurons, the discrete S_d^r-orbit of V is replaced by a p-dimensional connected S_{d+p}-invariant simplicial complex $Z \subset M(p+d,d)$ consisting of all points giving the global minimum. Necessarily, the Hessian, where defined on Z, will have zero eigenvalues; that does not preclude Z from being an attractor under gradient descent. Suppose instead that $\mathfrak{c} \in M(k,d)$, $k \geq d$, is any (non-degenerate) critical point of $\mathcal{L}$. The addition of a p neurons will replace $\mathfrak{c}$ by a connected p-dimensional simplex, invariant by the action of S_d^r. Often (not always) many new non-degenerate critical points will be created as biproducts of the fossilization process.

For completeness, we give precise statements and proof of these results, starting with the case when $\mathfrak{c} = V$ and $k = d$. We start with an extension of the result on the uniqueness of critical points defining global minima [30, Prop. 4.14] to the over-parameterized case. See [7] for related results and discussions.

Assume $k \geq d$, set $m = k - d$, and $\Gamma = S_k \times S_d$. Let $S_m = \{e_d\} \times S_m$ denote the subgroup of S_k permuting the last m rows of matrices in $M(k,d)$. Define $\Delta_m S_d = \{(hg, g) \mid gS_d,\ h \in S_m\}$. Note that $\Delta_m S_d \approx \Delta S_d \times S_m$.

Let $\mathfrak{K}^\star$ denote the set of all partitions of $[k]$ such that each $\mathcal{K} \in \mathfrak{K}^\star$ has exactly d parts, $K_1, \cdots K_d$ and $j \in K_j$, for all $j \in [d]$. If $j \in [d]$, then

$$K_j \cap [d] = \{j\}, \quad \text{and } K_j \smallsetminus \{j\} \subset [k] \smallsetminus [d].$$

Clearly, $1 \leq |K_j| \leq m+1$ for all $j \in [d]$.

If $\mathcal{K} \in \mathfrak{K}^\star$, let $M_{\mathcal{K}} = [k_{ij}] \in M(k,d)$ be the matrix defined by

$$k_{ij} = 0,\ i \notin K_j \tag{12}$$

$$= 1,\ i \in K_j \tag{13}$$

For $j \in [d]$, define $\Delta_j(\mathcal{K}) \subset \mathbb{R}^k$ by

$$\Delta_j(\mathcal{K}) = \{(t_1, \cdots, t_k) \in \prod_{i \in K_j} [0, k_{ij}] \mid \sum_{i \in K_j} t_i = 1\}$$

and, viewing $M(k,d)$ as $(\mathbb{R}^k)^d$, define

$$\Delta(\mathcal{K}) = \prod_{j\in[d]} \Delta_j(\mathcal{K}) \subset M(k,d).$$

Clearly $\Delta(\mathcal{K})$ is a simplicial complex of dimension m and if $m=0$, $\Delta(\mathcal{K}) = \{V\}$. If $\boldsymbol{\delta} \in \Delta(\mathcal{K})$, then

$$\boldsymbol{\delta}^\Sigma = V^\Sigma = \mathcal{I}_{1,d},$$

where for $W \in M(k,d)$, W^Σ is the $1\times d$-row matrix defined by the column sums of W. Define

$$\boldsymbol{\Delta}(k,d) = \bigcup_{\mathcal{K}\in\mathfrak{K}^\star} \Delta(\mathcal{K}). \tag{14}$$

Suppose $W \in M(k,d)$ and $\boldsymbol{\delta} \in \Delta(\mathcal{K})$. Define $W_{\boldsymbol{\delta}} \in M(k,d)$ by

$$\begin{aligned} \boldsymbol{w}_i^{\boldsymbol{\delta}} &= \delta_{ii}\boldsymbol{w}_i,\ i\in[d] \\ &= \boldsymbol{w}_i + \sum_{j\in[d]} \delta_{ij}\boldsymbol{w}_j,\ i>d \end{aligned}$$

Observe that

$$W^\Sigma = W_{\boldsymbol{\delta}}^\Sigma.$$

The case of most interest will be when $W=V$ and so the last m rows of W will be zero.

