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
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.

large dimensional data $x_1, \ldots, x_n \in \mathbb{R}^p$
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 $x_i \sim \mathcal{N}(0, C)$, estimate population covariance $C$ from $n$ data samples $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$.

- Maximum likelihood sample covariance matrix:

$$
\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} XX^T \in \mathbb{R}^{p \times p}
$$

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 $C = I_p$ with $n < p$, $\hat{C}$ has at least $p - n$ zero eigenvalues.

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

$\Rightarrow$ 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

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

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

![Empirical eigenvalues of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500, n = 50000$.](image)

**Figure**: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500, n = 50000$.

- eigenvalues span on $[a = (1 - \sqrt{c})^2, b = (1 + \sqrt{c})^2]$.
- for $n = 100p$, on a range of $\pm 2\sqrt{c} = \pm 0.2$ around the population eigenvalue 1.
"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification:
  \[ C_1 : x \sim \mathcal{N}(\mu, I_p), \quad x = \mu + z; \]
  \[ C_2 : x \sim \mathcal{N}(-\mu, I_p + E), \quad x = -\mu + (I_p + E)^{1/2}z. \]
  for \( z \sim \mathcal{N}(0, I_p) \).

- Neyman-Pearson test: classification is possible only when
  \[
  \|\mu\| \geq O(1), \quad \|E\| \geq O(p^{-1/2}), \quad |\text{tr} E| \geq O(\sqrt{p}), \quad \|E\|_F \geq O(1).
  \]

- In this non-trivial setting, for \( x_i \in C_a, x_j \in C_b \),
  \[
  \frac{1}{p} \|x_i - x_j\|^2 = \frac{1}{p} \|z_i - z_j\|^2 + O(p^{-1/2})
  \]
  regardless of the classes \( C_a, C_b \)!

- Indeed,
  \[
  \max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|x_i - x_j\|^2 - 2 \right\} \to 0
  \]
  almost surely as \( n, p \to \infty \) (for \( n \sim p \) and even \( n = p^m \)).
Objective: “cluster” data $x_1, \ldots, x_n \in \mathbb{R}^p$ into $C_1$ or $C_2$.
Consider kernel matrix $K_{ij} = \exp \left( -\frac{1}{2p} \|x_i - x_j\|^2 \right)$ and the second top eigenvectors $v_2$ 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 $\mu, E$).

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

with $Z = [z_1, \ldots, z_n] \in \mathbb{R}^{p \times n}$ and $j = [1_{n/2}; -1_{n/2}]$, the class-information vector.

Therefore

- **entry-wise:** for $K_{ij} = \exp \left( -\frac{1}{2} \frac{1}{p} \|x_i - x_j\|^2 \right)$,

  $$K_{ij} = \exp(-1) \left( 1 + \frac{1}{p} z_i^T z_j \right) \pm \frac{1}{p} g(\mu, E) + \ast$$

  $O(p^{-1/2})$ $O(p^{-1})$

  so that $\frac{1}{p} g(\mu, E) \ll \frac{1}{p} z_i^T z_j$.

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

⇒ With RMT, we understand kernel spectral clustering for large dimensional data!
Reminder: random feature maps

\[ \Sigma \equiv \sigma(WX) \in \mathbb{R}^{N \times n} \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]

\[ \sigma \]
Random feature maps for large dimensional data

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

**Eigenspectrum of \( \frac{1}{N} \Sigma^T \Sigma \) [Louart, Liao, Couillet’18]**

For all Lipschitz function \( \sigma \), spectrum of \( \frac{1}{N} \Sigma^T \Sigma \) asymptotically determined by \( \bar{Q} \) via the fixed-point equation

\[
Q(z) \equiv \left( \frac{1}{N} \Sigma^T \Sigma - zI_n \right)^{-1} \leftrightarrow \bar{Q}(z) = \left( \frac{K}{1 + \delta(z)} - zI_n \right)^{-1}, \quad \delta(z) = \frac{1}{N} \text{tr} K \bar{Q}(z)
\]

for \( z \in \mathbb{C} \) not an eigenvalue of \( \frac{1}{N} \Sigma^T \Sigma \).

