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

Outline

Context

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

Context

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

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

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

Context

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

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

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

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

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

  or equivalently, in spectral norm

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

Context

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

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

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

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

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

or equivalently, in spectral norm

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

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

Context

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

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

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

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

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

or equivalently, in spectral norm

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

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

- For practical p, n with $p \simeq n$, leads to dramatically wrong conclusions

Context

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

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

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

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

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

or equivalently, in spectral norm

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

Random Matrix Regime

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

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

- 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., $x_1 \sim \mathcal{CN}(0, I_p)$

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

- assume $p = p(n)$ such that $p/n \to c > 1$

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

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

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

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

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

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

- however, eigenvalue mismatch

$$0 = \lambda_1(\hat{C}_p) = \ldots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \ldots \leq \lambda_p(\hat{C}_p)$$
$$1 = \lambda_1(I_p) = \ldots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \ldots = \lambda_p(I_p)$$

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

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

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

- however, eigenvalue mismatch

$$0 = \lambda_1(\hat{C}_p) = \ldots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \ldots \leq \lambda_p(\hat{C}_p)$$
$$1 = \lambda_1(I_p) = \ldots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \ldots = \lambda_p(I_p)$$

$\Rightarrow$ no convergence in spectral norm.

The Marčenko–Pastur law

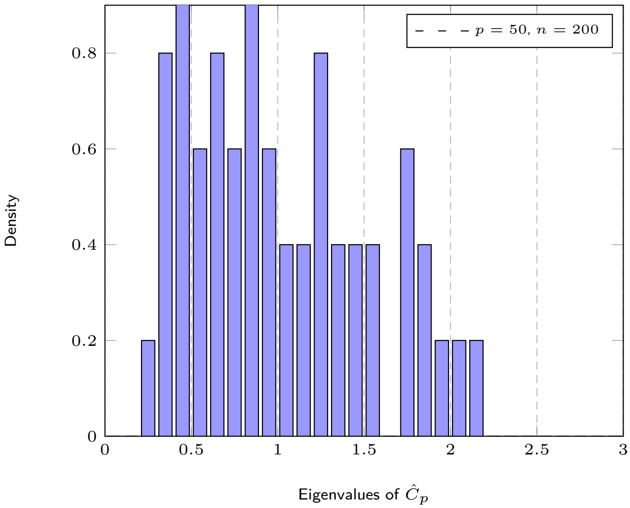


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

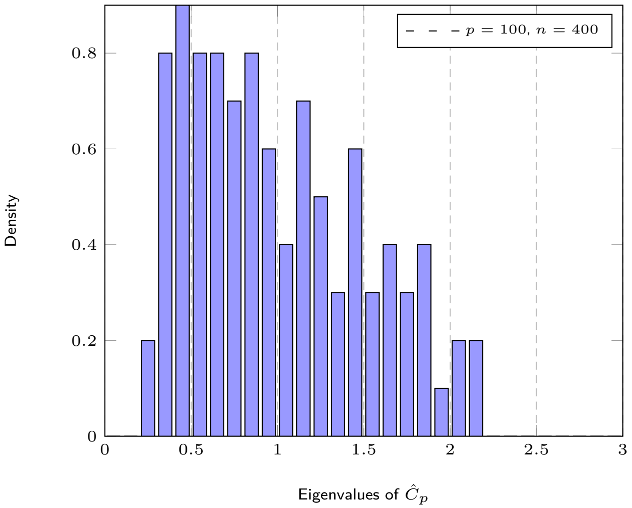


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

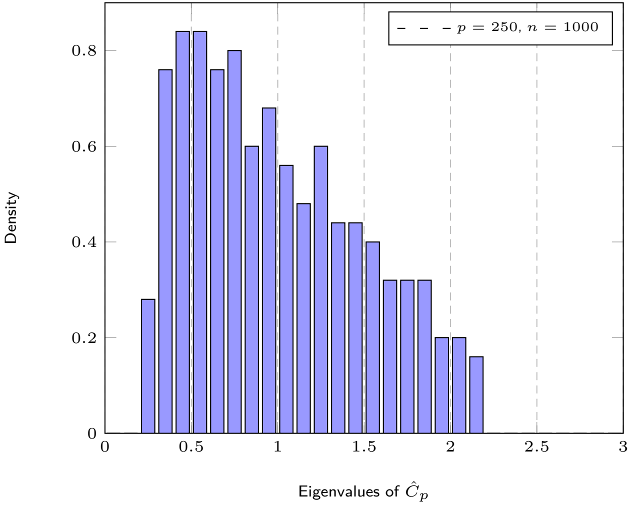


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

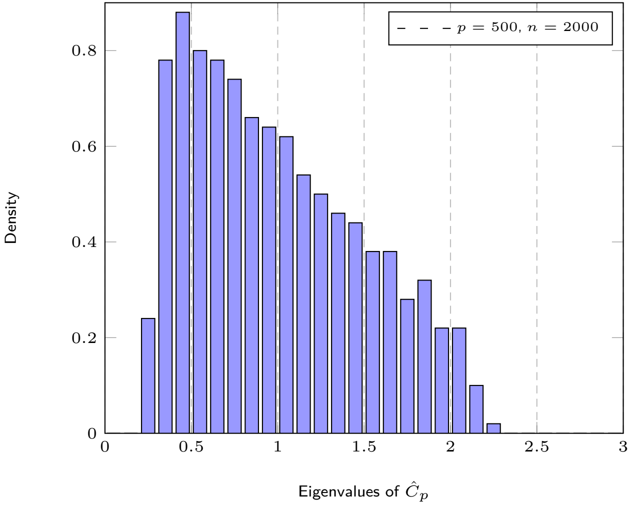


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

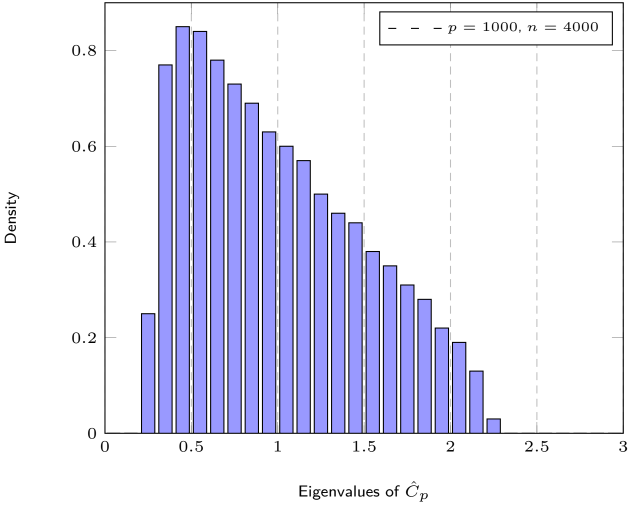


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

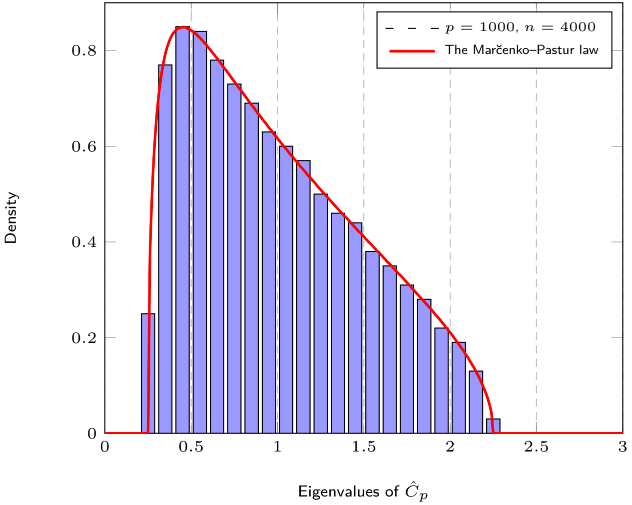


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

Definition (Empirical Spectral Distribution)

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

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

Definition (Empirical Spectral Distribution)

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

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

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 $p,n\to\infty$ with $p/n \to c \in (0,\infty)$, e.s.d. μ_p of $\frac{1}{n}X_pX_p^\mathsf{T}$ 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

- $\mu_c(\{0\}) = \max\{0, 1-c^{-1}\}$

Definition (Empirical Spectral Distribution)

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

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

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

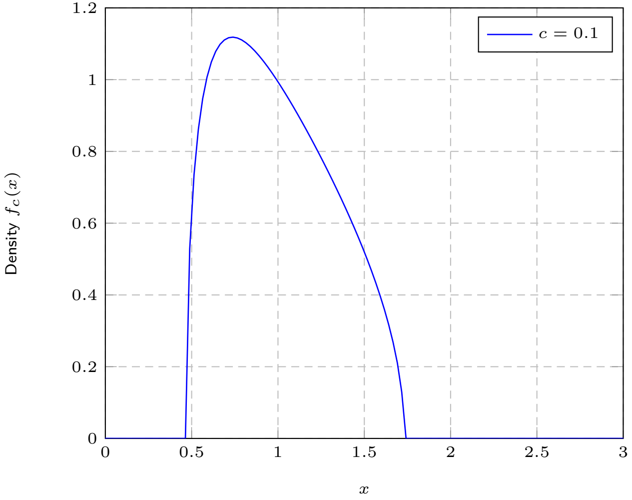


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

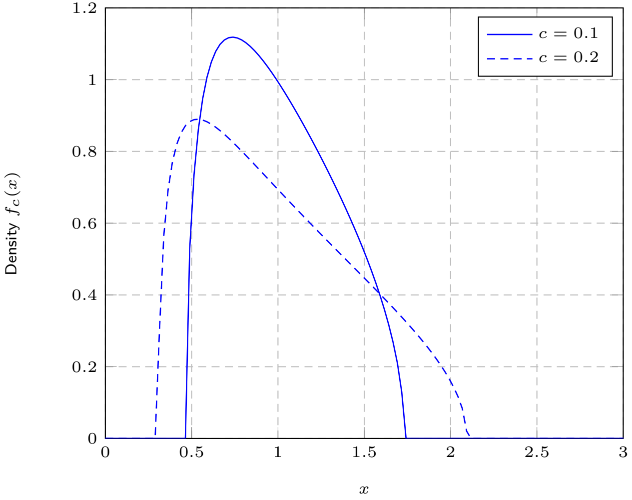


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

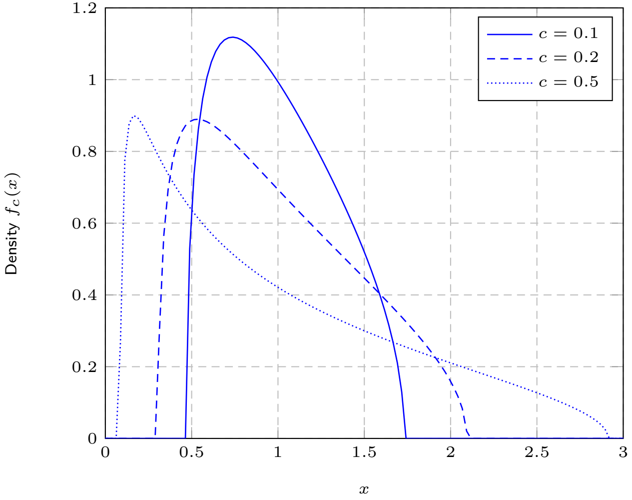


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Outline

Definition (Stieltjes Transform)

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

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

The Stieltjes transform

Definition (Stieltjes Transform)

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

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

Property (Inverse Stieltjes Transform)

For $a < b$ continuity points of μ,

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

Definition (Stieltjes Transform)

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

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

Property (Inverse Stieltjes Transform)

For $a < b$ continuity points of μ,

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

Besides, if μ has a density f at x,

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

Property (Relation to e.s.d.)

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

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

Property (Relation to e.s.d.)

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

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

Proof:

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

Property (Relation to e.s.d.)

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

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

Proof:

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

Fundamental object: the resolvent of A

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

Property (Stieltjes transform of Gram matrices)

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

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

Then

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

Property (Stieltjes transform of Gram matrices)

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

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

Then

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

Proof:

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

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.

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 $A + txx^{\mathsf{T}}$ invertible,

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

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

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, $t > 0$, $x \in \mathbb{R}^p$, $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

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

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, $t > 0$, $x \in \mathbb{R}^p$, $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

$$\left| \frac{1}{p} \operatorname{tr} B \left(A + t x x^{\mathsf{T}} - z I_p \right)^{-1} - \frac{1}{p} \operatorname{tr} B \left(A - z I_p \right)^{-1} \right| \leq \frac{1}{p} \frac{\|B\|}{\operatorname{dist}(z, \operatorname{supp}(\mu))}$$

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

$$\frac{1}{p} \operatorname{tr} B \left(A + t x x^{\mathsf{T}} - z I_p \right)^{-1} - \frac{1}{p} \operatorname{tr} B \left(A - z I_p \right)^{-1} \to 0.$$

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, $t > 0$, $x \in \mathbb{R}^p$, $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

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

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

$$\frac{1}{p} \operatorname{tr} B \left(A + txx^{\mathsf{T}} - zI_p \right)^{-1} - \frac{1}{p} \operatorname{tr} B \left(A - zI_p \right)^{-1} \to 0.$$

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

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

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

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

$$E\left[\left|\frac{1}{p}x^{\mathsf{T}}Ax - \frac{1}{p}\mathrm{tr}\,A\right|^k\right] \leq K\frac{\|A\|^p}{p^{k/2}}.$$

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

$$\frac{1}{p}x^{\mathsf{T}}Ax - \frac{1}{p}\mathrm{tr}\,A \xrightarrow{\text{a.s.}} 0$$

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

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

weakly, where

- $\mu_{(c)}(\{0\}) = \max\{0, 1 - c^{-1}\}$

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 $p, n \to \infty$ with $p/n \to c \in (0,\infty)$, e.s.d. μ_p of $\frac{1}{n}X_pX_p^{\mathsf{T}}$ 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 $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$*

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

Stieltjes transform approach.

Stieltjes transform approach.

Proof

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

$$m_{\mu_p}(z) = \frac{1}{p}\mathrm{tr}\left(\frac{1}{n}X_pX_p^{\mathsf{T}} - zI_p\right)^{-1} = \frac{1}{p}\sum_{i=1}^{p}\left[\left(\frac{1}{n}X_pX_p^{\mathsf{T}} - zI_p\right)^{-1}\right]_{ii}.$$

Stieltjes transform approach.

Proof

- With μ_p e.s.d. of $\frac{1}{n}X_p X_p^\mathsf{T}$,

$$m_{\mu_p}(z) = \frac{1}{p}\mathrm{tr}\left(\frac{1}{n}X_p X_p^\mathsf{T} - zI_p\right)^{-1} = \frac{1}{p}\sum_{i=1}^{p}\left[\left(\frac{1}{n}X_p X_p^\mathsf{T} - zI_p\right)^{-1}\right]_{ii}.$$

- Write

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

Stieltjes transform approach.

Proof

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

$$m_{\mu_p}(z) = \frac{1}{p}\mathrm{tr}\left(\frac{1}{n}X_pX_p^\mathsf{T} - zI_p\right)^{-1} = \frac{1}{p}\sum_{i=1}^{p}\left[\left(\frac{1}{n}X_pX_p^\mathsf{T} - zI_p\right)^{-1}\right]_{ii}.$$

- Write

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

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

$$\left(\frac{1}{n}X_pX_p^\mathsf{T} - zI_p\right)^{-1} = \begin{pmatrix} \frac{1}{n}y^\mathsf{T}y - z & \frac{1}{n}y^\mathsf{T}Y_{p-1} \\ \frac{1}{n}Y_{p-1}y & \frac{1}{n}Y_{p-1}Y_{p-1}^\mathsf{T} - zI_{p-1} \end{pmatrix}^{-1}.$$

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

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

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

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

- By Trace Lemma, as $p, n \to \infty$

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

Proof (continued)

- By Rank-1 Perturbation Lemma ($X_p^\mathsf{T} X_p = Y_{p-1}^\mathsf{T} Y_{p-1} + yy^\mathsf{T}$), as $p, n \to \infty$

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

Proof (continued)

- By Rank-1 Perturbation Lemma ($X_p^\mathsf{T} X_p = Y_{p-1}^\mathsf{T} Y_{p-1} + yy^\mathsf{T}$), as $p, n \to \infty$

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

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

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

Proof (continued)

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

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

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

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

- Repeating for entries $(2,2), \ldots, (p,p)$, 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 (continued)

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

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

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 $m(z) \to 0$ as $|z| \to \infty$)

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

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 $m(z) \to 0$ as $|z| \to \infty$)

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

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

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

And for $x = 0$,

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

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.* $\nu_p \to \nu$ *weakly,*
- $X_p \in \mathbb{C}^{p \times n}$ *has i.i.d. entries of zero mean and unit variance.*

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

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

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.* $\nu_p \to \nu$ *weakly,*
- $X_p \in \mathbb{C}^{p \times n}$ *has i.i.d. entries of zero mean and unit variance.*

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

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

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

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.* $\nu_p \to \nu$ *weakly,*
- $X_p \in \mathbb{C}^{p \times n}$ *has i.i.d. entries of zero mean and unit variance.*

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

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

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

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

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

weakly, with $\tilde{\mu} = c\mu + (1 - c)\boldsymbol{\delta}_0$.

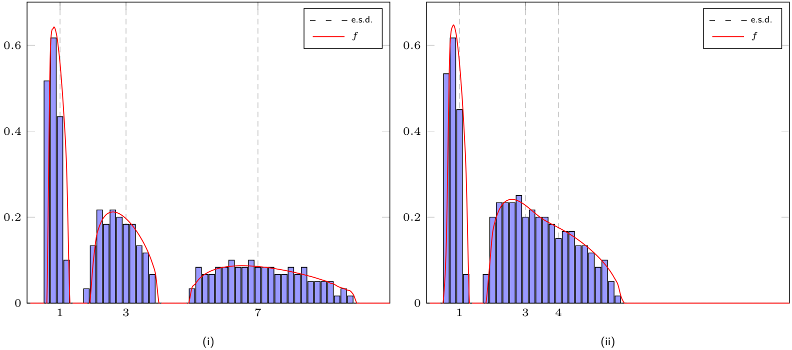


Figure: Histogram of the eigenvalues of $\frac{1}{n} Y_p Y_p^\mathsf{T}$, $n = 3000$, $p = 300$, with C_p diagonal with evenly weighted masses in (i) $1, 3, 7$, (ii) $1, 3, 4$.

Sometimes, μ_p does not converge!

Sometimes, μ_p does not converge!

- if ν_p does not converge

Sometimes, μ_p does not converge!

- if ν_p does not converge
- if p/n does not converge

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!

