On the Spectrum of Random Features Maps of High Dimensional Data

ICML 2018, Stockholm, Sweden

Zhenyu Liao, Romain Couillet

L2S, CentraleSupélec, Université Paris-Saclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, Université Grenoble-Alpes, France.

Outline

Problem Setup

Random projection/random feature maps for feature extraction:

$$\text{data vectors} \xrightarrow[\sigma(\cdot)\ \text{entry-wise}]{\text{random } \mathbf{W} \in \mathbb{R}^{n \times p}} \text{feature vectors}$$

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$ $\qquad$ $\mathbf{\Sigma} = \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{n \times T}$

Figure: Illustration of random feature maps

Objective

Gram matrix of random features $\mathbf{G} \equiv \frac{1}{n}\mathbf{\Sigma}^{\mathsf{T}}\mathbf{\Sigma}$ (sample covariance matrix in feature space):

- what kind of data information are extracted?
- what is the impact of different nonlinearities?
- how to perform clustering with $\mathbf{G}$, what do its eigenvectors "look like"?

With RMT: for large n, p, T, eigenspectrum of $\mathbf{G}$ is determined only by[1]

- the average kernel matrix $\mathbf{\Phi}_{i,j} \equiv \mathbb{E}_{\mathbf{w}}\, \mathbf{G}_{i,j} = \mathbb{E}_{\mathbf{w}}\, \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i)\sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}_j)$ (function of $\mathbf{X}$)
- the ratios between n, p, T.

[1] Louart Cosme, **Zhenyu Liao**, and Romain Couillet. "A Random Matrix Approach to Neural Networks." The Annals of Applied Probability 28, no. 2 (2018): 1190-1248.

Some Known Facts

Objective: spectral characterization of $\mathbf{\Phi}$, with $\mathbf{\Phi}_{i,j} = \mathbb{E}_{\mathbf{w}}\, \sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)$:
For standard Gaussian $\mathbf{W} \Rightarrow$ integral calculus on $\mathbb{R}^p$.

Table: $\mathbf{\Phi}_{i,j}$ for commonly used $\sigma(\cdot)$, $\angle \equiv \frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$.

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

$\Rightarrow$(still) highly nonlinear functions of the data $\mathbf{x}$!

Dig Deeper into the Average Kernel $\mathbf{\Phi}$

Data Model

Consider data from a K-class Gaussian mixture model: $\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \mathbf{x}_i = \boldsymbol{\mu}_a/\sqrt{p} + \boldsymbol{\omega}_i$, with $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_a/p)$, $a = 1, \ldots, K$ of statistical mean $\boldsymbol{\mu}_a$ and covariance $\mathbf{C}_a$.

Non-trivial Classification [Neyman-Pearson Minimal]

For p large, we have $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$, $\|\mathbf{C}_a\| = O(1)$ and $\operatorname{tr}(\mathbf{C}_a - \mathbf{C}_b)/\sqrt{p} = O(1)$.

As a consequence,

$$\|\mathbf{x}_i\|^2 = \underbrace{\|\boldsymbol{\omega}_i\|^2}_{O(1)} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})} = \underbrace{\operatorname{tr}\mathbf{C}_a/p}_{O(1)} + \underbrace{\|\boldsymbol{\omega}_i\|^2 - \operatorname{tr}\mathbf{C}_a/p}_{O(p^{-1/2})} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}$$

- if relaxed, classification too easy: it suffices to compare the norm $\|\mathbf{x}_i\|^2$ and $\|\mathbf{x}_j\|^2$!
- in fact reveals a more intrinsic property of high dimensional data:

Curse of dimensionality: little difference in Euclidean distance between pairs!

Denote $\mathbf{C}^\circ = \sum_{i=1}^K \frac{T_i}{T}\mathbf{C}_a$ and $\mathbf{C}_a = \mathbf{C}_a^\circ + \mathbf{C}^\circ$ for $a = 1, \ldots, K$.
Then $\|\mathbf{x}_i\|^2 = \tau + O(p^{-1/2})$ with $\tau \equiv \operatorname{tr}(\mathbf{C}^\circ)/p$, $\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - \mathbf{x}_i^\mathsf{T}\mathbf{x}_j \approx 2\tau$:

$\Rightarrow$ Almost constant distance no matter from the same or different classes!

