Random Matrix Advances in Large Dimensional Machine Learning

Random Matrices and Complex Data Analysis Workshop

Zhenyu Liao, Romain Couillet

CentraleSupélec, Université Paris-Saclay, France
G-STATS IDEX DataScience Chair, GIPSA-lab, Université Grenoble-Alpes, France.

Dec 12, 2019, Shanghai

Outline

The pitfalls of large dimensional statistics: sample covariances in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.
- Maximum likelihood sample covariance matrix:

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p \times p}$$

  of rank at most n: optimal for $n \gg p$ (or, for p "small").
- When $n \sim p$, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has $\geq p - n$ zero eigenvalues.

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

  $\Rightarrow$ eigenvalue mismatch and **not** consistent!

When is one under the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: the Marčenko–Pastur law

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

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

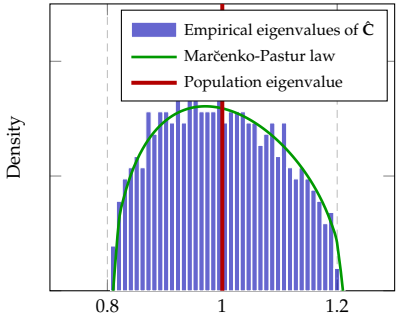


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm 2\sqrt{c} = \pm \mathbf{0.2}$ around population eigenvalue $\mathbf{1}$.

Take-away message:

- Counterintuitive phenomena in the large n, p regime
- RMT as a tool to **assess**, **understand** and **improve** large dimensional machine learning

Challenges:

- entry-wise **nonlinearity**: kernel function in kernel methods, activation function in neural networks
- (convex or non-convex) optimization-based methods with **implicit** solution
- more **realistic** data modeling

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification:

$$\mathcal{C}_1 : \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1) \quad \text{versus} \quad \mathcal{C}_2 : \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$$

- Neyman-Pearson test: classification possible only when

$$\boxed{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq O(1), \quad \|\mathbf{C}_1 - \mathbf{C}_2\| \geq O(p^{-1/2}), \quad |\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)| \geq O(\sqrt{p}), \quad \|\mathbf{C}_1 - \mathbf{C}_2\|_F^2 \geq O(1).}$$

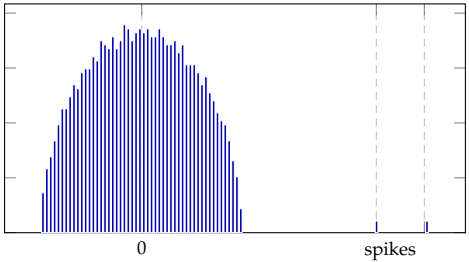
- In this non-trivial setting, for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$,

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

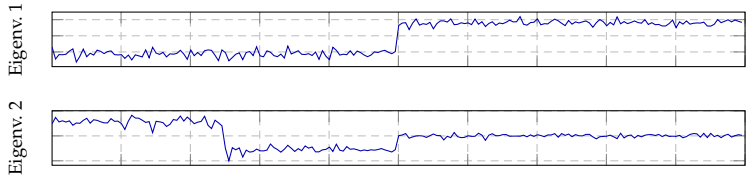
  as $n, p \to \infty$, regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$!
- direct consequence to large dimensional (distance-based) kernel matrices!

Reminder on kernel spectral clustering

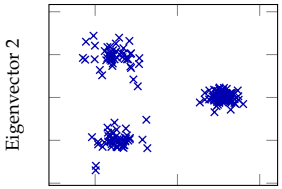
Two-step classification of n data points based on similarity kernel $\mathbf{K} \in \mathbb{R}^{n \times n}$, e.g., $\mathbf{K}_{ij} = \exp\left(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$:



Reminder on kernel spectral clustering



$\Downarrow$ **K-dimensional representation** $\Downarrow$



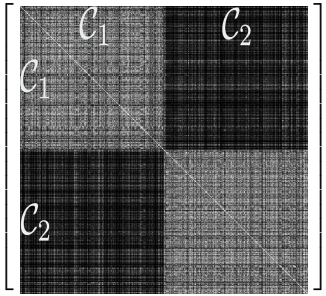
$\Downarrow$

EM or K-means clustering.

