Hessian Eigenspectra of More Realistic Nonlinear Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given an optimization problem, the Hessian matrix and its eigenspectrum can be used in many ways, ranging from designing more efficient second-order algorithms to performing model analysis and regression diagnostics. When nonlinear models and non-convex problems are considered, strong simplifying assumptions are often made to make Hessian spectral analysis more tractable. This leads to the question of how relevant the conclusions of such analyses are for realistic nonlinear models. In this paper, we exploit tools from random matrix theory to make a *precise* characterization of the Hessian eigenspectra for a broad family of nonlinear models that generalizes the classical generalized linear models, without relying on strong simplifying assumptions used previously. We show that, depending on the data properties, the nonlinear response model, and the loss function, the Hessian can have *qualitatively* different spectral behaviors: of bounded or unbounded support, with single- or multi-bulk, and with isolated eigenvalues on the left- or right-hand side of the main eigenvalue bulk. By focusing on such a simple but nontrivial model, our analysis takes a step forward to unveil the theoretical origin of many visually striking features observed in more realistic machine learning models.

1 Introduction

The Hessian is ubiquitous in applied mathematics, statistics, and machine learning (ML). Given a (loss) function $L(\mathbf{w})$ of some parameters $\mathbf{w} \in \mathbb{R}^p$, the Hessian $\mathbf{H}(\mathbf{w}) \in \mathbb{R}^{p \times p}$ is defined as the second derivative of the objective function with respect to the model parameter, i.e., $\mathbf{H}(\mathbf{w}) = \partial L(\mathbf{w})/(\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T})$. When a ML model is being trained, it is common to parameterize that model by $\mathbf{w}$, and then train that model by minimizing some (smooth) loss function $L(\mathbf{w})$, with the associated Hessian $\mathbf{H}(\mathbf{w})$, e.g., by backpropagating the error to improve $\mathbf{w}$ [25]. Alternatively, once a ML model is trained, the Hessian (and the related Fisher information matrix [60, 62]) can be examined to identify outliers, perform diagnostics, and/or engage in other sorts of model validation [29, 72, 57].

For convex problems, the Hessian $\mathbf{H}(\mathbf{w})$ provides detailed information on how to adjust the gradient to achieve improved convergence, e.g., in Newton-like methods. For non-convex problems, the properties of the local loss "landscape" around a given point $\mathbf{w}$ in the parameter space is of central significance [17, 34, 12, 37, 70, 71, 72]. In this case, most obviously, the signs of the smallest and largest Hessian eigenvalue can be used to test whether a given $\mathbf{w}$ is a local maximum, local minimum, or a saddle point. More subtly, the Hessian eigenvalue distribution characterizes the local curvature of the loss function and provides direct access to, for instance, the faction of negative Hessian eigenvalues that determines the number of (local) descent directions, a quantity that is directly connected to the rates of convergence of various optimization algorithms [31].

For theoretical analysis of neural network (NN) models, Hessian eigenspectra are often assumed to follow well-known random matrix distributions such the Marčenko–Pastur law [42] or the Wigner's

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

semicircle law [64]. This enables one to use Random Matrix Theory (RMT), but it involves (for NNs, at least) making relatively strong simplifying assumptions (e.g., the Hessian can be decomposed as the sum of the two freely independent matrices, the residual error, data feature, and weights are all composed of i.i.d. zero mean normal random variables) [52, 53, 14]. A somewhat more realistic setup involves using a so-called *spiked model* (or a spiked covariance model) [2, 4, 39]. In this case, the matrix follows a *signal-plus-noise* model and consists of *full rank* random noise matrix and *low rank* statistical information structure.[1] The "signal" eigenvalues are generally larger than the noisy "bulk" eigenvalues; and the maximum eigenvalues, when isolated from the bulk, are referred to as the "spikes." A substantial theory-practice gap exists, however. In both toy examples [26] and state-of-the-art NN models [70, 71, 72, 73, 58, 19], the strong simplifying assumptions are far from satisfactory. (A similar theory-practice gap has been observed for other NN matrices to which RMT has been applied, perhaps most notably weight matrices [43, 44].) A more precise understanding of the Hessian eigenspectra (and its dependence on input data structure, the underlying response model and model parameters, as well as the loss function) for more practical models is needed.

1.1 Our approach

In this article, we address these issues, in a setting that is simple enough to be analytically tractable but complex enough to shed light on realistic large-scale models. We consider a family of generalized generalized linear models (G-GLMs) that extends the popular generalized linear model (GLM) [18, 29]; and we show that, even for such simple models, the key simplifying assumptions used in previous theoretical analyses of Hessian can be very inexact. In particular, apart from a few special cases (including linear least squares and logistic regression with homogeneous features), most Hessians of G-GLM are *not* close to the Marčenko–Pastur and/or the semicircle law. Instead, the corresponding Hessian depends on the input feature structure, the underlying response model, and the loss function, in a more involved fashion that can be precisely characterized by the proposed analysis.

The G-GLM describes a generalized linear relation between the input feature $\mathbf{x}_i \in \mathbb{R}^p$ and the corresponding response y_i, in the sense that there exists some parameters $\mathbf{w}_* \in \mathbb{R}^p$ such that for given $\mathbf{w}_*^\mathsf{T}\mathbf{x}_i$, the response y_i is independently drawn from

$$y_i \sim f(y \mid \mathbf{w}_*^\mathsf{T}\mathbf{x}_i) \tag{1}$$

for some conditional density function $f(\cdot \mid \cdot)$. This extends the classical GLM such as

$$\text{logistic model: } \mathbb{P}(y = 1 \mid \mathbf{w}_*^\mathsf{T}\mathbf{x}) = (1 + e^{-\mathbf{w}_*^\mathsf{T}\mathbf{x}})^{-1}, \quad y \in \{-1, 1\}, \tag{2}$$

and covers a large family of models in applications in statistics and ML. Other examples include: i) the (noisy) nonlinear factor model [13] where $y \sim \mathcal{N}(g(\mathbf{w}_*^\mathsf{T}\mathbf{x}), \sigma^2)$ for some nonlinear $g : \mathbb{R} \to \mathbb{R}$ and $\sigma > 0$; ii) the (noiseless) phase retrieval model [20] with $y = |\mathbf{w}_*^\mathsf{T}\mathbf{x}|^2$, in which case one wishes to reconstruct $\mathbf{w}_*$ from its (squared) magnitude measurements; and iii) the single-layer NN model $y = \sigma(\mathbf{w}_*^\mathsf{T}\mathbf{x})$ for some nonlinear activation function $\sigma(t)$ such as the tanh-sigmoid $\sigma(t) = \tanh(t)$.

