We thank all the reviewers for their time and the positive feedback acknowledging the power of the methods to analytically characterize aspects of the Hessian which previously were only accessible numerically. The ReLU model we analyse has been the object of study of many works in recent years. We felt that pursuing generalizations in this work would have made the paper hard to read, especially in view of our introduction of novel mathematical techniques, largely unfamiliar in the field of machine learning. We put a lot of thought into how to present the new methods and appreciate the generally positive response of the reviewers to our efforts.

Other choices of distributions, activation functions and architectures are certainly of interest. Indeed, the methods described in the paper apply more broadly [23] and yield different spectral properties for the Hessian. Adding bias (R2) to the activation or addressing a non-fixed second-layer (R2,R4) increases the technical complexity but is quite tractable. The extension of the methods and results to multi-layers and over-parametrization is very much a topic of our current research. The methods and results we present follow in the tradition of mathematics and physics in that we start with a symmetric model, for which we can prove detailed analytic results, and subsequently break symmetry to get insight into the general theory (the results we obtain are robust under symmetry breaking perturbation, see [22]). We agree with the reviewers that a discussion regarding applicability to more general settings would improve the manuscript and we will revise accordingly.

Reviewer #1 *Classification w.r.t number of classes.* Although the number of distinct dominating eigenvalues is of fixed cardinality, their total number (counting multiplicity) grows linearly with k (in Sagun, Bottou and LeCun 2016, the data generating model is based on two Gaussians of varying degree of separability (Figure 9), which could reflect on the effective number of classes in our setting). **35**. The assumption that the distribution is orthogonally invariant, together with permutation symmetries, determine the invariants of the loss and objective function. Choosing a different activation, underlying distribution or architecture may lead to different groups of invariants. **69**. A fair request but difficult to satisfy completely as Theorem 2 involves precise asymptotic estimates on eigenvalues, in terms of $1/\sqrt{k}$, in a setting of symmetric maps and non-trivial representation theory. In the revision we will highlight the spectral computation for $\mathbf{W} = \mathbf{V}$ (Section 3.3), illustrating the role played by representation theory, and also include a new *toy (polynomial) model* in the appendix that illuminates other aspects of Theorem 2. **192**. Symmetry properties of the hessian are inherited from the symmetry (isotropy) of the associated critical point. The hessian spectrum is constant along the group orbit of the critical point.

Reviewer #2 We directly address the generalization error and obtain an analytic description of spectral properties of the respective spurious minima. *Hessian at random weights.* As our intention was to study the spectrum at spurious minima, we view this as a virtue of the analysis. *Generalization and Training loss.* Following a recent series of works [15,17,18,19,21,22] and Hardt *et al.*, 2016; Hardt & Ma, 2016, we focus on the generalization error. Various properties of the training loss can be deduced by concentration of measure arguments, e.g., uniform convergence bounds for gradients (or higher-order derivatives) using generalized vector-valued Rademacher complexity (e.g., Foster *et al.* 2018, Mei *et al.* 2017).

Reviewer #3 **1**. The hessian of the objective function at a critical point is a $k^2 \times k^2$ symmetric matrix. The objective function is real analytic outside of a thin (measure zero) subset of parameter space [24,22]. Real analyticity is crucial for the analysis and for the power series representation in $1/\sqrt{k}$ for the critical points [24]. **2**. For the single neuron case the reviewer considers, $w, v \in \mathbb{R}$ are non-zero scalars and the angle between w, v is either 0 or π. A direct computation shows that the loss $\mathcal{L}(w, v) = (w - v)^2/4$ (resp. $(w^2 + v^2)/4$) if $wv > 0$ (resp. $wv < 0$) and so the *only* critical point is at $w = v$ where the Hessian is $1/2$—the angle zero case. The dependence is missed by differentiating under the integral sign. For $d = k > 1$, the explicit dependence of the hessian on the angle between parameters is given in the article (see also [15,22,24]). Some of the main implications for gradient-based methods based on our analysis: 1. Our spectral analysis *rigorously* establishes the spectrum of the Hessian at global and spurious minima is generically extremely skewed (indeed, highly ill-conditioned as $\kappa(\nabla^2\mathcal{F}) = \Omega(\#\mathrm{Neurons})$). This phenomenon challenges classical approaches to gradient-based methods w.r.t. the associated non-convex landscape [27,28] (reference to Chaudhari *et al.* 2017 and Lee *et al.* 2016 to be added). 2. Stability arguments imply that along the optimization process one should expect the formation of clusters of eigenvalues which drives most of the dynamics, as is indeed the case (note the follow-up discussion for Thm. 1, Section 2, and empirical corroboration of this phenomenon in Section F.1). 3. In terms of dynamical accessibility, our analysis shows that some minima are more likely to be detected by SGD than others. See the discussions under 'Flat minima conjecture and implicit bias'.

Reviewer #4 The issue of which local minima appear is interesting—minima of types I and A did not appear in [22]. The answer appears to involve a mix of initialization (e.g. Xavier) and the detailed critical point structure of the objective function; this is a topic of current work. *Non-fixed second layer.* Please see the comments in paragraph 2.