Kernel matrices of small and large dimensional Gaussian data

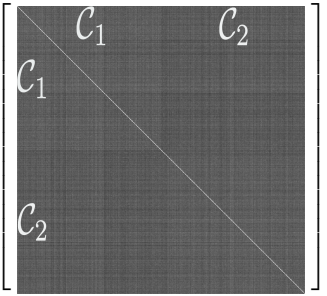
$\mathbf{K}_{ij} = \exp\left(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ and the second top eigenvectors $\mathbf{v}_2$ for small and large dimensional Gaussian data.

(a) $p = 5, n = 500$

$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

(b) $p = 250, n = 500$

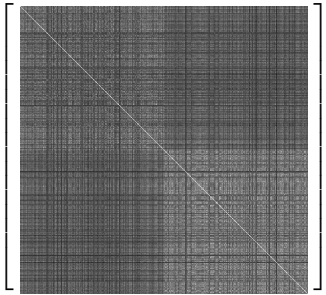
$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

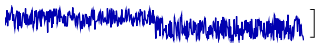
Kernel matrices of small and large dimensional real-world data

(a) MNIST

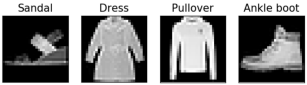
$\mathbf{K} =$



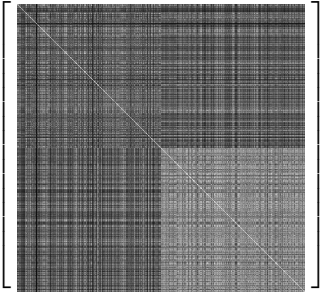
$\mathbf{v}_2 =$



(b) Fashion-MNIST



$\mathbf{K} =$



$\mathbf{v}_2 =$

A random matrix viewpoint of large kernel matrices

Asymptotic behavior of **K [Couillet and Benaych-Georges'16]**

For non-trivial classification of K-class Gaussian mixture $\mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, $a \in \{1, \ldots, K\}$, $\mathbf{K}_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ with f three-times differentiable around τ, we have, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \tilde{\mathbf{K}} \simeq f(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \mathbf{N} + \frac{1}{p}\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T} + *$$

with $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_K] \in \mathbb{R}^{n \times K}$, $\mathbf{j}_a = [\mathbf{0}, \mathbf{1}_{n_a}, \mathbf{0}]^\mathsf{T}$ (**class-info** vector!), $\mathbf{N}$ random **noise** and $\mathbf{A}$ function of

- $f(\tau), f'(\tau)$ and $f''(\tau)$
- **statistical information** $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2$, $\operatorname{tr}(\mathbf{C}_a - \mathbf{C}_b)$ and $\|\mathbf{C}_a - \mathbf{C}_b\|_F^2$, for $a, b \in \{1, \ldots, K\}$

- our low-dimensional machine learning intuitions collapse
- "curse of dimensionality" $\Rightarrow$ Taylor expansion and local linearization of f
- RMT provides a fully accessible **spiked-model** description of nonlinear $\mathbf{K}$

Intuition: from small to large dimensional kernels

Accumulated effect of small but structural "hidden" statistical information.

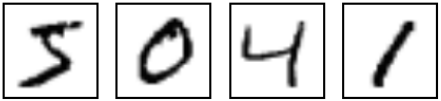
- **entry-wise**: for $\mathbf{K}_{ij} = \exp\left(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ with $\mathbf{x}_i = (-1)^a\boldsymbol{\mu} + \mathbf{z}_i$, $\mathbf{x}_j = (-1)^b\boldsymbol{\mu} + \mathbf{z}_j$, $\mathbf{z}_i, \mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$

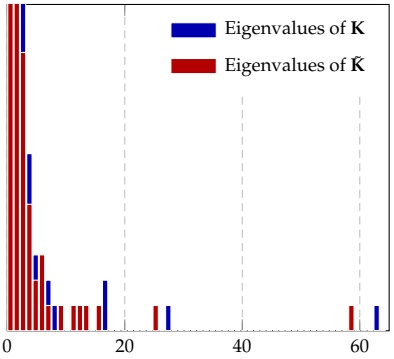
$$\mathbf{K}_{ij} \simeq \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}\|\boldsymbol{\mu}\|^2(-1)^{a+b}}_{O(p^{-1})} + *$$

  so that $\frac{1}{p}\|\boldsymbol{\mu}\|^2 \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j$;