For a given training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of size n, the standard approach to obtain/recover the parameter $\mathbf{w}_* \in \mathbb{R}^p$ of a G-GLM is to solve the following optimization problem

$$\min_{\mathbf{w}} L(\mathbf{w}) = \min_{\mathbf{w}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(y_i, \mathbf{w}^\mathsf{T}\mathbf{x}_i), \tag{3}$$

for some loss function $\ell(y, h) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, e.g., the negative log-likelihood of the observation model within the maximum likelihood estimation framework [29] such as the logistic loss $\ell(y, h) = \ln(1 + e^{-yh})$ in the case of logistic model (2). In many applications, however, the optimization problem in (3) may *not* be convex, for example to achieve superior robustness and/or accuracy [45, 65, 9], and can be NP-hard in general (the noiseless phase retrieval model $y = (\mathbf{w}_*^\mathsf{T}\mathbf{x})^2$ with the square loss $\ell(y, h) = (y - h^2)^2$ as an example [8]). As we shall see, in such non-convex G-GLMs, the dominant Hessian eigenvector can be shown, in some cases, to positively correlate with the sought-for parameter $\mathbf{w}_*$ and therefore be used as the initialization of gradient descent methods [8, 35, 32]. This particularly motivates our study of the possible isolated Hessian eigenvalue-eigenvector pairs.

1.2 Our main contributions

The main contribution of this work is the *exact* characterization of Hessian eigenspectra for the family of G-GLMs, in the high-dimensional regime where the feature dimension p and the sample size n are both large and comparable. Precisely, we establish:

[1] Since the *same* informative pattern is repeated in each row or column of the matrix.

1. the limiting eigenvalue distribution of the Hessian matrix (Theorem 1); and
2. the behavior of (possible) isolated eigenvalue-eigenvector pairs (Theorem 2 and 3),

as a function of the dimension ratio $c = \lim p/n$, feature statistics, the loss function ℓ in (3), and the underlying response model in (1). Our results are based on a technical result of independent interest:

3. a *deterministic equivalent* (Theorem 4) of the random *resolvent* $\mathbf{Q}(z) = (\mathbf{H} - z\mathbf{I}_p)^{-1}$, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{H}$, of the generalized sample covariance:

$$\mathbf{H} \equiv \mathbf{H}(\mathbf{w}) = \tfrac{1}{n}\sum_{i=1}^n \ell''(y_i, \mathbf{w}^\mathsf{T}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\mathsf{T} \equiv \tfrac{1}{n}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}, \quad \mathbf{D} \equiv \mathrm{diag}\{\ell''(y_i, \mathbf{w}^\mathsf{T}\mathbf{x}_i)\}_{i=1}^n \tag{4}$$

for $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ and $\ell''(y,h) \equiv \partial^2\ell(y,h)/\partial h^2$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, under the setting of *generic* Gaussian feature $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, for $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$. We also demonstrate our results empirically by showing that:

4. for a given response model (1), the Hessian eigenvalue distribution depends on the choice of loss function and the data/feature statistics in an intrinsic manner, e.g., bounded versus unbounded support and single- versus multi-bulk in Fig 2; and
5. there may exist two *qualitatively* different spikes—one due to data *signal* $\boldsymbol{\mu}$ and the other due to $\mathbf{w}_*$ or $\mathbf{w}$ and thus the *underlying model*—which may appear on different sides of the main bulk, and their associated phase transition behaviors are characterized (Fig 4 versus 5).

To have a more clear picture of our contribution, we compare, in Fig 1a and 1b, the Hessian eigenvalues for the logistic model (2) with the (maximum likelihood) logistic loss $\ell(y,h) = \ln(1+e^{-yh})$, for different choices of $\mathbf{w}$ in the parameter space. A nontrivial interplay between the response model, feature statistics and the parameter $\mathbf{w}$ is reflected by the range of the Hessian eigenvalue support and an additional right-hand spike in Fig 1b, as confirmed by our theory. For phase retrieval model $y = (\mathbf{w}_*^\mathsf{T}\mathbf{x})^2$ with square loss $\ell(y,h) = (y-h^2)^2/4$, the non-convex nature of the problem is reflected by a (relatively large) fraction of negative Hessian eigenvalues in Fig 1c. We also note that the top eigenvector (that corresponds to the largest eigenvalue) contains structural information of the underlying model, in the sense that it is positively correlated with $\mathbf{w}_*$, as predicted by our theory. This is indeed connected to the Hessian-based initialization scheme widely used in non-convex problems.

We conclude by emphasizing that, by focusing on the simple yet fundamental G-GLM, we obtain results that improve upon and are different than previous efforts in the following aspects:

i) We provide *precise* asymptotic characterizations of the Hessian eigenspectra that goes beyond, e.g., [6], where only Hessian lower bounds are given in the case of logistic model with logistic loss: our methodology and theoretical results hold much more generally for the family of G-GLM with arbitrary loss. As illustrating examples, we discuss linear least squares in Sec 3.1, logistic model with different choices of loss function in Fig 2, phase retrieval model in Figure 1c, and more in Sec 4 in the appendix.

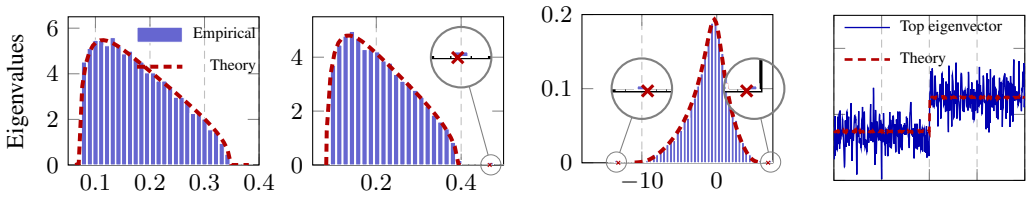
ii) We extend the results in [52, 48, 41, 46] to G-GLMs by considering *generic* data statistics and loss function, whereas in [52, 48, 41, 46] only much more homogeneous models are discussed, and sometimes under unrealistic assumption, e.g., the Hessian can be decomposed as the sum of the two freely independent matrices, the residual error, data feature, and weights are all composed of i.i.d. zero mean normal random variables [52, 53]).

iii) Instead of focusing solely on the main eigenvalue bulk as in [52, 53], our results also shed novel light on the isolated eigenvalues (above and/or below the bulk) that are empirically observed in the Hessian of modern NNs [55, 21, 40, 48, 47], as well as on the associated eigenvectors that are shown closely connected to NN training dynamics [27]. Also, relative to [52, 53], we show *qualitatively* different behaviors for the Hessian eigenspectra, e.g., bounded versus unbounded support, single- versus multi-bulk as in Figure 2. To our knowledge, these are *not* covered in the existing Hessian literature.

1.3 Related work

Here, we provide a brief review of related previous efforts, see more discussions in the appendix.