- for \( X = I_p \) and \( \sigma(t) = t \) ⇒ Marčenko-Pastur law
- access to asymptotic performance of e.g. random feature-based ridge regression

**Roadmap**

\[
X \to \Sigma(X) \equiv \sigma(WX), \quad \frac{1}{N} \Sigma^T \Sigma \overset{W \sim \mathcal{N}}{\to} N \to \infty K(X) = \mathbb{E}_w[\sigma(X^T w)\sigma(w^T X)].
\]
Application: large random feature-based ridge regression

\[ \mathbf{x} \in \mathbb{R}^{p \times n} \]
\[ \mathbf{\Sigma} \equiv \sigma(\mathbf{Wx}) \in \mathbb{R}^{N \times n} \]
\[ \mathbf{X} \in \mathbb{R}^{p \times n} \]

**Figure:** Illustration of a random feature-based ridge regression

- for a training set \((\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{p \times n} \times \mathbb{R}^{d \times n}\), \( \mathbf{\beta} = \frac{1}{n} \mathbf{\Sigma} \left( \frac{1}{n} \mathbf{\Sigma}^T \mathbf{\Sigma} + \gamma \mathbf{I}_n \right)^{-1} \mathbf{Y}^T \) with regularization factor \( \gamma > 0 \)
- training mean squared error (MSE) \( E_{\text{train}} = \frac{1}{n} \| \mathbf{Y} - \mathbf{\beta}^T \mathbf{\Sigma} \|_F^2 \)
- test error \( E_{\text{test}} = \frac{1}{\hat{n}} \| \hat{\mathbf{Y}} - \hat{\mathbf{\beta}}^T \sigma(\mathbf{W}\hat{\mathbf{x}}) \|_F^2 \) 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 ($N \gg p$), $\frac{1}{N} \Sigma^T \Sigma \to K$
- 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 $E_{\text{train}}$ 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

![Example of MNIST images](image1)

![Example of MNIST images](image2)

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

$\Rightarrow$ Theoretical understanding and fast tuning of hyperparameter $\gamma$!
From random feature maps to kernel matrices

\[ \Sigma \equiv \sigma(WX) \in \mathbb{R}^{N \times n} \]

random features

\[ W \in \mathbb{R}^{N \times p} \]

random \[ \sigma \]

\[ \Sigma \equiv \sigma(WX) \in \mathbb{R}^{N \times n} \]

\[ \mathbf{X} \in \mathbb{R}^{p \times n} \]

**Figure:** Illustration of random feature maps

- for \( W_{ij} \sim \mathcal{N}(0,1) \) and \( n, p, N \) large, \( \frac{1}{N} \Sigma^T \Sigma \) closely related to kernel matrix

\[
\mathbf{K}(\mathbf{X}) \equiv \mathbb{E}_{w \sim \mathcal{N}(0, I_p)} [\sigma(\mathbf{X}^T w)\sigma(w^T \mathbf{X})]
\]

- explicit \( \mathbf{K} \) for commonly used \( \sigma(\cdot) \): ReLU

\( (t) \equiv \max(t, 0) \), sigmoid, quadratic, and exponential

\( \sigma(t) = \exp(-t^2/2) \)

