Hessian Eigenspectra of More Realistic Nonlinear Models

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology, China
zhenyu_liao@hust.edu.cn

Michael W. Mahoney

ICSI and Department of Statistics
University of California, Berkeley, USA
mmahoney@stat.berkeley.edu

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 extends 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 \( L(w) \) of some parameters \( w \in \mathbb{R}^p \), the Hessian \( H(w) \in \mathbb{R}^{p \times p} \) is defined as the second derivative of the loss function with respect to the model parameter, i.e., \( H(w) = \partial L(w) / (\partial w \partial w^T) \). When a ML model is being trained, it is common to parameterize that model by \( w \), and then train that model by minimizing some (smooth) loss function \( L(w) \), with the associated Hessian \( H(w) \), e.g., by backpropagating the error to improve \( w \) \[^{26}\]. Alternatively, once a ML model is trained, the Hessian (and the related Fisher information matrix \[^{66}[68] \)) can be examined to identify outliers, perform diagnostics, and/or engage in other sorts of model validation \[^{31}[79][62] \).

For convex problems, the Hessian \( H(w) \) 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 \( w \) in the parameter space is of central

\[^{26}\] Work done at ICSI and Department of Statistics, University of California, Berkeley, USA.
\[^{66}\] We refer the readers to an extended version of this article \[^{43} \] for detailed proofs and more discussions.

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 linear models (G-GLMs) that extends the popular generalized linear model (GLM) [19, 21], 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-GLMs 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. (A similar theory-practice gap has been observed for other NN matrices to which RMT has been applied, perhaps most notably weight matrices [47, 48, 49].) A more precise understanding of the Hessian eigenspectra (and its dependence on the input data structure, the underlying response model and model parameters, as well as the loss function) for more realistic models is needed.

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

$$y_i \sim f(y \mid w_\star^T x_i)$$

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

logistic model: $\mathbb{P}(y = 1 \mid w_\star^T x) = (1 + e^{w_\star^T x})^{-1}$, $y \in \{-1, 1\}$, (2)

covers a large family of models in applications in statistics and ML. Other examples include: (i) the (noisy) nonlinear factor model where $y \sim \mathcal{N}(g(w_\star^T x), \sigma^2)$ for some nonlinear $g : \mathbb{R} \to \mathbb{R}$ and $\sigma > 0$ [14]; (ii) the (noiseless) phase retrieval model with $y = (w_\star^T x)^2$, in which case one wishes to reconstruct $w_\star$ from its (scaled) magnitude measurements [21]; and (iii) the single-layer NN model $y = \sigma(w_\star^T x)$ for some nonlinear activation function $\sigma(t)$ such as the tanh-sigmoid $\sigma(t) = \tanh(t)$.

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

$$\min_w L(w) = \min_w \frac{1}{n} \sum_{i=1}^n \ell(y_i, w^T x_i),$$

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 [31] such as the logistic loss $\ell(y, h) = issue, the signs of the smallest and largest Hessian eigenvalue can be used to test whether a given $w$ 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 fraction 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 [33].

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 [45] or the Wigner’s semicircle law [70]. 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) [57, 58, 15]. A somewhat more realistic setup involves using a so-called spiked model (or a spiked covariance model) [3, 4, 41]. 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. 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 [27] and state-of-the-art NN models [77, 78, 79, 80, 64, 20], 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 [47, 48, 49].) A more precise understanding of the Hessian eigenspectra (and its dependence on the input data structure, the underlying response model and model parameters, as well as the loss function) for more realistic 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 linear models (G-GLMs) that extends the popular generalized linear model (GLM) [19, 21], 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-GLMs 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 $x_i \in \mathbb{R}^p$ and the corresponding response $y_i$, in the sense that there exists some parameters $w_\star \in \mathbb{R}^p$ such that for given $w_\star^T x_i$, the response $y_i$ is independently drawn from

