Random Matrices in Machine Learning

Romain COUILLET

CentraleSupélec, University of Paris Saclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, University Grenoble–Alpes, France.

June 21, 2018
Basics of Random Matrix Theory
  Motivation: Large Sample Covariance Matrices
  Spiked Models

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case \( f'(\tau) = 0 \)
  Kernel Spectral Clustering: The case \( f'(\tau) = \frac{\alpha}{\sqrt{p}} \)
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Outline

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

Applications
- Reminder on Spectral Clustering Methods
- Kernel Spectral Clustering
- Kernel Spectral Clustering: The case $f'(\tau) = 0$
- Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
- Semi-supervised Learning
- Semi-supervised Learning improved
- Random Feature Maps, Extreme Learning Machines, and Neural Networks
- Community Detection on Graphs

Perspectives
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
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)

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

($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).
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 $E[y_1] = 0$, $E[y_1y_1^*] = C_p$:

- 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

$$
\left\| \hat{C}_p - C_p \right\| \overset{a.s.}{\to} 0.
$$

**Random Matrix Regime**

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

$$
\left\| \hat{C}_p - C_p \right\| \not\to 0.
$$
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

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

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 \( p \sim n \), 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 (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

$\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$
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
- Kernel Spectral Clustering: The case $f'(\tau) = 0$
- Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
- Semi-supervised Learning
- Semi-supervised Learning improved
- Random Feature Maps, Extreme Learning Machines, and Neural Networks
- Community Detection on Graphs

Perspectives
Spiked Models

Small rank perturbation: $C_p = I_p + P$, $P$ of low rank.

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 $Y_p = C_p^{\frac{1}{2}} X_p$, with

- $X_p$ with i.i.d. zero mean, unit variance, $E[|X_p|_{ij}^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^*$, 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 $Y_p = C_p^{\frac{1}{2}} X_p$, 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,

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

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

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

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

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

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

- $X_p$ with i.i.d. zero mean, unit variance, $E[|X_p|_4^4] < \infty$.

- $C_p = I_p + P$, $P = U\Omega U^* = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$.

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 $\lambda_i(\frac{1}{n} Y_p Y_p^*)$,

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

In particular,

$$|\hat{u}_i^* u_i|^2 \xrightarrow{a.s.} \frac{1 - c \omega_i^{-2}}{1 + c \omega_i^{-1}} \cdot 1_{\omega_i > \sqrt{c}}.$$
Spiked Models

Figure: Simulated versus limiting $|\hat{u}_1^*u_1|^2$ for $Y_p = C_p^{\frac{1}{2}}X_p$, $C_p = I_p + \omega_1u_1u_1^*$, $p/n = 1/3$, varying $\omega_1$. 
Other Spiked Models

Similar results for multiple matrix models:

\[ Y_p = \frac{1}{n} (I + P)^{\frac{1}{2}} X_p X_p^* (I + P)^{\frac{1}{2}} \]
\[ Y_p = \frac{1}{n} X_p X_p^* + P \]
\[ Y_p = \frac{1}{n} X_p^* (I + P) X \]
\[ Y_p = \frac{1}{n} (X_p + P)^* (X_p + P) \]
\[ \text{etc.} \]
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Outline

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

Applications
   Reminder on Spectral Clustering Methods
   Kernel Spectral Clustering
   Kernel Spectral Clustering: The case $f'(\tau) = 0$
   Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
   Semi-supervised Learning
   Semi-supervised Learning improved
   Random Feature Maps, Extreme Learning Machines, and Neural Networks
   Community Detection on Graphs

Perspectives
Reminder on Spectral Clustering Methods

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

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

\[ \Downarrow \text{Eigenvectors} \Downarrow \]

(in practice, shuffled)
Reminder on Spectral Clustering Methods

1. Eigenvector 1
2. Eigenvector 2
Reminder on Spectral Clustering Methods

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

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

(shuffling no longer matters)

\[ \quad \text{Eigenvector 1} \quad \text{Eigenvector 2} \]
Reminder on Spectral Clustering Methods

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

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

(shuffling no longer matters)

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

\[ \Downarrow \quad \text{EM or k-means clustering.} \]
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
    Kernel Spectral Clustering: The case $f'(\tau) = 0$
    Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Kernel Spectral Clustering

**Problem Statement**

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

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 $D^{-1/2}KD^{-1/2}$, $I_n - D^{-1}$, $I_n - D^{-1}K$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

- $K$ essentially low rank with class structure in eigenvectors.
- Ng–Weiss–Jordan key remark:
  $D^{-1/2}K D^{-1/2} (D^{-1})^a \approx D^{-1/2} (D^{-1})^a$ (canonical vector of $C_a$).
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

\[
K = \{\kappa(x_i, x_j)\}_{i,j=1}^n
\]
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 $D^{-1/2}K D^{-1/2}$, $I_n - D^{-1}K D^{-1}$, etc.
- several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

- $K$ essentially low rank with class structure in eigenvectors.
- Ng–Weiss–Jordan key remark:

\[
D^{-1/2}K D^{-1/2}(D_j a) \approx D_j a (j a \text{ canonical vector of } C_a)
\]
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.
**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.

**Intuition (from small dimensions)**

- $K$ essentially low rank with class structure in eigenvectors.
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)

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

- $K$ essentially low rank with class structure in eigenvectors.
- Ng–Weiss–Jordan key remark: $D^{-\frac{1}{2}} K D^{-\frac{1}{2}} (D^{\frac{1}{2}} j_a) \simeq D^{\frac{1}{2}} j_a$ ($j_a$ canonical vector of $C_a$)
Figure: Leading four eigenvectors of $D^{-1/2} K D^{-1/2}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

▶ Important Remark: Eigenvectors informative but far from $D^{1/2}$. 

-0.06
-0.07
-0.08
Kernel Spectral Clustering

Figure: Leading four eigenvectors of $D^{-1/2}K D^{-1/2}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

- Important Remark: eigenvectors informative but far from $D^{1/2}$. 

0 
1 
2
Important Remark: eigenvectors informative but far from $D^{1/2}j^a$!
Kernel Spectral Clustering

Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).
Kernel Spectral Clustering

Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data, RBF kernel $(f(t) = \exp(-t^2/2))$.

▶ **Important Remark:** eigenvectors informative **BUT** far from $D^{\frac{1}{2}} j_a$!
Model and Assumptions

Gaussian mixture model:

- \( x_1, \ldots, x_n \in \mathbb{R}^p \),
- \( k \) classes \( C_1, \ldots, C_k \),
- \( x_1, \ldots, x_{n_1} \in C_1, \ldots, x_{n-n_k+1}, \ldots, x_n \in C_k \),
- \( x_i \sim \mathcal{N}(\mu_i, C_i) \).
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 (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,

2. **Mean scaling**: with $\mu^\diamond \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ and $\mu_a^\diamond \triangleq \mu_a - \mu^\diamond$, then $\|\mu_a^\diamond\| = O(1)$

3. **Covariance scaling**: with $C^\diamond \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$ and $C_a^\diamond \triangleq C_a - C^\diamond$, then

$$\|C_a\| = O(1), \quad tr C_a^\diamond = O(\sqrt{p}), \quad tr C_a^\diamond C_b^\diamond = O(p)$$
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 (Growth Rate)

As \( n \to \infty \),

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

\[
\| C_a \| = O(1), \quad tr C^o_a = O(\sqrt{p}), \quad tr C^o_a C^o_b = O(p)
\]

For 2 classes, this is

\[
\| \mu_1 - \mu_2 \| = O(1), \quad tr (C_1 - C_2) = O(\sqrt{p}), \quad \| C_i \| = O(1), \quad tr ([C_1 - C_2]^2) = O(p).
\]
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 (Growth Rate)
As \( n \to \infty \),
1. Data scaling: \( \frac{p}{n} \to c_0 \in (0, \infty), \frac{n_a}{n} \to c_a \in (0, 1) \),
2. Mean scaling: with \( \mu^o \triangleq \sum_{a=1}^{k} \frac{n_a}{n} \mu_a \) and \( \mu_a^o \triangleq \mu_a - \mu^o \), then \( \|\mu_a^o\| = O(1) \)
3. Covariance scaling: with \( C^o \triangleq \sum_{a=1}^{k} \frac{n_a}{n} C_a \) and \( C_a^o \triangleq C_a - C^o \), then
   \[ \|C_a\| = O(1), \quad trC_a^o = O(\sqrt{p}), \quad trC_a^o C_b^o = O(p) \]

For 2 classes, this is
\[ \|\mu_1 - \mu_2\| = O(1), \quad tr(C_1 - C_2) = O(\sqrt{p}), \quad \|C_i\| = O(1), \quad tr([C_1 - C_2]^2) = O(p). \]

Remark: [Neyman–Pearson optimality]
- \( x \sim \mathcal{N}(\pm \mu, I_p) \) (known \( \mu \)) decidable if\( f \|\mu\| \geq O(1) \).
- \( x \sim \mathcal{N}(0, (1 \pm \varepsilon) I_p) \) (known \( \varepsilon \)) decidable if \( \|\varepsilon\| \geq O(p^{-\frac{1}{2}}) \).
Kernel Matrix:

- Kernel matrix of interest:

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

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

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).

- 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 \( d = K 1_n \), \( D = \text{diag}(d) \).

*(more stable both theoretically and in practice)*
Key Remark: Under growth rate assumptions,

\[
\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \| x_i - x_j \| - \tau \right| \right\} \xrightarrow{a.s.} 0.
\]

where \( \tau = \frac{1}{p} \text{tr} C^0 \).
**Key Remark:** Under growth rate assumptions,

\[
\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \| x_i - x_j \|_2^2 - \tau \right| \right\} \xrightarrow{\text{a.s.}} 0.
\]

where \( \tau = \frac{1}{p} \text{tr} C^\circ. \)

\[\Rightarrow\] Suggests that (up to diagonal) \( K \simeq f(\tau)1_n 1_n^T! \)
Random Matrix Equivalent

**Key Remark:** Under growth rate assumptions,

\[
\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \| x_i - x_j \|_2^2 - \tau \right\} \xrightarrow{\text{a.s.}} 0.
\]

where \( \tau = \frac{1}{p} \text{tr } C^\circ \).

\( \Rightarrow \) Suggests that (up to diagonal) \( K \simeq f(\tau) 1_n 1_n^T \!\). 

