A Random Matrix Framework for BigData Machine Learning
(Groupe Deep Learning, DigiCosme)

Romain COUILLET

CentraleSupélec, France

June, 2017
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Outline

Basics of Random Matrix Theory
- Motivation: Large Sample Covariance Matrices
- Spiked Models

Applications
- Reminder on Spectral Clustering Methods
- Kernel Spectral Clustering
- Semi-supervised Learning
- Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
    Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Context

**Baseline scenario:** \( y_1, \ldots, y_n \in \mathbb{C}^p \) (or \( \mathbb{R}^p \)) i.i.d. with \( E[y_1] = 0, \ E[y_1 y_1^*] = C_p \):
**Context**

**Baseline scenario:** $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0, E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for $C_p$ is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).
Context

**Baseline scenario:** $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for $C_p$ is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).
- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_p \xrightarrow{a.s.} C_p.$$  

or equivalently, in spectral norm

$$\|\hat{C}_p - C_p\| \xrightarrow{a.s.} 0.$$
Context

**Baseline scenario:** \( y_1, \ldots, y_n \in \mathbb{C}^p \) (or \( \mathbb{R}^p \)) i.i.d. with \( E[y_1] = 0, \ E[y_1 y_1^*] = C_p \):

- If \( y_1 \sim \mathcal{N}(0, C_p) \), ML estimator for \( C_p \) is the sample covariance matrix (SCM)

\[
\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^* 
\]

\( (Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^p \times n) \).

- If \( n \to \infty \), then, **strong law of large numbers**

\[
\hat{C}_p \overset{a.s.}{\to} C_p.
\]

or equivalently, in spectral norm

\[
\|\hat{C}_p - C_p\| \overset{a.s.}{\to} 0.
\]

**Random Matrix Regime**

- No longer valid if \( p, n \to \infty \) with \( p/n \to c \in (0, \infty) \),

\[
\|\hat{C}_p - C_p\| \not\to 0.
\]
Context

Baseline scenario: \( y_1, \ldots, y_n \in \mathbb{C}^p \) (or \( \mathbb{R}^p \)) i.i.d. with \( \mathbb{E}[y_1] = 0, \mathbb{E}[y_1 y_1^*] = C_p \):

- If \( y_1 \sim \mathcal{N}(0, C_p) \), ML estimator for \( C_p \) is the sample covariance matrix (SCM)

\[
\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*
\]

\((Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}).\)

- If \( n \to \infty \), then, strong law of large numbers

\[
\hat{C}_p \xrightarrow{a.s.} C_p.
\]

or equivalently, in spectral norm

\[
\|\hat{C}_p - C_p\| \xrightarrow{a.s.} 0.
\]

Random Matrix Regime

- No longer valid if \( p, n \to \infty \) with \( p/n \to c \in (0, \infty) \),

\[
\|\hat{C}_p - C_p\| \not\to 0.
\]

- For practical \( p, n \) with \( p \asymp n \), leads to dramatically wrong conclusions
The Marčenko–Pastur law

Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $p = 500$, $n = 2000$, $C_p = I_p$. 
The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) \( \mu_p \) of Hermitian matrix \( A_p \in \mathbb{C}^{p \times p} \) is

\[
\mu_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(A_p)}.
\]
The Marčenko–Pastur law

Definition (Empirical Spectral Density)
Empirical spectral density \(\mu_p\) of Hermitian matrix \(A_p \in \mathbb{C}^{p \times p}\) is

\[
\mu_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(A_p)}.
\]

Theorem (Marčenko–Pastur Law \[\text{[Marčenko,Pastur’67]}\])
\(X_p \in \mathbb{C}^{p \times n}\) with i.i.d. zero mean, unit variance entries.
As \(p, n \to \infty\) with \(p/n \to c \in (0, \infty)\), e.s.d. \(\mu_p\) of \(\frac{1}{n}X_pX_p^*\) satisfies

\[
\mu_p \xrightarrow{\text{a.s.}} \mu_c
\]
weakly, where

\[
\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}
\]
The Marčenko–Pastur law

Definition (Empirical Spectral Density)
Empirical spectral density (e.s.d.) $\mu_p$ of Hermitian matrix $A_p \in \mathbb{C}^{p \times p}$ is

$$\mu_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(A_p)}.$$ 

Theorem (Marčenko–Pastur Law [Marčenko, Pastur’67])
$X_p \in \mathbb{C}^{p \times n}$ with i.i.d. zero mean, unit variance entries.
As $p, n \to \infty$ with $p/n \to c \in (0, \infty)$, e.s.d. $\mu_p$ of $\frac{1}{n} X_p X_p^*$ satisfies $\mu_p \xrightarrow{a.s.} \mu_c$ weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$
- on $(0, \infty)$, $\mu_c$ has continuous density $f_c$ supported on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$

$$f_c(x) = \frac{1}{2\pi cx} \sqrt{(x - (1 - \sqrt{c})^2)((1 + \sqrt{c})^2 - x)}.$$
The Marčenko–Pastur law

Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p \to \infty} p/n$. 
The Marčenko–Pastur law

Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p \to \infty} p/n$. 
The Marčenko–Pastur law

Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p \to \infty} p/n$. 
Outline

Basics of Random Matrix Theory
- Motivation: Large Sample Covariance Matrices
- Spiked Models

Applications
- Reminder on Spectral Clustering Methods
- Kernel Spectral Clustering
- Semi-supervised Learning
- Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Spiked Models

If we break:

- **Small rank Perturbation**: \( C_p = I_P + P, \) \( P \) of low rank.

\[
\begin{align*}
1 + \omega_1 + c \frac{1 + \omega_1}{\omega_1}, \\
1 + \omega_2 + c \frac{1 + \omega_2}{\omega_2}
\end{align*}
\]

