Recent Advances in Random Matrix Theory for Neural Networks

Zhenyu Liao

L2S, CentraleSupélec, Université Paris-Saclay, France.

28/July/2018, Shanghai Jiao Tong University, Shanghai

CentraleSupélec

Outline

1. Motivation
2. Random Weights Ridge Regression
3. Random Weights Spectral Clustering
4. Random Matrix Analysis for Learning Dynamics of Neural Networks
5. Summary: Take-away Messages and References

Motivation: Deep Neural Networks in Double Asymptotic Regime

- Big Data era: both high dimensional and massive amount of data
- Understanding deep neural nets in the double asymptotic regime (random matrix regime): often have far more network parameters than needed, but still generalize well

  $\Rightarrow$ number of **network parameters** and number of **data instances** comparably large
- Counterintuitive phenomenon in random matrix regime:

Classical Statistics Break Down in Random Matrix Regime

- Estimating covariance matrix of data $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$, $x_i \sim \mathcal{N}(0, I_p)$ of true covariance I_p.
- Classical sample covariance matrix: $\mathrm{SCM} = \frac{1}{T} \sum_{i=1}^{T} x_i x_i^\mathsf{T} = \frac{1}{T} X X^\mathsf{T}$ of rank at most T!
- In random matrix regime where $T \sim p$, classical estimator breaks down!
  $\Rightarrow$ For example if $T < p$, SCM will never be correct (with at least $p - T$ zero eigenvalues)!

- Apply (classical) RMT to neural network analysis: remaining difficulty in nonlinearity!

Motivation: Nonlinearity in Random Matrix Theory

Objective: Random weights (untrained) neural networks, also called "random feature maps".

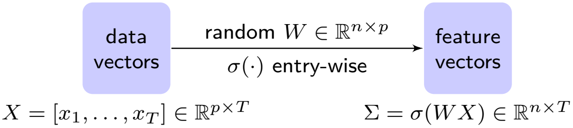


Figure: Illustration of random feature maps

Sample Covariance Matrix of data $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p\times T}$

$$\mathrm{SCM} \equiv \frac{1}{T} X X^{\mathsf{T}}.$$

SCM in feature space $\Rightarrow$ feature Gram matrix G:

$$G \equiv \frac{1}{T} \Sigma^{\mathsf{T}} \Sigma$$

with $\Sigma = [\sigma(x_1), \ldots, \sigma(x_T)]$ feature matrix of X.

Motivation: RMT for random feature maps

Example:

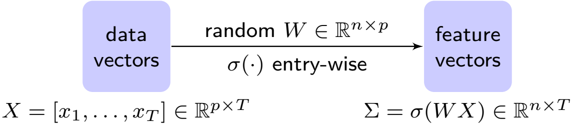


Figure: Illustration of random feature maps

MSE of random weights ridge regression (also called *extreme learning machines*):

$$\mathrm{E}_{\rm train} = \frac{1}{T}\|y - \beta^{\sf T}\Sigma\|_F^2 = \frac{\gamma^2}{T} y^{\sf T} Q^2(-\gamma) y, \quad \mathrm{E}_{\rm test} = \frac{1}{\hat{T}}\|\hat{y} - \beta^{\sf T}\hat{\Sigma}\|_F^2$$

with ridge regressor $\beta \equiv \frac{1}{T}\Sigma\left(G + \gamma I_T\right)^{-1} y^{\sf T} = \frac{1}{T}\Sigma Q(-\gamma) y^{\sf T}$ and regularization $\gamma > 0$. y associated target of training data X and $\hat{y}$ target of test data $\hat{X}$.

$\Rightarrow G$ determines training and test performance via its *resolvent*

$$Q(z) \equiv \left(G - zI_T\right)^{-1}.$$

Key Issue

(Classical) quadratic form $a^{\sf T} Q(z) b$ for nonlinear model $\Sigma = \sigma(WX)$!

Handle nonlinearity in RMT: concentration of measure approach

Recall:
For $\sigma(t) = t$, $G = \frac{1}{T} X^\mathsf{T} W^\mathsf{T} W X$ with random W: Sample Covariance Matrix Model. Proof essentially based on trace lemma: $w \in \mathbb{R}^n$ of i.i.d. entries and A of bound norm,

$$\left| \frac{1}{n} w^\mathsf{T} A w - \frac{1}{n} \operatorname{tr} A \right| \xrightarrow{\text{a.s.}} 0.$$

Nonlinearity

However, here for nonlinear $\sigma(\cdot)$, similar to the proof of Marčenko-Pastur law:

$$\Sigma = \sigma(WX) = \begin{bmatrix} \sigma_i^\mathsf{T} \\ \Sigma_{-i} \end{bmatrix} \in \mathbb{R}^{n \times T}$$

with $\sigma_i = \sigma(X^\mathsf{T} w_i) \in \mathbb{R}^T$, w_i the i-th row of W. Rank-one perturbation:

$$Q = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma - z I_T \right)^{-1} = \left(\frac{1}{T} \Sigma_{-i}^\mathsf{T} \Sigma_{-i} + \frac{1}{T} \sigma_i \sigma_i^\mathsf{T} - z I_T \right)^{-1}$$

$$= Q_{-i} - \frac{Q_{-i} \frac{1}{T} \sigma_i \sigma_i^\mathsf{T} Q_{-i}}{1 + \frac{1}{T} \sigma_i^\mathsf{T} Q_{-i} \sigma_i}$$

with $Q_{-i} \equiv \left(\frac{1}{T} \Sigma_{-i}^\mathsf{T} \Sigma_{-i} - z I_T \right)^{-1}$ independent of σ_i!

