

Random Matrix Advances in Machine Learning and Neural Nets

(EUSIPCO'2018, Rome, Italy)

Romain COUILLET, Zhenyu LIAO, Xiaoyi MAI

CentraleSupélec, L2S, University of ParisSaclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, University Grenoble-Alpes, France.

September 3rd, 2018



CentraleSupélec



Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Context

Baseline scenario: $x_1, \dots, x_n \in \mathbb{R}^p$ (or \mathbb{C}^p) i.i.d. with $E[x_1] = 0$, $E[x_1 x_1^\top] = C_p$:

- ▶ If $x_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

- ▶ If $n \rightarrow \infty$, then, **strong law of large numbers**

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

or equivalently, **in spectral norm**

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

Random Matrix Regime

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

$$\|\hat{C}_p - C_p\| \not\rightarrow 0.$$

- ▶ For practical p, n with $p \simeq n$, leads to dramatically wrong conclusions
- ▶ **Even for $n = 100 \times p$.**

The Large Dimensional Fallacies

Setting: $x_i \in \mathbb{R}^p$ i.i.d., $x_1 \sim \mathcal{CN}(0, I_p)$

- ▶ assume $p = p(n)$ such that $p/n \rightarrow c > 1$
- ▶ then, **joint point-wise convergence**

$$\max_{1 \leq i, j \leq p} \left| [\hat{C}_p - I_p]_{ij} \right| = \max_{1 \leq i, j \leq p} \left| \frac{1}{n} X_{j, \cdot} X_{i, \cdot}^\top - \delta_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- ▶ however, **eigenvalue mismatch**

$$\begin{aligned} 0 &= \lambda_1(\hat{C}_p) = \dots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \dots \leq \lambda_p(\hat{C}_p) \\ 1 &= \lambda_1(I_p) = \dots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \dots = \lambda_p(I_p) \end{aligned}$$

\Rightarrow **no convergence in spectral norm.**

The Marčenko–Pastur law

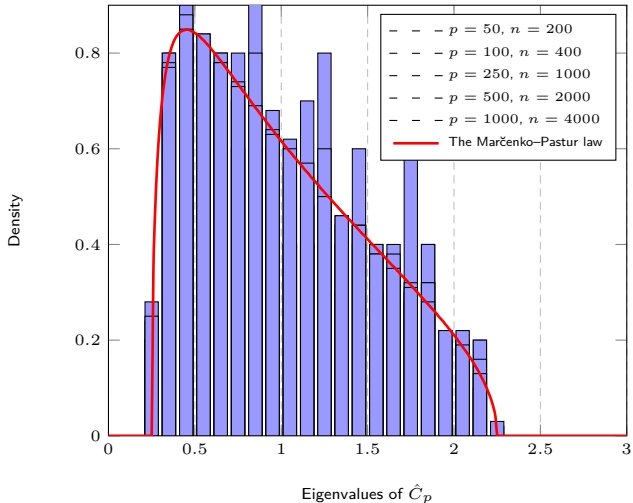


Figure: Histogram of the eigenvalues of \hat{C}_p for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

Definition (Empirical Spectral Distribution)

Empirical spectral distribution (e.s.d.) μ_p of Hermitian matrix $A_p \in \mathbb{R}^{p \times p}$ is

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

Theorem (Marčenko–Pastur Law [Marčenko, Pastur'67])

$X_p \in \mathbb{R}^{p \times n}$ with i.i.d. zero mean, unit variance entries.

As $p, n \rightarrow \infty$ with $p/n \rightarrow c \in (0, \infty)$, e.s.d. μ_p of $\frac{1}{n} X_p X_p^T$ satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

in distribution (i.e., $\int f(t) \mu_p(dt) \xrightarrow{\text{a.s.}} \int f(t) \mu_{(c)}(dt)$ for all bounded continuous f), where

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

$$f_c(x) = \frac{1}{2\pi c x} \sqrt{(x - (1 - \sqrt{c})^2)((1 + \sqrt{c})^2 - x)}.$$

The Marčenko–Pastur law

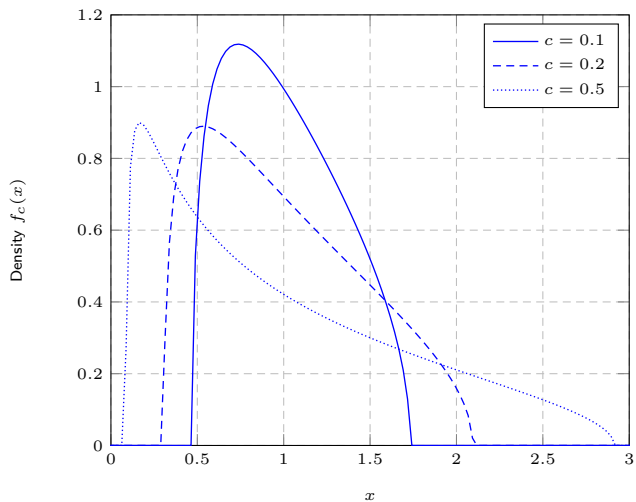


Figure: Marčenko–Pastur law for different limit ratios $c = \lim_{p \rightarrow \infty} p/n$.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

The Stieltjes transform

Definition (Stieltjes Transform)

For μ real probability measure of support $\text{supp}(\mu)$, Stieltjes transform m_μ defined, for $z \in \mathbb{C} \setminus \text{supp}(\mu)$, as

$$m_\mu(z) = \int \frac{1}{t - z} \mu(dt).$$

Property (Inverse Stieltjes Transform)

For $a < b$ continuity points of μ ,

$$\mu([a, b]) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + i\varepsilon)] dx$$

Besides, if μ has a density f at x ,

$$f(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m_\mu(x + i\varepsilon)].$$

The Stieltjes transform

Property (Relation to e.s.d.)

If μ e.s.d. of Hermitian $A \in \mathbb{R}^{p \times p}$, (i.e., $\mu = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(A)}$)

$$m_\mu(z) = \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}$$

Proof:

$$\begin{aligned} m_\mu(z) &= \int \frac{\mu(dt)}{t - z} = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i(A) - z} = \frac{1}{p} \operatorname{tr} (\operatorname{diag}\{\lambda_i(A)\} - zI_p)^{-1} \\ &= \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}. \end{aligned}$$

Fundamental object: the resolvent of A

$$Q_A(z) \equiv (A - zI_p)^{-1}.$$

The Stieltjes transform

Property (Stieltjes transform of Gram matrices)

For $X \in \mathbb{C}^{p \times n}$, and

- ▶ μ e.s.d. of XX^\top
- ▶ $\tilde{\mu}$ e.s.d. of $X^\top X$

Then

$$m_\mu(z) = \frac{n}{p} m_{\tilde{\mu}}(z) - \frac{p-n}{p} \frac{1}{z}.$$

Proof:

$$m_\mu(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i(XX^\top) - z} = \frac{1}{p} \sum_{i=1}^n \frac{1}{\lambda_i(X^\top X) - z} + \frac{1}{p} (p-n) \frac{1}{0-z}.$$

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Resolvent Identity)

For $A, B \in \mathbb{R}^{p \times p}$ invertible,

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

Proof: Simply left-multiply by A and right-multiply by B on both sides.

Corollary

For $t \in \mathbb{C}$, $x \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$, with A and $A + txx^T$ invertible,

$$(A + txx^T)^{-1}x = \frac{A^{-1}x}{1 + tx^T A^{-1}x}.$$

Proof Intuition: Left-multiply by $(A + tcc^T)$ on both sides.

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Rank-one perturbation)

For $A, B \in \mathbb{R}^{p \times p}$ Hermitian nonnegative definite, e.s.d. μ of A , $t > 0$, $x \in \mathbb{R}^p$, $z \in \mathbb{C} \setminus \text{supp}(\mu)$,

$$\left| \frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \right| \leq \frac{1}{p} \frac{\|B\|}{\text{dist}(z, \text{supp}(\mu))}$$

In particular, as $p \rightarrow \infty$, if $\limsup_p \|B\| < \infty$,

$$\frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \rightarrow 0.$$

Proof Intuition: Based on Weyl's interlacing identity (eigenvalues of A and $A + txx^T$ are interlaced).

The Stieltjes transform

Three fundamental lemmas in all proofs.

Lemma (Trace Lemma)

For

- ▶ $x \in \mathbb{R}^p$ with i.i.d. entries with zero mean, unit variance, finite $2k$ order moment,
- ▶ $A \in \mathbb{R}^{p \times p}$ deterministic (or independent of x),

then

$$E \left[\left| \frac{1}{p} x^\top A x - \frac{1}{p} \operatorname{tr} A \right|^k \right] \leq K \frac{\|A\|^p}{p^{k/2}}.$$

In particular, if $\limsup_p \|A\| < \infty$, and x has entries with finite eighth-order moment,

$$\frac{1}{p} x^\top A x - \frac{1}{p} \operatorname{tr} A \xrightarrow{\text{a.s.}} 0$$

(by Markov inequality and Borel Cantelli lemma).

Theorem (Marčenko–Pastur Law [Marčenko,Pastur'67])

$X_p \in \mathbb{R}^{p \times n}$ with i.i.d. zero mean, unit variance entries.

As $p, n \rightarrow \infty$ with $p/n \rightarrow c \in (0, \infty)$, e.s.d. μ_p of $\frac{1}{n} X_p X_p^T$ satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

weakly, where

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

$$f_c(x) = \frac{1}{2\pi c x} \sqrt{(x - (1 - \sqrt{c})^2)((1 + \sqrt{c})^2 - x)}.$$

Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- ▶ With μ_p e.s.d. of $\frac{1}{n}X_pX_p^\top$,

$$m_{\mu_p}(z) = \frac{1}{p} \operatorname{tr} \left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \frac{1}{p} \sum_{i=1}^p \left[\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{ii}.$$

- ▶ Write

$$X_p = \begin{bmatrix} y^\top \\ Y_{p-1} \end{bmatrix} \in \mathbb{R}^{p \times n}$$

so that, for $\Im[z] > 0$,

$$\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \begin{pmatrix} \frac{1}{n} y^\top y - z & \frac{1}{n} y^\top Y_{p-1} \\ \frac{1}{n} Y_{p-1} y & \frac{1}{n} Y_{p-1} Y_{p-1}^\top - z I_{p-1} \end{pmatrix}^{-1}.$$

Proof (continued)

- ▶ From block matrix inverse formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

we have

$$\left[\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} = \frac{1}{-z - z \frac{1}{n} \mathbf{y}^\top \left(\frac{1}{n} Y_{p-1}^\top Y_{p-1} - z I_n \right)^{-1} \mathbf{y}}.$$

- ▶ By **Trace Lemma**, as $p, n \rightarrow \infty$

$$\left[\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{-z - z \frac{1}{n} \text{tr} \left(\frac{1}{n} Y_{p-1}^\top Y_{p-1} - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- By **Rank-1 Perturbation Lemma** ($X_p^\top X_p = Y_{p-1}^\top Y_{p-1} + yy^\top$), as $p, n \rightarrow \infty$

$$\left[\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{-z - z \frac{1}{n} \operatorname{tr} \left(\frac{1}{n} X_p^\top X_p - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- Since $\frac{1}{n} \operatorname{tr} \left(\frac{1}{n} X_p^\top X_p - z I_n \right)^{-1} = \frac{1}{n} \operatorname{tr} \left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} - \frac{n-p}{n} \frac{1}{z}$,

$$\left[\left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{1 - \frac{p}{n} - z - z \frac{1}{n} \operatorname{tr} \left(\frac{1}{n} X_p X_p^\top - z I_p \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- Repeating for **entries** $(2, 2), \dots, (p, p)$, and averaging, we get (for $\Im[z] > 0$)

$$m_{\mu_p}(z) - \frac{1}{1 - \frac{p}{n} - z - z \frac{p}{n} m_{\mu_p}(z)} \xrightarrow{\text{a.s.}} 0.$$

Proof of the Marčenko–Pastur law

Proof (continued)

- Then $m_{\mu_p}(z) \xrightarrow{\text{a.s.}} m(z)$ solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

i.e., (with branch of $\sqrt{f(z)}$ such that $m(z) \rightarrow 0$ as $|z| \rightarrow \infty$)

$$m(z) = \frac{1-c}{2cz} - \frac{1}{2c} + \frac{\sqrt{(z - (1 + \sqrt{c})^2)(z - (1 - \sqrt{c})^2)}}{2cz}.$$

- Finally, by **inverse Stieltjes Transform**, for $x > 0$,

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m(x + i\varepsilon)] = \frac{\sqrt{((1 + \sqrt{c})^2 - x)(x - (1 - \sqrt{c})^2)}}{2\pi cx} \mathbf{1}_{\{x \in [(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]\}}.$$

And for $x = 0$,

$$\lim_{\varepsilon \downarrow 0} i\varepsilon \Im[m(i\varepsilon)] = (1 - c^{-1}) \mathbf{1}_{\{c > 1\}}.$$

Sample Covariance Matrices

Theorem (Sample Covariance Matrix Model [Silverstein, Bai'95])

Let $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$, with

- ▶ $C_p \in \mathbb{C}^{p \times p}$ nonnegative definite with e.s.d. $\nu_p \rightarrow \nu$ weakly,
- ▶ $X_p \in \mathbb{C}^{p \times n}$ has i.i.d. entries of zero mean and unit variance.

As $p, n \rightarrow \infty$, $p/n \rightarrow c \in (0, \infty)$, $\tilde{\mu}_p$ e.s.d. of $\frac{1}{n} Y_p^T Y_p \in \mathbb{R}^{n \times n}$ satisfies

$$\tilde{\mu}_p \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with $m_{\tilde{\mu}}(z)$, $\Im[z] > 0$, unique solution with $\Im[m_{\tilde{\mu}}(z)] > 0$ of

$$m_{\tilde{\mu}}(z) = \left(-z + c \int \frac{t}{1 + tm_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

Moreover, $\tilde{\mu}$ is continuous on \mathbb{R}^+ and real analytic wherever positive.