Random matrix theory. Random matrices of the type (4) are related to the *separable covariance model* [74, 16] in the RMT literature, which is of the form $\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\mathbf{D}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}}$, for random $\mathbf{Z}$ and $\mathbf{C}, \mathbf{D}$



(a) Logistic, $\mathbf{w} = \mathbf{w}^* = \boldsymbol{\mu}$ (b) Logistic, indep. $\mathbf{w}$ (c) Phase retrieval model (d) Eigenvector in Fig 1c

Figure 1: Illustration of our main results: eigenspectral properties of the Hessian of G-GLMs with $p = 800$, $n = 6\,000$ and $\mathbf{C} = \mathbf{I}_p$. **Fig 1a versus 1b**: absence versus presence of a right-hand side spike for different choices of $\mathbf{w}$, logistic model (2) with logistic loss, and $\mathbf{w}_* = \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/p)$. **Fig 1c versus 1d**: the Hessian eigenspectra have a rather different shape (as opposed to the Marčenko-Pastur-like in Fig 1a and 1b) for the (non-convex) phase retrieval model (**1c**) and the top eigenvector is known in this case to be a (noisy) estimate of $\mathbf{w}_*$ (**1d**), as confirmed by our theory. With square loss $\ell(y, h) = (y - h^2)^2/4$, $\mathbf{w}_* = [-2 \cdot \mathbf{1}_{p/2};\ 2 \cdot \mathbf{1}_{p/2}]/\sqrt{p}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/p)$ and $\boldsymbol{\mu} = \mathbf{0}$.

independent of $\mathbf{Z}$. Our results generalize this, in the sense that we allow $\mathbf{D}$ to *depend* on $\mathbf{Z}$, in a possibly nonlinear fashion, per (4). This is of direct interest for the Hessian of G-GLMs.

Hessian eigenspectra. The eigenspectra of Hessian matrices arising in ML models (in particular, for NNs) have attracted considerable interest recently [55, 56, 10, 21, 66, 23, 30, 19, 58, 72, 73, 24]. However, these investigations are either limited to empirical evaluation [55, 56] or built upon somewhat unrealistic simplifying assumptions and reduce to the "mixed" behavior of Marčenko–Pastur and semicircle law [52, 14]. In contrast, here we focus on the more tractable example of G-GLM and provide *precise* results on the Hessian eigenspectra for structural feature on arbitrary loss.

Spectral initialization in non-convex problems. A popular initialization scheme (of gradient-based methods) for non-convex problems is the *spectral initialization*, where the top eigenvectors of some Hessian-type matrices are used as gradient descent initialization [7, 35, 32, 38, 1]. In [41], which was generalized in [46], the authors evaluated the eigenspectrum asymptotics of $\frac{1}{n}\sum_{i=1}^n f(y_i)\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$, for some $f : \mathbb{R} \to \mathbb{R}$ and $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Their technical approach is, however, limited to the case of very homogeneous features. Here we generalize the analysis in [41, 46] to the Hessian of G-GLM, by developing a systematic approach to account for both feature structures and loss functions.

Scalable second-order methods. Second order methods are among the most powerful optimization methods that have been designed, and there have been several attempts to use their many advantages for machine learning applications [68, 63, 54], particularly for training NNs [72, 73, 58, 19, 67]. We expect that our precise characterization of the Hessian sheds new light on the understanding and improved design of (e.g., computationally) more efficient second-order methods.

2 Main results

In the section, we present our main results: on the limiting Hessian eigenspectrum (in Sec 2.1); and on the behavior of the (possible) isolated eigenvalue-eigenvector(s) (in Sec 2.2). These two main results depend on a technical deterministic equivalent result for the Hessian resolvent (in Sec 2.3), which is of independent interest. We position ourselves in the following high-dimensional regime.

Assumption 1 (High-dimensional asymptotics)**.** *As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, we have $\max\{\|\mathbf{w}\|, \|\mathbf{w}_*\|\} = O(1)$ and $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with $\max\{\|\boldsymbol{\mu}\|, \|\mathbf{C}\|\} = O(1)$.*

2.1 Limiting spectral measure

Our first result is the limiting Hessian eigenvalue distribution. This is a direct consequence of our main technical Theorem 4 and is proven in Sec A.2 of the appendix.

Theorem 1 (Limiting spectral measure)**.** *Let Assumption 1 hold, we have, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the empirical spectral measure[2] $\mu_\mathbf{H}$ of the Hessian matrix $\mathbf{H}$ defined in* (4)

[2]That is, the normalized counting measure of the eigenvalues $\lambda_i(\mathbf{H})$ of $\mathbf{H}$, i.e., $\mu_\mathbf{H} \equiv \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(\mathbf{H})}$.

converges weakly and almost surely to a probability measure μ, defined through its Stieltjes transform $m(z) = \int (t-z)^{-1}\mu(dt)$ as the unique solution to[3]

$$m(z) = \tfrac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}_b(z), \quad \delta(z) = \tfrac{1}{n}\operatorname{tr}\left(\mathbf{C}\bar{\mathbf{Q}}_b(z)\right), \quad \bar{\mathbf{Q}}_b^{-1}(z) \equiv \mathbb{E}\left[\tfrac{g\cdot\mathbf{C}}{1+g\cdot\delta(z)}\right] - z\mathbf{I}_p, \tag{5}$$

where

$$g \equiv \partial^2\ell(y,h)/\partial h^2, \quad \textit{for } h = \mathbf{w}^\mathsf{T}\mathbf{x} \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w}), \tag{6}$$

and y and ℓ defined respectively in (1) *and* (3)*. Moreover, if we denote ν the law of g and assume the empirical spectral measure of $\mathbf{C}$ converges to $\tilde{\nu}$ as $p \to \infty$, then* (5) *can be compactly written as*

$$m(z) = \int \left(-z + \tilde{t}\int \frac{t}{1+t\delta(z)}\,\nu(dt)\right)^{-1}\tilde{\nu}(d\tilde{t}), \quad \delta(z) = \int \frac{c\tilde{t}}{-z+\tilde{t}\int \frac{t}{1+t\delta(z)}\,\nu(dt)}\,\tilde{\nu}(d\tilde{t}). \tag{7}$$

In the form of (7), the (Stieltjes transform of the) limiting Hessian spectral measure μ is determined by the ratio $c = \lim p/n$ and the two measures ν and $\tilde{\nu}$. This formulation is closely connected to the separable covariance model [36, 5, 50, 16, 69] in RMT. Moreover, if $\nu(dt) = \delta_1(t)$ is a Dirac mass at one, this reduces to the popular sample covariance model [59]; taking further $\tilde{\nu}(dt) = \delta_1(t)$ gives the Marčenko-Pastur law. See Sec 3.1 for numerical evaluations of these special cases. In particular, the support of the (limiting) Hessian spectrum μ is directly linked to that of ν and $\tilde{\nu}$.