**In fact, information hidden in low order fluctuations!** from “matrix-wise” Taylor expansion of \( K \):

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

\( O_{\| \cdot \|}(n) \) \quad \text{low rank, } \( O_{\| \cdot \|}(\sqrt{n}) \) \quad \text{informative terms, } \( O_{\| \cdot \|}(1) \)
Key Remark: Under growth rate assumptions,

\[
\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \| x_i - x_j \|_2^2 - \tau \right| \right\} \xrightarrow{a.s.} 0.
\]

where \( \tau = \frac{1}{p} \text{tr} C^\circ \).

\[ \Rightarrow \text{Suggests that (up to diagonal) } K \simeq f(\tau)1_n 1_n^T ! \]

In fact, information hidden in low order fluctuations! from “matrix-wise” Taylor expansion of \( K \):

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

\[ O\|\cdot\|_2(n) \quad \text{low rank, } O\|\cdot\|_2(\sqrt{n}) \quad \text{informative terms, } O\|\cdot\|_2(1) \]

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

As \( n, p \to \infty \), \( \| L - \hat{L} \| \xrightarrow{a.s.} 0 \), where

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

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

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

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

As \( n, p \to \infty, \|L - \hat{L}\| \xrightarrow{a.s.} 0, \) where

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

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

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

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

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

Recall \( M = [\mu_1^\circ, \ldots, \mu_k^\circ], \) \( t = \left[ \frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \ldots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ \right]^T, \) \( T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a, b = 1}^k. \)
Random Matrix Equivalent

Theorem (Random Matrix Equivalent [Couillet, Benaych’2015])
As \( n, p \to \infty \), \( \| L - \hat{L} \| \xrightarrow{a.s.} 0 \), where

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

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

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

\[
J = [j_1, \ldots, j_k], \quad j_a^T = (0 \ldots 0, 1_{na}, 0, \ldots, 0)
\]

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

Recall \( M = [\mu_1^\circ, \ldots, \mu_k^\circ] \), \( t = \left[ \frac{1}{\sqrt{p}} tr C_1^\circ, \ldots, \frac{1}{\sqrt{p}} tr C_k^\circ \right]^T \), \( T = \left\{ \frac{1}{p} tr C_a^\circ C_b^\circ \right\}_{a,b=1}^k \).

Fundamental conclusions:
Random Matrix Equivalent

Theorem (Random Matrix Equivalent [Couillet, Benaych’2015])
As \( n, p \to \infty \), \( \|L - \hat{L}\| \xrightarrow{a.s.} 0 \), where

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

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

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

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

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

Recall \( M = [\mu_1^\circ, \ldots, \mu_k^\circ], \)
\( t = \left[ \frac{1}{\sqrt{p}} tr C_1^\circ, \ldots, \frac{1}{\sqrt{p}} tr C_k^\circ \right]^T, \)
\( T = \left\{ \frac{1}{p} tr C_a^\circ C_b^\circ \right\}_{a,b=1}^k \).

Fundamental conclusions:

- asymptotic kernel impact only through \( f'(\tau) \) and \( f'''(\tau) \), that’s all!
Random Matrix Equivalent

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

As \( n, p \to \infty \), \( \|L - \hat{L}\| \xrightarrow{a.s.} 0 \), where

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

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

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

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

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

Recall \( M = [\mu_1^\circ, \ldots, \mu_k^\circ] \), \( t = \left[ \frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \ldots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ \right]^T \), \( T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k \).

**Fundamental conclusions:**

- asymptotic **kernel impact** only through \( f'(\tau) \) and \( f''(\tau) \), that’s all!
- spectral clustering reads \( M^T M, tt^T \) and \( T \), that’s all!
Isolated eigenvalues: Gaussian inputs

Figure: Eigenvalues of $L$ and $\hat{L}$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\delta_{aj}$, $C_a = (1 + 2(a - 1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$. 
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, $p = 784$, $n = 192$. 
Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data (red) and theoretical findings (blue).
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).
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.
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:** Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_\alpha)$, with $C_1 = I_p$, $[C_2]_{i,j} = 0.4|i-j|$, $c_0 = \frac{1}{4}$. 
The surprising $f'(\tau) = 0$ case

Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_\alpha)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$. 
The surprising $f'(\tau) = 0$ case

Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_\alpha)$, with $C_1 = I_p$, $[C_2]_{i,j} = 4|i-j|$, $c_0 = \frac{1}{4}$. 
The surprising $f'(\tau) = 0$ case

Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = 0.4|i-j|$, $c_0 = \frac{1}{4}$.

▶ **Trivial classification** when $t = 0$, $M = 0$ and $\|T\| = O(1)$. 
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  **Kernel Spectral Clustering: The case** $f'_{\tau} = 0$
  Kernel Spectral Clustering: The case $f'_{\tau} = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on “spanned subspaces”.

Position of the Problem
Position of the Problem

**Problem:** Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on “spanned subspaces”.

**Method:**

- Still assume $x_1, \ldots, x_n$ belong to $k$ classes $C_1, \ldots, C_k$.
- Zero-mean Gaussian model for the data: for $x_i \in C_k$,

$$x_i \sim \mathcal{N}(0, C_k).$$
Position of the Problem

**Problem:** Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on “spanned subspaces”.

**Method:**

- Still assume $x_1, \ldots, x_n$ belong to $k$ classes $C_1, \ldots, C_k$.
- Zero-mean Gaussian model for the data: for $x_i \in C_k$,

$$x_i \sim \mathcal{N}(0, C_k).$$

- Performance of $L = nD^{-\frac{1}{2}} \left(K - \frac{1}{n} \frac{1^T}{1n}D1_n\right)D^{-\frac{1}{2}}$, with

$$K = \left\{ f \left(\|\bar{x}_i - \bar{x}_j\|^2\right)\right\}_{1 \leq i, j \leq n}, \quad \bar{x} = \frac{x}{\|x\|}$$

in the regime $n, p \to \infty$.

*(alternatively, we can ask $\frac{1}{p} \text{tr} C_i = 1$ for all $1 \leq i \leq k$)*
Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from $k$-class Gaussian mixture, with $x_i \in C_k \iff x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).
Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from $k$-class Gaussian mixture, with $x_i \in C_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Assumption 2a [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr } C_a = 1$ and $\text{tr } C_a^o C_b^o = O(p)$, with $C_a^o = C_a - C^o$, $C^o = \sum_{b=1}^{k} c_b C_b$. 

Theorem (Corollary of Previous Section)

Let $f$ smooth with $f' \neq 0$. Then, under Assumptions 2a, $L = nD - \frac{1}{2} \left( K - \frac{1}{n} T \right) n D^{- \frac{1}{2}}$, with $K = \{ f(\|\bar{x}_i - \bar{x}_j\|_2) \}_{i,j=1}^n (\bar{x} = x/\|x\|)$ exhibits phase transition phenomenon, i.e., leading eigenvectors of $L$ asymptotically contain structural information about $C_1, \ldots, C_k$ if and only if $T = \{ \frac{1}{p} \text{tr } C_a^o C_b^o \}_{a,b=1}^k$ has sufficiently large eigenvalues (here $M = 0$, $t = 0$).
Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from $k$-class Gaussian mixture, with $x_i \in C_k \iff x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

Assumption 2a [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,
1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr } C_a = 1$ and $\text{tr } C^o_a C^o_b = O(p)$, with $C^o_a = C_a - C^o$, $C^o = \sum_{b=1}^{k} c_b C_b$.

Theorem (Corollary of Previous Section)

Let $f$ smooth with $f'(2) \neq 0$. Then, under Assumptions 2a,

$$L = n D^{-\frac{1}{2}} \left( K - \frac{1_n 1_n^T}{1_n^T D 1_n} \right) D^{-\frac{1}{2}}, \text{ with } K = \left\{ f \left( \| \bar{x}_i - \bar{x}_j \| \right) \right\}_{i,j=1}^n \left( \bar{x} = x/\|x\| \right)$$

exhibits phase transition phenomenon.
Model and Reminders

**Assumption 1 [Classes].** Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ i.i.d. from $k$-class Gaussian mixture, with $x_i \in \mathcal{C}_k \iff x_i \sim \mathcal{N}(0, C_k)$ (sorted by class for simplicity).

**Assumption 2a [Growth Rates].** As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr} C_a = 1$ and $\text{tr} C^\circ_a C^\circ_b = O(p)$, with $C^\circ_a = C_a - C^\circ$, $C^\circ = \sum_{b=1}^k c_b C_b$.

**Theorem (Corollary of Previous Section)**

Let $f$ smooth with $f'(2) \neq 0$. Then, under Assumptions 2a,

$$L = n D^{-\frac{1}{2}} \left( K - \frac{1_n 1_n^T}{1_n^T D 1_n} \right) D^{-\frac{1}{2}}, \text{ with } K = \left\{ f \left( \|\bar{x}_i - \bar{x}_j\|^2 \right) \right\}_{i,j=1}^n (\bar{x} = x/\|x\|)$$

exhibits phase transition phenomenon, i.e., leading eigenvectors of $L$ asymptotically contain structural information about $\mathcal{C}_1, \ldots, \mathcal{C}_k$ if and only if

$$T = \left\{ \frac{1}{p} \text{tr} C^\circ_a C^\circ_b \right\}_{a,b=1}^k$$

has sufficiently large eigenvalues (here $M = 0, t = 0$).
The case $f'(2) = 0$

**Assumption 2b [Growth Rates].** As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr} C_a = 1$ and $\text{tr} C_a^\circ C_b^\circ = O(p)$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^{k} c_b C_b$. 

Remark: [Neyman–Pearson optimality]

if $C_i = I_p \pm E$ with $\|E\| \to 0$, detectability if $\frac{1}{p} \text{tr} (C_1 - C_2) \geq O(p)$.
The case $f'(2) = 0$

**Assumption 2b [Growth Rates].** As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr} C_a = 1$ and $\text{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^{k} c_b C_b$.

*(in this regime, previous kernels clearly fail)*

**Remark: [Neyman–Pearson optimality]**

- if $C_i = I_p \pm E$ with $\|E\| \to 0$, **detectability** *iff* $\frac{1}{p} \text{tr} (C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$. 
The case $f'(2) = 0$

Assumption 2b [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr} C_a = 1$ and $\text{tr} C_a C_b^\circ = O(\sqrt{p})$, with $C_a^\circ = C_a - C^\circ$, $C^\circ = \sum_{b=1}^{k} c_b C_b$.

(in this regime, previous kernels clearly fail)

Remark: [Neyman–Pearson optimality]

$\blacktriangleright$ if $C_i = I_p \pm E$ with $\|E\| \to 0$, detectability iif $\frac{1}{p} \text{tr} (C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$.