$$y_i \sim f(y \mid w_\star^T x_i)$$

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

logistic model: $\mathbb{P}(y = 1 \mid w_\star^T x) = (1 + e^{w_\star^T x})^{-1}$, $y \in \{-1, 1\}$, (2)

covers a large family of models in applications in statistics and ML. Other examples include: (i) the (noisy) nonlinear factor model where $y \sim \mathcal{N}(g(w_\star^T x), \sigma^2)$ for some nonlinear $g : \mathbb{R} \to \mathbb{R}$ and $\sigma > 0$ [14]; (ii) the (noiseless) phase retrieval model with $y = (w_\star^T x)^2$, in which case one wishes to reconstruct $w_\star$ from its (scaled) magnitude measurements [21]; and (iii) the single-layer NN model $y = \sigma(w_\star^T x)$ for some nonlinear activation function $\sigma(t)$ such as the tanh-sigmoid $\sigma(t) = \tanh(t)$.

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

$$\min_w L(w) = \min_w \frac{1}{n} \sum_{i=1}^n \ell(y_i, w^T x_i),$$

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 [31] such as the logistic loss $\ell(y, h) =

---

3Since the same informative pattern is repeated in each row or column of the matrix.
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
\cite{49, 71, 10}, and can be NP-hard in general (the noiseless phase retrieval model y = (w^∗x)^2 with the square loss \ell(y, h) = (y - h^2)^2 as an example \cite{9}). 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 w^∗, and therefore be used as the initialization of gradient descent methods \cite{9, 37, 34}. 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. 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 \ell 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 Q(z) = (H - z1_p)^{-1}, for z ∈ C not an eigenvalue of H, of the generalized sample covariance [4]

\[ H = H(w) = \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i, w^T x_i)x_i, x_i^T = \frac{1}{n} XD^T X^T, \]  

(4)

for X = [x_1, ..., x_n] ∈ \mathbb{R}^{p×n}, D = \text{diag}\{\ell''(y_i, w^T x_i)\}_{i=1}^{n} ∈ \mathbb{R}^{n×n}, and \ell''(y, h) ≡ \partial^2 \ell(y, h)/\partial h^2, as n, p → ∞ with p/n → c ∈ (0, ∞), under the setting of generic Gaussian feature x_i ~ N(μ, C), for μ ∈ \mathbb{R}^p and positive definite covariance C ∈ \mathbb{R}^{p×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 μ and the other due to w^∗ or 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 9 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 logistic loss \ell(y, h) = \ln(1 + e^{−yh}) (which is the loss of choice within the maximum likelihood framework), for different choices of w in the parameter space. A nontrivial interplay between the response model, feature statistics and the parameter w is reflected by the range of the Hessian eigenvalue support and an additional right-hand spike in Fig 1a as confirmed by our theory. For phase retrieval model y = (w^∗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 w^∗, as predicted by our theory. This is indeed connected to the Hessian-based initialization scheme widely used in non-convex problems \cite{8, 37, 34, 40, 2, 51}.

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 go beyond, e.g., \cite{7}, 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 1b, and more in \cite{43} Sec 4.

(ii) We extend the results in \cite{57, 53, 44, 50} to G-GLMs by considering generic data statistics and loss function, whereas in previous efforts only much more homogeneous models are

\footnote{Here we consider w independent of x_i, y_i. With some additional effort, our results extend to the case where w takes some simple explicit function form of X and y, e.g., w is the least square solution to (X, y).}
discussed, and sometimes under unrealistic assumption, e.g., the Hessian can be decomposed as the sum of two freely independent matrices, and the residual error, data feature, and weights are all composed of i.i.d. zero mean normal random variables [57, 58].

(iii) Instead of focusing solely on the main eigenvalue bulk as in [57, 58], 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 [60, 22, 42, 53, 52], as well as on the associated eigenvectors that are shown closely connected to NN training dynamics [28]. Also, relative to [57, 58], 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.

