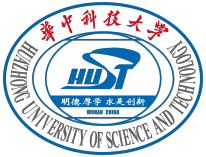
Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 1

CIMI Thematic School "Models & Methods for High-dimensional Inference and Learning"

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

October 17 and 18, 2024

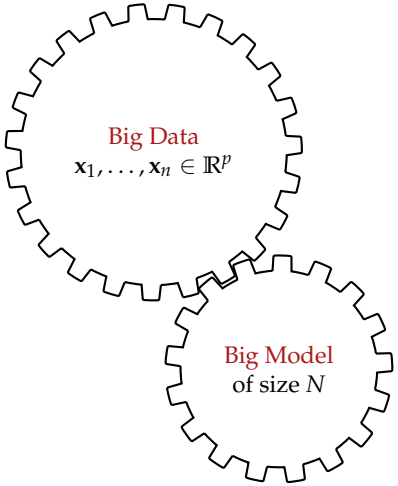


Schedule of the mini-course

1. **Part 1**: Motivation and Mathematical Background (concentration, resolvent-based approach to eigenspectral analysis, high-dimensional linearization, etc.)
2. **Part 2**: Four Ways to Characterize Sample Covariance Matrices and Some More Random Matrix Models (Wigner semicircle law, generalized sample covariance model, and separable covariance model)

Outline

Motivation: understanding large-dimensional machine learning



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics and machine learning (ML)
- **Random Matrix Theory (RMT)** provides the tools!

Sample covariance matrix in the large n, p regime

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,
- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

- due to **loss of matrix norm "equivalence"**: $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_{\max}$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_{\max} \equiv \max_{ij} |\mathbf{A}_{ij}|$.

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

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

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

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

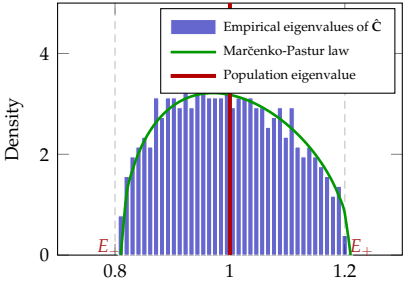


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

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

Noisy linear model

Consider a given set of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of size n, composed of the (random) input data $\mathbf{x}_i \in \mathbb{R}^p$ and its corresponding output target $y_i \in \mathbb{R}$, drawn from the following noisy linear model.

Definition (Noisy linear model)

We say a data-target pair $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ follows a noisy linear model if it satisfies

$$y = \boldsymbol{\beta}_*^\mathsf{T} \mathbf{x} + \epsilon \tag{1}$$

for some deterministic (ground-truth) vector $\boldsymbol{\beta}_* \in \mathbb{R}^p$, and random variable $\epsilon \in \mathbb{R}$ independent of $\mathbf{x} \in \mathbb{R}^p$, with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$.

- aim to find a regressor $\boldsymbol{\beta} \in \mathbb{R}^p$ that best describes the linear relation $y_i \approx \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i$, by minimizing the ridge-regularized mean squared error (MSE)

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i)^2 + \gamma \|\boldsymbol{\beta}\|^2 = \frac{1}{n} \|\mathbf{X}^\mathsf{T} \boldsymbol{\beta} - \mathbf{y}\|^2 + \gamma \|\boldsymbol{\beta}\|^2 \tag{2}$$

for $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, and some regularization penalty $\gamma \geq 0$

Out-of-sample prediction risk

- unique solution given by

$$\boldsymbol{\beta}_\gamma = \left(\mathbf{X}\mathbf{X}^\mathsf{T} + n\gamma \mathbf{I}_p\right)^{-1}\mathbf{X}\mathbf{y} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X} + n\gamma \mathbf{I}_n\right)^{-1}\mathbf{y}, \quad \gamma > 0 \tag{3}$$

- in the $\gamma = 0$ setting, the minimum ℓ_2 norm least squares solution

$$\boldsymbol{\beta}_0 = \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{+}\mathbf{X}\mathbf{y} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{+}\mathbf{y}, \tag{4}$$

  where $(\mathbf{A})^+$ denotes the Moore–Penrose pseudoinverse, also **"ridgeless" least squares solution**.
- statistical quality of $\boldsymbol{\beta}$, as a function of dimensions n, p, noise level σ^2, and the regularization γ
- evaluating the **out-of-sample prediction risk** (or simply, **risk**)

$$R_\mathbf{X}(\boldsymbol{\beta}) = \mathbb{E}[(\boldsymbol{\beta}^\mathsf{T}\hat{\mathbf{x}} - \boldsymbol{\beta}_*^\mathsf{T}\hat{\mathbf{x}})^2 \mid \mathbf{X}] = \underbrace{(\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*)^\mathsf{T}\mathbf{C}(\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*)}_{\equiv B_\mathbf{X}(\boldsymbol{\beta})} + \underbrace{\operatorname{tr}\left(\operatorname{Cov}[\boldsymbol{\beta} \mid \mathbf{X}]\mathbf{C}\right)}_{\equiv V_\mathbf{X}(\boldsymbol{\beta})} \tag{5}$$

  for an independent test data point. We denote $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{C}$, and $B_\mathbf{X}(\boldsymbol{\beta}), V_\mathbf{X}(\boldsymbol{\beta})$ the **bias** as well as **variance** of the solution $\boldsymbol{\beta}$.

Objects of interest

$$\boxed{\begin{aligned} B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) &= (\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*)^{\mathsf{T}} \mathbf{C} (\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{X}] - \boldsymbol{\beta}_*) \\ V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) &= \operatorname{tr}\left(\operatorname{Cov}[\boldsymbol{\beta} \mid \mathbf{X}] \mathbf{C}\right). \end{aligned}} \tag{6}$$

- Denote $\boxed{\mathbf{Q}(-\gamma) \equiv (\hat{\mathbf{C}} + \gamma \mathbf{I}_p)^{-1}}$ the resolvent of the SCM $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$. Write

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \boldsymbol{\beta}_*^{\mathsf{T}} \left(\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}}\right) \mathbf{C} \left(\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}}\right) \boldsymbol{\beta}_*, \quad V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \frac{\sigma^2}{n} \operatorname{tr}\left(\mathbf{Q}(-\gamma)\hat{\mathbf{C}}\mathbf{Q}(-\gamma)\mathbf{C}\right). \tag{7}$$

- For $\gamma > 0$, one has $\mathbf{I}_p - \mathbf{Q}(-\gamma)\hat{\mathbf{C}} = \mathbf{I}_p - \mathbf{Q}(-\gamma)(\hat{\mathbf{C}} + \gamma\mathbf{I}_p - \gamma\mathbf{I}_p) = \gamma\mathbf{Q}(-\gamma)$, so that

$$B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \gamma^2 \boldsymbol{\beta}_*^{\mathsf{T}} \mathbf{Q}^2(-\gamma) \boldsymbol{\beta}_* = \boxed{-\gamma^2 \frac{\partial \boldsymbol{\beta}_*^{\mathsf{T}} \mathbf{Q}(-\gamma) \boldsymbol{\beta}_*}{\partial \gamma}} \tag{8}$$

$$V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = \sigma^2 \left(\frac{1}{n} \operatorname{tr} \mathbf{Q}(-\gamma) - \frac{\gamma}{n} \operatorname{tr} \mathbf{Q}^2(-\gamma)\right) = \boxed{\sigma^2 \left(\frac{1}{n} \operatorname{tr} \mathbf{Q}(-\gamma) + \frac{\gamma}{n} \frac{\partial \operatorname{tr} \mathbf{Q}(-\gamma)}{\partial \gamma}\right)} \tag{9}$$

  where we used the fact that $\mathbf{C} = \mathbf{I}_p$ and $\partial \mathbf{Q}(-\gamma)/\partial\gamma = -\mathbf{Q}^2(-\gamma)$.
- suffice to evaluate quadratic and trace forms of the random resolvent matrix $\mathbf{Q}(-\gamma)$.

Numerical results

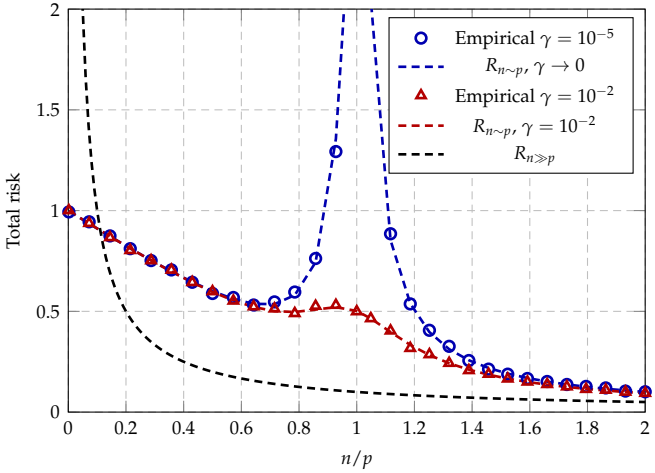


Figure: Out-of-sample risk $R_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) = B_{\mathbf{X}}(\boldsymbol{\beta}_\gamma) + V_{\mathbf{X}}(\boldsymbol{\beta}_\gamma)$ of the ridge regression solution $\boldsymbol{\beta}_\gamma$ as a function of the dimension ratio n/p, for fixed $p = 512$, $\|\boldsymbol{\beta}_*\| = 1$, and different regularization penalty $\gamma = 10^{-2}$ and $\gamma = 10^{-5}$, Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 0.1)$.

Linear model trained with gradient descent

- Consider again minimizing the following loss function to obtain the linear model parameter $\boldsymbol{\beta}$:

$$L(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)^2 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|^2 = \frac{1}{2n}\|\mathbf{X}^\mathsf{T}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|^2 \tag{10}$$

- but this time using gradient descent with infinitely small step size (i.e., gradient flow)

$$\frac{d\boldsymbol{\beta}(t)}{dt} = -\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Rightarrow \boldsymbol{\beta}(t) = e^{-(\hat{\mathbf{C}}+\gamma\mathbf{I}_p)t}\boldsymbol{\beta}(0) + \left(\mathbf{I}_p - e^{-(\hat{\mathbf{C}}+\gamma\mathbf{I}_p)t}\right)\boldsymbol{\beta}_{RR}, \tag{11}$$

  where we recall $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ the SCM and denote $\boldsymbol{\beta}_{\mathrm{RR}} = \left(\hat{\mathbf{C}} + \gamma\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}$ is the ridge regression solution (that corresponds to $\boldsymbol{\beta}(t)$ as $t \to \infty$)
- understand the interplay between **training dynamics** and **generalization performance**
- slightly more involved eigenspectral functional of $\hat{\mathbf{C}}$
- as well shall see below, writes as (complex counter) integration of the resolvent $\boxed{\mathbf{Q}(z) = (\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1}}$

Some numerical results

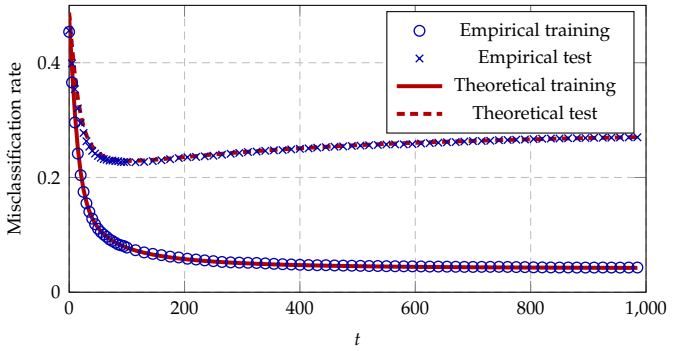


Figure: Training and test misclassification rates of a linear network as a function of the gradient descent training time t, for $p = 256$, $n = 512$, $\gamma = 0$, $\alpha = 10^{-2}$, $\sigma^2 = 0.1$ and $\boldsymbol{\mu} = [-\mathbf{1}_{p/2}, \mathbf{1}_{p/2}]/\sqrt{p}$. Empirical results averaged over 50 runs.

Scaling of sum of independent random variables: LLN and CLT

- **Strong law of large numbers (LLN)**: for a sequence of i.i.d. random variables $x_1, \ldots, x_n$ with the same expectation $\mathbb{E}[x_i] = \mu < \infty$, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to \mu, \tag{12}$$

  almost surely as $n \to \infty$.
- **Central limit theorem (CLT)**: for a sequence of i.i.d. random variables $x_1, \ldots, x_n$ with the same expectation $\mathbb{E}[x_i] = \mu$ and variance $\mathrm{Var}[x_i] = \sigma^2 < \infty$, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right) \to \mathcal{N}(0, \sigma^2), \tag{13}$$

  in distribution as $n \to \infty$.

Consequences of LLN and CLT

For i.i.d. random variables $x_1, \ldots, x_n$ of zero mean and unit variance, e.g., $x_i \sim \mathcal{N}(0,1)$, we have, for n large, the following scaling laws for the sum $\frac{1}{n}\sum_{i=1}^{n} x_i$:

- $\frac{1}{n}\sum_{i=1}^{n} x_i \simeq 0$ by LLN; and
- $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i = O(1)$ with high probability by CLT.

