A Random Matrix Framework for Large Dimensional Machine Learning and Neural Networks

Ph.D. defense

Zhenyu LIAO

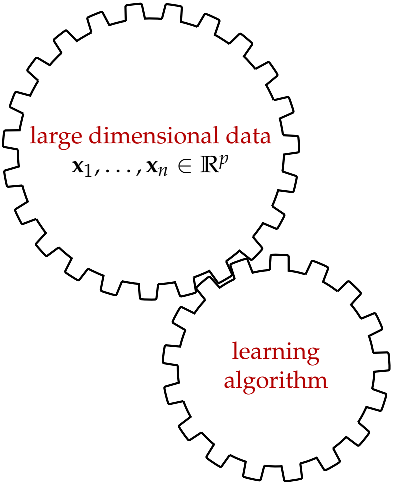
supervised by Romain COUILLET and Yacine CHITOUR

CentraleSupélec, Université Paris-Saclay, France.

September 30, 2019

CentraleSupélec

Understanding the mechanism of large dimensional machine learning



- big data era: exploit large n, p
- counterintuitive phenomena, e.g., the "*curse of dimensionality*"
- complete change of understanding of many algorithms
- **RMT** provides the tools.

Outline

1. Motivation
   - Sample covariance matrix for large dimensional data
   - A random matrix perspective of the “curse of dimensionality”

2. Main results: statistical behavior of large dimensional random feature maps
   - Random feature maps for large dimensional data
   - Application to random features-based ridge regression
   - Random feature maps for classifying Gaussian mixtures
   - Application to random-feature based spectral clustering

3. Conclusion
   - From toy to more realistic learning schemes
   - From toy to more realistic data models

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.
- Maximum likelihood sample covariance matrix:

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p \times p}$$

  of rank at most n: optimal for $n \gg p$ (or, for p "small").
- In the regime $n \sim p$, conventional wisdom breaks down:
  for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues.

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty}$$

  $\Rightarrow$ eigenvalue mismatch and not consistent!

When is one under the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x-a)^+(b-x)^+} dx$$

where $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

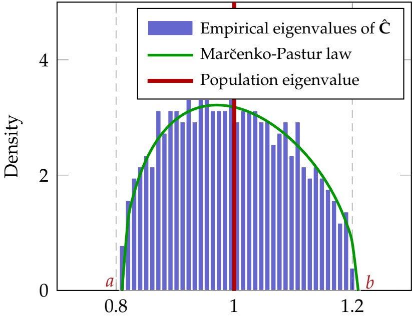


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[a = (1-\sqrt{\mathbf{c}})^2, b = (1+\sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm \mathbf{2}\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the *population* eigenvalue $\mathbf{1}$.

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p), \quad \mathbf{x} = \boldsymbol{\mu} + \mathbf{z};$$
$$\mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p + \mathbf{E}), \quad \mathbf{x} = -\boldsymbol{\mu} + (\mathbf{I}_p + \mathbf{E})^{\frac{1}{2}}\mathbf{z}.$$

  for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

- Neyman-Pearson test: classification is possible only when

$$\boxed{\|\boldsymbol{\mu}\| \geq O(1), \quad \|\mathbf{E}\| \geq O(p^{-1/2}), \quad |\operatorname{tr}\mathbf{E}| \geq O(\sqrt{p}), \quad \|\mathbf{E}\|_F^2 \geq O(1).}$$

- In this non-trivial setting, for $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b$,

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + O(p^{-1/2})$$

  regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$!

- Indeed,

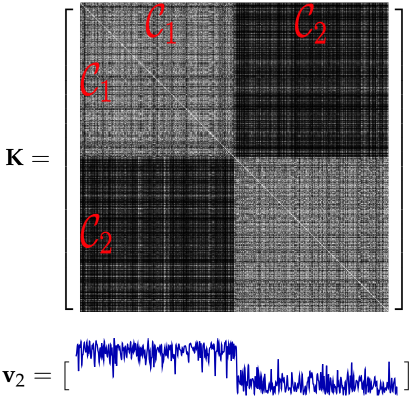
$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - 2 \right\} \to 0}$$

  almost surely as $n, p \to \infty$ (for $n \sim p$ and even $n = p^m$).