\( \lambda_i \) \( i = 1, \ldots, n \)

**Figure**: Eigenvalues of \( \frac{1}{n} Y_p Y_p^* \), \( C_p = \text{diag}(1, \ldots, 1, 2, 2, 3, 3) \), \( p = 500, \) \( n = 1500 \).
Spiked Models

Theorem (Eigenvalues [Baik, Silverstein’06])

Let $Y_p = C_p^{1/2} X_p$, with

- $X_p$ with i.i.d. zero mean, unit variance, $E[|X_p|_i^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^*$, where, for $K$ fixed,

  $$\Omega = \text{diag}(\omega_1, \ldots, \omega_K) \in \mathbb{R}^{K \times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$
Theorem (Eigenvalues [Baik,Silverstein’06])

Let $Y_p = C_p^{1/2} X_p$, with

- $X_p$ with i.i.d. zero mean, unit variance, $E[|X_p|_4^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^*$, where, for $K$ fixed,

$$\Omega = \text{diag} (\omega_1, \ldots, \omega_K) \in \mathbb{R}^{K \times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$ 

Then, as $p, n \to \infty$, $p/n \to c \in (0, \infty)$, denoting $\lambda_m = \lambda_m(\frac{1}{n}Y_p Y_p^*)$ ($\lambda_m > \lambda_{m+1}$),

$$\lambda_m \overset{a.s.}{\to} \begin{cases} 1 + \omega_m + c \frac{1+\omega_m}{\omega_m} > (1 + \sqrt{c})^2 & , \omega_m > \sqrt{c} \\ (1 + \sqrt{c})^2 & , \omega_m \in (0, \sqrt{c}] \end{cases}.$$
Theorem (Eigenvectors [Paul’07])

Let $Y_p = C_p^{1/2} X_p$, with

- $X_p$ with i.i.d. zero mean, unit variance, $E[|X_p|_i^4] < \infty$.
- $C_p = I_p + P$, $P = U \Omega U^* = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$. 
Theorem (Eigenvalues [Paul’07])

Let \( Y_p = C_p^{1/2} X_p \), with
- \( X_p \) with i.i.d. zero mean, unit variance, \( E[|X_p|_i^4] < \infty \).
- \( C_p = I_p + P \), \( P = U \Omega U^* = \sum_{i=1}^{K} \omega_i u_i u_i^* \), \( \omega_1 > \ldots > \omega_M > 0 \).

Then, as \( p, n \to \infty \), \( p/n \to c \in (0, \infty) \), for \( a, b \in \mathbb{C}^p \) deterministic and \( \hat{u}_i \) eigenvector of \( \lambda_i \left( \frac{1}{n} Y_p Y_p^* \right) \),

\[
    a^* \hat{u}_i \hat{u}_i^* b - \frac{1 - c \omega_i^{-2}}{1 + c \omega_i^{-1}} a^* u_i u_i^* b \cdot 1_{\omega_i > \sqrt{c}} \xrightarrow{a.s.} 0
\]

In particular,

\[
    |\hat{u}_i^* u_i|^2 \xrightarrow{a.s.} \frac{1 - c \omega_i^{-2}}{1 + c \omega_i^{-1}} \cdot 1_{\omega_i > \sqrt{c}}.
\]
Figure: Simulated versus limiting $|\hat{u}_1^* u_1|^2$ for $Y_p = C_p^{1/2} X_p$, $C_p = I_p + \omega_1 u_1 u_1^*$, $p/n = 1/3$, varying $\omega_1$. 

**Figure:** Simulated versus limiting $|\hat{u}_1^* u_1|^2$ for $Y_p = C_p^{1/2} X_p$, $C_p = I_p + \omega_1 u_1 u_1^*$, $p/n = 1/3$, varying $\omega_1$. 

Spiked Models
Other Spiked Models

Similar results for multiple matrix models:

- $Y_p = \frac{1}{n} X_p X_p^* + P$
- $Y_p = \frac{1}{n} X_p^* (I + P) X$
- $Y_p = \frac{1}{n} (X_p + P)^* (X_p + P)$
- $Y_p = \frac{1}{n} T X_p^* (I + P) X_p T$
- etc.
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
**Context:** Two-step classification of \( n \) objects based on similarity \( A \in \mathbb{R}^{n \times n} \):

1. extraction of eigenvectors \( U = [u_1, \ldots, u_\ell] \) with “dominant” eigenvalues
Reminder on Spectral Clustering Methods

**Context:** Two-step classification of $n$ objects based on similarity $A \in \mathbb{R}^{n \times n}$:

1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues
2. classification of $n$ rows $U_1, \ldots, U_n \in \mathbb{R}^\ell$ using k-means/EM.
**Context:** Two-step classification of $n$ objects based on similarity $A \in \mathbb{R}^{n \times n}$:

1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues
2. classification of $n$ rows $U_1, \ldots, U_n \in \mathbb{R}^\ell$ using k-means/EM.
Reminder on Spectral Clustering Methods

**Context:** Two-step classification of $n$ objects based on similarity $A \in \mathbb{R}^{n \times n}$:

1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with “dominant” eigenvalues
2. classification of $n$ rows $U_1, \ldots, U_n \in \mathbb{R}^{\ell}$ using k-means/EM.

\[ \downarrow \textbf{Eigenvectors} \downarrow \]

(in practice, shuffled)
Reminder on Spectral Clustering Methods

- **Eigenv. 1**
- **Eigenv. 2**

The plots show the spectral clustering results with shuffling no longer matters.
Reminder on Spectral Clustering Methods

\[
\begin{array}{c}
\text{Eigenv. 1} \\
\Downarrow \\
\text{Eigenv. 2}
\end{array}
\]

\[
\Downarrow \ell\text{-dimensional representation} \Downarrow
\]