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

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

<table>
<thead>
<tr>
<th>$\sigma(t)$</th>
<th>$K_{i,j} = f(x_i, x_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$\frac{x_i^T x_j}{|x_i| |x_j|}$</td>
</tr>
<tr>
<td>$\max(t, 0)$</td>
<td>$\frac{1}{2\pi} |x_i| |x_j| (\angle \arccos (-\angle) + \sqrt{1 - \angle^2})$</td>
</tr>
<tr>
<td>$</td>
<td>t</td>
</tr>
<tr>
<td>$\text{sign}(t)$</td>
<td>$\frac{2}{\pi} \arcsin (\angle)$</td>
</tr>
<tr>
<td>$\zeta_2 t^2 + \zeta_1 t + \zeta_0$</td>
<td>$\zeta_2^2 (2(x_i^T x_j)^2 + |x_i|^2 |x_j|^2) + \zeta_1^2 x_i^T x_j + \zeta_2 \zeta_0 (|x_i|^2 + |x_j|^2) + \zeta_0^2$</td>
</tr>
<tr>
<td>$\cos(t)$</td>
<td>$\exp (-\frac{1}{2} (|x_i|^2 + |x_j|^2)) \cosh(x_i^T x_j)$</td>
</tr>
<tr>
<td>$\sin(t)$</td>
<td>$\exp (-\frac{1}{2} (|x_i|^2 + |x_j|^2)) \sinh(x_i^T x_j)$</td>
</tr>
<tr>
<td>$\text{erf}(t)$</td>
<td>$\frac{2}{\pi} \arcsin \left( \frac{\sqrt{(1+2|x_i|^2)(1+2|x_j|^2)}}{2x_i^T x_j} \right)$</td>
</tr>
<tr>
<td>$\exp(-\frac{t^2}{2})$</td>
<td>$\frac{1}{\sqrt{(1+|x_i|^2)(1+|x_j|^2)-(x_i^T x_j)^2}}$</td>
</tr>
</tbody>
</table>

⇒(still) highly **nonlinear** functions of the data $x$!

Roadmap

$X \rightarrow \Sigma(X) \equiv \sigma(WX), \quad \frac{1}{N} \Sigma^T \Sigma \xrightarrow{N \rightarrow \infty} \mathbf{K}(X) = \{f(x_i, x_j)\}_{i,j=1}^n : \sigma \rightarrow f$. 

Z. Liao (CentraleSupélec)
Dig Deeper into $\mathbf{K}$

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

**Data: $K$-class Gaussian mixture model (GMM)**

$$x_i \in C_a \iff \sqrt{p}x_i \sim \mathcal{N}(\mu_a, C_a), \quad x_i = \mu_a / \sqrt{p} + z_i$$

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

**Non-trivial classification (again)**

$$\|\mu_a - \mu_b\| = O(1), \|C_a\| = O(1), |\text{tr}(C_a - C_b)| = O(\sqrt{p}), |C_a - C_b|_F^2 = O(p).$$

\[
\|x_i\|^2 = \underbrace{\|z_i\|^2}_{O(1)} + \frac{1}{p}\|\mu_a\|^2 + \frac{2}{p}\mu_a^T z_i = \frac{1}{p}\text{tr} C_a + \underbrace{\|z_i\|^2 - \frac{1}{p}\text{tr} C_a}_{O(p^{-1/2})} + \underbrace{\frac{1}{p}\|\mu_a\|^2}_{O(1)} + \frac{2}{\sqrt{p}}\mu_a^T z_i
\]

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

$$\Rightarrow \|x_i\|^2 = \tau + O(p^{-1/2}) \text{ with } \tau \equiv \frac{1}{p}\text{tr}(C^\circ), \quad \|x_i - x_j\|^2 \approx 2\tau \text{ again!}$$
Understand random feature nonlinearity in classifying GMM

Asymptotic behavior of $K$ [Liao, Couillet’18]

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

$$\|K - \tilde{K}\| \to 0, \quad \tilde{K} = d_1(\sigma) \left( Z + M \frac{J^T}{\sqrt{p}} \right)^T \left( Z + M \frac{J^T}{\sqrt{p}} \right) + d_2(\sigma)UBU^T + d_0I_n$$

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

- data structure: $J \equiv [j_1, \ldots, j_K]$, $j_a$ canonical vector of class $C_a$;
- randomness of data: $z, \phi = \{\|z_i\|^2 - \mathbb{E}[\|z_i\|^2]\}_{i=1}^n$;
- statistical info: $M \equiv [\mu_1, \ldots, \mu_K]$, $t \equiv \{\text{tr}\, C_a^T / \sqrt{p}\}_{a=1}^K$, $S \equiv \{\text{tr}(C_a C_b) / p\}_{a,b=1}^K$.

