Random Matrix Theory for Neural Networks

Ph.D. Mid-Term Evaluation

Zhenyu Liao

Laboratoire des Signaux et Systèmes
CentraleSupélec
Université Paris-Saclay

Salle sd.207, Bâtiment Bouygues
Gif-sur-Yvette, France

CentraleSupélec

Outline

Curriculum Vitae

- Education
  - Ph.D. in Statistics and Signal Processing, L2S, CentraleSupélec, France 2016-present
    - Thesis: Random Matrix Theory for Neural Networks.
    - Supervisor: **Prof. Romain Couillet**, **Prof. Yacine Chitour**.
  - M.Sc. in Signal and Image Processing, CentraleSupélec/Paris-Sud, France 2014-2016
  - B.Sc. in Electronic Engineering, Paris-Sud, France 2013-2014
  - B.Sc. in Optical & Electronic Information, HUST, Wuhan, China 2010-2014
- Ph.D. training
  - Scientific training [**completed**]
    - Random matrix theory and machine learning application: 27 hours.
    - Summer school of signal and image processing in Peyresq: 21 hours.
  - Professional training [**completed**]
    - Law and intellectual property: 18 hours.
    - European projects Horizon 2020: 18 hours.
    - Techniques for scientific writing and associated softwares: 15 hours.
- Teaching
  - 2017-2018: Lab work of Signal and System 1, with Prof. Laurent Le Brusquet, Department of Signal and Statistics, CentraleSupélec: **54 hours**.

Curriculum Vitae

- Review activities
  - IEEE Transactions on Signal Processing
  - Neural Processing Letters

- Publications
  - Conferences
    - **Z. Liao**, R. Couillet, "The Dynamics of Learning: A Random Matrix Approach", (submitted to) The 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 2018.
    - **Z. Liao**, R. Couillet, "On the Spectrum of Random Features Maps of High Dimensional Data", (submitted to) The 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 2018.
    - **Z. Liao**, R. Couillet, "Une Analyse des Méthodes de Projections Aléatoires par la Théorie des Matrices Aléatoires (in French)", Colloque GRETSI'17, Juan Les Pins, France, 2017.
    - **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.
  - Journals
    - C. Louart, **Z. Liao**, R. Couillet, "A Random Matrix Approach to Neural Networks", (in press) Annals of Applied Probability, 2017.
    - **Z. Liao**, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines", (submitted to) Journal of Machine Learning Research, 2017.

Motivation: features in machine learning

Learning = Representation + Evaluation + Optimization.[1]

Features: representation of the data that contains crucial information.

Various methods for feature extraction:

- feature selection by hand
- feature learned via backpropagation
- random feature maps

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

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

of some data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{p \times T}$.

SCM in feature space $\Rightarrow$ feature Gram matrix $\mathbf{G}$:

$$\mathbf{G} \equiv \frac{1}{T}\mathbf{F}^\mathsf{T}\mathbf{F}$$

with $\mathbf{F} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_T)]$ feature matrix of data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$.

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

SCM in feature space $\Rightarrow$ feature Gram matrix $\mathbf{G}$:

$$\mathbf{G} \equiv \frac{1}{T}\mathbf{F}^\mathsf{T}\mathbf{F}$$

with $\mathbf{F} \in \mathbb{R}^{n\times T}$ feature matrix of data $\mathbf{X}$.

In RFM, $\mathbf{G}$ determines training and test performance via its *resolvent*

$$\mathbf{Q}(z) \equiv (\mathbf{G} - z\mathbf{I}_T)^{-1}.$$

Example 1:
MSE of RFM-based ridge regression (also called *extreme learning machines*):

$$E_{\text{train}} = \frac{1}{T}\|\mathbf{y} - \boldsymbol{\beta}^\mathsf{T}\mathbf{F}\|_F^2 = \frac{\gamma^2}{T}\mathbf{y}^\mathsf{T}\mathbf{Q}^2(-\gamma)\mathbf{y}$$
$$E_{\text{test}} = \frac{1}{\hat{T}}\|\hat{\mathbf{y}} - \boldsymbol{\beta}^\mathsf{T}\hat{\mathbf{F}}\|_F^2$$

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

Motivation: gradient descent dynamics

Example 2:

Feature matrix $\mathbf{F}$ with associated labels $\mathbf{y}$, a classifier trained by GD is to minimize

$$L(\mathbf{w}) = \frac{1}{2T}\|\mathbf{y} - \mathbf{w}^\mathsf{T}\mathbf{F}\|^2$$

with constant small learning rate α, GD dynamics yields:

$$\frac{\partial \mathbf{w}(t)}{\partial t} = -\alpha \Delta_{\mathbf{w}} L = \frac{\alpha}{T}\mathbf{F}(\mathbf{y} - \mathbf{F}^\mathsf{T}\mathbf{w}(t)) \Rightarrow \mathbf{w}(t) = e^{-\frac{\alpha t}{T}\mathbf{F}\mathbf{F}^\mathsf{T}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{T}\mathbf{F}\mathbf{F}^\mathsf{T}}\right)(\mathbf{F}\mathbf{F}^\mathsf{T})^{-1}\mathbf{F}\mathbf{y}.$$

Generalization performance of new feature $\hat{\mathbf{f}}$ related to:

$$\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{f}} = -\frac{1}{2\pi i}\oint_\gamma f_t(z)\hat{\mathbf{f}}^\mathsf{T}\tilde{\mathbf{Q}}(z)\mathbf{w}_0 \, dz - \frac{1}{2\pi i}\oint_\gamma \frac{1 - f_t(z)}{z}\hat{\mathbf{f}}^\mathsf{T}\tilde{\mathbf{Q}}(z)\frac{1}{T}\mathbf{F}\mathbf{y} \, dz$$

with $f_t(z) = \exp(-\alpha t z)$, $\tilde{\mathbf{Q}}$ resolvent of $\tilde{\mathbf{G}} = \frac{1}{T}\mathbf{F}\mathbf{F}^\mathsf{T}$, γ contours all eigenvalues of $\tilde{\mathbf{G}}$.

Important remark

Both examples: study of eigenspectrum through resolvent of SCM-like matrices ($\mathbf{G}$ or $\tilde{\mathbf{G}}$), especially in high dimensional regime today where n, p, T are comparably large.

$\Rightarrow$ Random Matrix Theory (RMT) is the answer!

Motivation: difficulty from nonlinearity

However,

- highly nonlinear and abstract nature of real-world problems
  $\Rightarrow$ classical RMT results cannot apply directly
- in need of new tools to adapt to nonlinear models

Example: random feature maps

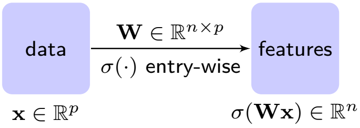

Figure: Illustration of random feature maps

with random matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$ with i.i.d. entries and (nonlinear) *activation function* $\sigma(\cdot)$.

Classical RMT results essentially based on trace lemma: for $\mathbf{A} \in \mathbb{R}^{n \times n}$ of bounded operator norm and random vector $\mathbf{w} \in \mathbb{R}^n$ with i.i.d. entries

$$\left| \frac{1}{n} \mathbf{w}^\mathsf{T} \mathbf{A} \mathbf{w} - \frac{1}{n} \operatorname{tr}(\mathbf{A}) \right| \to 0$$

almost surely as $n \to \infty$.

Motivation: difficulty from nonlinearity

Example: random feature maps

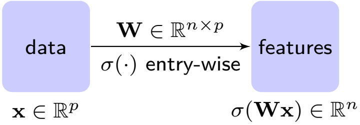

Figure: Illustration of random feature maps

Classical RMT results are essentially based on the trace lemma: for $\mathbf{A} \in \mathbb{R}^{n \times n}$ of bounded operator norm and random vector $\mathbf{w} \in \mathbb{R}^n$ with i.i.d. entries

$$\left| \frac{1}{n} \mathbf{w}^\mathsf{T} \mathbf{A} \mathbf{w} - \frac{1}{n} \operatorname{tr}(\mathbf{A}) \right| \to 0$$

almost surely as $n \to \infty$.

However, object under study $\frac{1}{n} \sigma(\mathbf{w}^\mathsf{T} \mathbf{X}) \mathbf{A} \sigma(\mathbf{X}^\mathsf{T} \mathbf{w})$:

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

$\Rightarrow$ One major objective of the thesis: handle nonlinearity in RMT

Handle nonlinearity in RMT: concentration

Lemma: Concentration of quadratic forms

For $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\|\mathbf{X}\|, \|\mathbf{A}\|$ bounded and $\sigma(\cdot)$ Lipschitz, as $n, p \to \infty$ with $\frac{p}{n} \to c$,

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

where $\mathbf{\Phi} \equiv \mathbb{E}_{\mathbf{w}} \left[\sigma(\mathbf{X}^\mathsf{T} \mathbf{w}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{X}) \right]$.