(Shuffling no longer matters)

\[
\begin{array}{c}
\text{Eigenvector 1} \\
\text{Eigenvector 2}
\end{array}
\]
Reminder on Spectral Clustering Methods

\[ \text{Eigenv. } 1 \]
\[ \text{Eigenv. } 2 \]

\[ \Downarrow \ell\text{-dimensional representation} \Downarrow \]
\(\text{(shuffling no longer matters)}\)

\[ \text{Eigenvector } 1 \]
\[ \text{Eigenvector } 2 \]

\[ \Downarrow \]
EM or k-means clustering.
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: “cluster” data in $k$ similarity classes $C_1, \ldots, C_k$. 
Problem Statement

- Dataset \( x_1, \ldots, x_n \in \mathbb{R}^p \)
- Objective: "cluster" data in \( k \) similarity classes \( C_1, \ldots, C_k \).

Kernel spectral clustering based on kernel matrix

\[
K = \{\kappa(x_i, x_j)\}_{i,j=1}^n
\]
Kernel Spectral Clustering

**Problem Statement**

- Dataset \( x_1, \ldots, x_n \in \mathbb{R}^p \)
- Objective: “cluster” data in \( k \) similarity classes \( C_1, \ldots, C_k \).

- **Kernel spectral clustering** based on kernel matrix

\[
K = \{ \kappa(x_i, x_j) \}_{i,j=1}^n
\]

- Usually, \( \kappa(x, y) = f(x^T y) \) or \( \kappa(x, y) = f(\|x - y\|^2) \)
Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: “cluster” data in $k$ similarity classes $C_1, \ldots, C_k$.

- Kernel spectral clustering based on kernel matrix

\[ K = \{\kappa(x_i, x_j)\}_{i,j=1}^{n} \]

- Usually, $\kappa(x,y) = f(x^T y)$ or $\kappa(x,y) = f(\|x - y\|^2)$
- Refinements:
  - instead of $K$, use $D - K$, $I_n - D^{-1} K$, $I_n - D^{-\frac{1}{2}} KD^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.
Kernel Spectral Clustering

Problem Statement

- Dataset \( x_1, \ldots, x_n \in \mathbb{R}^p \)
- Objective: “cluster” data in \( k \) similarity classes \( C_1, \ldots, C_k \).

- Kernel spectral clustering based on kernel matrix

\[
K = \{\kappa(x_i, x_j)\}_{i,j=1}^n
\]

- Usually, \( \kappa(x, y) = f(x^T y) \) or \( \kappa(x, y) = f(\|x - y\|^2) \)
- Refinements:
  - instead of \( K \), use \( D - K, I_n - D^{-1} K, I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}} \), etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

\[
K = \begin{pmatrix}
\kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\
\ll 1 & \ll 1 & \ll 1 \\
\kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\
\ll 1 & \gg 1 & \ll 1 \\
\kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\
\ll 1 & \ll 1 & \gg 1
\end{pmatrix}
\]

- \( K \) essentially low rank with class structure in eigenvectors.
Kernel Spectral Clustering

Figure: Leading four eigenvectors of \( D^{-1/2}K D^{-1/2} \) for MNIST data.
Kernel Spectral Clustering

Figure: Leading four eigenvectors of $D^{-1/2}KD^{-1/2}$ for MNIST data.
Kernel Spectral Clustering

Figure: Leading four eigenvectors of $D^{-1/2}K D^{-1/2}$ for MNIST data.
Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data.
Model and Assumptions

Gaussian mixture model:

- $x_1, \ldots, x_n \in \mathbb{R}^p$,
- $k$ classes $C_1, \ldots, C_k$,
- $x_1, \ldots, x_{n_1} \in C_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in C_k$,
- $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$. 

Assumption (Convergence Rate)

As $n \to \infty$,

1. Data scaling: $p_n \to c_0 \in (0, \infty)$,
2. Class scaling: $n_a \to c_a \in (0, 1)$,
3. Mean scaling: with $\mu^{\circ} \equiv \sum_{k=1}^{a} n_a \mu_a$ and $\mu^{\circ}_a \equiv \mu_a - \mu^{\circ}$, then $\|\mu^{\circ}_a\| = O(1)$,
4. Covariance scaling: with $C^{\circ} \equiv \sum_{k=1}^{a} n_a C_a$ and $C^{\circ}_a \equiv C_a - C^{\circ}$, then $\|C_a\| = O(1)$, $\text{tr} C^{\circ}_a = O(\sqrt{p})$, $\text{tr} [C_1 - C_2]^2 = O(p)$.

Remark: For 2 classes, this is $\|\mu_1 - \mu_2\| = O(1)$, $\text{tr} (C_1 - C_2) = O(\sqrt{p})$, $\|C_i\| = O(1)$, $\text{tr} (C_1 - C_2)^2 = O(p)$.
Gaussian mixture model:

- $x_1, \ldots, x_n \in \mathbb{R}^p$,
- $k$ classes $C_1, \ldots, C_k$,
- $x_1, \ldots, x_{n_1} \in C_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in C_k$,
- $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$.

Assumption (Convergence Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$,
2. **Class scaling**: $\frac{n_a}{n} \to c_a \in (0, 1)$,
3. **Mean scaling**: with $\mu^\circ \triangleq \sum_{a=1}^{k} \frac{n_a}{n} \mu_a$ and $\mu_a^\circ \triangleq \mu_a - \mu^\circ$, then
   \[ \|\mu_a^\circ\| = O(1) \]
4. **Covariance scaling**: with $C^\circ \triangleq \sum_{a=1}^{k} \frac{n_a}{n} C_a$ and $C_a^\circ \triangleq C_a - C^\circ$, then
   \[ \|C_a\| = O(1), \quad tr C_a^\circ = O(\sqrt{p}), \quad tr C_a^\circ C_b^\circ = O(p) \]
Gaussian mixture model:

- $x_1, \ldots, x_n \in \mathbb{R}^p$,
- $k$ classes $C_1, \ldots, C_k$,
- $x_1, \ldots, x_{n_1} \in C_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in C_k$,
- $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$.