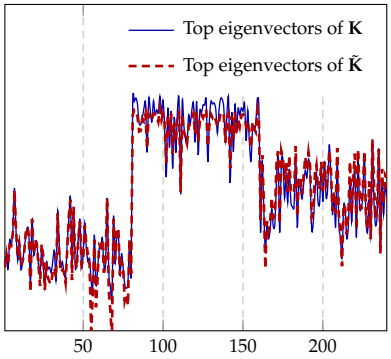
- **spectrum-wise**: $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|\boldsymbol{\mu}\|^2\|\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$ as well!
- we understand how kernel spectral clustering works for large dimensional data!

Experiments on real-world data: MNIST



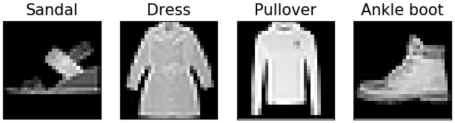


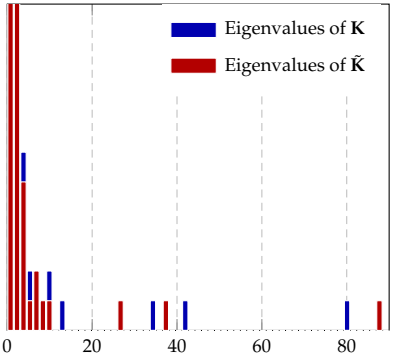
(a) Eigenvalues of $\mathbf{K}$ versus $\tilde{\mathbf{K}}$



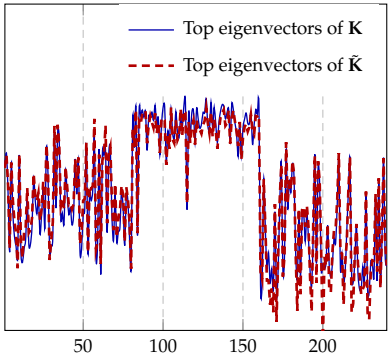
(b) Isolated eigenvectors of $\mathbf{K}$ and $\tilde{\mathbf{K}}$

Experiments on real-world data: Fashion-MNIST





(a) Eigenvalues of $\mathbf{K}$ versus $\tilde{\mathbf{K}}$



(b) Isolated eigenvectors of $\mathbf{K}$ and $\tilde{\mathbf{K}}$

How to better exploit the nonlinear f?

Due to "curse of dimensionality", $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \to \tau$.

- Taylor-expand $\mathbf{K}_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ around $f(\tau)$
- $\mathbf{K}$ depends only on **local** behavior of smooth f
- use a single point of nonlinear f

To exploit global information of $f \Rightarrow$ properly scaled kernel

$$f\left(\sqrt{p}\left(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau\right)\right) \quad \text{or} \quad \boxed{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})}$$

- $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} \to \mathcal{N}(0,1)$, use **whole** domain of f
- however, no "concentration" $\Rightarrow$ CLT expansion with orthogonal polynomials

Key object

$$\boxed{\mathbf{K} \equiv \{\delta_{i\neq j} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}\}_{i,j=1}^n}$$

Question: How f impacts the performance of kernel spectral clustering?

Orthogonal polynomials and inner-product kernels

$\sqrt{p}\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$: since $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} \to \mathcal{N}(0,1)$, essentially evaluate $f(\mathcal{N}(0,1))$.

Orthogonal polynomial decomposition

Any $f \in L^2(\mu_\mathcal{N})$ admits formal expansion (with Hermite polynomial)

$$f(x) \sim \sum_{l=0}^{\infty} a_l P_l(x), \quad \int P_{l_1}(x)P_{l_2}(x)\mu_\mathcal{N}(dx) = \delta_{l_1-l_2}, \quad a_l = \int f(x)P_l(x)\mu_\mathcal{N}(dx)$$

with $a_0 = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)} f(\xi) = 0$, and generalized moments $\boxed{a_1 = \mathbb{E}[\xi f(\xi)], \sqrt{2}a_2 = \mathbb{E}[\xi^2 f(x)], \nu = \mathbb{E}[f^2(x)]}$.