Immediate corollary: For μ_p e.s.d. of $\frac{1}{n} Y_p Y_p^T = \frac{1}{n} \sum_{i=1}^n C_p^{\frac{1}{2}} x_i x_i^T C_p^{\frac{1}{2}}$,

$$\mu_p \xrightarrow{\text{a.s.}} \mu$$

weakly, with $\tilde{\mu} = c\mu + (1 - c)\delta_0$.

Sample Covariance Matrices

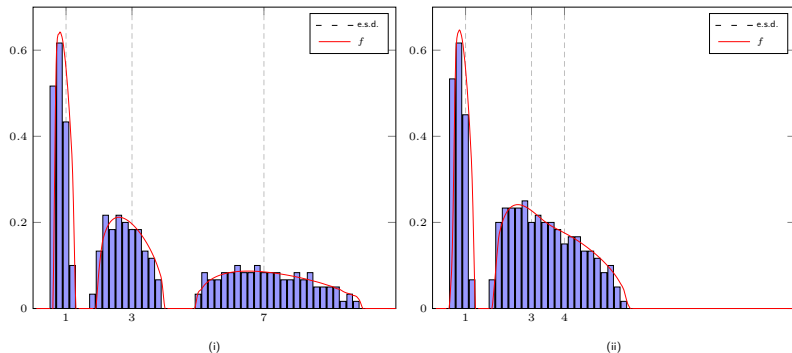


Figure: Histogram of the eigenvalues of $\frac{1}{n} Y_p Y_p^T$, $n = 3000$, $p = 300$, with C_p diagonal with evenly weighted masses in (i) 1, 3, 7, (ii) 1, 3, 4.

Further Models and Deterministic Equivalents

Sometimes, μ_p does not converge!

- ▶ if ν_p does not converge
- ▶ if p/n does not converge
- ▶ if eigenvectors of deterministic matrices play a role!

Deterministic equivalents: sequence $\bar{\mu}_p$ of **deterministic** measures, with

$$\mu_p - \bar{\mu}_p \xrightarrow{\text{a.s.}} 0$$

or equivalently, **deterministic** sequence of m_p with

$$m_{\mu_p} - m_p \xrightarrow{\text{a.s.}} 0.$$

Theorem (Doubly-correlated i.i.d. matrices)

Let $B_p = C_p^{\frac{1}{2}} X_p T_p X_p^T C_p^{\frac{1}{2}}$, with e.s.d. μ_p , $X_p \in \mathbb{R}^{p \times n}$ with i.i.d. entries of zero mean, variance $1/n$, C_p Hermitian nonnegative definite, T_p diagonal nonnegative, $\limsup_p \max(\|C_p\|, \|T_p\|) < \infty$. Denote $c = p/n$.

Then, as $p, n \rightarrow \infty$ with bounded ratio c , for $z \in \mathbb{C} \setminus \mathbb{R}^-$,

$$m_{\mu_p}(z) - m_p(z) \xrightarrow{\text{a.s.}} 0, \quad m_p(z) = \frac{1}{p} \text{tr} (-zI_p + \bar{e}_p(z)C_p)^{-1}$$

with $\bar{e}(z)$ unique solution in $\{z \in \mathbb{C}^+, \bar{e}_p(z) \in \mathbb{C}^+\}$ or $\{z \in \mathbb{R}^-, \bar{e}_p(z) \in \mathbb{R}^+\}$ of

$$e_p(z) = \frac{1}{p} \text{tr} C_p (-zI_p + \bar{e}_p(z)C_p)^{-1}$$

$$\bar{e}_p(z) = \frac{1}{n} \text{tr} T_p (I_n + ce_p(z)T_p)^{-1}.$$

Side note on other models.

Similar results for multiple matrix models:

- ▶ **Information-plus-noise:** $Y_p = A_p + X_p$, A_p deterministic
- ▶ **Variance profile:** $Y_p = P_p \odot X_p$ (entry-wise product)
- ▶ **Per-column covariance:** $Y_p = [y_1, \dots, y_n]$, $y_i = C_{p,i}^{\frac{1}{2}} x_i$
- ▶ etc.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support [Silverstein, Bai'98])

Let $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$, with

- ▶ $C_p \in \mathbb{R}^{p \times p}$ nonnegative definite with e.s.d. $\nu_p \rightarrow \nu$ weakly,
- ▶ $X_p \in \mathbb{R}^{p \times n}$ has i.i.d. entries of zero mean and unit variance,
- ▶ $E[|X_p|_{ij}^4] < \infty$,
- ▶ $\max_i \text{dist}(\lambda_i(C_p), \text{supp}(\nu)) \rightarrow 0$.

Let $\tilde{\mu}$ be the limiting e.s.d. of $\frac{1}{n} Y_p^T Y_p$ as before. Let $[a, b] \subset \mathbb{R}^T \setminus \text{supp}(\tilde{\nu})$. Then,

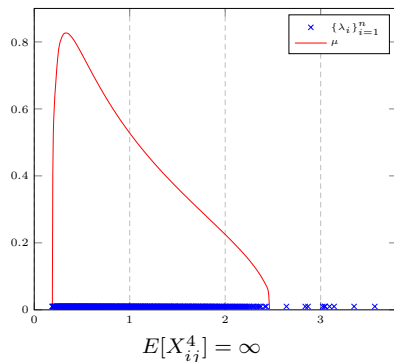
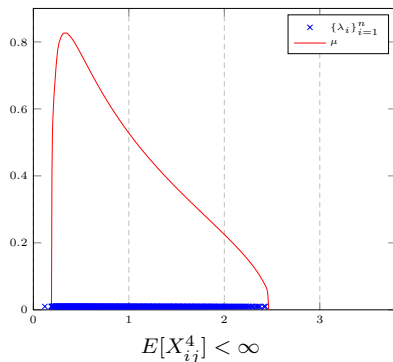
$$\left\{ \lambda_i \left(\frac{1}{n} Y_p^T Y_p \right) \right\}_{i=1}^n \cap [a, b] = \emptyset$$

for all large n , almost surely.

In practice: This means that eigenvalues of $\frac{1}{n} Y_p^T Y_p$ cannot be bound at macroscopic distance from the bulk, for p, n large.

Breaking the rules. If we break

- ▶ **Rule 1:** Infinitely many eigenvalues may wander away from $\text{supp}(\mu)$.



Spiked Models

If we break:

- ▶ **Rule 2:** C_p may create isolated eigenvalues in $\frac{1}{n} Y_p Y_p^T$, called spikes.

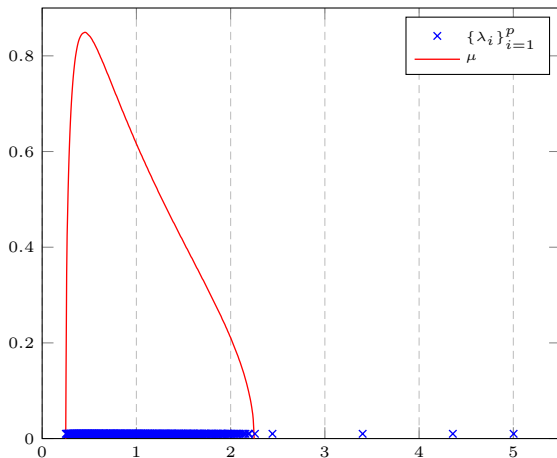


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^T$, $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$, $p = 500$, $n = 2000$.

Spiked Models: The phase transition phenomenon

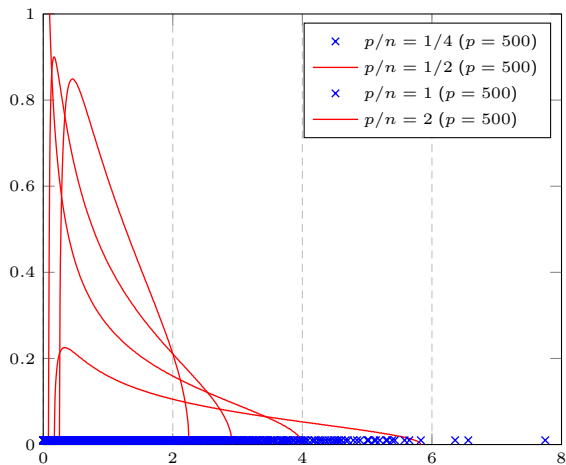


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^T$, $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$.

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, $E[|X_p|_{ij}^4] < \infty$.
- ▶ $C_p = I_p + P$, $P = U\Omega U^T$, where, for K fixed,

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

Then, as $p, n \rightarrow \infty$, $p/n \rightarrow c \in (0, \infty)$, denoting $\lambda_i = \lambda_i(\frac{1}{n} Y_p Y_p^T)$,

- ▶ if $\omega_m > \sqrt{c}$,

$$\lambda_m \xrightarrow{\text{a.s.}} 1 + \omega_m + c \frac{1 + \omega_m}{\omega_m} > (1 + \sqrt{c})^2$$

- ▶ if $\omega_m \in (0, \sqrt{c}]$,

$$\lambda_m \xrightarrow{\text{a.s.}} (1 + \sqrt{c})^2$$

Spiked Models

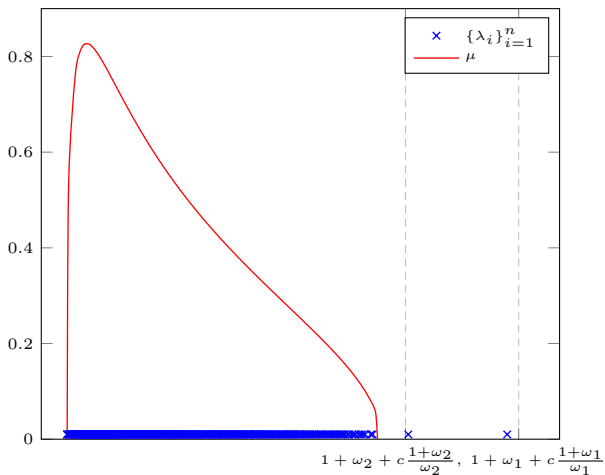


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^T$, $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-2}, 2, 3)$, $p = 500$, $n = 1500$.

Proof

- ▶ **Two ingredients:** Algebraic calculus + trace lemma
- ▶ **Find eigenvalues away from eigenvalues of $\frac{1}{n}X_pX_p^\top$:**

$$\begin{aligned}0 &= \det\left(\frac{1}{n}Y_pY_p^\top - \lambda I_p\right), \quad Y_p = C_p^{\frac{1}{2}}X_p \\&= \det(C_p) \det\left(\frac{1}{n}X_pX_p^\top - \lambda C_p^{-1}\right) \\&= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p + \lambda(I_p - C_p^{-1})\right) \\&= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda(I_p - C_p^{-1})\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).\end{aligned}$$

- ▶ **Use low rank property:** ($C_p = I_p + P = I_p + U\Omega U^\top$)

$$I_p - C_p^{-1} = I_p - (I_p + U\Omega U^\top)^{-1} = U(I_K + \Omega^{-1})^{-1}U^\top, \quad \Omega \in \mathbb{C}^{K \times K}.$$

Hence

$$0 = \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda U(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).$$

Proof (2)

- ▶ **Sylvester's identity** ($\det(I + AB) = \det(I + BA)$),

$$0 = \det\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right) \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right)^{-1}U\right)$$

- ▶ **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_p X_p^\top - \lambda I_p)$ has no zero beyond $(1 + \sqrt{c})^2$ for all large n a.s.
- ▶ **Extension of Trace Lemma**: for each $z \in \mathbb{C} \setminus \text{supp}(\mu)$,

$$U^\top\left(\frac{1}{n}X_p X_p^\top - z I_p\right)^{-1}U \xrightarrow{\text{a.s.}} m_\mu(z)I_K.$$

(X_p being “almost-unitarily invariant”, U made of “i.i.d.-like” random vectors)

- ▶ As a result, for all large n a.s.,

$$\begin{aligned} 0 &= \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right)^{-1}U\right) \\ &\simeq \prod_{k=1}^K \left(1 + \frac{\lambda}{1 + \omega_k^{-1}} m_\mu(\lambda)\right) = \prod_{k=1}^K \left(1 + \frac{\omega_k}{1 + \omega_k} \lambda m_\mu(\lambda)\right) \end{aligned}$$

Proof (3)

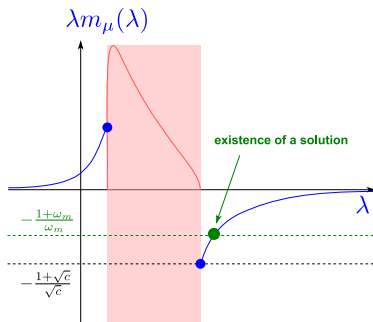
- ▶ **Limiting solutions:** zeros of

$$\lambda m_\mu(\lambda) = -\frac{1 + \omega_m}{\omega_m}.$$

- ▶ Marčenko–Pastur law properties ($m_\mu(z) = (1 - c - z - czm_\mu(z))^{-1}$):

- ▶ $\lambda \mapsto \lambda m_\mu(\lambda) = \int \frac{\lambda}{t-\lambda} \mu(dt)$ maps $((1 + \sqrt{c})^2, \infty)$ onto $(-\frac{1+\sqrt{c}}{\sqrt{c}}, 0^-)$
- ▶ Solution only when $\omega_m > \sqrt{c}$:

$$\lambda = 1 + \omega_m + c \frac{1 + \omega_m}{\omega_m}.$$



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, *finite fourth order moment entries*
- ▶ $C_p = I_p + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^\top$, $\omega_1 > \dots > \omega_M > 0$.