Handle nonlinearity in RMT: concentration of measure approach

Object under study $\frac{1}{n}\sigma(w^\mathsf{T} X)A\sigma(X^\mathsf{T} w)$: (compared to $\frac{1}{n}w^\mathsf{T} Aw$)

- loss of independence between entries
- more elusive due to $\sigma(\cdot)$

$\Rightarrow$ extend trace lemma to handle nonlinear case!

Lemma (Concentration of Quadratic Forms)

$w \in \mathbb{R}^n$ of i.i.d. standard Gaussian entries and $\sigma(\cdot)$ λ_σ-Lipschitz continuous. For $\|A\| \leq 1$ and X of bounded norm,

$$P\left(\left|\frac{1}{T}\sigma(w^\mathsf{T} X)A\sigma(X^\mathsf{T} w) - \frac{1}{T}\operatorname{tr}\Phi A\right| > t\right) \leq Ce^{-cn\min(t,t^2)}$$

for some $C, c > 0$ and $\Phi \equiv E_w\left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$ (function of data X).

Performance evaluation of random feature-based ridge regression

Theorem (Asymptotic Training Performance)

$W \sim \mathcal{N}(0, I_n)$ and $\sigma(\cdot)$ λ_σ-Lipschitz continuous and X of bounded norm. Then, as $n, p, T \to \infty$, $p/n \to c_p \in (0, \infty)$ and $T/n \to c_T \in (0, \infty)$,

$$\mathrm{E}_{\rm train} - \bar{\mathrm{E}}_{\rm train} \xrightarrow{\rm a.s.} 0$$

where $\bar{\mathrm{E}}_{\rm train} = \frac{\gamma^2}{T} y^\mathsf{T} \bar{Q} \left[\frac{\frac{1}{n} \operatorname{tr} \bar{Q} \Psi \bar{Q}}{1 - \frac{1}{n} \operatorname{tr} \Psi^2 \bar{Q}^2} + I_T \right] \bar{Q} y$ and $\bar{Q} = (\Psi + \gamma I_T)^{-1}$, $\Psi \equiv \frac{n}{T} \frac{\Phi}{1+\delta}$ with δ the unique solution of $\delta = \frac{1}{T} \operatorname{tr} \Phi \bar{Q}$ and $\Phi \equiv E_w \left[\sigma(X^\mathsf{T} w) \sigma(w^\mathsf{T} X) \right]$.

Several remarks:

- (asymptotic) training performance only depends on (the training data X via) the key averaged kernel matrix Φ and the dimension of problem
- similar results can be obtained for test performance
- $\Rightarrow$ remains to compute Φ on function of X

To evaluate the training and test performance, it remains to compute Φ for different σ:

$$\Phi(X) = E_w \left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$$

the (i,j)-th entry of which given by

$$\begin{aligned}\Phi_{i,j} &= (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(w^\mathsf{T} x_i)\sigma(w^\mathsf{T} x_j) dw \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \sigma(\tilde{w}^\mathsf{T}\tilde{x}_i)\sigma(\tilde{w}^\mathsf{T}\tilde{x}_j) e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w} \quad \text{(projection on } \operatorname{span}(x_i, x_j)).\end{aligned}$$

Example: for $\sigma(t) = \max(t, 0) = \operatorname{ReLU}(t)$,

$$\Phi_{i,j} = \frac{1}{2\pi} \int_S \sigma(\tilde{w}^\mathsf{T}\tilde{x}_i)\sigma(\tilde{w}^\mathsf{T}\tilde{x}_j) e^{-\frac{1}{2}\|\tilde{w}\|^2} d\tilde{w} = \frac{1}{2\pi}\|x_i\|\|x_j\| \left(\sqrt{1-\angle^2} + \angle \cdot \arccos(-\angle)\right)$$

with $S = \min(\tilde{w}^\mathsf{T}\tilde{x}_i, \tilde{w}^\mathsf{T}\tilde{x}_j) > 0$, $\angle \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$.

Results of Φ for commonly used $\sigma(\cdot)$

Table: $\Phi_{i,j}$ for commonly used $\sigma(\cdot)$, $\angle \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$.

$\sigma(t)$	$\Phi_{i,j}$
t	$x_i^\mathsf{T} x_j$
$\max(t,0)$	$\frac{1}{2\pi}\|x_i\|\|x_j\|\left(\angle\cdot\arccos(-\angle)+\sqrt{1-\angle^2}\right)$
$\|t\|$	$\frac{2}{\pi}\|x_i\|\|x_j\|\left(\angle\cdot\arcsin(\angle)+\sqrt{1-\angle^2}\right)$
$\varsigma_+\max(t,0)+$ $\varsigma_-\max(-t,0)$	$\frac{1}{2}(\varsigma_+^2+\varsigma_-^2)x_i^\mathsf{T} x_j+\frac{\|x_i\|\|x_j\|}{2\pi}(\varsigma_++\varsigma_-)^2\left(\sqrt{1-\angle^2}-\angle\cdot\arccos(\angle)\right)$
$1_{t>0}$	$\frac{1}{2}-\frac{1}{2\pi}\arccos(\angle)$
$\operatorname{sign}(t)$	$\frac{2}{\pi}\arcsin(\angle)$
$\varsigma_2 t^2+\varsigma_1 t+\varsigma_0$	$\varsigma_2^2\left(2(x_i^\mathsf{T} x_j)^2+\|x_i\|^2\|x_j\|^2\right)+\varsigma_1^2 x_i^\mathsf{T} x_j+\varsigma_2\varsigma_0\left(\|x_i\|^2+\|x_j\|^2\right)+\varsigma_0^2$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|x_i\|^2+\|x_j\|^2\right)\right)\cosh(x_i^\mathsf{T} x_j)$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|x_i\|^2+\|x_j\|^2\right)\right)\sinh(x_i^\mathsf{T} x_j)$
$\operatorname{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2x_i^\mathsf{T} x_j}{\sqrt{(1+2\|x_i\|^2)(1+2\|x_j\|^2)}}\right)$
$\exp(-\frac{t^2}{2})$	$\frac{1}{\sqrt{(1+\|x_i\|^2)(1+\|x_j\|^2)-(x_i^\mathsf{T} x_j)^2}}$

$\Rightarrow$ (Still) highly nonlinear function of data X!