Asymptotic behavior of $K$ [Liao, Couillet’18]

$$\|K - \tilde{K}\| \to 0, \quad \tilde{K} = d_1(\sigma)A_1(\mu_a - \mu_b, Z) + d_2(\sigma)A_2(C_a - C_b, \phi) + *$$

Roadmap

$$\Sigma = \sigma(WX), \quad \frac{1}{N} \Sigma^T \Sigma \xrightarrow{W \sim \mathcal{N}} K(X) = \{f(x_i, x_j)\} \quad \xrightarrow{X \sim \text{GMM}} \tilde{K}(d_1, d_2) : \sigma \to f \to (d_1, d_2).$$
\[
\tilde{K} = d_1(\sigma)A_1(\mu_a - \mu_b, Z) + d_2(\sigma)A_2(C_a - C_b, \phi) + *
\]

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

<table>
<thead>
<tr>
<th>(\sigma(t))</th>
<th>(d_1)</th>
<th>(d_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(\max(t, 0))</td>
<td>(\frac{1}{4})</td>
<td>(\frac{1}{8\pi \tau})</td>
</tr>
<tr>
<td>(</td>
<td>t</td>
<td>)</td>
</tr>
<tr>
<td>(\text{sign}(t))</td>
<td>(\frac{2}{\pi \tau})</td>
<td>0</td>
</tr>
<tr>
<td>(\zeta_2 t^2 + \zeta_1 t + \zeta_0)</td>
<td>(\zeta_1^2)</td>
<td>(\zeta_2^2)</td>
</tr>
<tr>
<td>(\cos(t))</td>
<td>0</td>
<td>(e^{-\tau}/4)</td>
</tr>
<tr>
<td>(\sin(t))</td>
<td>(e^{-\tau})</td>
<td>0</td>
</tr>
<tr>
<td>(\text{erf}(t))</td>
<td>(\frac{4}{\pi} \frac{1}{2\tau + 1})</td>
<td>0</td>
</tr>
<tr>
<td>(\exp(-t^2/2))</td>
<td>0</td>
<td>(\frac{1}{4(\tau+1)^3})</td>
</tr>
</tbody>
</table>

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

- **Mean-oriented:** \(d_1 \neq 0, d_2 = 0\) → separate with difference in \(M\);
- **Cov-oriented:** \(d_1 = 0, d_2 \neq 0\) → exploit differences in cov;
- **"Balanced," both \(d_1, d_2 \neq 0\):** \(\text{ReLU max}(t, 0)\) and quadratic → make use of both statistics!
Random-feature based spectral clustering: Gaussian data

**Setting:** Spectral clustering using $\frac{1}{n} \Sigma^T \Sigma$ on Gaussian mixture data of four classes: $C_1 : \mathcal{N}(\mu_1, C_1), C_2 : \mathcal{N}(\mu_1, C_2), C_3 : \mathcal{N}(\mu_2, C_1)$ and $C_4 : \mathcal{N}(\mu_2, C_2)$ with different $\sigma(\cdot)$.

**Mean-oriented:** linear map $\sigma(t) = t \Rightarrow \mathcal{N}(\mu_1, C_1), \mathcal{N}(\mu_1, C_2), \mathcal{N}(\mu_2, C_1), \mathcal{N}(\mu_2, C_2)$.

**Cov-oriented:** $\sigma(t) = |t| \Rightarrow \mathcal{N}(\mu_1, C_1), \mathcal{N}(\mu_1, C_2), \mathcal{N}(\mu_2, C_1), \mathcal{N}(\mu_2, C_2)$.
Random-feature based spectral clustering: Gaussian data

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

\[ \sigma(t) = \max(t, 0) \]

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.

<table>
<thead>
<tr>
<th></th>
<th>$|\mu_1 - \mu_2|^2$</th>
<th>$|C_1 - C_2|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST data</td>
<td>391.1</td>
<td>83.8</td>
</tr>
<tr>
<td>EEG data</td>
<td>2.4</td>
<td>14.5</td>
</tr>
</tbody>
</table>

Table: Clustering accuracies on MNIST.