Then, as $p, n \rightarrow \infty$, $p/n \rightarrow c \in (0, \infty)$, for $a, b \in \mathbb{R}^p$ deterministic and \hat{u}_i eigenvector of $\lambda_i(\frac{1}{n} Y_p Y_p^\top)$,

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

In particular,

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

Proof: Based on Cauchy integral + similar ingredients as eigenvalue proof

$$a^\top \hat{u}_i \hat{u}_i^\top b = \frac{1}{2\pi i} \oint_{C_i} a^\top \left(\frac{1}{n} Y_p Y_p^\top - z I_p \right)^{-1} b dz$$

for C_m contour circling around λ_i only.

Spiked Models

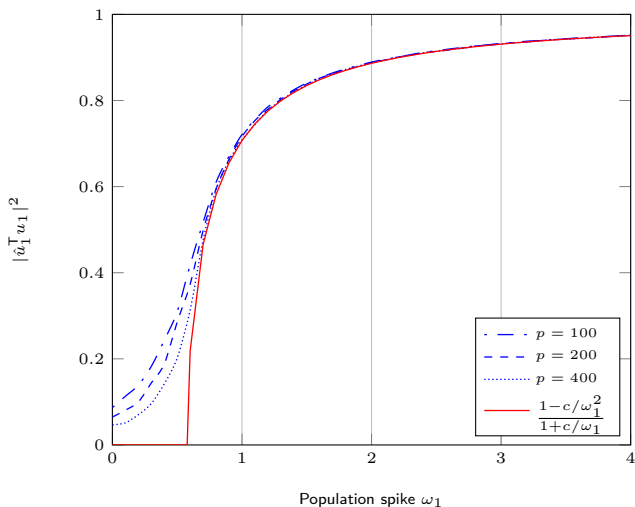


Figure: Simulated versus limiting $|\hat{u}_1^T u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^T$, $p/n = 1/3$, varying ω_1 .

Theorem (Fluctuations of Eigenvalues [Baik, BenArous, P  ch  '05])

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

- ▶ X_p with i.i.d. *real or complex Gaussian* zero mean, unit variance entries,
- ▶ $C_p = I_p + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^T$, $\omega_1 > \dots > \omega_K > 0$ ($K \geq 0$).

Then, as $p, n \rightarrow \infty$, $p/n \rightarrow c < 1$,

- ▶ If $\omega_1 < \sqrt{c}$ (or $K = 0$),

$$p^{\frac{2}{3}} \frac{\lambda_1 - (1 + \sqrt{c})^2}{(1 + \sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T, \text{ (real or complex Tracy–Widom law)}$$

- ▶ If $\omega_1 > \sqrt{c}$,

$$\left(\frac{(1 + \omega_1)^2}{c} - \frac{(1 + \omega_1)^2}{\omega_1^2} \right)^{\frac{1}{2}} p^{\frac{1}{2}} \left[\lambda_1 - \left(1 + \omega_1 + c \frac{1 + \omega_1}{\omega_1} \right) \right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Tracy–Widom Theorem

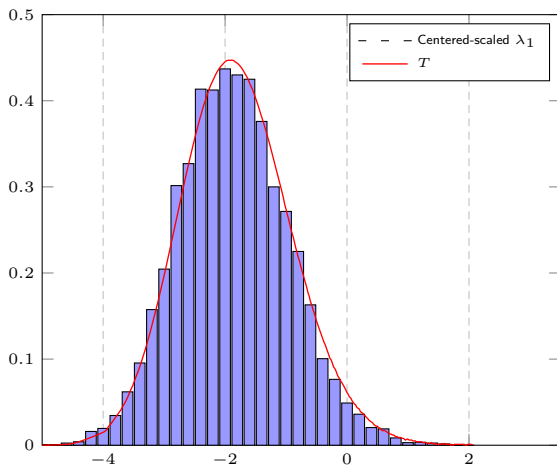


Figure: Distribution of $p^{\frac{2}{3}} c^{-\frac{1}{2}} (1 + \sqrt{c})^{-\frac{4}{3}} \left[\lambda_1 \left(\frac{1}{n} X_p X_p^T \right) - (1 + \sqrt{c})^2 \right]$ versus real Tracy–Widom (T), $p = 500$, $n = 1500$.

Similar results for multiple matrix models:

- ▶ $Y_p = \frac{1}{n}XX^T + P$, P deterministic and low rank
- ▶ $Y_p = \frac{1}{n}X^T(I + P)X$
- ▶ $Y_p = \frac{1}{n}(X + P)^T(X + P)$
- ▶ $Y_p = \frac{1}{n}TX^T(I + P)XT$
- ▶ etc.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

The Semi-circle law

Theorem

Let $X_n \in \mathbb{R}^{n \times n}$ Hermitian with e.s.d. μ_n such that $\frac{1}{\sqrt{n}}[X_n]_{i>j}$ are i.i.d. with zero mean and unit variance. Then, as $n \rightarrow \infty$,

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

with $\mu(dt) = \frac{1}{2\pi} \sqrt{(4-t^2)_+} dt$. In particular, m_μ satisfies

$$m_\mu(z) = \frac{1}{-z - m_\mu(z)}.$$

The Semi-circle law

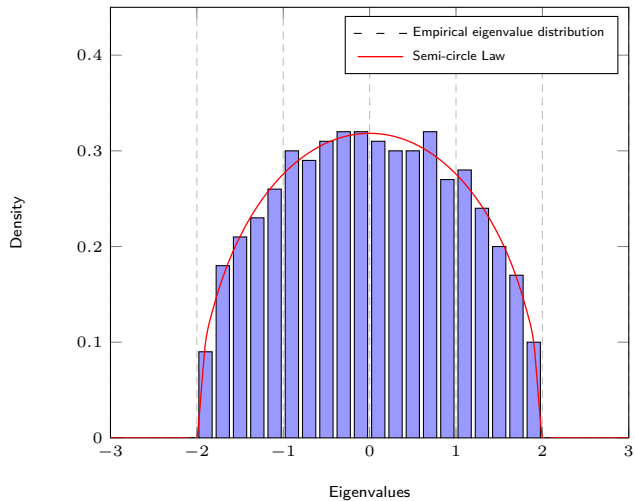


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $n = 500$

Theorem

Let $X_n \in \mathbb{C}^{n \times n}$ with e.s.d. μ_n be such that $\frac{1}{\sqrt{n}}[X_n]_{ij}$ are i.i.d. entries with zero mean and unit variance. Then, as $n \rightarrow \infty$,

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

with μ a complex-supported measure with $\mu(dz) = \frac{1}{2\pi} \delta_{|z| \leq 1} dz$.

The Circular law

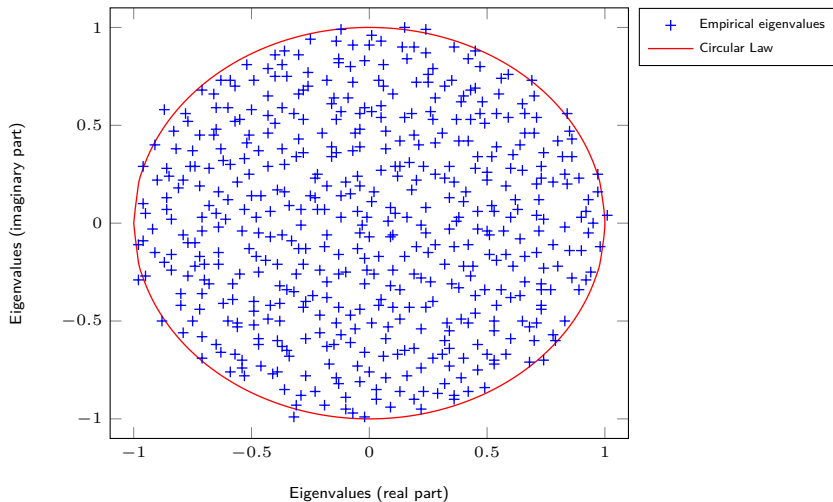







Figure: Eigenvalues of X_n with i.i.d. standard Gaussian entries, for $n = 500$.

From most accessible to least:

-  Couillet, R., & Debbah, M. (2011). Random matrix methods for wireless communications. Cambridge University Press.
-  Tao, T. (2012). Topics in random matrix theory (Vol. 132). Providence, RI: American Mathematical Society.
-  Bai, Z., & Silverstein, J. W. (2010). Spectral analysis of large dimensional random matrices (Vol. 20). New York: Springer.
-  Pastur, L. A., Shcherbina, M., & Shcherbina, M. (2011). Eigenvalue distribution of large random matrices (Vol. 171). Providence, RI: American Mathematical Society.
-  Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). An introduction to random matrices (Vol. 118). Cambridge university press.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

From classical applications...

Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels $H \in \mathbb{C}^{p \times n}$, optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ($p \sim n$), optimal beamforming (MVDR), detection filters (ANMF), etc.
- ▶ **Statistical finance:** portfolio optimization (Markowitz, GMVP) for large portfolios and short time windows.
- ▶ **Brain signal processing:** EEG covariance estimation on short windows.

Any application where $p \sim n$ “rather large”

(convergence speed in up to $O(n)$ and not $O(\sqrt{n})$ as usual!)

BUT mostly linear settings...

... to machine learning!

Specificities in statistical and machine learning:

- ▶ **Matrix of non-linear entries:** kernel matrices $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$, activation functions in neural nets $x_{l+1} = \sigma(Wx_l)$, non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

CENTRAL ISSUE: Given that basic sample covariance matrices are not consistent for large n, p , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

TUTORIAL: first answers to **understand, improve, and change paradigm**.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

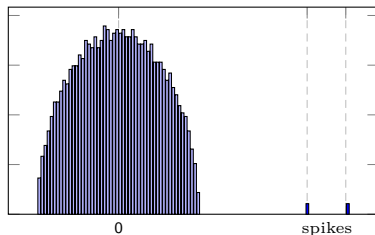
Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

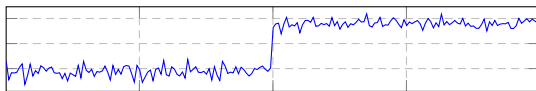
Reminder on Spectral Clustering Methods

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



↓ Eigenvectors ↓
(in practice, **shuffled**)

Eigenv. 1

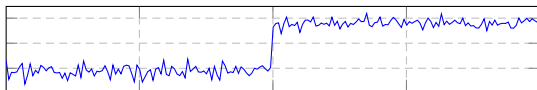


Eigenv. 2



Reminder on Spectral Clustering Methods

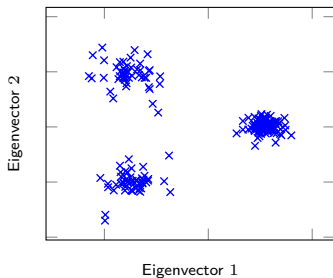
Eigenv. 1



Eigenv. 2



↓ ℓ -dimensional representation ↓
(shuffling no longer matters)



↓
EM or k-means clustering.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Kernel Spectral Clustering

Problem Statement

- ▶ Dataset $x_1, \dots, x_n \in \mathbb{R}^p$
- ▶ Objective: “cluster” data in k similarity classes $\mathcal{C}_1, \dots, \mathcal{C}_k$.
- ▶ **Kernel spectral clustering** based on kernel matrix

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

- ▶ Usually, $\kappa(x, y) = f(x^\top y)$ 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)

$$K = \begin{pmatrix} \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \gg 1 & \ll 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \gg 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \ll 1 & \gg 1 \\ \hline \end{array} \\ \hline \end{pmatrix} \begin{array}{l} \updownarrow \mathcal{C}_1 \\ \updownarrow \mathcal{C}_2 \\ \updownarrow \mathcal{C}_3 \end{array}$$

- ▶ K essentially low rank with class structure in eigenvectors.
- ▶ Ng–Weiss–Jordan key remark: $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}(D^{\frac{1}{2}}j_a) \simeq D^{\frac{1}{2}}j_a$ (j_a canonical vector of \mathcal{C}_a)

Kernel Spectral Clustering

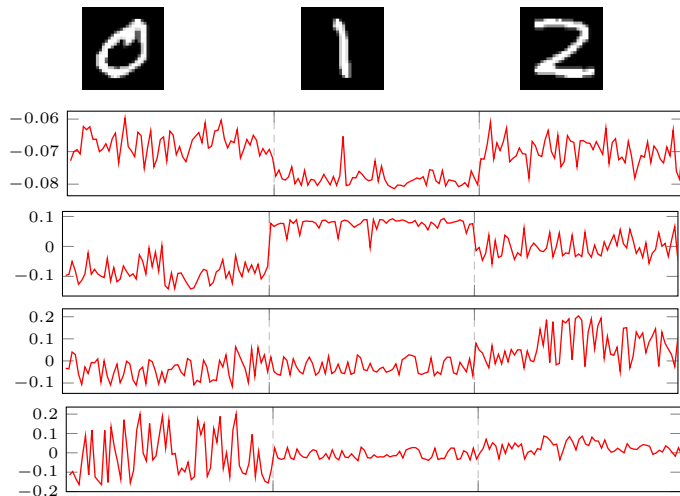


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

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

Model and Assumptions

Gaussian mixture model:

- ▶ $x_1, \dots, x_n \in \mathbb{R}^p$,
- ▶ k classes $\mathcal{C}_1, \dots, \mathcal{C}_k$,
- ▶ $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$,
- ▶ $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$.

Assumption (Growth Rate)

As $n \rightarrow \infty$,

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

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(p)$$

For 2 classes, this is

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

Remark: [Neyman–Pearson optimality]

- ▶ $x \sim \mathcal{N}(\pm\mu, I_p)$ (known μ) decidable iff $\|\mu\| \geq O(1)$.
- ▶ $x \sim \mathcal{N}(0, (1 \pm \varepsilon)I_p)$ (known ε) decidable iff $\|\varepsilon\| \geq O(p^{-\frac{1}{2}})$.

