Random Matrix Theory for Machine Learning

Introduction to Random Matrix Theory

Fabian Pedregosa, Courtney Paquette, Tom Trogdon, Jeffrey Pennington

https://random-matrix-learning.github.io

Table of contents

Stieltjes Transform

Maximum entropy principle

A disordered (real world) system will be random in all ways that are not explicitly prevented.

Conversely, a matrix is interesting only in those ways it fails to look like a random matrix.

Notes of Elliot Paquette and the thesis, *A random matrix framework for large dimensional machine learning and neural networks* by Zhenyu Liao

Example MNIST

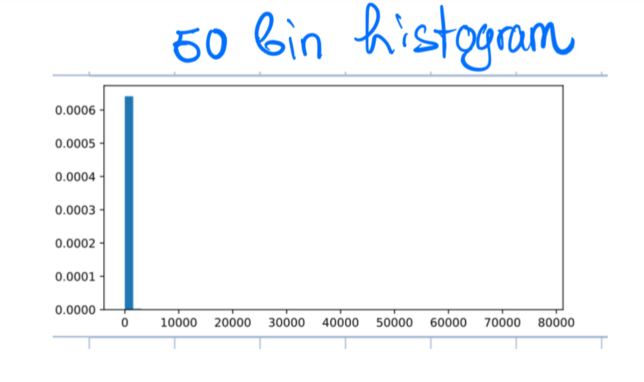
MNIST M $(60,000 \times 28 \times 28)$, form sample covariance matrix, $S = MM^T$

Does S look like a random matrix?

Example MNIST

MNIST M $(60,000 \times 28 \times 28)$, form sample covariance matrix, $S = MM^T$

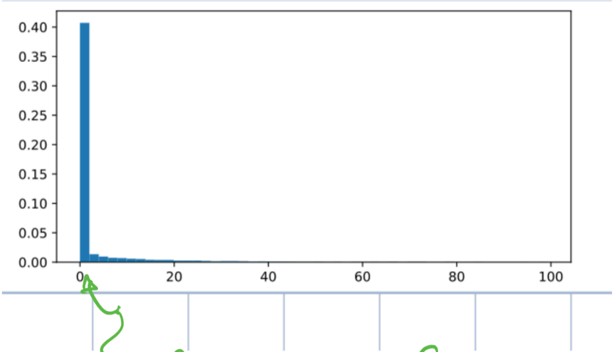
Does S look like a random matrix?



Observation 1: 1 giant eigenvalue, M has non-zero mean

Example MNIST

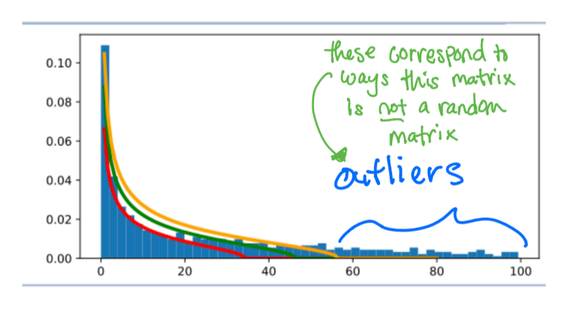
Remove large eigenvalue



we have a big δ_0 at 0.
we remove this for the picture

Example MNIST

Observation 2: There is a bulk component



- Fit bulk with Marchenko-Pastur(1) to correspond to point mass
- Large eigenvalue correspond to interesting outliers
- May be ”hidden” (weak) outliers in the bulk eigenvalues

Method 1: Stieltjes transform

Stieltjes transform

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{t-z}\,\mu(\mathrm{d}t)$$

$$z \in \mathbb{C}, \quad \Im(z) > 0, \quad \mu \text{ is } \mathbb{P}\text{-measure on } \mathbb{R}$$

Theorem (Stieltjes inversion)

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im\left(m_\mu(x + i\varepsilon)\right) \xrightarrow[\varepsilon \to 0]{} \mu$$