$$\mathbf{\Phi}(\mathbf{a}, \mathbf{b}) \equiv \mathbb{E}_{\mathbf{w}} \left[\sigma(\mathbf{w}^\mathsf{T} \mathbf{a}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{b}) \right] = (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(\mathbf{w}^\mathsf{T} \mathbf{a}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{b}) e^{-\frac{1}{2}\|\mathbf{w}\|^2} d\mathbf{w}$$

$$= \frac{1}{2\pi} \int_{\mathbb{R}^2} \sigma(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}}) \sigma(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}}) e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2} d\tilde{\mathbf{w}} \quad \text{(projection on span}(\mathbf{a}, \mathbf{b}))$$

For $\sigma(t) = \max(t, 0) \equiv \mathrm{ReLU}(t)$,

$$\mathbf{\Phi}(\mathbf{a}, \mathbf{b}) = \frac{1}{2\pi} \int_S \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}} \cdot \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}} \cdot e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2} d\tilde{\mathbf{w}} = \frac{1}{2\pi} \|\mathbf{a}\| \|\mathbf{b}\| \left(\sqrt{1 - \angle^2} + \angle \arccos(-\angle) \right)$$

where $S \equiv \min\left(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}}, \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}} \right) > 0$, $\angle \equiv \frac{\mathbf{a}^\mathsf{T} \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$.

Handle nonlinearity in RMT: concentration

Behavior of $\mathbf{\Phi}$ as a function of $\mathbf{X}$: "concentration" phenomenon for large n, p, T

For ReLU function,

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

Assumption: GMM data

For $i = 1, \ldots, T$, $\mathbf{x}_i \sim \mathcal{N}\left(\frac{1}{\sqrt{p}}\boldsymbol{\mu}_a, \frac{1}{p}\mathbf{C}_a\right)$ and $\mathbf{x}_i = \frac{1}{\sqrt{p}}\boldsymbol{\mu}_a + \boldsymbol{\omega}_i$, with $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{C}_a)$, class $\mathcal{C}_a$ has cardinality T_a, for $a = 1, \ldots, K$.

$$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j = \left(\boldsymbol{\mu}_a^\mathsf{T}/\sqrt{p} + \boldsymbol{\omega}_i^\mathsf{T}\right)\left(\boldsymbol{\mu}_b/\sqrt{p} + \boldsymbol{\omega}_j\right) = \boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b/p + (\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i)/\sqrt{p} + \boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j$$

$$\mathbf{x}_i^\mathsf{T}\mathbf{x}_i = \left(\boldsymbol{\mu}_a^\mathsf{T}/\sqrt{p} + \boldsymbol{\omega}_i^\mathsf{T}\right)\left(\boldsymbol{\mu}_a/\sqrt{p} + \boldsymbol{\omega}_i\right) = \|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p} + \|\boldsymbol{\omega}_i\|^2$$

Growth rate [Neyman-Pearson Minimal]

As $n, p, T \to \infty$, let $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{T_a}{T}\mathbf{C}_a$ and for $a = 1, \ldots, K$, $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$,

- the Euclidean norm $\|\boldsymbol{\mu}_a\| = O(1)$
- the operator norm $\|\mathbf{C}_a\| = O(1)$ and $\mathrm{tr}(\mathbf{C}_a^\circ) = O(\sqrt{p})$

Handle nonlinearity in RMT: concentration

Growth rate [Neyman-Pearson Minimal]

As $n, p, T \to \infty$, let $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{T_a}{T} \mathbf{C}_a$ and for $a = 1, \ldots, K$, $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$,

- the Euclidean norm $\|\boldsymbol{\mu}_a\| = O(1)$
- the operator norm $\|\mathbf{C}_a\| = O(1)$ and $\operatorname{tr}(\mathbf{C}_a^\circ) = O(\sqrt{p})$