Kernel Matrix:

- ▶ Kernel matrix of interest:

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

for some sufficiently smooth nonnegative f ($f(\frac{1}{p}x_i^\top x_j)$ simpler).

- ▶ We study the normalized Laplacian:

$$L = nD^{-\frac{1}{2}} \left(K - \frac{dd^\top}{d^\top 1_n} \right) D^{-\frac{1}{2}}$$

with $d = K1_n$, $D = \text{diag}(d)$.

(more stable both theoretically and in practice)

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

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

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

⇒ Suggests that (up to diagonal) $K \simeq f(\tau)1_n 1_n^\top$!

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

$$K = \underbrace{f(\tau)1_n 1_n^\top}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}K_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{K_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

Clearly not the (small dimension) expected behavior.

Random Matrix Equivalent

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

As $n, p \rightarrow \infty$, $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$, where

$$L = nD^{-\frac{1}{2}} \left(K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left(\frac{1}{p} \|x_i - x_j\|^2 \right)$$
$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \left[\frac{1}{p} PW^T WP + \frac{1}{p} JBJ^T + * \right]$$

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

$$J = [j_1, \dots, j_k], \quad j_a^T = (0 \dots 0, 1_{n_a}, 0, \dots, 0)$$
$$B = M^T M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) tt^T - \frac{f''(\tau)}{f'(\tau)} T + *.$$

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

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

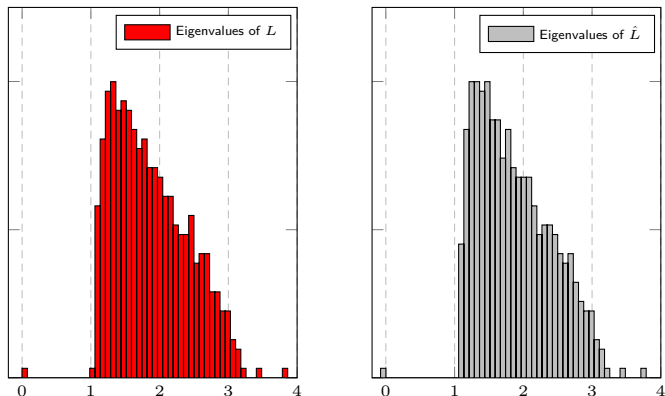


Figure: Eigenvalues of L and \hat{L} , $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\delta_{aj}$, $C_a = (1 + 2(a - 1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Theoretical Findings versus MNIST

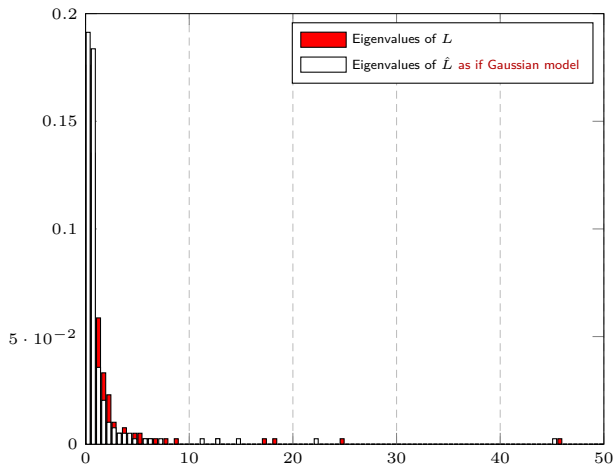


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) \hat{L} (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

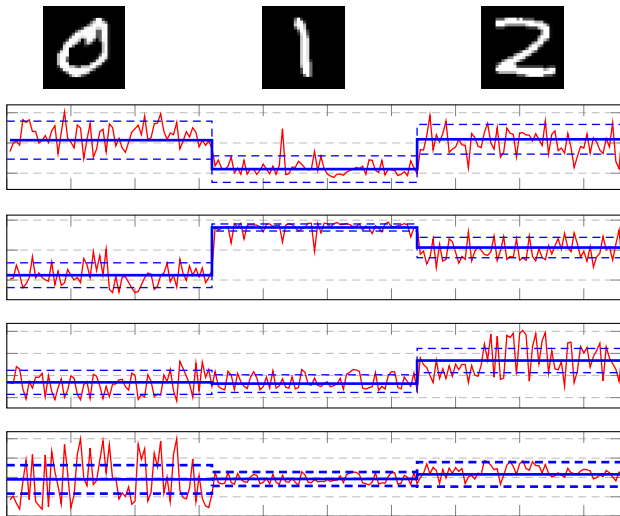


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data (red) and theoretical findings (blue).

Theoretical Findings versus MNIST

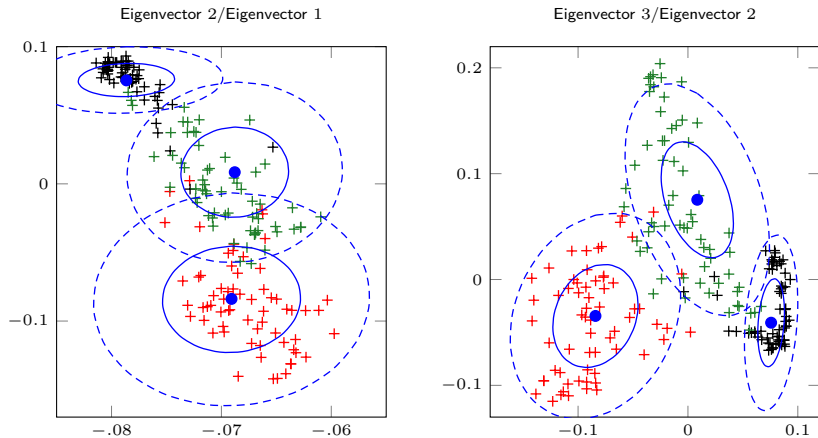


Figure: 2D representation of eigenvectors of L , for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

The surprising $f'(\tau) = 0$ case

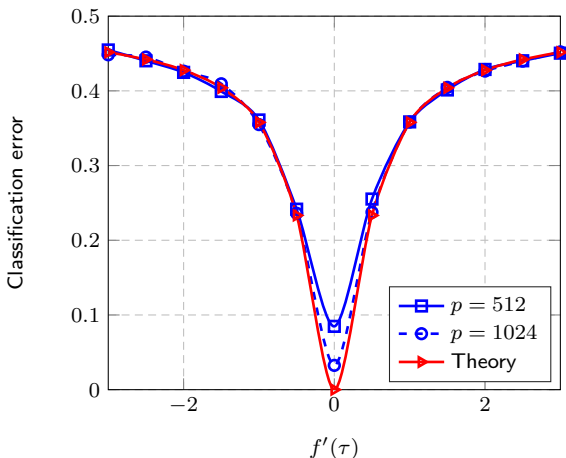


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Position of the Problem

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

Method:

- ▶ Still assume x_1, \dots, x_n belong to k classes $\mathcal{C}_1, \dots, \mathcal{C}_k$.
- ▶ Zero-mean Gaussian model for the data: for $x_i \in \mathcal{C}_k$,

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

- ▶ Performance of $L = nD^{-\frac{1}{2}} \left(K - \frac{1_n 1_n^\top}{1_n^\top D 1_n} \right) D^{-\frac{1}{2}}$, 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 $n, p \rightarrow \infty$.

(alternatively, we can ask $\frac{1}{p} \text{tr} C_i = 1$ for all $1 \leq i \leq k$)

Model and Reminders

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

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

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

Theorem (Corollary of Previous Section)

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

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

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

$$T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$$

has sufficiently large eigenvalues (here $M = 0$, $t = 0$).

The case $f'(2) = 0$

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

1. $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3. $\frac{1}{p} \text{tr} C_a = 1$ and $\text{tr} C_a^\circ C_b^\circ = O(p)$ ~~$\text{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$~~ , 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 $\|E\| \rightarrow 0$, **detectability** iff $\frac{1}{p} \text{tr}(C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$.

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

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

$$\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[L - \frac{f(0) - f(2)}{f(2)} P \right], \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

Then, under Assumptions 2b,

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

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

The case $f'(2) = 0$

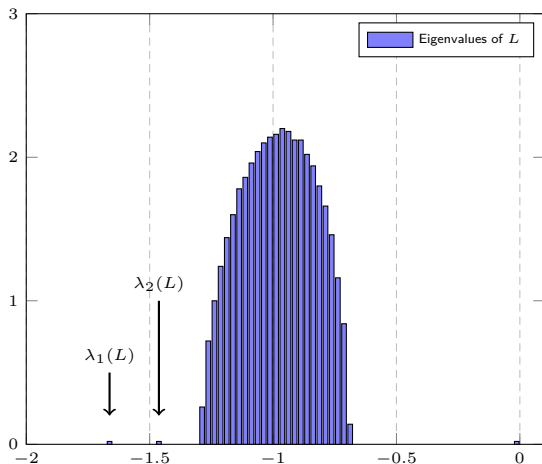
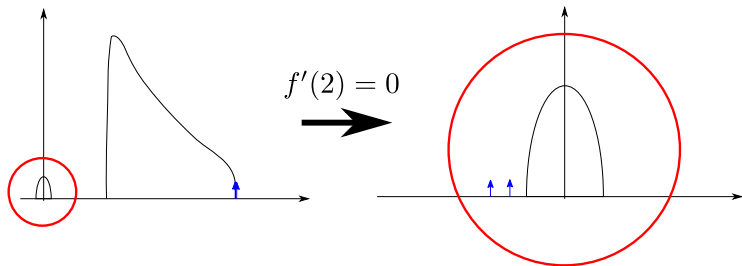


Figure: Eigenvalues of L , $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$,
 $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^T$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0, 1)$ entries, $f(t) = \exp(-(t - 2)^2)$.

\Rightarrow 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 Φ
- ▶ study the isolated eigenvalues of \mathcal{L} (and the phase transition)
- ▶ retrieve information from the eigenvectors.

Theorem (Semi-circle law for Φ)

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

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

with μ the semi-circle distribution

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2} \sqrt{(4c_0 \omega^2 - t^2)_+} dt, \quad \omega = \lim_{p \rightarrow \infty} \sqrt{2} \frac{1}{p} \text{tr}(C^\circ)^2.$$

The case $f'(2) = 0$

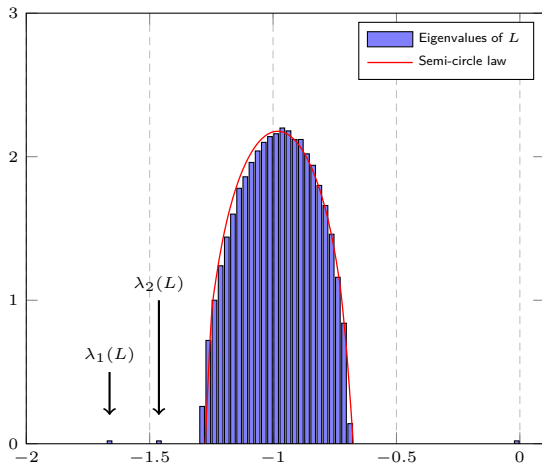


Figure: Eigenvalues of L , $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-5/4} W_i W_i^T$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0, 1)$ entries, $f(t) = \exp(-(t - 2)^2)$.

The case $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p \rightarrow \infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k.$$

Theorem (Isolated Eigenvalues)

Let $\nu_1 \geq \dots \geq \nu_k$ eigenvalues of \mathcal{T} . Then, if $\sqrt{c_0} |\nu_i| > \omega$, \mathcal{L} has an isolated eigenvalue λ_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 (λ_i, u_i) of \mathcal{L} corresponding to (ν_i, v_i) of \mathcal{T} , write

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

with $j_a = [0_{n_1}^\top, \dots, 1_{n_a}^\top, \dots, 0_{n_k}^\top]^\top$, $(w_i^a)^\top j_a = 0$, $\text{supp}(w_i^a) = \text{supp}(j_a)$, $\|w_i^a\| = 1$.
Then, under Assumptions 1–2b,

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

and the fluctuations of u_i, u_j , $i \neq j$, are asymptotically uncorrelated.

The case $f'(2) = 0$

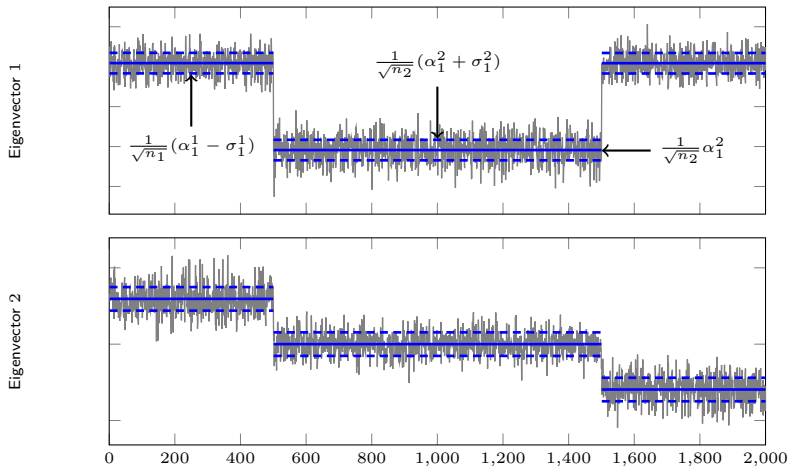


Figure: Leading two eigenvectors of \mathcal{L} (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

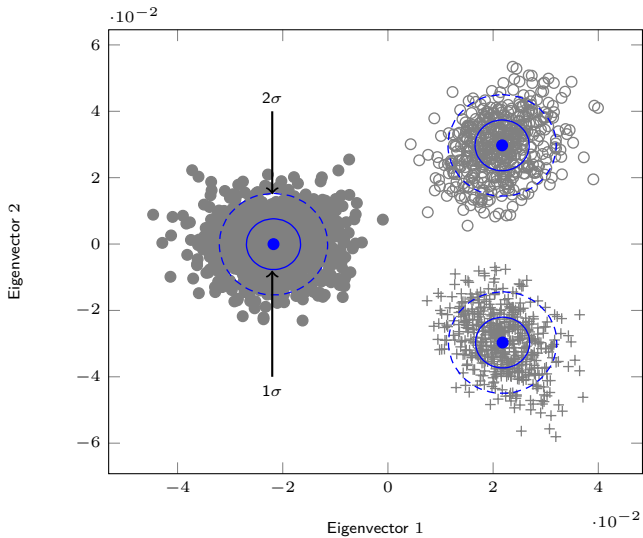


Figure: Leading two eigenvectors of \mathcal{L} (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

Application: Multiple-source Subspace Clustering

Setting.