We have known this a bit in the context of DNN

- DNNs involve linear (i.e., weights) and nonlinear (i.e., activation) transformation
- **Xavier initialization** [GB10]: for sigmoid-type activation, randomly initialize a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ having N neurons as

$$[\mathbf{W}]_{ij} \sim \mathcal{N}(0, N^{-1}). \tag{14}$$

```
torch.nn.init.xavier_normal_
```

- **He initialization** [He+15]: for ReLU-type activation, randomly initialize a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ having N neurons as

$$[\mathbf{W}]_{ij} \sim \mathcal{N}(0, 2N^{-1}). \tag{15}$$

```
torch.nn.init.kaiming_normal_
```

- derivation based on **forward propagation**
- similar considerations for CNN, RNN, ResNet, etc.

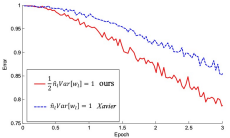


Figure 2. The convergence of a **22-layer** large model (B in Table 3). The x-axis is the number of training epochs. The y-axis is the top-1 error of 3,000 random val samples, evaluated on the center crop. We use ReLU as the activation for both cases. Both our initialization (red) and "Xavier" (blue) [7] lead to convergence, but ours starts reducing error earlier.

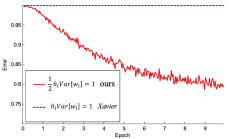


Figure 3. The convergence of a **30-layer** small model (see the main text). We use ReLU as the activation for both cases. Our initialization (red) is able to make it converge. But "Xavier" (blue) [7] completely stalls - we also verify that its gradients are all diminishing. It does not converge even given more epochs.

Figure: Numerical results in [He+15] for moderately deep NN.

Let us say more on the appropriate scaling of large and deep NNs

Setup and Notations:

- supervised training of an L-layer multi-layer perceptrons (MLP) with full batch gradient flow
- input data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, denote pre-activation vectors $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^N$ at layer $\ell \in \{1, \ldots, L\}$ as

$$\boxed{\mathbf{h}_i^{(1)} = \frac{1}{N^{a_1}\sqrt{p}}\mathbf{W}^{(1)}\mathbf{x}_i, \quad \mathbf{h}_i^{(\ell)} = \frac{1}{N^{a_\ell}}\mathbf{W}^{(\ell)}\sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)} \quad i \in \{1, \ldots, n\} \tag{16}$$

- scalar output $\boxed{f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \frac{1}{\gamma N^{a_L}}\left(\mathbf{w}^{(L)}\right)^{\mathsf{T}}\sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)}$ for trainable parameters $\boldsymbol{\theta} = \{\mathbf{W}^{(1)}, \ldots, \mathbf{w}^{(L)}\}$.
- for a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, train the above DNN on the loss function $L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n L(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$, with full-batch gradient flow

$$\frac{d\boldsymbol{\theta}}{dt} = -\eta\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \eta\frac{1}{n}\sum_{i=1}^n \Delta_i \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \boldsymbol{\theta}}, \quad \Delta_i \equiv -\frac{\partial L(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}, \tag{17}$$

  learning rate $\boxed{\eta = \eta_0\gamma^2 N^{-c}}$ and **feature learning parameter** $\boxed{\gamma = \gamma_0 N^d}$ for $\eta_0 = \Theta(1)$ and $\gamma_0 = \Theta(1)$
- **initialization scaling scheme**: $\boxed{w_i^{(L)} \sim \mathcal{N}(0, N^{-b_L}), W_{ij}^{(\ell)} \sim \mathcal{N}(0, N^{-b_\ell})}$ and $W_{ij}^{(1)} \sim \mathcal{N}(0, N^{-b_1})$

[1]This part is majorly borrowed from the Lecture Notes on Infinite-Width Limits of Neural Networks, by Cengiz Pehlevan and Blake Bordelon, *Princeton Machine Learning Theory Summer School*, 2023.

Appropriate scaling of large and deep NNs

Settings:

- **scaling of NN model**: $\mathbf{h}_i^{(1)} = \frac{1}{N^{a_1}\sqrt{p}}\mathbf{W}^{(1)}\mathbf{x}_i$, $\mathbf{h}_i^{(\ell)} = \frac{1}{N^{a_\ell}}\mathbf{W}^{(\ell)}\sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)$, $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \frac{1}{\gamma N^{a_L}}\left(\mathbf{w}^{(L)}\right)^\intercal \sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)$
- **initialization scaling**: $w_i^{(L)} \sim \mathcal{N}(0, N^{-b_L})$, $W_{ij}^{(\ell)} \sim \mathcal{N}(0, N^{-b_\ell})$, and $W_{ij}^{(1)} \sim \mathcal{N}(0, N^{-b_1})$
- **trained under full-batch gradient flow**: $\frac{d\boldsymbol{\theta}}{dt} = -\eta\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \eta\frac{1}{n}\sum_{i=1}^n \Delta_i \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \boldsymbol{\theta}}$ of learning rate $\eta = \eta_0\gamma^2 N^{-c}$ and feature learning parameter $\gamma = \gamma_0 N^d$ for $\eta_0 = \Theta(1)$ and $\gamma_0 = \Theta(1)$

Objective: for large p, N, achieve **appropriate scaling** on (a, b, c, d) so that

1. **pre-activations $\mathbf{h}^{(\ell)}$ have $\Theta(1)$ entries**:
   - computing the 1st and 2nd moments of $\mathbf{h}^{(1)}$: $\mathbb{E}[\mathbf{h}_i^{(1)}] = \mathbf{0}$, $\mathbb{E}[\mathbf{h}_i^{(1)}(\mathbf{h}_j^{(1)})^\intercal]_{kq} = \delta_{kq}N^{-(2a_1+b_1)} \cdot \frac{1}{p}\mathbf{x}_i^\intercal\mathbf{x}_j$; then of $\mathbf{h}^{(\ell)}$
   - we get $\boxed{2a_1 + b_1 = 1}$ and similarly $\boxed{2a_\ell + b_\ell = 1, \ell \in \{1, \ldots, L\}}$
2. **network prediction evolve in $\Theta(1)$ time**:
   - define **feature/conjugate kernel** as the Gram matrix at layer ℓ as $\boldsymbol{\Phi}^{(\ell)} \in \mathbb{R}^{n\times n}, \Phi_{ij}^{(\ell)} = \frac{1}{N}\sigma(\mathbf{h}_i^{(\ell)})^\intercal\sigma(\mathbf{h}_j^{(\ell)})$
   - under the condition of $\Theta(1)$ pre-activation, it can be shown that in the $N \to \infty$ limit that the pre-activations are **Gaussian process** of zero mean, and covariance given by the (expected) conjugate kernel
   - for $\partial_t f_{\boldsymbol{\theta}}(\cdot) = \Theta(1)$, we get $\boxed{2a_1 + c = 0}$ and $\boxed{2a_\ell + c = 1, \ell \in \{2, \ldots, L\}}$

Appropriate scaling of large and deep NNs

Settings:

- **scaling of NN model**: $\mathbf{h}_i^{(1)} = \frac{1}{N^{a_1}\sqrt{p}}\mathbf{W}^{(1)}\mathbf{x}_i$, $\mathbf{h}_i^{(\ell)} = \frac{1}{N^{a_\ell}}\mathbf{W}^{(\ell)}\sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)$, $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \frac{1}{\gamma N^{a_L}}\left(\mathbf{w}^{(L)}\right)^\top \sigma_\ell\left(\mathbf{h}_i^{(\ell-1)}\right)$
- **initialization scaling**: $w_i^{(L)} \sim \mathcal{N}(0, N^{-b_L})$, $W_{ij}^{(\ell)} \sim \mathcal{N}(0, N^{-b_\ell})$, and $W_{ij}^{(1)} \sim \mathcal{N}(0, N^{-b_1})$
- **trained under full-batch gradient flow**: $\frac{d\boldsymbol{\theta}}{dt} = -\eta\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \eta\frac{1}{n}\sum_{i=1}^n \Delta_i \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \boldsymbol{\theta}}$ of learning rate $\eta = \eta_0\gamma^2 N^{-c}$ and feature learning parameter $\gamma = \gamma_0 N^d$ for $\eta_0 = \Theta(1)$ and $\gamma_0 = \Theta(1)$

Objective: for large p, N, achieve **appropriate scaling** on (a, b, c, d) so that

3. **features evolve in $\Theta(1)$ time**:
   - by $\partial_t \mathbf{h}_i^{(\ell)} = \Theta(1)$ we have $2a_1 + c - d + 1/2 = 0$, recall that $2a_1 + c = 0$, this is $d = 1/2$, similarly $2a_\ell + c - d - 1/2 = 0$ so that $\boxed{d = 1/2}$
   - in fact, any $d < 1/2$ leads to kernel behavior, and $d = 0$ the **NTK parameterization**

- if further demand raw learning rate $\eta = \Theta(1)$, then parameterization is unique:

$$\boxed{d = 1/2, c = 1, a_\ell = 0, b_\ell = 1, a_1 = -1/2, b_1 = 1} \tag{18}$$

- this is equivalent to the muP parameterization in [YH21]

What is good about this appropriate scaling

- well, things (e.g., DNN pre-activation, evolution of prediction and feature/pre-activation with respect to time) do not scale with the network width N
- BTW, in the case of **ResNet**, a scaling scheme of a similar type can be obtained by considering the infinitely deep $L \to \infty$ limit [Bor+23]
- idea of **maximal update parameterization (muP)** for **hyperparameter transfer** in large models (G. Yang)
- in muP, "narrow" and wide neural networks **share the same set of optimal hyperparameters**, e.g., optimal learning rate (and decay), cross-entropy temperature, initialization scale, regularization, etc.
- one can tune the large model **by just tuning a tiny version** of it and copying over the hyperparameters

[2] Blake Bordelon et al. "Depthwise Hyperparameter Transfer in Residual Networks: Dynamics and Scaling Limit". In: *The Twelfth International Conference on Learning Representations*. Oct. 2023

Some experiments on muP and µTransfer

Show some simulations!

Some experiments on muP and µTransfer

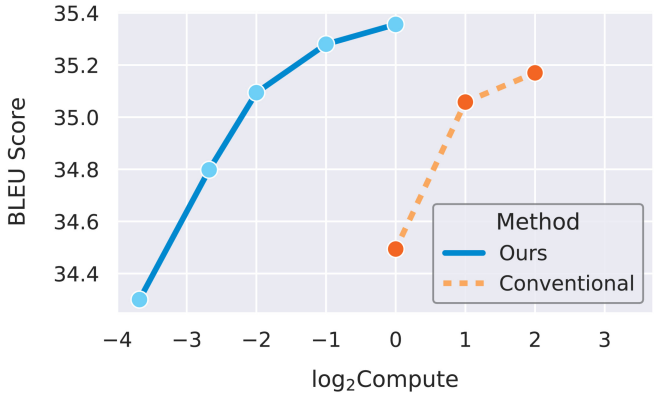


Figure: Comparison µTransfer, which transfers tuned hyperparameters from a small proxy model, with directly tuning the large target model, on IWSLT14 De-En, a machine translation dataset.

Take-away of this section

- sample covariance matrix $\hat{\mathbf{C}}$ have different behavior in the large n, p regime
- loss of matrix norm "equivalence" for large matrices $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_{\max}$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_{\max} \equiv \max_{ij} |\mathbf{A}_{ij}|$
- evaluation of linear regression model trained with gradient descent involves **eigenspectral functionals** of SCM, **RMT** provides **an analytic answer**
- further allows better **understanding and scaling** of large and deep neural networks

Summary: analyze and optimize large-scale ML models

Definition (High-dimensional Equivalent)

For a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ and a (possibly) nonlinear model of interest $f(\mathbf{X})$ of $\mathbf{X}$ for some $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times n}$, we are interested in the behavior of the scalar observation $g(f(\mathbf{X}))$ of the random model $f(\mathbf{X})$, via the observation map $g\colon \mathbb{R}^{p\times n} \to \mathbb{R}$.
We say that $\bar{\mathbf{X}}_f$ (which may be deterministic or random) is an High-dimensional Equivalent for the random model $f(\mathbf{X})$ with respect to the observation map g if we have, with probability at least $1-\delta(p,n)$ that

$$\left| \frac{g(f(\mathbf{X})) - g(\mathbf{X}_f)}{g(f(\mathbf{X}))} \right| \leq \varepsilon(n,p), \tag{19}$$

for some non-negative functions $\varepsilon(n,p)$ and $\delta(n,p)$ that decrease to zero as $n,p \to \infty$.