Figure 1: Illustration of our main results: eigenspectral properties of the Hessian of G-GLMs with $p = 800$, $n = 6000$ and $C = I_p$. (a) Logistic, $w = w^* = \mu$. (b) Logistic, indep. $w$. (c) Phase retrieval model $w_\ast = \mu \sim N(0, I_{p/p})$. (d) Eigenvector in Fig 1c.

1.3 Related work

Here, we provide a brief review of related previous efforts, see more discussions in [43].

Random matrix theory. Random matrices of the type (4) are related to the separable covariance model [81, 17] in the RMT literature, which is of the form $C^\frac{p}{2}ZDZ^T C^\frac{p}{2}$, for $Z$ random and $C, D$ deterministic or random but independent of $Z$. Our results generalize this, by allowing $D$ to depend on $Z$, 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 [60, 61, 11, 22, 24, 20, 64, 79, 80, 25]. However, these investigations are often built upon somewhat unrealistic simplifying assumptions and to the “mixed” behavior of Marčenko–Pastur and semicircle law [57, 15]. 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. Efforts have also been devoted to efficiently computing the eigenspectra of Hessian matrices, which, despite not enabling analytic results, allows for the computation or approximation of Hessian spectra for large-scale ML models such as DNNs, without any assumption on the data or underlying (response) model [30, 1, 79, 74, 25].

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 in [37, 40, 2, 51]. In [44], which was generalized in [50], the authors evaluated the eigenspectrum asymptotics of $\frac{1}{n} \sum_{i=1}^n f(y_i)x_i x_i^T$, for $x_i \sim N(0, I_p)$. Their technical approach is, however, limited to the case of very homogeneous features with $x_i \sim N(0, I_p)$. Here we generalize the analysis in [44, 50] 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 [75, 89, 59], particularly for training NNS [72, 80, 64, 20, 73]. 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.

**Spectra of realistic NN weight matrices.** Recent work demonstrated that the eigenvalue distribution of weight matrices of state-of-the-art DNN models in computer vision and natural language applications exhibit heavy-tailed (i.e., not Marčenko–Pastur) behavior, and that this can be used to define metrics to predict trends in the quality of state-of-the-art NNs without access to training or testing data [47, 48, 46]. It would be of interest to extend our approach to weight matrix analysis.

## 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 \(\max\{\|w\|, \|w_\ast\|\} = O(1)\) and \(x_i \overset{i.i.d.}{\sim} N(\mu, C)\) with \(\max\{\|\mu\|, \|C\|\} = O(1)\).

### 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 [43, Sec A.2].

**Theorem 1** (Limiting spectral measure). Let Assumption 7 hold, we have, as \(n, p \to \infty\) with \(p/n \to c \in (0, \infty)\), the empirical spectral measure \(\mu_{\xi_n}\) of the Hessian matrix \(H\) defined in (4) converges weakly and almost surely to a probability measure \(\mu\), defined through its Stieltjes transform

\[
m(z) = \frac{1}{p} \text{tr} \tilde{Q}_6(z), \quad \delta(z) = \frac{1}{n} \text{tr} \left(C\tilde{Q}_6(z)\right), \quad \tilde{Q}_6^{-1}(z) \equiv E \left[ \frac{g \cdot C}{1 + g \cdot \delta(z)} \right] - zI_p,
\]

where

\[
g \equiv \frac{\partial^2 \ell(y, h)}{\partial h^2}, \quad \text{for } h = w^T x \sim N(w^T \mu, w^T C w),
\]

and \(y\) and \(\ell\) defined respectively in (1) and (3). Moreover, if we denote \(\nu\) the law/distribution of \(g\) and assume the empirical spectral measure of \(C\) converges to \(\tilde{\nu}\) as \(p \to \infty\), then (5) writes

\[
m(z) = \int \left( -z + \tilde{\ell} \int \frac{t}{1 + \tilde{\ell}(z)} \nu(dt) \right)^{-1} \tilde{\nu}(dt), \quad \delta(z) = \int \left( -z + \tilde{\ell} \int \frac{t}{1 + \tilde{\ell}(z)} \nu(dt) \right) \tilde{\nu}(dt).
\]