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

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

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 $B_p = C_p^{\frac{1}{2}} X_p T_p X_p^\mathsf{T} C_p^{\frac{1}{2}}$*, 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 $p, n \to \infty$ *with bounded ratio* c*, for* $z \in \mathbb{C} \setminus \mathbb{R}^-$*,*

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

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*

$$e_p(z) = \frac{1}{p} \operatorname{tr} C_p \left(-zI_p + \bar{e}_p(z) C_p\right)^{-1}$$
$$\bar{e}_p(z) = \frac{1}{n} \operatorname{tr} T_p \left(I_n + c e_p(z) T_p\right)^{-1}.$$

Side note on other models.
Similar results for multiple matrix models:

Other Refined Sample Covariance Models

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**: $Y_p = [y_1, \ldots, y_n]$, $y_i = C_{p,i}^{\frac{1}{2}} x_i$
- etc.

Outline

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.* $\nu_p \to \nu$ *weakly,*

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.* $\nu_p \to \nu$ *weakly,*
- $X_p \in \mathbb{R}^{p \times n}$ *has i.i.d. entries of zero mean and unit variance,*

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.* $\nu_p \to \nu$ *weakly,*
- $X_p \in \mathbb{R}^{p \times n}$ *has i.i.d. entries of zero mean and unit variance,*
- $E[|X_p|_{ij}^4] < \infty$,

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

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

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

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

for all large n*, almost surely.*

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

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

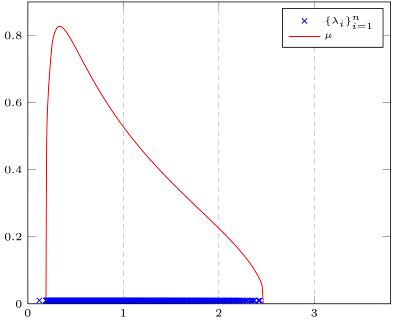
$$\left\{ \lambda_i \left(\frac{1}{n} Y_p^\mathsf{T} Y_p \right) \right\}_{i=1}^{n} \cap [a,b] = \emptyset$$

for all large n*, almost surely.*

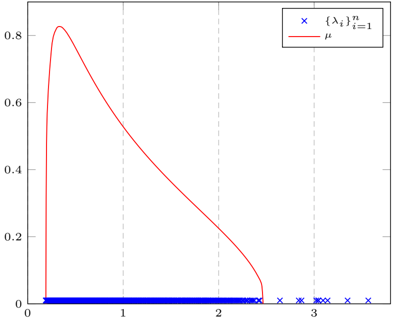
In practice: This means that eigenvalues of $\frac{1}{n} Y_p^\mathsf{T} Y_p$ cannot be bound at macroscopic distance from the bulk, for p, n large.

Breaking the rules. If we break

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



$E[X_{ij}^4] < \infty$



$E[X_{ij}^4] = \infty$

If we break:

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

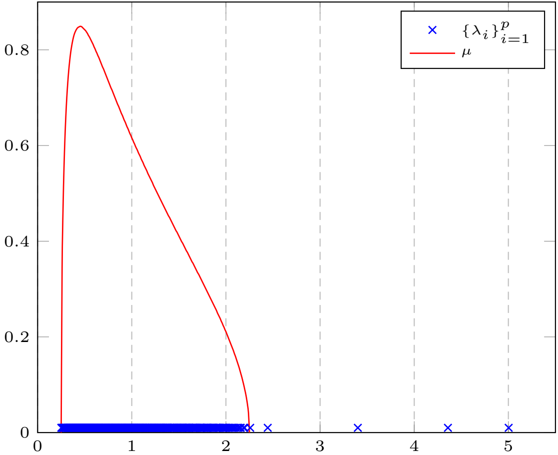


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^{\mathsf{T}}$, $C_p = \text{diag}(\underbrace{1,\ldots,1}_{p-4},2,3,4,5)$, $p = 500$, $n = 2000$.

Spiked Models: The phase transition phenomenon

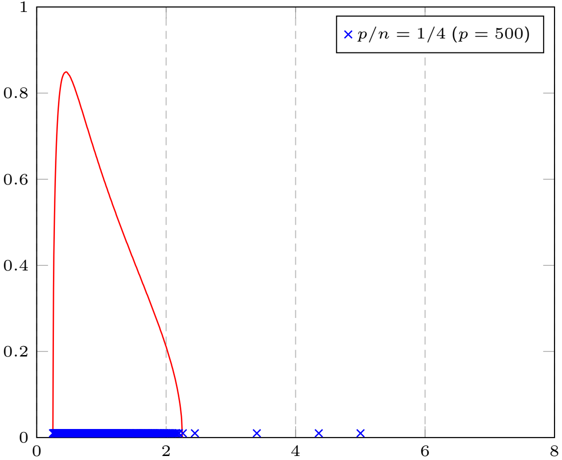


Figure: Eigenvalues of $\frac{1}{n}Y_p Y_p^{\mathsf{T}}$, $C_p = \text{diag}(\underbrace{1,\ldots,1}_{p-4}, 2, 3, 4, 5)$.

Spiked Models: The phase transition phenomenon

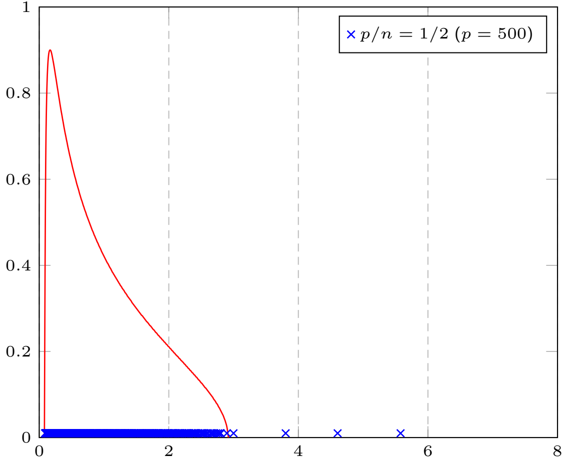


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^\mathsf{T}$, $C_p = \mathrm{diag}(\underbrace{1, \ldots, 1}_{p-4}, 2, 3, 4, 5)$.

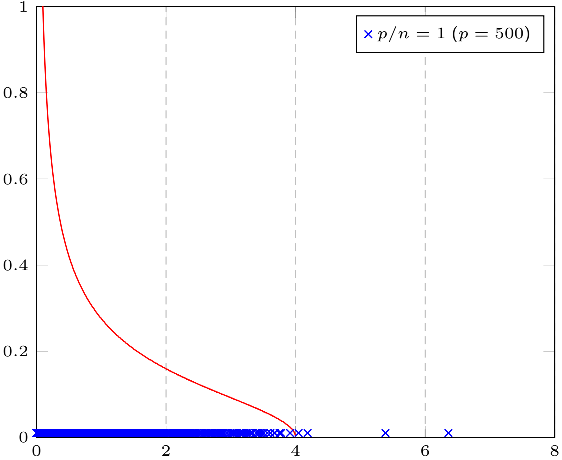


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^{\mathsf{T}}$, $C_p = \text{diag}(\underbrace{1,\ldots,1}_{p-4},2,3,4,5)$.

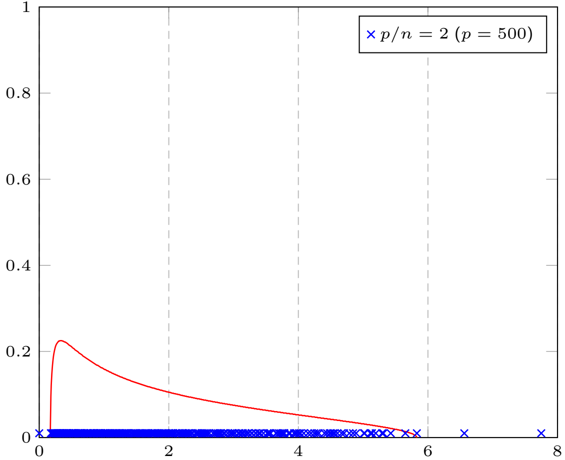


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^{\mathsf{T}}$, $C_p = \text{diag}(\underbrace{1, \ldots, 1}_{p-4}, 2, 3, 4, 5)$.

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

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

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

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

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

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

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

- *if* $\omega_m > \sqrt{c}$,

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

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

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

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

- *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* $\omega_m \in (0, \sqrt{c}]$,

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

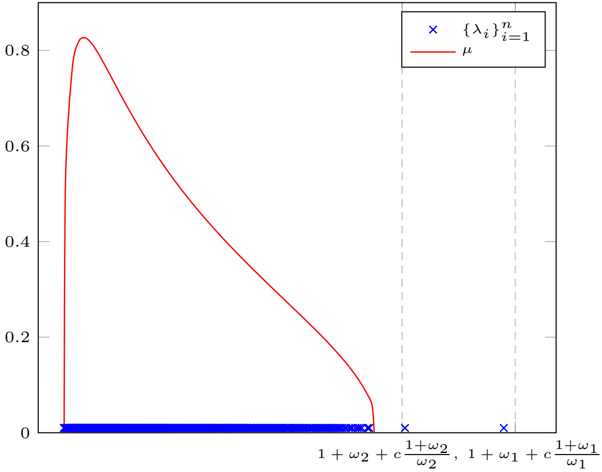


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^{\mathsf{T}}$, $C_p = \text{diag}(\underbrace{1,\ldots,1}_{p-2},2,3)$, $p=500$, $n=1500$.

Proof

- **Two ingredients**: Algebraic calculus + trace lemma

Proof

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

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

Proof

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

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

- **Use low rank property**: $(C_p = I_p + P = I_p + U\Omega U^\mathsf{T})$

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

Hence

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

Proof

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

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

- **Use low rank property**: $(C_p = I_p + P = I_p + U\Omega U^{\mathsf{T}})$

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

Hence

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

Proof (2)

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

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

Proof (2)

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

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

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_pX_p^\mathsf{T} - \lambda I_p)$ has no zero beyond $(1+\sqrt{c})^2$ for all large n a.s.

Proof (2)

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

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

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_pX_p^\mathsf{T} - \lambda I_p)$ has no zero beyond $(1+\sqrt{c})^2$ for all large n a.s.
- **Extension of Trace Lemma**: for each $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$,

$$U^\mathsf{T}\left(\frac{1}{n}X_pX_p^\mathsf{T} - zI_p\right)^{-1}U \xrightarrow{\text{a.s.}} m_\mu(z)I_K.$$

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

Proof (2)

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

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

- **No eigenvalue outside the support [Bai,Sil'98]**: $\det(\frac{1}{n}X_pX_p^\mathsf{T}-\lambda I_p)$ has no zero beyond $(1+\sqrt{c})^2$ for all large n a.s.
- **Extension of Trace Lemma**: for each $z\in\mathbb{C}\setminus\operatorname{supp}(\mu)$,

$$U^\mathsf{T}\left(\frac{1}{n}X_pX_p^\mathsf{T}-zI_p\right)^{-1}U\xrightarrow{\text{a.s.}}m_\mu(z)I_K.$$

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

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

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

Proof (3)

- **Limiting solutions:** zeros of

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

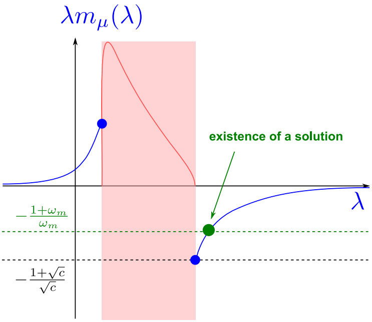
Proof (3)

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

Theorem (Eigenvectors **[Paul'07]**)

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

- X_p *with i.i.d. zero mean, unit variance, finite fourth order moment entries*
- $C_p = I_p + P$, $P = \sum_{i=1}^K \omega_i u_i u_i^\mathsf{T}$, $\omega_1 > \ldots > \omega_M > 0$.

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

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

In particular,

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

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

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

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

In particular,

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

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

$$a^\mathsf{T} \hat{u}_i \hat{u}_i^\mathsf{T} b = \frac{1}{2\pi\imath} \oint_{\mathcal{C}_i} a^\mathsf{T} \left(\frac{1}{n} Y_p Y_p^\mathsf{T} - z I_p \right)^{-1} b \, dz$$

for $\mathcal{C}_m$ contour circling around λ_i only.

Spiked Models

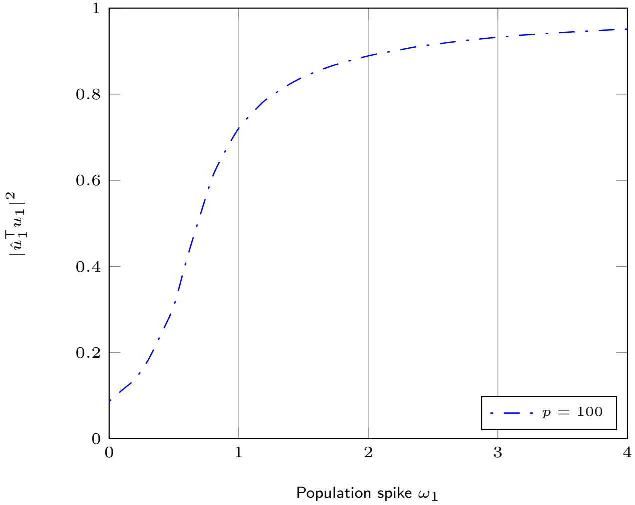


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

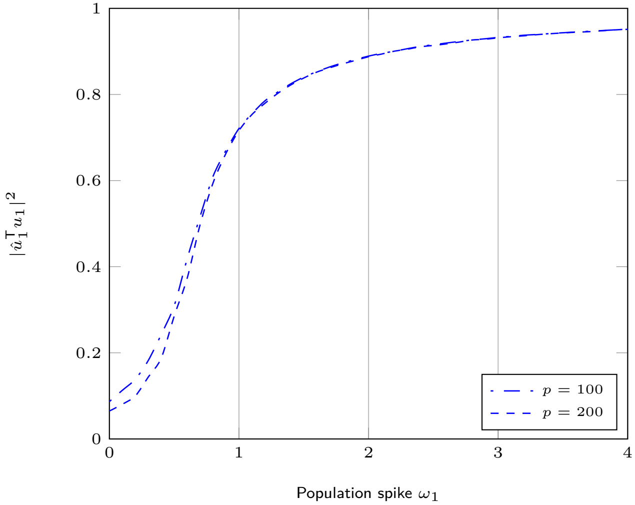


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

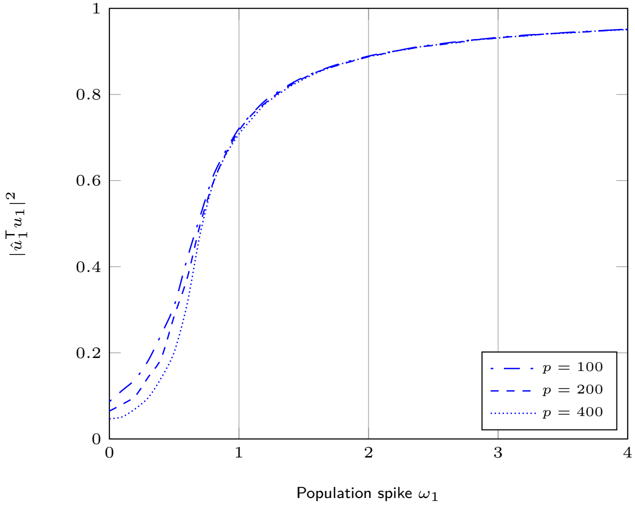


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

Spiked Models

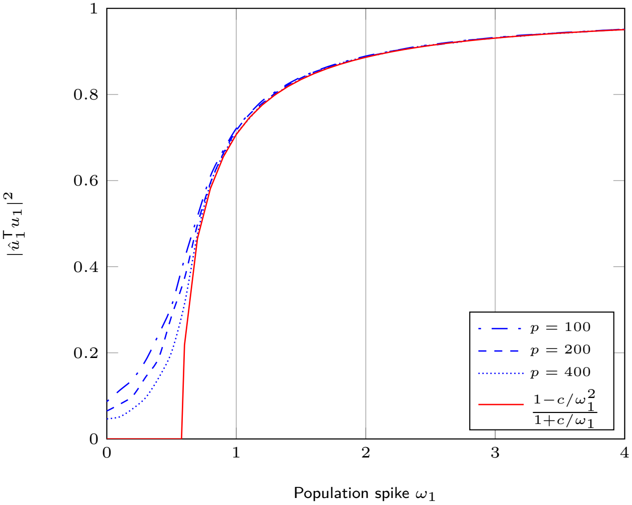


Figure: Simulated versus limiting $|\hat{u}_1^{\mathsf{T}} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^{\mathsf{T}}$, $p/n = 1/3$, varying ω_1.

Theorem (Fluctuations of Eigenvalues **[Baik,BenArous,Péché'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,*
- $C_p = I_p + P$*,* $P = \sum_{i=1}^{K} \omega_i u_i u_i^\mathsf{T}$*,* $\omega_1 > \ldots > \omega_K > 0$ *(*$K \geq 0$*).*

Theorem (Fluctuations of Eigenvalues [Baik,BenArous,Péché'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,*
- $C_p = I_p + P$, $P = \sum_{i=1}^{K} \omega_i u_i u_i^{\mathsf{T}}$, $\omega_1 > \ldots > \omega_K > 0$ *(*$K \geq 0$*).*

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

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

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

Theorem (Fluctuations of Eigenvalues **[Baik,BenArous,Péché'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,*
- $C_p = I_p + P$*,* $P = \sum_{i=1}^K \omega_i u_i u_i^\mathsf{T}$, $\omega_1 > \ldots > \omega_K > 0$ *(*$K \geq 0$*).*

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

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

$$p^{\frac{2}{3}} \frac{\lambda_1 - (1+\sqrt{c})^2}{(1+\sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T,$$ *(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

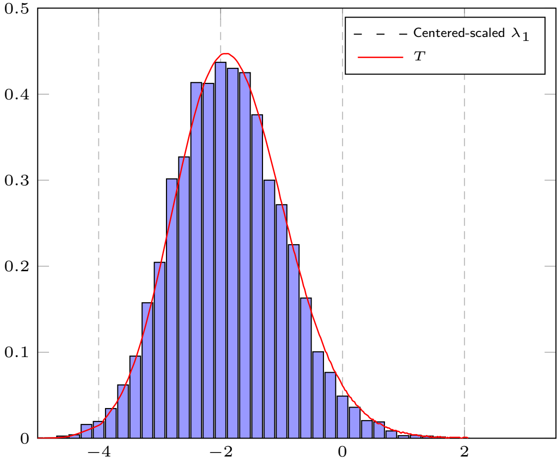


Figure: Distribution of $p^{\frac{2}{3}} c^{-\frac{1}{2}} (1+\sqrt{c})^{-\frac{4}{3}} \left[\lambda_1(\frac{1}{n} X_p X_p^{\mathsf{T}}) - (1+\sqrt{c})^2\right]$ versus real Tracy–Widom (T), $p = 500$, $n = 1500$.