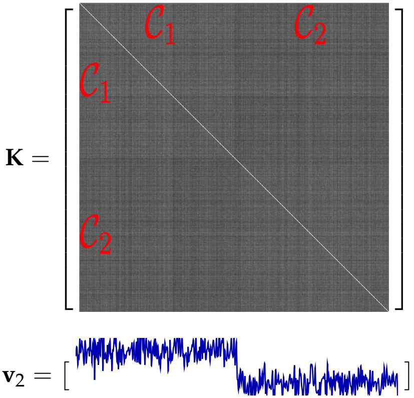
Visualization of kernel matrices for large dimensional data

Objective: "cluster" data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$.
Consider kernel matrix $\mathbf{K}_{ij} = \exp\left(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ and the second top eigenvectors $\mathbf{v}_2$ for small (left) and large (right) dimensional data.

(a) $p = 5, n = 500$



(b) $p = 250, n = 500$



A spectral viewpoint of large kernel matrices

Accumulated effect of small "hidden" statistical information (in $\boldsymbol{\mu}, \mathbf{E}$).

$$\mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\boldsymbol{\mu}, \mathbf{E})\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

with $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$, the class-information vector.

Therefore

- **entry-wise**: for $\mathbf{K}_{ij} = \exp\left(-\frac{1}{2}\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$,

$$\mathbf{K}_{ij} = \exp(-1)\left(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\right) \pm \underbrace{\frac{1}{p}g(\boldsymbol{\mu}, \mathbf{E})}_{O(p^{-1})} + *$$

  so that $\frac{1}{p}g(\boldsymbol{\mu}, \mathbf{E}) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j$;

- **spectrum-wise**: $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\boldsymbol{\mu}, \mathbf{E})\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$ as well!

$\Rightarrow$ With **RMT**, we understand kernel spectral clustering for large dimensional data!

Reminder: random feature maps

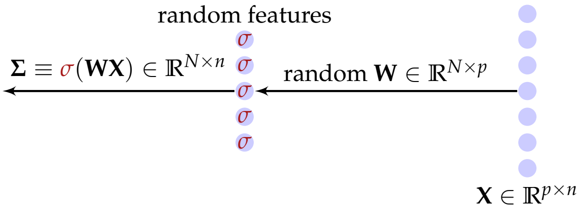


Figure: Illustration of random feature maps

- **Key object**: $\frac{1}{N}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$, correlation in the random feature space.
- **Setting**: $\mathbf{W}_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and n, p, N large.
- $\frac{1}{N}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^N \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i)\sigma(\mathbf{w}_i^\mathsf{T}\mathbf{X})$ for independent $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.
- **Performance guarantee**: if $N \to \infty$ alone, goes to the expected kernel matrix

$$\boxed{\mathbf{K}(\mathbf{X}) \equiv \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})] \in \mathbb{R}^{n \times n}}$$

- of practical (computational and storage) interests only for $N < p$.

Random feature maps for large dimensional data

For $n, p, N \to \infty$ with $n \sim p \sim N$, (again) closely related to $\mathbf{K} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^{\mathsf{T}}\mathbf{w})\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{X})]$.

Eigenspectrum of $\frac{1}{N}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$ [Louart, Liao, Couillet'18]

For all Lipschitz function σ, spectrum of $\frac{1}{N}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$ asymptotically determined by $\bar{\mathbf{Q}}$ via the fixed-point equation

$$\mathbf{Q}(z) \equiv \left(\frac{1}{N}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma} - z\mathbf{I}_n\right)^{-1} \leftrightarrow \bar{\mathbf{Q}}(z) = \left(\frac{\mathbf{K}}{1+\delta(z)} - z\mathbf{I}_n\right)^{-1}, \quad \delta(z) = \frac{1}{N}\operatorname{tr}\mathbf{K}\bar{\mathbf{Q}}(z)$$

for $z \in \mathbb{C}$ not an eigenvalue of $\frac{1}{N}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$.