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

$$\begin{aligned} \mathbf{x}_i^\mathsf{T} \mathbf{x}_i &= \left(\frac{1}{\sqrt{p}} \boldsymbol{\mu}_a^\mathsf{T} + \boldsymbol{\omega}_i^\mathsf{T} \right) \left(\frac{1}{\sqrt{p}} \boldsymbol{\mu}_a + \boldsymbol{\omega}_i \right) = \frac{1}{p} \|\boldsymbol{\mu}_a\|^2 + \frac{2}{\sqrt{p}} \boldsymbol{\mu}_a^\mathsf{T} \boldsymbol{\omega}_i + \|\boldsymbol{\omega}_i\|^2 \\ &= \underbrace{\frac{1}{p} \|\boldsymbol{\mu}_a\|^2 + \frac{2}{\sqrt{p}} \boldsymbol{\mu}_a^\mathsf{T} \boldsymbol{\omega}_i}_{O(p^{-1})} + \underbrace{\|\boldsymbol{\omega}_i\|^2 - \frac{1}{p} \operatorname{tr}(\mathbf{C}_a) + \frac{1}{p} \operatorname{tr}({\mathbf{C}_a}^\circ)}_{O(p^{-1/2})} + \underbrace{\tau}_{O(1)} \end{aligned}$$

with $\mathbb{E}_{\boldsymbol{\omega}_i} \|\boldsymbol{\omega}_i\|^2 = \operatorname{tr}(\mathbf{C}_a)/p$ and $\tau \equiv \operatorname{tr}(\mathbf{C}^\circ)/p$.

$\Rightarrow g(\mathbf{x}_i^\mathsf{T} \mathbf{x}_i)$ concentrates around $g(\tau) \Rightarrow$ Taylor expansion

Theorem (Asymptotic equivalent of $\mathbf{\Phi}_c$)

Recenter $\mathbf{\Phi} \equiv \mathbb{E}_{\mathbf{w}} \left[\sigma(\mathbf{X}^\mathsf{T} \mathbf{w}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{X}) \right]$ *to get* $\mathbf{\Phi}_c \equiv \mathbf{P} \mathbf{\Phi} \mathbf{P}$, *with* $\mathbf{P} \equiv \mathbf{I}_T - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\mathsf{T}$. *As* $T \to \infty$,

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

almost surely, with $\tilde{\mathbf{\Phi}}_c \equiv \mathbf{P} \tilde{\mathbf{\Phi}} \mathbf{P}$ *and*

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

as well as

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

where $\mathbf{\Omega} \equiv \left[\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_T \right]$, $\quad \phi \equiv \left\{ \|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_i\|^2 \right] \right\}_{i=1}^T$, and data statistics[2],

$$\mathbf{M} \equiv \left[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \right], \quad \mathbf{t} \equiv \{ \operatorname{tr} \mathbf{C}_a^\circ / \sqrt{p} \}_{a=1}^K, \quad \mathbf{S} \equiv \{ \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b)/p \}_{a,b=1}^K, \quad \mathbf{J} \equiv \left[\mathbf{j}_1, \ldots, \mathbf{j}_K \right]$$

where $\mathbf{j}_a \in \mathbb{R}^T$ denotes the canonical vector of class $\mathcal{C}_a$ such that $(\mathbf{j}_a)_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$.

[2] $\mathbf{M}$ for *means*, $\mathbf{t}$ for (difference in) *traces* while $\mathbf{S}$ for the "*shapes*" of covariances.

Coefficients d_i for different activation functions

Table: $\mathbf{\Phi}(\mathbf{a},\mathbf{b})$ for different $\sigma(\cdot)$, $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

$\sigma(t)$	$\mathbf{\Phi}(\mathbf{a},\mathbf{b})$
t	$\mathbf{a}^\mathsf{T}\mathbf{b}$
$\max(t,0) \equiv \mathrm{ReLU}(t)$	$\frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\|\left(\angle(\mathbf{a},\mathbf{b})\arccos\left(-\angle(\mathbf{a},\mathbf{b})\right)+\sqrt{1-\angle(\mathbf{a},\mathbf{b})^2}\right)$
$\|t\|$	$\frac{2}{\pi}\|\mathbf{a}\|\|\mathbf{b}\|\left(\angle(\mathbf{a},\mathbf{b})\arcsin\left(\angle(\mathbf{a},\mathbf{b})\right)+\sqrt{1-\angle(\mathbf{a},\mathbf{b})^2}\right)$
$1_{t>0}$	$\frac{1}{2}-\frac{1}{2\pi}\arccos\left(\angle(\mathbf{a},\mathbf{b})\right)$
$\mathrm{sign}(t)$	$\frac{2}{\pi}\arcsin\left(\angle(\mathbf{a},\mathbf{b})\right)$
$\varsigma_2 t^2+\varsigma_1 t+\varsigma_0$	$\varsigma_2^2\left(2\left(\mathbf{a}^\mathsf{T}\mathbf{b}\right)^2+\|\mathbf{a}\|^2\|\mathbf{b}\|^2\right)+\varsigma_1^2\mathbf{a}^\mathsf{T}\mathbf{b}+\varsigma_2\varsigma_0\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)+\varsigma_0^2$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)\right)\cosh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2\right)\right)\sinh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\mathrm{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{a}^\mathsf{T}\mathbf{b}}{\sqrt{(1+2\|\mathbf{a}\|^2)(1+2\|\mathbf{b}\|^2)}}\right)$
$\exp(-\frac{t^2}{2})$	$\frac{1}{\sqrt{(1+\|\mathbf{a}\|^2)(1+\|\mathbf{b}\|^2)-(\mathbf{a}^\mathsf{T}\mathbf{b})^2}}$