Similar results for multiple matrix models:

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

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* $n \to \infty$*,*

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

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

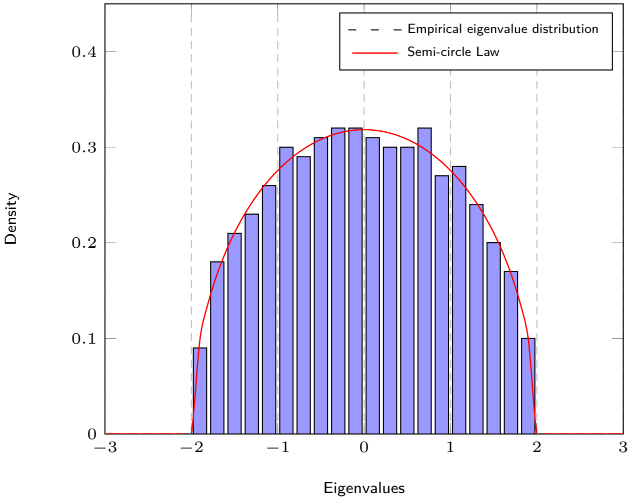


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for $n = 500$

The Circular law

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 $n \to \infty$,

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

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

The Circular law

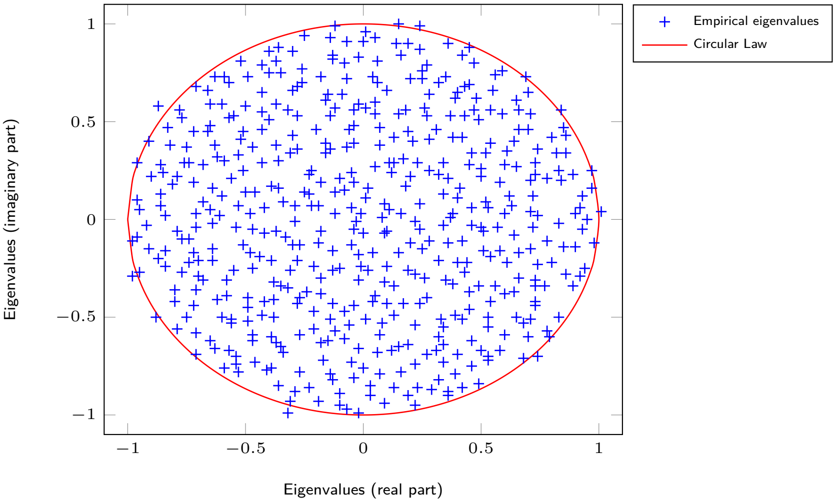


Figure: Eigenvalues of X_n with i.i.d. standard Gaussian entries, for $n = 500$.

Bibliographical references: Maths Book and Tutorial References I

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.

Outline

From classical applications...

Large range of 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.

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.

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.

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.

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!)

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...

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 $x_{l+1} = \sigma(W x_l)$, non-linear features, etc.

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.

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 $x_{l+1} = \sigma(W x_l)$, 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?

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**

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**

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.

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

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**,

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

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?

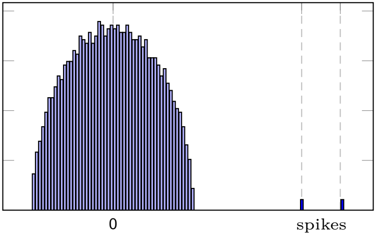
- we will see that **small-dimensional intuitions dramatically fail**
- some **classical and widely-used algorithms become ineffective**
- **BUT** random matrix theory provides a renewed understanding.

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

Outline

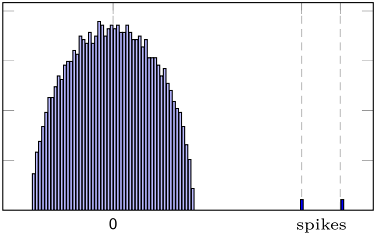
Reminder on Spectral Clustering Methods

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



Reminder on Spectral Clustering Methods

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

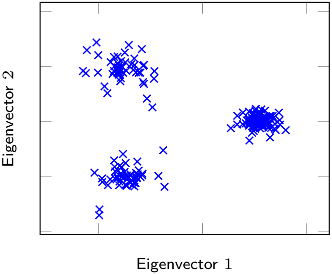


$\Downarrow$ **Eigenvectors** $\Downarrow$
(in practice, shuffled)


Eigenv. 1
Eigenv. 2

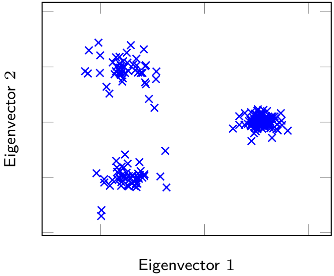
Reminder on Spectral Clustering Methods

⇓ **ℓ-dimensional representation** ⇓
(shuffling no longer matters)



Reminder on Spectral Clustering Methods

⇓ **ℓ-dimensional representation** ⇓
(shuffling no longer matters)



⇓
EM or k-means clustering.

Outline

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

- Kernel spectral clustering based on kernel matrix

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

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

- Kernel spectral clustering based on kernel matrix

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

- Usually, $\kappa(x, y) = f(x^\mathsf{T} y)$ or $\kappa(x, y) = f(\|x - y\|^2)$

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

- Kernel spectral clustering based on kernel matrix

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

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

Kernel Spectral Clustering

Problem Statement

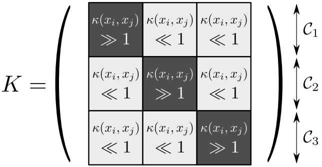
- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

- Kernel spectral clustering based on kernel matrix

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

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

Intuition (from small dimensions)



- K essentially low rank with class structure in eigenvectors.

Kernel Spectral Clustering

Problem Statement

- Dataset $x_1, \ldots, x_n \in \mathbb{R}^p$
- Objective: "cluster" data in k similarity classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$.

- Kernel spectral clustering based on kernel matrix

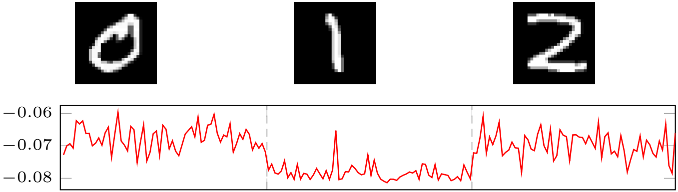
$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^{n}$$

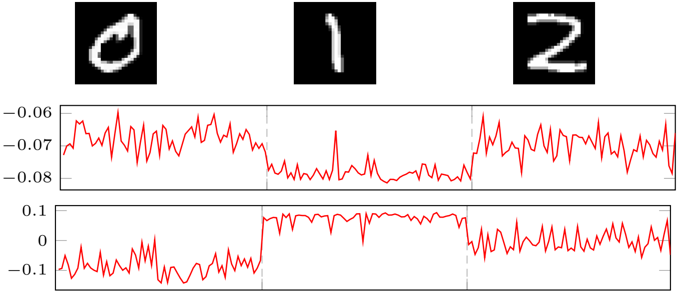
- Usually, $\kappa(x, y) = f(x^\mathsf{T} y)$ or $\kappa(x, y) = f(\|x - y\|^2)$
- Refinements:
  - instead of K, use $D - K$, $I_n - D^{-1}K$, $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

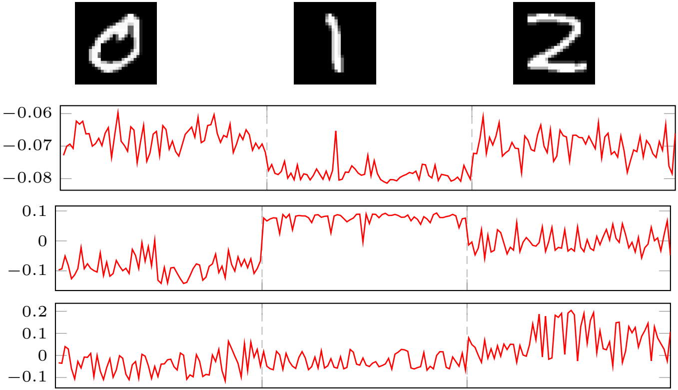
$$K = \begin{pmatrix} \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 \end{pmatrix} \begin{matrix} \updownarrow \mathcal{C}_1 \\ \updownarrow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

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


−0.06
−0.07
−0.08



−0.06
−0.07
−0.08
0.1
0
−0.1


−0.06
−0.07
−0.08
0.1
0
−0.1
0.2
0.1
0
−0.1

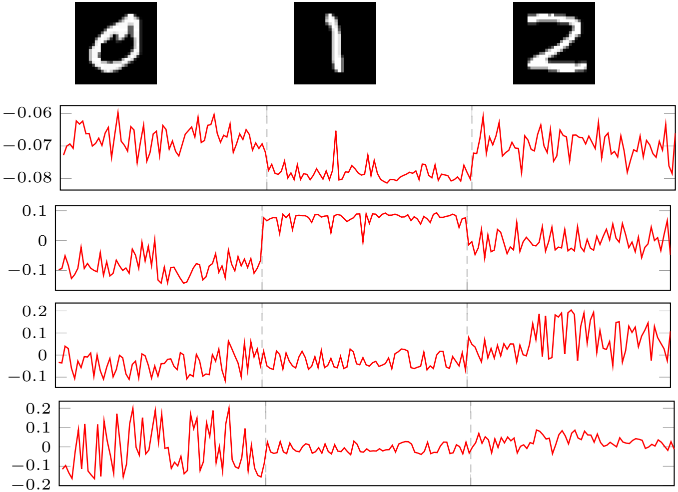


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

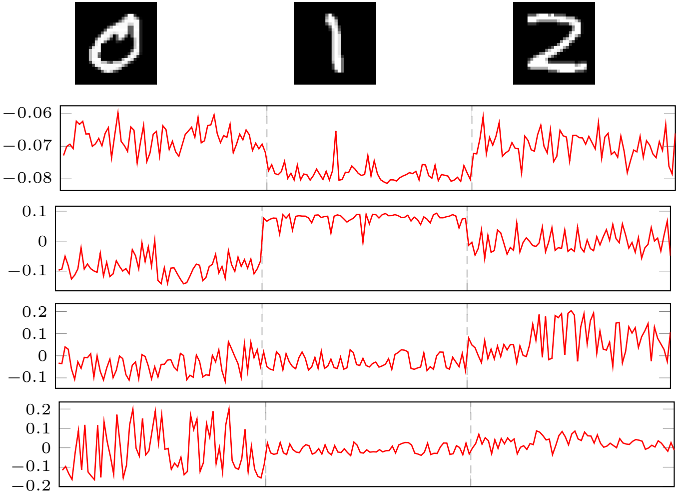


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

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

Model and Assumptions

Gaussian mixture model:

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

Model and Assumptions

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
2. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ *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*

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

Model and Assumptions

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
2. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ *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

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

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
2. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ *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*

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

For 2 classes, this is

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

Remark: [Neyman–Pearson optimality]

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

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 ($f(\frac{1}{p}x_i^\mathsf{T}x_j)$ simpler).

Kernel Matrix:

- Kernel matrix of interest:

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

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

- We study the normalized Laplacian:

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

with $d = K1_n$, $D = \text{diag}(d)$.
(more stable both theoretically and in practice)

Random Matrix Equivalent

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

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

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

Random Matrix Equivalent

- **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 $\tau=\frac{1}{p}\operatorname{tr}C^\circ$.

$\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

Random Matrix Equivalent

- **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 $\tau=\frac{1}{p}\operatorname{tr}C^\circ$.
  $\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

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

$$K=\underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)}+\underbrace{\sqrt{n}K_1}_{\text{low rank, }O_{\|\cdot\|}(\sqrt{n})}+\underbrace{K_2}_{\text{informative terms, }O_{\|\cdot\|}(1)}$$

Random Matrix Equivalent

- **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 $\tau=\frac{1}{p}\mathrm{tr}\, C^\circ$.
$\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

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

$$K=\underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)}+\underbrace{\sqrt{n}K_1}_{\text{low rank, }O_{\|\cdot\|}(\sqrt{n})}+\underbrace{K_2}_{\text{informative terms, }O_{\|\cdot\|}(1)}$$

Clearly not the (small dimension) expected behavior.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Fundamental conclusions:

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

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

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

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

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

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

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

Fundamental conclusions:

- asymptotic **kernel impact** only through $f'(\tau)$ and $f''(\tau)$, **that's all!**

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

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

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

et $W=[w_1,\ldots,w_n]\in\mathbb{R}^{p\times n}$ ($x_i=\mu_a+w_i$), $P=I_n-\frac{1}{n}1_n1_n^{\mathsf{T}}$,

$$J=[j_1,\ldots,j_k],\ j_a^{\mathsf{T}}=(0\ldots0,1_{n_a},0,\ldots,0)$$

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

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

Fundamental conclusions:

- asymptotic **kernel impact** only through $f'(\tau)$ and $f''(\tau)$, **that's all!**
- spectral clustering **reads** $M^{\mathsf{T}}M$, tt^{T} and T, **that's all!**

Isolated eigenvalues: Gaussian inputs

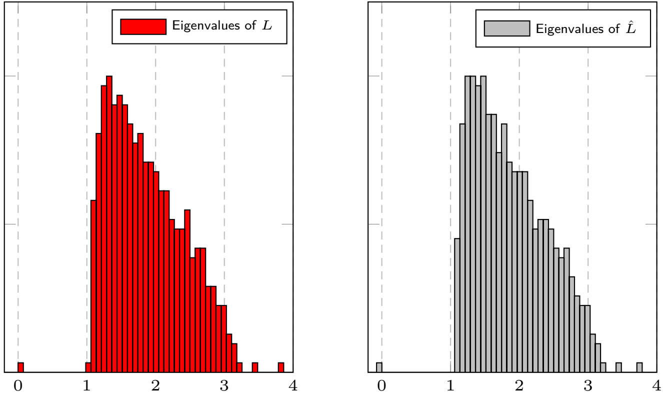


Figure: Eigenvalues of L and $\hat{L}$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\boldsymbol{\delta}_{aj}$, $C_a = (1 + 2(a-1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Theoretical Findings versus MNIST