Lemma 2 *(Notation and assumptions as above.)*

1. *For all $\mathcal{K}\in\mathfrak{K}^\star$, $V\in\Delta(\mathcal{K})$.*
2. *If $\mathcal{K},\mathcal{J}\in\mathfrak{K}^\star$, $\mathcal{K}\neq\mathcal{J}$, then $\Delta(\mathcal{K})\cap\Delta(\mathcal{J})$ is a simplicial complex which is the union of the common vertices and faces of $\Delta(\mathcal{K})$ and $\Delta(\mathcal{J})$.*
3. *$\boldsymbol{\Delta}(k,d)$ is a connected simplicial complex of dimension m.*
4. *$\Delta_m S_d(\boldsymbol{\Delta}(k,d)) = \boldsymbol{\Delta}(k,d)$ and $\Delta_m S_d$ is the maximal subgroup of Γ with this property.*
5. *If $g\in S_d^r\cup S_d^c$, $g\neq e$, then $g\boldsymbol{\Delta}(k,d)\cap\Gamma V = gV \neq V$.*
6. *$\mathcal{L}(W)=0$ if $W\in\boldsymbol{\Delta}(k,d)$.*

Statements (1–3) are all immediate from the definitions and the proofs of (4,5) are straightforward and omitted. It remains to prove (6). Recall that

$$\mathcal{L}(W) = \frac{1}{2}\sum_{i,j\in[k]} f(\boldsymbol{w}_i,\boldsymbol{w}_j) - \sum_{i\in[k],j\in[d]} f(\boldsymbol{w}_i,\boldsymbol{v}_j) + \frac{1}{2}\sum_{i,j\in[d]} f(\boldsymbol{v}_i,\boldsymbol{v}_j),$$

where

1. If $\boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^d$ are non-zero and we set $\theta_{\boldsymbol{w},\boldsymbol{v}} = \cos^{-1}\left(\frac{\langle \boldsymbol{w},\boldsymbol{v}\rangle}{\|\boldsymbol{w}\|\|\boldsymbol{v}\|}\right)$, then
$$f(\boldsymbol{w},\boldsymbol{v}) = \frac{1}{2\pi}\|\boldsymbol{w}\|\|\boldsymbol{v}\|(\sin(\theta_{\boldsymbol{w},\boldsymbol{v}}) + (\pi-\theta_{\boldsymbol{w},\boldsymbol{v}})\cos(\theta_{\boldsymbol{w},\boldsymbol{v}}))$$
2. $f(\boldsymbol{w},\boldsymbol{v})=0$ iff either $\boldsymbol{v}$ or $\boldsymbol{w}$ is zero or $\theta_{\boldsymbol{w},\boldsymbol{v}}=\pi$.

Clearly, f is positively homogeneous:

$$f(a\boldsymbol{w},b\boldsymbol{v}) = abf(\boldsymbol{w},\boldsymbol{v}),\ a,b\geq 0,\ \boldsymbol{w},\boldsymbol{v}\in\mathbb{R}^d.$$

If $W\in\boldsymbol{\Delta}(k,d)$, then there exist $\mathcal{K}\in\mathfrak{K}^\star$ and $\boldsymbol{\delta}\in\Delta(\mathcal{K})$ such that $W=V_{\boldsymbol{\delta}}$—$\boldsymbol{v}_j$ is zero for $j>d$. The result follows from the positive homogeneity of f and the formula for $\mathcal{L}$ in terms of f. □

Remark 6 *If $k=d$, V is the natural choice for a critical point on the group orbit of critical points giving the global minima. When $k>d$, the natural choice—at least from a symmetry perspective—is the set $\boldsymbol{\Delta}(k,d)$ which is invariant by $\Delta_m S_d$ (the isotropy group of V). It follows by (5) of the lemma that for all $g\in S_d\times S_d$, $g\boldsymbol{\Delta}(k,d)$ contains exactly one point in ΓV.*

Define $\mathbf{\Delta}^\star = \Gamma\mathbf{\Delta}(k,d)$ and note that $\Gamma[V] \subset \mathbf{\Delta}^\star$.

Theorem 5 *(Notation and assumptions as above.)*

1. $\mathbf{\Delta}^\star$ *is a* Γ*-invariant* m*-dimensional simplicial complex of* $M(k,d)$.
2. $\mathbf{\Delta}^\star$ *is connected.*
3. $\mathcal{L}(W) = 0$ *iff* $W \in \mathbf{\Delta}^\star$.