Numerical validations

Performance of random feature-based ridge regression:

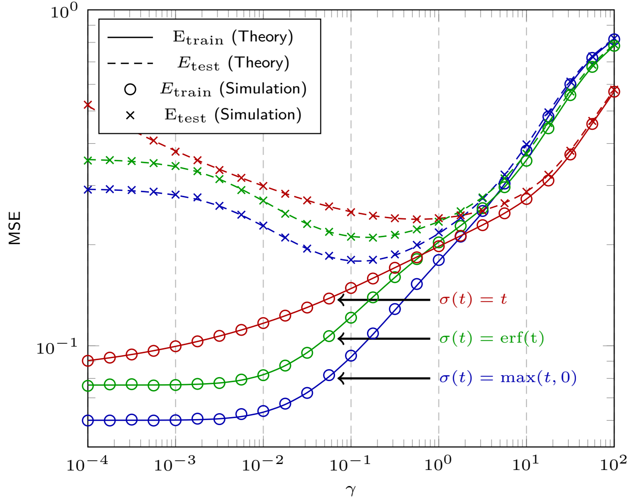


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Dig deeper into the averaged kernel Φ

For random feature maps:

- if deterministic data: performance determined by $\Phi(X)$ and problem dimension
- if data following certain distribution (statistical information+random fluctuation):
  $\Rightarrow$ what is the impact of nonlinearities on information extraction?

Data Model

Consider data from a K-class Gaussian mixture model:

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i = \mu_a/\sqrt{p} + \omega_i$$

with $\omega_i \sim \mathcal{N}(0, C_a/p)$, $a = 1, \ldots, K$ of statistical mean μ_a and covariance C_a.

Non-trivial Classification [Neyman-Pearson Minimal]

For p large, we have $\|\mu_a - \mu_b\| = O(1)$, $\|C_a\| = O(1)$ and $\operatorname{tr}(C_a - C_b)/\sqrt{p} = O(1)$.

$\Rightarrow$ how different nonlinearities influence statistical information contained in Φ (and thus G)?

Counterintuitive phenomenon for high dimensional data

Classification high dimensional Gaussian mixtures:

Non-trivial Classification [Neyman-Pearson Minimal]

For p large, we have $\|\mu_a - \mu_b\| = O(1)$, $\|C_a\| = O(1)$ and $\operatorname{tr}(C_a - C_b)/\sqrt{p} = O(1)$.

As a consequence,

$$\|x_i\|^2 = \underbrace{\|\omega_i\|^2}_{O(1)} + \underbrace{\|\mu_a\|^2/p + 2\mu_a^\mathsf{T}\omega_i/\sqrt{p}}_{O(p^{-1})} = \underbrace{\operatorname{tr} C_a/p}_{O(1)} + \underbrace{\|\omega_i\|^2 - \operatorname{tr} C_a/p}_{O(p^{-1/2})} + \underbrace{\|\mu_a\|^2/p + 2\mu_a^\mathsf{T}\omega_i/\sqrt{p}}_{O(p^{-1})}$$

- if relaxed, classification too easy: it suffices to compare the norm $\|x_i\|^2$ and $\|x_j\|^2$!
- in fact reveals a more intrinsic property of high dimensional data:

Curse of dimensionality: little difference in Euclidean distance between pairs!

Denote $C^\circ = \sum_{i=1}^K \frac{T_i}{T} C_a$ and $C_a = C_a^\circ + C^\circ$ for $a = 1, \ldots, K$.
Then $\|x_i\|^2 = \tau + O(p^{-1/2})$ with $\tau \equiv \operatorname{tr}(C^\circ)/p$, $\|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - x_i^\mathsf{T} x_j \approx 2\tau$:

$\Rightarrow$ Almost constant distance no matter from the same or different classes!

Counterintuitive phenomenon for high dimensional data

Why things are still working? $\Rightarrow$ statistical information are hidden in smaller order terms!

$$\Rightarrow \|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - x_i^\mathsf{T} x_j \approx 2\tau + \underbrace{\omega_i^\mathsf{T}\omega_j}_{O(p^{-1/2})} + \underbrace{\mu_a^\mathsf{T}\mu_b/p + \mu_a^\mathsf{T}\omega_j/\sqrt{p} + \mu_b^\mathsf{T}\omega_i/\sqrt{p}}_{O(p^{-1})}$$

Small entry-wise $\neq$ small in matrix form (in operator norm): repeated in $p \times p$ large matrix $\Rightarrow$ spectral clustering works! 😄

Moreover, "concentration" brings simplifications: for $\Phi_{i,j} = \mathbb{E}_w \, \sigma(w^\mathsf{T} x_i)\sigma(w^\mathsf{T} x_j)$ and ReLU,

$$\Phi_{i,j} = \frac{1}{2\pi}\|x_i\|\|x_j\|\left(\angle \arccos\left(-\angle\right) + \sqrt{1-\angle^2}\right)$$

with $\angle \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$. "Concentration": $\angle = \frac{0}{\tau^2} +$ information terms (μ_a, C_a)!

"Blessing" of Dimensionality

High dimensional "concentration" $\Rightarrow$ Taylor expansion to linearize Φ!