Theorem (Random Equivalent for $f'(2) = 0$)

Let $f$ be smooth with $f'(2) = 0$ and

$$
\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[ L - \frac{f(0) - f(2)}{f(2)} P \right], \quad P = I_n - \frac{1}{n} 1_n 1_n^T.
$$

Then, under Assumptions 2b,

$$
\mathcal{L} = P \Phi P + \left\{ \frac{1}{\sqrt{p}} \text{tr} (C_a^\circ C_b^\circ) \frac{1}{n_a} 1_{n_b}^T \right\}_{a, b=1}^{k} + o_{\|\cdot\|}(1)
$$

where $\Phi_{ij} = \delta_{i \neq j} \sqrt{p} \left[ (x_i^T x_j)^2 - E[(x_i^T x_j)^2] \right]$. 

The case $f'(2) = 0$

Figure: Eigenvalues of $L$, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$,
$C_i \propto I_p + (p/8)^{-5/4} W_i W_i^T$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0, 1)$ entries, $f(t) = \exp(-(t - 2)^2)$.

⇒ No longer a Marcenko–Pastur like bulk, but rather a semi-circle bulk!
The case $f'(2) = 0$
The case $f'(2) = 0$

**Roadmap.** We now need to:

- study the spectrum of $\Phi$

\[\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}(L)\]

Then, under Assumption 2b, $\mu_n \rightarrow \mu$ with $\mu$ the semi-circle distribution

\[\mu(dt) = \frac{1}{2\pi c_0} \sqrt{4c_0^2 - t^2} + dt, \quad \omega = \lim_{p \to \infty} \sqrt{\frac{1}{2p} \text{tr}(C \circ C)}^2.\]
The case $f'(2) = 0$

**Roadmap.** We now need to:

- study the spectrum of $\Phi$
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
The case \( f'(2) = 0 \)

**Roadmap.** We now need to:

- study the spectrum of \( \Phi \)
- study the isolated eigenvalues of \( \mathcal{L} \) (and the phase transition)
- retrieve information from the eigenvectors.
The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of $\Phi$
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
- retrieve information from the eigenvectors.

Theorem (Semi-circle law for $\Phi$)

Let $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\mathcal{L})}$. Then, under Assumption 2b,

$$\mu_n \xrightarrow{a.s.} \mu$$

with $\mu$ the semi-circle distribution

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2} \sqrt{(4c_0 \omega^2 - t^2)^+} dt, \quad \omega = \lim_{p \to \infty} \sqrt{2p} \frac{1}{p} tr(C^{\circ})^2.$$
The case $f'(2) = 0$

Figure: Eigenvalues of $L$, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$,

$C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^T$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0, 1)$ entries, $f(t) = \exp(-(t - 2)^2)$. 

$\lambda_1(L)$

$\lambda_2(L)$
The case $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p \to \infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k.$$
The case $f'(2) = 0$

Denote now

$$
\mathcal{T} \equiv \lim_{p \to \infty} \left\{ \frac{\sqrt{c_ac_b}}{\sqrt{p}} \text{tr} C_a^0 C_b^0 \right\}_{a,b=1}^k.
$$

Theorem (Isolated Eigenvalues)

Let $\nu_1 \geq \ldots \geq \nu_k$ eigenvalues of $\mathcal{T}$. Then, if $\sqrt{c_0}|\nu_i| > \omega$, $\mathcal{L}$ has an isolated eigenvalue $\lambda_i$ satisfying

$$
\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}.
$$
The case $f'(2) = 0$

**Theorem (Isolated Eigenvectors)**

For each isolated eigenpair $(\lambda_i, u_i)$ of $L$ corresponding to $(\nu_i, v_i)$ of $T$, write

$$u_i = \sum_{a=1}^{k} \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with $j_a = [0^T_{n_1}, \ldots, 1^T_{n_a}, \ldots, 0^T_{n_k}]^T$, $(w_i^a)^T j_a = 0$, $\text{supp}(w_i^a) = \text{supp}(j_a)$, $\|w_i^a\| = 1$.

Then, under Assumptions 1–2b,

$$\alpha_i^a \alpha_i^b \xrightarrow{a.s.} \left(1 - \frac{1}{c_0} \frac{\omega^2}{\nu_i^2}\right) [v_i v_i^T]_{ab}$$

$$(\sigma_i^a)^2 \xrightarrow{a.s.} \frac{c_a \omega^2}{c_0 \nu_i^2}$$

and the fluctuations of $u_i, u_j$, $i \neq j$, are asymptotically uncorrelated.
The case $f'(2) = 0$

Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of $L$) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$. 
The case $f'(2) = 0$

**Figure:** Leading two eigenvectors of $\mathcal{L}$ (or equivalently of $L$) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$. 
The case $f'(2) = 0$

Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of $L$) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$. 
Application to Massive MIMO UE Clustering
Massive MIMO UE Clustering

**Setting.** Massive MIMO cell with

- $p$ antenna elements
- $n$ users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE’s belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^T] = C_a \equiv C(\Theta_a)$.
Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- $p$ antenna elements
- $n$ users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE’s belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^T] = C_a \equiv C(\Theta_a)$. 
- $T$ independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE $i$. 

Algorithm.

1. Build kernel matrix $K$, then $L$, based on $nT$ vectors $x_{i}^{(1)}, \ldots, x_{i}^{(T)}$ as if $nT$ values to cluster.
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each $i$, create $\tilde{u}_i = 1^T (I_n \otimes 1^T) u_i$, i.e., average eigenvectors along time.
4. Perform $k$-class clustering on vectors $\tilde{u}_1, \ldots, \tilde{u}_\kappa$. 

Massive MIMO UE Clustering

**Setting.** Massive MIMO cell with

- \( p \) antenna elements
- \( n \) users equipments (UE) with channels \( x_1, \ldots, x_n \in \mathbb{R}^p \)
- UE’s belong to solid angle groups, i.e., \( E[x_i] = 0, E[x_i x_i^T] = C_a \equiv C(\Theta_a) \).
- \( T \) independent channel observations \( x_i^{(1)}, \ldots, x_i^{(T)} \) for UE \( i \).

**Objective.** Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination).*
Massive MIMO UE Clustering

**Setting.** Massive MIMO cell with
- $p$ antenna elements
- $n$ users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE’s belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^T] = C_a \equiv C(\Theta_a)$.
- $T$ independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE $i$.

**Objective.** Clustering users in same solid angle groups (*for scheduling reasons, to avoid pilot contamination*).

**Algorithm.**

1. Build kernel matrix $K$, then $L$, based on $nT$ vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if $nT$ values to cluster).
Massive MIMO UE Clustering

**Setting.** Massive MIMO cell with

- \( p \) antenna elements
- \( n \) users equipments (UE) with channels \( x_1, \ldots, x_n \in \mathbb{R}^p \)
- UE’s belong to solid angle groups, i.e., \( E[x_i] = 0, E[x_i x_i^T] = C_a \equiv C(\Theta_a) \)
- \( T \) independent channel observations \( x_i^{(1)}, \ldots, x_i^{(T)} \) for UE \( i \).

**Objective.** Clustering users in same solid angle groups (for scheduling reasons, to avoid pilot contamination).

**Algorithm.**

1. Build kernel matrix \( K \), then \( L \), based on \( nT \) vectors \( x_1^{(1)}, \ldots, x_n^{(T)} \) (as if \( nT \) values to cluster).
2. Extract dominant isolated eigenvectors \( u_1, \ldots, u_{\kappa} \)
Setting. Massive MIMO cell with

- $p$ antenna elements
- $n$ users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE’s belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^T] = C_a \equiv C(\Theta_a)$.
- $T$ independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE $i$.

Objective. Clustering users in same solid angle groups (for scheduling reasons, to avoid pilot contamination).

Algorithm.

1. Build kernel matrix $K$, then $L$, based on $nT$ vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if $nT$ values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each $i$, create $\bar{u}_i = \frac{1}{T} (I_n \otimes 1_T^T) u_i$, i.e., average eigenvectors along time.
Massive MIMO UE Clustering

**Setting.** Massive MIMO cell with

- $p$ antenna elements
- $n$ users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE’s belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^T] = C_a \equiv C(\Theta_a)$.
- $T$ independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE $i$.

**Objective.** Clustering users in same solid angle groups (*for scheduling reasons, to avoid pilot contamination*).

**Algorithm.**

1. Build kernel matrix $K$, then $L$, based on $nT$ vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if $nT$ values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each $i$, create $\tilde{u}_i = \frac{1}{T}(I_n \otimes 1_T)u_i$, i.e., average eigenvectors along time.
4. Perform $k$-class clustering on vectors $\tilde{u}_1, \ldots, \tilde{u}_\kappa$. 
Massive MIMO UE Clustering

Figure: Leading two eigenvectors before (left figure) and after (right figure) $T$-averaging. Setting: $p = 400$, $n = 40$, $T = 10$, $k = 3$, $c_1 = c_3 = 1/4$, $c_2 = 1/2$, angular spread model with angles $-\pi/30 \pm \pi/20$, $0 \pm \pi/20$, and $\pi/30 \pm \pi/20$. Kernel function $f(t) = \exp(-(t - 2)^2)$. 
Massive MIMO UE Clustering

**Figure:** Overlap for different $T$, using the k-means or EM starting from actual centroid solutions (oracle) or randomly.
Massive MIMO UE Clustering

Figure: Overlap for optimal kernel $f(t)$ (here $f(t) = \exp(-(t-2)^2)$) and Gaussian kernel $f(t) = \exp(-t^2)$, for different $T$, using the k-means or EM.
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel \( f(\frac{1}{p}\|x_i - x_j\|^2) \) with \( f'(\tau) \neq 0 \):
  - optimal in \( \|\mu^\circ_a\| = O(1), \frac{1}{p}\text{tr} C^\circ_a = O(p^{-\frac{1}{2}}) \)
  - suboptimal in \( \frac{1}{p}\text{tr} C^\circ_a C^\circ_b = O(1) \)

\[ \rightarrow \textbf{Model type: Marčenko–Pastur + spikes.} \]
Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel $f\left(\frac{1}{p} \|x_i - x_j\|^2\right)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu^o\| = O(1)$, $\frac{1}{p} \text{tr} C^o_a = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p} \text{tr} C^o_a C^o_b = O(1)$
  $\rightarrow$ **Model type:** Marčenko–Pastur + spikes.

- kernel $f\left(\frac{1}{p} \|x_i - x_j\|^2\right)$ with $f'(\tau) = 0$:
  - suboptimal in $\|\mu^o\| \gg O(1)$ (**kills the means**)
  - suboptimal in $\frac{1}{p} \text{tr} C^o_a C^o_b = O(p^{-\frac{1}{2}})$
  $\rightarrow$ **Model type:** smaller order semi-circle law + spikes.
Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel \( f(\frac{1}{p} \|x_i - x_j\|^2) \) with \( f'(\tau) \neq 0 \):
  - optimal in \( \|\mu_a^o\| = O(1) \), \( \frac{1}{p} \text{tr} \, C_a^o = O(p^{-\frac{1}{2}}) \)
  - suboptimal in \( \frac{1}{p} \text{tr} \, C_a^o C_b^o = O(1) \)
  \( \rightarrow \textbf{Model type}: \) Marčenko–Pastur + spikes.