By Lemma 2, $\mathbf{\Delta}(k,d)$ is a connected m-dimensional simplicial complex and it follows easily that $\mathbf{\Delta}^\star$ is a Γ-invariant m-dimensional simplicial complex, proving (1). (2) If $g \in \Gamma$, $g\mathbf{\Delta}(k,d) \cap \mathbf{\Delta}(k,d)$ may be empty. However, given any $g \in \Gamma$, it is easy to choose a sequence $g_1, \cdots g_{n+1} \in S_k$ such that such that $g_1 = e$, and $g_j\mathbf{\Delta}(k,d) \cap g_{j+1}\mathbf{\Delta}(k,d) \neq \emptyset$, $j \in [n]$, and $g_{n+1}\mathbf{\Delta}(k,d) \cap g\mathbf{\Delta}(k,d) \neq \emptyset$. Hence $\Gamma\mathbf{\Delta}(k,d) = \mathbf{\Delta}^\star$ is connected (see Example 3 below for more detail). It is straightforward to check that for all $g \in \Gamma$, and $\mathcal{K}, \mathcal{J} \in \mathfrak{K}^\star$, $g\Delta(\mathcal{K}) \cap \Delta(\mathcal{J})$ is a simplicial complex (possibly empty) and from this it follows that $\mathbf{\Delta}^\star$ is a simplicial complex.
(3) The 'if' implication is immediate from the Γ-invariance of $\mathcal{L}$, the connectedness of $\mathbf{\Delta}^\star$ and Lemma 2(6). For the converse, we show that if $\mathcal{L}(W) = 0$, then (a) $w_{ij} \in [0,1]$, $(i,j) \in [k] \times [d]$ and $\sum_{i\in[k]} w_{ij} = 1$, for all $j \in [d]$ (and so $W^\Sigma = \mathcal{I}_{1,d}$). The remainder of the proof follows along the same lines as that of [30, Prop. 4.14] except that now for each $j \in [d]$ we have to allow for several rows of W being strictly positive multiples of v^j since $k > d$. $\square$

Example 3 *Suppose* $m = 1$ *and* $d = 2$*. We claim that* $\mathbf{\Delta}^\star$ *is connected. Here the set* $\mathfrak{K}^\star$ *contains only two partitions:* $K = \{\{1,3\},\{2\}\}$*,* $J = \{\{1\},\{2,3\}\}$*. Hence there are two families of matrices in* $\mathbf{\Delta}(3,2)$

$$X(\alpha,\beta) = \begin{bmatrix} \alpha & 0 \\ 0 & 1 \\ \beta & 0 \end{bmatrix}, \quad Y(\gamma,\delta) = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \\ 0 & \delta \end{bmatrix},$$

where $\alpha + \beta = \gamma + \delta = 1$*,* $\alpha, \beta, \gamma, \delta \geq 0$*.*

Let $\mathbf{\Delta}^\star_{\mathbf{0}}$ *denote the connected component of* $\mathbf{\Delta}^\star$ *containing* X *(that is, the arc* $X(\alpha,\beta)$*,* $\alpha + \beta = 1$*,* $\alpha, \beta \geq 0$*). Use the symbol* $\sim$ *is signify that two families intersect. For example,* $X \sim Y$ *since* $X(1,0) = Y(1,0)$*. We claim* $(12)^r X \in \mathbf{\Delta}^\star_{\mathbf{0}}$*. This follows since* $X_1 = (13)^r X \sim X$*,* $X_2 = (12)^r X_1 \sim X_1$*,* $X_3 = (23)^r X_2 \sim X_2$ *and* $X_3 = (12)^r X$*. It is easy to see that* $\mathbf{\Delta}^\star$ *is isomorphic to a hexagon: 6 vertices, 6 edges and that* $\Delta S_2(\mathbf{\Delta}(3,2)) = \mathbf{\Delta}(3,2)$*, where* $\Delta S_2 = \Gamma_V$*. See Figure 1 where the vertices, connecting edges and symmetries of* $\mathbf{\Delta}^\star$ *are shown.*

The argument is general and applies when $k > d$*—that is, when* $V \in M(k,d)$ *has at least one row of zeros. The connection can always be made through* row *permutations.*

Fossilization of critical points, general case.