In the form of (7), the (Stieltjes transform of the) limiting Hessian spectral measure \(\mu\) is determined by the ratio \(c = \lim p/n\) and the two measures \(\nu\) and \(\tilde{\nu}\). This formulation is closely connected to the separable covariance model [33, 32, 35, 17, 76] in RMT. Moreover, if \(\nu(dt) = \delta_1(t)\) is a Dirac mass at one, this reduces to the popular sample covariance model [63]; 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 \(\mu\) is directly linked to that of \(\nu\) and \(\tilde{\nu}\).

**Remark 1** (Hessian eigen-support). Under Assumption 7, the (limiting) spectral measure \(\tilde{\nu}\) of \(C\) has bounded support. However, this may not be the case for \(\nu\), the law of \(g\) defined in (6). Since the Hessian eigenvalue distribution \(\mu\) is of compact support if and only if both \(\nu\) and \(\tilde{\nu}\) have compact support [17, Proposition 3.4], \(\mu\) may be of unbounded support, depending on the model and the loss.

An example of unbounded \(\mu\) is the phase retrieval model with \(y = (w_\ast^T x)^2\) and square loss \(\ell(y, h) = (y - h^2)^2/4\), for which we have \(g = 3(w_\ast^T x)^2 - (w_\ast^T x)^2\) for \(x \sim N(\mu, C)\). As a consequence, with say \(w_\ast = w\), \(g\) 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

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

\[\text{Uniqueness is ensured in such a way that } \Im[m(z)] : \Im[z] > 0 \text{ for } \Im[z] \neq 0 \text{ and } zm(z) < 0 \text{ for } \Im[z] = 0, \text{ so that } (z, m(z)) \text{ is a valid Stieltjes transform couple, see more details in [29]}.\]
As discussed in Remark 1, under Assumption 1, the Hessian has bounded (limiting) eigen-support. As a side remark, the "multi-bulk" behavior similar to Fig 2d has been empirically observed in which follows a log-normal distribution and has unbounded support. As such, the (limiting) Hessian spectrum can have very different forms. Here we compare the empirical Hessian eigenvalues with their limiting behaviors per Theorem 2. In the setting of Theorem 1, assume that the law of the random variable \( g \) (of spectrum of feature covariance \( C \)) describe 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.

Figure 2: Impact of loss function: bounded (2a) with logistic loss versus unbounded (2b) with exponential loss Hessian eigenvalues, with \( p = 800, n = 6000, \) logistic model in (2) with \( \mu = 0, C = I_p, w_* = 0 \) and \( w = [-1_{p/2}, 1_{p/2}] / \sqrt{p} \). Impact of feature covariance: Hessian spectrum of single (2c) with \( C = \text{diag}(1_{p/2}, 2 \cdot 1_{p/2}) \) versus multi-bulk (2d) with \( C = \text{diag}(1_{p/2}, 4 \cdot 1_{p/2}) \), with \( p = 800, n = 6000, \) logistic model with \( w^* = 0_p, w = \mu \sim \mathcal{N}(0, I_p/p) \).

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 \( w_* \) using the top Hessian eigenvector), some pre-processing function \( f \) must be applied. This has been discussed in previous work \([44, 50]\) and corresponds to the so-called trimming strategy in phase retrieval \([12]\), with for instance the truncation function \( f(t) = \delta_{[t]} \leq \epsilon \) for some \( \epsilon > 0 \).