Summary: analyze and optimize large-scale ML models

Analyze and Optimize Large-scale ML model $f(\mathbf{X}, \mathbf{\Theta})$

Objective: Evaluation of $f(\mathbf{X}, \mathbf{\Theta})$ via Performance Metric $g(\cdot)$

Technical Challenge 1 High-dimensionality in $\mathbf{X}, \mathbf{\Theta}$	**Key Idea 1** Concentration of $g(f(\mathbf{X}, \mathbf{\Theta})) \simeq \mathbb{E}[g(f(\mathbf{X}, \mathbf{\Theta}))]$
Technical Challenge 2 Analysis of Eigen-functional	**Key Idea 2** Leave-one-out + complex analysis
Technical Challenge 3 Non-linearity in ML model	**Key Idea 3** High-dimensional linearization of $f(\mathbf{X}, \mathbf{\Theta})$

Characterization of scalar random variables: from moments to tails

Definition (Moments and moment generating function, MGF)

For a scalar random variable x defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote

- $\mathbb{E}[x]$ the *expectation* of x;
- $\mathrm{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$ the *variance* of x;
- for $p > 0$, $\mathbb{E}[x^p]$ the p^{th} *moment* of x, and $\mathbb{E}[|x|^p]$ the p^{th} *absolute moment*;
- for $\lambda \in \mathbb{R}$, $M_x(\lambda) = \mathbb{E}[e^{\lambda x}] = \sum_{p=0}^{\infty} \frac{\lambda^p}{p!} \mathbb{E}[x^p]$ the *moment generating function (MGF)* of x.

Lemma (Moments versus tails)

For a scalar random variable x and fixed $p > 0$, we have

1. $\mathbb{E}[|x|^p] = \int_0^{\infty} p t^{p-1} \mathbb{P}(|x| \geq t)\, dt$
2. $\mathbb{P}(|x| \geq t) \leq \exp(-\lambda t) M_{|x|}(\lambda)$, *for $t > 0$ and MGF $M_{|x|}(\lambda)$ of $|x|$*

Definition (Sub-gaussian and sub-exponential distributions)

For a standard Gaussian random variable $x \sim \mathcal{N}(0,1)$, its law given by $\mu(dt) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$, so that $\mathbb{P}(x \geq X) = \mu([X,\infty)) = \frac{1}{\sqrt{2\pi}} \int_X^\infty \exp(-t^2/2)\, dt \leq \exp(-X^2/2)$.

- We say y is a *sub-gaussian random variable* if it has a tail that decays *as fast as* standard Gaussian random variables, that is
$$\mathbb{P}\left(|y| \geq t\right) \leq \exp(-t^2/\sigma_{\mathcal{N}}^2), \tag{20}$$
for some $\sigma_{\mathcal{N}} > 0$ (known as the *sub-gaussian norm* of y) for all $t > 0$.
- We can define a *sub-exponential random variable* z similarly via $\mathbb{P}(|z| \geq t) \leq \exp(-t/\sigma_{\mathcal{N}})$.

- for a sub-gaussian random variable x of mean $\mu = \mathbb{E}[x]$ and sub-gaussian norm $\sigma_{\mathcal{N}}$ that
$$\mathbb{P}\left(|x - \mu| \geq t\sigma_{\mathcal{N}}\right) \leq \exp(-t^2), \tag{21}$$
for all $t > 0$, in which the sub-gaussian norm $\sigma_{\mathcal{N}}$ of x acts as a scale parameter (that is similar, in spirit, to the variance parameter of Gaussian distribution).

A collection of scalar random variables: from LLN to CLT

For a collection of independent and identically distributed (i.i.d.) random variables $x_1, \ldots, x_n$ of mean μ and variance σ^2, we have, by independence, that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mu, \quad \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] = \frac{\sigma^2}{n}. \tag{22}$$

- for μ, σ^2 do *not* scale with n, the (random) sample mean strongly concentrates around its expectation μ.

Theorem (Weak and strong law of large numbers, LLN)

For a sequence of i.i.d. random variables $x_1, \ldots, x_n$ with finite expectation $\mathbb{E}[x_i] = \mu < \infty$, we have that the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to \mu, \tag{23}$$

*in probability/almost surely as $n \to \infty$, known as the **weak law/strong of large numbers (LLN)**.*

Theorem (Central limit theorem, CLT)

For a sequence of i.i.d. random variables $x_1, \ldots, x_n$ with $\mathbb{E}[x_i] = \mu$ and $\mathrm{Var}[x_i] = \sigma^2$, we have, for every $t \in \mathbb{R}$ that

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) \geq t\right) \to \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}\,dx \tag{24}$$

as $n \to \infty$. That is, as $n \to \infty$, the random variable $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) \to \mathcal{N}(0,1)$ in distribution.

Remark (Unified form of LLN and CLT)

The results of LLN and CLT can be compactly written as $\frac{1}{n}\sum_{i=1}^{n} x_i \simeq \underbrace{\mu}_{O(1)} + \underbrace{\mathcal{N}(0,1)\cdot\sigma/\sqrt{n}}_{O(n^{-1/2})}$, as $n \to \infty$, for μ, σ both of order $O(1)$.

(i) In the first order (of magnitude $O(1)$), it has an asymptotically deterministic behavior around the expectation μ; and

(ii) in the second order (of magnitude $O(n^{-1/2})$), it strongly concentrates around this deterministic quantity with a universal Gaussian fluctuation, regardless of the distribution of the component of x_i.

Concentration of random vectors in high dimensions?

- "concentration" for a random vector $\mathbf{x} \in \mathbb{R}^n$?

Observation (Random vectors do not "concentrate" around their means)

For two *independent* random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, having i.i.d. entries with zero mean and unit variance (that is, $\mu = 0$ and $\sigma = 1$), we have that

$$\mathbb{E}[\|\mathbf{x} - \mathbf{0}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] = \mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]) = n, \tag{25}$$

and further by independence that

$$\mathbb{E}[\|\mathbf{x} - \mathbf{y}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x} + \mathbf{y}^\mathsf{T}\mathbf{y}] = 2n. \tag{26}$$

- the origin $\mathbf{0}$ (and *mean* of $\mathbf{x}$) is always, in expectation, at the midpoint of two independent draws of random vectors in $\mathbb{R}^n$
- any random vector $\mathbf{x} \in \mathbb{R}^n$ with n large is not close to its mean
- $\mathbf{x}$ does not itself "concentrate" around any n-dimensional deterministic vector in any traditional sense.

Numerical illustration

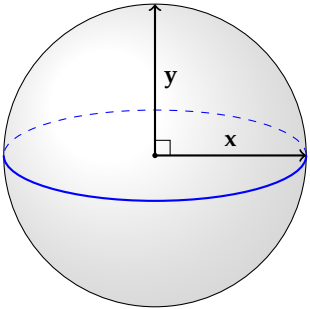


Figure: Visualization of "non-concentration" behavior of large-dimensional random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

- In spite of this, from the LLN and CLT one expects that some types of "observations" of $\mathbf{x} \in \mathbb{R}^n$ (e.g., averages over all the entries of $\mathbf{x}$, to retrieve the sample mean), must concentrate in some sense for n large
- we "interpret" the sample mean as a linear scalar observation of a vector $\mathbf{x} \in \mathbb{R}^n$.

Remark (Sample mean as a linear scalar observation)

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector having i.i.d. entries, then the sample mean of the entries of $\mathbf{x}$ can be rewritten as the following linear scalar observation $f\colon \mathbb{R}^n \to \mathbb{R}$ of $\mathbf{x}$ defined as

$$f(\mathbf{x}) = \mathbf{1}_n^{\mathsf{T}}\mathbf{x}/n = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ or } f(\cdot) = \mathbf{1}_n^{\mathsf{T}}(\cdot)/n. \tag{27}$$

- LLN and CLT are nothing but asymptotic characterization of the concentration behavior of the linear scalar observation $f(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathbb{R}^n$
- we can say things non-asymptotically as well, under two different assumptions on the tail of $\mathbf{x}$.
  - (i) are only assumed to have finite variance σ^2 (but nothing on its tail behavior or higher-order moments); and
  - (ii) have sub-gaussian tails with sub-gaussian norm $\sigma_{\mathcal{N}}$.

Table: Different types of characterizations of the linear scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ for $\mathbf{x} \in \mathbb{R}^n$, having i.i.d. entries with mean $\mathbb{E}[x_i] = \mu$ and variance σ^2 or sub-gaussian norm $\sigma_{\mathcal{N}}$.

	First-order behavior	Second-order behavior
Asymptotic	$f(\mathbf{x}) \to \mu$ via Law of Large Numbers	$\frac{\sqrt{n}}{\sigma}(f(\mathbf{x}) - \mu) \to \mathcal{N}(0,1)$ in law Central Limit Theorem
Non-asymptotic under finite variance	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma/\sqrt{n}\right) \leq t^{-2}$ via Chebyshev's inequality
Non-asymptotic under sub-gaussianity	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma_{\mathcal{N}}/\sqrt{n}\right) \leq \exp(-Ct^2)$ via sub-gaussian tail bound

Remark (Concentration of scalar observation of large random vectors)

A random vector $\mathbf{x} \in \mathbb{R}^n$, when "observed" via the linear scalar observation $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$:

$$f(\mathbf{x}) \simeq \underbrace{\mu}_{O(1)} + \underbrace{X/\sqrt{n}}_{O(n^{-1/2})}, \tag{28}$$

for n large, with some random X of order $O(1)$ that:

(i-i) has a tail that decays (at least) as t^{-2}, for finite n and $\mathbf{x}$ having entries of bounded variance;

(i-ii) has a sub-gaussian tail (at least) as $\exp(-t^2)$, for finite n and $\mathbf{x}$ having sub-gaussian entries;

(ii) has a precise Gaussian tail *independent* of the law of (the entries of) $\mathbf{x}$, but in the limit of $n \to \infty$ via CLT.

Lipschitz, quadratic concentration, and beyond

The concentration properties extend beyond the specific *linear* observation, $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$, to many types of (possibly) nonlinear observations.

Definition (Scalar observation maps)

For random vector $\mathbf{x} \in \mathbb{R}^n$, we say $f(\mathbf{x}) \in \mathbb{R}$ is a scalar observation of $\mathbf{x}$ with observation map $f\colon \mathbb{R}^n \to \mathbb{R}$.

Table: Different types of scalar observations $f(\mathbf{x})$ of random vector $\mathbf{x} \in \mathbb{R}^n$, having independent entries.

	Scalar observation	Characterization
Linear	sample mean $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$, and $f(\mathbf{x}) = \mathbf{a}^\mathsf{T}\mathbf{x}$ for $\mathbf{a} \in \mathbb{R}^n$	Table in last slide
Lipschitz	$f(\mathbf{x})$ for a Lipschitz map $f\colon \mathbb{R}^n \to \mathbb{R}$	Lipschitz concentration
Quadratic form	$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ for some $\mathbf{A} \in \mathbb{R}^{n\times n}$	Hanson–Wright inequality
Nonlinear quadratic form	$f(\mathbf{x}) = \sigma(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\sigma(\mathbf{Y}^\mathsf{T}\mathbf{x})$ for entry-wise $\sigma\colon \mathbb{R} \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{Y} \in \mathbb{R}^{p\times n}$	Nonlinear quadratic concentration, of direct use in NN

Theorem (Concentration of Lipschitz map of Gaussian random vectors, [Ver18, Theorem 5.2.2])

For a standard Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ *and a Lipschitz function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *that satisfies* $|f(\mathbf{y}_1) - f(\mathbf{y}_2)| \leq K_f \|\mathbf{y}_1 - \mathbf{y}_2\|_2$ *for any* $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$*, we have, for all* $t > 0$ *that*

$$\mathbb{P}\left(|f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]| \geq t\right) \leq \exp(-Ct^2/K_f^2), \tag{29}$$

for some universal constant $C > 0$*, with* $K_f > 0$ *known as the Lipschitz constant of* f*.*

Remark (Concentration of Lipschitz observation of large random vectors)

The Lipschitz scalar observations $f(\mathbf{x})$ *of the random vector* $\mathbf{x} \in \mathbb{R}^n$ *behave as*

$$f(\mathbf{x}) \simeq \underbrace{\mathbb{E}[f(\mathbf{x})]}_{O(1)} + \underbrace{K_f}_{O(n^{-1/2})}, \tag{30}$$

for n *large, where* K_f *is the Lipschitz constant of* f *that is, in general, of order* $O(n^{-1/2})$ *for* $\mathbb{E}[f(\mathbf{x})] = O(1)$*, for example for* $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$*.*

[3] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and

Concentration of quadratic forms

- intuitively expect that non-Lipschitz observation $f(\mathbf{x})$ still concentrates in some way, but "less so"

Theorem (Hanson–Wright inequality for quadratic forms, [Ver18, Theorem 6.2.1])

For a random vector $\mathbf{x} \in \mathbb{R}^n$ having independent, zero-mean, unit-variance, sub-gaussian entries with sub-gaussian norm bounded by $\sigma_{\mathcal{N}}$, and deterministic matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we have, for every $t > 0$, that

$$\mathbb{P}\left(\left|\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - \operatorname{tr}\mathbf{A}\right| \geq t\right) \leq \exp\left(-\frac{C}{\sigma_{\mathcal{N}}^2}\min\left(\frac{t^2}{\sigma_{\mathcal{N}}^2\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2}\right)\right), \tag{31}$$

for some universal constant $C > 0$.

