# A Large Dimensional Analysis of Kernel LS-SVM

## ED STIC reception meeting 2019-2020

**Zhenyu Liao**

joint work with Romain Couillet

CentraleSupélec, Université Paris-Saclay, France.

Nov 28, 2019

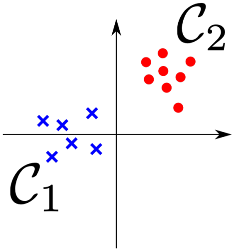
CentraleSupélec

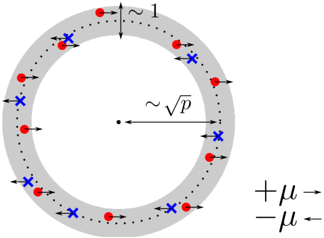
# Motivation: counterintuitive phenomena in large dimensional learning

- Big Data era: large dimensional and massive amount of data
- data number $n$ and dimension $p$ both **large** and **comparable**: analysis with **Random Matrix Theory**
- "curse of dimensionality" in large dimensional classification:

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

$\mathbf{x} \in \mathbb{R}^p$ has norm $\|\mathbf{x}\| = O(\sqrt{p})$ with spread $\|\mathbf{x}\| - \mathbb{E}[\|\mathbf{x}\|] = O(1)$.




- indeed, for $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b, a \in \{1,2\}$

$$\boxed{\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \simeq \tau}$$

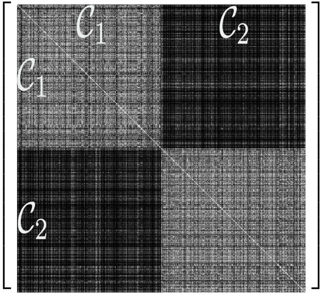
for $p$ large, regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$!

# Consequences to large kernel matrices: Gaussian mixture

Classify data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ into $\mathcal{C}_1$ or $\mathcal{C}_2$ with distance-based kernel $\mathbf{K}_{ij} = e^{-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2}$.

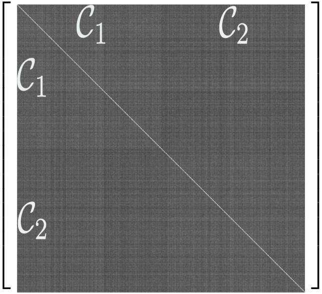
(a) $p = 5, n = 500$

$\mathbf{K} =$




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

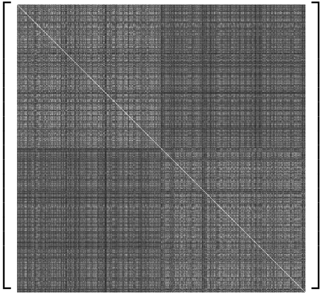
$\mathbf{K} =$



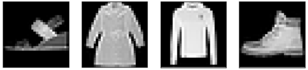
# Consequences to large kernel matrices: real-world datasets

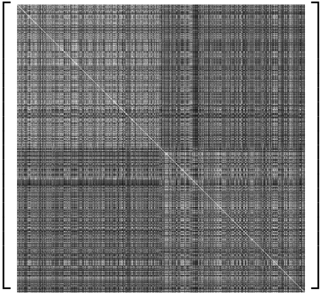
Distance-based kernel $\mathbf{K}_{ij} = e^{-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ on MNIST and Fashion-MNIST data.

(a) MNIST

$\mathbf{K} =$ 

(b) Fashion-MNIST




$\mathbf{K} =$ 

**Question**: impact of large $p$ on performance of kernel-based methods, e.g., LS-SVM?

# Reminder on least-squares support vector machine

- find classifier $g(\mathbf{x}) = \mathbf{w}^\mathsf{T}\varphi(\mathbf{x}) + b$ by minimizing

$$L(\mathbf{w}, b) = \frac{\gamma}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\mathsf{T}\varphi(\mathbf{x}_i) - b\right)^2 + \|\mathbf{w}\|^2$$

on training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, +1\}$.