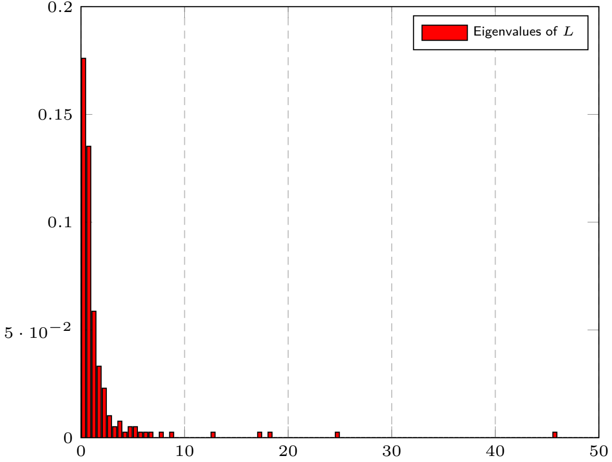


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

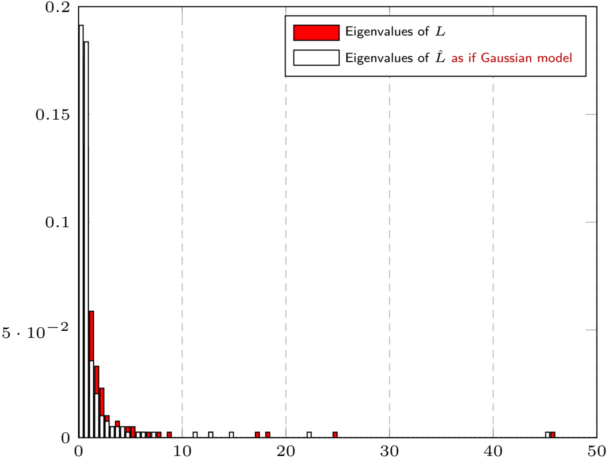


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

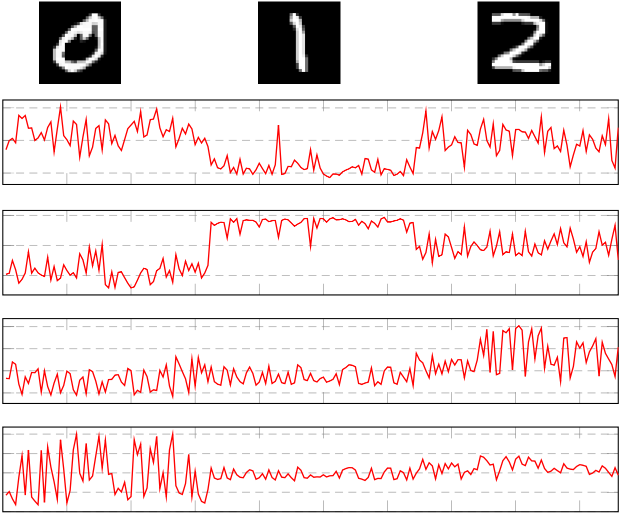


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

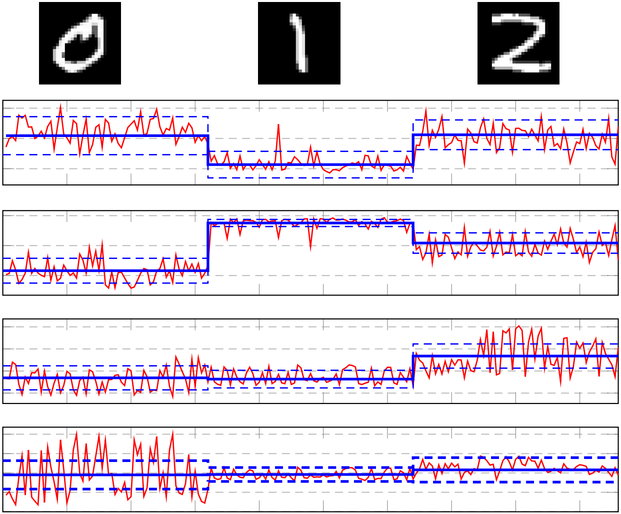


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

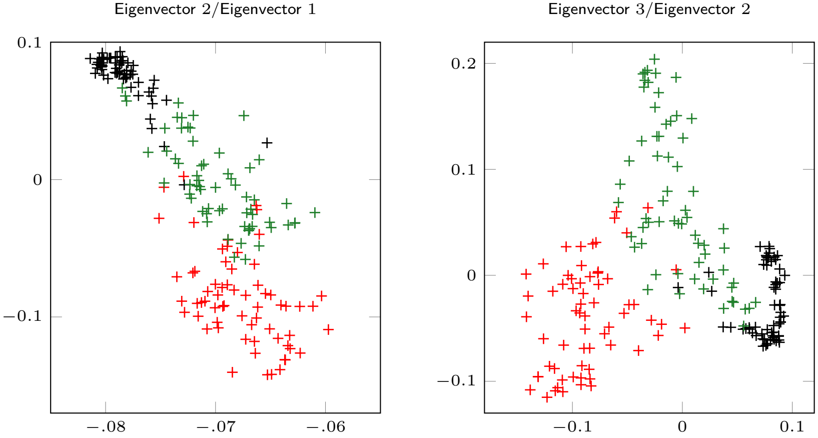


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

Theoretical Findings versus MNIST

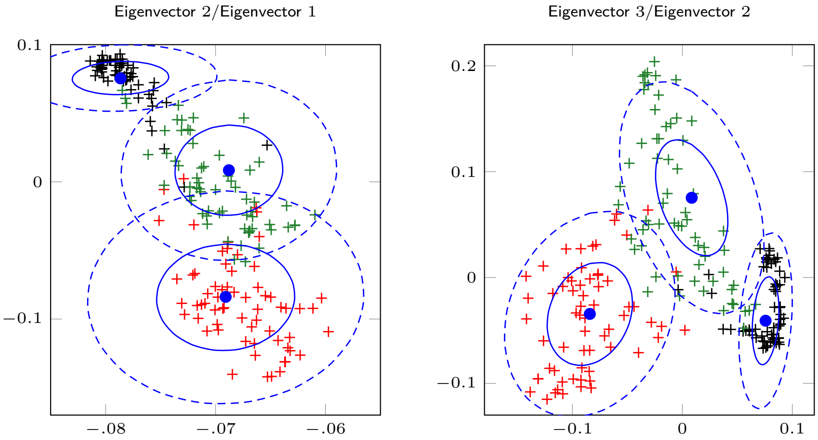


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

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

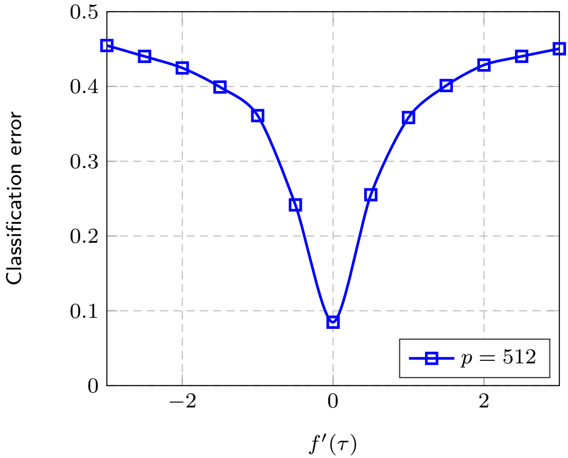


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

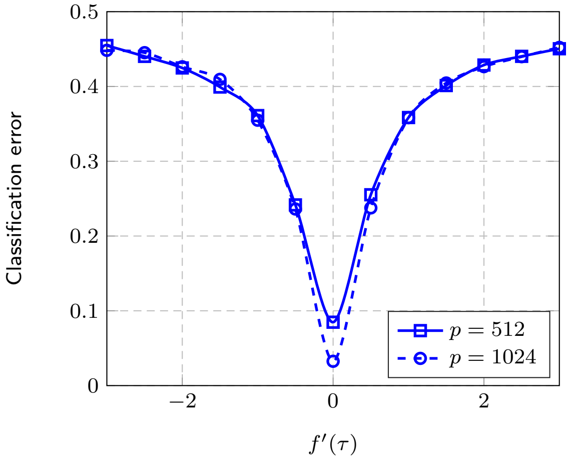


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

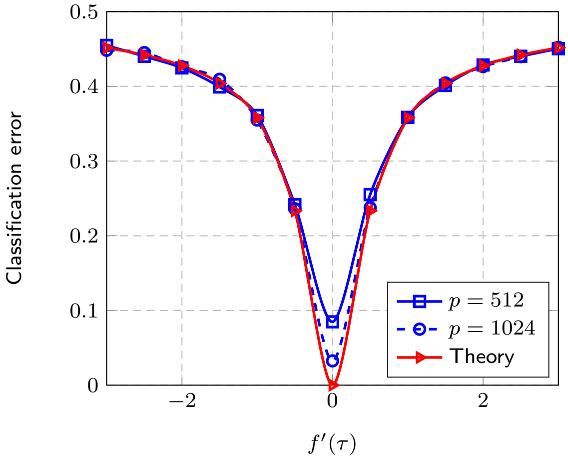


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

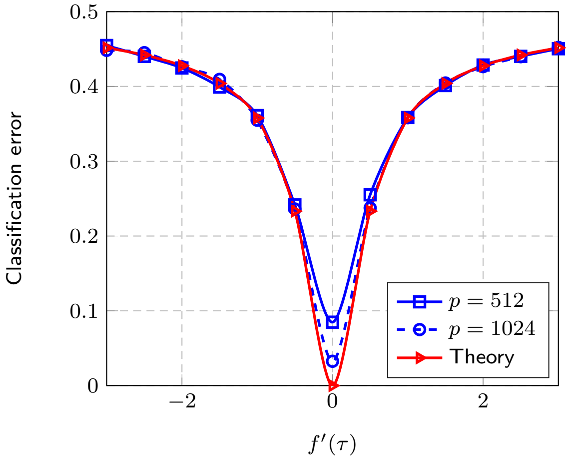


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

Outline

Position of the Problem

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

Position of the Problem

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

Method:

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

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

Position of the Problem

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

Method:

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

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

- Performance of $L=nD^{-\frac{1}{2}}\left(K-\frac{1_n1_n^\mathsf{T}}{1_n^\mathsf{T}D1_n}\right)D^{-\frac{1}{2}}$, with

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

  in the regime $n,p\to\infty$.
  (alternatively, we can ask $\frac{1}{p}$*tr* $C_i=1$ *for all* $1\leq i\leq k$*)*

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

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

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

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

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

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

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p}\operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, 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,

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

exhibits phase transition phenomenon

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

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

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, 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,

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

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

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

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

The case $f'(2) = 0$

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

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

The case $f'(2) = 0$

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

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

(in this regime, previous kernels clearly fail)

Remark: [Neyman–Pearson optimality]

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

The case $f'(2) = 0$

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

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

(in this regime, previous kernels clearly fail)

Remark: [Neyman–Pearson optimality]

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

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

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

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

Then, under Assumptions 2b,

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

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

The case $f'(2) = 0$

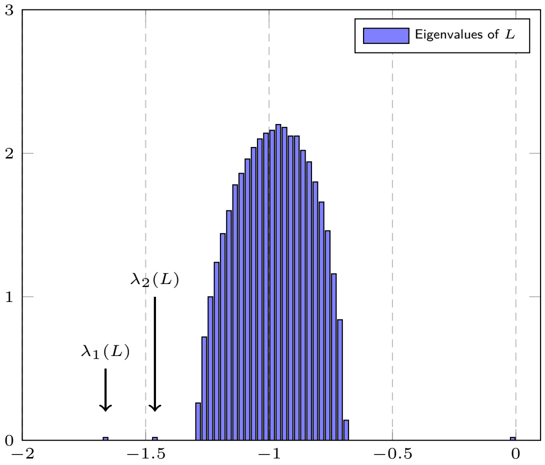
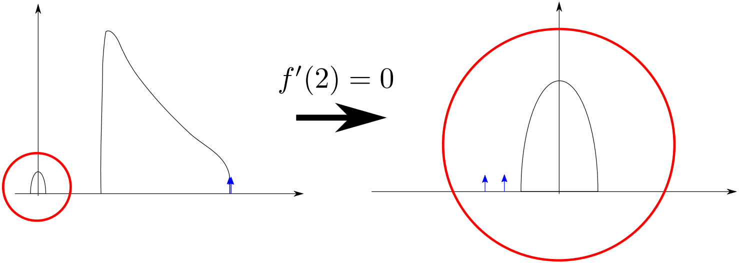


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^\mathsf{T}$, $W_i \in \mathbb{R}^{p\times(p/8)}$ of i.i.d. $\mathcal{N}(0,1)$ entries, $f(t) = \exp(-(t-2)^2)$.

⇒ No longer a Marcenko–Pastur like bulk, but rather a semi-circle bulk!

The case $f'(2) = 0$



The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ

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)

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.

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

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

with μ the semi-circle distribution

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

The case $f'(2) = 0$

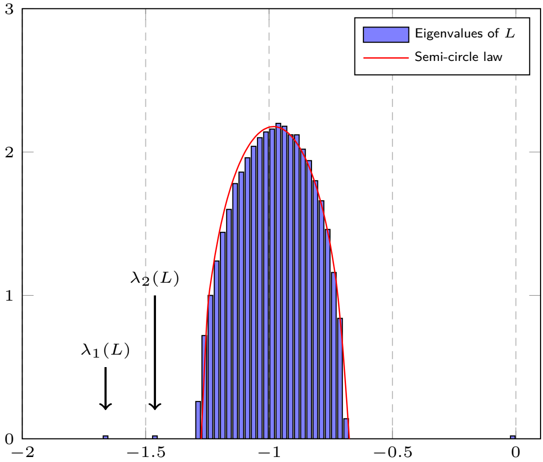


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^{\mathsf{T}}$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0, 1)$ entries, $f(t) = \exp(-(t-2)^2)$.

The case $f'(2) = 0$

Denote now

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

The case $f'(2) = 0$

Denote now

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

Theorem (Isolated Eigenvalues)

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

$$\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}.$$

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 $j_a = [0_{n_1}^\mathsf{T}, \ldots, 1_{n_a}^\mathsf{T}, \ldots, 0_{n_k}^\mathsf{T}]^\mathsf{T}$, $(w_i^a)^\mathsf{T} j_a = 0$, $\operatorname{supp}(w_i^a) = \operatorname{supp}(j_a)$, $\|w_i^a\| = 1$.
Then, under Assumptions 1–2b,

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

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

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

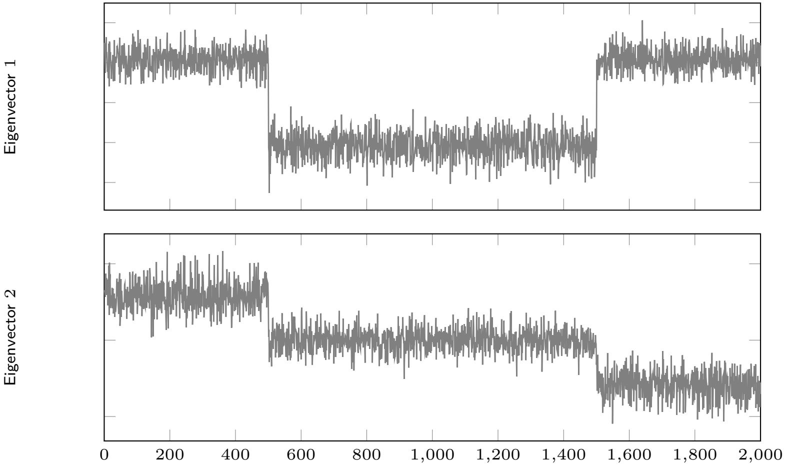


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

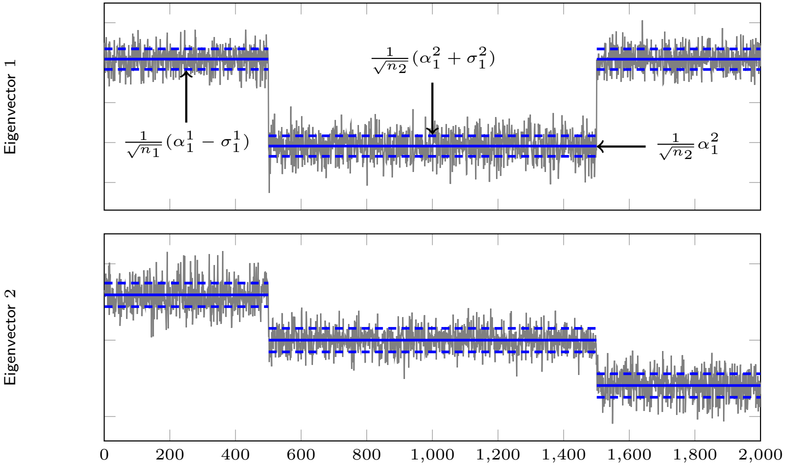


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

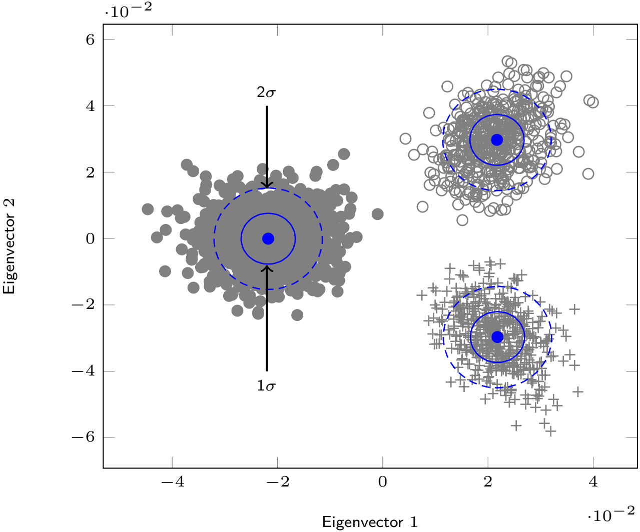


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

Application: Multiple-source Subspace Clustering

Setting.

- p dimensional vector observations.
- n data sources.
- $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a$.

Setting.

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

Application: Multiple-source Subspace Clustering

Setting.

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

Objective. Cluster sources based on spanned subspace.

Application: Multiple-source Subspace Clustering

Setting.

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

Objective. Cluster sources based on spanned subspace.

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

Application: Multiple-source Subspace Clustering

Setting.

- p dimensional vector observations.
- n data sources.
- $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a$.
- T independent observations $x_i^{(1)}, \ldots, x_i^{(T)}$ 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 $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).

Application: Multiple-source Subspace Clustering

Setting.

- p dimensional vector observations.
- n data sources.
- $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a$.
- T independent observations $x_i^{(1)}, \ldots, x_i^{(T)}$ 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 $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$

Application: Multiple-source Subspace Clustering

Setting.

- p dimensional vector observations.
- n data sources.
- $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a$.
- T independent observations $x_i^{(1)}, \ldots, x_i^{(T)}$ 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 $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each i, create $\tilde{u}_i = \frac{1}{T}(I_n \otimes 1_T^\mathsf{T})u_i$, i.e., average eigenvectors along time.

Application: Multiple-source Subspace Clustering

Setting.

- p dimensional vector observations.
- n data sources.
- $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a$.
- T independent observations $x_i^{(1)}, \ldots, x_i^{(T)}$ 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 $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each i, create $\tilde{u}_i = \frac{1}{T}(I_n \otimes 1_T^\mathsf{T})u_i$, i.e., average eigenvectors along time.
4. Perform k-class clustering on vectors $\tilde{u}_1, \ldots, \tilde{u}_\kappa$.

Application Example: Massive MIMO UE Clustering

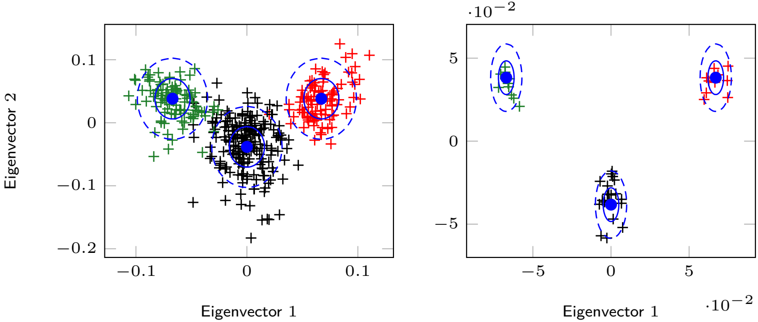


Figure: **Massive MIMO application**: Leading two eigenvectors before (left figure) and after (right figure) T-averaging. Setting: $p = 400$, $n = 40$, $T = 10$, $k = 3$, $c_1 = c_3 = 1/4$, $c_2 = 1/2$, angular spread model with angles $-\pi/30 \pm \pi/20$, $0 \pm \pi/20$, and $\pi/30 \pm \pi/20$. Kernel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

Application Example: Massive MIMO UE Clustering

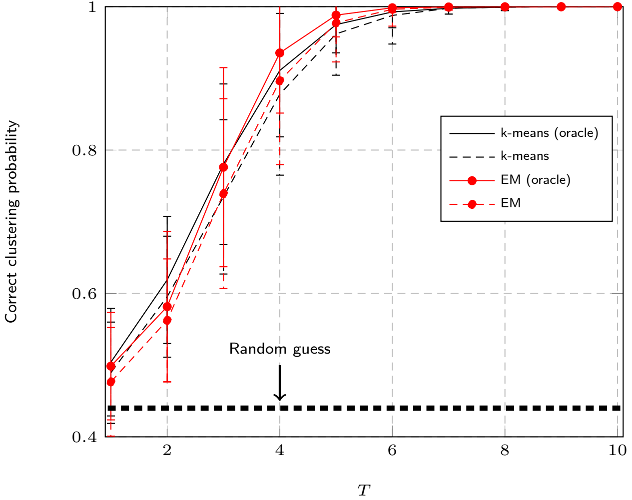


Figure: Overlap for different T, using the k-means or EM starting from actual centroid solutions (oracle) or randomly.

Application Example: Massive MIMO UE Clustering

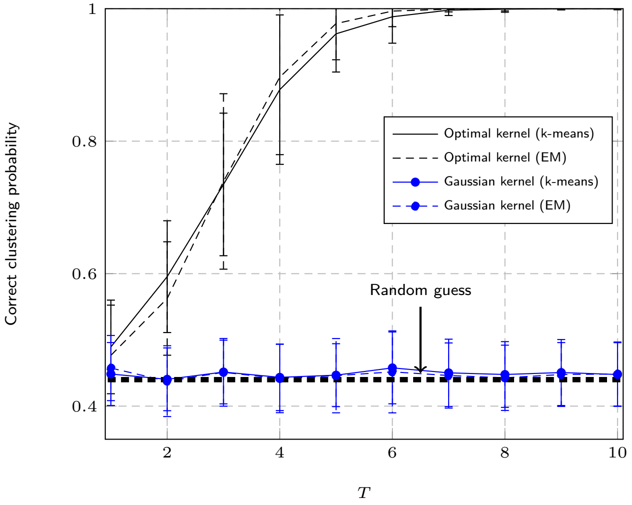


Figure: Overlap for optimal kernel $f(t)$ (here $f(t) = \exp(-(t-2)^2)$) and Gaussian kernel $f(t) = \exp(-t^2)$, for different T, using the k-means or EM.

Outline

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **Model type:** Marčenko–Pastur + spikes.

Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **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 $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

$\longrightarrow$ **Model type:** smaller order semi-circle law + spikes.

Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **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 $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

$\longrightarrow$ **Model type:** smaller order semi-circle law + spikes.