Another example of unbounded \( \mu \) is when the exponential loss \([23]\) is used. Precisely, consider the logistic model \([2]\) with \( \ell(y, h) = \exp(-y h) \) we have that \( g = \exp(-y h) \) for \( h \sim \mathcal{N}(0^T \mu, w^T C w) \) which follows a log-normal distribution and has unbounded support. As such, the (limiting) Hessian eigenvalue distribution \( \mu \) has also unbounded support. On the other hand, with logistic loss \( \ell(y, h) = \ln(1 + e^{-y h}) \), one has \( g \leq 1/4 \) and \( \mu \) 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 \( \nu \) (of \( g \), which depends on feature statistics, loss and underlying model) and \( \nu \) (of spectrum of feature covariance \( C \)), 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. In particular, one may observe a single main bulk with more "compact" Hessian spectrum as in Fig 2a or multiple bulks (two in the case of Fig 2d) with Hessian eigenvalues more "spread-out", depending on the feature covariance structure \( \nu \). In the form of (7), the condition for the existence of multi-bulk eigenspectrum has been thoroughly discussed in [17] 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 \([42, 53]\) 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 holds beyond the classification setting.

### 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 \( \nu \), 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 [43 Sec A.3].

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

\[
G(z) = I_3 + \Lambda(z) V^T Q_b(z) V \in \mathbb{R}^{3 \times 3},
\]

with \( Q_b(z), \delta(z) \) defined in (5), \( V = [\mu, Cw_*, Cw] \in \mathbb{R}^{p \times 3}, U \equiv C^{\frac{1}{2}} [w_*, w] \in \mathbb{R}^{p \times 2} \) and

\[
\Lambda(z) \equiv \mathbb{E} g \frac{1}{1 + g \cdot \delta(z)} \begin{bmatrix}
1 & \left(U^T z\right)^T \\
U^+ z(U^+ z)^T - (U^T U)^+ & 0
\end{bmatrix}, \quad z = C^{-\frac{1}{2}} (x - \mu) \sim \mathcal{N}(0, I_p),
\]

\(^1\)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.

6
where we denote $U^+$ the Moore–Penrose pseudoinverse of $U$. Then, for $\lambda$ such that $G(\lambda)$ has a zero eigenvalue (of multiplicity one), there exists an eigenvalue $\hat{\lambda}$ of $H$ 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 $G$ 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” $\mu$ 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 $H \xrightarrow{a.s.} E[H]$, with

$$E[H] = E[\ell''(y, w^T x) xx^T] = E[g] \cdot C + V \begin{pmatrix} 1 & E[g \cdot U^+ z] \\ E[g \cdot U^+ z] & V^T \end{pmatrix} U^+ E[g \cdot (zz^T - I_p)](U^+)^T$$

As a consequence, it is expected that in the large $n, p \to \infty$ limit, the top eigenvalues of $H$ could also be related to the columns of $V$. This is the case in Fig 1d where the top eigenvector is observed to be a “noisy” version of the model parameter $w_*$. More precisely, for $(\hat{\lambda}, \hat{u})$ an isolated eigenpair of $H$, the projection $V^T \hat{u} \hat{u}^T V \in R^{3 \times 3}$ can be shown to be asymptotically close to a deterministic matrix. This measures the “cosine-similarly” between the Hessian isolated eigenvector $\hat{u}$ with each column of $V$ and consequently the performance of using $\hat{u}$ as an estimate of, for instance the model parameter $w_*$ for $C = I_p$.

This result is given in the following theorem, which is proven in [43 Sec A.4].

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

$$V^T \hat{u}\hat{u}^TV = -V^T Q_0(\lambda) V \cdot \Xi(\lambda) + o(1),$$

$$\Xi(\lambda) = (v_{l,G}^T G'(\lambda) v_{r,G})^{-1} \cdot v_{r,G} v_{l,G}^T,$$

for $Q_0(z)$ and $G(z)$ defined in (5) and (8), respectively, $v_{l,G}, v_{r,G} \in R^3$ the left and right eigenvectors of $G(\lambda)$ associated with eigenvalue zero, and $G'(\lambda)$ the derivative of $G(z)$ with respect to $z$ evaluated at $z = \lambda$.