Stieltjes transform

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{t-z}\,\mu(\mathrm{d}t)$$

$$z \in \mathbb{C}, \quad \Im(z) > 0, \quad \mu \text{ is } \mathbb{P}\text{-measure on } \mathbb{R}$$

Theorem (Stieltjes inversion)

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im\left(m_\mu(x+i\varepsilon)\right) \xrightarrow[\varepsilon \to 0]{} \mu$$

Example: μ is law of Unif$([-1,1])$
Stieltjes transform:

$$m_\mu(z) = \frac{1}{2}\int_{\mathbb{R}} \frac{1}{t-z} 1\left(\{|t| \le 1\}\right)\,\mathrm{d}t = \frac{1}{2}\int_{-1}^{1} \frac{\mathrm{d}t}{t-z} = \frac{1}{2}\log\left(\frac{1-z}{-1-z}\right)$$

Inversion $\qquad \lim_{\varepsilon \downarrow 0} \frac{\Im}{\pi}\left(\frac{1}{2}\log\left(\frac{1-z}{-1-z}\right)\right)\Big|_{z=x+i\varepsilon} = \begin{cases} \frac{1}{2}, & \text{if } |x| < 1 \\ 0, & \text{if } |x| > 1 \\ \star, & |x| = 1 \end{cases}$

Empirical spectral distribution and Stieltjes

$$\text{ESD of } \boldsymbol{A}: \quad \mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{A})}$$

Stieltjes transform of ESD

$$m_{\mu_n}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i(\boldsymbol{A}) - z} = \frac{1}{n}\mathrm{tr}\left((\boldsymbol{A} - z\boldsymbol{I}_d)^{-1}\right)$$

Theorem

If for each $z \in \mathbb{C}$, $\Im(z) > 0$, and

$$m_{\mu_n}(z) \xrightarrow[n\to\infty]{} m_{\mu}(z) \quad \Rightarrow \quad \mu_n \xrightarrow[n\to\infty]{} \mu.$$

Resolvent

$$\boldsymbol{Q}(z) \stackrel{\text{def}}{=} \text{Resolvent of } \boldsymbol{A} = (\boldsymbol{A} - z\boldsymbol{I}_d)^{-1}$$

Remarks

For nice random matrices (GOE, Wishart, sample covariance),

$$\text{Resolvent of } \boldsymbol{A} \approx m_{\boldsymbol{A}}(z)\boldsymbol{I}_d$$

where $m_{\boldsymbol{A}}$ is the Stieltjes transform of $\boldsymbol{A}$. That is, for any unit vector $\boldsymbol{u}$ independent of $\boldsymbol{A}$,

$$\boldsymbol{u}^T(\boldsymbol{A} - z\boldsymbol{I}_d)^{-1}\boldsymbol{u} \cong m_{\boldsymbol{A}}(z) \qquad \text{(weak sense)}$$

*This gives not only eigenvalues but also eigenvectors

Lemma:
Suppose $\boldsymbol{x} \in \mathbb{R}^p$ has i.i.d. entries of mean zero, unit variance. Then

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - \mathrm{tr} \boldsymbol{A} \to 0$$

Wishart: $W = \frac{1}{n} X X^T, \quad X \in \mathbb{R}^{d \times n}, \quad d/n \to r \in (0 \infty)$

$$X = \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \end{bmatrix} \Rightarrow \text{Resolvent of } W = \boldsymbol{Q}_n(z) = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - z \boldsymbol{I}_d \right)^{-1}$$

Question: What is the Stieltjes transform of Wishart W?