Remark 1 (Hessian eigen-support)**.** *Under Assumption 1, the (limiting) spectral measure $\tilde{\nu}$ of $\mathbf{C}$ has bounded support. However, this may not be the case for ν, the law of g defined in* (6)*. Since the Hessian eigenvalue distribution μ is of compact support* if and only if *both ν and $\tilde{\nu}$ have compact support [16, Porposition 3.4], μ may be of unbounded support, depending on the model and the loss.*

An example of unbounded μ is the phase retrieval model with $y = (\mathbf{w}_*^\mathsf{T}\mathbf{x})^2$ and square loss $\ell(y,h) = (y-h^2)^2/4$, for which we have $g = 3(\mathbf{w}^\mathsf{T}\mathbf{x})^2 - (\mathbf{w}_*^\mathsf{T}\mathbf{x})^2$ for $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$. As a consequence, with say $\mathbf{w}_* = \mathbf{w}$, g follows a chi-square distribution with one degree of freedom and has thus *unbounded* support. This corresponds to Fig 1c, where the Hessian spectrum has a "heavier" tail compared to Fig 1a (logistic model), and the empirically observed "isolated" eigenvalue is due to a finite-dimensional effect and will be "buried" in the noisy main bulk for larger values of n, p. Therefore, aiming for an (almost surely) isolated eigenvalue-eigenvector (e.g., to recover the model parameter $\mathbf{w}_*$ using the top Hessian eigenvector), some preprocessing function f must be applied. This has been discussed in previous work [41, 46] and corresponds to the so-called trimming strategy in phase retrieval [11], with for instance the truncation function $f(t) = \delta_{|t|\le\epsilon}$ for some $\epsilon > 0$.

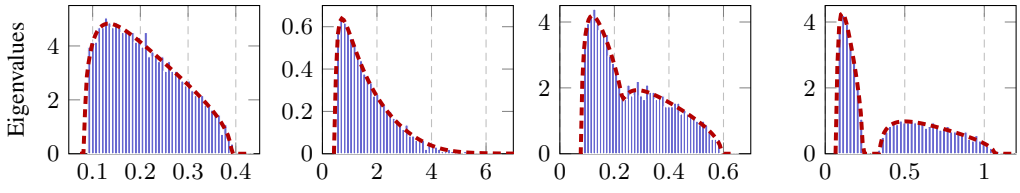
Another example of unbounded μ is when the exponential loss [22] is used. Precisely, consider the logistic model (2) with $\ell(y,h) = \exp(-yh)$, we have that $g = \exp(-yh)$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ which follows a log-normal distribution and has unbounded support. As such, the (limiting) Hessian eigenvalue distribution μ has also unbounded support. On the other hand, with logistic loss $\ell(y,h) = \ln(1+e^{-yh})$, one has $g \le 1/4$ and μ is guaranteed to have bounded support. In Fig 2a and 2b, the empirical Hessian eigenvalues and the limiting distributions are compared for logistic and exponential losses, with a more "heavy-tailed" behavior observed for the exponential loss.

Clearly, depending on the measures ν (of g, which depends on feature statistics, loss and underlying model) and $\tilde{\nu}$ (of spectrum of feature covariance $\mathbf{C}$), the Hessian spectrum can have very different forms. Here we compare the empirical Hessian eigenvalues with their limiting behaviors per Theorem 1 for different feature covariance structures[4]. In particular, one may observe a single main bulk with more "compact" Hessian spectrum as in Fig 2c or multiple bulks (two in the case of Fig 2d) with Hessian eigenvalues more "spread-out", depending on the feature covariance structure $\tilde{\nu}$. In the form of (7), the condition for the existence of multi-bulk eigenspectrum has been throughly discussed in [16, Sec 3.2–3.4] and can be numerically evaluated with ease.

As a side remark, the "multi-bulk" behavior similar to Fig 2d has been empirically observed in Hessians of NNs in [40, 48] and is believed to be due to the classification structure within data (i.e., the data vectors are drawn from a mixture of distributions). Here, we provide an alternative explanation via feature covariance structure that can be observed beyond the classification setting.

[3] Uniqueness is ensured in such a way that $\Im[m(z)]\cdot\Im[z] > 0$ for $\Im[z] \neq 0$ and $zm(z) < 0$ for $\Im[z] = 0$, so that $(z, m(z))$ is a valid Stieltjes transform couple, see more details in [28].

[4] Covariance describes the joint variability or the "correlation" between entries of the feature vector and it is of particular significance in the analysis of image (with local structure) and time series data.



(a) Logistic loss (b) Exponential loss (c) $\tilde{\nu} = \frac{1}{2}(\delta_1 + \delta_2)$ (d) $\tilde{\nu} = \frac{1}{2}(\delta_1 + \delta_4)$

Figure 2: Impact of loss function: bounded (**2a**, with logistic loss) versus unbounded (**2b**, with exponential loss) Hessian eigenvalues, with $p = 800$, $n = 6\,000$, logistic model in (2) with $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{C} = \mathbf{I}_p$, $\mathbf{w}_* = \mathbf{0}$ and $\mathbf{w} = [-\mathbf{1}_{p/2},\ \mathbf{1}_{p/2}]/\sqrt{p}$. Impact of feature covariance: Hessian spectrum of single- (**2c**, with $\mathbf{C} = \mathrm{diag}[\mathbf{1}_{p/2};\ 2 \cdot \mathbf{1}_{p/2}]$) versus multi-bulk (**2d**, with $\mathbf{C} = \mathrm{diag}[\mathbf{1}_{p/2};\ 4 \cdot \mathbf{1}_{p/2}]$), with $p = 800$, $n = 6\,000$, logistic model with $\mathbf{w}^* = \mathbf{0}_p$, $\mathbf{w} = \boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/p)$.

2.2 Isolated eigenvalues and eigenvectors

As discussed in Remark 1, under Assumption 1, the Hessian has bounded (limiting) eigen-support if and only if ν, the law of g, has bounded support. Under this condition (or, after the application of some function f so that $f(g)$ is bounded), we can then talk about the (possible) isolated Hessian eigenvalues, as in the following result, the proof of which is given in Sec A.3 of the appendix.

Theorem 2 (Isolated eigenvalues)**.** *In the setting of Theorem 1, assume that the law ν of the random variable g defined in* (6) *is of bounded support, define*

$$\mathbf{G}(z) = \mathbf{I}_3 + \boldsymbol{\Lambda}(z)\mathbf{V}^\mathsf{T}\bar{\mathbf{Q}}_b(z)\mathbf{V} \in \mathbb{R}^{3\times 3}, \tag{8}$$

with $\bar{\mathbf{Q}}_b(z), \delta(z)$ defined in (5)*, $\mathbf{V} \equiv [\boldsymbol{\mu},\ \mathbf{C}\mathbf{w}_*,\ \mathbf{C}\mathbf{w}] \in \mathbb{R}^{p\times 3}$, $\mathbf{U} \equiv \mathbf{C}^{\frac{1}{2}}[\mathbf{w}_*,\ \mathbf{w}] \in \mathbb{R}^{p\times 2}$ and*

$$\boldsymbol{\Lambda}(z) \equiv \mathbb{E}\frac{g}{1+g\cdot\delta(z)}\begin{bmatrix} 1 & (\mathbf{U}^+\mathbf{z})^\mathsf{T} \\ \mathbf{U}^+\mathbf{z} & \mathbf{U}^+\mathbf{z}(\mathbf{U}^+\mathbf{z})^\mathsf{T} - (\mathbf{U}^\mathsf{T}\mathbf{U})^+ \end{bmatrix}, \quad \mathbf{z} = \mathbf{C}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \tag{9}$$

where we denote $\mathbf{U}^+$ the Moore–Penrose pseudoinverse of $\mathbf{U}$. Then, for λ such that $\mathbf{G}(\lambda)$ has a zero eigenvalue (of multiplicity one), there exists an eigenvalue $\hat{\lambda}$ of $\mathbf{H}$ such that $\hat{\lambda} - \lambda \xrightarrow{a.s.} 0$.

