Hessian Eigenspectra of More Realistic Nonlinear Models

Zhenyu Liao, Michael W. Mahoney

EIC, Huazhong University of Science and Technology, China
and
ICSI and Department of Statistics, University of California, Berkeley, USA

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)
- for involved ML models (e.g., neural networks), characterization of Hessian often under strong simplifying assumptions, e.g., “mixed” behavior of Marčenko-Pastur and semicircle laws

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)
- for involved ML models (e.g., neural networks), characterization of Hessian often under strong simplifying assumptions, e.g., "mixed" behavior of Marčenko-Pastur and semicircle laws

In this work:

- focus on a large family extends generalized linear model: $y_i \sim f(y \mid \mathbf{w}_*^\top \mathbf{x}_i)$, **convex** or **non-convex**

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)
- for involved ML models (e.g., neural networks), characterization of Hessian often under strong simplifying assumptions, e.g., "mixed" behavior of Marčenko-Pastur and semicircle laws

In this work:

- focus on a large family extends generalized linear model: $y_i \sim f(y \mid \mathbf{w}_*^\top \mathbf{x}_i)$, **convex** or **non-convex**
- precise characterization of Hessian eigenvalue distribution and (possible) isolated eigenvalue-eigenvector pairs via **Random Matrix Theory (RMT)**

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)
- for involved ML models (e.g., neural networks), characterization of Hessian often under strong simplifying assumptions, e.g., “mixed” behavior of Marčenko-Pastur and semicircle laws

In this work:

- focus on a large family extends generalized linear model: $y_i \sim f(y \mid \mathbf{w}_*^\top \mathbf{x}_i)$, **convex** or **non-convex**
- precise characterization of Hessian eigenvalue distribution and (possible) isolated eigenvalue-eigenvector pairs via **Random Matrix Theory (RMT)**
- qualitatively different Hessian behavior depending on the response model, loss, and feature statistics

Motivation

- Hessian plays a crucial role in applied math, optimization, statistics, and machine learning (ML)
- for involved ML models (e.g., neural networks), characterization of Hessian often under strong simplifying assumptions, e.g., "mixed" behavior of Marčenko-Pastur and semicircle laws

In this work:

- focus on a large family extends generalized linear model: $y_i \sim f(y \mid \mathbf{w}_*^\top \mathbf{x}_i)$, **convex** or **non-convex**
- precise characterization of Hessian eigenvalue distribution and (possible) isolated eigenvalue-eigenvector pairs via **Random Matrix Theory (RMT)**
- qualitatively different Hessian behavior depending on the response model, loss, and feature statistics
- **application**: spectral initialization using top Hessian eigenvectors in non-convex models

System model

For input feature $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and response model $y_i \sim f(y \mid \mathbf{w}_*^\mathsf{T} \mathbf{x}_i)$, minimizing the empirical risk

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

for some loss $\ell(y, h) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

System model

For input feature $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and response model $y_i \sim f(y \mid \mathbf{w}_*^\mathsf{T} \mathbf{x}_i)$, minimizing the empirical risk

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

for some loss $\ell(y, h) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, associated Hessian

$$\mathbf{H} = \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \equiv \frac{1}{n} \mathbf{X} \mathbf{D} \mathbf{X}^\mathsf{T}, \quad \mathbf{D} = \operatorname{diag}\{\ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i)\}_{i=1}^{n}, \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n} \tag{2}$$

System model

For input feature $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and response model $y_i \sim f(y \mid \mathbf{w}_*^\mathsf{T} \mathbf{x}_i)$, minimizing the empirical risk

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

for some loss $\ell(y, h) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, associated Hessian

$$\mathbf{H} = \frac{1}{n} \sum_{i=1}^n \ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \equiv \frac{1}{n} \mathbf{X} \mathbf{D} \mathbf{X}^\mathsf{T}, \quad \mathbf{D} = \operatorname{diag}\{\ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i)\}_{i=1}^n, \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n} \tag{2}$$

(related to *separable covariance model* in RMT, but with **D dependent** on **X**!)

System model

For input feature $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and response model $y_i \sim f(y \mid \mathbf{w}_*^\mathsf{T} \mathbf{x}_i)$, minimizing the empirical risk

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

for some loss $\ell(y,h) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, associated Hessian

$$\mathbf{H} = \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \equiv \frac{1}{n} \mathbf{X} \mathbf{D} \mathbf{X}^\mathsf{T}, \quad \mathbf{D} = \operatorname{diag}\{\ell''(y_i, \mathbf{w}^\mathsf{T} \mathbf{x}_i)\}_{i=1}^{n}, \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n} \tag{2}$$

(related to *separable covariance model* in RMT, but with **D dependent** on **X**!)