Assumption (Convergence Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$,
2. **Class scaling**: $\frac{n_a}{n} \to c_a \in (0, 1)$,
3. **Mean scaling**: with $\mu^0 \triangleq \sum_{a=1}^{k} \frac{n_a}{n} \mu_a$ and $\mu_a^o \triangleq \mu_a - \mu^0$, then
   \[ \|\mu_a^0\| = O(1) \]
4. **Covariance scaling**: with $C^0 \triangleq \sum_{a=1}^{k} \frac{n_a}{n} C_a$ and $C_a^o \triangleq C_a - C^0$, then
   \[ \|C_a\| = O(1), \quad tr C_a^o = O(\sqrt{p}), \quad tr C_a^o C_b^o = O(p) \]

Remark: For 2 classes, this is

\[ \|\mu_1 - \mu_2\| = O(1), \quad tr (C_1 - C_2) = O(\sqrt{p}), \quad \|C_i\| = O(1), \quad tr ([C_1 - C_2]^2) = O(p). \]
Kernel Matrix:

- Kernel matrix of interest:

\[ K = \left\{ f \left( \frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^n \]

for some sufficiently smooth nonnegative \( f \) \((f(\frac{1}{p} x_i^T x_j) \) simpler).
Kernel Matrix:

- Kernel matrix of interest:

\[
K = \left\{ f \left( \frac{1}{p} \| x_i - x_j \|^2 \right) \right\}_{i,j=1}^n
\]

for some sufficiently smooth nonnegative \( f \) (\( f \left( \frac{1}{p} x_i^T x_j \right) \) simpler).

- We study the normalized Laplacian:

\[
L = n D^{-\frac{1}{2}} \left( K - \frac{d d^T}{d^T 1_n} \right) D^{-\frac{1}{2}}
\]

with \( d = K 1_n, D = \text{diag}(d) \).
Key Remark: Under our assumptions, uniformly on \( i, j \in \{1, \ldots, n\} \),

\[
\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{a.s.} \tau > 0.
\]
**Key Remark:** Under our assumptions, uniformly on $i, j \in \{1, \ldots, n\}$,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{a.s.} \tau > 0.$$ 

**Allows for Taylor expansion of $K$:**

$$K = f(\tau)1_n 1_n^T + \sqrt{n}K_1 + K_2$$

- $O_{\|\cdot\|}(n)$: low rank
- $O_{\|\cdot\|}(\sqrt{n})$: informative terms
- $O_{\|\cdot\|}(1)$
Random Matrix Equivalent

- **Key Remark:** Under our assumptions, uniformly on $i, j \in \{1, \ldots, n\}$,

$$\frac{1}{p} \| x_i - x_j \|^2 \xrightarrow{\text{a.s.}} \tau > 0.$$  

- Allows for Taylor expansion of $K$:

$$K = f(\tau)1_n1_n^T + \sqrt{n}K_1 + K_2$$

  - $O_{\| \cdot \| (n)}$ low rank,
  - $O_{\| \cdot \| (\sqrt{n})}$ informative terms,
  - $O_{\| \cdot \| (1)}$

However not the (small dimension) intuitive behavior.
Theorem (Random Matrix Equivalent [Couillet, Benaych’2015])

As $n, p \to \infty$, $\|L - \hat{L}\| \xrightarrow{a.s.} 0$, o

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} \Pi W^T W \Pi + \frac{1}{p} JBJ^T + * \right]$$

et $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n} (x_i = \mu_a + w_i)$, $\Pi = I_n - \frac{1}{n} 1_n 1_n^T$, 
Random Matrix Equivalent

Theorem (Random Matrix Equivalent [Couillet, Benaych’2015])

As $n, p \to \infty$, $\|L - \hat{L}\| \xrightarrow{a.s.} 0$, we have

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2\frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} \Pi W^T W \Pi + \frac{1}{p} JB J^T + * \right]$$

et $W = [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n}$ $(x_i = \mu_a + w_i)$, $\Pi = I_n - \frac{1}{n} 1_n 1_n^T$,

$$J = [j_1, \ldots, j_k], \quad j_a = (0, 0, \ldots, 1_{n_a}, 0, \ldots, 0)$$

$$B = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) tt^T - \frac{f''(\tau)}{f'(\tau)} T + *.$$

Recall $M = [\mu_1^0, \ldots, \mu_k^0]$, $t = \left[ \frac{1}{\sqrt{p}} tr C_1^0, \ldots, \frac{1}{\sqrt{p}} tr C_k^0 \right]$, $T = \left\{ \frac{1}{p} tr C^0_a C^0_b \right\}_{a,b=1}^k$. 
Isolated eigenvalues: Gaussian inputs

Figure: Eigenvalues of $L$ and $\hat{L}$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\delta_{aj}$, $C_a = (1 + 2(a - 1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$. 
Theoretical Findings versus MNIST

Figure: Eigenvalues of $L$ (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$. 
Theoretical Findings versus MNIST

Figure: Eigenvalues of $L$ (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$. 
Theoretical Findings versus MNIST

Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}K D^{-\frac{1}{2}}$ for MNIST data (red) and theoretical findings (blue).
Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} KD^{-\frac{1}{2}}$ for MNIST data (red) and theoretical findings (blue).
Theoretical Findings versus MNIST

Figure: 2D representation of eigenvectors of $L$, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in blue. Class 1 in red, Class 2 in black, Class 3 in green.
The surprising $f'(\tau) = 0$ case