- for $\mathbf{X} = \mathbf{I}_p$ and $\sigma(t) = t \Rightarrow$ Marčenko-Pastur law
- access to asymptotic performance of e.g. random feature-based ridge regression

Roadmap

$$\mathbf{X} \to \mathbf{\Sigma}(\mathbf{X}) \equiv \sigma(\mathbf{W}\mathbf{X}), \quad \frac{1}{N}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma} \xrightarrow[N\to\infty]{\mathbf{W}\sim\mathcal{N}} \mathbf{K}(\mathbf{X}) = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^{\mathsf{T}}\mathbf{w})\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{X})].$$

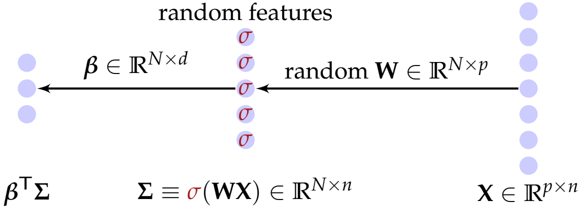


Figure: Illustration of a random feature-based ridge regression

- for a training set $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{p\times n} \times \mathbb{R}^{d\times n}$, $\beta = \frac{1}{n}\mathbf{\Sigma}(\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} + \gamma\mathbf{I}_n)^{-1}\mathbf{Y}^\mathsf{T}$ with regularization factor $\gamma > 0$
- training mean squared error (MSE) $E_{\text{train}} = \frac{1}{n}\|\mathbf{Y} - \beta^\mathsf{T}\mathbf{\Sigma}\|_F^2$
- test error $E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{Y}} - \beta^\mathsf{T}\sigma(\mathbf{W}\hat{\mathbf{X}})\|_F^2$ on a test set $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ of size $\hat{n}$
- can be as a single-hidden-layer neural network model with random weights

Large random feature-based ridge regression: performance mismatch

- if $N \to \infty$ alone ($N \gg p$), $\frac{1}{N}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} \to \mathbf{K}$
- not true for large dimensional data ($p \sim N$) **[Louart, Liao, Couillet'18]**
- $\Rightarrow$ mismatch in performance prediction for MNIST data!

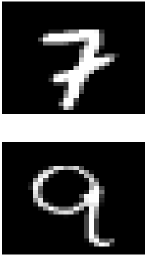


Figure: Example of MNIST images

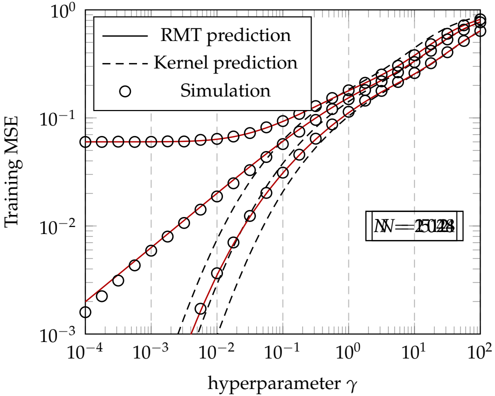
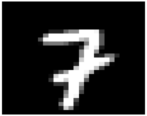


Figure: Training error E_{train} on MNIST data with ReLU activation $\sigma(t) = \max(t, 0)$, $n = \hat{n} = 1024$, $p = 784$.