Theorem 2 provides an asymptotic characterization of the possible isolated Hessian eigenvalues by computing the determinant of the much smaller (three-by-three) deterministic matrix $\mathbf{G}$ closely related to the key quantity $\delta(z)$ defined in Theorem 1. Note that, Theorem 2 does not provide, at least explicitly, the *phase transition* condition under which these spikes become "isolated" from the main bulk. As we shall see in more details in Sec 3.2, two types of quantitatively different phase transitions can be characterized, due to the data "signal" $\boldsymbol{\mu}$ and the underlying model, respectively.

We can also analyze the associated isolated eigenvectors. First note that, in the infinite data regime (i.e., for $n \to \infty$ with p fixed), we have, by the strong law of large numbers, that $\mathbf{H} \xrightarrow{a.s.} \mathbb{E}[\mathbf{H}]$, with

$$\mathbb{E}[\mathbf{H}] = \mathbb{E}[\ell''(y, \mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x}\mathbf{x}^\mathsf{T}] = \mathbb{E}[g]\cdot\mathbf{C} + \mathbf{V}\begin{bmatrix} 1 & \mathbb{E}[g\cdot\mathbf{U}^+\mathbf{z}]^\mathsf{T} \\ \mathbb{E}[g\cdot\mathbf{U}^+\mathbf{z}] & \mathbf{U}^+\mathbb{E}[g\cdot(\mathbf{z}\mathbf{z}^\mathsf{T} - \mathbf{I}_p)](\mathbf{U}^+)^\mathsf{T} \end{bmatrix}\mathbf{V}^\mathsf{T}.$$

As a consequence, it is expected that in the large $n, p \to \infty$ limit, the top eigenvectors of $\mathbf{H}$ could also be related to the columns of $\mathbf{V}$. This is the case in Fig 1d, where the top eigenvector is observed to be a "noisy" version of the model parameter $\mathbf{w}_*$. More precisely, for $(\hat{\lambda}, \hat{\mathbf{u}})$ an isolated eigenpair of $\mathbf{H}$, the projection $\mathbf{V}^\mathsf{T}\hat{\mathbf{u}}\hat{\mathbf{u}}^\mathsf{T}\mathbf{V} \in \mathbb{R}^{3\times 3}$ can be shown to be asymptotically close to a deterministic matrix. This measures the "cosine-similarly" between the Hessian isolated eigenvector $\hat{\mathbf{u}}$ with any column of $\mathbf{V}$ and consequently the performance of using $\hat{\mathbf{u}}$ as an estimate of, for instance the model parameter $\mathbf{w}_*$ for $\mathbf{C} = \mathbf{I}_p$. This result is given in the following theorem, which is proven in Appendix A.3.

Theorem 3 (Isolated eigenvectors)**.** *In the setting of Theorem 2, for an isolated eigenvalue-eigenvector pair $(\hat{\lambda}, \hat{\mathbf{u}})$ of $\mathbf{H}$ and λ the asymptotic position (of $\hat{\lambda}$) given in Theorem 2, then*

$$\mathbf{V}^\mathsf{T}\hat{\mathbf{u}}\hat{\mathbf{u}}^\mathsf{T}\mathbf{V} = -\mathbf{V}^\mathsf{T}\bar{\mathbf{Q}}_b(\lambda)\mathbf{V}\cdot\Xi(\lambda) + o(1), \quad \Xi(\lambda) = (\mathbf{v}_{l,\mathbf{G}}^\mathsf{T}\mathbf{G}'(\lambda)\mathbf{v}_{r,\mathbf{G}})^{-1}\cdot\mathbf{v}_{r,\mathbf{G}}\mathbf{v}_{l,\mathbf{G}}^\mathsf{T},$$

for $\bar{\mathbf{Q}}_b(z)$ and $\mathbf{G}(z)$ defined in (5) *and* (8)*, respectively, $\mathbf{v}_{l,\mathbf{G}}, \mathbf{v}_{r,\mathbf{G}} \in \mathbb{R}^3$ the left and right eigenvectors of $\mathbf{G}(\lambda)$ associated with eigenvalue zero, and $\mathbf{G}'(\lambda)$ the derivative of $\mathbf{G}(z)$ with respect to z evaluated at $z = \lambda$.*

2.3 Technical tool: deterministic equivalent

Our main technical tool to derive Theorem 1, 2 and 3 is a so-called deterministic equivalent [28, 15] result for the Hessian resolvent $\mathbf{Q}(z) = (\mathbf{H} - z\mathbf{I}_p)^{-1}$, that provides simultaneous access to the Hessian limiting eigenvalue distribution and the behavior of the possible isolated eigenpairs. Precisely, the normalized trace $\operatorname{tr} \mathbf{Q}(z)/p$ gives the Stieltjes transform $m_{\mathbf{H}}(z) = \int (t-z)^{-1} \mu_{\mathbf{H}}(dt)$ of the empirical spectral measure $\mu_{\mathbf{H}}$ of $\mathbf{H}$, from which $\mu_{\mathbf{H}}$ can be recovered. Also, for $(\hat{\lambda}, \hat{\mathbf{u}})$ an eigenpair of interest, with Cauchy's integral formula we have $|\mathbf{w}^\mathsf{T} \hat{\mathbf{u}}|^2 = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} \mathbf{w}^\mathsf{T} \mathbf{Q}(z) \mathbf{w} \, dz$, for a deterministic vector $\mathbf{w} \in \mathbb{R}^p$ and Γ_λ a positively oriented contour surrounding *only* $\hat{\lambda}$. As such, for $\bar{\mathbf{Q}}(z)$ a deterministic equivalent of $\mathbf{Q}(z)$, that is, $\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z)$ with $\operatorname{tr} \mathbf{A}(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z))/p \to 0$ and $\mathbf{a}^\mathsf{T}(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z))\mathbf{b} \to 0$ almost surely as $n, p \to \infty$, for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of bounded (Euclidean and spectral) norms, the limiting spectral measure (via the associated Stieltjes transform) and the isolated eigenpairs of $\mathbf{H}$ are directly accessible via the study of the deterministic equivalent $\bar{\mathbf{Q}}(z)$. This result is given as follows, with the proof deferred to Sec A.1 in the appendix.