### 2.3 Technical Tool: Deterministic Equivalent

Our main tool to derive Theorem 1 and 2 is a so-called deterministic equivalent [29, 16] result for the Hessian resolvent $Q(z) = (H - I_p)^{-1}$, that provides simultaneous access to the Hessian limiting eigenvalue distribution and the behavior of the possible isolated eigenpairs. Precisely, the normalized trace $\text{tr} Q(z)/p$ gives the Stieltjes transform $m_H(z) = \int (t - z)^{-1} \mu_H(dt)$ of the empirical spectral measure $\mu_H$ of $H$, from which $\mu_H$ can be recovered. Also, for $(\lambda, u)$ an eigenpair of interest, with Cauchy’s integral formula we have $|w^T \hat{u}|^2 = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} w^T Q(z) w dz$, for a deterministic vector $w \in R^p$ and $\Gamma_\lambda$ a positively oriented contour surrounding only $\lambda$. As such, for $\hat{Q}(z)$ a deterministic equivalent of $Q(z)$, that is, $\hat{Q}(z) \leftrightarrow Q(z)$ with $\text{tr} A(\hat{Q}(z) - Q(z))/p \to 0$ and $a^T (\hat{Q}(z) - Q(z)) b \to 0$ almost surely as $n, p \to \infty$, for $A \in R^{p \times p}$ and $a, b \in R^p$ of bounded (Euclidean and spectral) norms, the limiting spectral measure (via the associated Stieltjes transform) and the isolated eigenpairs of $H$ are directly accessible via the study of the deterministic equivalent $\hat{Q}(z)$. This result is given as follows, with the proof deferred to [43 Sec A.1].

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

$$Q(z) \leftrightarrow \hat{Q}(z),$$

with $\hat{Q}^{-1}(z) = E \left[ \frac{1}{1 + g(\delta(z))} (C \hat{z}^2 + P_U z \hat{z}_U) \hat{z}_U + \alpha \alpha^T \right] - zI_p,$

for random vector $\alpha \equiv \mu + C \hat{z}^2 z \in R^p$ and $g = \ell''(y, w^T x + w^T C \hat{z}^2 z) z \sim N(0, I_p)$ as defined in (6), $y$ and $\delta(z)$ defined in (11) and (5), respectively, and $P_U \in R^{p \times p}$ the projection onto the subspace spanned by the columns of $U \equiv C \hat{z}_U [w_*, w]$.

### 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
loose Mar˘cenko-Pastur approximation and acts as a more accurate first example in the understanding of concentrated (around some constant, see Fig 3b versus 3d for a comparison between different cases).

While Theorem 1 is proven here only for Gaussian features, we conjecture, as is the case for many g-As long as C

\( (ii) \) logistic loss and (iii) identity covariance

(2) with logistic loss, the Hessian spectrum is close, at least visually, to a (rescaled) Mar˘cenko-Pastur law. This “empirical fit” has been observed in [56, Fig 5], where acceleration methods proposed for 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 Mar˘cenko-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 identity covariance C \( = I_p \), 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 depend on \( w, w^* \) and the Stieltjes transform \( m(z) \) is the solution to \( zcm^2(z) - (1 - c - z)m(z) + 1 = 0 \) and corresponds to the Mar˘cenko-Pastur law.

As long as g is not constant, the limiting Hessian spectrum is, a priori, different from the Mar˘cenko-Pastur law, even in the C \( = I_p \) setting, since the associated Stieltjes transform \( m(z) \) is different from the solution to the Mar˘cenko-Pastur equation. However, we see in Fig 3a of Theorem 1 that, for the logistic model (2) with logistic loss, the Hessian spectrum is close, at least visually, to a (rescaled) Mar˘cenko-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 3a versus 3b 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 Mar˘cenko-Pastur-shape.