Figure: Classification performance, polynomial kernel with $f(\tau) = 4, f''(\tau) = 2, x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p, [C_2]_{i,j} = 4^{|i-j|}, c_0 = \frac{1}{4}$. 
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Problem Statement

Context: Similar to clustering:
- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k$ classes, with $n_l$ labelled and $n_u$ unlabelled data.
Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in $k$ classes, with $n_l$ labelled and $n_u$ unlabelled data.
- Problem statement: give scores $F_{ia}$ ($d_i = [K1n]_i$)

$$F = \arg\min_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that $F_{ia} = \delta\{x_i \in C_a\}$, for all labelled $x_i$. 
Problem Statement

**Context:** Similar to clustering:

- Classify \( x_1, \ldots, x_n \in \mathbb{R}^p \) in \( k \) classes, with \( n_l \) labelled and \( n_u \) unlabelled data.
- Problem statement: give scores \( F_{ia} \) \( (d_i = [K1n]_i) \)

\[
F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2
\]

such that \( F_{ia} = \delta \{ x_i \in C_a \} \), for all labelled \( x_i \).

- **Solution:** for \( F^{(u)} \in \mathbb{R}^{n_u \times k} \), \( F^{(l)} \in \mathbb{R}^{n_l \times k} \) scores of unlabelled/labelled data,

\[
F^{(u)} = \left( I_{n_u} - D^{(u)} K_{(u,u)} D^{(u)\alpha-1} \right)^{-1} D^{(u)} K_{(u,l)} D^{(l)} F^{(l)}
\]

where we naturally decompose

\[
K = \begin{bmatrix}
K_{(l,l)} & K_{(l,u)} \\
K_{(u,l)} & K_{(u,u)}
\end{bmatrix}
\]

\[
D = \begin{bmatrix}
D_{(l)} & 0 \\
0 & D_{(u)}
\end{bmatrix} = \text{diag} \{ K1n \}.
\]
Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Vectors $[F^{(u)}]_{:,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Vectors $[F(u)]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
**MNIST Data Example**

**Figure:** Centered Vectors $[F_{(u)}^\alpha] \cdot a = [F(u) \cdot - \frac{1}{k} \sum_{k=1}^k F(u)_k] \cdot a$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Centered Vectors $[F_{(u)}^\circ]_{:,a} = [F_{(u)} - \frac{1}{k} F_{(u)} 1_k 1_k^T]_{:,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Centered Vectors $[F^O(u)]_{\cdot,a} = [F(u) - \frac{1}{k} F(u) 1_k 1_k^T]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Main Results

Results: Assuming \( n_l/n \to c_l \in (0,1) \), by previous Taylor expansion,

- In the first order,

\[
F_{:,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ v + \frac{t_a 1 n_u}{\sqrt{n}} \right] + O(n^{-1})
\]

Informative terms

where \( v = O(1) \) random vector (entry-wise) and \( t_a = \frac{1}{\sqrt{p}} \text{tr} C_a^o \).
Main Results

**Results:** Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F_{\cdot,a}(u) = C \frac{n_l,a}{n} \left[ v + \alpha \frac{t_a 1_n u}{\sqrt{n}} \right] + O(n^{-1})$$

**Informative terms**

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}} \text{tr} C_a^o$.

- Consequences:
Main Results

**Results:** Assuming $n_l/n \to c_l \in (0, 1)$, by previous Taylor expansion,

- In the first order,

$$ F(u) = C \frac{n_{l,a}}{n} \left[ v + \alpha \frac{t_a 1_n u}{\sqrt{n}} \right] + O(n^{-1}) $$

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}} \text{tr} C^\circ_a$.

- **Consequences:**
  - Random non-informative bias $v$
Main Results

**Results:** Assuming \( n_l/n \to c_l \in (0, 1) \), by previous Taylor expansion,

- In the first order,

\[
F^{(u)}_{\cdot,a} = C \frac{n_{l,a}}{n} \left[ v + \alpha \frac{t_{a \perp n_u}}{\sqrt{n}} \right] + O(n^{-1})
\]

where \( v = O(1) \) random vector (entry-wise) and \( t_a = \frac{1}{\sqrt{p}} \text{tr} \, C^o_a \).

- **Consequences:**
  - Random non-informative bias \( v \)
  - Strong Impact of \( n_{l,a} \)

\( F^{(u)}_{\cdot,a} \) to be scaled by \( n_{l,a} \)
Main Results

**Results:** Assuming $n_l/n \to c_l \in (0, 1)$, by previous Taylor expansion,

- In the first order,

\[
F^{(u)}_{*, a} = C \frac{n_{l,a}}{n} \left[ v \frac{t_a 1_n u}{\sqrt{n}} + \alpha \right] + O(n^{-1})
\]

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}} \text{tr} C_a^\circ$.

- Consequences:
  - Random non-informative bias $v$
  - Strong Impact of $n_{l,a}$
  - Additional per-class bias $\alpha t_a 1_n u$

\[
\alpha = 0 + \frac{\beta}{\sqrt{p}}.
\]
Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{i,a}^{(u)}.$$
Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{i,a}^{(u)}.$$

**Theorem**

*For* $x_i \in C_b$ *unlabelled*,

$$\hat{F}_{i,.} - G_b \to 0, \ G_b \sim \mathcal{N}(m_b, \Sigma_b)$$

where $m_b \in \mathbb{R}^k$, $\Sigma_b \in \mathbb{R}^{k \times k}$ given by

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} M_{ab} + \frac{2f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_{l,a}} \frac{f'(\tau)}{f(\tau)} t_a + B_B$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2tr C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^T C_b M]_{a_1 a_2} + \frac{\delta_{a_2}}{n_{l,a_1}} T_{ba_1} \right)$$

with $t, T, M$ as before, $\tilde{X}_a = X_a - \sum_{d=1}^{k} \frac{n_{l,d}}{n_l} X_d^o$ and $B_B$ bias independent of $a$. 
Main Results

Corollary (Asymptotic Classification Error)

For \( k = 2 \) classes and \( a \neq b \),

\[
P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in C_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1, -1] \Sigma_b [1, -1]^T}}\right) \to 0.
\]
Main Results

Corollary (Asymptotic Classification Error)

For $k = 2$ classes and $a \neq b$,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in C_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1, -1] \Sigma_b [1, -1]^T}}\right) \to 0.$$

Some consequences:

- non obvious choices of appropriate kernels
- non obvious choice of optimal $\beta$ (induces a possibly beneficial bias)
- importance of $n_l$ versus $n_u$. 