Asymptotic behavior of inner-product K [Liao and Couillet'19]

Under non-trivial classification condition and $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$, for $f \in L^2(\mu_\mathcal{N})$, $n, p \to \infty$ with $p/n \to c \in (0, \infty)$

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \tilde{\mathbf{K}} \simeq \mathbf{N} + \frac{1}{p}\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T} + *$$

with **J** class-info vectors, **N** random noise and **A** function of f and statistical information

$$\boxed{\mathbf{A} = a_1 \cdot \{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2\}_{a,b=1}^K + a_2 \cdot g(\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b), \|\mathbf{C}_a - \mathbf{C}_b\|_F^2)_{a,b=1}^K.}$$

Consequences

$$\|\mathbf{K}-\tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \tilde{\mathbf{K}} \simeq \frac{1}{p}\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T} + \mathbf{N} + *, \qquad \boxed{\mathbf{A} = a_1 \cdot \{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2\}_{a,b=1}^K + a_2 \cdot g(\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b), \|\mathbf{C}_a - \mathbf{C}_b\|_F^2)_{a,b=1}^K}$$

- $\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T}$ of low rank, limiting spectrum measure $\mu_\mathbf{K}$ same as $\mu_\mathbf{N}$, characterized by unique $m(z) \in \mathbb{C}^+$ **[Cheng and Signer'13]**

$$-\frac{1}{m(z)} = z + \frac{a_1{}^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z) \qquad (1)$$

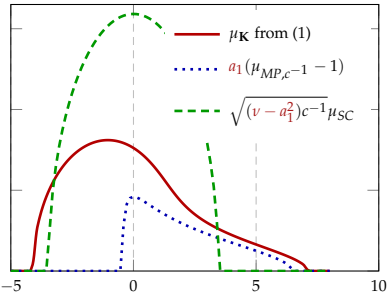
with $a_1 = \mathbb{E}[\xi f(\xi)]$, $\sqrt{2}a_2 = \mathbb{E}[\xi^2 f(x)]$ and $\nu = \mathbb{E}[f^2(x)]$.

- $\mu_\mathbf{K}$: "mix" of Marčenko-Pastur law μ_{MP} and semicircle law μ_{SC}

$$\mu_\mathbf{K} = a_1(\mu_{MP,c^{-1}} - 1) \boxtimes \sqrt{(\nu - a_1^2)c^{-1}}\mu_{SC}$$

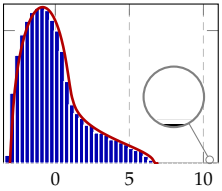
with $\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$ and $f(x) = a_1 x + \sqrt{\nu - a_1^2}\tilde{f}(x)$

- (a_1, ν) determines "noisy" **bulk**; (a_1, a_2) informative **spike**.

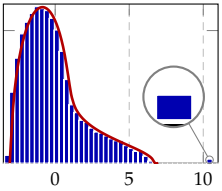


Practical consequences

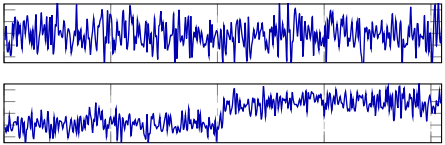
$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \tilde{\mathbf{K}} \simeq \frac{1}{p}\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T} + \mathbf{N} + *, \qquad \boxed{\mathbf{A} = a_1 \cdot \{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2\}_{a,b=1}^K + a_2 \cdot g(\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b), \|\mathbf{C}_a - \mathbf{C}_b\|_F^2)_{a,b=1}^K}$$



(a) Eigenvalues of $\mathbf{N}$

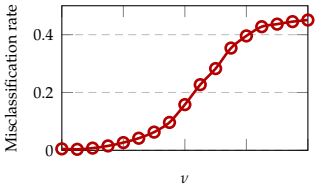


(b) Eigenvalues of $\mathbf{K}$



(c) Top eigenvectors of $\mathbf{N}$ (top) and $\mathbf{K}$ (bottom)