Dig deeper into the average kernel matrix Φ

Asymptotic Equivalent of Φ

For all $\sigma(\cdot)$ listed in the table above, we have, as $n \sim p \sim T \to \infty$,

$$\|\Phi - \tilde{\Phi}\| \to 0$$

almost surely, with

$$\tilde{\Phi} \equiv d_1 \left(\Omega + M \frac{J^\mathsf{T}}{\sqrt{p}} \right)^\mathsf{T} \left(\Omega + M \frac{J^\mathsf{T}}{\sqrt{p}} \right) + d_2 U B U^\mathsf{T} + d_0 I_T$$

and $U \equiv \left[\frac{J}{\sqrt{p}}, \phi \right]$, $\quad B \equiv \begin{bmatrix} tt^\mathsf{T} + 2S & t \\ t^\mathsf{T} & 1 \end{bmatrix}$.

Table: Coefficients d_i in $\tilde{\Phi}$ for different $\sigma(\cdot)$.

$\sigma(t)$	d_1	d_2
t	1	0
$\max(t,0)$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\lvert t\rvert$	0	$\frac{1}{2\pi\tau}$
$\varsigma_+ \max(t,0) + \varsigma_- \max(-t,0)$	$\frac{1}{4}(\varsigma_+ - \varsigma_-)^2$	$\frac{1}{8\tau\pi}(\varsigma_+ + \varsigma_-)^2$
$1_{t>0}$	$\frac{1}{2\pi\tau}$	0
$\operatorname{sign}(t)$	$\frac{2}{\pi\tau}$	0
$\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$	ς_1^2	ς_2^2
$\cos(t)$	0	$\frac{e^{-\tau}}{4}$
$\sin(t)$	$e^{-\tau}$	0
$\operatorname{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-\frac{t^2}{2})$	0	$\frac{1}{4(\tau+1)^3}$

With $J \equiv [j_1, \ldots, j_K]$, j_a canonical vector of $\mathcal{C}_a$: $(j_a)_i = \delta_{x_i \in \mathcal{C}_a}$ (for clustering), weighted by

- Ω, ϕ random fluctuations of data.
- $M \equiv [\mu_1, \ldots, \mu_K]$, $t \equiv \left\{ \operatorname{tr} C_a^\circ / \sqrt{p} \right\}_{a=1}^K$, $S \equiv \{\operatorname{tr}(C_a C_b)/p\}_{a,b=1}^K$ statistical information from data distribution.

Consequence

Table: Coefficients d_i in $\tilde{\Phi}$ for different $\sigma(\cdot)$.

$\sigma(t)$	d_1	d_2
t	1	0
$\max(t,0)$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\|t\|$	0	$\frac{1}{2\pi\tau}$
$\varsigma_+ \max(t,0) + \varsigma_- \max(-t,0)$	$\frac{1}{4}(\varsigma_+ - \varsigma_-)^2$	$\frac{1}{8\tau\pi}(\varsigma_+ + \varsigma_-)^2$
$1_{t>0}$	$\frac{1}{2\pi\tau}$	0
$\mathrm{sign}(t)$	$\frac{2}{\pi\tau}$	0
$\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$	ς_1^2	ς_2^2
$\cos(t)$	0	$\frac{e^{-\tau}}{4}$
$\sin(t)$	$e^{-\tau}$	0
$\mathrm{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-\frac{t^2}{2})$	0	$\frac{1}{4(\tau+1)^3}$

A natural classification of $\sigma(\cdot)$:

- *mean-oriented*, $d_1 \neq 0$, $d_2 = 0$: t, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$ $\Rightarrow$separate with differences in means M;
- *covariance-oriented*, $d_1 = 0$, $d_2 \neq 0$: $|t|$, $\cos(t)$ and $\exp(-t^2/2)$ $\Rightarrow$track differences in covariances t, S;
- *balanced*, both $d_1, d_2 \neq 0$:
  - ReLU function $\max(t,0)$,
  - Leaky ReLU function $\varsigma_+ \max(t,0) + \varsigma_- \max(-t,0)$,
  - quadratic function $\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$.

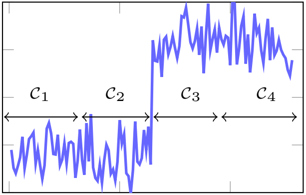
  $\Rightarrow$make use of both statistics!

Not freely tunable as in the case of spectral clustering or SSL!

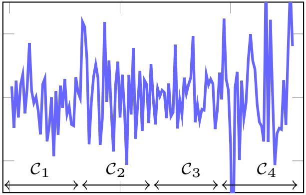
Numerical Validations: Gaussian Data

Example: Gaussian mixture data of four classes: $\mathcal{N}(\mu_1, C_1)$, $\mathcal{N}(\mu_1, C_2)$, $\mathcal{N}(\mu_2, C_1)$ and $\mathcal{N}(\mu_2, C_2)$ with Leaky ReLU function $\varsigma_+ \max(t,0) + \varsigma_- \max(-t,0)$.

Case 1: $\varsigma_+ = \varsigma_- = 1$ (equivalent to linear map $\sigma(t) = t$)

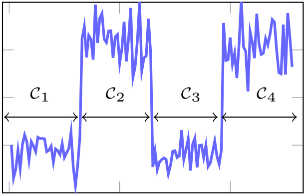


Eigenvector 1

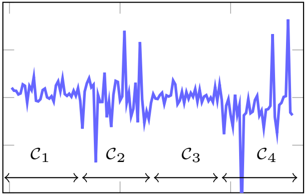


Eigenvector 2

Case 2: $\varsigma_+ = -\varsigma_- = 1$ (equivalent to $\sigma(t) = |t|$)

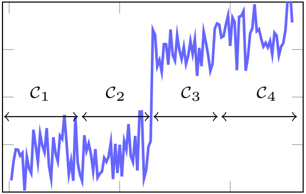


Eigenvector 1

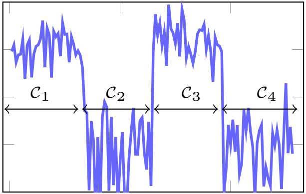


Eigenvector 2

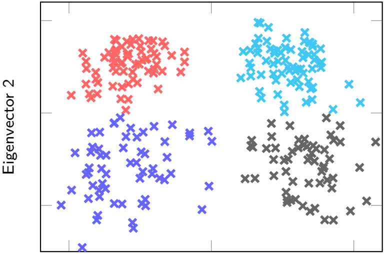
Case 3: $\varsigma_+ = 1$, $\varsigma_- = 0$ (the ReLU function)



Eigenvector 1



Eigenvector 2



Eigenvector 1

Numerical validations: real datasets

Figure: The MNIST image database.