The phenomenon described above occurs when the network is over-parameterized. In what follows we assume $V \in M(d,d)$ is a matrix with no zero or parallel rows, and extend in the usual way to $V \in M(k,d)$, $k > d$.

Suppose $W \in M(k,d)$ is a critical point of $\mathcal{L}$: in particular, assume that $\mathcal{L}$ is C^2 at W and so W has no zero rows. Typically, we assume that W is non-degenerate: all the eigenvalues of the Hessian are non-zero.

Let $\bar{k} > k$ and $\mathfrak{K}^\star$ denote the set of all partitions of $[\bar{k}]$ such that each $\mathcal{K} \in \mathfrak{K}^\star$ has exactly k parts, $K_1, \cdots K_k$ and $j \in K_j$, for all $j \in [k]$.

Just as we did previously, if $\mathcal{K} \in \mathfrak{K}^\star$ and $j \in [k]$, we define the simplex $\Delta_j(\mathcal{K}) \subset \mathbb{R}^{\bar{k}}$, and simplicial set $\Delta(\mathcal{K}) = \prod_{j\in[k]} \Delta_j(\mathcal{K}) \subset M(\bar{k},k)$ of dimension $\bar{k} - k$. Set $\mathbf{\Delta}(\bar{k},k) = \cup_{\mathcal{K}\in\mathfrak{K}^\star} \Delta(\mathcal{K})$.

Given $W \in M(k,d)$, $\mathcal{K} \in \mathfrak{K}^\star$ and $\boldsymbol{\delta} \in \Delta(\mathcal{K})$, define $W_\delta \in M(\bar{k},d)$ by

$$\begin{aligned} \boldsymbol{w}_i^{\boldsymbol{\delta}} &= \delta_{ii}\boldsymbol{w}_i, \ i \in [k] \\ &= \sum_{j\in[k]} \delta_{ij}\boldsymbol{w}_j, \ i > k \end{aligned}$$

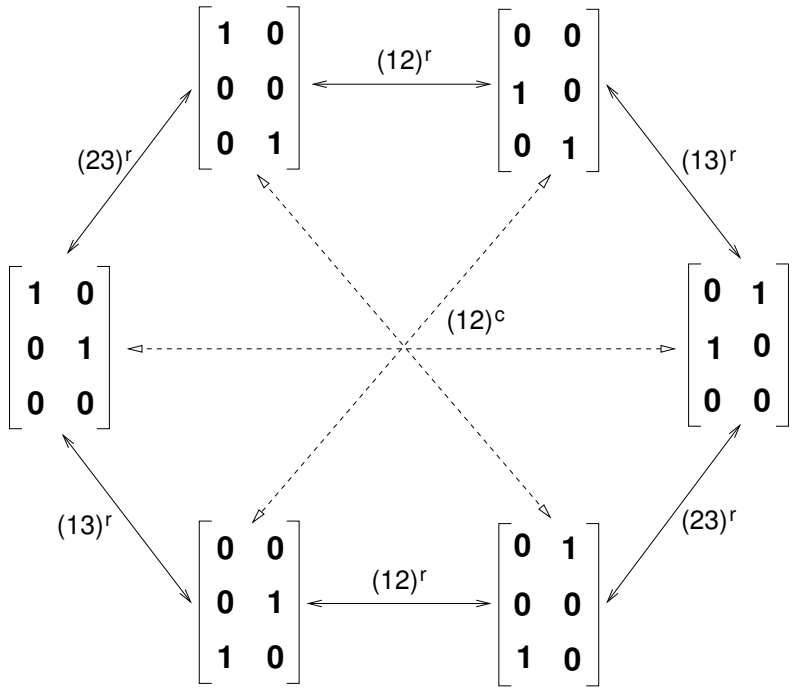

Figure 1: The simplex $\mathbf{\Delta}^{\bigstar}$ in case $k = 3, d = 2$. Connecting edges are shown using unbroken lines and are labelled by the row transposition that interchanges vertices. The simplicies $\mathbf{\Delta}(3,2), (12)^c\mathbf{\Delta}(3,2)$ are both invariant by ΔS_2.

As we did above when $W = V$, we have $W_{\delta}^{\Sigma} = W^{\Sigma}$. It follows from the definitions that if $\boldsymbol{\delta} \in \Delta(\mathcal{K})$, then $\sum_{\ell \in K_i} \boldsymbol{w}_{\ell}^{\boldsymbol{\delta}} = \boldsymbol{w}_i$, for all $i \in [k]$.