- depending on the interplay between the "range" t and the deterministic matrix $\mathbf{A}$, the random quadratic form $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ swings between a sub-gaussian ($\exp(-t^2)$) and a sub-exponential ($\exp(-t)$) tail

Remark (Concentration of Euclidean norm of large random vectors)

It follows that the squared Euclidean norm $\|\mathbf{x}\|_2^2$, as a (non-Lipschitz) quadratic observation of $\mathbf{x} \in \mathbb{R}^n$, behaves as

$$\frac{1}{n}\|\mathbf{x}\|_2^2 \simeq 1 + O(n^{-1/2}), \quad n \gg 1. \tag{32}$$

Concentration of nonlinear quadratic forms

- nonlinear quadratic forms $\frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x})$ for Gaussian $\mathbf{x} \in \mathbb{R}^p$ and deterministic $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{Y} \in \mathbb{R}^{p\times n}$

Theorem (Concentration of nonlinear quadratic forms, [LLC18, Lemma 1])

For a standard Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and deterministic $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{Y} \in \mathbb{R}^{p\times n}$ such that $\|\mathbf{A}\|_2 \le 1, \|\mathbf{Y}\|_2 = 1$, we have, for Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant K_f and any $t > 0$ that

$$\mathbb{P}\left(\left| \frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x}) - \frac{1}{n} \operatorname{tr} \mathbf{A}\mathbf{K}_f(\mathbf{Y}) \right| \ge \frac{t}{\sqrt{n}} \right) \le \exp\left(-\frac{C}{K_f^2} \min\left(\frac{t^2}{(|f(0)| + K_f\sqrt{p/n})^2}, \sqrt{n}t \right) \right), \tag{33}$$

with $\mathbf{K}_f(\mathbf{Y}) = \mathbb{E}_\mathbf{x}[f(\mathbf{Y}^\mathsf{T}\mathbf{x}) f(\mathbf{x}^\mathsf{T}\mathbf{Y})] \in \mathbb{R}^{n\times n}$, for some universal constant $C > 0$.

- a nonlinear extension of the Hanson–Wright inequality (consider, e.g., $\mathbf{Y} = \mathbf{I}_n$ with $p = n$)

Remark (Concentration of nonlinear quadratic form observation of large random vectors):

$$\frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x}) \simeq \frac{1}{n} \operatorname{tr} \mathbf{A}\mathbf{K}_f(\mathbf{Y}) + O(n^{-1/2}), \tag{34}$$

for n large, with $\max\{f(0), K_f, p/n\} = O(1)$, and similar first and second order behavior as above.

[4] Cosme Louart, Zhenyu Liao, and Romain Couillet. "A random matrix approach to neural networks". In: *Annals of Applied Probability* 28.2 (2018), pp. 1190–1248

Take-away of this section

- high-dimensional random vectors are **not** "concentrating", but orthogonal
- scalar observation $f(\mathbf{x})$ of large random vector $\mathbf{x}$ **does concentrate**: linear, Lipschitz, quadratic form, and nonlinear quadratic forms, etc.
- same holds for random matrices, leads to **Deterministic Equivalent** for random matrices with respect to observation $g(\cdot)$

Definition (**High-dimensional Deterministic Equivalent**)

We say that $\bar{\mathbf{Q}} \in \mathbb{R}^{p\times p}$ is an $(\varepsilon_1, \varepsilon_2, \delta)$-**Deterministic Equivalent** for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ if, for a deterministic matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of unit norms (spectral and Euclidean, respectively), we have, with probability at least $1-\delta(p)$ that

$$\left|\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}})\right| \leq \varepsilon_1(p), \quad \left|\mathbf{a}^{\mathsf{T}}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b}\right| \leq \varepsilon_2(p), \tag{35}$$

for some non-negative functions $\varepsilon_1(p), \varepsilon_2(p)$ and $\delta(p)$ that decrease to zero as $p \to \infty$. To denote this relation, we use the notation

$$\mathbf{Q} \overset{\varepsilon_1,\varepsilon_2,\delta}{\longleftrightarrow} \bar{\mathbf{Q}}, \text{ or simply } \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}. \tag{36}$$

A quick recap on linear algebra: matrices

Definition (Matrix inner product and Frobenius norm)

Given matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$,

- $\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{Y}]_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ji}Y_{ji}$ is the matrix inner product between $\mathbf{X}$ and $\mathbf{Y}$, where $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$; and
- $\|\mathbf{X}\|_F^2 = \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{X}]_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ji}^2$ denotes the (squared) Frobenius norm of $\mathbf{X}$, which is also the sum of the squared entries of $\mathbf{X}$.

Definition (Matrix norm)

For $\mathbf{X} \in \mathbb{R}^{p \times n}$, the following "entry-wise" extension of the p-norms of vectors.

1. matrix Frobenius norm $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2} = \|\mathrm{vec}(\mathbf{X})\|_2$ that extends the vector ℓ_2 Euclidean norm; and
2. matrix maximum norm $\|\mathbf{X}\|_{\max} = \max_{i,j} |X_{ij}| = \|\mathrm{vec}(\mathbf{X})\|_\infty$ that extends the vector ℓ_∞ norm.

and also matrix norm induced by vectors: $\|\mathbf{X}\|_p \equiv \sup_{\|\mathbf{v}\|_p = 1} \|\mathbf{X}\mathbf{v}\|_p$.

- taking $p = 2$ is the spectral norm: $\|\mathbf{X}\|_2 = \sqrt{\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})} = \sigma_{\max}(\mathbf{X})$, with $\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})$ and $\sigma_{\max}(\mathbf{X})$ the maximum eigenvalue and singular of $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}$, respectively.

Remark (Matrix norm "equivalence")

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$*, one has the following*

1. $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{\operatorname{rank}(\mathbf{A})} \cdot \|\mathbf{A}\|_2 \leq \sqrt{\max(m,n)} \cdot \|\mathbf{A}\|_2$*, so that the control of the spectral norm via the Frobenius norm can be particularly loose for matrices of large rank; and*
2. $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_2 \leq \sqrt{mn} \cdot \|\mathbf{A}\|_{\max}$*, with* $\|\mathbf{A}\|_{\max} \equiv \max_{i,j} |A_{ij}|$ *the max norm of* $\mathbf{A}$*, so that the max and spectral norm can be significantly different for matrices of large size.*

- matrix norm "equivalence" holds only up to **dimensional factors** (e.g., rank and size)

Definition (Eigen-decomposition of symmetric matrices)

A symmetric real matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ admits the following eigen-decomposition

$$\mathbf{X} = \mathbf{U}_{\mathbf{X}}\mathbf{\Lambda}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^{\mathsf{T}} = \sum_{i=1}^{n} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}, \tag{37}$$

for diagonal $\mathbf{\Lambda}_{\mathbf{X}} = \operatorname{diag}\{\lambda_i(\mathbf{X})\}_{i=1}^{n}$ containing $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$ the real eigenvalues of $\mathbf{X}$, and orthonormal $\mathbf{U}_{\mathbf{X}} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ containing the corresponding eigenvectors. In particular,

$$\mathbf{X}\mathbf{u}_i = \lambda_i(\mathbf{X})\mathbf{u}_i. \tag{38}$$

- interested in a single eigenvalue of a symmetric real matrix, $\mathbf{X} \in \mathbb{R}^{n\times n}$, one may either resort to the eigenvalue-eigenvector equation in (38) or the determinant equation $\det(\mathbf{X} - \lambda\mathbf{I}_n) = 0$
- classical RMT is interested in the *joint* behavior of all eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$, e.g., the (empirical) eigenvalue distribution of $\mathbf{X}$

Definition (Empirical Spectral Distribution, ESD)

For a real symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, the *empirical spectral distribution (ESD)* or *empirical spectral measure* $\mu_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$ of $\mathbf{X}$,

$$\mu_{\mathbf{X}} \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{X})}, \tag{39}$$

where δ_x represents the Dirac measure at x. Since $\int \mu_{\mathbf{X}}(dx) = 1$, the spectral measure $\mu_{\mathbf{X}}$ of a matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ (which may be random or not) is a probability measure.

- $\int t\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{X})$ is the first moment of $\mu_{\mathbf{X}}$, and gives the average of all eigenvalues of $\mathbf{X}$; and
- $\int t^2\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2(\mathbf{X})$ is the second moment of $\mu_{\mathbf{X}}$, so that $\int t^2\mu_{\mathbf{X}}(dt) - \left(\int t\mu_{\mathbf{X}}(dt)\right)^2$ gives the variance of the eigenvalues of $\mathbf{X}$.

A unified spectral analysis approach via the resolvent

- **Note**: here everything hold deterministically, not necessarily random **yet**
- combined with **Deterministic Equivalent** and concentration, gives the whole picture

Definition (Resolvent)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as

$$\mathbf{Q}_{\mathbf{X}}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}. \tag{40}$$

Proposition (Properties of resolvent)

For $\mathbf{Q}_{\mathbf{X}}(z)$ the resolvent of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$ with ESD $\mu_{\mathbf{X}}$ with supported on $\mathrm{supp}(\mu_{\mathbf{X}})$, then

(i) $\mathbf{Q}_{\mathbf{X}}(z)$ is complex analytic on its domain of definition $\mathbb{C} \setminus \mathrm{supp}(\mu_{\mathbf{X}})$;

(ii) it is bounded in the sense that $\|\mathbf{Q}_{\mathbf{X}}(z)\|_2 \leq 1/\operatorname{dist}(z, \mathrm{supp}(\mu_{\mathbf{X}}))$;

(iii) $x \mapsto \mathbf{Q}_{\mathbf{X}}(x)$ for $x \in \mathbb{R} \setminus \mathrm{supp}(\mu_{\mathbf{X}})$ is an increasing matrix-valued function with respect to symmetric matrix partial ordering (i.e., $\mathbf{A} \succeq \mathbf{B}$ whenever $\mathbf{z}^\mathsf{T}(\mathbf{A} - \mathbf{B})\mathbf{z} \geq 0$ for all $\mathbf{z}$).

A unified spectral analysis approach via the resolvent

- for real z, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ is nothing but a regularized inverse of $\mathbf{X}$
- when interested in the eigenvalues and eigenvectors of $\mathbf{X} \in \mathbb{R}^{p \times p}$, consider the eigenvalue and eigenvector equation

$$\mathbf{X}\mathbf{v} = \lambda \mathbf{v} \Leftrightarrow (\mathbf{X} - \lambda \mathbf{I}_p)\mathbf{v} = \mathbf{0}, \quad \lambda \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^p, \tag{41}$$

  for an eigenvalue-eigenvector pair $(\lambda, \mathbf{v})$ of $\mathbf{X}$ with $\mathbf{v} \neq \mathbf{0}$
- again a linear system, but solving for a pair of eigenvalue and eigenvector $(\lambda, \mathbf{v})$ for which the inverse/resolvent $(\mathbf{X} - \lambda \mathbf{I}_p)^{-1}$ does not exist
- while seemingly less convenient at first sight, turns out to be very efficient in providing a unified assess to general spectral functionals of $\mathbf{X}$, by taking z to be complex and exploiting tools from **complex analysis**

Theorem (Cauchy's integral formula)

For $\Gamma \subset \mathbb{C}$ a positively (i.e., counterclockwise) oriented simple closed curve and a complex function $f(z)$ analytic in a region containing Γ and its inside, then

(i) *if $z_0 \in \mathbb{C}$ is enclosed by Γ, $f(z_0) = -\frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z_0 - z} dz$;*

(ii) *if not, $\frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z_0 - z} dz = 0$.*

$$(\mathbf{X} - \lambda \mathbf{I}_p)\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{Q}_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I}_n)^{-1} \tag{42}$$

- let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ be the spectral decomposition of $\mathbf{X}$, with $\mathbf{\Lambda} = \{\lambda_i(\mathbf{X})\}_{i=1}^p$ eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p \times p}$ the associated eigenvectors, then

$$\mathbf{Q}(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^\mathsf{T} = \sum_{i=1}^{p} \frac{\mathbf{u}_i \mathbf{u}_i^\mathsf{T}}{\lambda_i(\mathbf{X}) - z}. \tag{43}$$

- thus, same eigenspace as $\mathbf{X}$, but maps the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ to $1/(\lambda_i(\mathbf{X}) - z)$.

Applying Cauchy's integral formula to the resolvent matrix $\mathbf{Q}_{\mathbf{X}}(z)$ allows one to (somewhat magically!) assess the **eigenvalue** and **eigenvector** behavior of $\mathbf{X}$:

- characterize the eigenvalues of $\mathbf{X}$, one needs to determine a $z \in \mathbb{R}$ such that $\mathbf{Q}_{\mathbf{X}}(z)$ does *not* exist.
- can be done by directly calling the Cauchy's integral formula, which allows to determine the value of a (sufficiently nice) function f at a point of interest $z_0 \in \mathbb{R}$, by integrating its "inverse" $g_f(z) = f(z)/(z_0 - z)$ on the complex plane.
- this "inverse" $g_f(z)$ is akin to the resolvent and does not, *by design*, exist at the point of interest z_0.
- in the following example, we compare the two approaches of

(i) **directly solving** the determinantal equation; and
(ii) use resolvent + Cauchy's integral formula.