- kernel \( f(\frac{1}{p} \|x_i - x_j\|^2) \) with \( f'(\tau) = 0 \):
  - suboptimal in \( \|\mu_a^o\| \gg O(1) \) (kills the means)
  - suboptimal in \( \frac{1}{p} \text{tr} \, C_a^o C_b^o = O(p^{-\frac{1}{2}}) \)
  \( \rightarrow \textbf{Model type}: \) smaller order semi-circle law + spikes.

Jointly optimal solution:

- evenly weighing Marčenko–Pastur and semi-circle laws
Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel \( f\left( \frac{1}{p} \| x_i - x_j \|^2 \right) \) with \( f'(\tau) \neq 0 \):
  - optimal in \( \| \mu_a^o \| = O(1), \frac{1}{p} \text{tr} C_a^o = O(p^{-\frac{1}{2}}) \)
  - suboptimal in \( \frac{1}{p} \text{tr} C_a^o C_b^o = O(1) \)
  \( \rightarrow \) **Model type:** Marčenko–Pastur + spikes.

- kernel \( f\left( \frac{1}{p} \| x_i - x_j \|^2 \right) \) with \( f'(\tau) = 0 \):
  - suboptimal in \( \| \mu_a^o \| \gg O(1) \) (**kills the means**)
  - suboptimal in \( \frac{1}{p} \text{tr} C_a^o C_b^o = O(p^{-\frac{1}{2}}) \)
  \( \rightarrow \) **Model type:** smaller order semi-circle law + spikes.

Jointly optimal solution:

- evenly weighing Marčenko–Pastur and semi-circle laws
- the “\( \alpha-\beta \)” kernel:
  \[
  f'(\tau) = \frac{\alpha}{\sqrt{p}}, \quad \frac{1}{2} f''(\tau) = \beta.
  \]
We consider now a **fully optimal growth rate setting**

**Assumption (Optimal Growth Rate)**

As \( n \to \infty \),

1. **Data scaling:** \( \frac{p}{n} \to c_0 \in (0, \infty) \), \( \frac{n_a}{n} \to c_a \in (0, 1) \),

2. **Mean scaling:** with \( \mu^o \triangleq \sum_{a=1}^{k} \frac{n_a}{n} \mu_a \) and \( \mu_a^o \triangleq \mu_a - \mu^o \), then \( \| \mu_a^o \| = O(1) \)

3. **Covariance scaling:** with \( C^o \triangleq \sum_{a=1}^{k} \frac{n_a}{n} C_a \) and \( C_a^o \triangleq C_a - C^o \), then

\[
\| C_a \| = O(1), \quad tr C_a^o = O(\sqrt{p}), \quad tr C_a^o C_b^o = O(\sqrt{p}).
\]
New assumption setting

- We consider now a fully optimal growth rate setting

**Assumption (Optimal Growth Rate)**

As \( n \to \infty \),

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

**Kernel:**

- For technical simplicity, we consider

\[
\tilde{K} = P K P = P \left\{ f \left( \frac{1}{p} (x^\circ)^T (x^\circ_j) \right) \right\}_{i,j=1}^n P
\]

\[P = I_n - \frac{1}{n} 1_n 1_n^T.\]

i.e., \( \tau \) replaced by 0.
Main Results

Theorem

As $n \to \infty$,

$$
\left\| \sqrt{p} \left( PKP + (f(0) + \tau f'(0)) P \right) - \hat{K} \right\| \xrightarrow{a.s.} 0
$$

with, for $\alpha = \sqrt{p} f'(0) = O(1)$ and $\beta = \frac{1}{2} f''(0) = O(1)$,

$$
\hat{K} = \alpha PW^T WP + \beta P\Phi P + UAU^T
$$

$$
A = \begin{bmatrix}
\alpha M^T M + \beta T & \alpha I_k \\
\alpha I_k & 0
\end{bmatrix}
$$

$$
U = \begin{bmatrix}
\frac{J}{\sqrt{p}}, PW^T M
\end{bmatrix}
$$

$$
\frac{\Phi}{\sqrt{p}} = \left\{ \left( (\omega_i^o)^T \omega_j^o \right)^2 \delta_{i \neq j} \right\}_{i,j=1}^{n} - \left\{ \frac{\text{tr}(C_a C_b)}{p^2} 1_{n_a} 1_{n_b}^T \right\}_{a,b=1}^{k}.
$$
Main Results

**Theorem**

As $n \to \infty$,

$$\left\| \sqrt{p} \left( PKP + (f(0) + \tau f'(0)) P \right) - \hat{K} \right\| \xrightarrow{a.s.} 0$$

with, for $\alpha = \sqrt{p} f'(0) = O(1)$ and $\beta = \frac{1}{2} f''(0) = O(1)$,

$$\hat{K} = \alpha PW^T WP + \beta P \Phi P + UAU^T$$

$$A = \begin{bmatrix} \alpha M^T M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} J \\ \sqrt{p} \end{bmatrix}, PW^T M$$

$$\Phi = \left\{ \left( (\omega_i^\circ)^T \omega_j^\circ \right)^2 \delta_{i \neq j} \right\}_{i,j=1}^n - \left\{ \frac{tr(C_a C_b)}{p^2} 1_{n_a} 1_{n_b}^T \right\}_{a,b=1}^k.$$

**Role of $\alpha, \beta$:**

- Weighs Marčenko–Pastur versus semi-circle parts.
Limiting eigenvalue distribution

Theorem (Eigenvalues Bulk)
As \( p \to \infty \),

\[
\nu_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\hat{K})} \xrightarrow{a.s.} \nu
\]

with \( \nu \) having Stieltjes transform \( m(z) \) solution of

\[
\frac{1}{m(z)} = -z + \frac{\alpha}{p} \text{tr} C^o \left( I_k + \frac{\alpha m(z)}{c_0} C^o \right)^{-1} - \frac{2\beta^2}{c_0} \omega^2 m(z)
\]

where \( \omega = \lim_{p \to \infty} \frac{1}{p} \text{tr} (C^o)^2 \).
Limiting eigenvalue distribution

Figure: Eigenvalues of $K$ (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\mu_i = 3\delta_i$, $f(x) = \frac{1}{2} \beta \left( x + \frac{1}{\sqrt{p}} \frac{\alpha}{\beta} \right)^2$. (Top left): $\alpha = 8$, $\beta = 1$, (Top right): $\alpha = 4$, $\beta = 3$, (Bottom left): $\alpha = 3$, $\beta = 4$, (Bottom right): $\alpha = 1$, $\beta = 8$. 
Asymptotic performances: MNIST

▶ MNIST is “means-dominant” but not that much!

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$|\mu^1_\circ - \mu^2_\circ|^2$</th>
<th>$\frac{1}{\sqrt{p}} \text{TR} \left( C_1 - C_2 \right)^2$</th>
<th>$\frac{1}{p} \text{TR} \left( C_1 - C_2 \right)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST (digits 1, 7)</td>
<td>612.7</td>
<td>71.1</td>
<td>2.5</td>
</tr>
<tr>
<td>MNIST (digits 3, 6)</td>
<td>441.3</td>
<td>39.9</td>
<td>1.4</td>
</tr>
<tr>
<td>MNIST (digits 3, 8)</td>
<td>212.3</td>
<td>23.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Asymptotic performances: MNIST

MNIST is “means-dominant” but not that much!

| Datasets          | $||\mu_1^0 - \mu_2^0||^2$ | $\frac{1}{\sqrt{p}} \text{tr} (C_1 - C_2)^2$ | $\frac{1}{p} \text{tr} (C_1 - C_2)^2$ |
|-------------------|---------------------------|-----------------------------------------------|------------------------------------------|
| MNIST (digits 1, 7) | 612.7                     | 71.1                                          | 2.5                                      |
| MNIST (digits 3, 6) | 441.3                     | 39.9                                          | 1.4                                      |
| MNIST (digits 3, 8) | 212.3                     | 23.5                                          | 0.8                                      |

Figure: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$. 
Asymptotic performances: EEG data

- EEG data are “variance-dominant”

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$|\mu_1^o - \mu_2^o|^2$</th>
<th>$\frac{1}{\sqrt{p}} \text{Tr} (C_1 - C_2)^2$</th>
<th>$\frac{1}{p} \text{Tr} (C_1 - C_2)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEG (sets A, E)</td>
<td>2.4</td>
<td>10.9</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Asymptotic performances: EEG data

- EEG data are “variance-dominant”

<table>
<thead>
<tr>
<th>DATASETS</th>
<th>$|\mu_1^0 - \mu_2^0|^2$</th>
<th>$\frac{1}{\sqrt{p}} \text{TR} (C_1 - C_2)^2$</th>
<th>$\frac{1}{p} \text{TR} (C_1 - C_2)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEG (sets A, E)</td>
<td>2.4</td>
<td>10.9</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Figure**: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$. 
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case \( f'(\tau) = 0 \)
  Kernel Spectral Clustering: The case \( f'(\tau) = \frac{\alpha}{\sqrt{p}} \)

Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

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 $F_{ia}$ ($d_i = [K1n]_i$)

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

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 \( F_{ia} \) (\( d_i = [K1n]_i \))

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

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

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,

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

where we naturally decompose

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

\[
D = \begin{bmatrix} D_{(l)} & 0 \\ 0 & D_{(u)} \end{bmatrix} = \text{diag}\{K1n\}.
\]
The finite-dimensional intuition: What we expect

Figure: Typical expected performance output
The finite-dimensional intuition: What we expect

Figure: Typical expected performance output
The finite-dimensional intuition: What we expect

Figure: Typical expected performance output
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.
Figure: Centered Vectors \([F_{(u)}^0]., a = [F(u) - \frac{1}{k} F(u) 1_k 1_k^T]., a\), 3-class MNIST data (zeros, ones, twos), \(\alpha = 0\), \(n = 192\), \(p = 784\), \(n_l/n = 1/16\), Gaussian kernel.
MNIST Data Example

\[ F(u) - \frac{1}{k} \sum_{k=1}^{k} F(u) \cdot \alpha = \left[ F(u) - \frac{1}{k} F(u) 1_k 1_k^T \right] \cdot \alpha, \] 3-class MNIST data (zeros, ones, twos), \( \alpha = 0, n = 192, p = 784, n_l/n = 1/16, \) Gaussian kernel.