- ▶ p dimensional vector observations.
- ▶ n data sources.
- ▶ $E[x_i] = 0$, $E[x_i x_i^T] = C_a$.
- ▶ T independent observations $x_i^{(1)}, \dots, x_i^{(T)}$ for source i .

Objective. Cluster sources based on spanned subspace.

Applications examples. Massive MIMO scheduling / EEG classification / etc.

Algorithm.

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

Application Example: Massive MIMO UE Clustering

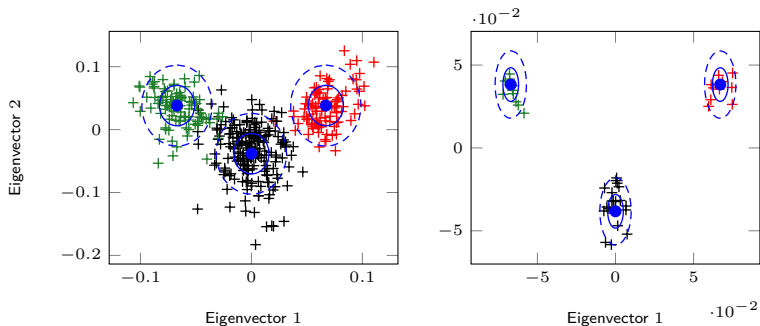


Figure: Massive MIMO application: Leading two eigenvectors before (left figure) and after (right figure) T -averaging. Setting: $p = 400$, $n = 40$, $T = 10$, $k = 3$, $c_1 = c_3 = 1/4$, $c_2 = 1/2$, angular spread model with angles $-\pi/30 \pm \pi/20$, $0 \pm \pi/20$, and $\pi/30 \pm \pi/20$. Kernel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

Application Example: Massive MIMO UE Clustering

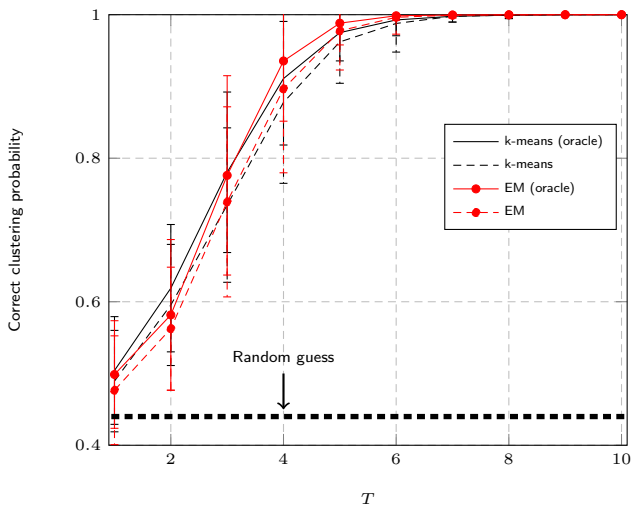


Figure: Overlap for different T , using the k-means or EM starting from actual centroid solutions (oracle) or randomly.

Application Example: Massive MIMO UE Clustering

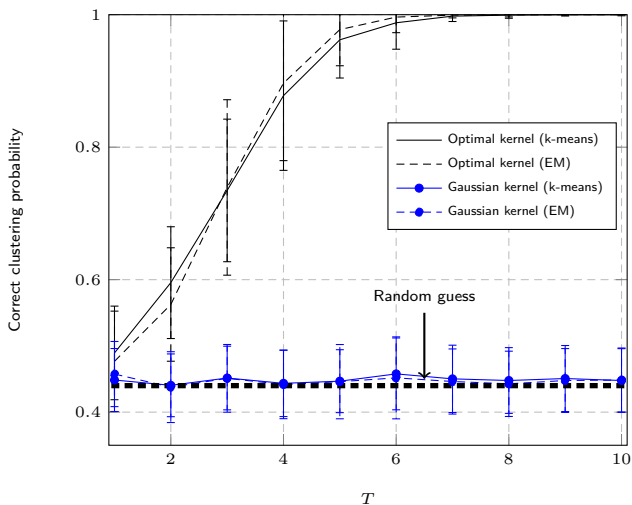


Figure: Overlap for optimal kernel $f(t)$ (here $f(t) = \exp(-(t - 2)^2)$) and Gaussian kernel $f(t) = \exp(-t^2)$, for different T , using the k-means or EM.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

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 $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\text{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
 - ▶ suboptimal in $\frac{1}{p}\text{tr} C_a^\circ C_b^\circ = O(1)$→ **Model type:** Marčenko–Pastur + spikes.

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

Jointly optimal solution:

- ▶ evenly weighing Marčenko–Pastur and semi-circle laws
- ▶ the “ α - β ” kernel:

$$f'(\tau) = \frac{\alpha}{\sqrt{p}}, \quad \frac{1}{2}f''(\tau) = \beta.$$

New assumption setting

- ▶ We consider now an **improved growth rate setting**.

Assumption (Optimal Growth Rate)

As $n \rightarrow \infty$,

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

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(\sqrt{p}).$$

Kernel:

- ▶ For technical simplicity, we consider

$$\tilde{K} = PKP = P \left\{ f \left(\frac{1}{p} (x^\circ)^\top (x_j^\circ) \right) \right\}_{i,j=1}^n P, \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

i.e., τ replaced by 0.

Main Results

Theorem

As $n \rightarrow \infty$,

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

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

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

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

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

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

Role of α, β :

- ▶ Weighs Marčenko–Pastur versus semi-circle parts.

Theorem (Eigenvalues Bulk)

As $p \rightarrow \infty$,

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

with ν having Stieltjes transform $m(z)$ solution of

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

where $\omega = \lim_{p \rightarrow \infty} \frac{1}{p} \text{tr}(C^\circ)^2$.

Limiting eigenvalue distribution

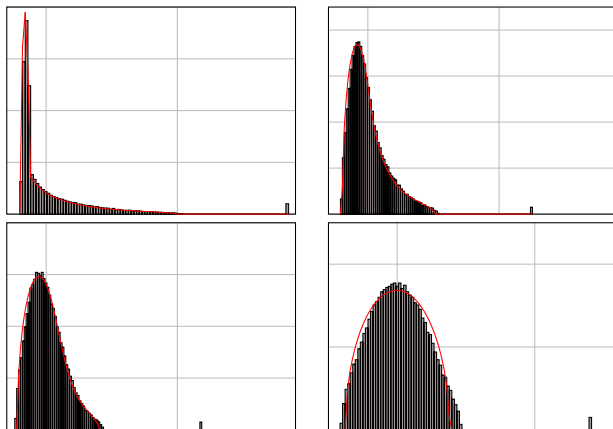


Figure: Eigenvalues of K (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\mu_i = 3\delta_i$, $f(x) = \frac{1}{2}\beta \left(x + \frac{1}{\sqrt{p}} \frac{\alpha}{\beta}\right)^2$. **(Top left):** $\alpha = 8, \beta = 1$, **(Top right):** $\alpha = 4, \beta = 3$,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

Asymptotic performances: MNIST

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

DATASETS	$\ \mu_1^\circ - \mu_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (DIGITS 1, 7)	613	1990	71.1
MNIST (DIGITS 3, 6)	441	1119	39.9
MNIST (DIGITS 3, 8)	212	652	23.5

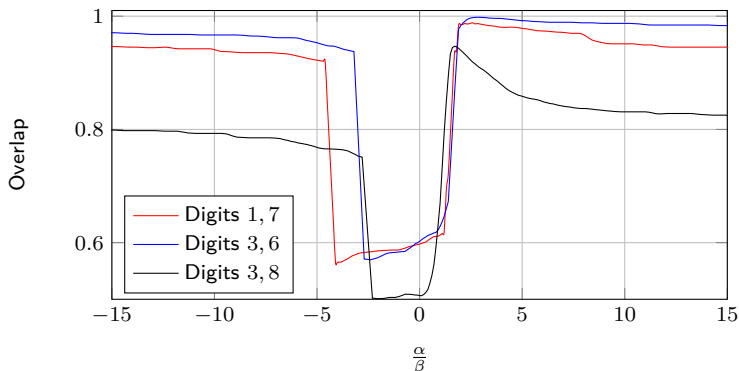


Figure: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$.

Asymptotic performances: EEG data

- ▶ EEG data are “variance-dominant”

DATASETS	$\ \boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (SETS A, E)	2.4	10.9	1.1

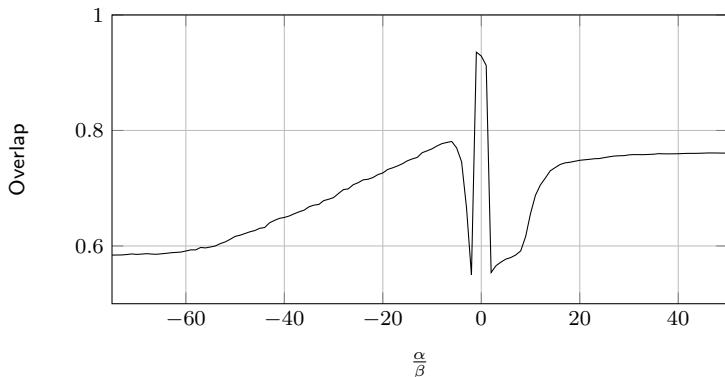


Figure: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Laplacian Regularization

Context: Similar to clustering:

- ▶ Classify $x_1, \dots, x_n \in \mathbb{R}^p$ in K classes, with $n_{[l]}$ labelled ($n_{[l]k}$ in class \mathcal{C}_k) and $n_{[u]}$ unlabelled data ($n_{[u]k}$ in class \mathcal{C}_k).
- ▶ Problem statement: give scores F_{ia} ($d_i = [K1_n]_i$)

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

such that $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$, 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,

$$F_{[u]} = \left(L_{[uu]}^{(\alpha)} \right)^{-1} L_{[ul]}^{(\alpha)} F_{[l]}$$

where

$$L^{(\alpha)} = I - D^{-1-\alpha} K D^\alpha = \begin{bmatrix} L_{[ll]}^{(\alpha)} & L_{[lu]}^{(\alpha)} \\ L_{[ul]}^{(\alpha)} & L_{[uu]}^{(\alpha)} \end{bmatrix}$$

with $D = \operatorname{diag} \{K1_n\}$.

- ▶ Three common choices of α :
 - ▶ $\alpha = 0$: Standard Laplacian Regularization
 - ▶ $\alpha = -1/2$: Symmetric Normalized Laplacian Regularization
 - ▶ $\alpha = -1$: Random Walk Normalized Laplacian Regularization

The finite-dimensional intuition: What we expect

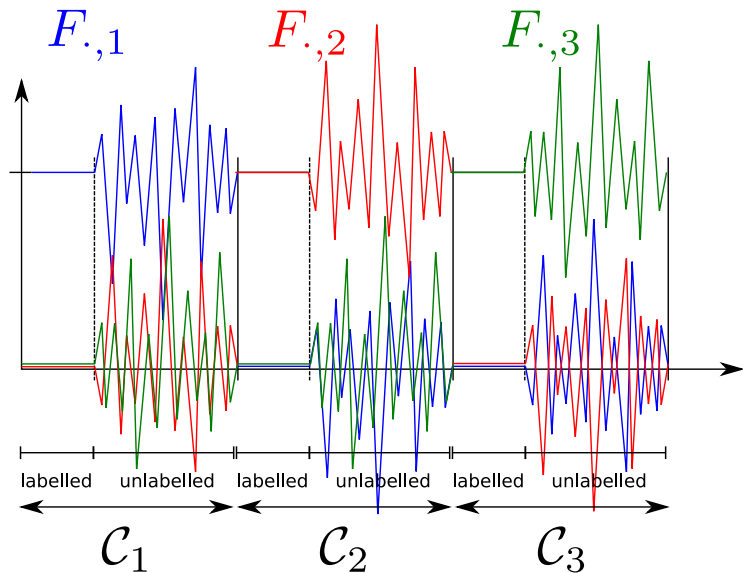


Figure: Typical expected performance output

MNIST Data Example

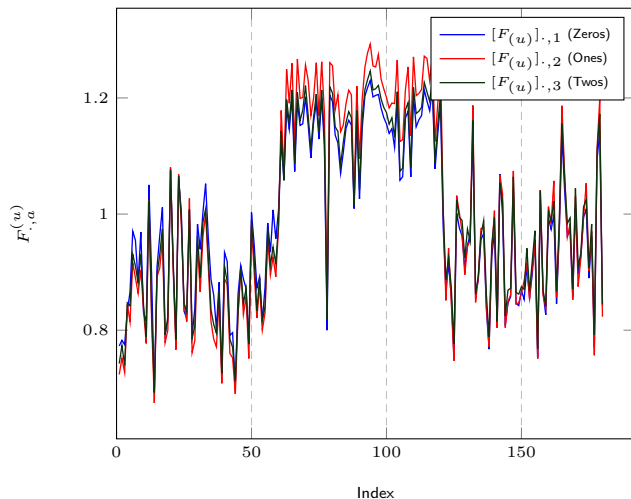


Figure: Vectors $[F^{(u)}]_{:,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