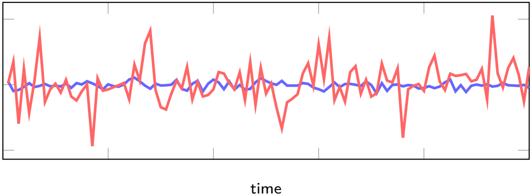


Figure: The epileptic EEG datasets.[1]

Reproducibility: codes available at https://github.com/Zhenyu-LIAO/RMT4RFM.

[1] http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html.

Table: Empirical estimation of differences in means and covariances of MNIST and EEG datasets.

	$\|M^\mathsf{T} M\|$	$\|tt^\mathsf{T} + 2S\|$
MNIST data	**172.4**	86.0
EEG data	1.2	**182.7**

Table: Clustering accuracies on MNIST dataset.

	$\sigma(t)$	$T = 64$	$T = 128$
mean-oriented	t	**88.94%**	87.30%
	$1_{t>0}$	82.94%	85.56%
	$\mathrm{sign}(t)$	83.34%	85.22%
	$\sin(t)$	87.81%	**87.50%**
	$\mathrm{erf}(t)$	87.28%	86.59%
cov-oriented	$\|t\|$	60.41%	57.81%
	$\cos(t)$	59.56%	57.72%
	$\exp(-\frac{t^2}{2})$	60.44%	58.67%
balanced	$\mathrm{ReLU}(t)$	85.72%	82.27%

Table: Clustering accuracies on EEG dataset.

	$\sigma(t)$	$T = 64$	$T = 128$
mean-oriented	t	70.31%	69.58%
	$1_{t>0}$	65.87%	63.47%
	$\mathrm{sign}(t)$	64.63%	63.03%
	$\sin(t)$	70.34%	68.22%
	$\mathrm{erf}(t)$	70.59%	67.70%
cov-oriented	$\|t\|$	99.69%	99.50%
	$\cos(t)$	99.38%	99.36%
	$\exp(-\frac{t^2}{2})$	**99.81%**	**99.77%**
balanced	$\mathrm{ReLU}(t)$	87.91%	90.97%

Numerical Validations: Real Datasets

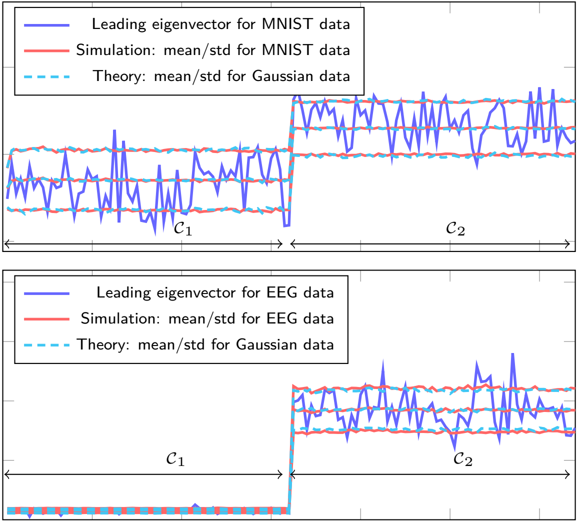


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix Φ
  $\Rightarrow$ random feature-based spectral clustering

Take-away messages:

- fast tuning of hyperparameters
- nonlinearities into three attributes: means-, covariance-oriented and "balanced"
- optimize the choice of nonlinearity as a function of data for quadratic and LReLU

$\Rightarrow$ What happens if weights W are not i.i.d. but depend on data
(in the case of backpropagation)?

Motivation: learning dynamics of neural networks

About neural networks and deep learning:

- Some known facts:
  - trained with backpropagation (gradient-based method)
  - highly over-parameterized, but some still generalize remarkably well
- and some (more) mysteries:
  - how do neural networks learn from training data? what kind of features are learned?
  - how they generalize on unseen data of similar nature? why they do not over-fit?
  - can the network performance be guaranteed or . . . even predicted?

$\Rightarrow$ The learning dynamics of neural networks!

With RMT:

A general framework for studying learning dynamics of a single-layer network!

In particular, under the appropriate double asymptotic regime: number of network parameters and number of data instances comparably large!

As a consequence, more insights on:

- (random) initialization of training
- overfitting in neural networks
- (explicit or implicit) regularization: early stopping, l_2-penalization

Problem setup

Toy model of binary classification:

Gaussian Mixture Data

Consider data x_i drawn from a two-class Gaussian mixture model: for $a = 1, 2$

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i = (-1)^a \mu + \omega_i$$

with ω_i of i.i.d. $\mathcal{N}(0,1)$ entries, label $y_i = -1$ for $\mathcal{C}_1$ and $+1$ for $\mathcal{C}_2$.

Objective: Learning Dynamics

Gradient descent on loss $L(w) = \frac{1}{2n}\|y^\mathsf{T} - w^\mathsf{T} X\|^2$ with $X = [x_1, \ldots, x_n]$. For small learning rate α, with continuous-time approximation:

$$\frac{dw(t)}{dt} = -\alpha \frac{\partial L(w)}{\partial w} = \frac{\alpha}{n} X \left(y - X^\mathsf{T} w(t) \right)$$

of explicit solution $w(t) = e^{-\frac{\alpha t}{n} X X^\mathsf{T}} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} X X^\mathsf{T}} \right) (X X^\mathsf{T})^{-1} X y$ if $X X^\mathsf{T}$ invertible and w_0 the initialization.