This “empirical fit” has been observed in [56, Fig 5], where acceleration methods proposed for a Mar˘cenko-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 Mar˘cenko-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 identity covariance C \( = I_p \); 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 Mar˘cenko-Pastur law. In this vein, our Theorem 1 goes beyond such “loose” Mar˘cenko-Pastur approximation and acts as a more accurate first example in the understanding 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 [43, Fig 5] and more discussions therein.

### 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 [3, 5, 41], for which (i) the isolated spike appears due to the presence of Hessian asymptotics, that it holds more generally beyond Gaussian distribution, see [43, Fig 5] and more discussions therein.
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., \( w_* \) and \( w \)) and exhibits a rather different behavior.

Figure 4: Spike due to data signal in Corollary 1; theory versus practice of (left) Hessian eigen-spectrum with \( \|\mu\|^2 = 0.8 \), (middle) eigengap \( \text{dist}(\lambda_{\mu}, \text{supp}(\mu)) \), and (right) top eigenvector alignment (with \( w \)), as a function of the signal strength \( \|\mu\|^2 \), on logistic model with logistic loss, for \( \mu \approx [-1/p/2, 1/p/2], w = w_* = 0, C = I_p, p = 512 \) and \( n = 2048 \). 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 \( \|w\| = 2 \), with a left isolated eigenvalue \( \hat{\lambda}_{w} \). (middle) eigengap \( \text{dist}(\lambda_{w}, \text{supp}(\mu)) \), and (right) dominant eigenvector alignment (with \( w \)), as a function \( \|w\| \) with \( w \approx [-1/p/2, 1/p/2], w_* = \mu = 0, C = I_p, p = 800 \) and \( n = 8000 \). Results averaged over 50 runs.

3.2.1 Spike due to data signal

To study the spike due to data “signal” \( \mu \) and its phase transition behavior, we focus here on the case \( w_* = w = 0 \). 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 [43, Sec B.3].

**Corollary 1** (Spike due to data signal: logistic model). Consider the logistic model in (2) with logistic loss, for \( w = w_* = 0 \) and \( C = I_p \), 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 \( (\hat{\lambda}_{\mu}, \hat{u}_{\mu}) \) of \( H \) and it satisfies

\[
\hat{\lambda}_{\mu} \xrightarrow{a.s.} \begin{cases} \lambda_{\mu} = \frac{1}{4}(1 + \rho + c \cdot \sqrt{\rho}) & \rho > \sqrt{c}, \\ \frac{1}{4}(1 + \sqrt{\rho})^2 & \rho \leq \sqrt{c}; \end{cases}, \quad \frac{\|\mu^T \hat{u}_{\mu}\|^2}{\|\mu\|^2} \xrightarrow{a.s.} \begin{cases} \alpha = \frac{\rho^2 + \rho}{p^2 + \rho} & \rho > \sqrt{c}, \\ 0 & \rho \leq \sqrt{c}; \end{cases}
\]

with the signal strength \( \rho = \lim_{p \to \infty} \|\mu\|^2 \) and \( c = \lim_{p \to \infty} p/n \).

The behavior of the isolated eigen-pairs described in Corollary 1 follows the “classical” phase transition [43][54]: (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 \( \|\mu\|^2 \) in the sense that, for a fixed dimension ratio \( c \), the largest Hessian eigenvalue is bound to become asymptotically isolated once \( \|\mu\|^2 \) exceeds \( \sqrt{c} \) and its value, as well as the eigenvector alignment, increase monotonically as \( \|\mu\|^2 \) 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 \( w \).
3.2.2 Spike due to model

To investigate the spike due to the underlying model (i.e., \( w, \tilde{w} \), and \( w \)), we position ourselves in the situation where \( \mu = 0 \), that is, in the absence of data “signal”. This leads to the following corollary, the proof of which is given in [43] Sec B.4.