Given $W \in M(k,d)$, define $\Delta(\bar{k},k)(W) = \{W_{\boldsymbol{\delta}} \mid \boldsymbol{\delta} \in \Delta(\bar{k},k)\}$ and

$$\mathbf{\Delta}^{\bigstar}(W) = \Gamma\Delta(\bar{k},k)(W) \subset M(\bar{k},d).$$

Set $\mathbf{\Delta}^{\bigstar\bigstar}(W) = \mathbf{\Delta}^{\bigstar}(W) \smallsetminus \partial\mathbf{\Delta}^{\bigstar}(W)$

Proposition 1 *(Notation and assumptions as above.)*

1. $\mathbf{\Delta}^{\bigstar}(W)$ *is a connected* S_d^r*-invariant simplical complex of dimension* $\bar{k} - k$.
2. $\mathcal{L}$ *is* C^2 *at all points of* $\mathbf{\Delta}^{\star\star}(W)$ *and* $\nabla\mathcal{L}|\mathbf{\Delta}^{\star\star}(W) \equiv 0$
3. $\mathcal{L}$ *is constant on* $\mathbf{\Delta}^{\bigstar}(W)$.

The proof of (1) is similar to that of Theorem 5 (1,2). Since $\mathbf{\Delta}^{\bigstar}(W)$ is connected, and $\mathcal{L}$ is continuous, (3) follows from (2) ((3) may be proved directly as in Theorem 5 and then (2) follows using the regularity of $\mathcal{L}$ on $\mathbf{\Delta}^{\star\star}(W)$). For completeness, we give a direct proof that $\nabla\mathcal{L}$ vanishes on $\mathbf{\Delta}^{\star\star}(W)$. By an obvious induction, we can reduce to showing that if the partition $\mathcal{K} \in \mathfrak{K}^{\star}$ satisfies $|K_i| = 1$, $i < k$ and $|K_k| - 1 = n = \bar{k} - k \geq 1$, then $\nabla\mathcal{L}|\Delta(\mathcal{K}) \smallsetminus \partial\Delta(\mathcal{K}) \equiv 0$. Set $m = n + 1 = |K_k|$.

By standard results on $\nabla\mathcal{L}$ [1], W is a critical point of $\mathcal{L}$ if for $i \in [k]$

$$\sum_{j\in[k]} \left(\frac{\|\boldsymbol{w}_j\| \sin(\theta_{\boldsymbol{w}_i,\boldsymbol{w}_j})}{\|\boldsymbol{w}_i\|} \boldsymbol{w}_i - \theta_{\boldsymbol{w}_i,\boldsymbol{w}_j} \boldsymbol{w}_j \right) - \sum_{j\in[d]} \left(\frac{\sin(\theta_{\boldsymbol{w}_i,\boldsymbol{v}_j})}{\|\boldsymbol{w}_i\|} \boldsymbol{w}_i - \theta_{\boldsymbol{w}_i,\boldsymbol{v}_j} \boldsymbol{v}_j \right) + \pi(W - V)^{\Sigma} = 0 \tag{15}$$

Suppose that $U \in \Delta(\mathcal{K}) \smallsetminus \partial\Delta(\mathcal{K})$. In order to check whether or not U is a critical point, $\boldsymbol{w}_k$ will be replaced in (15) by m non-zero parallel rows $\boldsymbol{u}_\ell = \alpha_\ell \boldsymbol{w}_k$, where $\sum_{\ell\in[m]} \alpha_\ell = 1$ and $\alpha_\ell > 0$, $\ell \in [m]$. Since the new rows $\boldsymbol{u}_\ell$ are all non-zero and strictly positive multiples of $\boldsymbol{w}_k$,

(A) $\theta_{\boldsymbol{w}_k,\boldsymbol{w}_j} = \theta_{\boldsymbol{u}_\ell,\boldsymbol{w}_j},\ j \in [k]$, and $\theta_{\boldsymbol{u}_\ell,\boldsymbol{u}_{\ell'}} = 0$ for all $\ell, \ell' \in [n]$.

(B) $\|\boldsymbol{u}_\ell\| = \alpha_\ell \|\boldsymbol{w}_k\|,\ \ell \in [m]$.