Figure: Centered Vectors \([F(u)]_{\cdot 1} (Zeros)\) and \([F(u)]_{\cdot 2} (Ones)\)
MNIST Data Example

Figure: Centered Vectors $[F^O_{(u)}]_{:,a} = [F(u) - \frac{1}{k} \sum_{b=1}^{k} F(u)^T]_{:,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Theoretical Findings

Method: Assume $n_l/n \rightarrow c_l \in (0, 1)$

- We aim at characterizing

$$F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$
Theoretical Findings

**Method:** Assume \( n_l/n \to c_l \in (0, 1) \)

- We aim at characterizing

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

- Taylor expansion of \( K \) as \( n, p \to \infty \),

\[
K(u,u) = f(\tau) 1_{n_u} 1_{n_u}^T + O\|\cdot\|(n^{-\frac{1}{2}})
\]

\[
D_u = nf(\tau) I_{n_u} + O(n^{\frac{1}{2}})
\]

and similarly for \( K(u,l), D(l) \).
Theoretical Findings

**Method:** Assume $n_l/n \to c_l \in (0, 1)$

- We aim at characterizing

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

- Taylor expansion of $K$ as $n, p \to \infty$,

\[
K_{(u,u)} = f(\tau) n_u 1_n u^T 1_n u + O \| \cdot \| (n^{-\frac{1}{2}})
\]

\[
D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})
\]

and similarly for $K_{(u,l)}$, $D_{(l)}$.

- So that

\[
\left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} = \left( I_{n_u} - \frac{1_n u 1_n u^T}{n} + O \| \cdot \| (n^{-\frac{1}{2}}) \right)^{-1}
\]

easily Taylor expanded.
Main Results

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

- In the first order,

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

where \( v = O(1) \) random vector (entry-wise) and \( t_a = \frac{1}{\sqrt{p}} \text{tr} C_\alpha^0 \).

Informative terms
Main Results

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

- In the first order,

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

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

- Consequences:
**Main Results**

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

- In the first order,

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

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,

\[
F_{:,a}^{(u)} = C \frac{n_l,a}{n} \left[ v + \alpha \frac{t_a 1_n u}{\sqrt{n}} \right] + O(n^{-1}) + O(n^{-1/2})
\]

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 $n_l,a$

\[
F_{:,a}^{(u)} \text{ to be scaled by } n_l,a
\]
Main Results

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

- In the first order,

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

Informative terms

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

- Consequences:
  - Random non-informative bias $v$
  - Strong Impact of $n_l,a$

  $F_{\cdot,a}(u)$ to be scaled by $n_l,a$

  Additional per-class bias $\alpha t_a 1_{n_u}$

  $\alpha = 0 + \beta \sqrt{p}$. 

Main Results

As a consequence of the remarks above, we take

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

and define

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

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f'''(\tau)}{f(\tau)^2} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \frac{\beta}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2trC_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left[ [M^T C_b M]_{a_1 a_2} + \frac{\delta_{a_2} a_2 p}{n_l a_1} T_{b a_1} \right]$$

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 \( a \neq b \),

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

Some consequences:

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

![Graph showing 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), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.
Figure: Performance as a function of $\alpha$, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_1/n = 1/16$, Gaussian kernel.
Figure: Performance as a function of $\alpha$, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_t/n = 1/16$, Gaussian kernel.
Reminder:
For \( x_i \in C_b \) unlabelled, \( \hat{F}_{i,.} - G_b \to 0 \), \( G_b \sim \mathcal{N}(m_b, \Sigma_b) \) with

\[
(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b
\]

\[
(\Sigma_b)_{a1a2} = \frac{2\text{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a1} t_{a2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^T C_b M]_{a1a2} + \frac{\delta_{a1}^2 p}{n_l,a1} T_{ba1} \right)
\]

with \( t, T, M \) as before, \( \tilde{X}_a = X_a - \sum_{d=1}^{k} \frac{n_l,d}{n_l} X_d^c \) and \( B_b \) bias independent of \( a \).
Is semi-supervised learning really semi-supervised?

Reminder:
For \( x_i \in C_b \) unlabelled, \( \hat{F}_{i,.} - G_b \to 0 \), \( G_b \sim \mathcal{N}(m_b, \Sigma_b) \) with

\[
(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b
\]

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

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 \( a \).

The problem with unlabelled data:

- Result does not depend on \( n_u \)!
  \( \longrightarrow \) increasing \( n_u \) asymptotically non beneficial.
Is semi-supervised learning really semi-supervised?

Reminder:
For $x_i \in C_b$ unlabelled, $\hat{F}_{i,.} - G_b \to 0$, $G_b \sim N(m_b, \Sigma_b)$ with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$

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

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

The problem with unlabelled data:

- Result does not depend on $n_u$!
  $\longrightarrow$ increasing $n_u$ asymptotically non beneficial.

- Even best Laplacian regularizer brings SSL to be merely supervised learning.
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
Resurrecting SSL by centering

Reminder:

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

with \( F_{ia} = \delta\{x_i \in \mathcal{C}_a\} \)

\[ \Leftrightarrow F^{(u)} = \left( I_n - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}. \]
Resurrecting SSL by centering

Reminder:

\[ F = \arg\min_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with} \quad F_{ia}^{(l)} = \delta \{ x_i \in C_a \} \]

\[ \iff F^{(u)} = \left( I_{nu} - D_{(u)}^{-\alpha} K(u,u) D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K(u,l) D_{(l)}^{\alpha-1} F^{(l)}. \]

Domination of score flattening:

- Finite-dimensional intuition imposes \( K_{ij} \) decreasing with \( \|x_i - x_j\| \Rightarrow \) solutions \( F_{ia} \) tend to “flatten”
Reminder:

\[ F = \arg\min_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with} \quad F_{ia} = \delta\{x_i \in C_a\} \]

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

Domination of score flattening:

- Finite-dimensional intuition imposes \( K_{ij} \) decreasing with \( \|x_i - x_j\| \Rightarrow \) solutions \( F_{ia} \) tend to “flatten”

- **Consequence:** \( D^{-\alpha} (u) K(u,u) D^{\alpha-1} (u) \approx \frac{1}{n} 1_{n_u} 1_{n_u}^T \) and clustering information vanishes (not so obvious but can be shown).
Resurrecting SSL by centering

Reminder:

\[ F = \arg\min_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with} \quad F_{ia}^{(l)} = \delta\{x_i \in C_a\} \]

\[ \Leftrightarrow F^{(u)} = \left( I_{nu} - D^{-\alpha}(u) K(u,u) D^{\alpha-1}(u) \right)^{-1} D^{-\alpha}(u) K(u,l) D^{\alpha-1}(l) F^{(l)}. \]

Domination of score flattening:

- Finite-dimensional intuition imposes \( K_{ij} \) decreasing with \( \|x_i - x_j\| \) ⇒ solutions \( F_{ia} \) tend to “flatten”

- **Consequence:** \( D^{-\alpha}(u) K(u,u) D^{\alpha-1}(u) \approx \frac{1}{n} 1_{nu} 1_{nu}^T \) and clustering information vanishes (not so obvious but can be shown).

Solution:

- Forgetting finite-dimensional intuition: “recenter” \( K \) to kill flattening, i.e., use

\[ \tilde{K} = PKP, \quad P = I_n - \frac{1}{n} 1_n 1_n^T. \]
Theoretical results

Setting

- $K = 2$, $x_i \sim \mathcal{N}(\pm \mu, I_p)$
- scores $f_u = (\alpha I_{n_u} - \tilde{K}_{uu})^{-1} \tilde{K}_{ul} f_l$.  

\textbf{Theorem (Asymptotic mean and variance)}

As $n \to \infty$, $j_u(f_u - m_{1n_u})_T (f_u - m_{1n_u}) \to 0$, $(f_u - m_{1n_u})_T D_j (f_u - m_{1n_u}) \to 0$

where, for $i = 1, 2$, $m_i \equiv -c_{ul}(1 - \left[1 + c_{cl}c_1 \parallel \mu \parallel^2 c_{00} \delta_1 + \delta_2 \right]^{-1})$  

$\sigma_i^2 \equiv s_i^2 c_{li}^2 \parallel \mu \parallel^2 c_{00} (1 + \delta_2)^2 - c_{ul} \delta_2^2 (1 + c_{cl}c_2 \parallel \mu \parallel^2 c_{00} \delta_1 + \delta_2)^2 + s_i^2 c_{li}^2 \parallel \mu \parallel^2 c_{00} (1 + c_{cl}c_2 \parallel \mu \parallel^2 c_{00} \delta_1 + \delta_2)^2$ 

\textbf{with} $\delta$ defined as $\delta \equiv -\frac{1}{2} + c_{ul} - c_{00} + \text{sign}(\alpha) \sqrt{\left(\alpha - \alpha - \alpha + \alpha\right)\left(\alpha - \alpha + \alpha\right)^2} \alpha$.  

70 / 113
Theoretical results

Setting

- $K = 2$, $x_i \sim \mathcal{N}(\pm \mu, I_p)$
- scores $f_u = (\alpha I_{n_u} - \tilde{K}_{uu})^{-1} \tilde{K}_{ul} f_l$.

Theorem (Asymptotic mean and variance)

As $n \to \infty$,

$$\frac{j_i^{(u)\top} f_u}{n_{ui}} - m_i \xrightarrow{a.s.} 0, \quad \frac{(f_u - m_i 1_{n_u})\top D_i^{(u)} (f_u - m_i 1_{n_u})}{n_{ui}} - \sigma_i^2 \xrightarrow{a.s.} 0$$

where, for $i = 1, 2$,

$$m_i \equiv -\frac{c_l}{c_u} s_i \left(1 - \left[1 + \frac{c_u c_1 c_2 \|\mu\|^2 c_0}{\delta (1 + \delta)}\right]^{-1}\right)$$

$$\sigma_i^2 \equiv \frac{s_i^2 c_l^2 c_i^2 \|\mu\|^2 \delta^2}{c_0^2 (1 + \delta)^2 - c_u c_0 \delta^2} \frac{1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta^2}{(1 + \delta)^2}}{\left(1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta}{1 + \delta}\right)^2} + \frac{s_i^2 c_l c_i}{1 - c_i} \frac{\delta^2}{c_0 (1 + \delta)^2 - c_u \delta^2}$$

with $\delta$ defined as

$$\delta \equiv -\frac{1}{2} + \frac{c_u - c_0 + \text{sign}(\alpha) \sqrt{(\alpha - \alpha_-)(\alpha - \alpha_+)}}{2\alpha}.$$
Performance as a function of $n_u, n_l$

**Figure:** Correct classification rate, at optimal $\alpha$, as a function of (i) $n_u$ for fixed $p/n_l = 5$ (blue) and (ii) $n_l$ for fixed $p/n_u = 5$ (black); $c_1 = c_2 = \frac{1}{2}$; different values for $\|\mu\|$. Comparison to optimal Neyman–Pearson performance for known $\mu$ (in red).
The spike case or not (1)

\textbf{Marčenko–Pastur + spike limit}

- limiting eigenvalue distribution is \textit{Marčenko–Pastur law}

\[
\text{If } \|\mu\|_2 > 1, \text{ then there is a leading isolated eigenvalue.}
\]

In the presence of a spike, the empirical eigenvalues can diverge from the \textit{Marčenko–Pastur law}.
The spike case or not (1)

Marčenko–Pastur + spike limit

- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of isolated spike if

\[ \|\mu\|^2 > \frac{1}{c_1 c_2} \sqrt{\frac{c_0}{c_u}}. \]
The spike case or not (1)

**Marčenko–Pastur + spike limit**

- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of **isolated spike iif**

\[
\|\mu\|^2 > \frac{1}{c_1 c_2} \sqrt{\frac{c_0}{c_u}}.
\]

- determines **existence or not of unsupervised spectral clustering solution**.
The spike case or not (1)

Marčenko–Pastur + spike limit
- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of isolated spike iff

\[
\|\mu\|^2 > \frac{1}{c_1 c_2} \sqrt{\frac{c_0}{c_u}}.
\]

- determines existence or not of unsupervised spectral clustering solution.