High dimensional asymptotics

As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, we have

1. $\max\{\|\mathbf{w}\|, \|\mathbf{w}_*\|\} = O(1)$
2. $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with $\max\{\|\boldsymbol{\mu}\|, \|\mathbf{C}\|\} = O(1)$

(Limiting) Hessian eigenvalue distribution

Theorem (Limiting spectral measure)

As $n, p \to \infty$, the empirical Hessian eigenvalue distribution $\mu_{\mathbf{H}}$ 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

$$m(z) = \int \left(-z + \tilde{t} \int \frac{t}{1 + t\delta(z)} \nu(dt) \right)^{-1} \tilde{\nu}(d\tilde{t}), \quad \delta(z) = \int \frac{c\tilde{t}}{-z + \tilde{t} \int \frac{t}{1+t\delta(z)} \nu(dt)} \tilde{\nu}(d\tilde{t}). \tag{3}$$

for ν the law/distribution of g with

$$g \equiv \partial^2 \ell(y,h) / \partial h^2, \quad h = \mathbf{w}^\mathsf{T} \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\mathsf{T} \boldsymbol{\mu}, \mathbf{w}^\mathsf{T} \mathbf{C} \mathbf{w}), \tag{4}$$

and $\tilde{\nu}$ the (limiting) eigenvalue distribution of $\mathbf{C}$.

(Limiting) Hessian eigenvalue distribution

Theorem (Limiting spectral measure)

As $n, p \to \infty$, the empirical Hessian eigenvalue distribution $\mu_{\mathbf{H}}$ 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

$$m(z) = \int \left(-z + \tilde{t} \int \frac{t}{1 + t\delta(z)} \nu(dt) \right)^{-1} \tilde{\nu}(d\tilde{t}), \quad \delta(z) = \int \frac{c\tilde{t}}{-z + \tilde{t} \int \frac{t}{1+t\delta(z)} \nu(dt)} \tilde{\nu}(d\tilde{t}). \tag{3}$$

for ν the law/distribution of g with

$$g \equiv \partial^2 \ell(y,h) / \partial h^2, \quad h = \mathbf{w}^\mathsf{T} \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\mathsf{T} \boldsymbol{\mu}, \mathbf{w}^\mathsf{T} \mathbf{C} \mathbf{w}), \tag{4}$$

and $\tilde{\nu}$ the (limiting) eigenvalue distribution of $\mathbf{C}$.

Looks complicated but

(Limiting) Hessian eigenvalue distribution

Theorem (Limiting spectral measure)

As $n, p \to \infty$, the empirical Hessian eigenvalue distribution $\mu_{\mathbf{H}}$ 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

$$m(z) = \int \left(-z + \tilde{t} \int \frac{t}{1 + t\delta(z)} \nu(dt) \right)^{-1} \tilde{\nu}(d\tilde{t}), \quad \delta(z) = \int \frac{c\tilde{t}}{-z + \tilde{t} \int \frac{t}{1+t\delta(z)} \nu(dt)} \tilde{\nu}(d\tilde{t}). \tag{3}$$

for ν the law/distribution of g with

$$g \equiv \partial^2 \ell(y,h)/\partial h^2, \quad h = \mathbf{w}^\mathsf{T} \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\mathsf{T} \boldsymbol{\mu}, \mathbf{w}^\mathsf{T} \mathbf{C} \mathbf{w}), \tag{4}$$

and $\tilde{\nu}$ the (limiting) eigenvalue distribution of $\mathbf{C}$.