- $\mathbf{K}$ depend on f **only** via (a_1, a_2, ν):
  (a_1, a_2) impacts info and (a_1, ν) impacts noise
- minimize ν for larger **eigengap** and better performance!
- $\nu_{\min} = a_1^2 + a_2^2$, **quadratic** f is optimal among $L^2(\mu_{\mathcal{N}})$



Practical consequences

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \tilde{\mathbf{K}} \simeq \frac{1}{p}\mathbf{J}\mathbf{A}\mathbf{J}^\mathsf{T} + \mathbf{N} + *, \qquad \boxed{\mathbf{A} = a_1 \cdot \{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2\}_{a,b=1}^K + a_2 \cdot g(\mathrm{tr}(\mathbf{C}_a - \mathbf{C}_b), \|\mathbf{C}_a - \mathbf{C}_b\|_F^2)_{a,b=1}^K}$$

- a_1 control info in means and a_2 info in covariances $\Rightarrow$ **tuning a_1/a_2 with data**!

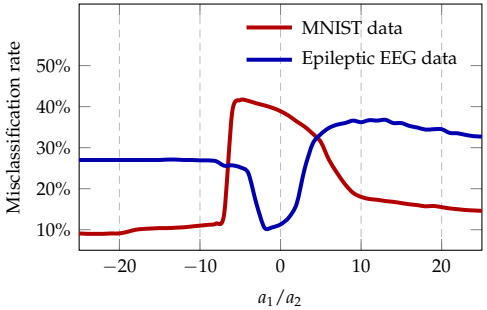


Table: Empirical estimation of differences in means and covariances of MNIST and EEG data

	$\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|^2$	$\|\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\|$
MNIST	**464.17**	166.35
EEG	2.41	**14.90**

[1] `http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html`.

Conclusion and limitations

Conclusion on large dimensional kernel methods:

- "curse of dimensionality" $\Rightarrow \mathbf{K}_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ depends on f in a **local** manner
  - proof based on Taylor expansion, holds only for **smooth** f
  - no **interpretation** for key parameters $f(\tau), f'(\tau), f''(\tau)$
  - **close match** between theory and real-world data experiments
- exploit **global** information with properly scaled kernel $\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$
  - proof based on orthogonal polynomials, holds for **any** $f \in L^2(\mu_\mathcal{N})$
  - better **interpretation** for key parameters (a_1, a_2, ν)
  - **only** tunning a_1/a_2 in real-world data experiments
- allow for "plug-and-play" analyses of most kernel-based methods: e.g., kernel ridge regression, support vector machines, graph-based semi-supervised methods, as well as random neural networks etc.

Limitations:

? optimization-based problems with implicit solution

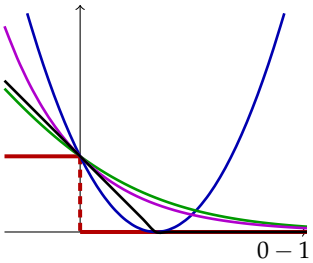
? limited to Gaussian data

Large dimensional optimization-based learning problems

Empirical risk minimization: for $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, find classifier $\boldsymbol{\beta} \in \mathbb{R}^p$ that minimizes

$$\frac{1}{n}\sum_{i=1}^n \ell(y_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

for some nonnegative (possibly non-smooth) convex loss ℓ and regularization $\lambda \geq 0$.



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

- **existence**: for different ℓ and $\mathbf{x}$, when such $\boldsymbol{\beta}$ exists? (function of $\lambda, \lim p/n$)
- **optimality**: how to choose loss function ℓ with respect to distribution of $\mathbf{x}_i$?
- **difficulty**: no **closed-form** solution: depend on all $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in a more involved manner.

Some related works on optimization-based learning problems

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \quad \text{or} \quad \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i) + \lambda R(\boldsymbol{\beta})$$

Different approaches:

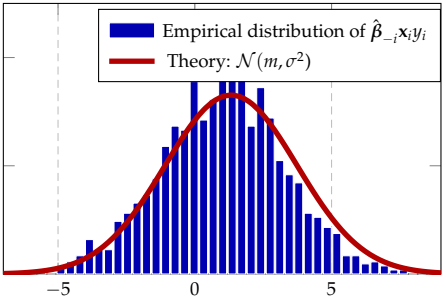
- approximate message passing and state-evolution analysis **[Donoho and Montanari'16]**
- convex Gaussian min-max theorem **[Taheri, Pedarsani and Thrampoulidis'19]**
- double "leave-one-out" approach **[El Karoui *et al.*'13]**: not suitable for data with pattern!