Jointly optimal solution:

- evenly weighing Marčenko–Pastur and semi-circle laws

Optimal growth rates and optimal kernels

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **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 $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

$\longrightarrow$ **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 $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
2. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ *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*

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

New assumption setting

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

Assumption (Optimal Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
2. **Mean scaling**: *with* $\mu^\circ \triangleq \sum_{a=1}^{k} \frac{n_a}{n} \mu_a$ *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*

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

Kernel:

- For technical simplicity, we consider

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

i.e., τ replaced by 0.

Theorem

As $n \to \infty$,

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

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

$$\hat{\mathcal{K}} = \alpha P W^{\mathsf{T}} W P + \beta P \Phi P + U A U^{\mathsf{T}}$$

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

$$U = \left[\frac{J}{\sqrt{p}}, P W^{\mathsf{T}} M \right]$$

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

Theorem

As $n \to \infty$,

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

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

$$\hat{\mathcal{K}} = \alpha P W^{\mathsf{T}} W P + \beta P \Phi P + U A U^{\mathsf{T}}$$

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

$$U = \left[\frac{J}{\sqrt{p}}, P W^{\mathsf{T}} M \right]$$

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

Role of α, β:

- Weighs Marčenko–Pastur versus semi-circle parts.

Theorem (Eigenvalues Bulk)

As $p \to \infty$,

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

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

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

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

Limiting eigenvalue distribution

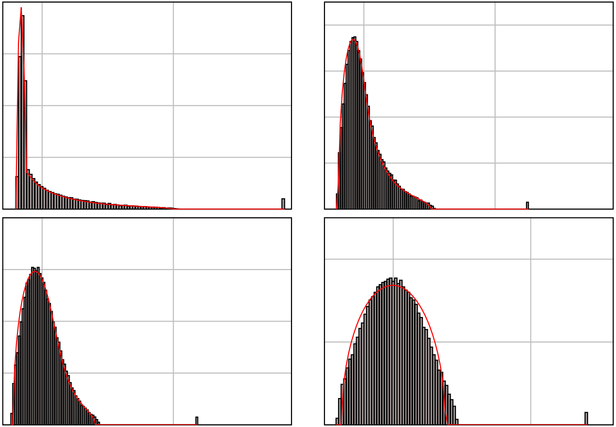


Figure: Eigenvalues of K (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\boldsymbol{\mu}_i = 3\boldsymbol{\delta}_i$, $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$. **(Top left)**: $\alpha = 8, \beta = 1$, **(Top right)**: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

- MNIST is "means-dominant" but not that much!

Datasets	$\\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\\|^2$	$\frac{1}{\sqrt{p}}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (digits 1, 7)	613	1990	71.1
MNIST (digits 3, 6)	441	1119	39.9
MNIST (digits 3, 8)	212	652	23.5

Asymptotic performances: MNIST

- MNIST is "means-dominant" but not that much!

Datasets	$\|\mu_1^\circ - \mu_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (digits $1,7$)	613	1990	71.1
MNIST (digits $3,6$)	441	1119	39.9
MNIST (digits $3,8$)	212	652	23.5

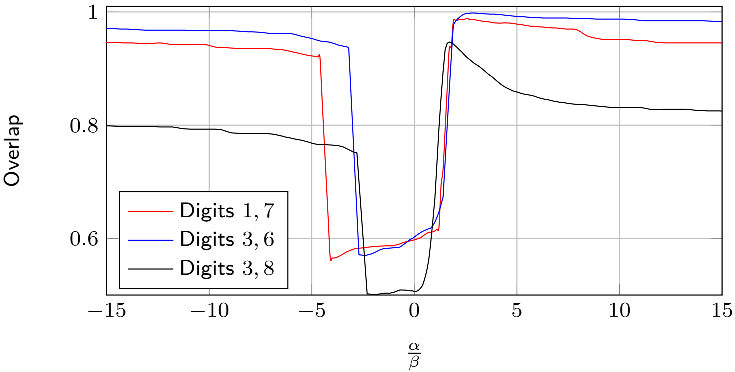


Figure: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$.

- EEG data are "variance-dominant"

Datasets	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (sets A, E)	2.4	10.9	1.1

- EEG data are "variance-dominant"

Datasets	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (sets A, E)	2.4	10.9	1.1

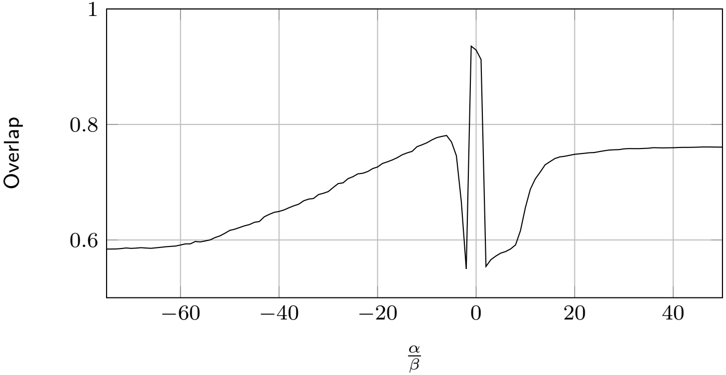


Figure: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$.

Outline

Laplacian Regularization

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $\mathcal{C}_k$).

Laplacian Regularization

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $F_{ia} = \boldsymbol{\delta}_{\{x_i \in \mathcal{C}_a\}}$, for all labelled x_i.

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $F_{ia} = \boldsymbol{\delta}_{\{x_i \in \mathcal{C}_a\}}$, 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 $D = \operatorname{diag}\{K1_n\}$.

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $F_{ia} = \boldsymbol{\delta}_{\{x_i \in \mathcal{C}_a\}}$, 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

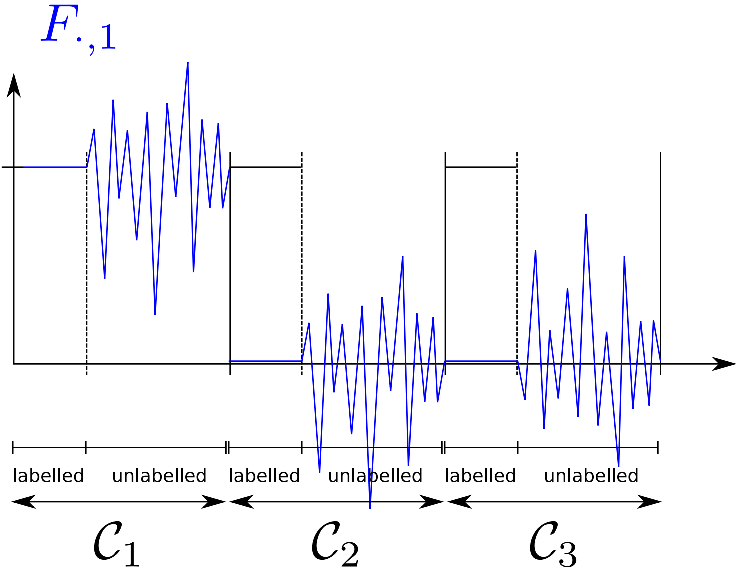


Figure: Typical expected performance output

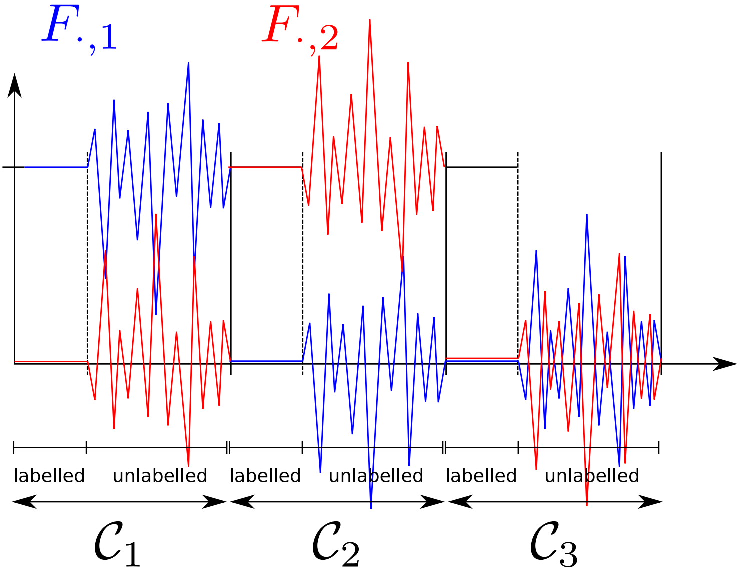


Figure: Typical expected performance output

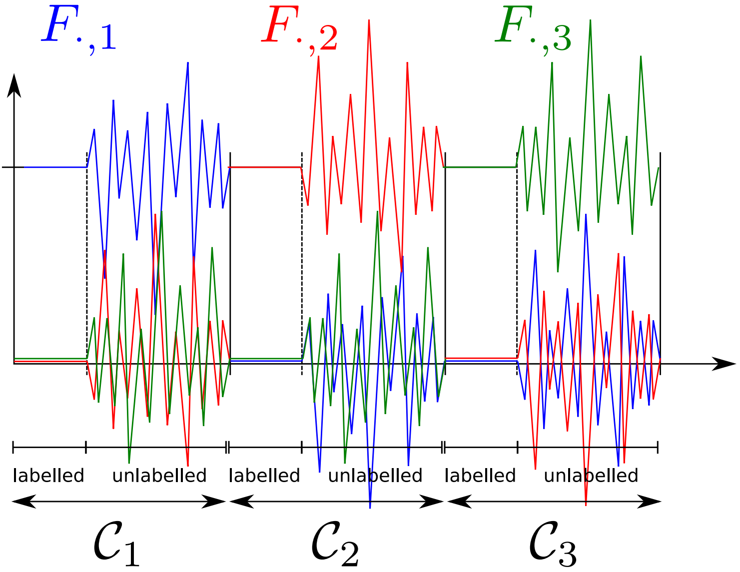


Figure: Typical expected performance output

MNIST Data Example

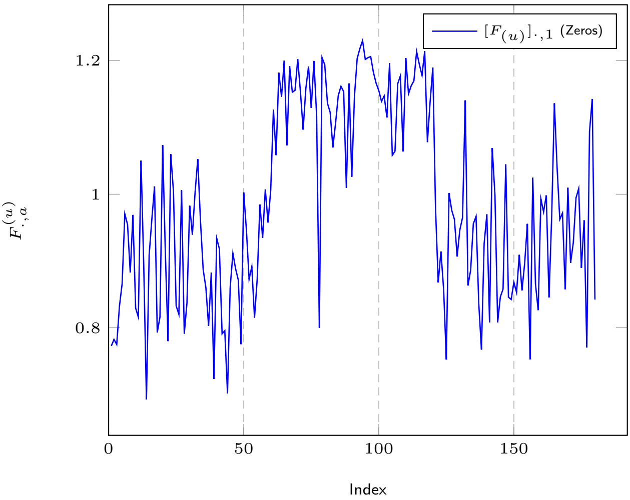


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

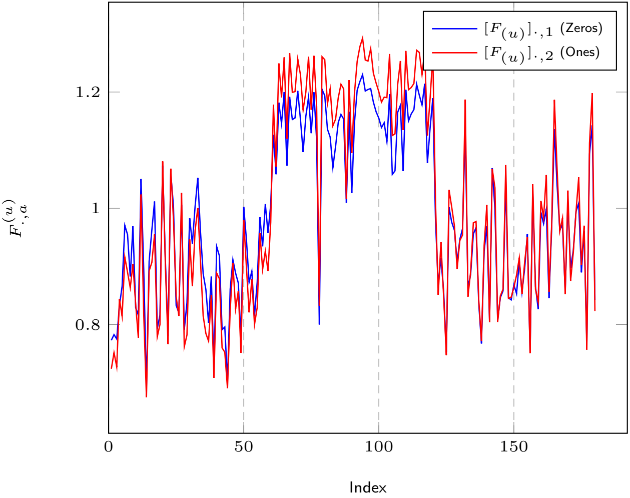


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

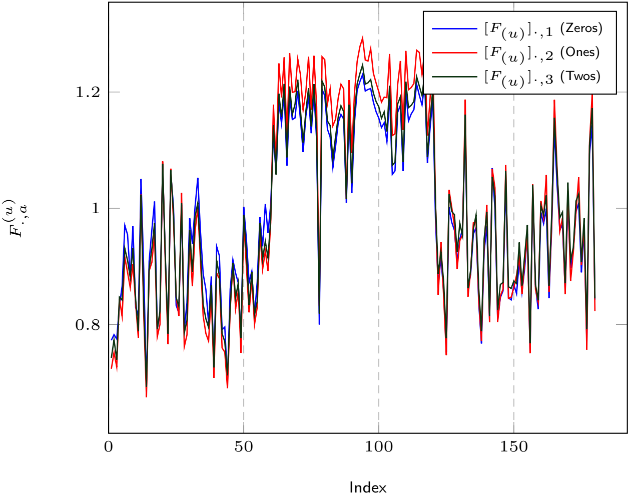


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

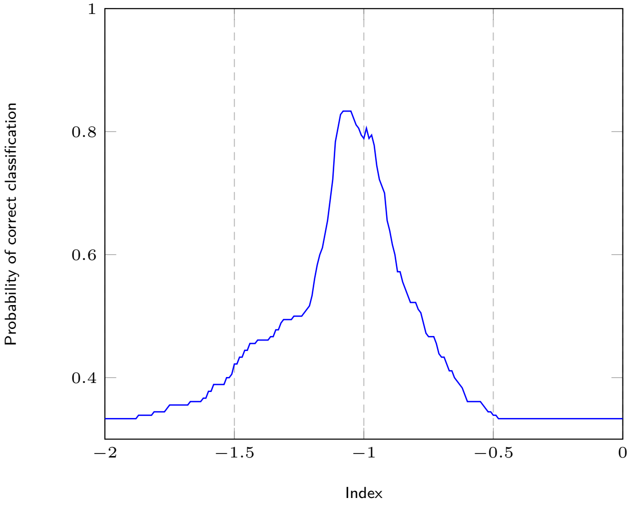


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

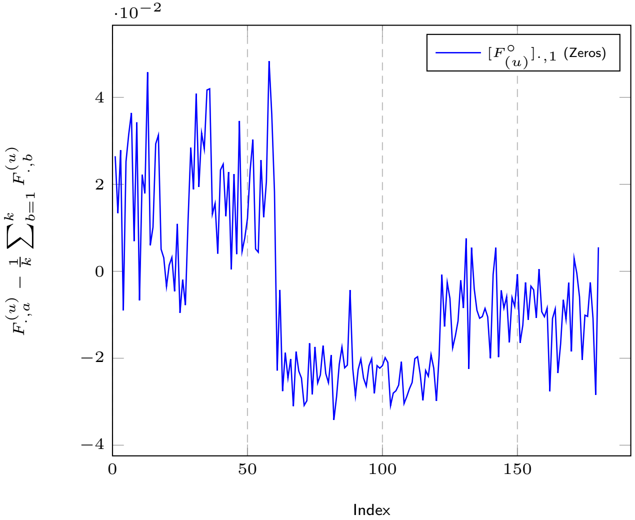


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = -1$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

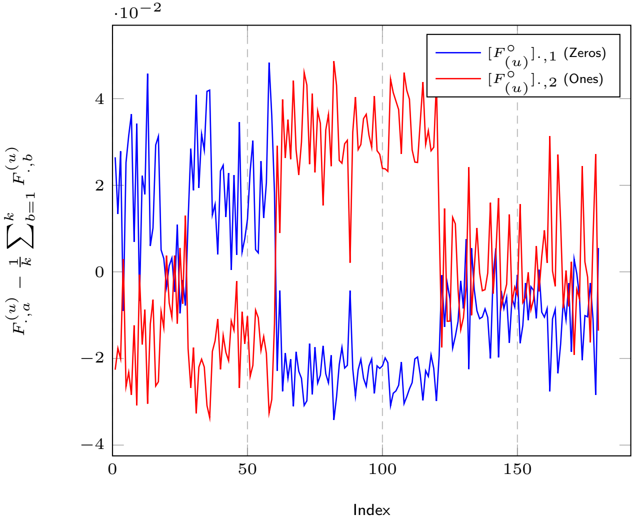


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = -1$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

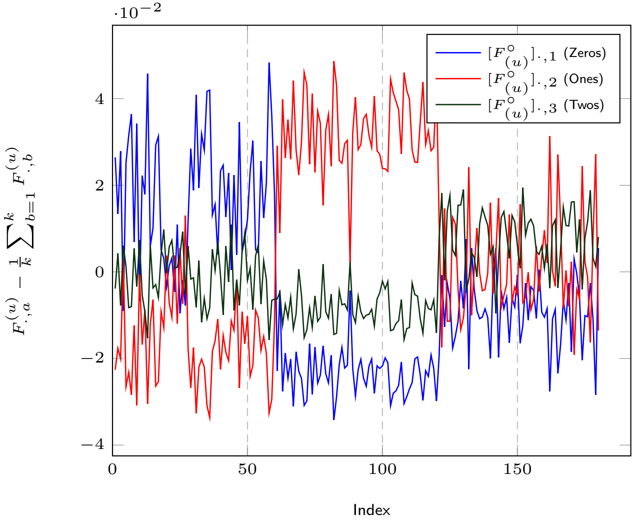


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = -1$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Motivation

Empirical observations:

- Troubling *flat* classification scores!

Motivation

Empirical observations:

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

Motivation

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

Motivation

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

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

Motivation

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

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

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

Motivation

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

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

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

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

Main Results

We can show that, for x_i unlabelled,

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

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 $c_{[l]2} \neq c_{[l]1}$.

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 $\mathcal{C}_2$).

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 $\mathcal{C}_2$).

$$\Downarrow$$

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

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 $\mathcal{C}_2$).

$$\Downarrow$$

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

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

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 $\mathcal{C}_2$).

$$\Downarrow$$

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

Consequence: All f_i have the same sign if $t_2 \neq t_1$.
Amendment: Take $\alpha = -1 + \frac{\beta}{\sqrt{p}}$, $\beta = O(1)$.

$$\Downarrow$$

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

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

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

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

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

Performance: Theoretical versus Empirical

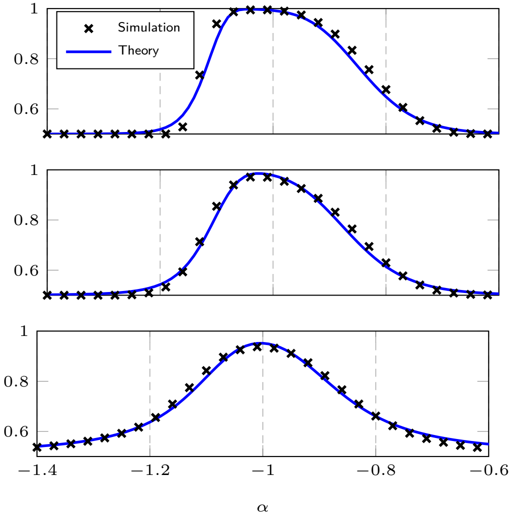


Figure: Theoretical and empirical accuracy as a function of α for 2-class MNIST data (**top:** digits (0,1), **middle:** digits (1,7), **bottom:** digits (8,9)), $n = 1024$, $p = 784$, $n_{[l]}/n = 1/16$, $n_{[u]1} = n_{[u]2}$, Gaussian kernel. Averaged over 50 iterations.