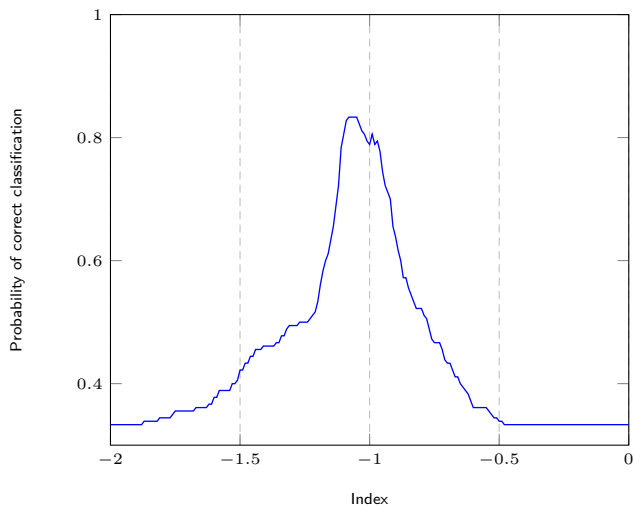


Figure: Performance as a function of α , for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

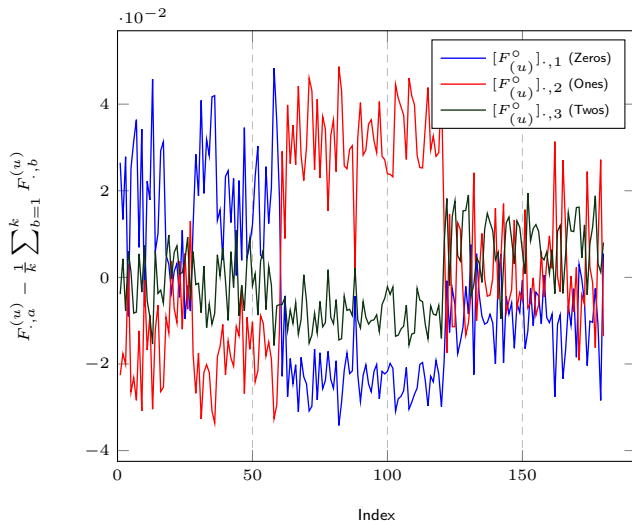


Figure: Centered Vectors $[F_{(u)}^{\circ}]_{.,a} = [F_{(u)} - \frac{1}{k} F_{(u)} \mathbf{1}_k \mathbf{1}_k^T]_{.,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = -1$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Empirical observations:

- ▶ Troubling *flat* classification scores!
- ▶ Only random walk normalized Laplacian regularization ($\alpha = -1$) works!.

Analysis to understand:

- ▶ Consider binary classification for simplicity of notations (easy to generalize to 'one-versus-all' case), and define

$$f_i = F_{i2} - F_{i1}$$

Then x_i is classified in \mathcal{C}_1 if f_i negative, otherwise x_i in \mathcal{C}_2 .

- ▶ Assume $n_{[l]k}/p \rightarrow c_{[l]k} \in (0, 1)$ and $n_{[u]k}/p \rightarrow c_{[u]k} \in (0, 1)$. $c_{[l]} = \sum_k c_{[l]k}$, $c_{[u]} = \sum_k c_{[u]k}$. Under the previous Gaussian mixture data model.

Main Results

We can show that, for x_i unlabelled,

$$f_i = c_0(c_{[l]2} - c_{[l]1}) + o(1)$$

Consequence: All f_i have the same sign if $c_{[l]2} \neq c_{[l]1}$.

Amendment: Use a normalized labelling $y_{[l]}$ ($-1/c_{[l]1}$ for \mathcal{C}_1 , $-1/c_{[l]2}$ for \mathcal{C}_2).

↓

$$f_i = \eta(1 + \alpha)(t_2 - t_1) + o(1/\sqrt{p})$$

Consequence: All f_i have the same sign if $t_2 \neq t_1$.

Amendment: Take $\alpha = -1 + \frac{\beta}{\sqrt{p}}$, $\beta = O(1)$.

↓

$$f_i = g_i + o(1/p)$$

where $g_i \sim \mathcal{N}(m_k, \sigma_k^2)$ for $x_i \in \mathcal{C}_k$ with

$$m_k = \frac{c_{[l]} - c_{[l]k}}{c_{[l]}} (-1)^k \left[-\frac{2f'(\tau)}{pf(\tau)} \|\Delta\mu\|^2 + \frac{f''(\tau)}{pf(\tau)} \Delta t + \frac{2f''(\tau)}{pf(\tau)} \text{tr} \Delta C^2 \right] + (-1)^k \beta \frac{n}{n_l} \frac{f'(\tau)}{pf(\tau)} \Delta t$$

$$\sigma_k^2 = \frac{2\text{tr} C_k^2}{p} \left(\frac{f'(\tau)^2}{pf(\tau)^2} - \frac{f''(\tau)}{pf(\tau)} \right)^2 \Delta t^2 + \frac{4f'(\tau)^2}{p^2 f(\tau)^2} \left[\Delta\mu^\top C_k \Delta\mu + \sum_{a=1}^2 \text{tr} C_k C_a / c_{[l]a} \right]$$

where $\Delta\mu = \mu_2 - \mu_1$, $\Delta t = t_2 - t_1$, $\Delta C = C_2 - C_1$.

Performance: Theoretical versus Empirical

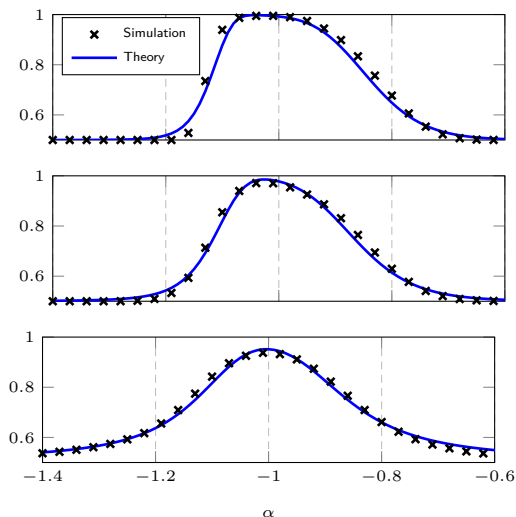


Figure: Theoretical and empirical accuracy as a function of α for 2-class MNIST data (**top:** digits (0,1), **middle:** digits (1,7), **bottom:** digits (8,9)), $n = 1024$, $p = 784$, $n_{[l]}/n = 1/16$, $n_{[u]1} = n_{[u]2}$, Gaussian kernel. Averaged over 50 iterations.

Main Results

↓

$$f_i = g_i + o(1/p)$$

where $g_i \sim \mathcal{N}(m_k, \sigma_k^2)$ for $x_i \in \mathcal{C}_k$ with

$$m_k = \frac{c_{[l]} - c_{[l]k}}{c_{[l]}} (-1)^k \left[-\frac{2f'(\tau)}{pf(\tau)} \|\Delta\mu\|^2 + \frac{f''(\tau)}{pf(\tau)} \Delta t + \frac{2f''(\tau)}{pf(\tau)} \text{tr} \Delta C^2 \right] + (-1)^k \beta \frac{n}{n_l} \frac{f'(\tau)}{pf(\tau)} \Delta t$$

$$\sigma_k^2 = \frac{2\text{tr} C_k^2}{p} \left(\frac{f'(\tau)^2}{pf(\tau)^2} - \frac{f''(\tau)}{pf(\tau)} \right)^2 \Delta t^2 + \frac{4f'(\tau)^2}{p^2 f(\tau)^2} \left[\Delta\mu^\top C_k \Delta\mu + \sum_{a=1}^2 \text{tr} C_k C_a / c_{[l]a} \right]$$

where $\Delta\mu = \mu_2 - \mu_1$, $\Delta t = t_2 - t_1$, $\Delta C = C_2 - C_1$.

m_k, σ_k^2 independent of $c_{[u]}$

Consequence: Learning dominated by labelled data with negligible contribution from unlabelled data. **Not actual semi-supervised learning!**

MNIST Data Example

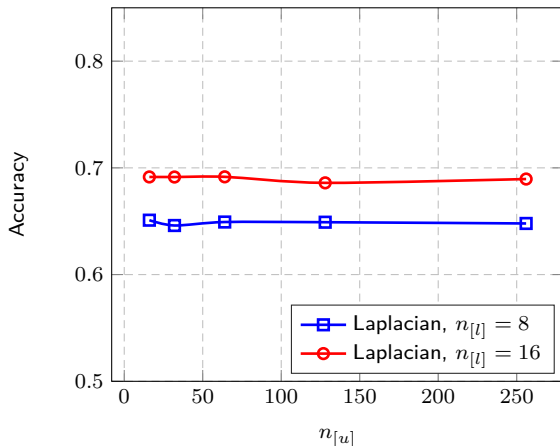


Figure: Classification accuracy as a function of n_u with fixed n_l for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

Main Results

↓

$$f_i = g_i + o(1/p)$$

where $g_i \sim \mathcal{N}(m_k, \sigma_k^2)$ for $x_i \in \mathcal{C}_k$ with

$$m_k = \frac{c_{[l]} - c_{[l]k}}{c_{[l]}} (-1)^k \left[-\frac{2f'(\tau)}{pf(\tau)} \|\Delta\mu\|^2 + \frac{f''(\tau)}{pf(\tau)} \Delta t + \frac{2f''(\tau)}{pf(\tau)} \text{tr} \Delta C^2 \right] + (-1)^k \beta \frac{n}{n_l} \frac{f'(\tau)}{pf(\tau)} \Delta$$

$$\sigma_k^2 = \frac{2\text{tr} C_k^2}{p} \left(\frac{f'(\tau)^2}{pf(\tau)^2} - \frac{f''(\tau)}{pf(\tau)} \right)^2 \Delta t^2 + \frac{4f'(\tau)^2}{p^2 f(\tau)^2} \left[\Delta\mu^\top C_k \Delta\mu + \sum_{a=1}^2 \text{tr} C_k C_a / c_{[l]a} \right]$$

where $\Delta\mu = \mu_2 - \mu_1$, $\Delta t = t_2 - t_1$, $\Delta C = C_2 - C_1$.

m_k, σ_k^2 independent of $c_{[u]}$

Consequence: Learning only from labelled data, **not actual semi-supervised learning!**

Amendment: **No direct solution**, motivating the proposition of **centered kernel regularization**, presented in the following section.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Link between scores flatness and non-expressive unlabelled data:

- ▶ The optimization solution same as **stationary point of label propagation**:

$$f_{[u]} \leftarrow L_{[uu]}^{(\alpha)} f_{[u]} + L_{[ul]}^{(\alpha)} y_{[l]}$$

with $y_{[l]}$ composed of -1 and 1 for respectively labelled data in \mathcal{C}_1 and in \mathcal{C}_2 .

- ▶ **negligible** contribution of $L_{[uu]} f_{[u]}$ if $f_{[u]}$ **flat**.

Cause of flat scores: In high dimensional regime, $K_{ij} \simeq f(\tau)$ for all $i \neq j$, i.e.,

$$(\mathbb{E}\{K_{a_1 a_2}\} - \mathbb{E}\{K_{a_1 b_1}\}) / |\mathbb{E}\{K_{a_1 a_2}\}| |\mathbb{E}\{K_{a_1 b_1}\}| \simeq \epsilon / f(\tau)^2 = o(1)$$

where $x_{a_1}, x_{a_2} \in \mathcal{C}_a$ and $x_{b_1} \in \mathcal{C}_b$ for $a \neq b \in \{1, 2\}$.

Resurrecting SSL by centering

Solution:

- ▶ **“Recenter” K to kill flattening**, i.e., use

$$\tilde{K} = PKP, \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

The recentering imposes $\mathbb{E}\{\hat{K}_{a_1 a_2}\} + \mathbb{E}\{\hat{K}_{a_1 b_1}\} = 0$ (in the case of balanced datasets).

- ▶ Since $\mathbb{E}\{\hat{K}_{a_1 a_2}\} - \mathbb{E}\{\hat{K}_{a_1 b_1}\} = \mathbb{E}\{K_{a_1 a_2}\} - \mathbb{E}\{K_{a_1 b_1}\} = \epsilon$,
 $\mathbb{E}\{\hat{K}_{a_1 a_2}\} = -\mathbb{E}\{\hat{K}_{a_1 b_1}\} = \epsilon/2$.

- ▶ Hence,

$$(\mathbb{E}\{\hat{K}_{a_1 a_2}\} - \mathbb{E}\{\hat{K}_{a_1 b_1}\}) / |\mathbb{E}\{\hat{K}_{a_1 a_2}\}| |\mathbb{E}\{\hat{K}_{a_1 b_1}\}| = 4 = O(1)$$

↓

Non flat scores!

Centered Kernel Regularization

Method:

- ▶ Same loss function as Laplacian regularization, but with **centered similarities** \tilde{K}_{ij} .
- ▶ Optimization problem:

$$\min_f \sum_{i,j=1}^n \tilde{K}_{ij} |f_i - f_j|^2$$

s.t. $\|f_{[u]}\| = t$

with $f_{[l]} = y_{[l]}$.

- ▶ Solution obtained by the Lagrange multipliers method (α being the Lagrange multiplier):

$$f_{[u]} = (\alpha I - \tilde{K}_{[uu]})^{-1} \tilde{K}_{[ul]} y_{[l]} \quad (1)$$

with α determined by $\alpha > \|\tilde{K}_{[uu]}\|$ and $\|f_{[u]}\| = t$.