Suppose $\exists\ \boldsymbol{Q}(z) \in \mathbb{C}^{d \times d}$ s.t. $\frac{1}{d} \mathrm{tr}(\boldsymbol{Q}_n(z) - \boldsymbol{Q}(z)) \to 0$ (Stieltjes of MP $= \mathrm{tr}(\boldsymbol{Q}(z))$)

Fact: $$\left| \frac{1}{d} \mathrm{tr} \left(\boldsymbol{Q}_n(z) (\boldsymbol{Q}(z)^{-1} + z \boldsymbol{I}_d) \boldsymbol{Q}(z) \right) - \frac{1}{n} \frac{1}{d} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{Q}(z) \boldsymbol{Q}_n(z) \boldsymbol{x}_i \right| \to 0$$

Marchenko-Pastur and Stieltjes

Linear algebra to construct self-consistent equation for $\mathrm{tr}(\mathbf{Q}_n(z))$:

Remove 1 column and 1 row:

$$\mathbf{Q}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T - z\mathbf{I}_d \right)^{-1} = \left(\frac{1}{n} \mathbf{x}_1 \mathbf{x}_1^T + \overbrace{\frac{1}{n} \sum_{i=2}^{n} \mathbf{x}_i \mathbf{x}_i^T - z\mathbf{I}_d}^{\mathbf{Q}_n^{(1)}} \right)^{-1}$$

$$= \mathbf{Q}_n^{(1)} - \frac{n^{-1} \mathbf{Q}_n^{(1)} \mathbf{x}_1 \mathbf{x}_1^T \mathbf{Q}_n^{(1)}}{1 + n^{-1} \mathbf{x}_1^T \mathbf{Q}_n^{(1)} \mathbf{x}_1} \quad \text{(Sherman-Morrison)}$$

Marchenko-Pastur and Stieltjes

Linear algebra to construct self-consistent equation for $\mathrm{tr}(Q_n(z))$:

Remove 1 column and 1 row:

$$Q_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - zI_d\right)^{-1} = \Bigg(\frac{1}{n}x_1x_1^T + \overbrace{\frac{1}{n}\sum_{i=2}^{n} x_i x_i^T - zI_d}^{Q_n^{(1)}}\Bigg)^{-1}$$

$$= Q_n^{(1)} - \frac{n^{-1}Q_n^{(1)}x_1x_1^TQ_n^{(1)}}{1+n^{-1}x_1^TQ_n^{(1)}x_1} \quad \text{(Sherman-Morrison)}$$

$Q_n^{(1)}$, x_1 independent $\Rightarrow x_1^T Q Q_n^{(1)} x_1 = tr(QQ_n^{(1)})$ and $x_1^T Q_n^{(1)} x_1 = tr(Q_n^{(1)})$

$$d^{-1}x_1^TQQ_nx_1 \quad = \quad \frac{d^{-1}\mathrm{tr}(QQ_n^{(1)})}{1+n^{-1}\mathrm{tr}(Q_n^{(1)})} \approx \frac{d^{-1}\mathrm{tr}(QQ_n)}{1+n^{-1}\mathrm{tr}(Q_n)}$$

Marchenko-Pastur and Stieltjes

Linear algebra to construct self-consistent equation for $\mathrm{tr}(Q_n(z))$:

Remove 1 column and 1 row:

$$Q_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - zI_d\right)^{-1} = \left(\frac{1}{n}x_1 x_1^T + \overbrace{\frac{1}{n}\sum_{i=2}^{n} x_i x_i^T - zI_d}^{Q_n^{(1)}}\right)^{-1}$$

$$= Q_n^{(1)} - \frac{n^{-1}Q_n^{(1)} x_1 x_1^T Q_n^{(1)}}{1 + n^{-1}x_1^T Q_n^{(1)} x_1} \quad \text{(Sherman-Morrison)}$$