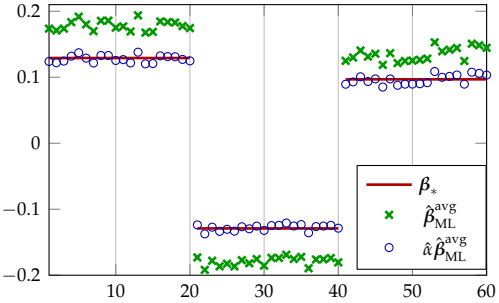
Intuition of improved "leave-one-out" approach [Mai and Liao'19]: single "leave-one-out" + RMT

- binary classification of $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{C})$
- denote $\hat{\boldsymbol{\beta}}_{-i} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{j\neq i} \ell(y_j \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_j) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$: $\hat{\boldsymbol{\beta}}_{-i}$ independent of $(\mathbf{x}_i, y_i)$ and $\hat{\boldsymbol{\beta}}_{-i} \simeq \hat{\boldsymbol{\beta}}$
- but $\hat{\boldsymbol{\beta}}_{-i}^\mathsf{T}\mathbf{x}_i y_i \not\simeq \hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i y_i \Rightarrow$ characterize $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T}\mathbf{x}_i y_i$: e.g., $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T}\mathbf{x}_i y_i \xrightarrow{\text{a.s.}} \kappa$ as $n, p \to \infty$
- since $\hat{\boldsymbol{\beta}}_{-i}^\mathsf{T}\mathbf{x}_i y_i \to \mathcal{N}(m, \sigma^2)$ by CLT $\Rightarrow$ form a system of (fixed-point) equations for (m, σ^2, κ)
- understand limiting performance for any three-times differentiable and convex $\ell(\cdot)$, as a function of p/n, and data statistics $(\boldsymbol{\mu}, \mathbf{C})$

Some simulations

Setting: binary classification $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{C})$ with $y_i = \pm 1$, logistic regression $\ell(t) = \ln(1 + e^{-t})$ known to be the maximum likelihood solution, with optimal Bayes solution (i.e., "true" parameter vector) $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$.





Remarks:

- can we get $\boldsymbol{\beta}_*$ in the large n, p limit? Not true even in expectation! $\hat{\boldsymbol{\beta}} \simeq \alpha^{-1}\boldsymbol{\beta}_* + \text{Gaussian noise}$
- maximum likelihood not optimal in high dimension: least-squares $\ell(t) = (t-1)^2$ always better!

Limitations:

- ✔ optimization-based problems with implicit solution: yes if convex!
- ? limited to Gaussian data

From theory to practice: concentrated random vectors

RMT often assumes $\mathbf{x}$ are affine maps $\mathbf{Az}+\mathbf{b}$ of $\mathbf{z}\in\mathbb{R}^p$ with i.i.d. entries.

Concentrated random vectors [Ledoux'05]

For Lipschitz function $f:\mathbb{R}^p\mapsto\mathbb{R}$, there exists deterministic $m_f\in\mathbb{R}$

$$P\left(|f(\mathbf{x})-m_f|>\epsilon\right)\leq e^{-g(\epsilon)},\quad \text{for some strictly increasing function } g.$$

Asymptotic behavior of $\mathbf{K}$ for concentrated data [Seddik, Tamaazousti, Couillet'19]

For non-trivial classification of K-class (**universal**) concentrated random mixture of mean $\boldsymbol{\mu}_a$ and covariance $\mathbf{C}_a$, $a\in\{1,\ldots,K\}$, $\mathbf{K}_{ij}=f(\|\mathbf{x}_i-\mathbf{x}_j\|^2/p)$ with f three-times differentiable around τ,

$$\|\mathbf{K}-\tilde{\mathbf{K}}\|\xrightarrow{\text{a.s.}}0,\quad \tilde{\mathbf{K}}\simeq f(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}+\mathbf{N}+\frac{1}{p}\mathbf{JAJ}^\mathsf{T}+*$$

with $\mathbf{J}$ **class-info** vectors, $\mathbf{N}$ random **noise** and $\mathbf{A}$ function of $f(\tau), f'(\tau), f''(\tau)$ and **statistical information**.