Dig Deeper into the Average Kernel Φ

Why things are still working? $\Rightarrow$ statistical information are hidden in smaller order terms!

$$\Rightarrow \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - \mathbf{x}_i^\mathsf{T}\mathbf{x}_j \approx 2\tau + \underbrace{\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b/p + \boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j/\sqrt{p} + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}$$

Small entry-wise $\neq$ small in matrix form (in operator norm): repeated in $p \times p$ large matrix $\Rightarrow$ spectral clustering works! 😄

Moreover, "concentration" brings simplifications: for $\mathbf{\Phi}_{i,j} = \mathbb{E}_{\mathbf{w}}\, \sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_j)$ and ReLU,

$$\mathbf{\Phi}_{i,j} = \frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\angle \arccos\left(-\angle\right) + \sqrt{1-\angle^2}\right)$$

with $\angle \equiv \frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$. "Concentration": $\angle = \frac{0}{\tau^2} +$ information terms $(\boldsymbol{\mu}_a, \mathbf{C}_a)$!

"Blessing" of Dimensionality

High dimensional "concentration" $\Rightarrow$ Taylor expansion to linearize $\mathbf{\Phi}$!

Main Results

Asymptotic Equivalent of $\mathbf{\Phi}$

For all $\sigma(\cdot)$ listed in the table above, we have, as $n \sim p \sim T \to \infty$,

$$\|\mathbf{\Phi} - \tilde{\mathbf{\Phi}}\| \to 0$$

almost surely, with

$$\tilde{\mathbf{\Phi}} \equiv d_1 \left(\mathbf{\Omega} + \mathrm{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right)^\mathsf{T} \left(\mathbf{\Omega} + \mathrm{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}} \right) + d_2 \mathbf{U}\mathbf{B}\mathbf{U}^\mathsf{T} + d_0 \mathbf{I}_T$$

and $\mathbf{U} \equiv \left[\frac{\mathbf{J}}{\sqrt{p}}, \phi \right]$, $\quad \mathbf{B} \equiv \begin{bmatrix} \mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S} & \mathbf{t} \\ \mathbf{t}^\mathsf{T} & 1 \end{bmatrix}$.

Table: Coefficients d_i in $\tilde{\mathbf{\Phi}}$ 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	$\frac{1}{4(\tau+1)^3}$

With $\mathbf{J} \equiv [\mathbf{j}_1, \ldots, \mathbf{j}_K]$, $\mathbf{j}_a$ canonical vector of $\mathcal{C}_a$: $(\mathbf{j}_a)_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$ (for clustering), weighted by

- $\mathbf{\Omega}$, ϕ random fluctuations of data.
- $\mathbf{M} \equiv [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$, $\mathbf{t} \equiv \left\{ \operatorname{tr} \mathbf{C}_a^\circ / \sqrt{p} \right\}_{a=1}^K$, $\mathbf{S} \equiv \{\operatorname{tr}(\mathbf{C}_a\mathbf{C}_b)/p\}_{a,b=1}^K$ statistical information from data distribution.

Consequence

Table: Coefficients d_i in $\tilde{\mathbf{\Phi}}$ 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
$\operatorname{sign}(t)$	$\frac{2}{\pi\tau}$	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
$\operatorname{erf}(t)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-\frac{t^2}{2})$	0	$\frac{1}{4(\tau+1)^3}$

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

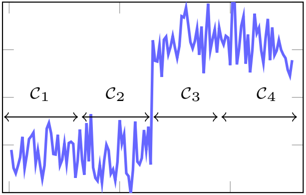
- *mean-oriented*, $d_1 \neq 0$, $d_2 = 0$: t, $1_{t>0}$, $\operatorname{sign}(t)$, $\sin(t)$ and $\operatorname{erf}(t)$ $\Rightarrow$ separate with difference in means $\mathbf{M}$;
- *covariance-oriented*, $d_1 = 0$, $d_2 \neq 0$: $|t|$, $\cos(t)$ and $\exp(-t^2/2)$ $\Rightarrow$ track differences in covariances $\mathbf{t}$, $\mathbf{S}$;
- "*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!

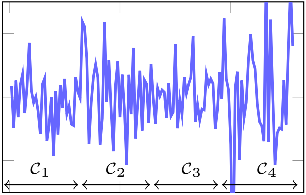
Numerical Validations: Gaussian Data

Example: Gaussian mixture data of four classes: $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2)$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ with Leaky ReLU function $\varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$.