Asymptotic performance of random feature-based ridge regression



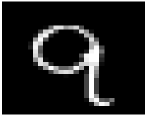


Figure: Example of MNIST images

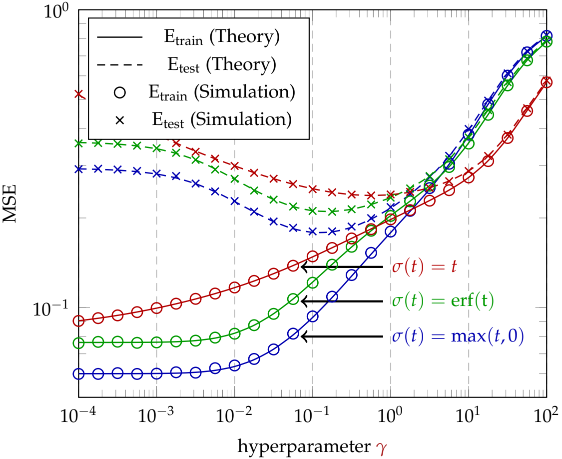


Figure: Performance on MNIST data, $N = 512$, $n = \hat{n} = 1024$, $p = 784$.

$\Rightarrow$ Theoretical understanding and fast tuning of hyperparameter γ!

From random feature maps to kernel matrices

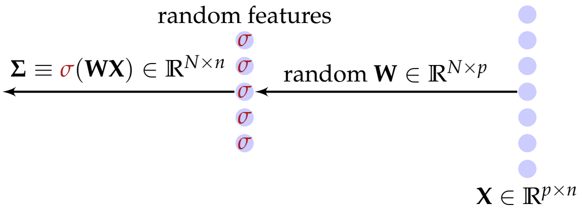


Figure: Illustration of random feature maps

- for $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$ and n,p,N large, $\frac{1}{N}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ closely related to kernel matrix

$$\boxed{\mathbf{K}(\mathbf{X}) \equiv \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]}$$

- explicit $\mathbf{K}$ for commonly used $\sigma(\cdot)$: $\mathrm{ReLU}(t) \equiv \max(t,0)$, sigmoid, quadratic, and exponential $\sigma(t) = \exp(-t^2/2)$

$$\mathbf{K}_{ij} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)] = (2\pi)^{-\frac{p}{2}}\int_{\mathbb{R}^p}\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)e^{-\frac{\|\mathbf{w}\|^2}{2}}d\mathbf{w} \equiv f(\mathbf{x}_i,\mathbf{x}_j).$$

Nonlinearity in simple random neural networks

Table: $\mathbf{K}_{i,j}$ for commonly used $\sigma(\cdot)$, $\angle \equiv \frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$.

$\sigma(t)$	$\mathbf{K}_{i,j} = f(\mathbf{x}_i, \mathbf{x}_j)$
t	$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$
$\max(t,0)$	$\frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\angle \arccos\left(-\angle\right) + \sqrt{1-\angle^2}\right)$
$\|t\|$	$\frac{2}{\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\angle \arcsin\left(\angle\right) + \sqrt{1-\angle^2}\right)$
$\mathrm{sign}(t)$	$\frac{2}{\pi}\arcsin\left(\angle\right)$
$\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$	$\varsigma_2^2(2(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2 + \|\mathbf{x}_i\|^2\|\mathbf{x}_j\|^2) + \varsigma_1^2\mathbf{x}_i^\mathsf{T}\mathbf{x}_j + \varsigma_2\varsigma_0(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2) + \varsigma_0^2$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2\right)\right)\cosh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2\right)\right)\sinh(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$
$\mathrm{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\sqrt{(1+2\|\mathbf{x}_i\|^2)(1+2\|\mathbf{x}_j\|^2)}}\right)$
$\exp(-\frac{t^2}{2})$	$\frac{1}{\sqrt{(1+\|\mathbf{x}_i\|^2)(1+\|\mathbf{x}_j\|^2)-(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2}}$

$\Rightarrow$(still) highly nonlinear functions of the data $\mathbf{x}$!