MNIST Data Example

Figure: Performance as a function of $\alpha$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Performance as a function of $\alpha$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_1/n = 1/16$, Gaussian kernel.
Figure: Performance as a function of $\alpha$, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
MNIST Data Example

Figure: Performance as a function of $\alpha$, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Outline

Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Semi-supervised Learning
  Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Random Feature Maps and Extreme Learning Machines

**Context:** Random Feature Map

- (large) input $x_1, \ldots, x_T \in \mathbb{R}^p$
- random $W = \begin{bmatrix} w_1^T \\ \vdots \\ w_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$
- non-linear activation function $\sigma$.

$$X = [x_1, \ldots, x_T]$$

$$\sigma(Wx_t)$$
Random Feature Maps and Extreme Learning Machines

**Context:** Random Feature Map

- (large) input $x_1, \ldots, x_T \in \mathbb{R}^p$
- random $W = \begin{bmatrix} w_1^T \\ \vdots \\ w_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$
- non-linear activation function $\sigma$.

**Neural Network Model (extreme learning machine):** Ridge-regression learning

- small output $y_1, \ldots, y_T \in \mathbb{R}^d$
- ridge-regression output $\beta \in \mathbb{R}^{n \times d}$

![Diagram of neural network model and random feature map](image-url)
Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$
Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

**Training MSE:**

$$E_{\text{train}} = \frac{1}{T} \sum_{i=1}^{T} \| y_i - \beta^T \sigma(W x_i) \|^2 = \frac{1}{T} \| Y - \beta^T \Sigma \|^2_F$$

with

$$\Sigma = \sigma(W X) = \left\{ \sigma(w_i^T x_j) \right\}_{1 \leq i \leq n, 1 \leq j \leq T}$$

$$\beta = \frac{1}{T} \Sigma \left( \frac{1}{T} \Sigma^T \Sigma + \gamma I_T \right)^{-1} Y.$$
Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- **Training MSE:**

$$E_{\text{train}} = \frac{1}{T} \sum_{i=1}^{T} \| y_i - \beta^T \sigma(W x_i) \|^2 = \frac{1}{T} \| Y - \beta^T \Sigma \|^2_F$$

with

$$\Sigma = \sigma(W X) = \left\{ \sigma(w_i^T x_j) \right\}_{1 \leq i \leq n, 1 \leq j \leq T}$$

$$\beta = \frac{1}{T} \Sigma \left( \frac{1}{T} \Sigma^T \Sigma + \gamma I_T \right)^{-1} Y.$$  

- **Testing MSE:** upon new pair $(\hat{X}, \hat{Y})$ of length $\hat{T}$,

$$E_{\text{test}} = \frac{1}{\hat{T}} \| \hat{Y} - \beta^T \hat{\Sigma} \|^2_F.$$

where $\hat{\Sigma} = \sigma(W \hat{X})$. 

---
Technical Aspects

Preliminary observations:

- Link to resolvent of $\frac{1}{T} \Sigma^T \Sigma$:

$$E_{\text{train}} = \frac{\gamma^2}{T} \text{tr} \ Y^T Y Q^2 = -\gamma^2 \frac{\partial}{\partial \gamma} \frac{1}{T} \text{tr} \ Y^T Y Q$$

where $Q = Q(\gamma)$ is the resolvent

$$Q \equiv \left( \frac{1}{T} \Sigma^T \Sigma + \gamma I_T \right)^{-1}$$

with $\Sigma_{ij} = \sigma(w_i^T x_j)$. 
Technical Aspects

Preliminary observations:

- Link to resolvent of $\frac{1}{T} \Sigma^T \Sigma$:

$$E_{\text{train}} = \frac{\gamma^2}{T} \text{tr} Y^T Y Q^2 = -\gamma^2 \frac{\partial}{\partial \gamma} \frac{1}{T} \text{tr} Y^T Y Q$$

where $Q = Q(\gamma)$ is the resolvent

$$Q \equiv \left( \frac{1}{T} \Sigma^T \Sigma + \gamma I_T \right)^{-1}$$

with $\Sigma_{ij} = \sigma(w_i^T x_j)$.