$Q_n^{(1)}$, x_1 independent $\Rightarrow x_1^T Q Q_n^{(1)} x_1 = tr(QQ_n^{(1)})$ and $x_1^T Q_n^{(1)} x_1 = tr(Q_n^{(1)})$

$$d^{-1}x_1^T Q Q_n x_1 = \frac{d^{-1}\mathrm{tr}(QQ_n^{(1)})}{1+n^{-1}\mathrm{tr}(Q_n^{(1)})} \approx \frac{d^{-1}\mathrm{tr}(QQ_n)}{1+n^{-1}\mathrm{tr}(Q_n)}$$

$$\begin{array}{ccc} \frac{1}{nd}\sum_{i=1}^{n} x_i^T Q Q_n x_i & \approx & \frac{d^{-1}\mathrm{tr}(QQ_n)}{1+n^{-1}\mathrm{tr}(Q_n)} \\ \downarrow & & \downarrow \\ d^{-1}\mathrm{tr}\big(Q_n(Q^{-1}+zI_d)Q\big) & \approx & \frac{d^{-1}\mathrm{tr}(QQ_n)}{1+n^{-1}\mathrm{tr}(Q_n)} \end{array} \Rightarrow \quad Q^{-1} + zI_d = \frac{1}{1+n^{-1}\mathrm{tr}(Q_n)} I_d$$

$$Q = \left(-z + \frac{1}{1+n^{-1}\mathrm{tr}(Q_n)}\right)^{-1} I_d$$

Marchenko-Pastur and Stieltjes

Stieltjes of $W_n = m_{W_n}(z) = d^{-1}\mathrm{tr}(Q_n) \approx d^{-1}\mathrm{tr}(Q) = \left(-z + \frac{1}{1 + r \cdot d^{-1}\mathrm{tr}(Q_n)}\right)^{-1}$

Marchenko-Pastur and Stieltjes

Stieltjes of $W_n = m_{W_n}(z) = d^{-1}\mathrm{tr}(Q_n) \approx d^{-1}\mathrm{tr}(Q) = \left(-z + \frac{1}{1 + r \cdot d^{-1}\mathrm{tr}(Q_n)}\right)^{-1}$

Fixed point equation for trace ($r = \lim_{n\to\infty} \frac{d}{n}$):

$$m_{W_n}(z) = \frac{1}{-z + \frac{1}{1+r\cdot m_{W_n}(z)}} + \varepsilon(n,z), \qquad \varepsilon(n,z) \xrightarrow[n\to\infty]{} 0$$

Fixed point for Marchenko-Pastur, MP:

$$m_{\mathrm{MP}}(z) = \frac{1}{-z + \frac{1}{1+r\cdot m_{\mathrm{MP}}(z)}}, \qquad \text{Note: } m(z) \sim \frac{-1}{z} \text{ as } z \to \infty$$

✓ solve numerically ✓ many dist. satisfy fixed point eqn.

Marchenko-Pastur and Stieltjes

Stieltjes of $\boldsymbol{W}_n = m_{\boldsymbol{W}_n}(z) = d^{-1}\mathrm{tr}(\boldsymbol{Q}_n) \approx d^{-1}\mathrm{tr}(\boldsymbol{Q}) = \left(-z + \frac{1}{1 + r \cdot d^{-1}\mathrm{tr}(\boldsymbol{Q}_n)}\right)^{-1}$

Fixed point equation for trace ($r = \lim_{n\to\infty} \frac{d}{n}$):

$$m_{\boldsymbol{W}_n}(z) = \frac{1}{-z + \frac{1}{1+r\cdot m_{\boldsymbol{W}_n}(z)}} + \varepsilon(n,z), \qquad \varepsilon(n,z) \xrightarrow[n\to\infty]{} 0$$

Fixed point for Marchenko-Pastur, MP:

$$m_{\mathrm{MP}}(z) = \frac{1}{-z + \frac{1}{1+r\cdot m_{\mathrm{MP}}(z)}}, \qquad \text{Note: } m(z) \sim \frac{-1}{z} \text{ as } z \to \infty$$

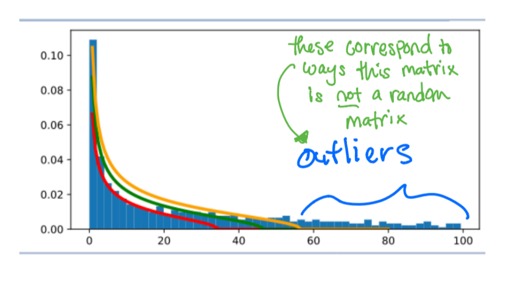
✓ solve numerically ✓ many dist. satisfy fixed point eqn.