Roadmap

$$\mathbf{X} \to \mathbf{\Sigma}(\mathbf{X}) \equiv \sigma(\mathbf{W}\mathbf{X}), \quad \frac{1}{N}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} \xrightarrow[N\to\infty]{\mathbf{W}\sim\mathcal{N}} \mathbf{K}(\mathbf{X}) = \{f(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j=1}^n : \quad \boxed{\sigma \to f}.$$

Dig Deeper into $\mathbf{K}$

Objective: simpler and better interpretation of σ (thus f) in $\frac{1}{N}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ (and $\mathbf{K}$).

Data: K-class Gaussian mixture model (GMM)

$$\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \sqrt{p}\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a), \quad \mathbf{x}_i = \boldsymbol{\mu}_a/\sqrt{p} + \mathbf{z}_i$$

with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_a/p)$, $a = 1, \ldots, K$ of statistical mean $\boldsymbol{\mu}_a$ and covariance $\mathbf{C}_a$.

Non-trivial classification (again)

$\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$, $\|\mathbf{C}_a\| = O(1)$, $|\operatorname{tr}(\mathbf{C}_a - \mathbf{C}_b)| = O(\sqrt{p})$, $\|\mathbf{C}_a - \mathbf{C}_b\|_F^2 = O(p)$.

$$\|\mathbf{x}_i\|^2 = \underbrace{\|\mathbf{z}_i\|^2}_{O(1)} + \underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a\|^2 + \frac{2}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\mathbf{z}_i}_{O(p^{-1})} = \underbrace{\frac{1}{p}\operatorname{tr}\mathbf{C}_a}_{O(1)} + \underbrace{\|\mathbf{z}_i\|^2 - \frac{1}{p}\operatorname{tr}\mathbf{C}_a}_{O(p^{-1/2})} + \underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a\|^2 + \frac{2}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\mathbf{z}_i}_{O(p^{-1})}$$

Then for $\mathbf{C}^\circ = \sum_{a=1}^K \frac{n_a}{n}\mathbf{C}_a$ and $\mathbf{C}_a = \mathbf{C}_a^\circ + \mathbf{C}^\circ$, $a = 1, \ldots, K$,

$$\Rightarrow \|\mathbf{x}_i\|^2 = \tau + O(p^{-1/2}) \text{ with } \tau \equiv \frac{1}{p}\operatorname{tr}(\mathbf{C}^\circ), \boxed{\|\mathbf{x}_i - \mathbf{x}_j\|^2 \approx 2\tau} \text{ again!}$$

Understand random feature nonlinearity in classifying GMM

Asymptotic behavior of K [Liao, Couillet'18]

For all σ (and f) listed, we have, as $n \sim p \to \infty$,

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0, \quad \tilde{\mathbf{K}} = d_1(\sigma) \left(\mathbf{Z} + \mathbf{M} \frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right)^\mathsf{T} \left(\mathbf{Z} + \mathbf{M} \frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right) + d_2(\sigma) \mathbf{U} \mathbf{B} \mathbf{U}^\mathsf{T} + d_0 \mathbf{I}_n$$

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

- data structure: $\mathbf{J} \equiv [\mathbf{j}_1, \ldots, \mathbf{j}_K]$, $\mathbf{j}_a$ canonical vector of class $\mathcal{C}_a$;
- randomness of data: $\mathbf{z}$, $\boldsymbol{\phi} = \{\|\mathbf{z}_i\|^2 - \mathbb{E}[\|\mathbf{z}_i\|^2]\}_{i=1}^n$;
- statistical info: $\mathbf{M} \equiv [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$, $\mathbf{t} \equiv \left\{ \operatorname{tr} \mathbf{C}_a^\circ / \sqrt{p} \right\}_{a=1}^K$, $\mathbf{S} \equiv \{ \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b)/p \}_{a,b=1}^K$.