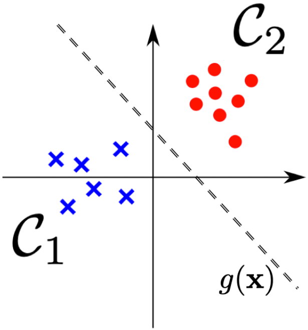
- "kernel trick": $g(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\{k(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^n + b$ with

$$\boldsymbol{\alpha} = \mathbf{Q}(\mathbf{y} - b\mathbf{1}_n), \quad b = \frac{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{y}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n}$$

where $\mathbf{Q} \equiv \left(\mathbf{K} + \frac{\gamma}{n}\mathbf{I}_n\right)^{-1}$ resolvent of kernel matrix

$$\boxed{\mathbf{K} \equiv \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n = \left\{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\right\}_{i,j=1}^n.}$$

- for new $\mathbf{x}$, assign to $\mathcal{C}_1$ if $g(\mathbf{x}) < 0$ and $\mathcal{C}_2$ otherwise.




**Key observation**: $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \simeq \tau$ for large $p$, $\mathbf{K}$ only depends on $f(\tau), f'(\tau)$ and $f''(\tau)$!

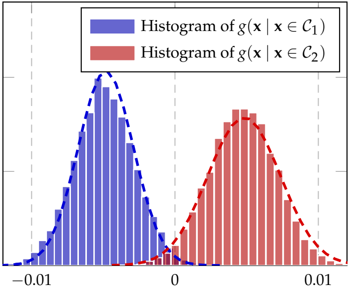
# Main result: exact performance of LS-SVM

## Main result

Under a binary Gaussian mixture model $\mathcal{C}_1 : \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$ vs. $\mathcal{C}_2 : \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$, **decision function** $g(\mathbf{x})$ is asymptotically Gaussian

$$g(\mathbf{x} \mid \mathbf{x} \in \mathcal{C}_a) \sim \mathcal{N}(E_a, V_a), \quad a = \{1, 2\}$$

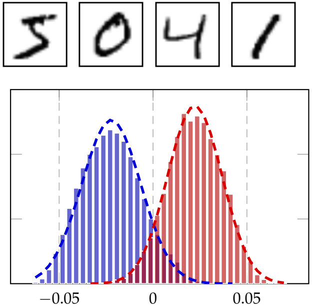
that depends on data statistics $(\boldsymbol{\mu}_a, \mathbf{C}_a)$, hyperparameter $\gamma$ and kernel function $f$ "locally".



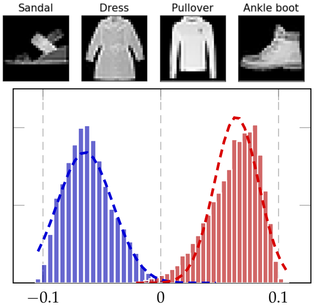

$\Rightarrow$ direct access to classification performance via Gaussian tail $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$.

# When applied to real world datasets

(a) MNIST




(b) Fashion-MNIST




**Why?**

- MNIST and Fashion-MINIT data are clearly NOT mixture of Gaussian vectors
- when $n, p$ large, algorithms tend to work AS IF they were: use only 1st and 2nd order statistical info.

[1]Means and covariances of data empirically estimated from the whole database.

# Conclusion and take-away message

- counterintuitive phenomena in real-world large dimensional learning
- RMT as a tool to assess exact performance, understand and improve large dimensional learning
- **in this work**: "curse of dimensionality" $\Rightarrow$ exact performance of kernel LS-SVM
- more to be done in the general context of large dimensional learning!

Some references and related works:


- Zhenyu Liao and Romain Couillet. "A Large Dimensional Analysis of Least Squares Support Vector Machines". In: *IEEE Transactions on Signal Processing* 67.4 (2019), pp. 1065–1074
- Cosme Louart, Zhenyu Liao, and Romain Couillet. "A Random Matrix Approach to Neural Networks". In: *The Annals of Applied Probability* 28.2 (2018), pp. 1190–1248
- 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
- Xiaoyi Mai and Romain Couillet. "A Random Matrix Analysis and Improvement of Semi-supervised Learning for Large Dimensional Data". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 3074–3100
- 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!