To evaluate the learning dynamics:

- depends only on the projection of eigenvector weighted by $\exp(-\alpha t \lambda)$ of associated eigenvalue λ
- functional of sample covariance matrix $\frac{1}{n} X X^\mathsf{T}$ (again): RMT is the answer!

Objective: Generalization Performance

Generalization performance for a new datum $\hat{x}$: $P(w(t)^\mathsf{T}\hat{x} > 0 \mid \hat{x} \in \mathcal{C}_1)$, or $P(w(t)^\mathsf{T}\hat{x} < 0 \mid \hat{x} \in \mathcal{C}_2)$. Since $\hat{x}$ Gaussian and independent of $w(t)$:

$$w(t)^\mathsf{T}\hat{x} \sim \mathcal{N}(\pm w(t)^\mathsf{T}\mu, \|w(t)\|^2)$$

for $w(t) = e^{-\frac{\alpha t}{n}XX^\mathsf{T}}w_0 + \left(I_p - e^{-\frac{\alpha t}{n}XX^\mathsf{T}}\right)(XX^\mathsf{T})^{-1}Xy$.

With RMT:

- although X random: $w(t)^\mathsf{T}\mu$ and $\|w(t)\|^2$ have asymptotically deterministic behavior (only depends on data statistics and problem dimension):
  $\Rightarrow$ the technique of **deterministic equivalent**
- **Cauchy's integral formula** to express the functional $\exp(\cdot)$ via contour integration

$\Rightarrow$ Network performance at any time is in fact deterministic and predictable!

Resolvent and deterministic equivalents

Consider an $n \times n$ Hermitian random matrix M. Define its resolvent $Q_M(z)$, for $z \in \mathbb{C}$ not eigenvalue of M

$$Q_M(z) = (M - zI_n)^{-1}.$$

For a family of M, define a so-called deterministic equivalent $\bar{Q}_M$ of Q_M: a deterministic matrix so that as $n \to \infty$,

- $\frac{1}{n} \operatorname{tr} AQ_M - \frac{1}{n} \operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T} \left(Q_M - \bar{Q}_M \right) b \xrightarrow{\text{a.s.}} 0$

with A, a, b of bounded norm (operator and Euclidean).

$\Rightarrow$ Study $\bar{Q}_M$ instead of the random Q_M for n large!

However, for more sophisticated functionals of M (than $\frac{1}{n} \operatorname{tr} AQ_M$ and $a^\mathsf{T} Q_M b$):

Cauchy's integral formula

Example: for $f(M) = a^\mathsf{T} e^M b dz$,

$$f(M) = -\frac{1}{2\pi i} \oint_\gamma \exp(z) a^\mathsf{T} Q_M(z) b dz \approx -\frac{1}{2\pi i} \oint_\gamma \exp(z) a^\mathsf{T} \bar{Q}_M(z) b dz.$$

with γ a positively oriented path circling around all the eigenvalues of M.

Generalization performance

To evaluate generalization performance: $w(t)^\mathsf{T}\hat{x} \sim \mathcal{N}(\pm w(t)^\mathsf{T}\mu, \|w(t)\|^2)$ with $w(t) = e^{-\frac{\alpha t}{n}XX^\mathsf{T}}w_0 + \left(I_p - e^{-\frac{\alpha t}{n}XX^\mathsf{T}}\right)(XX^\mathsf{T})^{-1}Xy$.

- **Cauchy's integral formula**: for $w(t)^\mathsf{T}\mu$:

$$\mu^\mathsf{T}w(t) = -\frac{1}{2\pi i}\oint_\gamma \mu^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}\left(f_t(z)w_0 + \frac{1-f_t(z)}{z}\frac{1}{n}Xy\right)dz$$

  with $f_t(x) \equiv \exp(-\alpha t x)$. Since $X = -\mu j_1^\mathsf{T} + \mu j_2^\mathsf{T} + \Omega = \mu y^\mathsf{T} + \Omega$, with $\Omega \equiv \left[\omega_1, \ldots, \omega_n\right] \in \mathbb{R}^{p\times n}$ of i.i.d. $\mathcal{N}(0,1)$ entries and $j_a \in \mathbb{R}^n$ the canonical vectors of class $\mathcal{C}_a$, With Woodbury's identity,

$$\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1} = Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}$$

$$\begin{bmatrix}\mu^\mathsf{T}Q(z)\mu & 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y \\ 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y & -1+\frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}Q(z)\frac{1}{n}\Omega y\end{bmatrix}^{-1}\begin{bmatrix}\mu^\mathsf{T} \\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}Q(z)$$

  where $Q(z) = \left(\frac{1}{n}\Omega\Omega^\mathsf{T} - zI_p\right)^{-1}$ and its **deterministic equivalent**:

$$Q(z) \leftrightarrow \bar{Q}(z) = m(z)I_p$$

  with $m(z)$ given by Marčenko-Pastur equation $m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2-4cz}}{2cz}$.
- "replace" the random $Q(z)$ by its **deterministic equivalent** $\bar{Q}(z) = m(z)I_p$.