Asymptotic behavior of K [Liao, Couillet'18]

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0, \quad \boxed{\tilde{\mathbf{K}} = d_1(\sigma) \mathbf{A}_1(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b, \mathbf{Z}) + d_2(\sigma) \mathbf{A}_2(\mathbf{C}_a - \mathbf{C}_b, \boldsymbol{\phi}) + *}$$

Roadmap

$$\boldsymbol{\Sigma} = \sigma(\mathbf{W}\mathbf{X}), \frac{1}{N} \boldsymbol{\Sigma}^\mathsf{T} \boldsymbol{\Sigma} \xrightarrow[N \to \infty]{\mathbf{W} \sim \mathcal{N}} \mathbf{K}(\mathbf{X}) = \{ f(\mathbf{x}_i, \mathbf{x}_j) \} \xrightarrow[n,p \to \infty]{\mathbf{X} \sim GMM} \tilde{\mathbf{K}}(d_1, d_2) : \boxed{\sigma \to f \to (d_1, d_2)}.$$

Consequence

$$\boxed{\tilde{\mathbf{K}} = d_1(\sigma)\mathbf{A}_1(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b, \mathbf{Z}) + d_2(\sigma)\mathbf{A}_2(\mathbf{C}_a - \mathbf{C}_b, \boldsymbol{\phi}) + *}$$

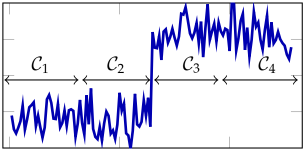
Table: Coefficients (d_1, d_2) in $\tilde{\mathbf{K}}$ for different $\sigma(\cdot)$

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

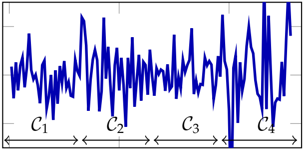
Table: Coefficients (d_1, d_2) in $\tilde{\mathbf{K}}$ for different $\sigma(\cdot)$

Setting: Spectral clustering using $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ on Gaussian mixture data of four classes: $\mathcal{C}_1 : \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, $\mathcal{C}_2 : \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2)$, $\mathcal{C}_3 : \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1)$ and $\mathcal{C}_4 : \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ with different $\sigma(\cdot)$.

Mean-oriented: linear map $\sigma(t) = t \Rightarrow \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2), \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1), \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$.

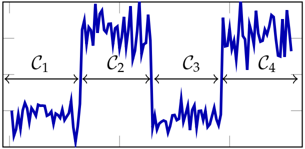


Eigenvector 1

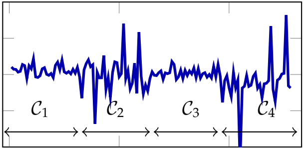


Eigenvector 2

Cov-oriented: $\sigma(t) = |t| \Rightarrow \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2), \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1), \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$.



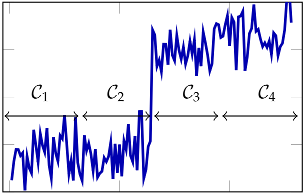
Eigenvector 1



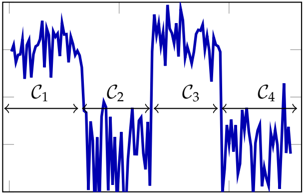
Eigenvector 2

Random-feature based spectral clustering: Gaussian data

"Balanced": the ReLU function $\sigma(t) = \max(t, 0)$.

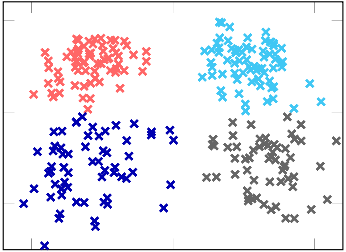


Eigenvector 1



Eigenvector 2

Eigenvector 2



Eigenvector 1

Random-feature based spectral clustering: real datasets

Figure: The MNIST image database.