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

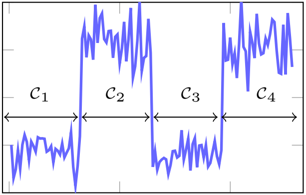


Eigenvector 1

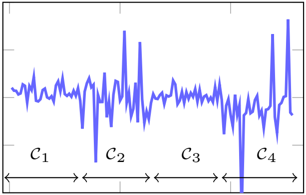


Eigenvector 2

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



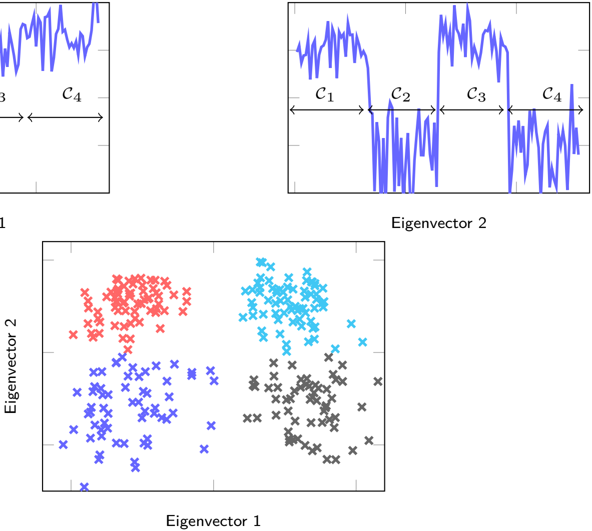
Eigenvector 1



Eigenvector 2

Numerical Validations: Gaussian Data

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



Numerical Validations: Real Datasets

Figure: The MNIST image database.

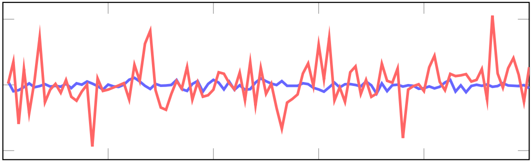


Figure: The epileptic EEG datasets.[2]

Reproducibility: codes available at `https://github.com/Zhenyu-LIAO/RMT4RFM`.

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

Numerical Validations: Real Datasets

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

	$\|\mathbf{M}^{\mathsf{T}}\mathbf{M}\|$	$\|\mathbf{t}\mathbf{t}^{\mathsf{T}} + 2\mathbf{S}\|$
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%

Numerical Validations: Real Datasets

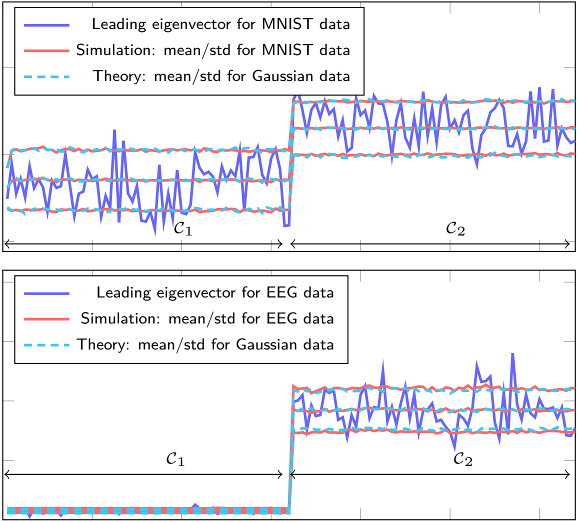


Figure: Leading eigenvector of $\mathbf{\Phi}$ for the MNIST (top) and EEG (bottom) with Gaussian mixture data (of same statistics) with a width of ± 1 standard deviations.

Summary

Take-away message:

- "concentration" of high dimensional data to handle the nonlinearity
- different nonlinearities into three attributes: mean-, covariance-oriented and "balanced"
- optimize the choice of nonlinearity as a function of data (quadratic and LReLU)
- novel insight into understanding of neural networks for high dimensional data

Future work:

- study of the eigenvalue distribution $\Rightarrow$ the (asymptotic) behavior of leading eigenvectors
- combination of different type of nonlinearities, e.g., $\sin + \cos \Rightarrow$ Gaussian kernel
- directly linking $\sigma(\cdot)$ and the coefficients d_0, d_1 and d_2

Thank you!

Poster # 62