Central object: resolvent $E[Q]$. 
Main Technical Result

Theorem [Asymptotic Equivalent for $E[Q]$]

For Lipschitz $\sigma$, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with $Z$ standard Gaussian, we have, for all $\varepsilon > 0$,

$$\|E[Q] - \bar{Q}\| < Cn^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left( \frac{n}{T} \Phi \frac{1}{1 + \delta} + \gamma I_T \right)^{-1}$$

$\Phi \equiv E \left[ \sigma(X^Tw)\sigma(w^TX) \right]$ with $w = f(z), z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T} \text{tr} \Phi \bar{Q}.$$
Main Technical Result

**Theorem [Asymptotic Equivalent for $E[Q]$]**

For Lipschitz $\sigma$, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with $Z$ standard Gaussian, we have, for all $\varepsilon > 0$,

$$\| E[Q] - \bar{Q} \| < C n^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left( \frac{n}{T} \frac{\Phi}{1 + \delta} + \gamma I_T \right)^{-1}$$

$$\Phi \equiv E \left[ \sigma(X^T w) \sigma(w^T X) \right]$$

with $w = f(z), \ z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T} \text{tr} \Phi \bar{Q}.$$

**Proof arguments:**

- $\sigma(W X)$ has independent rows but dependent columns
- breaks the “trace lemma” argument (i.e., $\frac{1}{p} w^T X A X^T w \simeq \frac{1}{p} \text{tr} X A X^T$)
Main Technical Result

**Theorem [Asymptotic Equivalent for $E[Q]$]**

For Lipschitz $\sigma$, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with $Z$ standard Gaussian, we have, for all $\varepsilon > 0$,

$$\|E[Q] - \bar{Q}\| < Cn^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left( \frac{n}{T} \frac{\Phi}{1 + \delta} + \gamma I_T \right)^{-1}$$

$$\Phi \equiv E \left[ \sigma(X^T w) \sigma(w^T X) \right]$$

with $w = f(z), z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T} \text{tr} \Phi \bar{Q}.$$

**Proof arguments:**

- $\sigma(WX)$ has independent rows but dependent columns
- breaks the “trace lemma” argument (i.e., $\frac{1}{p} w^T X A X^T w \simeq \frac{1}{p} \text{tr} X A X^T$)

Concentration of measure lemma: $\frac{1}{p} \sigma(w^T X) A \sigma(X^T w) \simeq \frac{1}{p} \text{tr} \Phi A$
Main Technical Result

- Values of $\Phi(a, b)$ for $w \sim \mathcal{N}(0, I_p)$,

<table>
<thead>
<tr>
<th>$\sigma(t)$</th>
<th>$\Phi(a, b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>max($t, 0$)</td>
<td>$\frac{1}{2\pi}</td>
</tr>
<tr>
<td>$</td>
<td>t</td>
</tr>
<tr>
<td>erf($t$)</td>
<td>$\frac{2}{\pi} \arcsin \left( \frac{2a^T b}{\sqrt{(1 + 2</td>
</tr>
<tr>
<td>$1{t&gt;0}$</td>
<td>$\frac{1}{2} - \frac{1}{2\pi} \arccos(\angle(a, b))$</td>
</tr>
<tr>
<td>sign($t$)</td>
<td>$1 - \frac{2}{\pi} \arccos(\angle(a, b))$</td>
</tr>
<tr>
<td>cos($t$)</td>
<td>$\exp\left(-\frac{1}{2}(</td>
</tr>
</tbody>
</table>

where $\angle(a, b) \equiv \frac{a^T b}{||a|| ||b||}$. 
Main Technical Result

Values of $\Phi(a, b)$ for $w \sim \mathcal{N}(0, I_p)$,

<table>
<thead>
<tr>
<th>$\sigma(t)$</th>
<th>$\Phi(a, b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\max(t, 0)$</td>
<td>$\frac{1}{2\pi}</td>
</tr>
<tr>
<td>$</td>
<td>t</td>
</tr>
<tr>
<td>$\text{erf}(t)$</td>
<td>$\frac{2}{\pi} \sin \left( \frac{2a^Tb}{\sqrt{(1+2</td>
</tr>
<tr>
<td>$1{t&gt;0}$</td>
<td>$\frac{1}{2} - \frac{1}{2\pi} \cos(\angle(a, b))$</td>
</tr>
<tr>
<td>$\text{sign}(t)$</td>
<td>$1 - \frac{\pi}{2} \cos(\angle(a, b))$</td>
</tr>
<tr>
<td>$\cos(t)$</td>
<td>$\exp(-\frac{1}{2}(</td>
</tr>
</tbody>
</table>

where $\angle(a, b) \equiv \frac{a^Tb}{||a||||b||}$.

Value of $\Phi(a, b)$ for $w_i$ i.i.d. with $E[w_i^k] = m_k$ ($m_1 = 0$), $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$

$$\Phi(a, b) = \zeta_2^2 \left[ m_2^2 \left( 2(a^Tb)^2 + ||a||^2 ||b||^2 \right) + (m_4 - 3m_2^2)(a^2)^T(b^2) \right] + \zeta_1^2 m_2 a^Tb$$

$$+ \zeta_2 \zeta_1 m_3 \left[ (a^2)^Tb + a^T(b^2) \right] + \zeta_2 \zeta_0 m_2 \left[ ||a||^2 + ||b||^2 \right] + \zeta_0^2$$

where $(a^2) \equiv [a_1^2, \ldots, a_p^2]^T$. 
Main Results

Theorem [Asymptotic $E_{\text{train}}$]

For all $\varepsilon > 0$,

$$n^{\frac{1}{2} - \varepsilon} \left( E_{\text{train}} - \bar{E}_{\text{train}} \right) \to 0$$

almost surely, where

$$E_{\text{train}} = \frac{1}{T} \left\| Y^T - \Sigma^T \beta \right\|_F^2 = \frac{\gamma^2}{T} \text{tr} Y^T Y Q^2$$

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T} \text{tr} Y^T Y \bar{Q} \left[ \frac{1}{n} \text{tr} \Psi \bar{Q}^2 \Psi + I_T \right] \bar{Q}$$

with $\Psi \equiv \frac{n}{T} \frac{\Phi}{1+\delta}$. 
Main Results

Letting $\hat{X} \in \mathbb{R}^{p \times \hat{T}}$, $\hat{Y} \in \mathbb{R}^{d \times \hat{T}}$ satisfy “similar properties” as $(X, Y)$,

Claim [Asymptotic $E_{\text{test}}$]