A resolvent approach to spectral analysis: an example

Consider the following two-by-two real symmetric random matrix

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \tag{44}$$

for (say independent) random variables x_1, x_2, x_3. For $\lambda_1(\mathbf{X})$ and $\lambda_2(\mathbf{X})$ the two (random) eigenvalues of $\mathbf{X}$ with associated (random) eigenvectors $\mathbf{u}_1(\mathbf{X}), \mathbf{u}_2(\mathbf{X}) \in \mathbb{R}^2$, we are interested in

$$f_{\mathbf{X}} = \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right], \qquad g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b},\ i \in \{1,2\}, \tag{45}$$

for some function $f\colon \mathbb{R} \to \mathbb{R}$ and deterministic $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$.

(i) **Directly solve** for the eigenvalues from the determinantal equation as

$$0 = \det(\mathbf{X} - \lambda \mathbf{I}_2) \Leftrightarrow \lambda(\mathbf{X}) = \frac{1}{2}\left(x_1 + x_3 \pm \sqrt{(x_1+x_3)^2 - 4(x_1x_3 - x_2^2)}\right), \tag{46}$$

and the associated eigenvectors from $\mathbf{X}\mathbf{u}_i(\mathbf{X}) = \lambda_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})$, $i \in \{1,2\}$. Then compute $f_{\mathbf{X}} = \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right]$, $g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b}$

- needs to **re-compute** of the expectation for a different choice of function f and the eigen-pair $(\lambda_1(\mathbf{X}), \mathbf{u}_1(\mathbf{X}))$ or $(\lambda_2(\mathbf{X}), \mathbf{u}_2(\mathbf{X}))$ of interest.

(ii) The **resolvent** approach:

$$\begin{aligned} f_{\mathbf{X}} &= \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right] \\ &= \mathbb{E}\left[-\frac{1}{2\pi\imath}\oint_\Gamma \left(\frac{f(z)}{\lambda_1(\mathbf{X})-z} + \frac{f(z)}{\lambda_2(\mathbf{X})-z}\right) dz\right] \\ &= -\frac{1}{2\pi\imath}\oint_\Gamma \mathbb{E}\left[f(z)\,\mathrm{tr}\,\mathbf{Q}_{\mathbf{X}}(z)dz\right] = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)\,\mathrm{tr}\left(\mathbb{E}[\mathbf{Q}_{\mathbf{X}}(z)]\right) dz, \end{aligned}$$

for Γ a positively-oriented contour that circles around both (random) eigenvalues of $\mathbf{X}$.

- a much more **unified approach** to the quantity $f_{\mathbf{X}}$ for different choices of f
- compute the expected resolvent once (which is much simpler in the case of large random matrices)
- then perform **contour integration** with the function f of interest.
- similarly, for $g_{i,\mathbf{X}}$, it follows that

$$g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b} = -\frac{1}{2\pi\imath}\oint_{\Gamma_i} \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{Q}_{\mathbf{X}}(z)]\mathbf{b}\,dz \tag{47}$$

for some contour Γ_i that circles around only $\lambda_i(\mathbf{X}), i \in \{1,2\}$

- given the expected resolvent $\mathbb{E}[\mathbf{Q}(z)]$, it suffices to choose the specific contour Γ_i to get the different expressions of $g_{1,\mathbf{X}}$ and $g_{2,\mathbf{X}}$

Resolvent as the core object

Objects of interest	Functionals of resolvent $\mathbf{Q}_{\mathbf{X}}(z)$
ESD $\mu_{\mathbf{X}}$ of $\mathbf{X}$	Stieltjes transform $m_{\mu_{\mathbf{X}}}(z) = \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)$
Linear spectral statistics (LSS): $f(\mathbf{X}) \equiv \frac{1}{p} \sum_i f(\lambda_i(\mathbf{X}))$	Integration of trace of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi i} \oint_\Gamma f(z) \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)\, dz$ (via Cauchy's integral)
Projections of eigenvectors $\mathbf{v}^\mathsf{T}\mathbf{u}(\mathbf{X})$ and $\mathbf{v}^\mathsf{T}\mathbf{U}(\mathbf{X})$ onto some given vector $\mathbf{v} \in \mathbb{R}^p$	Bilinear form $\mathbf{v}^\mathsf{T}\mathbf{Q}_{\mathbf{X}}(z)\mathbf{v}$ of $\mathbf{Q}_{\mathbf{X}}$
General matrix functional $F(\mathbf{X}) = \sum_i f(\lambda_i(\mathbf{X}))\mathbf{v}_1^\mathsf{T}\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}\mathbf{v}_2$ involving both eigenvalues and eigenvectors	Integration of bilinear form of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi i} \oint_\Gamma f(z)\mathbf{v}_1^\mathsf{T}\mathbf{Q}_{\mathbf{X}}(z)\mathbf{v}_2\, dz$

Using the resolvent to access eigenvalue distribution

Definition (Resolvent)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as

$$\mathbf{Q}_{\mathbf{X}}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}. \tag{48}$$

- let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$ be the spectral decomposition of $\mathbf{X}$, with $\mathbf{\Lambda} = \{\lambda_i(\mathbf{X})\}_{i=1}^{p}$ eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p\times p}$ the associated eigenvectors, then

$$\mathbf{Q}(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^{\mathsf{T}} = \sum_{i=1}^{p} \frac{\mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}}{\lambda_i(\mathbf{X}) - z}. \tag{49}$$

- thus, same eigenspace as $\mathbf{X}$, but maps the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ to $1/(\lambda_i(\mathbf{X}) - z)$.
- eigenvalue of $\mathbf{Q}_{\mathbf{X}}(z)$, and the resolvent matrix itself, must explode as z approaches any eigenvalue of $\mathbf{X}$.
- take the trace $\operatorname{tr}\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{Q}_{\mathbf{X}}(z)$ as the quantity to "locate" the eigenvalues of the matrix $\mathbf{X}$ of interest
- for $\mu_{\mathbf{X}} \equiv \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\mathbf{X})}$ the ESD of $\mathbf{X}$,

$$\boxed{\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i(\mathbf{X}) - z} = \int \frac{\mu_{\mathbf{X}}(dt)}{t - z} \equiv m_{\mu_{\mathbf{X}}}(z)}. \tag{50}$$

Definition (Stieltjes transform)

For a real probability measure μ with support $\text{supp}(\mu)$, the *Stieltjes transform* $m_\mu(z)$ is defined, for all $z \in \mathbb{C} \setminus \text{supp}(\mu)$, as

$$m_\mu(z) \equiv \int \frac{\mu(dt)}{t - z}. \tag{51}$$

Proposition (Properties of Stieltjes transform, [HLN07])

For m_μ the Stieltjes transform of a probability measure μ, it holds that

(i) m_μ is complex analytic on its domain of definition $\mathbb{C} \setminus \text{supp}(\mu)$;

(ii) it is bounded $|m_\mu(z)| \leq 1/\operatorname{dist}(z, \text{supp}(\mu))$;

(iii) it is an increasing function on all connected components of its restriction to $\mathbb{R} \setminus \text{supp}(\mu)$ (since $m'_\mu(x) = \int (t-x)^{-2} \mu(dt) > 0$) with $\lim_{x \to \pm\infty} m_\mu(x) = 0$ if $\text{supp}(\mu)$ is bounded; and

(iv) $m_\mu(z) > 0$ for $z < \inf \text{supp}(\mu)$, $m_\mu(z) < 0$ for $z > \sup \text{supp}(\mu)$ and $\Im[z] \cdot \Im[m_\mu(z)] > 0$ if $z \in \mathbb{C} \setminus \mathbb{R}$; and

BTW, for any $\mathbf{u} \in \mathbb{R}^p$ and matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ so that $\operatorname{tr}(\mathbf{A}) = 1$, $\mathbf{u}^\mathsf{T} \mathbf{Q}_\mathbf{X}(z) \mathbf{u}$, $\operatorname{tr}(\mathbf{A}\mathbf{Q}_\mathbf{X}(z))$ are STs.

[5] Walid Hachem, Philippe Loubaton, and Jamal Najim. "Deterministic equivalents for certain functionals of large random matrices". In: *The Annals of Applied Probability* 17.3 (2007), pp. 875–930

Definition (Inverse Stieltjes transform)

For a, b continuity points of the probability measure μ, we have

$$\mu([a,b]) = \frac{1}{\pi} \lim_{y \downarrow 0} \int_a^b \Im \left[m_\mu(x + \imath y) \right] \, dx. \tag{52}$$

Besides, if μ admits a density f at x (i.e., $\mu(x)$ is differentiable in a neighborhood of x and $\lim_{\epsilon \to 0} (2\epsilon)^{-1} \mu([x-\epsilon, x+\epsilon]) = f(x)$),

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im \left[m_\mu(x + \imath y) \right]. \tag{53}$$

Definition (Linear Spectral Statistic, LSS)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the *linear spectral statistics* (LSS) $f_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the averaged statistics of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ of $\mathbf{X}$ via some function $f\colon \mathbb{R} \to \mathbb{R}$, that is

$$f(\mathbf{X}) = \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(\mathbf{X})). \tag{54}$$

In particular, we have $= \int f(t)\mu_{\mathbf{X}}(dt)$, for $\mu_{\mathbf{X}}$ the ESD of $\mathbf{X}$.

LSS via contour integration: For $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ eigenvalues of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, some function $f\colon \mathbb{R} \to \mathbb{R}$ that is complex analytic in a compact neighborhood of the support $\mathrm{supp}(\mu_{\mathbf{X}})$ (of the ESD $\mu_{\mathbf{X}}$ of $\mathbf{X}$), then

$$f(\mathbf{X}) = \int f(t)\mu_{\mathbf{X}}(dt) = -\int \frac{1}{2\pi\imath}\oint_{\Gamma} \frac{f(z)\,dz}{t-z}\mu_{\mathbf{X}}(dt) = -\frac{1}{2\pi\imath}\oint_{\Gamma} f(z) m_{\mu_{\mathbf{X}}}(z)\,dz, \tag{55}$$

for *any* contour Γ that encloses $\mathrm{supp}(\mu_{\mathbf{X}})$, i.e., all the eigenvalues $\lambda_i(\mathbf{X})$.

Remark (LSS to retrieve the inverse Stieltjes transform formula):

$$\frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\delta_{\lambda_i(\mathbf{X})} = -\frac{1}{2\pi\imath}\oint_\Gamma 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$
$$= -\frac{1}{2\pi\imath}\int_{a-\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$
$$-\frac{1}{2\pi\imath}\int_{a-\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz.$$

- Since $\Re[m(x+\imath y)] = \Re[m(x-\imath y)], \Im[m(x+\imath y)] = -\Im[m(x-\imath y)]$;
- we have $\int_{a-\varepsilon_x}^{b+\varepsilon_x} m_{\mu_\mathbf{X}}(x-\imath\varepsilon_y)\,dx + \int_{b+\varepsilon_x}^{a-\varepsilon_x} m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)dx = -2\imath\int_{a-\varepsilon_x}^{b+\varepsilon_x}\Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$;
- and consequently $\mu([a,b]) = \frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\lambda_i(\mathbf{X}) = \frac{1}{\pi}\lim_{\varepsilon_y\downarrow 0}\int_a^b \Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$.

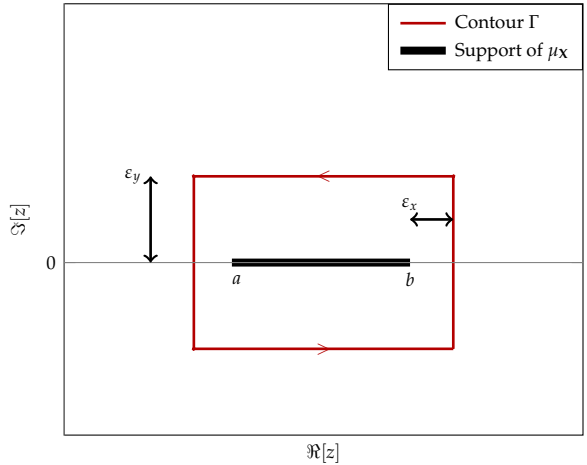


Figure: Illustration of a rectangular contour Γ and support of $\mu_{\mathbf{X}}$ on the complex plane.