$\Rightarrow$Theory remains valid for concentrated vectors and almost real images!

From concentrated random vectors to GANs

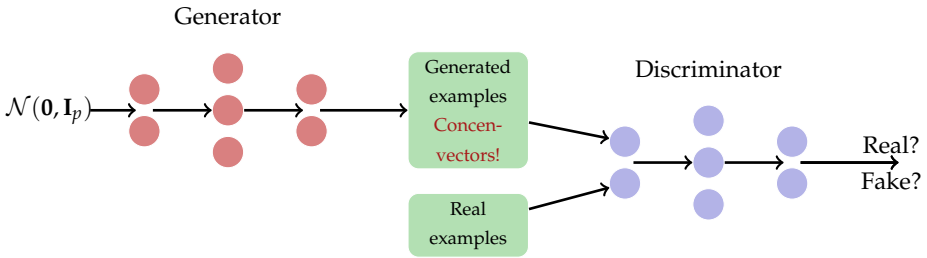


Figure: Illustration of a generative adversarial network (GAN).

Figure: Images samples generated by BigGAN **[Brock *et al.*'18]**.

Some clues . . . and much more can be done!

RMT as a tool to **analyze**, **understand** and **improve**
large dimensional machine learning methods.

- powerful and flexible tool to assess matrix-based machine learning systems;
- study (convex) optimization-based learning methods, e.g., logistic regression;
- understand impact of optimization methods, the dynamics of gradient descent;
- non-convex problems (e.g., deep neural nets) are more difficult, but accessible in some cases, e.g., low rank matrix recovery, phase retrieval, etc;
- more to be done: transfer learning, generative models, graph-based methods, robust statistics, etc.

References

- Noureddine El Karoui. “The spectrum of kernel random matrices”. In: *The Annals of Statistics* 38.1 (2010), pp. 1–50
- Xiuyuan Cheng and Amit Singer. “The spectrum of random inner-product kernel matrices”. In: *Random Matrices: Theory and Applications* 2.04 (2013), p. 1350010
- Romain Couillet and Florent Benaych-Georges. “Kernel spectral clustering of large dimensional data”. In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1393–1454
- Zhenyu Liao and Romain Couillet. “Inner-product Kernels are Asymptotically Equivalent to Binary Discrete Kernels”. In: *arXiv preprint arXiv:1909.06788* (2019)
- David Donoho and Andrea Montanari. “High dimensional robust M-estimation: asymptotic variance via approximate message passing”. In: *Probability Theory and Related Fields* 166.3-4 (2016), pp. 935–969
- Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. “Sharp Guarantees for Solving Random Equations with One-Bit Information”. In: *arXiv preprint arXiv:1908.04433* (2019)
- Noureddine El Karoui et al. “On robust regression with high-dimensional predictors”. In: *Proceedings of the National Academy of Sciences* 110.36 (2013), pp. 14557–14562
- Emmanuel J Candès and Pragya Sur. “The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression”. In: *arXiv preprint arXiv:1804.09753* (2018)
- Xiaoyi Mai and Zhenyu Liao. “High Dimensional Classification via Empirical Risk Minimization: Improvements and Optimality”. In: *arXiv preprint arXiv:1905.13742* (2019)
- Zhenyu Liao and Romain Couillet. “On the Spectrum of Random Features Maps of High Dimensional Data”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, 2018, pp. 3063–3071
- Zhenyu Liao and Romain Couillet. “The Dynamics of Learning: A Random Matrix Approach”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, 2018, pp. 3072–3081
- Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. “Kernel Random Matrices of Large Concentrated Data: the Example of GAN-Generated Images”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 7480–7484

Thank you!

More info: `https://romaincouillet.hebfree.org` and `https://zhenyu-liao.github.io`!