(C) $\theta_{\boldsymbol{u}_\ell,\boldsymbol{v}_j} = \theta_{\boldsymbol{w}_k,\boldsymbol{v}_j}$, for all $\ell \in [m]$, $j \in [d]$.

(D) $(W - V)^\Sigma = (U - V)^\Sigma$ (since $\sum_{\ell\in[m]} \alpha_\ell = 1$).

We have m new expressions replacing the right hand side of (15) in case $i = k$:

$$\sum_{j\in[\bar{k}]} \left(\frac{\|\boldsymbol{w}_j\| \sin(\theta_{\boldsymbol{u}_\ell,\boldsymbol{w}_j})}{\|\boldsymbol{u}_\ell\|} \boldsymbol{u}_\ell - \theta_{\boldsymbol{u}_\ell,\boldsymbol{w}_j} \boldsymbol{w}_j \right) - \sum_{j\in[\bar{k}]} \left(\frac{\sin(\theta_{\boldsymbol{u}_\ell,\boldsymbol{v}_j})}{\|\boldsymbol{u}_\ell\|} \boldsymbol{u}_\ell - \theta_{\boldsymbol{u}_\ell,\boldsymbol{v}_j} \boldsymbol{v}_j \right) + \pi(U - V)^\Sigma,$$

where by $\sum_{j\in[\bar{k}]}$, we mean the sum over $\boldsymbol{w}_j$ terms, $j \in [k-1]$, and $\boldsymbol{w}_{\ell'}$ terms, $\ell' \in [m]$. If $\ell \in [m]$, it follows from (A,B) that

$$\sum_{j\in[\bar{k}]} \left(\frac{\|\boldsymbol{w}_j\| \sin(\theta_{\boldsymbol{u}_\ell,\boldsymbol{w}_j})}{\|\boldsymbol{u}_\ell\|} \boldsymbol{u}_\ell - \theta_{\boldsymbol{u}_\ell,\boldsymbol{w}_j} \boldsymbol{w}_j \right) = \sum_{j\in[d]} \left(\frac{\|\boldsymbol{w}_j\| \sin(\theta_{\boldsymbol{w}_k,\boldsymbol{w}_j})}{\|\boldsymbol{w}_k\|} \boldsymbol{w}_k - \theta_{\boldsymbol{w}_\ell,\boldsymbol{w}_j} \boldsymbol{w}_j \right),$$

and from (B,C) that

$$\sum_{j\in[k]} \left(\frac{\sin(\theta_{\boldsymbol{u}_\ell,\boldsymbol{v}_j})}{\|\boldsymbol{u}_\ell\|} \boldsymbol{u}_\ell - \theta_{\boldsymbol{u}_\ell,\boldsymbol{v}_j} \boldsymbol{v}_j \right) = \sum_{j\in[k]} \left(\frac{\sin(\theta_{\boldsymbol{w}_k,\boldsymbol{v}_j})}{\|\boldsymbol{w}_i\|} \boldsymbol{w}_i - \theta_{\boldsymbol{w}_k,\boldsymbol{v}_j} \boldsymbol{v}_j \right)$$

Noting (D), it follows that U satisfies the critical point equations for $\ell \in [m]$. Along the same lines, but now using the convexity condition $\sum_{\ell\in[m]} \alpha_\ell = 1$, we verify that U satisfies the critical point equations for $i \in [k-1]$. Hence U is a critical point of $\mathcal{L}$. □

B Supplementary material for Section 4

We give a brief review of the technique used in this work to compute the Hessian spectrum. The introduction follows [31, 32] verbatim and is provided here for completeness.

Suppose $V \subset \mathbb{R}^m$ is a linear subspace, with Euclidean inner product induced from $\mathbb{R}^m$, and (V, G) is an orthogonal G-representation.