MNIST Data Example

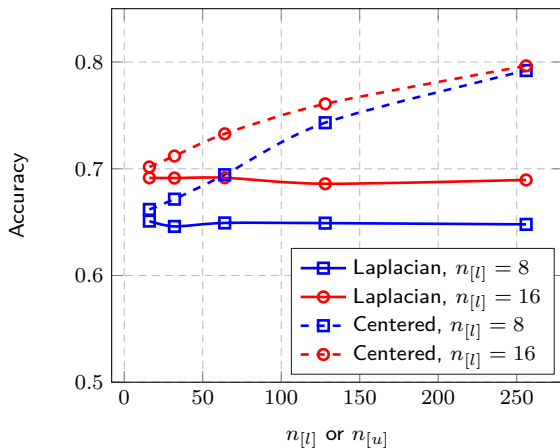


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

Theoretical results

Effective learning from labelled and unlabelled data

- ▶ $m_1 < 0$ and $m_2 > 0$ for all α . (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \text{Var}\{f_i\}$ with $x_i \in \mathcal{C}_k$)
- ▶ $\frac{\sigma_k^2}{m_k^2} = s_k + \frac{\gamma_{[u]k}}{c_{[u]}} + \frac{h(\gamma_{[l]k})}{c_{[l]}}$ where s_k , $\gamma_{[u]k}$ and $\gamma_{[l]k}$ upper-bounded positive values dependent of α .
- ▶ $\gamma_{[u]k}$ is a **decreasing** function of $\gamma_{[l]k}$ which has a minimal value of **zero**. $\gamma_{[l]k}$ can also achieve **zero**, buy only for a **sufficiently large** $c_{[u]}$.

Formula for special cases

- ▶ Setting: $x_i \sim \mathcal{N}(\pm\mu, I_p)$, with balanced data for each class.
- ▶ Formula:

$$\frac{\sigma_1^2}{m_1^2} = \frac{\sigma_2^2}{m_2^2} = \left(1 - \frac{g^2}{\|\mu\|^4 c_{[u]}}\right)^{-1} \left(\frac{1}{\|\mu\|^2} + \frac{g^2}{\|\mu\|^4 c_{[u]}} + \frac{(1-g)^2}{\|\mu\|^4 c_{[l]}}\right)$$

where $g(\alpha) \in (0, q)$ with $q = \min\{1, \sqrt{\|\mu\|^4 c_{[u]}}\}$.

- ▶ Optimal performance of Laplacian regularization (random walk normalized Laplacian):

$$\frac{\sigma_1^2}{m_1^2} = \frac{\sigma_2^2}{m_2^2} = \frac{1}{\|\mu\|^2} + \frac{1}{\|\mu\|^4 c_{[l]}}$$

Performance as a function of $n_{[u]}$, $n_{[l]}$

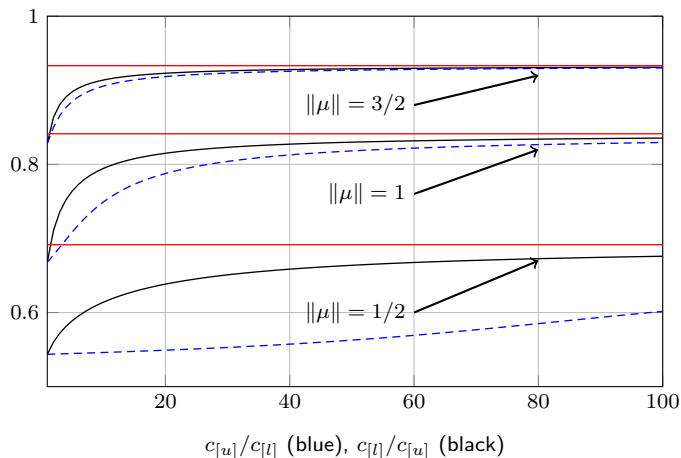


Figure: Correct classification rate, at optimal α , as a function of (i) $n_{[u]}$ for fixed $p/n_{[l]} = 5$ (blue) and (ii) $n_{[l]}$ for fixed $p/n_{[u]} = 5$ (black); $c_1 = c_2 = \frac{1}{2}$; different values for $\|\mu\|$. Comparison to optimal Neyman–Pearson performance for known μ (in red).

SSL: the road from supervised to unsupervised

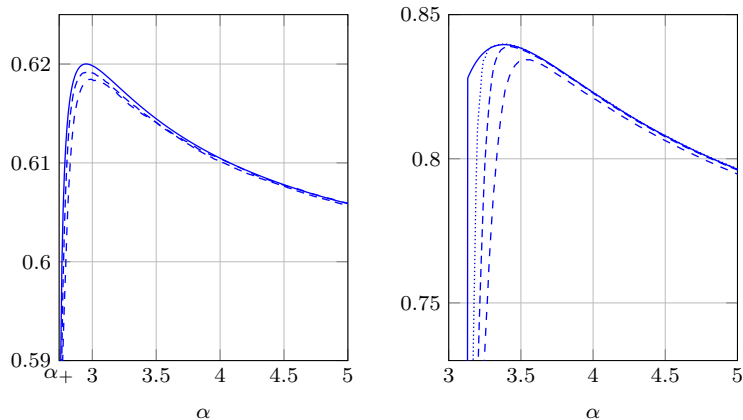


Figure: Theory (solid) versus practice (dashed; from right to left: $n = 400, 1000, 4000$): correct classification probability as a function of α for $c_{[u]} = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, and **left:** $\|\mu\| = 0.75$ (below phase transition); **right:** $\|\mu\| = 1.25$ (above phase transition). Different values of n .

Experimental evidence: MNIST

Digits	(0,8)	(2,7)	(6,9)
$n_{[u]} = 100$			
Centered kernel	89.5±3.6	89.5±3.4	85.3±5.9
Iterated centered kernel	89.5±3.6	89.5±3.4	85.3±5.9
Laplacian	75.5±5.6	74.2±5.8	70.0±5.5
Iterated Laplacian	87.2±4.7	86.0±5.2	81.4±6.8
Manifold	88.0±4.7	88.4±3.9	82.8±6.5
$n_{[u]} = 500$			
Centered kernel	91.7±1.3	92.2±1.3	91.6±2.2
Iterated centered kernel	91.8±1.4	92.2±1.3	92.0±2.1
Laplacian	75.6±4.1	74.4±4.0	69.5±3.7
Iterated Laplacian	91.6±1.5	91.9±1.4	90.6±2.7
Manifold	90.7±2.1	91.2±1.9	90.1±3.7

Table: Comparison of classification accuracy (%) on MNIST datasets with $n_{[u]} = 10$. Computed over 1000 random iterations for $n_{[u]} = 100$ and 500 for $n_{[u]} = 500$.

Experimental evidence: Traffic signs (HOG features)

Class ID	(2,7)	(9,10)	(11,18)
$n_{[u]} = 100$			
Centered kernel	79.0±10.4	77.5±9.2	78.5±7.1
Iterated centered kernel	85.3±5.9	89.2±5.6	90.1±6.7
Laplacian	73.8±9.8	77.3±9.5	78.6±7.2
Iterated Laplacian	83.7±7.2	88.0±6.8	87.1±8.8
Manifold	77.6±8.9	81.4±10.4	82.3±10.8
$n_{[u]} = 500$			
Centered kernel	82.5±4.0	82.6±6.4	79.2±18.0
Iterated centered kernel	84.4±4.2	88.9±5.7	95.8±3.2
Laplacian	72.7±8.9	77.6±8.3	79.1±6.3
Iterated Laplacian	82.7±5.7	88.1±7.4	92.4±6.7
Manifold	77.4±5.9	83.5±10.4	89.3±9.2

Table: Comparison of classification accuracy (%) on German Traffic Sign datasets with $n_{[l]} = 10$. Computed over 1000 random iterations for $n_{[u]} = 100$ and 500 for $n_{[u]} = 500$.

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Motivation: Feature extraction in machine learning

Learning = Representation + Evaluation + Optimization.¹

Features: representation of the data that contains crucial information for the given task.

Various methods for feature extraction:

- ▶ feature selection by hand (expert system)
- ▶ feature learned via backpropagation
- ▶ **random projections/random feature maps:**
 - ▶ simple, fast and tractable theoretical analysis
 - ▶ **early stage** of gradient-based methods (with random initialization)
 - ▶ remaining difficulty: handle the **nonlinearity!**

How to study and understand these features? \Rightarrow Sample Covariance Matrix

$$\text{SCM} \equiv \frac{1}{T} X X^T$$

of data $X = [x_1, \dots, x_T] \in \mathbb{R}^{p \times T}$. SCM in **feature space** \Rightarrow feature Gram matrix G :

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

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

¹Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

Motivation: RMT for random feature maps

Recall: G determines training and test performance via its *resolvent*

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

Example:

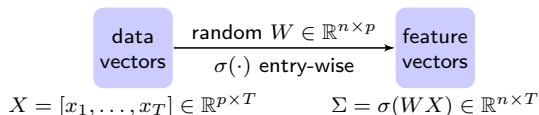


Figure: Illustration of random feature maps

MSE of random feature-based ridge regression (also called *extreme learning machines*):

$$E_{\text{train}} = \frac{1}{T} \|y - \beta^\top \Sigma\|_F^2 = \frac{\gamma^2}{T} y^\top Q^2(-\gamma) y, \quad E_{\text{test}} = \frac{1}{\hat{T}} \|\hat{y} - \beta^\top \hat{\Sigma}\|_F^2$$

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

Key Issue

(Classical) quadratic form $a^\top Q(z)b$ for **nonlinear** model $\Sigma = \sigma(WX)$!

Handle nonlinearity in RMT: concentration of measure approach

Recall:

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

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

Nonlinearity

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

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

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

$$\begin{aligned} Q &= \left(\frac{1}{T} \Sigma^T \Sigma - z I_T \right)^{-1} = \left(\frac{1}{T} \Sigma_{-i}^T \Sigma_{-i} + \frac{1}{T} \sigma_i \sigma_i^T - z I_T \right)^{-1} \\ &= Q_{-i} - \frac{Q_{-i} \frac{1}{T} \sigma_i \sigma_i^T Q_{-i}}{1 + \frac{1}{T} \sigma_i^T Q_{-i} \sigma_i} \end{aligned}$$

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

Handle nonlinearity in RMT: concentration of measure approach

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

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

⇒ extend **trace lemma** to handle nonlinear case!

Lemma (Concentration of Quadratic Forms)

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

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

for some $C, c > 0$ and $\Phi \equiv E_w [\sigma(X^\top w)\sigma(w^\top X)]$ (function of data X).

Theorem (Asymptotic Training Performance)

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

$$E_{\text{train}} - \bar{E}_{\text{train}} \xrightarrow{\text{a.s.}} 0$$

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

Several remarks:

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

Computation of averaged kernel Φ

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

$$\Phi(X) = E_w [\sigma(X^\top w)\sigma(w^\top X)]$$

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

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

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

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

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

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

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

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

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

Numerical validations

Performance of random feature-based ridge regression:

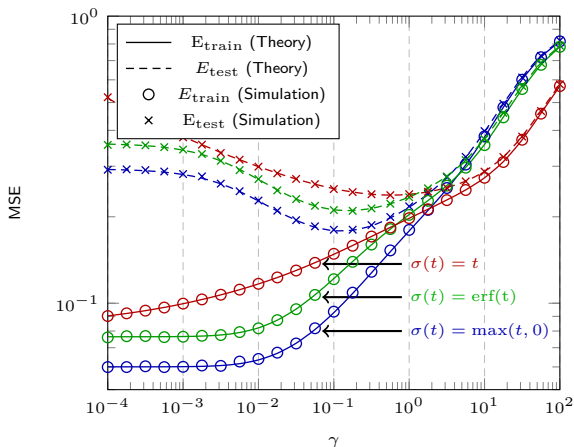


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

⇒ Theoretical performance understanding and **fast tuning** of hyperparameter γ !

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Dig deeper into the averaged kernel Φ

For random feature maps:

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

Data Model (same as for kernel clustering)

Consider data drawn from a K -class Gaussian mixture model (GMM):

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

with $\omega_i \sim \mathcal{N}(0, \frac{1}{p}C_a)$, $a = 1, \dots, K$ of statistical **means** $\mu_a \in \mathbb{R}^p$ and **covariance** $C_a \in \mathbb{R}^{p \times p}$. Class \mathcal{C}_a has cardinality T_a . For $T \rightarrow \infty$, we have

- ▶ $p/T \rightarrow c_0 \in (0, \infty)$
- ▶ $T_a/T \rightarrow c_a \in (0, 1)$
- ▶ let $\mu^\circ \equiv \sum_{i=1}^K \frac{T_i}{T} \mu_i$ and $\mu_a^\circ \equiv \mu_a - \mu^\circ$, then $\|\mu_a^\circ\| = O(1)$
- ▶ let $C^\circ \equiv \sum_{i=1}^K \frac{T_i}{T} C_i$ and $C_a^\circ \equiv C_a - C^\circ$, then $\|C_a\| = O(1)$, $\text{tr} C_a^\circ / \sqrt{p} = O(1)$.

⇒ how different **nonlinearities** influence **statistical information** in Φ (and thus G)?

Analysis of (averaged) kernel matrix Φ (revisit)

Similar to the analysis of kernel matrix $K \equiv f\left(\frac{1}{p}\|x_i - x_j\|^2\right)$, for $\sigma(t) = \text{ReLU}(t)$,

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

with $\angle(x_i, x_j) \equiv \frac{x_i^\top x_j}{\|x_i\| \|x_j\|}$. To understand Φ :

- ▶ Taylor-expand nonlinear functions of x_i, x_j to get **entry-wise approximation** of $\Phi_{i,j}$
- ▶ assembling in matrix form with careful control on **operator norm**

Theorem (Asymptotic Equivalent of Φ)

For all $\sigma(\cdot)$ listed, we have, as $T \rightarrow \infty$,

$$\|\Phi - \tilde{\Phi}\| \xrightarrow{\text{a.s.}} 0$$

with

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

and $U = \left[\frac{J}{\sqrt{p}}, \phi \right]$, $B = \begin{bmatrix} t t^\top + 2S & t \\ t^\top & 1 \end{bmatrix}$, where $J = [j_1, \dots, j_K]$, j_a canonical vector of class C_a (**for clustering**), weighted by two key parameters d_1, d_2 and

- ▶ Ω, ϕ **random** fluctuations of data
- ▶ $M = [\mu_1^\circ, \dots, \mu_K^\circ]$ containing differences in **means**, $t = \left\{ \frac{1}{\sqrt{p}} \text{tr} C_a^\circ \right\}_{a=1}^K$ and $S = \left\{ \frac{1}{p} \text{tr} C_a C_b \right\}_{a,b=1}^K$ differences in **traces** and **shapes** of **covariances**.