Theorem 4 (Deterministic equivalent)**.** *Let* $\mathbf{Q}(z) \equiv (\mathbf{H} - z\mathbf{I}_p)^{-1}$ *be the resolvent of* $\mathbf{H}$ *defined in* (4)*. Then, under Assumption 1, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$*,*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \text{with } \bar{\mathbf{Q}}^{-1}(z) = \mathbb{E}\left[\tfrac{g}{1+g\cdot\delta(z)} (\mathbf{C}^{\frac{1}{2}} (\mathbf{I}_p - \mathbf{P}_{\mathbf{U}}) \mathbf{C}^{\frac{1}{2}} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T}) \right] - z\mathbf{I}_p,$$

for random vector $\boldsymbol{\alpha} \equiv \boldsymbol{\mu} + \mathbf{C}^{\frac{1}{2}} \mathbf{P}_{\mathbf{U}} \mathbf{z} \in \mathbb{R}^p$ *and* $g = \ell''(y, \mathbf{w}^\mathsf{T}\boldsymbol{\mu} + \mathbf{w}^\mathsf{T}\mathbf{C}^{\frac{1}{2}}\mathbf{z})$ *for* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ *as defined in* (6)*,* y *and* $\delta(z)$ *defined in* (1) *and* (5)*, respectively, and* $\mathbf{P}_{\mathbf{U}} \in \mathbb{R}^{p \times p}$ *the projection onto the subspace spanned by the columns of* $\mathbf{U} \equiv \mathbf{C}^{\frac{1}{2}} [\mathbf{w}_*, \ \mathbf{w}]$*.*

3 Evaluations and Discussions

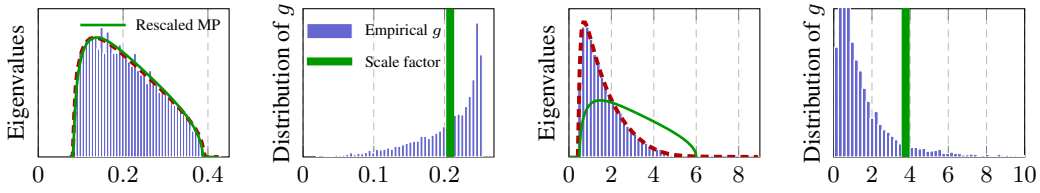
In this section, we provide further discussions on the consequences of Theorem 1, 2 and 3, together with numerical evaluations. Implications of Theorem 1 on the Hessian eigenvalue distribution is discussed in Sec 3.1. In Sec 3.2, we discuss the consequences of Theorem 2 and 3 on the possible isolated eigenpairs, for which two fundamentally different phase transitions are characterized.

3.1 Hessian eigenvalues distribution

For a better interpretation of Theorem 1 on the Hessian eigenspectrum, we consider here the special case of $\mathbf{C} = \mathbf{I}_p$, and start with the simple setting where the random variable g in (6) is constant, say $g = 1$: this happens, e.g., when the square loss $\ell(y, h) = (y - h)^2/2$ is employed. In this case, the Hessian does *not* dependent on $\mathbf{w}, \mathbf{w}_*$ and the Stieltjes transform $m(z)$ is the solution to $zcm^2(z) - (1 - c - z)m(z) + 1 = 0$ and corresponds to the Marčenko-Pastur law.

As long as g is *not* constant, the limiting Hessian spectrum is, a priori, different from the Marčenko-Pastur law, even in the $\mathbf{C} = \mathbf{I}_p$ setting, since the associated Stieltjes transform $m(z)$ is different from the solution to the Marčenko-Pastur equation. However, we see in Fig 3a that, for the logistic model (2) with logistic loss, the Hessian spectrum is close, at least visually, to a (rescaled) Marčenko-Pastur law. This can be understood with Theorem 1 and is due to the fact that, the distribution of g is more "concentrated" (around some constant, see Fig 3b versus 3d for a comparison between different cases). This is in sharp contrast to Fig 3c where with the exponential loss, the law of g has a much larger spread and the Hessian is therefore away from a Marčenko-Pastur-shape.

This "empirical fit" has been observed in [51, Fig 5], where acceleration methods proposed for a Marčenko-Pastur distributed Hessian (in linear least squares) work reasonably well on logistic regression models. Our theory proposes a convincing theoretical explanation of this empirical observation on logistic regression, and possibly for others more involved ML models. Nonetheless, it must be pointed out that this "visual approximation" by Marčenko-Pastur law is *not robust*, in the sense that it "visually" holds only for, yet formally different from, the case of (i) logistic model with (ii) logistic loss and (iii) identity covariance $\mathbf{C} = \mathbf{I}_p$: any change in the response model (e.g., the phase retrieval model in Fig 1c), in the choice of loss function (e.g., the exponential loss in Fig 2b), or beyond the identity covariance setting (as in Fig 2c and 2d) would induce a Hessian spectrum that is very different from the Marčenko-Pastur law. In this vein, our Theorem 1 goes beyond such "loose" Marčenko-Pastur approximation and acts as a more accurate first example in the understanding



(a) Logistic: Hessian (b) Logistic: g_is (c) Exp: Hessian (d) Exp: g_is

Figure 3: Comparison of Hessian eigenspectra with (rescaled and shifted) Marčenko-Pastur laws in the setting of Fig 2. **Fig 3a versus 3b**: Marčenko-Pastur-like Hessian with logistic loss, Hessian eigenvalues (**3a**) and empirical distribution of the g_is versus the scaling factor (**3b**, empirically obtained by matching the minimal and maximal empirical Hessian eigenvalues to the Marčenko-Pastur law). **Fig 3c versus 3d**: an example of non-Marčenko-Pastur-like Hessian with exponential loss and the associated g_is. Note that the scales of the axes are different in different subfigures.

of Hessian in more involved ML models beyond linear least squares that accounts for nonlinear transformations (such as activation function in NNs) and feature statistics.

While Theorem 1 is proven here only for Gaussian features, we conjecture, as is the case for many random matrix asymptotics, that it holds more generally beyond Gaussian distribution, see Fig 5 in the appendix for more discussions on this point.

3.2 Isolated eigenvalues and their phase transitions

In this section, we discuss the implications of Theorem 2 and 3 on the possible isolated eigenvalue-eigenvector pairs. More precisely, we show that, different from the classical spiked models extensively studied in RMT literature [2, 4, 39], for which (i) the isolated spike appears due to the presence of some statistical "signal" in the data and (ii) a "monotonic" phase transition behavior can be characterized as a function of the signal strength; here another type of Hessian spike arises due to the underlying G-GLM model (i.e., $\mathbf{w}_*$ and $\mathbf{w}$) and exhibits a rather different behavior.

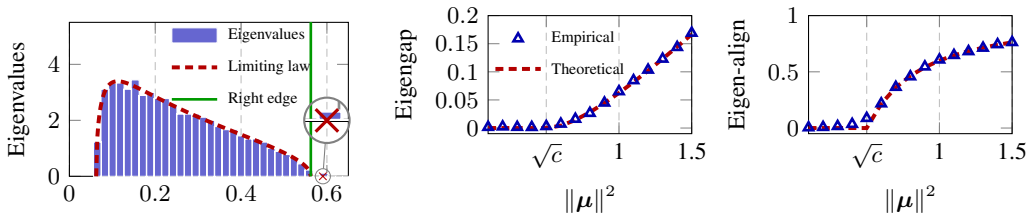


Figure 4: Spike due to data signal in Corollary 1: theory versus practice of **(left)** Hessian eigenspectrum with $\|\boldsymbol{\mu}\|^2 = 0.8$, **(middle)** eigengap $\mathrm{dist}(\lambda_{\boldsymbol{\mu}}, \mathrm{supp}(\mu))$, and **(right)** top eigenvector alignment α in (10), as a function of the signal strength $\|\boldsymbol{\mu}\|^2$, on logistic model with logistic loss, for $\boldsymbol{\mu} \propto [-\mathbf{1}_{p/2}, \mathbf{1}_{p/2}]$, $\mathbf{w} = \mathbf{w}_* = \mathbf{0}$, $\mathbf{C} = \mathbf{I}_p$, $p = 512$ and $n = 2\,048$. Results averaged over 50 runs.