<table>
<thead>
<tr>
<th></th>
<th>$n = 64$</th>
<th>$n = 128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean-oriented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>88.94%</td>
<td>87.30%</td>
</tr>
<tr>
<td>$1_{t&gt;0}$</td>
<td>82.94%</td>
<td>85.56%</td>
</tr>
<tr>
<td>sign($t$)</td>
<td>83.34%</td>
<td>85.22%</td>
</tr>
<tr>
<td>sin($t$)</td>
<td>87.81%</td>
<td>87.50%</td>
</tr>
<tr>
<td>cov-oriented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>t</td>
<td>$</td>
</tr>
<tr>
<td>cos($t$)</td>
<td>59.56%</td>
<td>57.72%</td>
</tr>
<tr>
<td>exp($-t^2/2$)</td>
<td>60.44%</td>
<td>58.67%</td>
</tr>
<tr>
<td>balanced</td>
<td>ReLU($t$)</td>
<td>85.72%</td>
</tr>
</tbody>
</table>

Table: Clustering accuracies on EEG.

<table>
<thead>
<tr>
<th></th>
<th>$n = 64$</th>
<th>$n = 128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean-oriented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>70.31%</td>
<td>69.58%</td>
</tr>
<tr>
<td>$1_{t&gt;0}$</td>
<td>65.87%</td>
<td>63.47%</td>
</tr>
<tr>
<td>sign($t$)</td>
<td>64.63%</td>
<td>63.03%</td>
</tr>
<tr>
<td>sin($t$)</td>
<td>70.34%</td>
<td>68.22%</td>
</tr>
<tr>
<td>cov-oriented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>t</td>
<td>$</td>
</tr>
<tr>
<td>cos($t$)</td>
<td>99.38%</td>
<td>99.36%</td>
</tr>
<tr>
<td>exp($-t^2/2$)</td>
<td>99.81%</td>
<td>99.77%</td>
</tr>
<tr>
<td>balanced</td>
<td>ReLU($t$)</td>
<td>87.91%</td>
</tr>
</tbody>
</table>
Conclusion and limitations

**Conclusion** on large dimensional random feature maps:

**Roadmap**

\[
\sigma \sum^{T} \sum \xrightarrow{\text{W}} \mathcal{N}, N \rightarrow \infty \xrightarrow{\text{K}(X)} \mathcal{X} \sim \text{GMM}, n, p \rightarrow \infty \xrightarrow{\tilde{K}} (d_1, d_2)
\]

\[
\text{W} \sim \mathcal{N}, n \sim p \sim N \quad \text{RF-based ridge regression} \quad \text{RF-based spectral clustering}
\]

**Limitations:**

- optimization-based problems with implicit solution
- limited to Gaussian data
Problem of \textit{empirical risk minimization}: for \( \{(x_i, y_i)\}_{i=1}^{n}, x_i \in \mathbb{R}^p, y_i \in \{-1, +1\} \), find classifier \( \beta \) such that

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \beta^T x_i)
\]

for some nonnegative convex loss \( \ell \).

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

No \text{closed-form solution}, \text{RMT provides tools to assess the performance} [\textit{Mai, Liao’19}].

\textbf{Limitations:}

- \text{optimization-based problems with \textit{implicit} solution: yes if convex!}
- \text{limited to Gaussian data}
From theory to practice: concentrated random vectors

RMT often assumes $\mathbf{x}$ are affine maps $A\mathbf{z} + \mathbf{b}$ 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 $m_f \in \mathbb{R}$

$$P \left( |f(\mathbf{x}) - m_f| > \epsilon \right) \leq e^{-g(\epsilon)},$$

for some strictly increasing function $g$.

$\sqrt{p}S^{p-1} \subset \mathbb{R}^p$

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

Generator

$\mathcal{N}(0, I_p)$

Discriminator

Real?
Fake?

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:


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


Reviewing activities:

- ICML, NeurIPS, AAAI, IEEE-TSP.
Thank you!

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