Main result

Theorem (Generalization Performance)

Let $p/n \to c \in (0,\infty)$ and the initialization w_0 be a random vector with i.i.d. entries of zero mean, variance σ^2/p and finite fourth moment. Then, as $n \to \infty$,

$$P(w(t)^{\mathsf{T}}\hat{x} > 0 \mid \hat{x} \in \mathcal{C}_1) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0,$$

$$P(w(t)^{\mathsf{T}}\hat{x} < 0 \mid \hat{x} \in \mathcal{C}_2) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0$$

with the Q-function: $Q(x) \equiv \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) du$ and

$$\mathrm{E} \equiv -\frac{1}{2\pi i} \oint_\gamma \frac{1 - f_t(z)}{z} \frac{\|\mu\|^2 m(z)\ dz}{(\|\mu\|^2 + c)\, m(z) + 1}$$

$$\mathrm{V} \equiv \frac{1}{2\pi i} \oint_\gamma \left[\frac{\frac{1}{z^2} (1 - f_t(z))^2}{(\|\mu\|^2 + c)\, m(z) + 1} - \sigma^2 f_t^2(z) m(z) \right] dz.$$

γ a closed positively oriented path containing all eigenvalues of $\frac{1}{n}XX^{\mathsf{T}}$ and origin.

Contour integration: hard to understand/interpret $\Rightarrow$ can we further simplify?

Simplification: "break" the contour integration

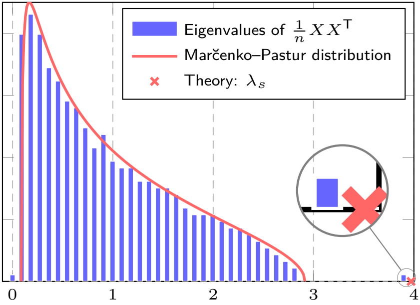


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

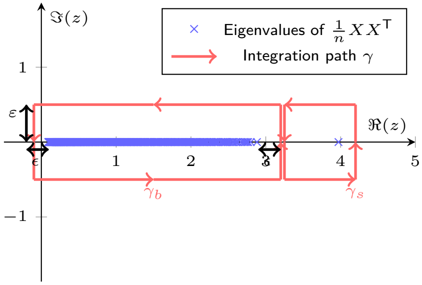


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

Two types of eigenvalues:

- "main bulk" ($[\lambda_-, \lambda_+]$): sum of real integrals
- isolated eigenvalue (λ_s): residue theorem.

Computation of λ_s (Spike model)

- find λ eigenvalue of $\frac{1}{n}XX^\mathsf{T}$ outside $[\lambda_-, \lambda_+]$ (i.e., not eigenvalue of $\frac{1}{n}\Omega\Omega^\mathsf{T}$),

$$\begin{aligned}
&\det\left(\frac{1}{n}XX^\mathsf{T} - \lambda I_p\right) = 0\\
\Leftrightarrow &\det\left(\frac{1}{n}\Omega\Omega^\mathsf{T} - \lambda I_p + \begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^\mathsf{T}\\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}\right) = 0\\
\Leftrightarrow &\det\left(I_2 + \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^\mathsf{T}\\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix} Q(\lambda)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}\right) = 0\\
\Leftrightarrow &1 + (\|\mu\|^2 + c)m(\lambda) + o(1) = 0
\end{aligned}$$

Discussions

(Simplified) generalization performance

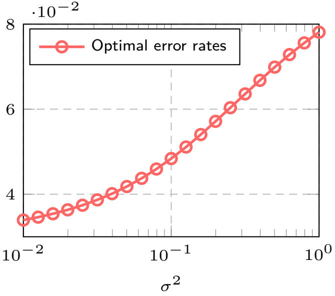
$$\mathrm{E} = \int \frac{1 - f_t(x)}{x} \eta(dx), \ \mathrm{V} = \frac{\|\mu\|^2 + c}{\|\mu\|^2} \int \frac{(1 - f_t(x))^2 \mu(dx)}{x^2} + \sigma^2 \int f_t^2(x) \nu(dx)$$

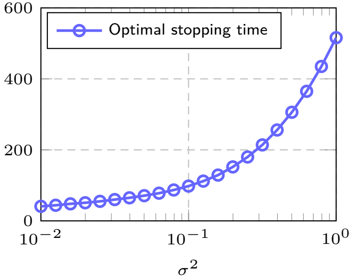
with Marčenko–Pastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+-x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$ with $\lambda_- \equiv (1 - \sqrt{c})^2$, $\lambda_+ \equiv (1 + \sqrt{c})^2$, $\lambda_s = c + 1 + \|\mu\|^2 + c/\|\mu\|^2$ and the measure

$$\eta(dx) \equiv \frac{\sqrt{(x - \lambda_-)^+(\lambda_+ - x)^+}}{2\pi(\lambda_s - x)} dx + \frac{(\|\mu\|^4 - c)^+}{\|\mu\|^2} \delta_{\lambda_s}(x).$$

Some remarks:

- $\eta(dx)$: continuous distribution $[\lambda_-, \lambda_+]$ ($p - 1$ eigenvalues) + Dirac measure at λ_s (one single eigenvalue): contains comparable information!
- $\int \eta(dx) = \|\mu\|^2$, together with Cauchy Schwarz inequality: $\mathrm{E}^2 \leq \int \frac{(1-f_t(x))^2}{x^2} d\mu(x) \cdot \int d\mu(x) \leq \frac{\|\mu\|^4}{\|\mu\|^2+c} \mathrm{V}$, with equality if and only if the (initialization) variance $\sigma^2 = 0$: $\Rightarrow$ Performance drop due to large σ^2!
- How much we over-fit? As $t \to \infty$, performance drop by $\sqrt{1 - \min(c, c^{-1})}$





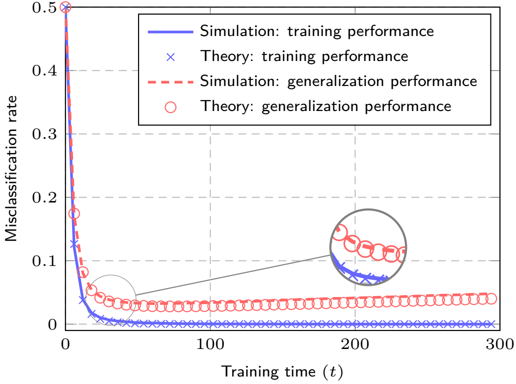


Figure: Training and generalization performance for MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 0.1$. Results averaged over 100 runs.