Main Results

$$\Downarrow$$

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

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

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

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

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

Main Results

$$\Downarrow$$

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

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

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

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

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 $c_{[u]}$

$$\Downarrow$$

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

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

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

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

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!

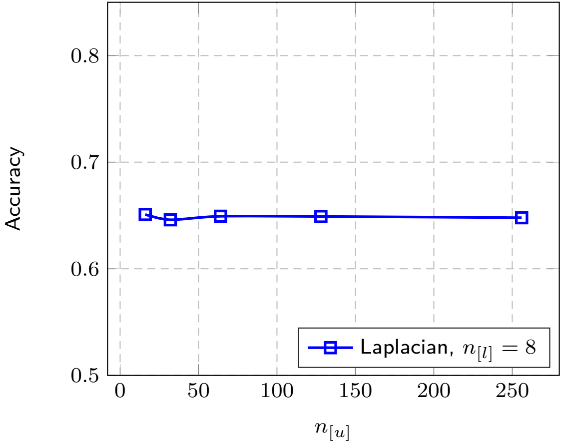


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

MNIST Data Example

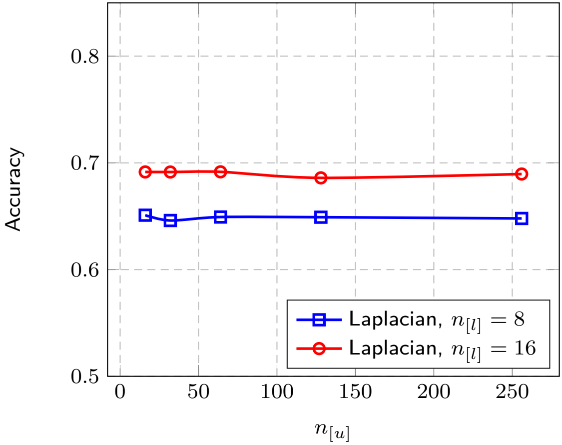


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

Main Results

$$\Downarrow$$

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

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

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

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

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!

$$\Downarrow$$

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

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

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

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

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.

Link between scores flatness and non-expressive unlabelled data:

Resurrecting SSL by centering

Link between scores flatness and non-expressive unlabelled data:

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

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

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

Link between scores flatness and non-expressive unlabelled data:

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

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

  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.

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

$$\boxed{\tilde{K} = PKP}, \quad P = I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}.$$

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).

Resurrecting SSL by centering

Solution:

- **"Recenter" K to kill flattening**, i.e., use

$$\boxed{\tilde{K} = PKP}, \quad P = I_n - \frac{1}{n} 1_n 1_n^{\mathsf{T}}.$$

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

Solution:

- **"Recenter" K to kill flattening**, i.e., use

$$\boxed{\tilde{K} = PKP}, \quad P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}.$$

The recentering imposes $\mathbb{E}\{\hat{K}_{a_1a_2}\} + \mathbb{E}\{\hat{K}_{a_1b_1}\} = 0$ (in the case of balanced datasets).

- Since $\mathbb{E}\{\hat{K}_{a_1a_2}\} - \mathbb{E}\{\hat{K}_{a_1b_1}\} = \mathbb{E}\{K_{a_1a_2}\} - \mathbb{E}\{K_{a_1b_1}\} = \epsilon$, $\mathbb{E}\{\hat{K}_{a_1a_2}\} = -\mathbb{E}\{\hat{K}_{a_1b_1}\} = \epsilon/2$.

- Hence,

$$\boxed{(\mathbb{E}\{\hat{K}_{a_1a_2}\} - \mathbb{E}\{\hat{K}_{a_1b_1}\})/|\mathbb{E}\{\hat{K}_{a_1a_2}\}||\mathbb{E}\{\hat{K}_{a_1b_1}\}| = 4 = O(1)}$$

$$\Downarrow$$

Non flat scores!

Method:

Centered Kernel Regularization

Method:

- Same loss function as Laplacian regularization, but with centered similarities $\tilde{K}_{ij}$.

Method:

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

$$\min_f \sum_{i,j=1}^{n} \tilde{K}_{ij} |f_i - f_j|^2$$
$$\text{s.t.} \quad \|f_{[u]}\| = t$$

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

Method:

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

$$\min_f \sum_{i,j=1}^{n} \tilde{K}_{ij}|f_i - f_j|^2$$
$$\text{s.t.} \quad \|f_{[u]}\| = t$$

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

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

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

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

MNIST Data Example

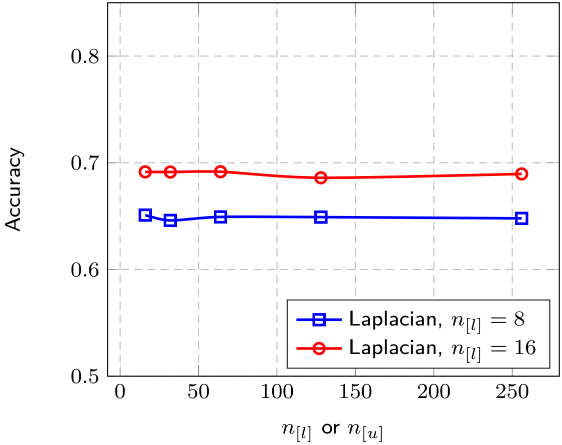


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

MNIST Data Example

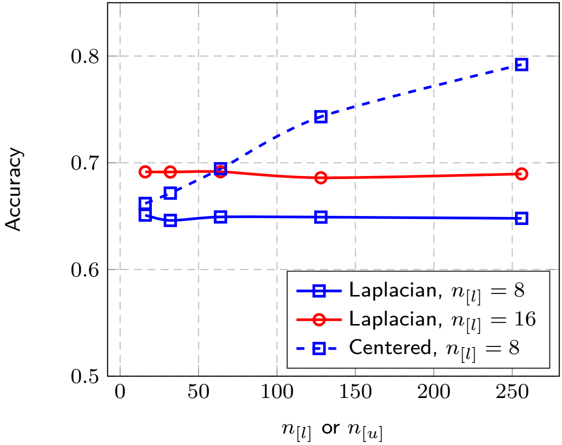


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

MNIST Data Example

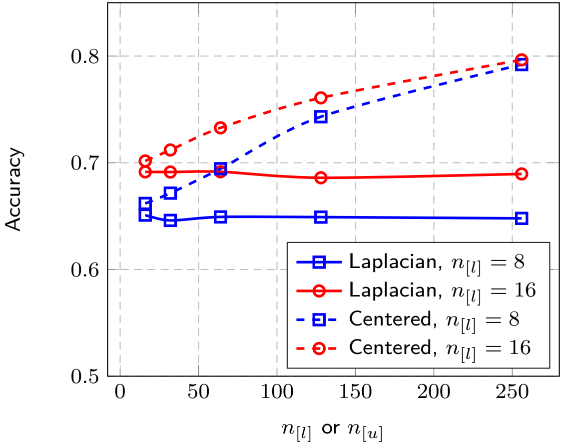


Figure: Classification accuracy as a function of $n_{[u]}$ with fixed $n_{[l]}$ for 2-class MNIST data (8,9), Gaussian kernel. Optimal average results over 200 iterations.

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ with $x_i \in \mathcal{C}_k$)

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ 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.

Theoretical results

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ 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 $c_{[u]}$.

Theoretical results

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ 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.

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ 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 $q = \min\{1, \sqrt{\|\mu\|^4 c_{[u]}}\}$.

Effective learning from labelled and unlabelled data

- $m_1 < 0$ and $m_2 > 0$ for all α. (recall that $m_k = \mathbb{E}\{f_i\}$, $\sigma_k^2 = \mathrm{Var}\{f_i\}$ 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 $n_{[u]}$, $n_{[l]}$

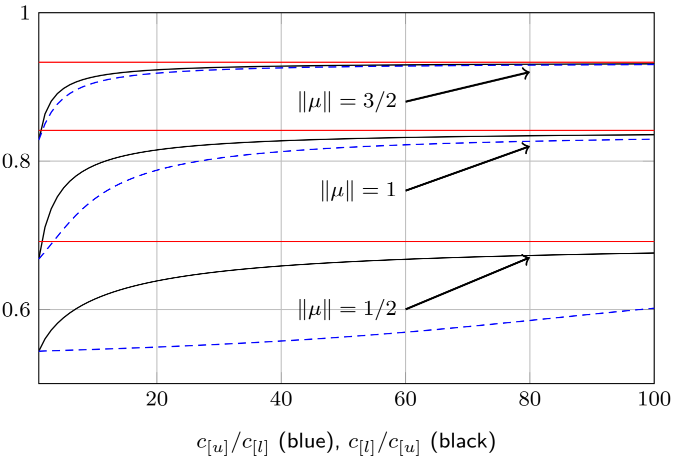


Figure: Correct classification rate, at optimal α, as a function of (i) $n_{[u]}$ for fixed $p/n_{[l]} = 5$ (blue) and (ii) $n_{[l]}$ for fixed $p/n_{[u]} = 5$ (black); $c_1 = c_2 = \frac{1}{2}$; different values for $\|\mu\|$. Comparison to optimal Neyman–Pearson performance for known μ (in red).

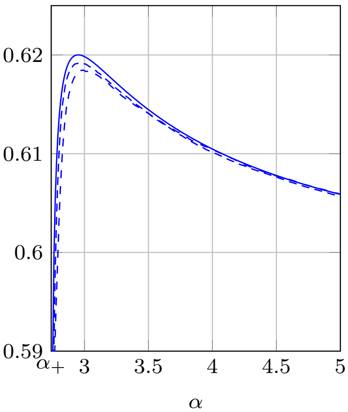


Figure: Theory (solid) versus practice (dashed; from right to left: $n = 400, 1000, 4000$): correct classification probability as a function of α for $c_{[u]} = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, and **left**: $\|\mu\| = 0.75$ (below phase transition); **right**: $\|\mu\| = 1.25$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

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 $n_{[l]} = 10$. 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 $n_{[u]} = 500$.

Outline

Motivation: Feature extraction in machine learning

Learning = Representation + Evaluation + Optimization.[1]

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

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

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:

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

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)

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

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

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

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:

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

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

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

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)

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

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!

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

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?

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

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? ⇒ Sample Covariance Matrix

$$\mathrm{SCM} \equiv \frac{1}{T} X X^{\mathsf{T}}$$

of data $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$.

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

Motivation: Feature extraction in machine learning

$$\text{Learning} = \text{Representation} + \text{Evaluation} + \text{Optimization.}^{1}$$

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

$$\mathrm{SCM} \equiv \frac{1}{T} X X^{\mathsf{T}}$$

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

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

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

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

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

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

Motivation: RMT for random feature maps

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

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

Example:

$$X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$$

Figure: Illustration of random feature maps

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

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

Example:

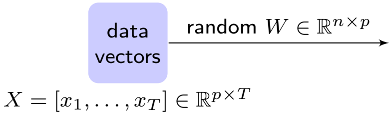


Figure: Illustration of random feature maps

Motivation: RMT for random feature maps

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

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

Example:

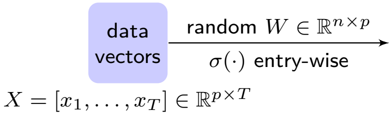


Figure: Illustration of random feature maps

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

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

Example:

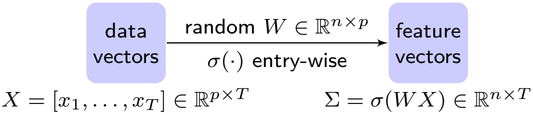


Figure: Illustration of random feature maps

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

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

Example:

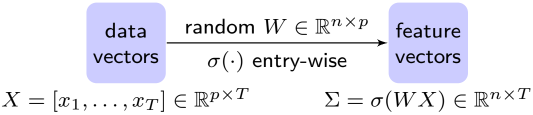


Figure: Illustration of random feature maps

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

$$\mathrm{E}_{\mathrm{train}} = \frac{1}{T}\|y - \beta^{\mathsf{T}}\Sigma\|_F^2 = \frac{\gamma^2}{T} y^{\mathsf{T}} Q^2(-\gamma) y, \quad \mathrm{E}_{\mathrm{test}} = \frac{1}{\hat{T}}\|\hat{y} - \beta^{\mathsf{T}}\hat{\Sigma}\|_F^2$$

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

Motivation: RMT for random feature maps

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

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

Example:

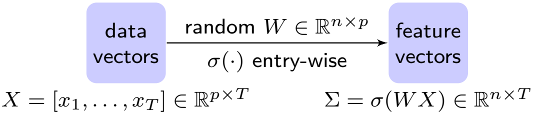


Figure: Illustration of random feature maps

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

$$\mathrm{E}_{\rm train} = \frac{1}{T}\|y - \beta^\mathsf{T}\Sigma\|_F^2 = \frac{\gamma^2}{T} y^\mathsf{T} Q^2(-\gamma) y, \quad \mathrm{E}_{\rm test} = \frac{1}{\hat{T}}\|\hat{y} - \beta^\mathsf{T}\hat{\Sigma}\|_F^2$$

with ridge regressor $\beta \equiv \frac{1}{T}\Sigma\left(G + \gamma I_T\right)^{-1} y^\mathsf{T} = \frac{1}{T}\Sigma Q(-\gamma) y^\mathsf{T}$ 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 $a^\mathsf{T} Q(z) b$ for nonlinear model $\Sigma = \sigma(WX)$!

Handle nonlinearity in RMT: concentration of measure approach

Recall:
For $\sigma(t) = t$, $G = \frac{1}{T} X^\mathsf{T} W^\mathsf{T} W X$ 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,

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

Handle nonlinearity in RMT: concentration of measure approach

Recall:
For $\sigma(t) = t$, $G = \frac{1}{T} X^\mathsf{T} W^\mathsf{T} W X$ 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,

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

Nonlinearity

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

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

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

Recall:
For $\sigma(t) = t$, $G = \frac{1}{T} X^\mathsf{T} W^\mathsf{T} W X$ 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,

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

Nonlinearity

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

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

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

$$Q = \left(\frac{1}{T} \Sigma^\mathsf{T} \Sigma - z I_T \right)^{-1} = \left(\frac{1}{T} \Sigma_{-i}^\mathsf{T} \Sigma_{-i} + \frac{1}{T} \sigma_i \sigma_i^\mathsf{T} - z I_T \right)^{-1}$$

Recall:
For $\sigma(t) = t$, $G = \frac{1}{T} X^\mathsf{T} W^\mathsf{T} W X$ 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,

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

Nonlinearity

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

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

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

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

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

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

- loss of independence between entries

Handle nonlinearity in RMT: concentration of measure approach

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

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

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

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

$\Rightarrow$ extend trace lemma to handle nonlinear case!

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

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

$\Rightarrow$ 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,

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

for some $C, c > 0$ and $\Phi \equiv E_w\left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$ (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 $n, p, T \to \infty$, $p/n \to c_p \in (0, \infty)$ and $T/n \to c_T \in (0, \infty)$,

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

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

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 $n, p, T \to \infty$, $p/n \to c_p \in (0, \infty)$ and $T/n \to c_T \in (0, \infty)$,

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

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

Several remarks:

- (asymptotic) training performance only depends on (the training data X via) the key averaged kernel matrix Φ and the dimension of problem

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 $n, p, T \to \infty$, $p/n \to c_p \in (0, \infty)$ and $T/n \to c_T \in (0, \infty)$,

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

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

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

Performance evaluation of random feature-based ridge regression

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 $n, p, T \to \infty$, $p/n \to c_p \in (0, \infty)$ and $T/n \to c_T \in (0, \infty)$,

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

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

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:

$$\Phi(X) = E_w \left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$$

Computation of averaged kernel Φ

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

$$\Phi(X) = E_w \left[\sigma(X^\mathsf{T} w) \sigma(w^\mathsf{T} X) \right]$$

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

$$\Phi_{i,j} = (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(w^\mathsf{T} x_i) \sigma(w^\mathsf{T} x_j) dw$$

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

$$\Phi(X) = E_w \left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$$

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

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

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

$$\Phi(X) = E_w \left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$$

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

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

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

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

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

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

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

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

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

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

Performance of random feature-based ridge regression:

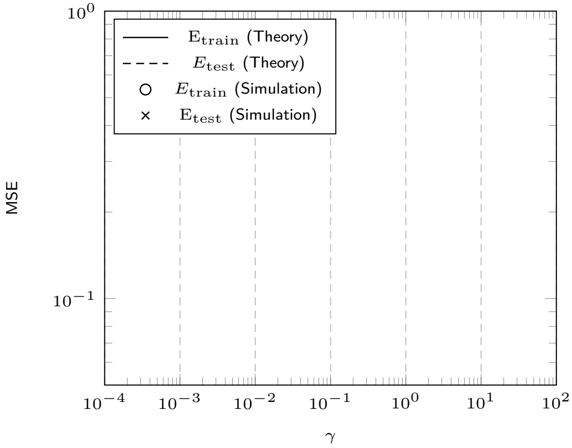


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

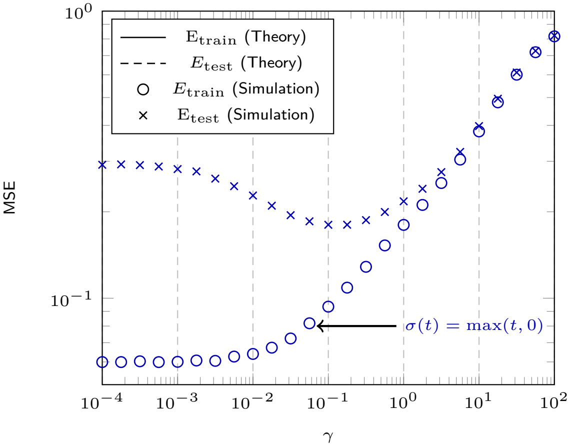


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

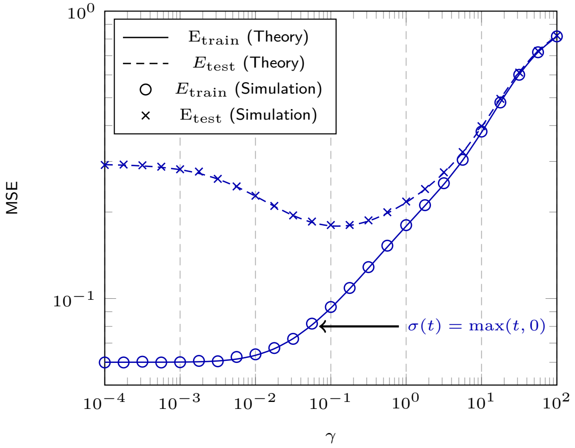


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Numerical validations

Performance of random feature-based ridge regression:

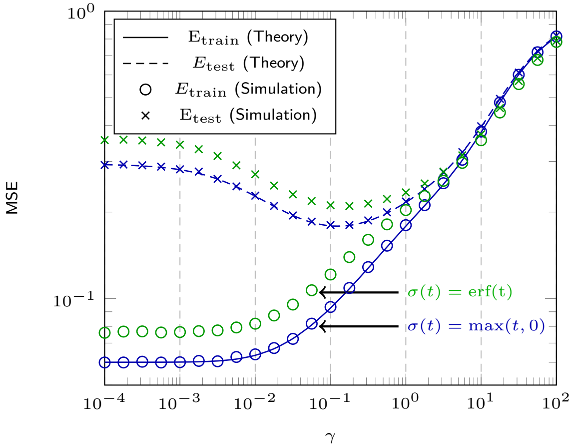


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

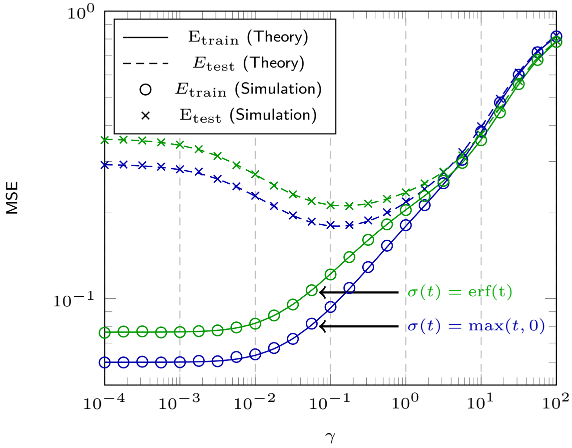


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

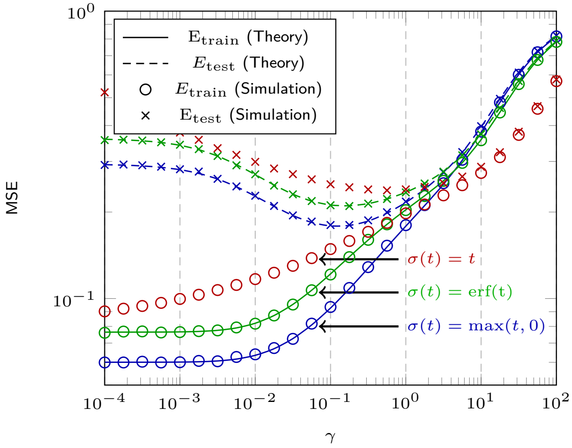


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

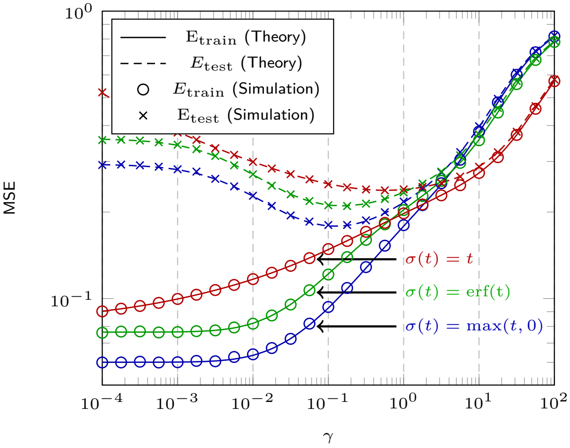


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Performance of random feature-based ridge regression:

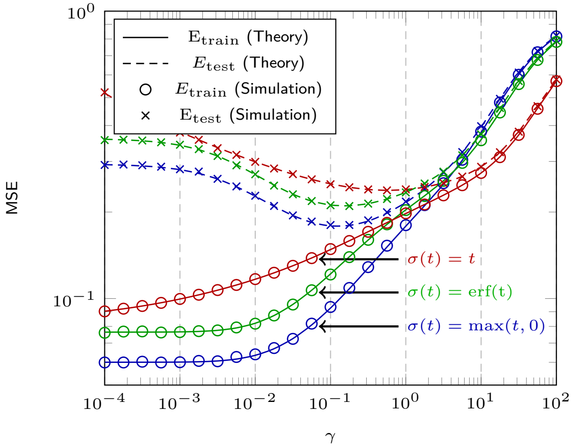


Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

Outline

Dig deeper into the averaged kernel Φ

For random feature maps:

- if deterministic data: performance determined by $\Phi(X)$ and problem dimension

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):
  $\Rightarrow$ what is the impact of nonlinearities on information extraction?

For random feature maps:

- if deterministic data: performance determined by $\Phi(X)$ and problem dimension
- if data following certain distribution (statistical information+random fluctuation): $\Rightarrow$ 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 $\omega_i \sim \mathcal{N}(0, \frac{1}{p}C_a)$, $a = 1, \ldots, K$ 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.

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):
  $\Rightarrow$ 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 $\omega_i \sim \mathcal{N}(0, \frac{1}{p} C_a)$, $a = 1, \ldots, K$ 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 $T \to \infty$, we have

- $p/T \to c_0 \in (0, \infty)$
- $T_a/T \to c_a \in (0, 1)$
- 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 $\|C_a\| = O(1)$, $\operatorname{tr} C_a^\circ / \sqrt{p} = O(1)$.

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):
  $\Rightarrow$ 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 $\omega_i \sim \mathcal{N}(0, \frac{1}{p}C_a)$, $a = 1, \ldots, K$ 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 $T \to \infty$, we have

- $p/T \to c_0 \in (0, \infty)$
- $T_a/T \to c_a \in (0, 1)$
- 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 $\|C_a\| = O(1)$, $\operatorname{tr} C_a^\circ/\sqrt{p} = O(1)$.

$\Rightarrow$ 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 $\sigma(t) = \mathrm{ReLU}(t)$,

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

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

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 $\sigma(t) = \mathrm{ReLU}(t)$,

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

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

- Taylor-expand nonlinear functions of x_i, x_j to get entry-wise approximation of $\Phi_{i,j}$

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 $\sigma(t) = \mathrm{ReLU}(t)$,

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

with $\angle(x_i, x_j) \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$. 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

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

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

with $\angle(x_i,x_j) \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$. 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 $T \to \infty$,

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

with

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

and $U = [\frac{J}{\sqrt{p}}, \phi]$, $B = \begin{bmatrix} tt^\mathsf{T} + 2S & t \\ t^\mathsf{T} & 1 \end{bmatrix}$,

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 $\sigma(t) = \mathrm{ReLU}(t)$,

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

with $\angle(x_i,x_j) \equiv \frac{x_i^\mathsf{T} x_j}{\|x_i\|\|x_j\|}$. 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 $T \to \infty$,

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

with

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

and $U = [\frac{J}{\sqrt{p}}, \phi]$, $B = \begin{bmatrix} tt^\mathsf{T} + 2S & t \\ t^\mathsf{T} & 1 \end{bmatrix}$, where $J = [j_1, \ldots, j_K]$, j_a canonical vector of class $\mathcal{C}_a$ (for clustering), weighted by two key parameters d_1, d_2 and

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

Consequence

Table: Coefficients d_i in $\tilde{\Phi}$ 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 $\sigma(\cdot)$:

Consequence

Table: Coefficients d_i in $\tilde{\Phi}$ 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
$\mathrm{sign}(t)$	$\frac{2}{\pi\tau}$	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
$\mathrm{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	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, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$ ⇒separate with differences in means M;

Consequence

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

$\sigma(t)$	d_1	d_2
t	1	0
$\max(t,0)$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\lvert t\rvert$	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
$\mathrm{sign}(t)$	$\frac{2}{\pi\tau}$	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
$\mathrm{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	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, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$ ⇒separate with differences in means M;
- *covariance-oriented*, $d_1 = 0$, $d_2 \neq 0$: $\lvert t\rvert$, $\cos(t)$ and $\exp(-t^2/2)$ ⇒track differences in covariances t, S;

Consequence

Table: Coefficients d_i in $\tilde{\Phi}$ 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
$\mathrm{sign}(t)$	$\frac{2}{\pi\tau}$	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
$\mathrm{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	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, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$ $\Rightarrow$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!

Consequence

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

$\sigma(t)$	d_1	d_2
t	1	0
$\max(t,0)$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\lvert t\rvert$	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
$\mathrm{sign}(t)$	$\frac{2}{\pi\tau}$	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
$\mathrm{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	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, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$ ⇒separate with differences in means M;
- *covariance-oriented*, $d_1 = 0$, $d_2 \neq 0$: $\lvert t\rvert$, $\cos(t)$ and $\exp(-t^2/2)$ ⇒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 $\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$.

⇒make use of both statistics!

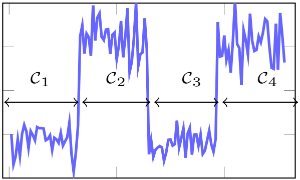
Not freely tunable as in the case of spectral clustering or SSL!

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 $\varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$.

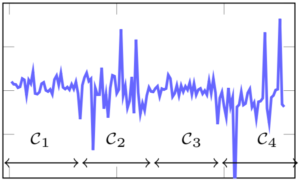
Numerical validations: Gaussian data

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

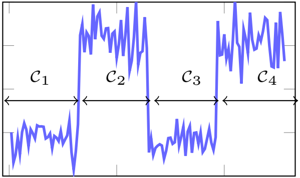


Eigenvector 2

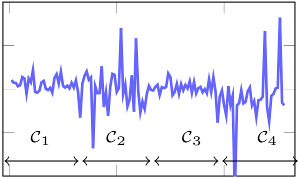
Numerical validations: Gaussian data

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 $\sigma(t) = |t|$)

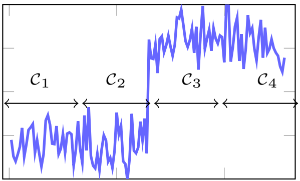


Eigenvector 1

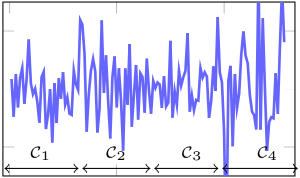


Eigenvector 2

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



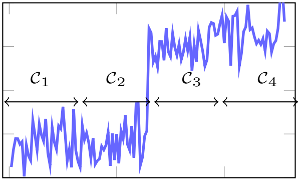
Eigenvector 1



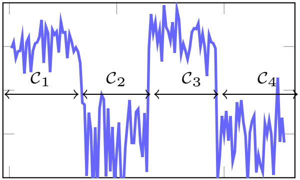
Eigenvector 2

Numerical validations: Gaussian data

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



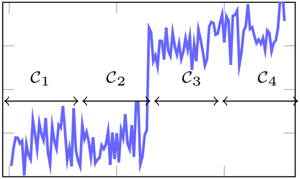
Eigenvector 1



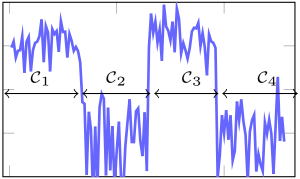
Eigenvector 2

Numerical validations: Gaussian data

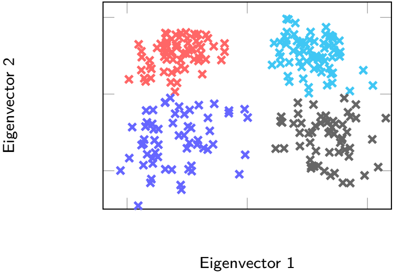
Case 3: $\varsigma_+ = 1$, $\varsigma_- = 0$ (the ReLU function)



Eigenvector 1



Eigenvector 2



Numerical validations: real datasets

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

	$\|M^\mathsf{T} M\|$	$\|tt^\mathsf{T} + 2S\|$
MNIST data	$\mathbf{172.4}$	86.0
EEG data	1.2	$\mathbf{182.7}$

Numerical validations: real datasets

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

	$\|M^\mathsf{T} M\|$	$\|tt^\mathsf{T} + 2S\|$
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%
	$\mathrm{sign}(t)$	83.34%	85.22%
	$\sin(t)$	87.81%	**87.50%**
	$\mathrm{erf}(t)$	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	$\mathrm{ReLU}(t)$	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%
	$\mathrm{sign}(t)$	64.63%	63.03%
	$\sin(t)$	70.34%	68.22%
	$\mathrm{erf}(t)$	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	$\mathrm{ReLU}(t)$	87.91%	90.97%

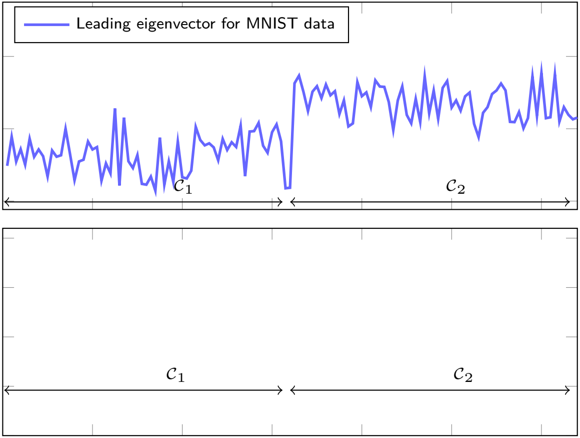


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Numerical validations: real datasets

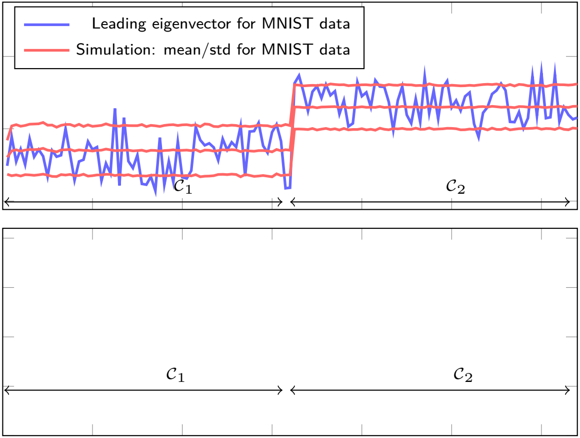


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Numerical validations: real datasets

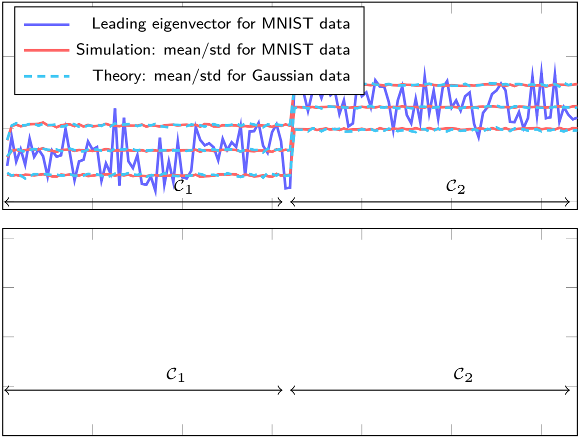


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

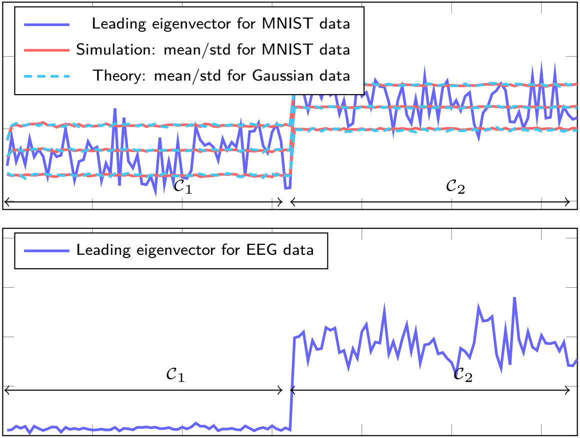


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

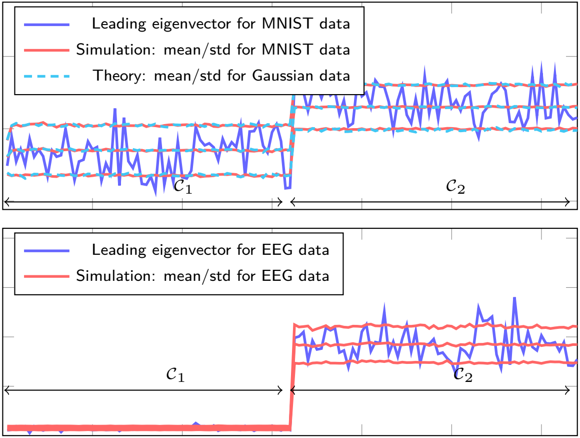
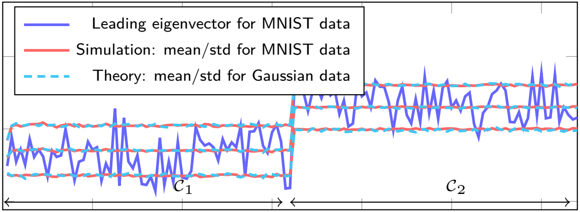


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Numerical validations: real datasets



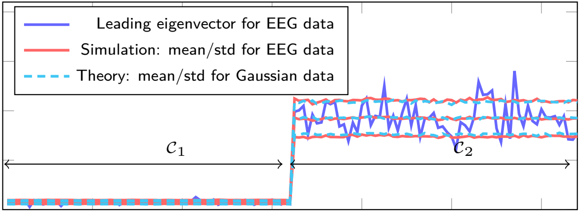


Figure: Leading eigenvector of Φ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix $\Phi \Rightarrow$ random feature-based spectral clustering

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix $\Phi \Rightarrow$ random feature-based spectral clustering

Take-away messages:

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix Φ $\Rightarrow$ random feature-based spectral clustering

Take-away messages:

- fast tuning of hyperparameters

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix Φ $\Rightarrow$ random feature-based spectral clustering

Take-away messages:

- fast tuning of hyperparameters
- nonlinearities into three attributes: means-, covariance-oriented and "balanced"

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix $\Phi \Rightarrow$ 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!)

Summary: random feature maps

Summary for random feature maps:

- concentration of measure helps extend trace lemma to nonlinear case
  $\Rightarrow$ asymptotic training/test performance of random feature-based ridge regression
- "concentration" of high dimensional data helps understand the key averaged kernel matrix Φ $\Rightarrow$ 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!)

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

Motivation: learning dynamics of neural networks

About neural networks and deep learning:

- Some known facts:

Motivation: learning dynamics of neural networks

About neural networks and deep learning:

- Some known facts:
  - trained with backpropagation (gradient-based method)

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

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:

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?

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?

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 . . .

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?

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?

$\Rightarrow$ The learning dynamics of neural networks!

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?

$\Rightarrow$ The learning dynamics of neural networks!

With RMT:

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

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?

$\Rightarrow$ 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!

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?

$\Rightarrow$ 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:

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?

$\Rightarrow$ 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

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?