Figure: Eigenvalue distribution of $K_{uu}$ versus the (scaled) Marčenko–Pastur law with Stieltjes transform $\delta$, for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$. The value $\|\mu\| = 2.5$ ensures the presence of a leading isolated eigenvalue (spike).
The spike case or not (2)

\[ \delta(\alpha) = -\left(1 + c_u c_1 c_2 c_0^{-1} \|\mu\|^2\right)^{-1} \]

**Figure:** Asymptotic correct classification probability \( \Phi \left( \frac{m_1}{\sigma_1} \right) \) as a function of \( \alpha \) for \( c_u = \frac{9}{10} \), \( c_0 = \frac{1}{2} \), \( c_1 = \frac{1}{2} \), two different values of \( \|\mu\| \), below and above phase transition.
SSL: the road from supervised to unsupervised

Figure: Theory (solid) versus practice (dashed; from right to left: $n = 400, 1000, 4000$): correct classification probability as a function of $\alpha$ for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, and left: $\|\mu\| = 1.5$ (below phase transition); right: $\|\mu\| = 2.5$ (above phase transition). Different values of $n$. 
## Experimental evidence: MNIST

<table>
<thead>
<tr>
<th>Digits</th>
<th>(0,8)</th>
<th>(2,7)</th>
<th>(6,9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_u = 100$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centered kernel</td>
<td>89.5±3.6</td>
<td>89.5±3.4</td>
<td>85.3±5.9</td>
</tr>
<tr>
<td>Iterated centered kernel</td>
<td>89.5±3.6</td>
<td>89.5±3.4</td>
<td>85.3±5.9</td>
</tr>
<tr>
<td>Laplacian</td>
<td>75.5±5.6</td>
<td>74.2±5.8</td>
<td>70.0±5.5</td>
</tr>
<tr>
<td>Iterated Laplacian</td>
<td>87.2±4.7</td>
<td>86.0±5.2</td>
<td>81.4±6.8</td>
</tr>
<tr>
<td>Manifold</td>
<td>88.0±4.7</td>
<td>88.4±3.9</td>
<td>82.8±6.5</td>
</tr>
<tr>
<td></td>
<td>$n_u = 1000$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centered kernel</td>
<td>92.2±0.9</td>
<td>92.5±0.8</td>
<td>92.6±1.6</td>
</tr>
<tr>
<td>Iterated centered kernel</td>
<td><strong>92.3±0.9</strong></td>
<td><strong>92.5±0.8</strong></td>
<td><strong>92.9±1.4</strong></td>
</tr>
<tr>
<td>Laplacian</td>
<td>65.6±4.1</td>
<td>74.4±4.0</td>
<td>69.5±3.7</td>
</tr>
<tr>
<td>Iterated Laplacian</td>
<td><strong>92.2±0.9</strong></td>
<td>92.4±0.9</td>
<td>92.0±1.6</td>
</tr>
<tr>
<td>Manifold</td>
<td>91.1±1.7</td>
<td>91.4±1.9</td>
<td>91.4±2.0</td>
</tr>
</tbody>
</table>

**Table:** Comparison of classification accuracy (%) on MNIST datasets with $n_l = 10$. Computed over 1000 random iterations for $n_u = 100$ and 100 for $n_u = 1000$. 
### Experimental evidence: Traffic signs (HOG features)

<table>
<thead>
<tr>
<th>Class ID</th>
<th>(2,7)</th>
<th>(9,10)</th>
<th>(11,18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_u = 100$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centered kernel</td>
<td>79.0±10.4</td>
<td>77.5±9.2</td>
<td>78.5±7.1</td>
</tr>
<tr>
<td>Iterated centered kernel</td>
<td><strong>85.3±5.9</strong></td>
<td><strong>89.2±5.6</strong></td>
<td><strong>90.1±6.7</strong></td>
</tr>
<tr>
<td>Laplacian</td>
<td>73.8±9.8</td>
<td>77.3±9.5</td>
<td>78.6±7.2</td>
</tr>
<tr>
<td>Iterated Laplacian</td>
<td>83.7±7.2</td>
<td>88.0±6.8</td>
<td>87.1±8.8</td>
</tr>
<tr>
<td>Manifold</td>
<td>77.6±8.9</td>
<td>81.4±10.4</td>
<td>82.3±10.8</td>
</tr>
<tr>
<td>$n_u = 1000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centered kernel</td>
<td>83.6±2.4</td>
<td>84.6±2.4</td>
<td>88.7±9.4</td>
</tr>
<tr>
<td>Iterated centered kernel</td>
<td><strong>84.8±3.8</strong></td>
<td><strong>88.0±5.5</strong></td>
<td><strong>96.4±3.0</strong></td>
</tr>
<tr>
<td>Laplacian</td>
<td>72.7±4.2</td>
<td>88.9±5.7</td>
<td>95.8±3.2</td>
</tr>
<tr>
<td>Iterated Laplacian</td>
<td>83.0±5.5</td>
<td>88.2±6.0</td>
<td>92.7±6.1</td>
</tr>
<tr>
<td>Manifold</td>
<td>77.7±5.8</td>
<td>85.0±9.0</td>
<td>90.6±8.1</td>
</tr>
</tbody>
</table>

**Table:** Comparison of classification accuracy (%) on German Traffic Sign datasets with $n_l = 10$. Computed over 1000 random iterations for $n_u = 100$ and 100 for $n_u = 1000$. 
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
**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 $\sigma$. 

\[ X = [x_1, \ldots, x_T] \]
\[ \sigma(Wx_t) \]
**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 $\beta \in \mathbb{R}^{n \times d}$
Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$
Random Feature Maps and Extreme Learning Machines

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- **Training MSE:**

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

  with

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

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

**Training MSE**:

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

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 \(\hat{T}\),

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

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

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

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

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

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

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

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

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., $\frac{1}{p} w^T XAX^T w \simeq \frac{1}{p} \text{tr} XAX^T$)
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:

$$P \left( \left| \frac{1}{p} \sigma(w^T X) A \sigma(X^T w) - \frac{1}{p} \text{tr} \Phi A \right| > t \right) \leq Ce^{-cn \min(t, t^2)}$$
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} \arcsin \left( \frac{2 a^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>$\text{sign}(t)$</td>
<td>$1 - \frac{1}{\pi} \arccos(\angle(a, b))$</td>
</tr>
<tr>
<td>$\cos(t)$</td>
<td>$\exp(-\frac{1}{2} (</td>
</tr>
</tbody>
</table>

where $\angle(a, b) \equiv \frac{a^T b}{||a|| ||b||}$.
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} |a| |b| \left( \angle(a, b) \cos(-\angle(a, b)) + \sqrt{1 - \angle(a, b)^2} \right) )</td>
</tr>
<tr>
<td>(</td>
<td>t</td>
</tr>
<tr>
<td>( \text{erf}(t) )</td>
<td>( \frac{2}{\pi} \arcsin \left( \frac{2a^Tb}{\sqrt{(1+2|a|^2)(1+2|b|^2)}} \right) )</td>
</tr>
<tr>
<td>( \mathbb{1}_{{t&gt;0}} )</td>
<td>( \frac{1}{2} - \frac{1}{2\pi} \arccos(\angle(a, b)) )</td>
</tr>
<tr>
<td>( \text{sign}(t) )</td>
<td>( 1 - \frac{1}{2\pi} \arccos(\angle(a, b)) )</td>
</tr>
<tr>
<td>( \cos(t) )</td>
<td>( \exp(-\frac{1}{2}(|a|^2 + |b|^2)) \cosh(a^Tb) )</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** \( \mathbb{E}[w_i^k] = m_k \ (m_1 = 0) \), \( \sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0 \)

\[
\Phi(a, b) = \zeta_2^2 \left[ m_k^2 \left( 2(a^Tb)^2 + \|a\|^2 \|b\|^2 \right) + (m_4 - 3m_2^2)(a^2)^T(b^2) \right] + \zeta_1^2 m_2 a^T b \\
+ \zeta_2 \zeta_1 m_3 \left[ (a^2)^T b + a^T(b^2) \right] + \zeta_0^2 m_2 \left[ \|a\|^2 + \|b\|^2 \right] + \zeta_0^2
\]

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

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

For all $\varepsilon > 0$,

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

almost surely, where

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

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

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

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

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

$$+ \frac{\frac{1}{n} \text{tr} Y^T Y \bar{Q} \Psi \bar{Q}}{1 - \frac{1}{n} \text{tr} (\Psi \bar{Q})^2} \left[ \frac{1}{\hat{T}} \text{tr} \Psi \hat{X} \hat{X} - \frac{1}{\hat{T}} \text{tr} (I_T + \gamma \bar{Q})(\Psi_{X\hat{X}} \Psi \hat{X} \hat{X} \bar{Q}) \right]$$

with $\Psi_{AB} = \frac{n}{\hat{T}} \Phi_{AB}^1 + \delta$, $\Phi_{AB} = E[\sigma(A^T w)\sigma(w^T B)]$. 
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), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 

$$
\sigma(t) = t
$$
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), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Simulations on MNIST: Lipschitz $\sigma(\cdot)$

![Graph showing neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of $\gamma$, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.](image)

**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), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 

$\sigma(t) = t \quad \sigma(t) = |t| \quad \sigma(t) = \max(t, 0) \quad \sigma(t) = \text{erf}(t)$
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), $n = 512$, $T = \hat{T} = 1024$, $p = 784$. 
Deeper investigation on $\Phi$

**Statistical Assumptions on $X$**

- **Gaussian mixture model**

\[
x_i \in C_a \iff x_i \sim \mathcal{N}(\frac{1}{\sqrt{p}}\mu_a, \frac{1}{p}C_a).
\]

- **Growth rate:** $\|\mu_a\| = O(1)$, $\frac{1}{\sqrt{p}}\text{tr} C_a = O(1)$. 
Deeper investigation on $\Phi$

Statistical Assumptions on $X$

- Gaussian mixture model

$$x_i \in C_a \Leftrightarrow x_i \sim \mathcal{N}\left( \frac{1}{\sqrt{p}}\mu_a, \frac{1}{p}C_a \right).$$

- **Growth rate:** $\|\mu_a^o\| = O(1)$, $\frac{1}{\sqrt{p}} \text{tr} C_a^o = O(1)$.