Take-away messages:

- RMT framework to understand and predict learning dynamics:

  Cauchy's integral formula + technique of deterministic equivalent

- easily extended to more elaborate data models: e.g., Gaussian mixture model with different means and covariances
- byproduct: take initialization variance σ^2 even smaller (than classical $1/p$)!

Take-away messages

- Asymptotic "**concentration effect**" for large $n, p \Rightarrow$ simplification in analyses **and** models.
- Non-trivial **phase transition** phenomena (ability to detect, estimate) when $p, n \to \infty$.
- Access to **limiting performances** and not only bounds! $\Rightarrow$ hyperparameter optimization, algorithm improvement.
- **Complete intuitive change** $\Rightarrow$ opens way to renewed methods.
- **Strong coincidence with real datasets** $\Rightarrow$ easy link between theory and practice.

Perspectives and Open Problems

- Neural nets: loss landscape, gradient descent dynamics and deep learning!
- Generalized linear models
- More general problems from convex optimization (often of *implicit solution*)
- More difficult: problem raised from *non-convex* optimization problems
- Transfer learning, active learning, generative networks (GAN)
- Robust statistics in machine learning
- . . .

N. El Karoui, "The spectrum of kernel random matrices", The Annals of Statistics, 38(1), 1-50, 2010.

C. Xiuyuan, A. Singer, "The spectrum of random inner-product kernel matrices", Random Matrices: Theory and Applications 2.04 (2013): 1350010.

R. Couillet, F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data", Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.

R. Couillet, A. Kammoun, "Random Matrix Improved Subspace Clustering", Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2016.

Z. Liao, R. Couillet, "Random matrices meet machine learning: a large dimensional analysis of LS-SVM", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

X. Mai, R. Couillet, "The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

X. Mai, R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data", (under review) Journal of Machine Learning Research, 2017.

Z. Liao, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines", (under review) Journal of Machine Learning Research, 2017.

K. Elkhalil, A. Kammoun, R. Couillet, T. Al-Naffouri, M.-S. Alouini, “Asymptotic Performance of Regularized Quadratic Discriminant Analysis Based Classifiers”, IEEE International Workshop on Machine Learning for Signal Processing (MLSP’17), Roppongi, Tokyo, Japan, 2017.

H. Tiomoko Ali, A. Kammoun, R. Couillet, “Random matrix-improved kernels for large dimensional spectral clustering”, Statistical Signal Processing Workshop (SSP’18), Freiburg, Germany, 2018.

Summary of Results and Perspectives I

Feature Maps and Neural Networks: References

- C. Williams, "Computation with infinite neural networks", Neural Computation, 10(5), 1203-1216, 1998.
- A. Rahimi, B. Recht, "Random features for large-scale kernel machines", Advances in neural information processing systems pp. 1177-1184, 2007.
- N. El Karoui, "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond", The Annals of Applied Probability, 19(6), 2362-2405, 2009.
- A. Saxe, J. McClelland, S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks", arXiv:1312.6120, 2013.
- A. Choromanska, M. Henaff, M. Mathieu, G. Arous, Y. LeCun, "The loss surfaces of multilayer networks", In Artificial Intelligence and Statistics (pp. 192-204), 2015.
- R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "The asymptotic performance of linear echo state neural networks", Journal of Machine Learning Research, vol. 17, no. 178, pp. 1-35, 2016.
- C. Louart, R. Couillet, "Harnessing neural networks: a random matrix approach", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.
- C. Louart, R. Couillet, "A Random Matrix and Concentration Inequalities Framework for Neural Networks Analysis", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18), Calgary, Canada, 2018.

C. Louart, Z. Liao, R. Couillet, "A Random Matrix Approach to Neural Networks", The Annals of Applied Probability, vol. 28, no. 2, pp. 1190-1248, 2018.

J. Pennington, Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory", In International Conference on Machine Learning, pp. 2798-2806. 2017.

Z. Liao, R. Couillet, "The Dynamics of Learning: A Random Matrix Approach", International Conference on Machine Learning, Stockholm, Sweden, 2018.

Z. Liao, R. Couillet, "On the Spectrum of Random Features Maps of High Dimensional Data", International Conference on Machine Learning, Stockholm, Sweden, 2018.

- N. El Karoui, Noureddine, et al. "On robust regression with high-dimensional predictors", Proceedings of the National Academy of Sciences 110.36 (2013): 14557-14562.
- R. Couillet, M. McKay, "Large Dimensional Analysis and Optimization of Robust Shrinkage Covariance Matrix Estimators", Elsevier Journal of Multivariate Analysis, vol. 131, pp. 99-120, 2014.
- R. Couillet, "Robust spiked random matrices and a robust G-MUSIC estimator", Elsevier Journal of Multivariate Analysis, vol. 140, pp. 139-161, 2015.
- D. Morales-Jimenez, R. Couillet, M. McKay, "Large Dimensional Analysis of Robust M-Estimators of Covariance with Outliers", IEEE Transactions on Signal Processing, vol. 63, no. 21, pp. 5784-5797, 2015.
- D. Donoho, A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing", Probability Theory and Related Fields 166.3-4 (2016): 935-969.
- R. Couillet, A. Kammoun, F. Pascal, "Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals", Elsevier Journal of Multivariate Analysis, vol. 143, pp. 249-274, 2016.
- A. Kammoun, R. Couillet, F. Pascal, M.-S. Alouini, "Optimal Design of the Adaptive Normalized Matched Filter Detector using Regularized Tyler Estimator", IEEE Transactions on Aerospace and Electronic Systems, 2017.

Figure: Related topic on ZhiHu: `https://zhuanlan.zhihu.com/RandomMatrixTheory`.

Thank you.