Definition (Matrix spectral functionals)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, we say $F\colon \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ is a matrix spectral functional of $\mathbf{X}$,

$$F(\mathbf{X}) = \sum_{i\in\mathcal{I}\subseteq\{1,\ldots,p\}} f(\lambda_i(\mathbf{X}))\mathbf{u}_i\mathbf{u}_i^\mathsf{T}, \quad \mathbf{X} = \sum_{i=1}^{p} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^\mathsf{T}. \tag{56}$$

Spectral functional via contour integration: For $\mathbf{X} \in \mathbb{R}^{p\times p}$, resolvent $\mathbf{Q}_\mathbf{X}(z) = (\mathbf{X} - z\mathbf{I}_p)^{-1}$, $z \in \mathbb{C}$, and $f\colon \mathbb{R} \to \mathbb{R}$ analytic in a neighborhood of the contour $\Gamma_\mathcal{I}$ that circles around the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ with their indices in the set $\mathcal{I} \subseteq \{1,\ldots,p\}$,

$$F(\mathbf{X}) = -\frac{1}{2\pi\imath} \oint_{\Gamma_\mathcal{I}} f(z)\mathbf{Q}_\mathbf{X}(z)\, dz. \tag{57}$$

Example: access to the i-th eigenvector $\mathbf{u}_i$ of $\mathbf{X}$ through

$$\mathbf{u}_i\mathbf{u}_i^\mathsf{T} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{Q}_\mathbf{X}(z)\, dz, \tag{58}$$

for $\Gamma_{\lambda_i(\mathbf{X})}$ a contour circling around $\lambda_i(\mathbf{X})$ only, so eigenvector projection $(\mathbf{v}^\mathsf{T}\mathbf{u}_i)^2 = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{v}^\mathsf{T}\mathbf{Q}_\mathbf{X}(z)\mathbf{v}\, dz$.

Example: training linear model with gradient descent

Note that

$$\begin{aligned}\boldsymbol{\beta}_*^{\mathsf{T}}\boldsymbol{\beta}(t) &= \boldsymbol{\beta}_*^{\mathsf{T}} e^{-t\hat{\mathbf{C}}}\boldsymbol{\beta}(0) + \boldsymbol{\beta}_*^{\mathsf{T}}\left(\mathbf{I}_p - e^{-t\hat{\mathbf{C}}}\right)\boldsymbol{\beta}_{RR} \\ &= \boldsymbol{\beta}_*^{\mathsf{T}} e^{-t\hat{\mathbf{C}}}\boldsymbol{\beta}(0) + \boldsymbol{\beta}_*^{\mathsf{T}}\left(\mathbf{I}_p - e^{-t\hat{\mathbf{C}}}\right)\hat{\mathbf{C}}^{-1}\frac{1}{n}\mathbf{X}y \\ &= -\frac{1}{2\pi\imath}\oint_\Gamma \left(\exp(-tz)\cdot\boldsymbol{\beta}_*^{\mathsf{T}}\mathbf{Q}(z)\boldsymbol{\beta}(0) + \frac{1-\exp(-zt)}{z}\cdot\frac{1}{n}\boldsymbol{\beta}_*^{\mathsf{T}}\mathbf{Q}(z)\mathbf{X}y\right)dz,\end{aligned}$$

for Γ a positively oriented contour that circles around all eigenvalues of $\hat{\mathbf{C}}$, and resolvent

$$\mathbf{Q}(z) = (\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1} = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} - z\mathbf{I}_p\right)^{-1}. \tag{59}$$

Take-away messages of this section

- "basic" probability: **concentration** of scalar observations of large random vectors: simple and involved, linear and nonlinear objects
- boils down to **expectation computation/evaluation**
- **same** holds for scalar observations of large random matrices
- linear algebra: matrix norm "equivalence" but up to **dimensional factors**
- **resolvent** (i.e., regularized inverse) naturally appears in eigenvalue/eigenvector assessment
- a unified **resolvent**-based to eigenspectral analysis of (not necessarily random) matrices: Cauchy's integral formula, Stieltjes transform (and its inverse), Linear Spectral Statistic, and generic matrix spectral functionals, etc.

Two different scaling regimes

Example (Nonlinear objects in two scaling regimes)

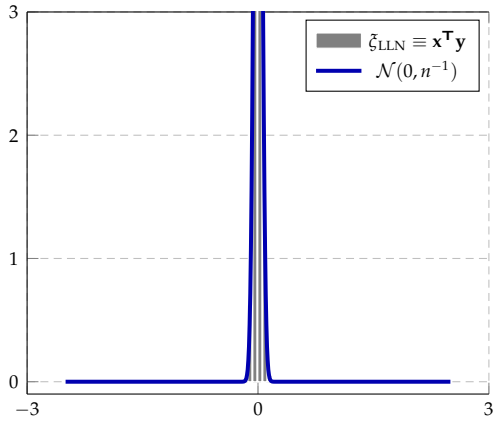
Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector so that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries with zero mean and unit variance, and $\mathbf{y} \in \mathbb{R}^n$ be a deterministic vector of unit norm $\|\mathbf{y}\| = 1$; and consider the following two families of nonlinear objects of interest with a nonlinear function f acting on different regimes:

(i) **LLN regime**: here we are interested in $f(\|\mathbf{x}\|^2)$ and $f(\mathbf{x}^\mathsf{T}\mathbf{y})$; and

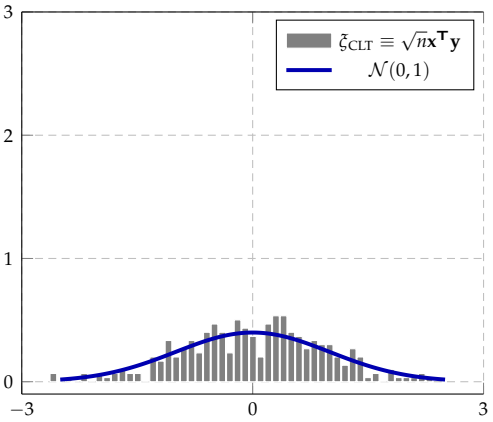
(ii) **CLT regime**: here we are interested in $f(\sqrt{n}(\|\mathbf{x}\|^2 - 1))$ and $f(\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y})$.

- the (strong) **law of large numbers (LLN)** implies that
$$\|\mathbf{x}\|^2 \to \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] = 1 \text{ and } \mathbf{x}^\mathsf{T}\mathbf{y} \to \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{y}] = 0$$
almost surely as $n \to \infty$; and
- the **central limit theorem (CLT)** implies that
$$\sqrt{n}(\|\mathbf{x}\|^2 - 1) \to \mathcal{N}(0, 2) \text{ and } \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \to \mathcal{N}(0, 1)$$
in law as $n \to \infty$
- leads to the more compact form, for n large,
$$\|\mathbf{x}\|^2 \simeq 1 + \mathcal{N}(0, 2)/\sqrt{n} \text{ and } \mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + \mathcal{N}(0, 1)/\sqrt{n}. \tag{60}$$

Illustration of the two scaling regime



(a) LLN regime



(b) CLT regime

Figure: Illustrations of random variables in LLN (**left**) and CLT (**right**) regime, with $n = 500$.

Two different scaling regimes and their corresponding linearization

Table: Comparison between two different high-dimensional linearization approaches.

Scaling regime	LLN type	CLT type
Object of interest	$f(\xi)$ for (almost) deterministic $\xi = \tau + o(1)$	$f(\xi)$ for random ξ, e.g., $\xi \sim \mathcal{N}(0, 1)$
Linearization technique	Taylor expansion	Orthogonal polynomial
Smoothness of f	Locally smooth f	Possibly non-smooth f

Theorem (Taylor's theorem for deterministic single-variable functions)

Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function that is at least k times continuously differentiable in a neighborhood of a given point $\tau \in \mathbb{R}$. Then, there exists a function $h_k\colon \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(\tau) + f'(x-\tau) + \frac{f''(\tau)}{2}(x-\tau)^2 + \ldots + \frac{f^{(k)}(\tau)}{k!}(x-\tau)^k + h_k(x)(x-\tau)^k, \tag{61}$$

with $\lim_{x\to\tau} h_k(x) = 0$ so that $h_k(x)(x-\tau)^k = o(|x-\tau|^k)$ as $x \to \tau$.

What makes the Taylor expansion approach work for random nonlinear functions $f(x)$?

- **Smoothness.** nonlinear f should be smooth, at least in the neighborhood of the point τ of interest, so that the derivatives $f'(\tau), f''(\tau), \ldots$ make sense.
- **Concentration.** variable of interest x is sufficiently close to (or, concentrates around, when being random) the point τ so that the higher orders terms are neglectable

Proposition (Taylor expansion of high-dimensional random functions in the LLN regime)

For random variable $\xi = \|\mathbf{x}\|^2$ with $\sqrt{n}\mathbf{x} \in \mathbb{R}^n$ having i.i.d. standard Gaussian entries, in the LLN regime, it follows from LLN and CLT that $\|\mathbf{x}\|^2 - 1 = O(n^{-1/2})$ with high probability for n large, so that one can apply Taylor theorem to write

$$f(\|\mathbf{x}\|^2) = f(1) + f'(1) \underbrace{(\|\mathbf{x}\|^2 - 1)}_{O(n^{-1/2})} + \frac{1}{2} f''(1) \underbrace{(\|\mathbf{x}\|^2 - 1)^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{62}$$

with high probability. Similarly,

$$f(\mathbf{x}^\mathsf{T}\mathbf{y}) = f(0) + f'(0) \underbrace{\mathbf{x}^\mathsf{T}\mathbf{y}}_{O(n^{-1/2})} + \frac{1}{2} f''(0) \underbrace{(\mathbf{x}^\mathsf{T}\mathbf{y})^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{63}$$

again as a consequence of $\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \xrightarrow{d} \mathcal{N}(0,1)$ in distribution as $n \to \infty$, where the orders $O(n^{-\ell})$ hold with high probability for n large.

A functional analysis perspective of expectation of nonlinear random function

- Consider the following functional analysis perspective of the expectation $\mathbb{E}[f(\xi)]$
- For a random variable ξ following some law μ, the expectation $\mathbb{E}[f(\xi)]$ of the nonlinear transformation $f(\xi)$ can be expresses as

$$\mathbb{E}_{\xi\sim\mu}[f(\xi)] = \int f(t)\mu(dt). \tag{64}$$

- In the case of Euclidean space, the canonical vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ form an orthonormal basis of $\mathbb{R}^n$; and thus any vector $\mathbf{x}$ living in the Euclidean space $\mathbb{R}^n$ can be decomposed as

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x}^\mathsf{T}\mathbf{e}_i)\mathbf{e}_i = \sum_{i=1}^{n} x_i\mathbf{e}_i, \tag{65}$$

  with the inner product $\mathbf{x}^\mathsf{T}\mathbf{e}_i = x_i$ the ith coordinate of $\mathbf{x}$.
- A similar result holds more generally, e.g., or a function f living in some (infinite dimensional) function space, can be decomposed into the sum of "orthonormal" basis functions, weighted by the projection (i.e., inner product) of f onto these basis functions

Definition (Orthogonal polynomials and orthogonal polynomial expansion)

For a probability measure μ, define the inner product

$$\langle f, g \rangle \equiv \int f(\xi) g(\xi) \mu(d\xi) = \mathbb{E}[f(\xi) g(\xi)], \tag{66}$$

for $\xi \sim \mu$. We say that $\{P_\ell(\xi), \ell \geq 0\}$ is a family of orthogonal polynomials with respect to this inner product, obtained by the Gram-Schmidt procedure on the monomials $\{1, \xi, \xi^2, \ldots\}$, with $P_0(\xi) = 1$, where P_ℓ is a polynomial function of degree ℓ that satisfies

$$\langle P_{\ell_1}, P_{\ell_2} \rangle = \mathbb{E}[P_{\ell_1}(\xi) P_{\ell_2}(\xi)] = \delta_{\ell_1 = \ell_2}. \tag{67}$$

Then, for any function $f \in L^2(\mu)$, the orthogonal polynomial expansion of f is

$$f(\xi) \sim \sum_{\ell=0}^{\infty} a_\ell P_\ell(\xi), \quad a_\ell = \int f(\xi) P_\ell(\xi) \mu(d\xi) \tag{68}$$

- denote "$f \sim \sum_{l=0}^{\infty} a_\ell P_\ell$" to denote that $\|f - \sum_{\ell=0}^{L} a_\ell P_\ell\|_\mu \to 0$ as $L \to \infty$ with $\|f\|_\mu^2 = \langle f, f \rangle$, or equivalently $\int \left(f(\xi) - \sum_{\ell=0}^{L} a_\ell P_\ell(\xi)\right)^2 \mu(d\xi) = \mathbb{E}\left[\left(f(\xi) - \sum_{\ell=0}^{L} a_\ell P_\ell(\xi)\right)^2\right] \to 0$.