Stieltjes transform of Marchenko-Pastur

$$m_{\mathrm{MP}}(z) = \frac{1 - r - z + \sqrt{((1+\sqrt{r})^2 - z)((1-\sqrt{r})^2 - z)}}{2rz}$$

Bulk + outliers

How do we model this?



	Bulk	+	Outliers
	$\downarrow$		$\downarrow$
Model:	Marchenko-Pastur	+	Spikes

Sample covariance matrices

Set-up

- Covariance matrix, C, (symmetric, positive semi-definite matrix)
- Noise matrix, $Z \in \mathbb{R}^{d\times n}$ (mean 0, variance 1, i.i.d.)
- $n^{1/\delta} \leq d \leq n^{\delta}$ for some $\delta > 0$
- $X = C^{1/2}Z$; Form $\frac{XX^T}{n}$
- $\|C\|_2 \leq$ constant, independent of n

Sample covariance matrices

Set-up

- Covariance matrix, C, (symmetric, positive semi-definite matrix)
- Noise matrix, $Z \in \mathbb{R}^{d\times n}$ (mean 0, variance 1, i.i.d.)
- $n^{1/\delta} \leq d \leq n^{\delta}$ for some $\delta > 0$
- $X = C^{1/2}Z$; Form $\frac{XX^T}{n}$
- $\|C\|_2 \leq$ constant, independent of n

Theorem (Bai, Krishnaiah, Silverstein, Yin, '80s-'90s)

$$\text{Stieltjes of } \frac{XX^T}{n} = \operatorname{tr}\left[\left(I_d - z\frac{XX^T}{n}\right)^{-1}\right] \approx \frac{n}{d}\widetilde{m}(z) + \frac{1-\frac{d}{n}}{\frac{d}{n}z}$$

$$\text{where} \qquad \widetilde{m}(z) = \left(-z + \frac{1}{n}\operatorname{tr}[C(I_d - \widetilde{m}(z)C)^{-1}]\right)^{-1}$$

✓ $\widetilde{m}(z) \approx$ Stieltjes transform $\frac{X^TX}{n}$ ✓ implicit eqn, solved numerically

Examples of Sample Covariances

See Colab for details

R-Transform

Example Hessian of 2-layer Network Model

Setup

- $W^{(1)} \in \mathbb{R}^{n_1 \times n_0}$ and $W^{(2)} \in \mathbb{R}^{n_2 \times n_1}$ weight matrices, i.i.d. $N(0, 1)$
- $\boldsymbol{x} \in \mathbb{R}^{n_0 \times m}$ is input data and $\boldsymbol{y} \in \mathbb{R}^{n_2 \times m}$ targets
- $g : \mathbb{R} \to \mathbb{R}$ activation function
- $n = n_0 = n_1 = n_2$ and $\phi = \frac{2n}{m}$

$$\text{outputs:} \quad \widehat{\boldsymbol{y}} = W^{(2)} g(W^{(1)} \boldsymbol{x}) \qquad \text{residuals:} \quad e_{ij} = \widehat{y}_{ij} - y_{ij}$$

Goal:
$$\min_{\boldsymbol{\theta} = [W^{(1)}, W^{(2)}]} \left\{ f(\boldsymbol{\theta}) = \frac{1}{2m} \| W^{(2)} g(W^{(1)} \boldsymbol{x}) - \boldsymbol{y} \|^2 \right\}$$