Looks complicated but

- capture the interplay between **loss function** (via ν), **feature statistics** (via $\tilde{\nu}$) and **dimensionality** $c = p/n$

(Limiting) Hessian eigenvalue distribution

Theorem (Limiting spectral measure)

As $n, p \to \infty$, the empirical Hessian eigenvalue distribution $\mu_{\mathbf{H}}$ 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

$$m(z) = \int \left(-z + \tilde{t} \int \frac{t}{1 + t\delta(z)} \nu(dt) \right)^{-1} \tilde{\nu}(d\tilde{t}), \quad \delta(z) = \int \frac{c\tilde{t}}{-z + \tilde{t} \int \frac{t}{1+t\delta(z)} \nu(dt)} \tilde{\nu}(d\tilde{t}). \tag{3}$$

for ν the law/distribution of g with

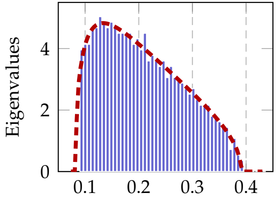
$$g \equiv \partial^2 \ell(y,h)/\partial h^2, \quad h = \mathbf{w}^\mathsf{T} \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\mathsf{T} \boldsymbol{\mu}, \mathbf{w}^\mathsf{T} \mathbf{C} \mathbf{w}), \tag{4}$$

and $\tilde{\nu}$ the (limiting) eigenvalue distribution of $\mathbf{C}$.

Looks complicated but

- capture the interplay between **loss function** (via ν), **feature statistics** (via $\tilde{\nu}$) and **dimensionality** $c = p/n$
- can be (analytically) evaluated with ease and lead to qualitatively different Hessian behavior

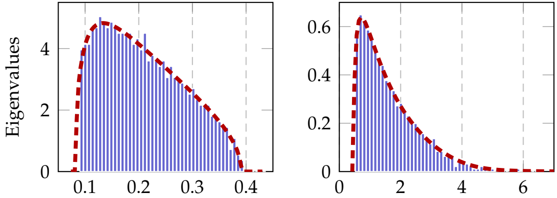
Hessian eigenvalue distribution: implications



(a) Logistic loss

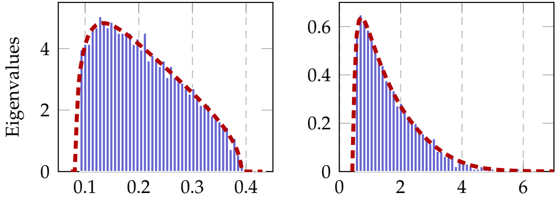
- **impact of loss function**: bounded (**a**)

Hessian eigenvalue distribution: implications



(a) Logistic loss (b) Exponential loss

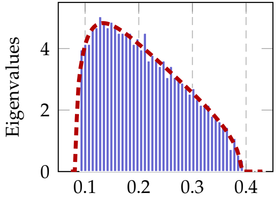
- **impact of loss function**: bounded (**a**) versus unbounded (**b**) Hessian eigenvalues



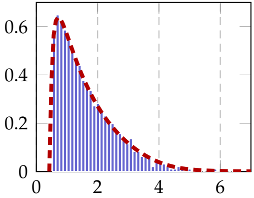
(a) Logistic loss (b) Exponential loss

- **impact of loss function**: bounded (**a**) versus unbounded (**b**) Hessian eigenvalues
  - Hessian has unbounded eigen-support if and only if $g \equiv \partial^2 \ell(y,h)/\partial h^2$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ is bounded

Hessian eigenvalue distribution: implications



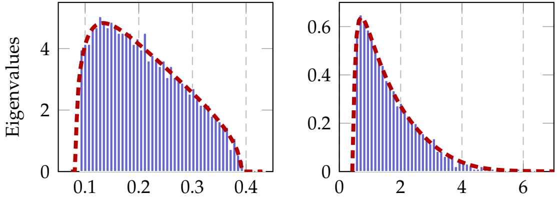
(a) Logistic loss



(b) Exponential loss

- **impact of loss function**: bounded (**a**) versus unbounded (**b**) Hessian eigenvalues
  - Hessian has unbounded eigen-support if and only if $g \equiv \partial^2 \ell(y,h)/\partial h^2$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ is bounded
  - ✓ logistic model with logistic loss

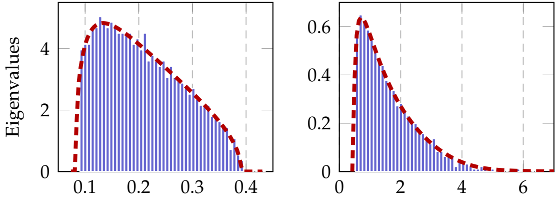
Hessian eigenvalue distribution: implications



(a) Logistic loss   (b) Exponential loss

- **impact of loss function**: bounded (**a**) versus unbounded (**b**) Hessian eigenvalues
  - – Hessian has unbounded eigen-support if and only if $g \equiv \partial^2 \ell(y,h)/\partial h^2$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ is bounded
  - ✓ logistic model with logistic loss
  - x logistic model with exponential loss