**Corollary 2 (Spike due to model: logistic model).** Consider the logistic model in (2), with logistic loss, \( \mu = 0 \) and \( C = I_p \), then the Stieltjes transform \( m(z) \) satisfies \( m(z) = 1/(|\mathbb{E}[f(r, z)] - z|) \) for \( f(r, z) = 1/(\text{student}(z) + 2 + e^{-r} + e^r) \) and \( r \sim \mathcal{N}(0, |w|^2) \) that depends on \( w \) but not on \( w_\ast \). Moreover, there is at most one isolated eigenvalue \( \lambda_{w} \) of \( H \) that is due to \( w \) and satisfies \( \lambda_{w} \rightarrow 0 \) with \( \lambda_{w} \) solution to \( 0 = \det G(\lambda_{w}) = 1 + m(\lambda_{w})E_{\mathbb{R}}[f(r, \lambda_{w})/|r^2 - |w|^2|] \).

The situation here is more subtle (than the spike due to data signal discussed in Sec 3.2.1): as the model parameter \( w \) changes (e.g., as the “energy” \( |w| \) grows), both the Hessian (limiting) eigenvalue distribution and the possible spike location are impacted. Fig 5 illustrates the behavior of the spike due to \( w \) in the setting of Corollary 2. Note first that, different from the case of spike due to data signal \( \mu \), 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. Also, note from Fig 5 (middle) that, different from the spike due to \( \mu \), the spike due to \( w \) exhibits here a “non-monotonic” behavior in the sense that, it is absent for small values of \( |w| \) (as for small \( \|\mu\| \) in Fig 3 (middle) and becomes “isolated” as \( |w| \) increases, but then again “merges into” the main bulk as \( |w| \) 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 \( w \) is, unlike the eigengap in Fig 5 (middle), monotonically increasing as \( |w| \) grows large, as in the case of Fig 3 (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 [67] [35]. It is worthy mentioning that, while, technically speaking, the proposed analysis is not capable of characterizing the behavior as \( |w| \rightarrow \infty \) under Assumption 1, empirical results suggest that for extremely large \( |w| \), the eigengap tends to “vanish”, the associated dominant eigenvector can still be used to recover \( w \) almost perfectly, see [43] Fig 8 as an example.

4 Conclusion

In this article, we provided a precise asymptotic characterization of Hessian eigenspectra (including the limiting eigenvalue distribution and the behavior of possible isolated eigenvalue-eigenvector pairs) for the family of G-GLMs via a random matrix analysis. By assuming generic Gaussian input features, our results show that the Hessian can have a “heavy-tailed” or multi-bulk spectral behavior, and with isolated eigenvalues of the left- or right-hand side of the main bulk, depending on the underlying GLM and the choice of loss function.

Extending our approach to more involved ML models such as multi-layer neural networks is challenging, e.g., one will need to first study the layer-wise Hessian and evaluate how the objective of interest (e.g., the eigenvalue distribution) of the full Hessian depends on each of its (layer-wise) blocks. Simplification can be made by specifying the choice of loss function and/or using different approximations (such as Gauss-Newton-type approximation [63] or working in the infinitely-wide but less-realistic NTK regime [32]), so as to facilitate analysis. We leave this for future work.

**Acknowledgments and Disclosure of Funding**

ZL would like to acknowledge the Fundamental Research Funds for the Central Universities of China (2021XXJS110) and CCF-Hikvision Open Fund (20210008) for providing partial support. MWM would like to acknowledge DARPA, IARPA (contract W911NF20C0035), NSF, and ONR via its BRC on RandNLA for providing partial support. Our conclusions do not necessarily reflect the position or policy of our sponsors, and no official endorsement should be inferred.

---

*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 5 and establish a non-monotonic behavior as a function of \(|w|\) or \(|w_\ast|\); see [43] Sec 4.3.*
References