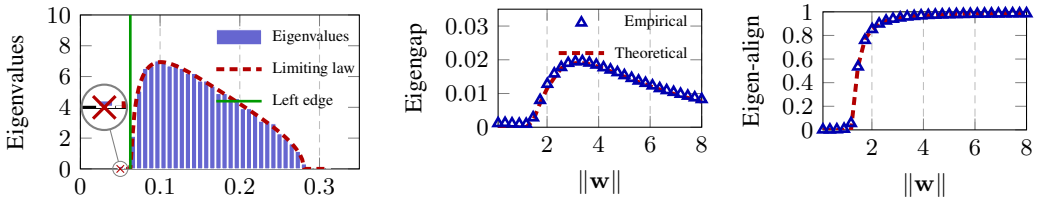


Figure 5: Left-hand side spike due to response model in Corollary 2 in the *absence of data signal*: **(left)** Hessian spectrum for $\|\mathbf{w}\| = 2$, with a *left* isolated eigenvalue $\hat{\lambda}_{\mathbf{w}}$, **(middle)** eigengap $\mathrm{dist}(\lambda_{\mathbf{w}}, \mathrm{supp}(\mu))$, and **(right)** dominant eigenvector alignment (with $\mathbf{w}$), as a function $\|\mathbf{w}\|$ with $\mathbf{w} \propto [-\mathbf{1}_{p/2}, \mathbf{1}_{p/2}]$, $\mathbf{w}_* = \boldsymbol{\mu} = \mathbf{0}$, $\mathbf{C} = \mathbf{I}_p$, $p = 800$ and $n = 8\,000$. Results averaged over 50 runs.

3.2.1 Spike due to data signal

To study the spike due to data "signal" $\boldsymbol{\mu}$ and its phase transition behavior, we focus here on the case $\mathbf{w}_* = \mathbf{w} = \mathbf{0}$. This, in the case of logistic model (2) for example, gives rise to a much simpler form of limiting spectrum (per Theorem 1) and possible isolated eigenpairs (per Theorem 2 and 3), as summarized in the following corollary, with detailed derivations given in Sec B.3 of the appendix.

Corollary 1 (Spike due to data signal: logistic model)**.** *Consider the logistic model in* (2) *with logistic loss, for* $\mathbf{w} = \mathbf{w}_* = \mathbf{0}$ *and* $\mathbf{C} = \mathbf{I}_p$*, the limiting Hessian eigenvalue distribution is the Marčenko-Pastur law, but rescaled by a factor of* $g = 1/4$*. Moreover, there is* at most one *isolated eigenpair* $(\hat{\lambda}_{\boldsymbol{\mu}}, \hat{\mathbf{u}}_{\boldsymbol{\mu}})$ *of* $\mathbf{H}$ *and it satisfies*

$$\hat{\lambda}_{\boldsymbol{\mu}} \xrightarrow{a.s.} \begin{cases} \lambda_{\boldsymbol{\mu}} = \frac{1}{4}(1+\rho+c\cdot\frac{\rho+1}{\rho}) & \rho > \sqrt{c}, \\ \frac{1}{4}(1+\sqrt{c})^2 & \rho \le \sqrt{c}; \end{cases}, \quad \frac{|\boldsymbol{\mu}^\mathsf{T}\hat{\mathbf{u}}_{\boldsymbol{\mu}}|^2}{\|\boldsymbol{\mu}\|^2} \xrightarrow{a.s.} \begin{cases} \alpha = \frac{\rho^2-c}{\rho^2+c\rho} & \rho > \sqrt{c}, \\ 0 & \rho \le \sqrt{c}; \end{cases} \tag{10}$$

with the signal strength $\rho = \lim_{p\to\infty} \|\boldsymbol{\mu}\|^2$ *and* $c = \lim p/n$*.*

The behavior of the isolated eigen-pairs described in Corollary 1 follows the "classical" phase transition [3, 2, 49]: (i) the isolated eigenvalue always appears on the right-hand side of the main (Marčenko-Pastur) bulk and (ii) the eigenvalue amplitude and eigenvector alignment is "monotonic" with respect to the signal strength $\|\boldsymbol{\mu}\|^2$ in the sense that, for a fixed dimension ratio c, the largest Hessian eigenvalue is bound to become asymptotically isolated once $\|\boldsymbol{\mu}\|^2$ exceeds $\sqrt{c}$ and its value, as well as the eigenvector alignment, increase monotonically as $\|\boldsymbol{\mu}\|^2$ grows. This is confirmed in Fig 4. As we shall see below, this is *not* the case for, e.g., the spike due to model parameter $\mathbf{w}$.

3.2.2 Spike due to model

To investigate the spike due to the underlying model (i.e., $\mathbf{w}_*$ and $\mathbf{w}$), we position ourselves in the situation where $\boldsymbol{\mu} = \mathbf{0}$, that is, in the absence of data "signal". This leads to the following corollary, the proof of which is given in Sec B.4 in the appendix.

Corollary 2 (Spike due to model: logistic model)**.** *Consider the logistic model in* (2) *with logistic loss,* $\boldsymbol{\mu} = \mathbf{0}$ *and* $\mathbf{C} = \mathbf{I}_p$*, then the Stieltjes transform* $m(z)$ *satisfies* $m(z) = 1/(\mathbb{E}[f(r,z)] - z)$ *for* $f(r,z) = 1/(cm(z) + 2 + e^{-r} + e^{r})$ *and* $r \sim \mathcal{N}(0, \|\mathbf{w}\|^2)$ *that* depends *on* $\mathbf{w}$ *but* not *on* $\mathbf{w}_*$*. Moreover, there is* at most one *isolated eigenvalue* $\hat{\lambda}_{\mathbf{w}}$ *of* $\mathbf{H}$ *that is due to* $\mathbf{w}$ *and satisfies* $\hat{\lambda}_{\mathbf{w}} - \lambda_{\mathbf{w}} \xrightarrow{a.s.} 0$ *with* $\lambda_{\mathbf{w}}$ *solution to* $0 = \det \mathbf{G}(\lambda_{\mathbf{w}}) = 1 + m(\lambda_{\mathbf{w}})\frac{\mathbb{E}[f(r,\lambda_{\mathbf{w}})(r^2 - \|\mathbf{w}\|^2)]}{\|\mathbf{w}\|^2}$.