For all $\varepsilon > 0$,

$$n^{\frac{1}{2} - \varepsilon} (E_{\text{test}} - \bar{E}_{\text{test}}) \to 0$$

almost surely, where

$$E_{\text{test}} = \frac{1}{\hat{T}} \left\| \hat{Y}^T - \hat{\Sigma}^T \beta \right\|_F^2$$

$$\bar{E}_{\text{test}} = \frac{1}{\hat{T}} \left\| \hat{Y}^T - \Psi^T \hat{X} \hat{X} \bar{Q} Y^T \right\|_F^2$$

$$+ \frac{1}{n} \text{tr} Y^T Y \bar{Q} \Psi \bar{Q}$$

$$\frac{1}{1 - \frac{1}{n} \text{tr} (\Psi \bar{Q})} \left[ \frac{1}{\hat{T}} \text{tr} \Psi \hat{X} \hat{X} - \frac{1}{\hat{T}} \text{tr} (I_{\hat{T}} + \gamma \bar{Q}) (\Psi \hat{X} \hat{X} \Psi \hat{X} \hat{X} \bar{Q}) \right]$$

with $\Psi_{AB} = \frac{n}{\hat{T}} \frac{\Phi_{AB}}{1 + \delta}$, $\Phi_{AB} = E[\sigma(A^T w)\sigma(w^T B)]$. 
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

**Figure:** Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 

$\sigma(t) = t \quad \sigma(t) = |t| \quad \sigma(t) = \text{erf}(t)$
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 

$\sigma(t) = t$, $\sigma(t) = |t|$, $\sigma(t) = \text{erf}(t)$, $\sigma(t) = \max(t, 0)$
Simulations on MNIST: non Lipschitz $\sigma(\cdot)$

Figure: Neural network performance for $\sigma(\cdot)$ either discontinuous or non Lipschitz, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Simulations on “tuned” Gaussian mixture

**Gaussian mixture classification**

- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\text{tr } C_1 = \text{tr } C_2$
Gaussian mixture classification

- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\text{tr } C_1 = \text{tr } C_2$
- We can prove that, for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $E[W_{ij}^k] = m_k$, classification only possible if $m_4 \neq m_2^2$
Gaussian mixture classification

- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\text{tr} \ C_1 = \text{tr} \ C_2$
- We can prove that, for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $E[W_{ij}^k] = m_k$,

$\rightarrow$ Classification only possible if $m_4 \neq m_2^2$
Simulations on “tuned” Gaussian mixture

- Interpretation in eigenstructure of $\Phi$: no information carried in dominant eigenmodes if $m_4 = m_2^2$.

\[ W_{ij} \sim \mathcal{N}(0, 1) \quad W_{ij} \sim \text{Bern} \quad W_{ij} \sim \text{Stud} \]
Outline

Basics of Random Matrix Theory
Motivation: Large Sample Covariance Matrices
Spiked Models

Applications
Reminder on Spectral Clustering Methods
Kernel Spectral Clustering
Semi-supervised Learning
Random Feature Maps, Extreme Learning Machines, and Neural Networks

Perspectives
Summary of Results and Perspectives I

Random Neural Networks.

✓ Extreme learning machines (one-layer random NN)
✓ Linear echo-state networks (ESN)
⇠ Logistic regression and classification error in extreme learning machines (ELM)
⇠ Further random feature maps characterization
⇠ Generalized random NN (multiple layers, multiple activations)
⇠ Random convolutional networks for image processing
✓ Non-linear ESN

Deep Neural Networks (DNN).

⇠ Backpropagation in NN ($\sigma(WX)$ for random $X$, backprop. on $W$)
✓ Statistical physics-inspired approaches (spin-glass models, Hamiltonian-based models)
✓ Non-linear ESN

DNN performance of physics-realistic models (4th-order Hamiltonian, locality)
References.


Summary of Results and Perspectives III

Summary of Results and Perspectives I

**Kernel methods.**
- ✔ Spectral clustering
- ✔ Subspace spectral clustering \((f'(\tau) = 0)\)
- ✡ Spectral clustering with outer product kernel \(f(x^T y)\)
- ✔ Semi-supervised learning, kernel approaches.
- ✔ Least square support vector machines (LS-SVM).
- ✡ Support vector machines (SVM).
- ✁ Kernel matrices based on Kendall \(\tau\), Spearman \(\rho\).

**Applications.**
- ✔ Massive MIMO user subspace clustering (patent proposed)
- ✁ Kernel correlation matrices for biostats, heterogeneous datasets.
- ✡ Kernel PCA.
- ✁ Kendall \(\tau\) in biostats.

**References.**


Community detection.

✓ Heterogeneous dense network clustering.
☞ Semi-supervised clustering.
💡 Sparse network extensions.
💡 Beyond community detection (hub detection).

Applications.

✓ Improved methods for community detection.
☞ Applications to distributed optimization (network diffusion, graph signal processing).

References.


Summary of Results and Perspectives I

Robust statistics.

✓ Tyler, Maronna (and regularized) estimators
✓ Elliptical data setting, deterministic outlier setting
✓ Central limit theorem extensions
💡 Joint mean and covariance robust estimation
💡 Robust regression (preliminary works exist already using strikingly different approaches)

Applications.

✓ Statistical finance (portfolio estimation)
✓ Localisation in array processing (robust GMUSIC)
✓ Detectors in space time array processing
💡 Correlation matrices in biostatistics, human science datasets, etc.

References.


Summary of Results and Perspectives I

Other works and ideas.

- ✔ Spike random matrix sparse PCA
- ✱ Non-linear shrinkage methods
- ✱ Sparse kernel PCA
- ✱ Random signal processing on graph methods.

Applications.

- ✔ Spike factor models in portfolio optimization
- ✱ Non-linear shrinkage in portfolio optimization, biostats

References.

The End

Thank you.