Theorem (Hermite polynomial decomposition)

For $\xi \in \mathbb{R}$, the ℓ^{th} order normalized Hermite polynomial, denoted $P_\ell(\xi)$, is given by given by

$$P_0(\xi) = 1, \text{ and } P_\ell(\xi) = \frac{(-1)^\ell}{\sqrt{\ell!}} e^{\frac{\xi^2}{2}} \frac{d^n}{d\xi^n}\left(e^{-\frac{\xi^2}{2}}\right), \text{ for } \ell \geq 1. \tag{69}$$

and the family of (normalized) Hermite polynomials

(i) *being orthogonal polynomials and (as the name implies) are orthonormal with respect the standard Gaussian measure: $\int P_m(\xi)P_n(\xi)\mu(d\xi) = \delta_{nm}$, for $\mu(dt) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ the standard Gaussian measure; and*

(ii) *form an orthonormal basis of $L^2(\mu)$, the Hilbert space consist of all square-integrable functions with respect to the inner product $\langle f, g\rangle \equiv \int f(\xi)g(\xi)\mu(d\xi)$, and that one can formally expand any $f \in L^2(\mu)$ as*

$$f(\xi) \sim \sum_{\ell=0}^{\infty} a_{\ell,f} P_\ell(\xi), \quad a_{\ell,f} = \int f(\xi)P_\ell(\xi)\mu(d\xi) = \mathbb{E}[f(\xi)P_\ell(\xi)], \tag{70}$$

where we use '$f \sim \sum_{\ell=0}^{\infty} a_{\ell,f} P_\ell$' for standard Gaussian $\xi \sim \mathcal{N}(0,1)$. The coefficients $a_{\ell,f}$s are generalized moments of the standard Gaussian measure μ involving f, and we have

$$a_{0,f} = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[f(\xi)], \quad a_{1,f} = \mathbb{E}[\xi f(\xi)], \quad \sqrt{2}a_{2,f} = \mathbb{E}[\xi^2 f(\xi)] - a_{0,f}, \quad \nu_f = \mathbb{E}[f^2(\xi)] = \sum_{\ell=0} a_{\ell,f}^2. \tag{71}$$

Numerical illustration of Hermite polynomials

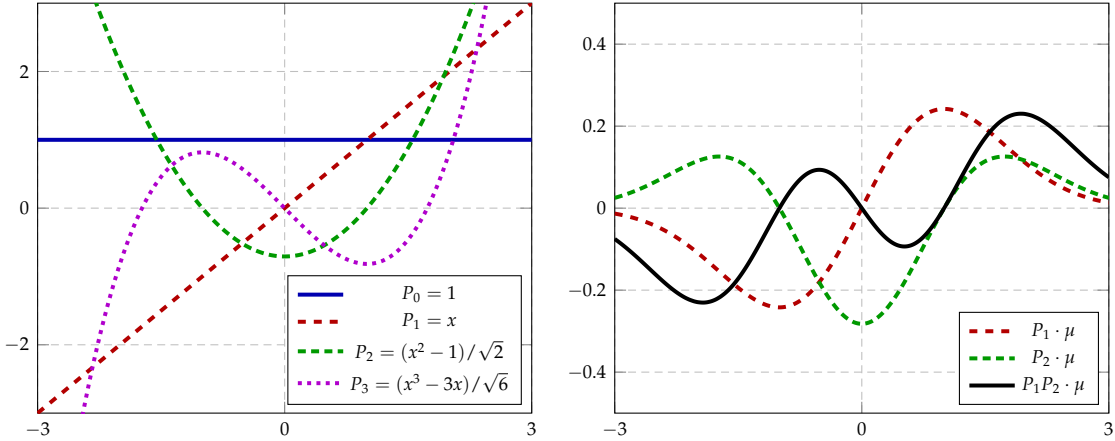


Figure: Illustration of the first four Hermite polynomials (**left**) and of the first- and second-order Hermite polynomial (P_1 and P_2) weighted by the Gaussian mixture $\mu(dx) = \exp(-x^2/2)/\sqrt{2\pi}$ (**right**).

Hermite polynomial "expansion" in the CLT regime

Proposition (Hermite polynomial "expansion" in the CLT regime)

For random variable $\xi_{\mathrm{CLT}} = \sqrt{n} \cdot (\|\mathbf{x}\|^2 - 1)$ *with* $\sqrt{n}\mathbf{x} \in \mathbb{R}^n$ *having i.i.d. standard Gaussian entries, in the CLT regime, it follows from the CLT that* $\xi_{\mathrm{CLT}} \sim \mathcal{N}(0,1)$ *in the* $n \to \infty$ *limit, so that one can write*

$$\mathbb{E}[f(\sqrt{n} \cdot (\|\mathbf{x}\|^2 - 1))] = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[f(\xi)] + o(1) = a_{0,f} + o(1), \tag{72}$$

as $n \to \infty$*; and similarly*

$$\mathbb{E}[f(\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y})] = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[f(\xi)] + o(1) = a_{0,f} + o(1). \tag{73}$$

- looks **not** extremely insightful
- makes a lot more sense for scalar nonlinear observations of random vectors and random matrices, e.g., $\mathbf{K} = f(\mathbf{X}^\mathsf{T}\mathbf{X}/\sqrt{p})/\sqrt{p} - \mathrm{diag}(\cdot)$, for random matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$

Example (Nonlinear behaviors of tanh in two scaling regimes)

Consider the hyperbolic tangent function $f(t) = \tanh(t)$. This nonlinear function is "close" to different quadratic functions in *different* regimes of interest. More precisely, we have the following.

(i) **In the LLN regime**, we have

$$\tanh(\xi_{\text{LLN}}) \simeq g(\xi_{\text{LLN}}),$$

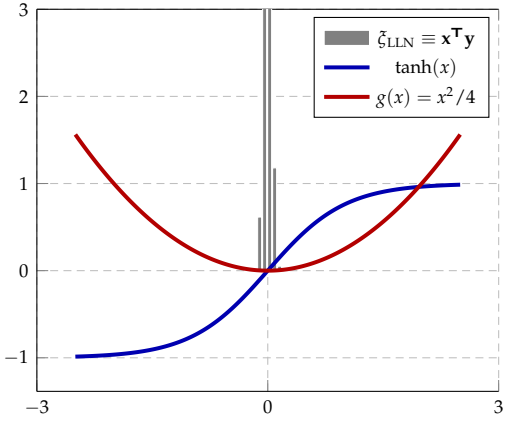
with $g(t) = t^2/4$. This is as a consequence of $\tanh(x) = g(x) = 0$. In particular, $\mathbb{E}[\tanh(\xi_{\text{LLN}})] \simeq \mathbb{E}[g(\xi_{\text{LLN}})]$.

(ii) **In the CLT regime**, we have

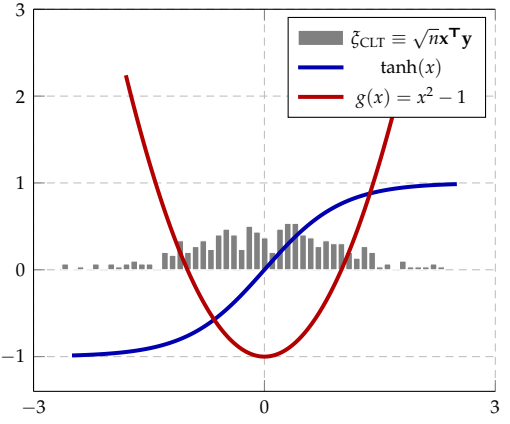
$$\mathbb{E}[\tanh(\xi_{\text{LLN}})] = \mathbb{E}[g(\xi_{\text{LLN}})]$$

in expectation, with now $g(t) = t^2 - 1$, i.e., with a different function. This is a consequence of the fact that their zeroth-order Hermite coefficient $a_0 = 0$.

Numerical illustration of two high-dimensional linearization technique



(a) LLN regime



(b) CLT regime

Figure: Different behavior of nonlinear $f(\xi_{\text{LLN}})$ and $f(\xi_{\text{CLT}})$ for $f(t) = \tanh(t)$ in the LLN and CLT regime, with $n = 500$. We have in particular $\tanh(\xi_{\text{LLN}}) \simeq g(\xi_{\text{LLN}})$ in the LLN regime and $\mathbb{E}[\tanh(\xi_{\text{CLT}})] = \mathbb{E}[g(\xi_{\text{CLT}})]$ in the CLT regime with *different* g.

Definition (High-dimensional Linear Equivalent)

For a random vector $\mathbf{x} \in \mathbb{R}^n$, its nonlinear transformation $f(\mathbf{x}) \in \mathbb{R}^n$ is obtained by applying $f\colon \mathbb{R} \to \mathbb{R}$ entry-wise on $\mathbf{x}$. Consider $g(f(\mathbf{x}))$ a scalar observation of $f(\mathbf{x})$ via observation function $g\colon \mathbb{R}^n \to \mathbb{R}$, we say that the random vector $\tilde{\mathbf{x}}_f$ (defined on an extended probability space if necessary) is an (ε, δ)-Linear Equivalent to $f(\mathbf{x})$ if, with probability at least $1 - \delta(n)$ that

$$\left| g(f(\mathbf{x})) - g(\tilde{\mathbf{x}}_f) \right| \leq \varepsilon(n), \tag{74}$$

for some non-negative functions $\varepsilon(n)$ and $\delta(n)$ that decrease to zero as $n \to \infty$. This, in the limit of $n \to \infty$, leads to

$$g(f(\mathbf{x})) - g(\tilde{\mathbf{x}}_f) \to 0, \tag{75}$$

in probability or almost surely for the observation function $g(\cdot)$, and we denote

$$f(\mathbf{x}) \overset{g}{\leftrightarrow} \tilde{\mathbf{x}}_f. \tag{76}$$

And similarly for a random matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$.

Example: Nonlinear random vectors in two scaling regimes

Example (Nonlinear random vectors in two scaling regimes)

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a *random* matrix so that $\sqrt{n}\mathbf{X}$ has i.i.d. standard Gaussian entries with zero mean and unit variance, and $\mathbf{y} \in \mathbb{R}^n, \boldsymbol{\alpha} \in \mathbb{R}^p$ be *deterministic* vectors of unit norm such that $\|\mathbf{y}\| = 1$ and $\|\boldsymbol{\alpha}\| = 1$; consider the following two families of *scalar* observations of *nonlinear* random vectors with observation function $g\colon \mathbb{R}^p \to \mathbb{R}$ and a nonlinear function f acting on different regimes:

(i) **LLN regime**: $g(f(\mathbf{X}\mathbf{y})) = \frac{1}{\sqrt{n}}\boldsymbol{\alpha}^\mathsf{T} f(\mathbf{X}\mathbf{y})$; and

(ii) **CLT regime**: $g(f(\sqrt{n} \cdot \mathbf{X}\mathbf{y})) = \frac{1}{\sqrt{n}}\boldsymbol{\alpha}^\mathsf{T} f(\sqrt{n} \cdot \mathbf{X}\mathbf{y})$.

Proposition (Taylor expansion of nonlinear random vector in the LLN regime)

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix so that $\sqrt{n}\mathbf{X}$ has i.i.d. standard Gaussian entries with zero mean and unit variance, and $\mathbf{y} \in \mathbb{R}^n, \boldsymbol{\alpha} \in \mathbb{R}^p$ be deterministic vectors of unit norm such that $\|\mathbf{y}\| = 1$ and $\|\boldsymbol{\alpha}\| = 1$, in the LLN regime, the following Linear Equivalent holds

$$f(\mathbf{X}\mathbf{y}) \overset{g}{\leftrightarrow} \underbrace{f(0)\cdot \mathbf{1}_p}_{O_{\|\cdot\|_\infty}(1)} + \underbrace{f'(0)\cdot \mathbf{X}\mathbf{y}}_{O_{\|\cdot\|_\infty}(n^{-1/2})}, \tag{77}$$

for the scalar observation function $g(\cdot) = \boldsymbol{\alpha}^\mathsf{T}(\cdot)/\sqrt{n}$, up to some approximation error $\varepsilon = O(n^{-1})$.

Proposition (Hermite polynomial expansion in the CLT regime.)

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix so that $\sqrt{n}\mathbf{X}$ has i.i.d. standard Gaussian entries with zero mean and unit variance, and $\mathbf{y} \in \mathbb{R}^n, \boldsymbol{\alpha} \in \mathbb{R}^p$ be deterministic vectors of unit norm such that $\|\mathbf{y}\| = 1$ and $\|\boldsymbol{\alpha}\| = 1$, in the CLT regime, if the nonlinear $f : \mathbb{R} \to \mathbb{R}$ and $g(\cdot) = \boldsymbol{\alpha}^\mathsf{T}(\cdot)/\sqrt{n}$ are such that $g(f(\sqrt{n}\mathbf{X}\mathbf{y}))$ strongly concentrates, i.e.,

$$g(f(\sqrt{n}\mathbf{X}\mathbf{y})) = \frac{1}{\sqrt{n}}\boldsymbol{\alpha}^\mathsf{T} f(\sqrt{n}\mathbf{X}\mathbf{y}) = \frac{1}{\sqrt{n}}\mathbb{E}[\boldsymbol{\alpha}^\mathsf{T} f(\sqrt{n}\mathbf{X}\mathbf{y})] + \varepsilon(n,p), \tag{78}$$

with high probability for n, p large, so $f(\sqrt{n}\mathbf{X}\mathbf{y}) \overset{g}{\leftrightarrow} a_{0,f}\cdot \mathbf{1}_p$, for the observation function $g(\cdot) = \boldsymbol{\alpha}^\mathsf{T}(\cdot)/\sqrt{n}$.

An additional example in the CLT regime

Example (Hermite polynomial expansion in the CLT regime)

Under the same notations and settings as above but for random observation function

$$g(\cdot) = \frac{1}{\sqrt{n}} \mathbf{y}^\mathsf{T} \mathbf{X}^\mathsf{T} (\cdot), \tag{79}$$

that is assumed to strongly concentrate around its expectation up to some $\varepsilon(n,p)$ for n,p large, then, the following Linear Equivalent holds

$$f(\sqrt{n}\mathbf{X}\mathbf{y}) \overset{g}{\leftrightarrow} a_{1,f} \cdot \sqrt{n}\mathbf{X}\mathbf{y}, \tag{80}$$

up to some approximation error $\varepsilon(n,p)$.