The situation here is more subtle (than the spike due to data signal discussed in Sec 3.2.1): as the model parameter $\mathbf{w}$ changes (e.g., as the "energy" $\|\mathbf{w}\|$ grows), both the Hessian (limiting) eigenvalue distribution and the possible spike location are impacted. Fig 5 illustrates the behavior of the spike due to $\mathbf{w}$ in the setting of Corollary 2. Note first that, different from the case of spike due to data signal $\boldsymbol{\mu}$, the spike in Fig 5-(left) appears on the *left-hand side* of the main bulk: this particularly means that the Hessian may admit an eigenvalue that is *significantly smaller* than all the other eigenvalues.[5] Also, note from Fig 5-(middle) that, different from the spike due to $\boldsymbol{\mu}$, the spike due to $\mathbf{w}$ exhibits here a "non-monotonic" behavior in the sense that, it is absent for small values of $\|\mathbf{w}\|$ (as for small $\|\boldsymbol{\mu}\|$ in Fig 4-middle) and becomes "isolated" as $\|\mathbf{w}\|$ increases, but then again "merges into" the main bulk as $\|\mathbf{w}\|$ continues to increase, resulting an eigengap that falls back to zero.

It is perhaps even more surprising to observe in Fig 5-(right) that, the alignment between the associated isolated eigenvector and the parameter $\mathbf{w}$ is, unlike the eigengap in Fig 5-(middle), monotonically increasing as $\|\mathbf{w}\|$ grows large, as in the case of Fig 4-(right). This suggests that, in the case of spike due to model, a *smaller* eigengap may not always imply *less* statistical "information" contained in the associated eigenvector, which somehow goes against the conventional eigengap heuristic [61, 33]. It is worthy mentioning that, while, technically speaking, the proposed analysis is not capable of charactering the behavior as $\|\mathbf{w}\| \to \infty$ under Assumption 1, empirical results suggest that for extremely large $\|\mathbf{w}\|$, the eigengap tends to vanish, the associated "dominant" eigenvector can still be used to recover $\mathbf{w}$ almost perfectly, see Fig 8 in the appendix as an example.

[5] Depending on the response model and loss, the spike due to model may also appear on the right-hand side of the bulk or even establish a left-to-right transition. The eigenvector alignment can behave differently from Fig 4-5 and establish a non-monotonic behavior as a function of $\|\mathbf{w}\|$ or $\|\mathbf{w}_*\|$; see Sec 4.3 in the appendix.

References

[1] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *Journal of Machine Learning Research*, 40(2015), 2015.

[2] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[3] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[4] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[5] Zdzisław Burda, Jerzy Jurkiewicz, and Bartłomiej Wacław. Spectral moments of correlated wishart matrices. *Physical Review E*, 71(2):026111, 2005.

[6] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

[7] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[8] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[9] Olivier Chapelle, Choon Teo, Quoc Le, and Alex Smola. Tighter bounds for structured estimation. *Advances in neural information processing systems*, 21:281–288, 2008.

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

[11] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[12] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2018.

[13] Dennis Child. *The essentials of factor analysis*. Cassell Educational, 1990.

[14] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204, 2015.

[15] Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[16] Romain Couillet and Walid Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

[17] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 27 of *NIPS'14*, pages 2933–2941. Curran Associates, Inc., 2014.

[18] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. CRC press, 2018.

[19] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 293–302, 2019.

[20] James R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758, 1982.

[21] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

[22] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[23] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

[24] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241, 2019.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[26] Diego Granziol. Beyond random matrix theory for deep networks. *arXiv preprint arXiv:2006.07721*, 2020.

[27] Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace, 2018.

[28] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[30] Arthur Jacot, Franck Gabriel, and Clement Hongler. The asymptotic spectrum of the hessian of dnn throughout training. In *International Conference on Learning Representations*, 2019.

[31] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[32] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

[33] Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

[34] Kenji Kawaguchi. Deep Learning without Poor Local Minima. In *Advances in Neural Information Processing Systems*, volume 29 of *NIPS'16*, pages 586–594. Curran Associates, Inc., 2016.

[35] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[36] A Khorunzhy. Eigenvalue distribution of large random matrices with correlated entries. *Mat. Fiz. Anal. Geom*, 3(1-2):80–101, 1996.

[37] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1-2):311–337, 2019.

[38] Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2016.

[39] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse pca. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1635–1639. IEEE, 2015.

[40] Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sSGD for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2020.

[41] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 2019.

[42] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[43] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. Technical Report Preprint: arXiv:1810.01075, 2018.

[44] C. H. Martin, T. S. Peng, and M. W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Technical Report Preprint: arXiv:2002.06716, 2020.

[45] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[46] Marco Mondelli and Andrea Montanari. Fundamental Limits of Weak Recovery with Applications to Phase Retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, 2019.

[47] Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5012–5021. PMLR, 09–15 Jun 2019.

[48] Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

[49] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[50] Debashis Paul and Jack W Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

[51] Fabian Pedregosa and Damien Scieur. Average-case acceleration through spectral density estimation. *arXiv preprint arXiv:2002.04756*, 2020.

[52] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2798–2806. JMLR. org, 2017.

[53] Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. *Advances in Neural Information Processing Systems*, 31:5410–5419, 2018.

[54] Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174(1-2):293–326, 2019.

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

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

[57] Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. *arXiv preprint arXiv:2012.03801*, 2020.

[58] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. In *AAAI*, pages 8815–8821, 2020.

[59] Jack W Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

[60] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2000.

[61] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[62] Martin J. Wainwright. *High-Dimensional Statistics: : A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[63] Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W. Mahoney. GIANT: Globally Improved Approximate Newton Method for Distributed Optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 2332–2342. Curran Associates, Inc., 2018.

[64] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

[65] Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

[66] Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

[67] Peng Xu and Michael W. Mahoney. Second-order Optimization for Non-convex Machine Learning: an Empirical Study. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 199–207, 2020.

[68] Peng Xu, Jiyan Yang, Farbod Roosta-Khorasani, Christopher Ré, and Michael W. Mahoney. Sub-sampled Newton Methods with Non-uniform Sampling. In *Advances in Neural Information Processing Systems*, volume 29, pages 3000–3008. Curran Associates, Inc., 2016.

[69] Fan Yang. Edge universality of separable covariance matrices. *Electronic Journal of Probability*, 24, 2019.

[70] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. Technical report, 2018. Preprint: arXiv:1802.08241.

[71] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. W. Mahoney. Trust region based adversarial attack on neural networks. Technical report, 2018. Preprint: arXiv:1812.06371.

[72] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. PyHessian: Neural networks through the lens of the Hessian. *arXiv preprint arXiv:1912.07145*, 2019.

[73] Zhewei Yao, Amir Gholami, Sheng Shen, Kurt Keutzer, and Michael W Mahoney. ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning. *arXiv preprint arXiv:2006.00719*, 2020.

[74] Lixin Zhang. Spectral analysis of large dimensional random matrices. *National University of Singapore PHD Thesis*, 2006.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section ??
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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] Theorem 1-3 is stated under the technical Assumption 1, Theorem 2 and 3 are stated under the additional assumption that the measure ν is of bounded support.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work is theoretical.
   (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] In the appendix.
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)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] For the ease of visualization, we only report one-shot results or the empirical means of multiple trials (versus our theoretical predictions).

(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)? [No]

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]