Example Hessian of 2-layer Network Model

Setup

- $W^{(1)} \in \mathbb{R}^{n_1 \times n_0}$ and $W^{(2)} \in \mathbb{R}^{n_2 \times n_1}$ weight matrices, i.i.d. $N(0,1)$
- $\boldsymbol{x} \in \mathbb{R}^{n_0 \times m}$ is input data and $\boldsymbol{y} \in \mathbb{R}^{n_2 \times m}$ targets
- $g : \mathbb{R} \to \mathbb{R}$ activation function
- $n = n_0 = n_1 = n_2$ and $\phi = \frac{2n}{m}$

outputs: $\widehat{\boldsymbol{y}} = W^{(2)} g(W^{(1)} \boldsymbol{x})$ residuals: $e_{ij} = \widehat{y}_{ij} - y_{ij}$

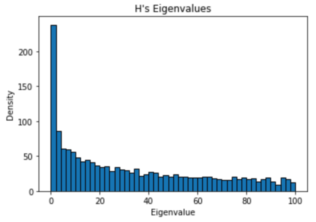
Goal: $$\min_{\boldsymbol{\theta}=[W^{(1)},W^{(2)}]} \left\{ f(\boldsymbol{\theta}) = \frac{1}{2m} \| W^{(2)} g(W^{(1)} \boldsymbol{x}) - \boldsymbol{y} \|^2 \right\}$$

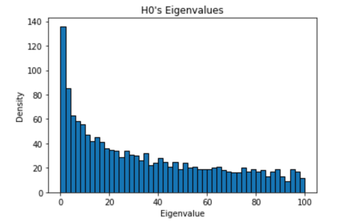
Hessian: $H = H_0 + H_1$

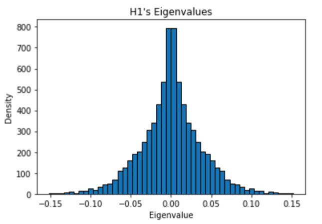
$$[H_0]_{\alpha\beta} = \frac{1}{m} \sum_{i,j=1}^{n_2,m} \frac{\partial \hat{y}_{ij}}{\partial \theta_\alpha} \frac{\partial \hat{y}_{ij}}{\partial \theta_\beta} \quad \text{and} \quad [H_1]_{\alpha\beta} = \frac{1}{m} \sum_{i,j}^{n_2,m} e_{ij} \frac{\partial^2 \hat{y}_{ij}}{\partial \theta_\alpha \partial \theta_\beta}$$

Snapshots of the Hessian during Training

Hessian: $H = H_0 + H_1$







Question: How do you model eigenvalues of H from H_0 and H_1?

Images by McGill undergraduate, Ria Stevens

R-Transform

Tool for writing simple formulas for densities from known densities

Definitions

Free convolution of measures, $\mu_A \boxplus \mu_B$

If A, B two random matrices with ESD, μ_A and μ_B,

$$\text{ESD of } A + B = \mu_A \boxplus \mu_B$$

provided matrix sizes large and matrices asymptotically free.

(side note: product of matrices version)

Asymptotically free matrices

Orthogonal invariance

An A random symmetric matrix is orthogonally invariance if for any fixed orthogonal matrix O

$$O^T A O \overset{\text{law}}{=} A$$

Examples

- Multiples of the Identity
- Wishart with Gaussian entries
- GOE with Gaussian entries

Asymptotically free matrices

Sufficient condition

Suppose $\{A_n\}$ and $\{B_n\}$ are $n \times n$ random matrices. If the following

- A_n and B_n are independent
- A_n is orthogonally invariant

Then A_n and B_n are asymptotically free and $\mu_{A_n} \boxplus \mu_{B_n} \cong \mu_{A_n+B_n}$

Intuition: Eigenvectors of A_n are completely unaligned from B_n

Stieltjes and R-transform

R-transform

R-transform is inverse of Stieltjes of m

$$R(-m(z)) - \frac{1}{m(z)} = z.$$

Examples

- $\beta \boldsymbol{u}\boldsymbol{u}^T$ is $R_{\beta \boldsymbol{u}\boldsymbol{u}^T} = \frac{\beta}{n(1-s\beta)}$
- $R_{\text{Marchenko-Pastur(r)}}(s) = \frac{1}{1-sr}$
- $R_{\text{semicircle}}(s) = s$