Theorem

As $p, T \to \infty$, for all $\sigma(\cdot)$ given in next table,

$$\|P\Phi P - P\tilde{\Phi}P\| \overset{a.s.}{\to} 0$$

with

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

$$U \equiv \left[ \frac{J}{\sqrt{p}}, \phi \right]$$

$$B \equiv \begin{bmatrix} tt^T + 2T & t \\ t^T & 1 \end{bmatrix}$$

and $d_0, d_1, d_2$ given in next table ($\phi_i = \|w_i\|^2 - E[\|w_i\|^2]$ for $x_i = \frac{1}{\sqrt{p}}\mu_a + w_i$).
Deeper investigation on $\Phi$

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

<table>
<thead>
<tr>
<th>$\sigma(t)$</th>
<th>$d_0$</th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>0</td>
<td>$\frac{1}{4} \left( 1 - \frac{1}{2\pi} \right) \tau$</td>
<td>$\frac{1}{8\pi \tau}$ $(\zeta_+ + \zeta_-)^2$</td>
</tr>
<tr>
<td>ReLU$(t)$</td>
<td>$\frac{1}{4} - \frac{1}{2\pi} \tau$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{8\pi \tau}$ $(\zeta_+ + \zeta_-)^2$</td>
</tr>
<tr>
<td>$</td>
<td>t</td>
<td>$</td>
<td>$1 - \frac{2}{\pi}$</td>
</tr>
<tr>
<td>LReLU$(t)$</td>
<td>$\frac{\pi - 2}{4\pi} (\zeta_+ + \zeta_-)^2 \tau$</td>
<td>$\frac{1}{4} (\zeta_+ - \zeta_-)^2$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1_{t&gt;0}$</td>
<td>$\frac{1}{4} - \frac{1}{2\pi}$</td>
<td>$\frac{1}{8\tau \pi}$ $(\zeta_+ + \zeta_-)^2$</td>
<td>$0$</td>
</tr>
<tr>
<td>sign$(t)$</td>
<td>$\frac{1}{4}$ - $\frac{1}{2\pi}$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\zeta_2 t^2 + \zeta_1 t + \zeta_0$</td>
<td>$2\tau^2 \zeta_2$</td>
<td>$\frac{1}{2} \frac{1}{2} + e^{-2\tau} - e^{-\tau}$</td>
<td>$\frac{1}{4} \zeta_2^2$</td>
</tr>
<tr>
<td>cos$(t)$</td>
<td>$\frac{1}{2} + e^{-2\tau} - e^{-\tau}$</td>
<td>$0$</td>
<td>$\frac{1}{4}$ $e^{-\tau}$</td>
</tr>
<tr>
<td>sin$(t)$</td>
<td>$\frac{1}{2} - e^{-2\tau} - \tau e^{-\tau}$</td>
<td>$e^{-\tau}$</td>
<td>$0$</td>
</tr>
<tr>
<td>erf$(t)$</td>
<td>$\frac{2}{\pi} \left( \arccos \left( \frac{2\tau}{2\tau + 1} \right) - \frac{2\tau}{2\tau + 1} \right)$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>exp$(-\frac{t^2}{2})$</td>
<td>$\frac{1}{\sqrt{2\tau + 1}} - \frac{1}{\tau + 1}$</td>
<td>$0$</td>
<td>$\frac{1}{4(\tau + 1)^3}$</td>
</tr>
</tbody>
</table>

where

- ReLU$(t) = \max(t, 0)$
- LReLU$(t) = \zeta_+ \max(t, 0) + \zeta_- \max(-t, 0)$. 

Deeper investigation on $\Phi$

Three groups of functions $\sigma(\cdot)$ emerge:
- “means-oriented”: $d_2 = 0$
- “covariance-oriented”: $d_1 = 0$
- “balanced”: $d_1, d_2 \neq 0$
Deeper investigation on $\Phi$

Three groups of functions $\sigma(\cdot)$ emerge:

- "means-oriented": $d_2 = 0$
- "covariance-oriented": $d_1 = 0$
- "balanced": $d_1, d_2 \neq 0$

Case of the Leaky–ReLU

- $\sigma(t) = \varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$
Deeper investigation on $\Phi$

Three groups of functions $\sigma(\cdot)$ emerge:
- “means-oriented”: $d_2 = 0$
- “covariance-oriented”: $d_1 = 0$
- “balanced”: $d_1, d_2 \neq 0$

Case of the Leaky–ReLU
- $\sigma(t) = \varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$

![Figure: Eigenvectors 1 and 2 of $P\Phi P$ for: $\mathcal{N}(\mu_1, C_1)$, $\mathcal{N}(\mu_1, C_2)$, $\mathcal{N}(\mu_2, C_1)$, $\mathcal{N}(\mu_2, C_2)$]
Deeper investigation on $\Phi$: Simulation results

Table: Clustering accuracies for different $\sigma(t)$ on MNIST dataset ($n = 32$).

<table>
<thead>
<tr>
<th>$\sigma(t)$</th>
<th>$T = 32$</th>
<th>$T = 64$</th>
<th>$T = 128$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>85.31%</td>
<td>88.94%</td>
<td>87.30%</td>
</tr>
<tr>
<td>$1_{t&gt;0}$</td>
<td>86.00%</td>
<td>82.94%</td>
<td>85.56%</td>
</tr>
<tr>
<td>$\text{sign}(t)$</td>
<td>81.94%</td>
<td>83.34%</td>
<td>85.22%</td>
</tr>
<tr>
<td>$\sin(t)$</td>
<td>85.31%</td>
<td>87.81%</td>
<td>87.50%</td>
</tr>
<tr>
<td>$\text{erf}(t)$</td>
<td>86.50%</td>
<td>87.28%</td>
<td>86.59%</td>
</tr>
</tbody>
</table>

Mean-oriented

| $|t|$ | 62.81% | 60.41% | 57.81% |
| $\cos(t)$ | 62.50% | 59.56% | 57.72% |
| $\exp(-\frac{t^2}{2})$ | 64.00% | 60.44% | 58.67% |

Cov-oriented

Balanced | $(t)$ | 82.87% | 85.72% | 82.27% |
Deeper investigation on $\Phi$: Simulation results

**Table:** Clustering accuracies for different $\sigma(t)$ on epileptic EEG dataset ($n = 32$).

<table>
<thead>
<tr>
<th></th>
<th>$\sigma(t)$</th>
<th>$T = 32$</th>
<th>$T = 64$</th>
<th>$T = 128$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MEAN-ORIENTED</strong></td>
<td>$t$</td>
<td>71.81%</td>
<td>70.31%</td>
<td>69.58%</td>
</tr>
<tr>
<td></td>
<td>$1_{t&gt;0}$</td>
<td>65.19%</td>
<td>65.87%</td>
<td>63.47%</td>
</tr>
<tr>
<td></td>
<td>sign($t$)</td>
<td>67.13%</td>
<td>64.63%</td>
<td>63.03%</td>
</tr>
<tr>
<td></td>
<td>sin($t$)</td>
<td>71.94%</td>
<td>70.34%</td>
<td>68.22%</td>
</tr>
<tr>
<td></td>
<td>erf($t$)</td>
<td>69.44%</td>
<td>70.59%</td>
<td>67.70%</td>
</tr>
<tr>
<td><strong>COV-ORIENTED</strong></td>
<td>$</td>
<td>t</td>
<td>$</td>
<td>99.69%</td>
</tr>
<tr>
<td></td>
<td>cos($t$)</td>
<td>99.00%</td>
<td>99.38%</td>
<td>99.36%</td>
</tr>
<tr>
<td></td>
<td>$\exp(-\frac{t^2}{2})$</td>
<td><strong>99.81%</strong></td>
<td><strong>99.81%</strong></td>
<td><strong>99.77%</strong></td>
</tr>
<tr>
<td><strong>BALANCED</strong></td>
<td>$(t)$</td>
<td>84.50%</td>
<td>87.91%</td>
<td>90.97%</td>
</tr>
</tbody>
</table>
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$
  Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

Perspectives
System Setting

Undirected graph with $n$ nodes, $m$ edges:
- “intrinsic” average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.

- “intrinsic” average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
System Setting

Undirected graph with $n$ nodes, $m$ edges:

- “intrinsic” average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- $k$ classes $C_1, \ldots, C_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment $C_{ab}$ between $C_a$ and $C_b$
System Setting

Undirected graph with $n$ nodes, $m$ edges:

- “intrinsic” average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- $k$ classes $C_1, \ldots, C_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment $C_{ab}$ between $C_a$ and $C_b$
- Edge probability for nodes $i \in C_{g_i}$:

$$P(i \sim j) = q_i q_j C_{g_i g_j}.$$
System Setting

Undirected graph with $n$ nodes, $m$ edges:

- “intrinsic” average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- $k$ classes $C_1, \ldots, C_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment $C_{ab}$ between $C_a$ and $C_b$
- edge probability for nodes $i \in C_{g_i}$:
  \[ P(i \sim j) = q_i q_j C_{g_i g_j}. \]

- adjacency matrix $A$ with
  \[ A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j}) \]
Limitations of Classical Methods

- 3 classes with $\mu$ bi-modal ($\mu = \frac{3}{4} \delta_{0.1} + \frac{1}{4} \delta_{0.5}$)
Limitations of Classical Methods

- 3 classes with $\mu$ bi-modal ($\mu = \frac{3}{4}\delta_{0.1} + \frac{1}{4}\delta_{0.5}$)

(Modularity $A - \frac{dd^T}{2m}$)  (Bethe Hessian $D - rA$)
Recall: \[ P(i \sim j) = q_i q_j C_{gi gj}. \]
Proposed Regularized Modularity Approach

Recall: \( P(i \sim j) = q_i q_j C_{g_i g_j} \).