Coefficients d_i for different activation functions

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

$\sigma(t)$	d_0	d_1	d_2
t	0	1	0
$\max(t,0) \equiv \mathrm{ReLU}(t)$	$\left(\frac{1}{4}-\frac{1}{2\pi}\right)\tau$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\|t\|$	$\left(1-\frac{2}{\pi}\right)\tau$	0	$\frac{1}{2\pi\tau}$
$1_{t>0}$	$\frac{1}{4}-\frac{1}{2\pi}$	$\frac{1}{2\pi\tau}$	0
$\mathrm{sign}(t)$	$1-\frac{2}{\pi}$	$\frac{2}{\pi\tau}$	0
$\varsigma_2 t^2+\varsigma_1 t+\varsigma_0$	$2\tau^2\varsigma_2^2$	ς_1^2	ς_2^2
$\cos(t)$	$\frac{1}{2}+\frac{e^{-2\tau}}{2}-e^{-\tau}$	0	$\frac{e^{-\tau}}{4}$
$\sin(t)$	$\frac{1}{2}-\frac{e^{-2\tau}}{2}-\tau e^{-\tau}$	$e^{-\tau}$	0
$\mathrm{erf}(t)$	$\frac{2}{\pi}\left(\arccos\left(\frac{2\tau}{2\tau+1}\right)-\frac{2\tau}{2\tau+1}\right)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-\frac{t^2}{2})$	$\frac{1}{\sqrt{2\tau+1}}-\frac{1}{\tau+1}$	0	$\frac{1}{4(\tau+1)^3}$

$\Rightarrow$ three types of function $\sigma(\cdot)$:

- *mean-oriented*, $d_1 \neq 0$ while $d_2 = 0$: t, $1_{t>1}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$
- *covariance-oriented*, $d_1 = 0$ while $d_2 \neq 0$: $|t|$, $\cos(t)$, $\exp(-t^2/2)$
- *balanced*, $d_1, d_2 \neq 0$: the ReLU function and quadratic function

Example 1: MSE of the random feature-based ridge regression:

$$E_{\rm train} = \frac{1}{T}\|\mathbf{y} - \beta^{\sf T}\mathbf{F}\|_F^2 = \frac{\gamma^2}{T}\mathbf{y}^{\sf T}\mathbf{Q}^2(-\gamma)\mathbf{y}, \quad E_{\rm test} = \frac{1}{\hat{T}}\|\hat{\mathbf{y}} - \beta^{\sf T}\hat{\mathbf{F}}\|_F^2$$

with the ridge regressor $\beta \equiv \frac{1}{T}\mathbf{F}\left(\mathbf{G} + \gamma\mathbf{I}_T\right)^{-1}\mathbf{y}^{\sf T} = \frac{1}{T}\mathbf{F}\mathbf{Q}(-\gamma)\mathbf{y}^{\sf T}$.

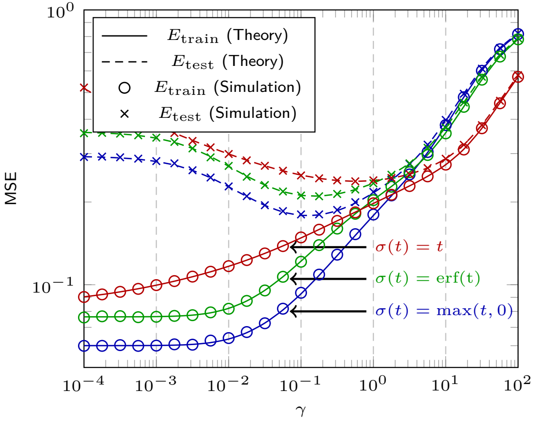

Figure: Performance for MNIST data (number 7 and 9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Example 2: Random-feature based spectral clustering