Lemma 3 *[31, Lemma 7, Setion B.1] The representation (V, G) may be written as an orthogonal direct sum $\bigoplus_{i=1}^m (\oplus_{j=1}^{p_i} V_{ij})$ where $V_{ij} \subset V$, (V_{ij}, G) is irreducible, and (V_{ij}, G) is isomorphic to $(V_{\ell k}, G)$ iff $i = \ell$, and $j, k \in [p_i]$. The subspaces $\oplus_{j=1}^{p_i} V_{ij}$ are unique, $i \in [m]$.*

If $p_i = 1$, for all $i \in m$, the orthogonal decomposition given by the lemma is unique, up to order; otherwise the decomposition is not unique. In spite of the lack of uniqueness of Lemma 3, in some cases there may be *natural* choices of invariant subspace for the irreducible components. This is exactly the situation for the isotypic decomposition of $(M(k,k), G)$, $G = S_p \times S_{k-p}$. This naturality allows us to give natural constructions of the matrices M_i, $i \in [m]$, used for determining the spectrum of G-maps $A : M(k,k) \to M(k,k)$.

The isotypic decomposition for $(M(k,k), S_k)$ is $2\mathfrak{t} + 3\mathfrak{s} + \mathfrak{x} + \mathfrak{y}$, $k \geq 4$ (see Section 4). The subspace of $M(k,k)$ determined by $2\mathfrak{t}$ is the set of all $k \times k$ matrices $\mathcal{T} = \{T_{a,b} \mid a, b \in \mathbb{R}\}$ where the diagonal entries of $T_{a,b}$ all equal a and the off-diagonal entries all equal b. There are many ways to write $\mathcal{T}$ as an orthogonal direct sum. For example, $\mathcal{T} = T_{1,1}\mathbb{R} \oplus T_{\frac{2}{k}, -\frac{1}{k(k-1)}}\mathbb{R}$. However, there is only one natural way: $\mathcal{T} = T_{1,0}\mathbb{R} \oplus T_{0,1}\mathbb{R}$. Define $\mathfrak{D}_1^k = T_{1,0}$, $\mathfrak{D}_2^k = T_{0,1}$. If we take the *standard* realization of $(\mathfrak{t}, S_k)$ to be $(\mathbb{R}, S_k)$, where S_k acts trivially on $\mathbb{R}$, then we have natural S_k-maps

$\alpha_1, \alpha_2 : \mathbb{R} \to M(k,k)$ defined by $\alpha_i(t) = t\mathfrak{D}_i^k$, $i = 1, 2$. If $A : M(k,k) \to M(k,k)$ is an S_k-map, then A restricts to the S_k-map $A_{\mathfrak{t}} : \mathcal{T} \to \mathcal{T}$ and $A_{\mathfrak{t}}$ uniquely determines a 2×2-matrix $[a_{ij}]$ by $A_{\mathfrak{t}}(\mathfrak{D}_i^k) = a_{i1}\mathfrak{D}_1^k + a_{i2}\mathfrak{D}_2^k$, $i = 1, 2$. The eigenvalues (and multiplicities in this case) of $A_{\mathfrak{t}} : \mathcal{T} \to \mathcal{T}$ are the same as the eigenvalues of $[a_{ij}]$. If we choose a different orthogonal decomposition of $\mathcal{T}$, we get a different 2×2-matrix that is similar to $[a_{ij}]$ and so has the same eigenvalues. The computation of the rest of the eigenvalues follows similarly.

B.1 Proof of Theorem 3

In Section A.5 algebraic relations between the FPS coefficients were shown to reveal important information on the structure of regular families of critical points. In this section we show how these relations can be further used to evaluate the $\mathfrak{x}$- and the $\mathfrak{y}$-eigenvalues. The method is illustrated for regular families with $b = 4$, $k = d + 1$ and isotropy $\Delta(S_{d-1} \times S_1)$.

Referring to notation and results given in Section A.5, any type I or type II family of $\Delta(S_{d-1} \times S_1)$-critical points must satisfy $c_0 \in \{\pm 1\}$ and $c_1 = e_0 = e_1 = e_2 = e_3 = f_0 = f_1 = f_2 = f_3 = 0$. Below we shall assume that $c_0 = -1$. The assumption is not needed, and is only introduced for ease of presentation. For non-zero g_0 and i_0, the Puiseux series of the eigenvalue associated to the $\mathfrak{x}$-representation is given by:

$$\begin{aligned}\lambda_{\mathfrak{x}} = &-\frac{c_2^2}{2\pi} - \frac{c_2\sqrt{g_0^2}}{2\pi} - \frac{c_2\sqrt{p_0^2}}{2\pi} + \frac{c_4}{2\pi} + \frac{e_4^2}{4\pi} - \frac{1}{\pi} + \frac{1}{4} + \frac{p_2\mathrm{sgn}(p_0)}{2\pi} + \frac{g_2\mathrm{sgn}(g_0)}{2\pi} \\ &+ d^{\frac{1}{4}}\left(\frac{c_3}{2\pi} + \frac{p_1\mathrm{sgn}(p_0)}{2\pi} + \frac{g_1\mathrm{sgn}(g_0)}{2\pi}\right) + d^{\frac{1}{2}}\left(\frac{c_2}{2\pi} + \frac{\sqrt{g_0^2}}{2\pi} + \frac{\sqrt{p_0^2}}{2\pi}\right) + O(d^{-\frac{1}{4}}).\end{aligned} \tag{16}$$

Thus, the expression $\lambda_{\mathfrak{x}}$ depends on FPS coefficients not determined in Section A.5. We show that $\lambda_{\mathfrak{x}}$ can be evaluated nonetheless, independently of the unknown coefficients.

With FPS coefficients as above,

$$\begin{aligned}[d^0]F_{d,1} &= \frac{c_4}{2\pi} + \frac{e_4^2}{4\pi} + \frac{e_4}{4} + \frac{g_0}{4} + \frac{p_0}{4} - \frac{1}{2\pi} + \frac{p_2\mathrm{sgn}(p_0)}{2\pi} + \frac{g_2\mathrm{sgn}(g_0)}{2\pi}, \\ [d^{\frac{1}{4}}]F_{d,1} &= \frac{c_3}{2\pi} + \frac{p_1\mathrm{sgn}(p_0)}{2\pi} + \frac{g_1\mathrm{sgn}(g_0)}{2\pi}, \\ [d^{\frac{1}{2}}]F_{d,1} &= \frac{c_2}{2\pi} + \frac{\sqrt{g_0^2}}{2\pi} + \frac{\sqrt{p_0^2}}{2\pi}, \\ [d^0]F_{d,2} &= \frac{e_4}{4} + \frac{g_0}{4} + \frac{p_0}{4}.\end{aligned} \tag{17}$$

Substituting into (16) gives

$$\lambda_{\mathfrak{x}} = \frac{1}{4} - \frac{1}{2\pi} + [d^0]F_{d,1} - c_2[d^{\frac{1}{2}}]F_{d,1} - [d^0]F_{d,2} + [d^{\frac{1}{4}}]F_{d,1}d^{\frac{1}{4}} + [d^{\frac{1}{2}}]F_{d,1}d^{\frac{1}{2}} + O(d^{-\frac{1}{4}}). \tag{18}$$

In particular, $\lambda_{\mathfrak{x}}$ can be expressed in terms of $[d^0]F_{d,1}, [d^{\frac{1}{4}}]F_{d,1}, [d^{\frac{1}{2}}]F_{d,1}$ and $[d^0]F_{d,2}$. Therefore, by continuity, Equation 18 also applies for $g_0 = p_0 = 0$. Since gradient entries vanish at critical points, so do their Puiseux coefficients and so $[d^0]F_{d,1}, [d^{\frac{1}{4}}]F_{d,1}, [d^{\frac{1}{2}}]F_{d,1}$ and $[d^0]F_{d,2}$ vanish, giving $\lambda_{\mathfrak{x}} = \frac{1}{4} - \frac{1}{2\pi} + O(d^{\frac{1}{4}})$. Equation 18 not only gives the exact value of the $\mathfrak{x}$-eigenvalue to $O(d^{\frac{1}{4}})$-order but also describes its sensitivity to variations in the FPS coefficients. For example, it is seen that varying c_2 accounts for perturbations of order $O(d^{\frac{1}{2}})$. The derivation of the eigenvalue associated to the $\mathfrak{y}$-representation follows along the same lines, giving

$$\lambda_{\mathfrak{y}} = \frac{1}{4} + \frac{1}{2\pi} + [d^0]F_{d,1} - c_2[d^{\frac{1}{2}}]F_{d,1} - [d^0]F_{d,2} + [d^{\frac{1}{4}}]F_{d,1}d^{\frac{1}{4}} + [d^{\frac{1}{2}}]F_{d,1}d^{\frac{1}{2}} + O(d^{-\frac{1}{4}}). \tag{19}$$

Similar relations exist between *criticality* and *loss*.