$\Rightarrow$ 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

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?

$\Rightarrow$ 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:

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 $\mathcal{C}_2$.

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 $L(w) = \frac{1}{2n}\|y^\mathsf{T} - w^\mathsf{T} X\|^2$ with $X = [x_1, \ldots, x_n]$. For small learning rate α, with continuous-time approximation:

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

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

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 $L(w) = \frac{1}{2n}\|y^\mathsf{T} - w^\mathsf{T} X\|^2$ with $X = [x_1, \ldots, x_n]$. For small learning rate α, with continuous-time approximation:

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

of explicit solution $w(t) = e^{-\frac{\alpha t}{n} X X^\mathsf{T}} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} X X^\mathsf{T}} \right) (XX^\mathsf{T})^{-1} X y$ if XX^T 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

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 $L(w) = \frac{1}{2n}\|y^\mathsf{T} - w^\mathsf{T} X\|^2$ with $X = [x_1, \ldots, x_n]$. For small learning rate α, with continuous-time approximation:

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

of explicit solution $w(t) = e^{-\frac{\alpha t}{n} X X^\mathsf{T}} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} X X^\mathsf{T}} \right) (X X^\mathsf{T})^{-1} X y$ if $X X^\mathsf{T}$ 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 $\frac{1}{n} X X^\mathsf{T}$ (again): RMT is the answer!

Objective: Generalization Performance

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

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

for $w(t) = e^{-\frac{\alpha t}{n} XX^\mathsf{T}} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} XX^\mathsf{T}} \right) (XX^\mathsf{T})^{-1} Xy$.

Objective: Generalization Performance

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

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

for $w(t) = e^{-\frac{\alpha t}{n} XX^\mathsf{T}} w_0 + \left(I_p - e^{-\frac{\alpha t}{n} XX^\mathsf{T}} \right) (XX^\mathsf{T})^{-1} Xy$.

With RMT:

- although X random: $w(t)^\mathsf{T}\mu$ and $\|w(t)\|^2$ have asymptotically deterministic behavior (only depends on data statistics and problem dimension):
  $\Rightarrow$ the technique of **deterministic equivalent**

Objective: Generalization Performance

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

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

for $w(t) = e^{-\frac{\alpha t}{n}XX^\mathsf{T}}w_0 + \left(I_p - e^{-\frac{\alpha t}{n}XX^\mathsf{T}}\right)(XX^\mathsf{T})^{-1}Xy.$

With RMT:

- although X random: $w(t)^\mathsf{T}\mu$ and $\|w(t)\|^2$ have asymptotically deterministic behavior (only depends on data statistics and problem dimension):
  $\Rightarrow$ the technique of **deterministic equivalent**
- **Cauchy's integral formula** to express the functional $\exp(\cdot)$ via contour integration

Objective: Generalization Performance

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

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

for $w(t) = e^{-\frac{\alpha t}{n}XX^\mathsf{T}}w_0 + \left(I_p - e^{-\frac{\alpha t}{n}XX^\mathsf{T}}\right)(XX^\mathsf{T})^{-1}Xy.$

With RMT:

- although X random: $w(t)^\mathsf{T}\mu$ and $\|w(t)\|^2$ have asymptotically deterministic behavior (only depends on data statistics and problem dimension):
  $\Rightarrow$ the technique of **deterministic equivalent**
- **Cauchy's integral formula** to express the functional $\exp(\cdot)$ via contour integration

$\Rightarrow$ Network performance at any time is in fact deterministic and predictable!

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

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

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 $n \to \infty$,

- $\frac{1}{n} \operatorname{tr} AQ_M - \frac{1}{n} \operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T} \left(Q_M - \bar{Q}_M \right) b \xrightarrow{\text{a.s.}} 0$

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

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 $n \to \infty$,

- $\frac{1}{n}\operatorname{tr} AQ_M - \frac{1}{n}\operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T} \left(Q_M - \bar{Q}_M\right) b \xrightarrow{\text{a.s.}} 0$

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

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

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 $n \to \infty$,

- $\frac{1}{n}\operatorname{tr} AQ_M - \frac{1}{n}\operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T}\left(Q_M - \bar{Q}_M\right) b \xrightarrow{\text{a.s.}} 0$

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 $\frac{1}{n}\operatorname{tr} AQ_M$ and $a^\mathsf{T} Q_M b$):

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 $n \to \infty$,

- $\frac{1}{n}\operatorname{tr} AQ_M - \frac{1}{n}\operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T}\left(Q_M - \bar{Q}_M\right) b \xrightarrow{\text{a.s.}} 0$

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 $\frac{1}{n}\operatorname{tr} AQ_M$ and $a^\mathsf{T} Q_M b$):

Cauchy's integral formula

Example: for $f(M) = a^\mathsf{T} e^M bdz$,

$$f(M) = -\frac{1}{2\pi i}\oint_\gamma \exp(z) a^\mathsf{T} Q_M(z) bdz$$

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 $n \to \infty$,

- $\frac{1}{n}\operatorname{tr} AQ_M - \frac{1}{n}\operatorname{tr} A\bar{Q}_M \xrightarrow{\text{a.s.}} 0$
- $a^\mathsf{T}\left(Q_M - \bar{Q}_M\right) b \xrightarrow{\text{a.s.}} 0$

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 $\frac{1}{n}\operatorname{tr} AQ_M$ and $a^\mathsf{T} Q_M b$):

Cauchy's integral formula

Example: for $f(M) = a^\mathsf{T} e^M b dz$,

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

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

Generalization performance

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

Generalization performance

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

- **Cauchy's integral formula**: for $w(t)^\mathsf{T}\mu$:

$$\mu^\mathsf{T}w(t) = -\frac{1}{2\pi i}\oint_\gamma \mu^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}\left(f_t(z)w_0 + \frac{1-f_t(z)}{z}\frac{1}{n}Xy\right)dz$$

with $f_t(x) \equiv \exp(-\alpha t x)$.

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

- **Cauchy's integral formula**: for $w(t)^\mathsf{T}\mu$:

$$\mu^\mathsf{T}w(t) = -\frac{1}{2\pi i}\oint_\gamma \mu^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}\left(f_t(z)w_0 + \frac{1-f_t(z)}{z}\frac{1}{n}Xy\right)dz$$

with $f_t(x) \equiv \exp(-\alpha t x)$. Since $X = -\mu j_1^\mathsf{T} + \mu j_2^\mathsf{T} + \Omega = \mu y^\mathsf{T} + \Omega$, with $\Omega \equiv \left[\omega_1, \ldots, \omega_n\right] \in \mathbb{R}^{p\times n}$ 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,

$$\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1} = Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}$$

$$\begin{bmatrix}\mu^\mathsf{T}Q(z)\mu & 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y \\ 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y & -1+\frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}Q(z)\frac{1}{n}\Omega y\end{bmatrix}^{-1}\begin{bmatrix}\mu^\mathsf{T} \\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}Q(z)$$

where $Q(z) = \left(\frac{1}{n}\Omega\Omega^\mathsf{T} - zI_p\right)^{-1}$

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

- **Cauchy's integral formula**: for $w(t)^\mathsf{T}\mu$:

$$\mu^\mathsf{T}w(t) = -\frac{1}{2\pi i}\oint_\gamma \mu^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}\left(f_t(z)w_0 + \frac{1-f_t(z)}{z}\frac{1}{n}Xy\right)dz$$

with $f_t(x) \equiv \exp(-\alpha t x)$. Since $X = -\mu j_1^\mathsf{T} + \mu j_2^\mathsf{T} + \Omega = \mu y^\mathsf{T} + \Omega$, with $\Omega \equiv \left[\omega_1, \ldots, \omega_n\right] \in \mathbb{R}^{p\times n}$ 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,

$$\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1} = Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}$$
$$\begin{bmatrix}\mu^\mathsf{T}Q(z)\mu & 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y \\ 1+\frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y & -1+\frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}Q(z)\frac{1}{n}\Omega y\end{bmatrix}^{-1}\begin{bmatrix}\mu^\mathsf{T} \\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}Q(z)$$

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

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

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

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

- **Cauchy's integral formula**: for $w(t)^\mathsf{T}\mu$:

$$\mu^\mathsf{T}w(t) = -\frac{1}{2\pi i}\oint_\gamma \mu^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}\left(f_t(z)w_0 + \frac{1-f_t(z)}{z}\frac{1}{n}Xy\right)dz$$

with $f_t(x) \equiv \exp(-\alpha t x)$. Since $X = -\mu j_1^\mathsf{T} + \mu j_2^\mathsf{T} + \Omega = \mu y^\mathsf{T} + \Omega$, with $\Omega \equiv \left[\omega_1, \ldots, \omega_n\right] \in \mathbb{R}^{p\times n}$ 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,

$$\left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1} = Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}$$

$$\begin{bmatrix}\mu^\mathsf{T}Q(z)\mu & 1 + \frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y \\ 1 + \frac{1}{n}\mu^\mathsf{T}Q(z)\Omega y & -1 + \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}Q(z)\frac{1}{n}\Omega y\end{bmatrix}^{-1}\begin{bmatrix}\mu^\mathsf{T} \\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}Q(z)$$

where $Q(z) = \left(\frac{1}{n}\Omega\Omega^\mathsf{T} - zI_p\right)^{-1}$ 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 $p/n \to c \in (0,\infty)$ 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 $n \to \infty$,

$$P(w(t)^\mathsf{T}\hat{x} > 0 \mid \hat{x} \in \mathcal{C}_1) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0,$$
$$P(w(t)^\mathsf{T}\hat{x} < 0 \mid \hat{x} \in \mathcal{C}_2) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0$$

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

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

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

Theorem (Generalization Performance)

Let $p/n \to c \in (0,\infty)$ 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 $n \to \infty$,

$$P(w(t)^\mathsf{T}\hat{x} > 0 \mid \hat{x} \in \mathcal{C}_1) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0,$$
$$P(w(t)^\mathsf{T}\hat{x} < 0 \mid \hat{x} \in \mathcal{C}_2) - Q\left(\frac{\mathrm{E}}{\sqrt{\mathrm{V}}}\right) \xrightarrow{\text{a.s.}} 0$$

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

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

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

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

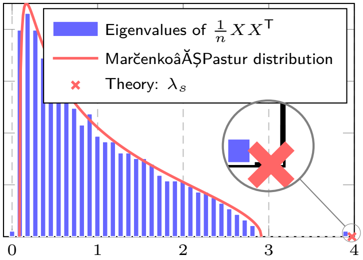


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

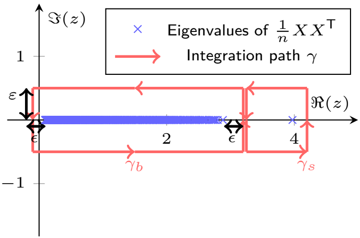


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

Simplification: “break” the contour integration

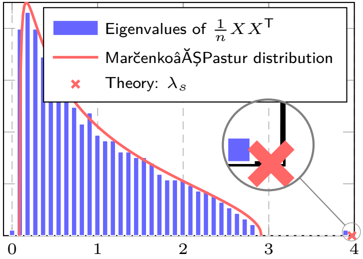


Figure: Eigenvalue distribution of $\frac{1}{n}XX^{\mathsf{T}}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

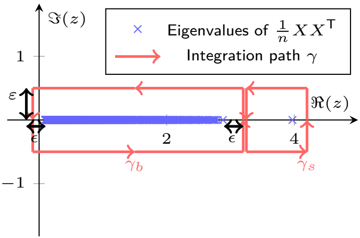


Figure: Eigenvalue distribution of $\frac{1}{n}XX^{\mathsf{T}}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

Two types of eigenvalues:

Simplification: "break" the contour integration

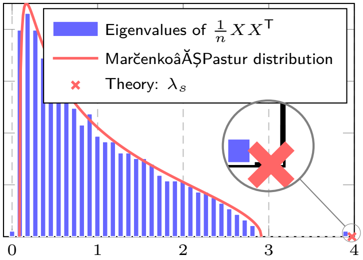


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

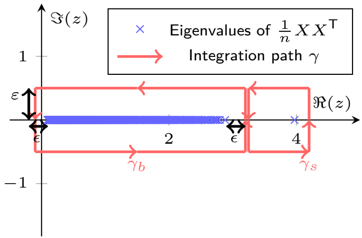


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

Two types of eigenvalues:

- "main bulk" ($[\lambda_-, \lambda_+]$): sum of real integrals

Simplification: “break” the contour integration

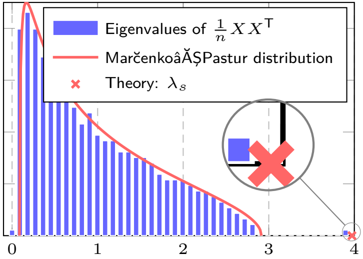


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

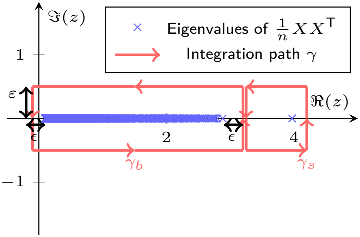


Figure: Eigenvalue distribution of $\frac{1}{n}XX^\mathsf{T}$ for $\mu = [1.5; 0_{p-1}]$, $p = 512$, $n = 1\,024$.

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

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

Computation of λ_s (Spike model)

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

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

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

Computation of λ_s (Spike model)

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

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

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

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

Computation of λ_s (Spike model)

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

$$\begin{aligned}
&\det\left(\frac{1}{n}XX^\mathsf{T} - \lambda I_p\right) = 0 \\
&\Leftrightarrow \det\left(\frac{1}{n}\Omega\Omega^\mathsf{T} - \lambda I_p + \begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^\mathsf{T}\\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\left(I_2 + \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^\mathsf{T}\\ \frac{1}{n}y^\mathsf{T}\Omega^\mathsf{T}\end{bmatrix} Q(\lambda)\begin{bmatrix}\mu & \frac{1}{n}\Omega y\end{bmatrix}\right) = 0 \\
&\Leftrightarrow 1 + (\|\mu\|^2 + c)m(\lambda) + o(1) = 0
\end{aligned}$$

(Simplified) generalization performance

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

with MarčenkoâĂŞPastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+-x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$ 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

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

(Simplified) generalization performance

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

with MarčenkoâĂŞPastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+-x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$ 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!

(Simplified) generalization performance

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

with MarčenkoâĂŞPastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+-x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$
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: $\mathrm{E}^2 \leq \int \frac{(1-f_t(x))^2}{x^2} d\mu(x) \cdot \int d\mu(x) \leq \frac{\|\mu\|^4}{\|\mu\|^2 + c} \mathrm{V}$, with equality if and only if the (initialization) variance $\sigma^2 = 0$: $\Rightarrow$ Performance drop due to large σ^2!

(Simplified) generalization performance

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

with MarčenkoâĂŞPastur distribution $\nu(dx) \equiv \frac{\sqrt{(x-\lambda_-)^+(\lambda_+ - x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)$ 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: $\mathrm{E}^2 \leq \int \frac{(1-f_t(x))^2}{x^2} d\mu(x) \cdot \int d\mu(x) \leq \frac{\|\mu\|^4}{\|\mu\|^2 + c} \mathrm{V}$, 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 $t \to \infty$, performance drop by $\sqrt{1 - \min(c, c^{-1})}$

Numerical validations

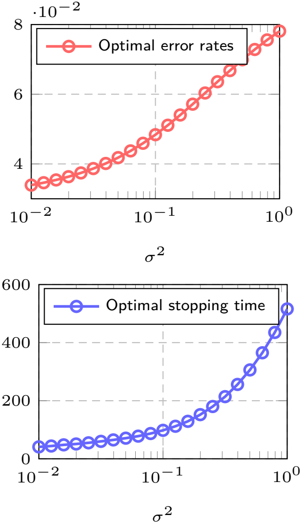


Figure: Optimal performance and stopping time as functions of σ^2 with $c = 1/2$, $\|\mu\|^2 = 4$ and $\alpha = 0.01$.

Numerical validations

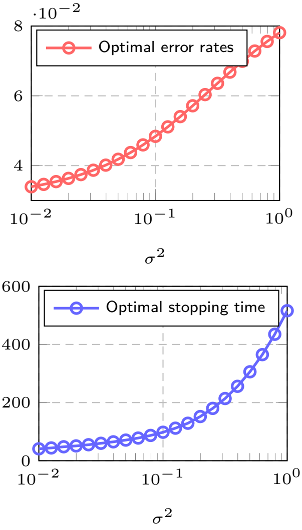


Figure: Optimal performance and stopping time as functions of σ^2 with $c = 1/2$, $\|\mu\|^2 = 4$ and $\alpha = 0.01$.

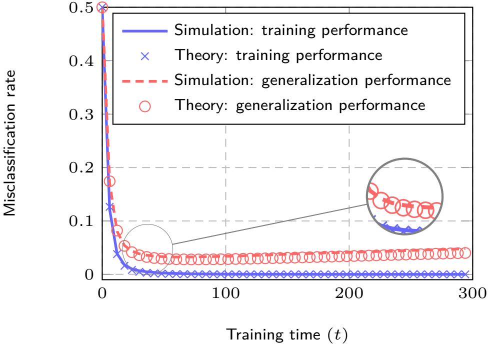


Figure: Training and generalization performance for MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 0.1$. Results averaged over 100 runs.

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

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

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 $1/p$)!

Outline

Take-away messages

- Asymptotic "**concentration effect**" for large n, p

Take-away messages

- Asymptotic "**concentration effect**" for large $n, p \Rightarrow$ simplification in analyses **and** models.

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 $p, n \to \infty$.

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 $p, n \to \infty$.
- Access to **limiting performances** and not only bounds!

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 $p, n \to \infty$.
- Access to **limiting performances** and not only bounds! $\Rightarrow$ hyperparameter optimization, algorithm improvement.

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 $p, n \to \infty$.
- Access to **limiting performances** and not only bounds! $\Rightarrow$ hyperparameter optimization, algorithm improvement.
- **Complete intuitive change**

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 $p, n \to \infty$.
- Access to **limiting performances** and not only bounds! $\Rightarrow$ hyperparameter optimization, algorithm improvement.
- **Complete intuitive change** $\Rightarrow$ opens way to renewed methods.

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 $p, n \to \infty$.
- 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**

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 $p, n \to \infty$.
- 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

Perspectives and Open Problems

- Neural nets: loss landscape, gradient descent dynamics and deep learning!

Perspectives and Open Problems

- Neural nets: loss landscape, gradient descent dynamics and deep learning!
- Generalized linear models

- Neural nets: loss landscape, gradient descent dynamics and deep learning!
- Generalized linear models
- More general problems from convex optimization (often of *implicit solution*)

Perspectives and Open Problems

- 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

Perspectives and Open Problems

- 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)

Perspectives and Open Problems

- 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

Perspectives and Open Problems

- 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
- ...

- 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.

K. Elkhalil, 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.

- 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.

- 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.

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.