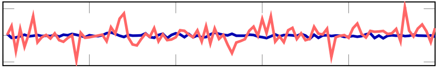


Figure: The epileptic EEG datasets.[1]

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

Random-feature based spectral clustering: real datasets

Table: Empirical estimation of statistical information of the MNIST and EEG datasets.

	$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$	$\|\mathbf{C}_1 - \mathbf{C}_2\|$
MNIST data	**391.1**	83.8
EEG data	2.4	**14.5**

Table: Clustering accuracies on MNIST.

	$\sigma(t)$	$n = 64$	$n = 128$
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**%
cov-oriented	$\|t\|$	60.41%	57.81%
	$\cos(t)$	59.56%	57.72%
	$\exp(-t^2/2)$	60.44%	58.67%
balanced	$\mathrm{ReLU}(t)$	85.72%	82.27%

Table: Clustering accuracies on EEG.

	$\sigma(t)$	$n = 64$	$n = 128$
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%
cov-oriented	$\|t\|$	99.69%	99.50%
	$\cos(t)$	99.38%	99.36%
	$\exp(-t^2/2)$	**99.81**%	**99.77**%
balanced	$\mathrm{ReLU}(t)$	87.91%	90.97%

Conclusion and limitations

Conclusion on large dimensional random feature maps:

Roadmap

$$
\begin{tikzcd}[column sep=huge, row sep=large]
\underbrace{\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}}_{\sigma} \arrow[r, "\mathbf{W}\sim\mathcal{N},N\to\infty"] & \underbrace{\mathbf{K}(\mathbf{X})}_{f} \arrow[r, "\mathbf{X}\sim GMM,n,p\to\infty"] \arrow[d, "\mathbf{W}\sim\mathcal{N},n\sim p\sim N"] & \underbrace{\tilde{\mathbf{K}}}_{(d_1,d_2)} \arrow[d] \\
& \text{RF-based ridge regression} & \text{RF-based spectral clustering}
\end{tikzcd}
$$

Limitations:

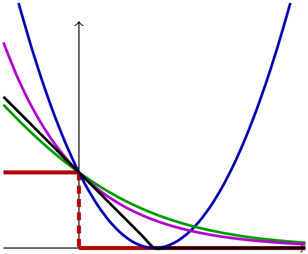
- optimization-based problems with implicit solution
- limited to Gaussian data

A random matrix framework to optimization-based learning problem

Problem of *empirical risk minimization*: for $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, find classifier $\boldsymbol{\beta}$ such that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i)$$

for some nonnegative convex loss ℓ.



- logistic regression: $\ell(t) = \log(1 + e^{-t})$
- least squares: $\ell(t) = (t-1)^2$
- boosting algorithm: $\ell(t) = e^{-t}$
- SVM: $\ell(t) = \max(1-t, 0)$

No closed-form solution, **RMT** provides tools to assess the performance **[Mai, Liao'19]**.

Limitations:

✔ optimization-based problems with implicit solution: yes if convex!

? limited to Gaussian data

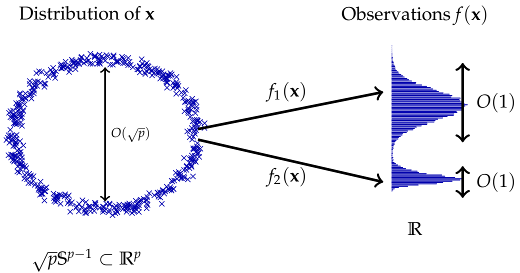
From theory to practice: concentrated random vectors

RMT often assumes $\mathbf{x}$ are affine maps $\mathbf{Az} + \mathbf{b}$ of $\mathbf{z} \in \mathbb{R}^p$ with i.i.d. entries.

Concentrated random vectors

For a certain family of functions $f : \mathbb{R}^p \mapsto \mathbb{R}$, there exists deterministic $m_f \in \mathbb{R}$

$$P\left(|f(\mathbf{x}) - m_f| > \epsilon\right) \leq e^{-g(\epsilon)}, \quad \text{for some strictly increasing function } g.$$



$\Rightarrow$The theory remains valid for concentrated random vectors and for almost real images **[Seddik, Tamaazousti, Couillet'19]**!