Dense Regime Assumptions: Non trivial regime when, \( \forall a, b \), as \( n \to \infty \),

\[
C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \quad M_{ab} = O(1).
\]
Proposed Regularized Modularity Approach

**Recall:** \( P(i \sim j) = q_i q_j C_{g_i g_j} \).

**Dense Regime Assumptions:** Non trivial regime when, \( \forall a, b \), as \( n \to \infty \),

\[
C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \quad M_{ab} = O(1).
\]

Community information is **weak but highly redundant**
Proposed Regularized Modularity Approach

Recall: \( P(i \sim j) = q_i q_j C_{g_i g_j} \).

Dense Regime Assumptions: Non trivial regime when, \( \forall a, b \) as \( n \to \infty \),

\[
C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}, \quad M_{ab} = O(1).
\]

Community information is weak but highly redundant

Considered Matrix:

\[
L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^T}{2m} \right] D^{-\alpha}.
\]
Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)
As $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \xrightarrow{a.s.} 0$, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^T}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D^{-\alpha} q X D^{-\alpha} + U \Lambda U^T$$

with $D_q = \text{diag}(\{q_i\})$, $X$ zero-mean random matrix with variance profile,

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D^{-\alpha} X 1_n \end{bmatrix}, \quad \text{rank } k + 1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^T) M (I_k - c 1_k^T) & -1_k \\ 1_k^T & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^T, 0, \ldots, 0]^T \in \mathbb{R}^n$. 
Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)

As $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \xrightarrow{a.s.} 0$, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^T}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} XD_q^{-\alpha} + U \Lambda U^T$$

with $D_q = \text{diag}(\{q_i\})$, $X$ zero-mean random matrix with variance profile,

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \text{ rank } k + 1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^T) M (I_k - c 1_k^T) & -1_k \\ 1_k^T & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1^T_{n_a}, 0, \ldots, 0]^T \in \mathbb{R}^n$.

Consequences:

- isolated eigenvalues beyond phase transition $\Leftrightarrow \lambda(M) > \text{“spectrum edge”}$

Optimal choice $\alpha_{\text{opt}}$ of $\alpha$ from study of limiting spectrum.
Asymptotic Equivalence

Theorem (Limiting Random Matrix Equivalent)

As \( n \to \infty \), \( \| \mathcal{L}_\alpha - \tilde{\mathcal{L}}_\alpha \| \xrightarrow{a.s.} 0 \), where

\[
\mathcal{L}_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^T}{2m} \right] D^{-\alpha}
\]

\[
\tilde{\mathcal{L}}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} XD_q^{-\alpha} + U\Lambda U^T
\]

with \( D_q = \text{diag}(\{q_i\}) \), \( X \) zero-mean random matrix with variance profile,

\[
U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} XD_q^{-\alpha} \end{bmatrix}, \quad \text{rank } k + 1
\]

\[
\Lambda = \begin{bmatrix} (I_k - 1_k c^T)M(I_k - c1_k^T) & -1_k \\ 1_k^T & 0 \end{bmatrix}
\]

and \( J = [j_1, \ldots, j_k] \), \( j_a = [0, \ldots, 0, 1_{n_a}^T, 0, \ldots, 0]^T \in \mathbb{R}^n \).

Consequences:

▶ isolated eigenvalues beyond phase transition \( \Leftrightarrow \lambda(M) > \text{“spectrum edge”} \)

Optimal choice \( \alpha_{\text{opt}} \) of \( \alpha \) from study of limiting spectrum.

▶ eigenvectors correlated to \( D_q^{1-\alpha} J \)

Necessary regularization by \( D^{\alpha-1} \).
Eigenvalue Spectrum

Figure: 3 classes, \( c_1 = c_2 = 0.3, c_3 = 0.4, \mu = \frac{1}{2} \delta_{0.4} + \frac{1}{2} \delta_{0.9}, \) \( M = 4 \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix} \).
Theorem (Phase Transition)

*Isolated eigenvalue* $\lambda_i(L_\alpha)$ if $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (D(c) - cc^T)M$, where

$$
\tau^\alpha = \lim_{x \downarrow S_+^\alpha} \frac{1}{g^\alpha(x)}, \text{ phase transition threshold}
$$

with $[S_-^\alpha, S_+^\alpha]$ limiting eigenvalue support of $L_\alpha$ and $g^\alpha(x)$ ($|x| > S_+^\alpha$) solution of

$$
\begin{align*}
\hat{f}^\alpha(x) &= \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq) \\
\hat{g}^\alpha(x) &= \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq).
\end{align*}
$$

In this case, $\lambda_i(L_\alpha) \xrightarrow{a.s.} (g^\alpha)^{-1} (-1/\lambda_i(\bar{M})).$
Theorem (Phase Transition)

Isolated eigenvalue \( \lambda_i(L_\alpha) \) if \(|\lambda_i(\bar{M})| > \tau_\alpha\), \( \bar{M} = (\mathcal{D}(c) - cc^T)M \), where

\[
\tau_\alpha = \lim_{x \downarrow S_+^\alpha} \frac{1}{g_\alpha(x)}, \quad \text{phase transition threshold}
\]

with \([S_-^\alpha, S_+^\alpha]\) limiting eigenvalue support of \(L_\alpha\) and \(g_\alpha(x)\) \((|x| > S_+^\alpha)\) solution of

\[
f_\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f_\alpha(x) + q^{2-2\alpha} g_\alpha(x)} \mu(dq)
\]

\[
g_\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f_\alpha(x) + q^{2-2\alpha} g_\alpha(x)} \mu(dq).
\]

In this case, \( \lambda_i(L_\alpha) \overset{a.s.}{\to} (g_\alpha)^{-1} \left(-1/\lambda_i(\bar{M})\right)\).

Clustering possible when \( \lambda_i(\bar{M}) > (\min_\alpha \tau_\alpha)\):

- “Optimal” \( \alpha_{opt} \equiv \arg\min_\alpha \{\tau_\alpha\} \).
Phase Transition

**Theorem (Phase Transition)**

*Isolated eigenvalue* $\lambda_i(L_\alpha)$ if $|\lambda_i(\bar{M})| > \tau_\alpha$, $\bar{M} = (D(c) - cc^T)M$, where

$$\tau_\alpha = \lim_{x \downarrow \alpha} \left( S^\alpha_+ - \frac{1}{g^\alpha(x)} \right), \text{ phase transition threshold}$$

with $[S^-_\alpha, S^+_\alpha]$ limiting eigenvalue support of $L_\alpha$ and $g^\alpha(x) (|x| > S^\alpha_+)$ solution of

$$f^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq)$$

$$g^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq).$$

In this case, $\lambda_i(L_\alpha) \xrightarrow{a.s.} (g^\alpha)^{-1} \left( -1/\lambda_i(\bar{M}) \right)$.

**Clustering possible** when $\lambda_i(\bar{M}) > (\min_\alpha \tau_\alpha)$:

- “Optimal” $\alpha_{\text{opt}} \equiv \arg\min_\alpha \{\tau_\alpha\}$.
- From $\hat{q}_i \equiv \frac{d_i}{\sqrt{d_i^T 1_n}} \xrightarrow{a.s.} q_i$, $\mu \simeq \hat{\mu} \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{\hat{q}_i}$ and thus:

  **Consistent estimator** $\hat{\alpha}_{\text{opt}}$ of $\alpha_{\text{opt}}$. 
Simulated Performance Results (2 masses of $q_i$)

(Modularity $A - \frac{dd^T}{2m}$)  
(Bethe Hessian $D - rA$)
Simulated Performance Results (2 masses of $q_i$)

Figure: 3 classes, $\mu = \frac{3}{4} \delta_{0.1} + \frac{1}{4} \delta_{0.5}, c_1 = c_2 = \frac{1}{4}, c_3 = \frac{1}{2}, M = 100I_3.$
Simulated Performance Results (2 masses of $q_i$)

Figure: 3 classes, $\mu = \frac{3}{4}\delta_{0.1} + \frac{1}{4}\delta_{0.5}$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $M = 100I_3$. 
Simulated Performance Results (2 masses for $q_i$)

**Figure:** Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta q_{(1)} + \frac{1}{4}\delta q_{(2)}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{opt} = 0.07$. 
Simulated Performance Results (2 masses for $q_i$)

Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta q(1) + \frac{1}{4}\delta q(2)$ with $q(1) = 0.1$ and $q(2) = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{opt} = 0.07$. 
Simulated Performance Results (2 masses for $q_i$)

Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4} \delta q(1) + \frac{1}{4} \delta q(2)$ with $q(1) = 0.1$ and $q(2) = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{opt} = 0.07$. 
Simulated Performance Results (2 masses for $q_i$)

Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4} \delta q(1) + \frac{1}{4} \delta q(2)$ with $q(1) = 0.1$ and $q(2) = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$. 
Simulated Performance Results (2 masses for $q_i$)

Figure: Overlap performance for $n = 3000$, $K = 3$, $\mu = \frac{3}{4}\delta q_1 + \frac{1}{4}\delta q_2$ with $q_1 = 0.1$ and $q_2 \in [0.1, 0.9]$, $M = 10(2I_3 - 1_31_3^T)$, $c_i = \frac{1}{3}$. 
Real Graph Example: PolBlogs ($n = 1490$, two classes)

\[ L_0 \]

\[ (\lambda_{\text{max}} \simeq 1.75) \]

\[ L_1 \]

\[ (\lambda_{\text{max}} \simeq 483) \]

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Overlap</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{\text{opt}} \ (\simeq 0)$</td>
<td>0.897</td>
<td>0.4246</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td>0.035</td>
<td>$\simeq 0$</td>
</tr>
<tr>
<td>$\alpha = 1$</td>
<td>0.040</td>
<td>$\simeq 0$</td>
</tr>
<tr>
<td>BH</td>
<td>0.304</td>
<td>0.2723</td>
</tr>
</tbody>
</table>
Outline

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

Applications
  Reminder on Spectral Clustering Methods
  Kernel Spectral Clustering
  Kernel Spectral Clustering: The case $f'(\tau) = 0$
  Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Semi-supervised Learning
  Semi-supervised Learning improved
  Random Feature Maps, Extreme Learning Machines, and Neural Networks
  Community Detection on Graphs

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 I

Kernel methods.

✓ Spectral clustering
✓ Subspace spectral clustering \( (f'(\tau) = 0) \)
◆ Spectral clustering with outer product kernel \( f(x^Ty) \)
✓ 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.


Summary of Results and Perspectives I

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 II


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.
Thank you.