Consequence

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

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

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

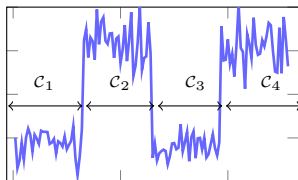
- ▶ **mean-oriented**, $d_1 \neq 0, d_2 = 0$: $t, 1_{t>0}, \text{sign}(t), \sin(t)$ and $\text{erf}(t)$
 \Rightarrow separate with differences in means M ;
- ▶ **covariance-oriented**, $d_1 = 0, d_2 \neq 0$: $|t|, \cos(t)$ and $\exp(-t^2/2)$
 \Rightarrow track differences in covariances t, S ;
- ▶ **balanced**,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

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

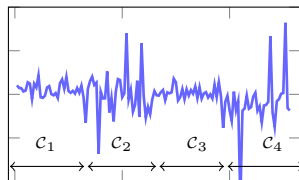
Numerical validations: Gaussian data

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

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

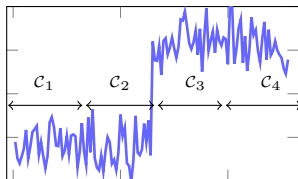


Eigenvector 1

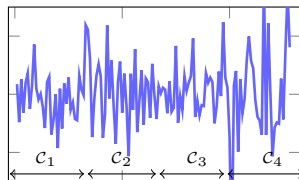


Eigenvector 2

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



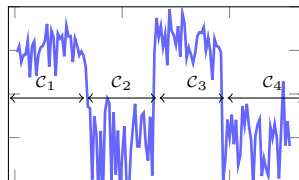
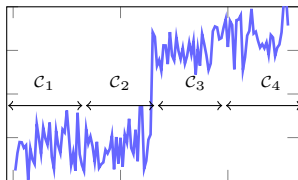
Eigenvector 1



Eigenvector 2

Numerical validations: Gaussian data

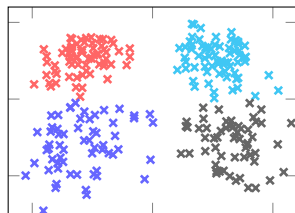
Case 3: $\varsigma_+ = 1$, $\varsigma_- = 0$ (the ReLU function)



Eigenvector 1

Eigenvector 2

Eigenvector 2



Eigenvector 1

Numerical validations: real datasets

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

	$\ M^T M\ $	$\ tt^T + 2S\ $
MNIST data	172.4	86.0
EEG data	1.2	182.7

Table: Clustering accuracies on MNIST dataset.

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

Table: Clustering accuracies on EEG dataset.

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

Numerical validations: real datasets

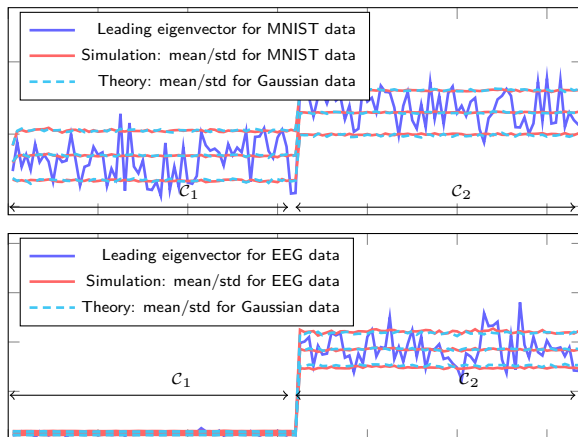


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Summary: random feature maps

Summary for random feature maps:

- ▶ **concentration of measure** helps extend **trace lemma** to nonlinear case
⇒ asymptotic training/test performance of random feature-based ridge regression
- ▶ “concentration” of high dimensional data helps understand the key **averaged kernel** matrix Φ ⇒ random feature-based spectral clustering

Take-away messages:

- ▶ fast tuning of hyperparameters
- ▶ nonlinearities into three attributes: **means-**, **covariance-oriented** and “**balanced**”
- ▶ **optimize** the choice of nonlinearity as a function of data for quadratic and LReLU (similar to the “ α - β ” kernel!)

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

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

Motivation: learning dynamics of neural networks

About neural networks and deep learning:

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

⇒ The learning dynamics of neural networks!

With RMT:

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

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

As a consequence, more insights on:

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

Problem setup

A toy model of binary classification:

Gaussian Mixture Data

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

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

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

Objective: Learning Dynamics

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

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

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

To evaluate the learning dynamics:

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

Objective: Generalization Performance

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

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

$$\text{for } w(t) = e^{-\frac{\alpha t}{n} X X^\top} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} X X^\top} \right) (X X^\top)^{-1} X y.$$

With RMT:

- ▶ although X random: $w(t)^\top \mu$ and $\|w(t)\|^2$ have **asymptotically** deterministic behavior (only depends on **data statistics** and problem dimension):
⇒ the technique of **deterministic equivalent**
- ▶ **Cauchy's integral formula** to express the functional $\exp(\cdot)$ via contour integration
⇒ Network performance at **any** time is in fact **deterministic** and **predictable!**

Proposed analysis framework

Resolvent and deterministic equivalents

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

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

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

$$\blacktriangleright \frac{1}{n} \operatorname{tr} A Q_M - \frac{1}{n} \operatorname{tr} A \bar{Q}_M \xrightarrow{\text{a.s.}} 0$$

$$\blacktriangleright a^\top (Q_M - \bar{Q}_M) b \xrightarrow{\text{a.s.}} 0$$

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

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

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

Cauchy's integral formula

Example: for $f(M) = a^\top e^M b dz$,

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

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

Generalization performance

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

► **Cauchy's integral formula:** for $w(t)^\top \mu$:

$$\mu^\top w(t) = -\frac{1}{2\pi i} \oint_{\gamma} \mu^\top \left(\frac{1}{n} X X^\top - z I_p \right)^{-1} \left(f_t(z) w_0 + \frac{1 - f_t(z)}{z} \frac{1}{n} X y \right) dz$$

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

$$\begin{aligned} \left(\frac{1}{n} X X^\top - z I_p \right)^{-1} &= Q(z) - Q(z) \begin{bmatrix} \mu & \frac{1}{n} \Omega y \end{bmatrix} \\ \begin{bmatrix} \mu^\top Q(z) \mu & 1 + \frac{1}{n} \mu^\top Q(z) \Omega y \\ 1 + \frac{1}{n} \mu^\top Q(z) \Omega y & -1 + \frac{1}{n} y^\top \Omega^\top Q(z) \frac{1}{n} \Omega y \end{bmatrix}^{-1} &\begin{bmatrix} \mu^\top \\ \frac{1}{n} y^\top \Omega^\top \end{bmatrix} Q(z) \end{aligned}$$

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

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

with $m(z)$ given by Marčenko-Pastur equation $m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2 - 4cz}}{2cz}$.

► “replace” the random $Q(z)$ by its **deterministic equivalent** $\bar{Q}(z) = m(z) I_p$.

Theorem (Generalization Performance)

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

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

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

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

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

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

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

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

Simplification: “break” the contour integration

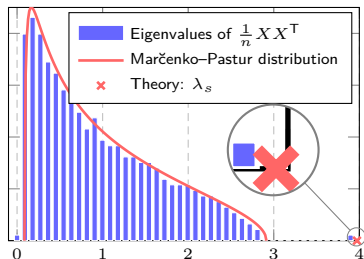


Figure: Eigenvalue distribution of $\frac{1}{n}XX^T$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1024$.

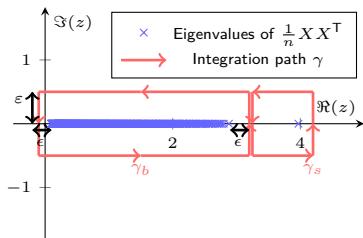


Figure: Eigenvalue distribution of $\frac{1}{n}XX^T$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1024$.

Two types of eigenvalues:

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

Computation of λ_s (Spike model)

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

$$\det\left(\frac{1}{n}XX^T - \lambda I_p\right) = 0$$

$$\Leftrightarrow \det\left(\frac{1}{n}\Omega\Omega^T - \lambda I_p + \begin{bmatrix} \mu & \frac{1}{n}\Omega y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \mu^T \\ \frac{1}{n}y^T\Omega^T \end{bmatrix}\right) = 0$$

$$\Leftrightarrow \det\left(I_2 + \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \mu^T \\ \frac{1}{n}y^T\Omega^T \end{bmatrix} Q(\lambda) \begin{bmatrix} \mu & \frac{1}{n}\Omega y \end{bmatrix}\right) = 0$$

$$\Leftrightarrow 1 + (\|\mu\|^2 + c)m(\lambda) + o(1) = 0$$

(Simplified) generalization performance

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

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

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

Some remarks:

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

Numerical validations

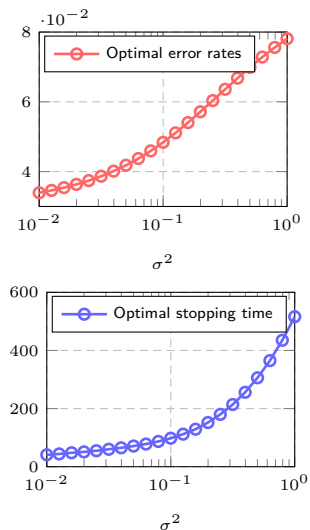


Figure: Optimal performance and stopping time as functions of σ^2 with $c = 1/2$, $\|\mu\|^2 = 4$ and $\alpha = 0.01$.

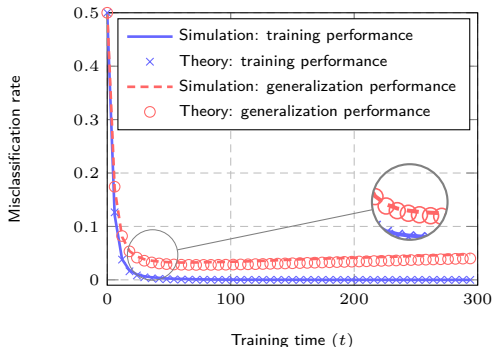


Figure: Training and generalization performance for MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 0.1$. Results averaged over 100 runs.

Summary: RMT for network learning dynamics

Take-away messages:

- ▶ RMT framework to understand and **predict** learning dynamics:
 - Cauchy's integral formula + technique of deterministic equivalent
- ▶ easily extended to more elaborate data models: e.g., Gaussian mixture model with different means and covariances
- ▶ a byproduct: choose the initialization variance σ^2 **even smaller** (than classical normalization of $1/p$)!

Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

Applications to Machine Learning (**Xiaoyi MAI**)

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

Improved Semi-supervised Learning

Applications to Random Projections and Neural Networks (**Zhenyu LIAO**)

Random Projections-based Ridge Regression

Random Projections-based Spectral Clustering

Random Matrix Analysis for Learning Dynamics of Neural Networks

Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)









Take-away messages

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

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

Summary of Results and Perspectives I

Kernel Methods: References

-  N. El Karoui, "The spectrum of kernel random matrices", *The Annals of Statistics*, 38(1), 1-50, 2010.
-  C. Xiuyuan, A. Singer, "The spectrum of random inner-product kernel matrices", *Random Matrices: Theory and Applications* 2.04 (2013): 1350010.
-  R. Couillet, F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data", *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393-1454, 2016.
-  R. Couillet, A. Kammoun, "Random Matrix Improved Subspace Clustering", *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2016.
-  Z. Liao, R. Couillet, "Random matrices meet machine learning: a large dimensional analysis of LS-SVM", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, USA, 2017.
-  X. Mai, R. Couillet, "The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, USA, 2017.
-  X. Mai, R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data", (under review) *Journal of Machine Learning Research*, 2017.
-  Z. Liao, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines", (under review) *Journal of Machine Learning Research*, 2017.

Summary of Results and Perspectives II

Kernel Methods: References



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



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

Summary of Results and Perspectives I

Feature Maps and Neural Networks: References



C. Williams, "Computation with infinite neural networks", *Neural Computation*, 10(5), 1203-1216, 1998.



A. Rahimi, B. Recht, "Random features for large-scale kernel machines", *Advances in neural information processing systems* pp. 1177-1184, 2007.



N. El Karoui, "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond", *The Annals of Applied Probability*, 19(6), 2362-2405, 2009.



A. Saxe, J. McClelland, S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks", arXiv:1312.6120, 2013.



A. Choromanska, M. Henaff, M. Mathieu, G. Arous, Y. LeCun, "The loss surfaces of multilayer networks", In *Artificial Intelligence and Statistics* (pp. 192-204), 2015.



R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "The asymptotic performance of linear echo state neural networks", *Journal of Machine Learning Research*, vol. 17, no. 178, pp. 1-35, 2016.



C. Louart, R. Couillet, "Harnessing neural networks: a random matrix approach", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, USA, 2017.

Summary of Results and Perspectives II

Feature Maps and Neural Networks: References



C. Louart, R. Couillet, "A Random Matrix and Concentration Inequalities Framework for Neural Networks Analysis", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18), Calgary, Canada, 2018.



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



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










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



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

Summary of Results and Perspectives I

Robust Statistics: References

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

Thank you.