From concentrated random vectors to GANs

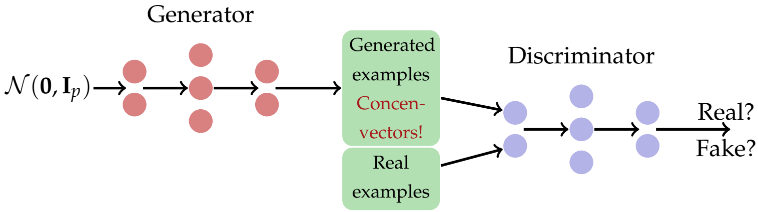


Figure: Illustration of a generative adversarial network (GAN).

Figure: Images samples generated by BigGAN [Brock *et al.*'18].

Limitations:

✔ optimization-based problems with implicit solution: yes if convex!

✔ limited to Gaussian data: to concentrated vectors and almost real images!

Some clues . . . and much more can be done!

RMT as a tool to **analyze**, **understand** and **improve**
large dimensional machine learning methods.

- powerful and flexible tool to assess matrix-based machine learning systems;
- study (convex) optimization-based learning methods, e.g., logistic regression;
- understand impact of optimization methods, the dynamics of gradient descent;
- non-convex problems (e.g, deep neural nets) are more difficult, but accessible in some cases, e.g., low rank matrix recovery, phase retrieval, etc;
- even **more** to be done: transfer learning, active learning, generative models, graph-based methods, robust statistics, etc.

Contributions during Ph.D.

Publications:

J1 C. Louart, **Z. Liao**, and R. Couillet. "A Random Matrix Approach to Neural Networks". *The Annals of Applied Probability*, 28(2) :1190–1248, 2018.

J2 **Z. Liao**, R. Couillet. "A Large Dimensional Analysis of Least Squares Support Vector Machines", *IEEE Transactions on Signal Processing* 67 (4), 1065-1074, 2019.

J3 X. Mai and **Z. Liao**. High Dimensional Classification via Empirical Risk Minimization: Improvements and Optimality. (submitted to) *IEEE Transactions on Signal Processing*, 2019.

J4 Y. Chitour, **Z. Liao**, R. Couillet. "A Geometric Approach of Gradient Descent Algorithms in Neural Networks", (submitted to) *Journal of Differential Equations*, 2019.

C1 **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.

C2 **Z. Liao**, R. Couillet. "On the Spectrum of Random Features Maps of High Dimensional Data". *International Conference on Machine Learning (ICML'18)*, Stockholm, Sweden, 2018.

C3 **Z. Liao**, R. Couillet, "The Dynamics of Learning: A Random Matrix Approach", *International Conference on Machine Learning (ICML'18)*, Stockholm, Sweden, 2018.

C4 X. Mai, **Z. Liao**, R. Couillet. "A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*, Brighton, UK, 2019.

C5 **Z. Liao**, R. Couillet. "On Inner-Product Kernels of High Dimensional Data", *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'19)*, Guadeloupe, France, 2019.

Contributions during Ph.D.

Invited talks and tutorials:

- Invited talks at
  - DIMACS center, Rutgers University, USA
  - Matrix series conference, Krakow, Poland
  - iCODE institute, Paris-Saclay, France
  - Shanghai Jiao Tong University, China
  - HUAWEI
- Tutorial on “Random Matrix Advances in Machine Learning and Neural Nets” (with R. Couillet and X. Mai), *The 26th European Signal Processing Conference (EUSIPCO'18)*, Roma, Italy, 2018.

Reviewing activities:

- ICML, NeurIPS, AAAI, IEEE-TSP.

Thank you!

For more information, visit `https://zhenyu-liao.github.io`!