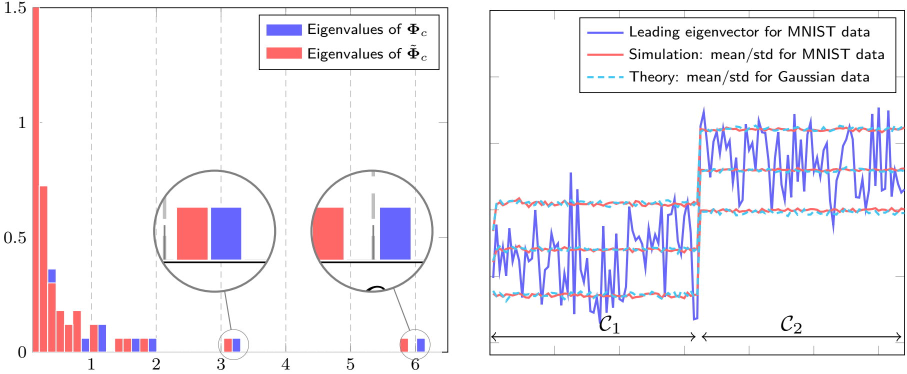

Figure: Eigenvalue distribution of $\mathbf{\Phi}_c$ and $\tilde{\mathbf{\Phi}}_c$ and leading eigenvector for MNIST data, with ± 1 standard deviations (from 500 trials). With the ReLU function, $p = 784, T = 128$ and $c_1 = c_2 = 1/2$.

Example 2: Random-feature based spectral clustering

Table: Empirical estimation of (normalized) statistics 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: Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the 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%
	$\exp(-t^2/2)$	60.44%	58.67%
balanced	$\mathrm{ReLU}(t)$	85.72%	82.27%

Table: Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the epileptic 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%
	$\exp(-t^2/2)$	**99.81%**	**99.77%**
balanced	$\mathrm{ReLU}(t)$	87.91%	90.97%

$\Rightarrow$Much better than ReLU in both cases

Example 3: Temporal evolution of training and generalization performance

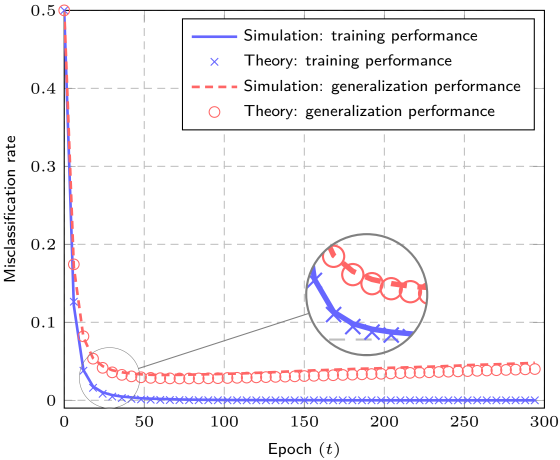

Figure: Training and generalization performance for MNIST data (number 1 and 7) with $n = p = 784$, $c_1 = c_2 = 1/2$, $\alpha = 0.01$ and $\sigma^2 = 0.1$. Simulation results obtained by averaging over 100 runs.

Conclusion

Take-away messages:

- SCMs or Gram-like matrices naturally appear in neural nets
- RMT is a powerful tool in the double-asymptotic regime
- difficulty from the nonlinear structure:

  $\Rightarrow$ handled with the "concentration" approach
- performance analyses and improvements of different algos
  - fast tunning of hyper-parameters
  - wise choice of activation vis-à-vis the data structure
  - more insights into training neural nets

Future work:

- deeper study of inner product kernels and activations in neural nets (relate $\sigma(\cdot)$ to d_1, d_2)
- the loss landscape of deep networks (beyond single-layer)

References

- On random feature maps and neural nets:
  - C. Louart, **Z. Liao**, R. Couillet, "A Random Matrix Approach to Neural Networks", (in press) Annals of Applied Probability, 2017.
  - **Z. Liao**, R. Couillet, "On the Spectrum of Random Features Maps of High Dimensional Data", (submitted to) The 35th International Conference on Machine Learning (ICML'18), Stockholm, Sweden, 2018.
- On the learning dynamics of neural nets:
  - **Z. Liao**, R. Couillet, "The Dynamics of Learning: A Random Matrix Approach", (submitted to) The 35th International Conference on Machine Learning (ICML'18), Stockholm, Sweden, 2018.
  - a journal in preparation with the co-supervisor Prof. Yacine CHITOUR
- On the kernel methods:
  - **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.
  - **Z. Liao**, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines", (submitted to) Journal of Machine Learning Research, 2017

Merci!