Calculus for spectral densities

Theorem

If A and B are asymptotically free,

$$R_{\mu_{A+B}} = R_{\mu_A \boxplus \mu_B} = R_{\mu_A}(s) + R_{\mu_B}(s).$$

Remark

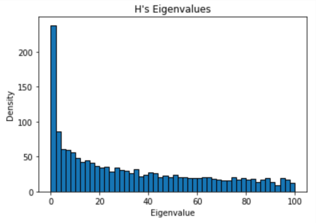
R-transform $\Leftrightarrow$ Stieltjes transform $\Leftrightarrow$ Spectral density

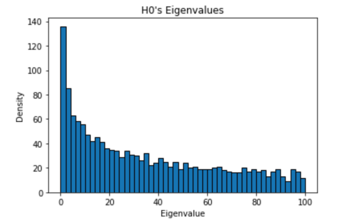
Example

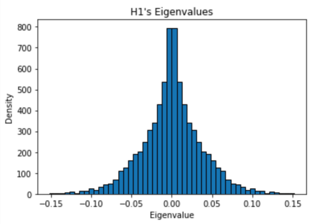
$$R_{\text{GOE + Wishart}} = R_{\text{semicircle}} + R_{\text{MP}} = \underbrace{s}_{\text{GOE}} + \underbrace{\frac{1}{1 - sr}}_{\text{Wishart}}$$

Snapshots of the Hessian during Training

Hessian: $H = H_0 + H_1$







Question: How do you model eigenvalues of H from H_0 and H_1?

Images by McGill undergraduate, Ria Stevens

Model for the Hessian

$$n_2 \varepsilon \stackrel{\text{def}}{=} f(\boldsymbol{\theta}) \quad \text{Hessian:} \quad \boldsymbol{H} = \boldsymbol{H}_0 + \boldsymbol{H}_1$$

Modeling Assumptions

- $\boldsymbol{H}_0$ is Marchenko-Pastur
- $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are asymptotically free
- $n = n_0 = n_1 = n_2$ and $\phi = \frac{2n}{m}$

R-transform of $\boldsymbol{H}_1$: $$R_{\boldsymbol{H}_1}(s) = \frac{\varepsilon \phi s}{2 - \varepsilon \phi^2 s^2}$$

R-transform of $\boldsymbol{H}$

$$R_{\boldsymbol{H}}(s) = \frac{1}{1 - s\phi} + \frac{\varepsilon \phi s}{2 - \varepsilon \phi^2 s^2}$$

Questions?

References i

Z. Bai and J. Silverstein.

Spectral analysis of large dimensional random matrices, volume 20.

Springer, 2010.

Z. D. Bai and Jack W. Silverstein.

CLT for linear spectral statistics of large-dimensional sample covariance matrices.

Ann. Probab., 32(1A):553–605, 2004.

ISSN 0091-1798.

doi: 10.1214/aop/1078415845.

URL *https://10.1214/aop/1078415845*.

Jinho Baik and Jack W. Silverstein.
Eigenvalues of large sample covariance matrices of spiked population models.
J. Multivariate Anal., 97(6):1382–1408, 2006.
doi: 10.1016/j.jmva.2005.08.003.
URL *https://10.1016/j.jmva.2005.08.003*.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters.
Cleaning large correlation matrices: tools from random matrix theory.
Phys. Rep., 666:1–109, 2017.
ISSN 0370-1573.
doi: 10.1016/j.physrep.2016.10.005.
URL *https://10.1016/j.physrep.2016.10.005*.

Zhenyu Liao.
A random matrix framework for large dimensional machine learning and neural networks.
PhD thesis, Université Paris-Saclay, 2019.

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 of *Proceedings of Machine Learning Research*, pages 2798–2806. PMLR, 2017.
URL
http://proceedings.mlr.press/v70/pennington17a.html.