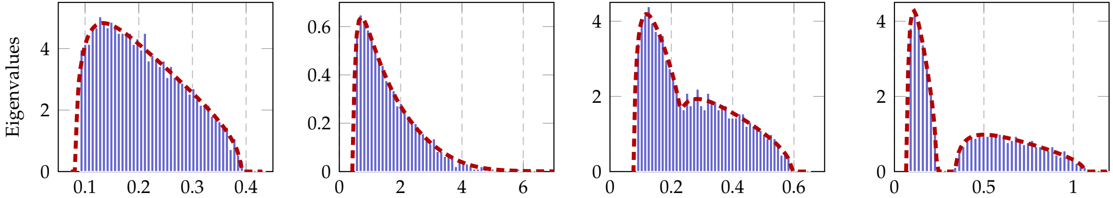
Hessian eigenvalue distribution: implications



(a) Logistic loss (b) Exponential loss

- **impact of loss function**: bounded (**a**) versus unbounded (**b**) Hessian eigenvalues
  - – Hessian has unbounded eigen-support if and only if $g \equiv \partial^2 \ell(y,h)/\partial h^2$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ is bounded
  - ✓ logistic model with logistic loss
  - x logistic model with exponential loss
  - x phase retrieval model with square loss

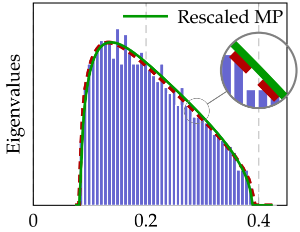
Hessian eigenvalue distribution: implications



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

- **impact of loss function**: bounded **(a)** versus unbounded **(b)** Hessian eigenvalues
  - – Hessian has unbounded eigen-support if and only if $g \equiv \partial^2 \ell(y,h)/\partial h^2$ for $h \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\boldsymbol{\mu}, \mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w})$ is bounded
  - ✓ logistic model with logistic loss
  - x logistic model with exponential loss
  - x phase retrieval model with square loss
- **impact of feature covariance C**: Hessian spectra of single- **(c)** versus multi-bulk **(d)**

Marčenko-Pastur-shaped Hessian?

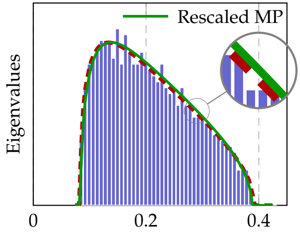


(a) Logistic loss

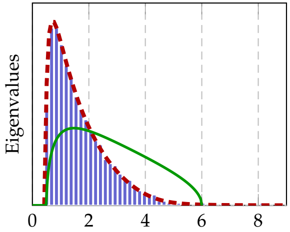
Hessian eigenvalues (empirical in **blue**, theory in **red**) versus rescaled and shifted Marčenko-Pastur (**green**):

(a) Marčenko-Pastur-like Hessian with logistic loss

Marčenko-Pastur-shaped Hessian?



(a) Logistic loss



(b) Exponential loss

Hessian eigenvalues (empirical in **blue**, theory in **red**) versus rescaled and shifted Marčenko-Pastur (**green**):

(a) Marčenko-Pastur-like Hessian with logistic loss

(b) an example of **non**-Marčenko-Pastur-like Hessian with exponential loss

Marčenko-Pastur-shaped Hessian?

(a) Logistic loss (b) Exponential loss

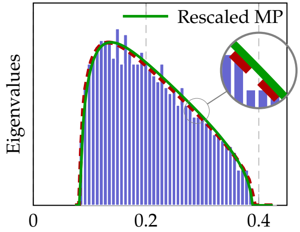
Hessian eigenvalues (empirical in **blue**, theory in **red**) versus rescaled and shifted Marčenko-Pastur (**green**):

(a) Marčenko-Pastur-like Hessian with logistic loss

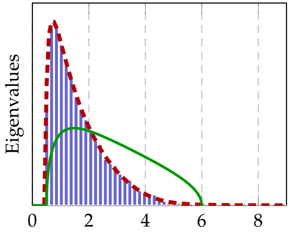
(b) an example of **non**-Marčenko-Pastur-like Hessian with exponential loss

$\Rightarrow$ this "visual approximation" with Marčenko-Pastur law is not robust!

Marčenko-Pastur-shaped Hessian? Yes but only visually in some cases!



(a) Logistic loss



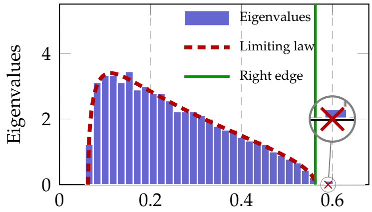
(b) Exponential loss

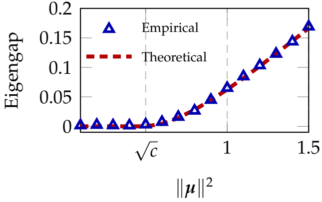
Hessian eigenvalues (empirical in **blue**, theory in **red**) versus rescaled and shifted Marčenko-Pastur (**green**):

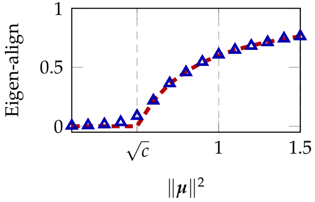
- **(a)** Marčenko-Pastur-like Hessian with logistic loss
- **(b)** an example of **non**-Marčenko-Pastur-like Hessian with exponential loss
- $\Rightarrow$ this "visual approximation" with Marčenko-Pastur law is not robust!

Isolated eigenvalue-eigenvectors pairs and their phase transitions

- spike due to feature signal **on the right-hand side**: classical BBP phase transition in RMT



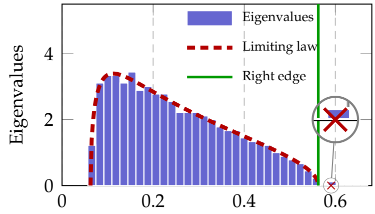


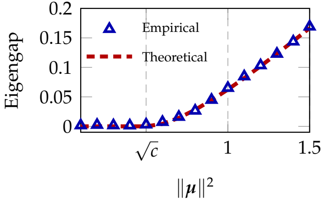


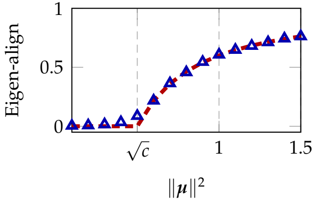
[0] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697

Isolated eigenvalue-eigenvectors pairs and their phase transitions

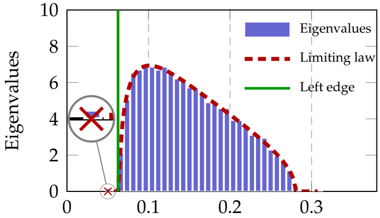
- spike due to feature signal **on the right-hand side**: classical BBP phase transition in RMT

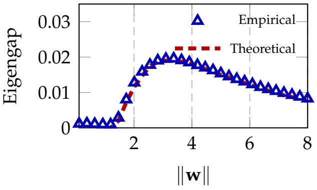


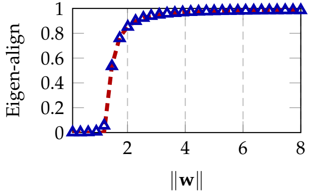




- spike due to model **on the left- or right-hand side**: **novel** phase transition!







[0] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697

Conclusion and take-away message

- Hessian eigenspectra in generally do **NOT** take a Marčenko-Pastur form, even for simple GLMs!

Conclusion and take-away message

- Hessian eigenspectra in generally do **NOT** take a Marčenko-Pastur form, even for simple GLMs!
- very different behavior for this simple model: bounded versus unbounded support, single- versus multi-bulk, different phase transitions, etc.

Conclusion and take-away message

- Hessian eigenspectra in generally do **NOT** take a Marčenko-Pastur form, even for simple GLMs!
- very different behavior for this simple model: bounded versus unbounded support, single- versus multi-bulk, different phase transitions, etc.

Check our paper

`https://arxiv.org/abs/2103.01519`

and my homepage `https://zhenyu-liao.github.io/` for more information!

Conclusion and take-away message

- Hessian eigenspectra in generally do **NOT** take a Marčenko-Pastur form, even for simple GLMs!
- very different behavior for this simple model: bounded versus unbounded support, single- versus multi-bulk, different phase transitions, etc.

Check our paper

`https://arxiv.org/abs/2103.01519`

and my homepage `https://zhenyu-liao.github.io/` for more information!

Thank you!