- we also have $f(\sqrt{n}\mathbf{X}\mathbf{y}) \overset{g}{\leftrightarrow} a_{1,f} \cdot \sqrt{n}\mathbf{X}\mathbf{y} + \mathbf{z}$, and Linear Equivalents are not unique
- in some cases we care joint behavior of multiple observation functions, etc.

An additional example of joint behavior in the CLT regime

Example (Hermite polynomial expansion in the CLT regime: joint behavior)

Consider random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ having i.i.d. standard Gaussian entries, and nonlinear random vector $f(\mathbf{x})$ with nonlinear $f\colon \mathbb{R} \to \mathbb{R}$ applied entry-wise on $\mathbf{x}$, in the CLT regime. Then, for the *joint* behavior of the two *scalar* observation of $f(\mathbf{x})$,

$$(g_1(f(\mathbf{x})), g_2(f(\mathbf{x}))) = \left(\frac{1}{p} \mathbf{x}^\mathsf{T} f(\mathbf{x}), \frac{1}{p} f(\mathbf{x})^\mathsf{T} f(\mathbf{x}) \right), \tag{81}$$

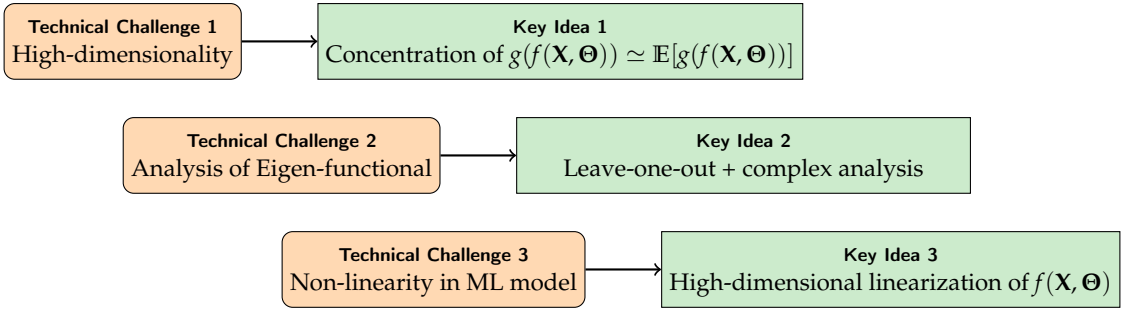
the following asymptotic equivalent linear model holds

$$f(\mathbf{x}) \overset{(g_1, g_2)}{\leftrightarrow} a_{0,f} \cdot \mathbf{1}_p + a_{1,f} \cdot \mathbf{x} + \sqrt{\nu_f - a_{0,f}^2 - a_{1,f}^2} \cdot \mathbf{z}, \tag{82}$$

with $a_{0,f}, a_{1,f}, \nu_f$ the Hermite coefficients of f, and standard Gaussian random vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ that is *independent* of $\mathbf{x}$.

Take-away messages of this section

- two different scaling regimes: LLN versus CLT
- high-dimensional linearizations of nonlinear random functions via **Taylor Expansion** and **Orthogonal Polynomial**
- Taylor Expansion can be performed in a close-to-deterministic fashion
- Orthogonal Polynomial is more tricky and depends on the form of the observation map



Thank you! Q & A?

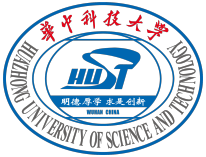
Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 2

CIMI Thematic School "Models & Methods for High-dimensional Inference and Learning"

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

October 18, 2024



Outline

Four ways to characterize sample covariance matrices

Definition (Sample Covariance Matrix, SCM)

The SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p\times p}$ of data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ composed of n independent data samples $\mathbf{x}_i \in \mathbb{R}^p$ of zero mean is given by

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}. \tag{1}$$

Definition (Classical versus proportional regimes)

For SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p\times p}$ from n samples of dimension p, consider the following two regimes.

1. **Classical regime** with $n \gg p$, this includes both asymptotic ($n \to \infty$ with p fixed) and non-asymptotic characterizations ($n \gg p$ for large but finite n).
2. **Proportional regime** with $n \sim p$, this includes both asymptotic ($n, p \to \infty$ with $p/n \to c \in (0, \infty)$, also known as thermodynamic limit in the statistical physics literature) and non-asymptotic characterizations ($n \sim p \gg 1$ both large but finite).

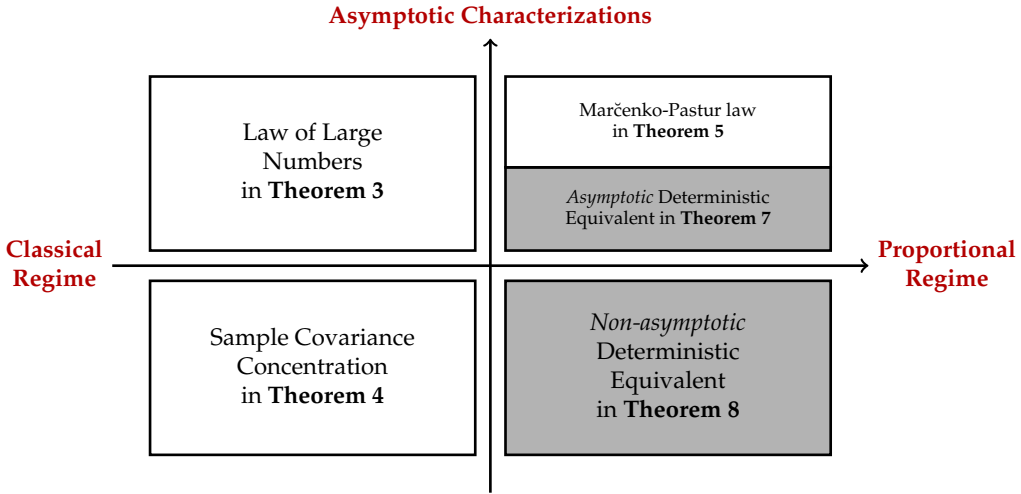


Figure: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$.

Asymptotic behavior of SCM in the classical regime via law of large numbers

Theorem (Asymptotic Law of Large Numbers for SCM)

Let p be fixed, and let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix with independent sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then one has,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0, \tag{2}$$

almost surely, as $n \to \infty$.

- LLN is "parameterized" to hold only in the **classical limit**, not the **proportional limit**
- many variants and extensions of the LLN exist, but become vacuous when applied to the **proportional regime** $n, p \to \infty$ and $p/n \to c \in (0, \infty)$, see below for an example

Non-asymptotic behavior of SCM in the classical regime via matrix concentration

Theorem (Non-asymptotic matrix concentration for SCM, [Ver18, Theorem 4.6.1])

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ *be a random matrix with independent sub-gaussian columns* $\mathbf{x}_i \in \mathbb{R}^p$ *such that* $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ *and* $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. *Then, one has, with probability at least* $1 - 2\exp(-t^2)$, *for any* $t \geq 0$, *that*

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n}), \tag{3}$$

for some constants $C_1, C_2 > 0$, *independent of* n, p.

Proof: combines Bernstein's concentration inequality with ϵ-net argument, see [Ver18] for details.

- can reproduce the LLN asymptotic result by taking $n \to \infty$ with Borel–Cantelli lemma

(i) **Classical regime.** Here, $n \gg p$, **say that** $n \sim p^2$. Then with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(n^{-1/4})$ and conveys a similar intuition to the asymptotic LLN result

(ii) **Proportional regime.** Here, n, p **are both large and** $n \sim p$. Then, with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(\sqrt{p/n}) = O(1)$, and qualitatively different LLN with a vacuous $\sim 100\%$ relative error, e.g., as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

Proportional regime: eigenvalues via traditional RMT and the Marčenko-Pastur law

Theorem (Limiting spectral distribution for SCM: Marčenko-Pastur law, [MP67])

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix with i.i.d. sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then, as $n, p \to \infty$ with $p/n \to c \in (0,\infty)$, with probability one, the empirical spectral measure (ESD) $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ converges weakly to a probability measure μ given explicitly by

$$\boxed{\mu(dx) = (1-c^{-1})^+\delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x-E_-)^+(E_+-x)^+}\,dx,} \tag{4}$$

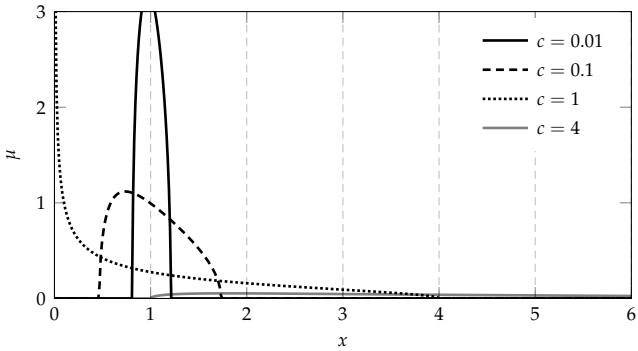
where $E_\pm = (1\pm\sqrt{c})^2$ and $(x)^+ = \max(0,x)$, which is known as the Marčenko-Pastur distribution.

- provides a more refined characterization of the eigenspectrum of $\hat{\mathbf{C}}$ (than, e.g., matrix concentration):

(i) **Classical regime.** Here, $n \gg p$ **so that** $c = p/n \to 0$, the Marčenko-Pastur law in Equation (4) shrinks to a Dirac mass, in agreement with $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \sim 0$

(ii) **Proportional regime.** Here, $n \sim p \gg 1$, and by the (true but vacuous) matrix concentration result $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(p/n) = O(1)$, and, depending on the ratio $c = p/n$, the eigenvalues of $\hat{\mathbf{C}}$ can be very different from one, and takes the form of the Marčenko-Pastur law

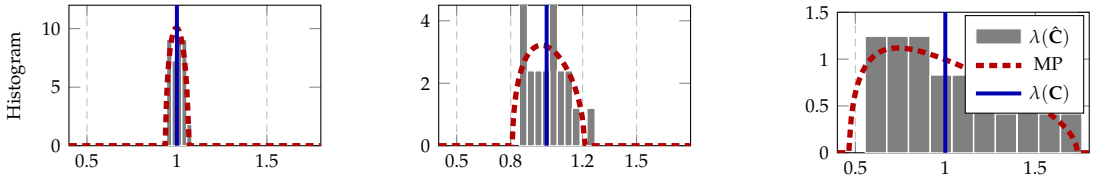
- we have in fact $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \simeq c + 2\sqrt{c}$ as $n, p \to \infty$ with $p/n \to c \in (0,\infty)$



- averaged amount of eigenvalues of $\hat{\mathbf{C}}$ lying within the interval $[1-\delta, 1+\delta]$, for $\delta \ll 1$, as

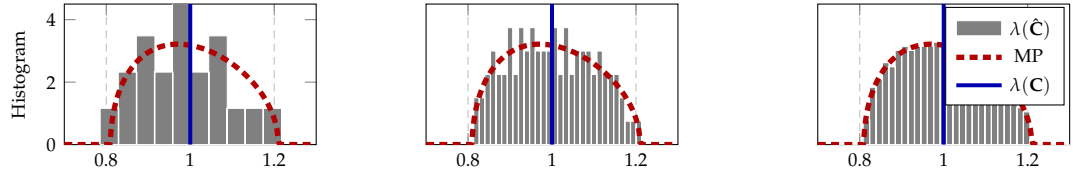
$$\mu([1-\delta,1+\delta]) = \int_{1-\delta}^{1+\delta} \frac{1}{2\pi c x} \sqrt{\left(x-(1-\sqrt{c})^2\right)^+ \left((1+\sqrt{c})^2 - x\right)^+}\, dx$$
$$= \frac{1}{2\pi c} \int_{-\delta}^{\delta} \left(\sqrt{4c-c^2} + O(\varepsilon)\right) d\varepsilon = \frac{\sqrt{4c^{-1}-1}}{\pi}\delta + O(\delta^2).$$

- for $p \approx 4n$ there is asymptotically no eigenvalue of $\hat{\mathbf{C}}$ close to one!
- in accordance with the shape of the limiting Marčenko-Pastur law with $c = 4$ above



(a) $n = 1\,000p$ (b) $n = 100p$ (c) $n = 10p$

Figure: **Varying n and $c = p/n$ for fixed p.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the limiting Marčenko-Pastur law in Theorem 5, for $\mathbf{X}$ having standard Gaussian entries with $p = 20$ and different $n = 1\,000p, 100p, 10p$ from left to right.



(a) $p = 20$ (b) $p = 100$ (c) $p = 500$

Figure: **Varying n and p for fixed $c = p/n$.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the Marčenko-Pastur law, for $\mathbf{X}$ having standard Gaussian entries with $n = 100p$ and different $p = 20, 100, 500$ from left to right.

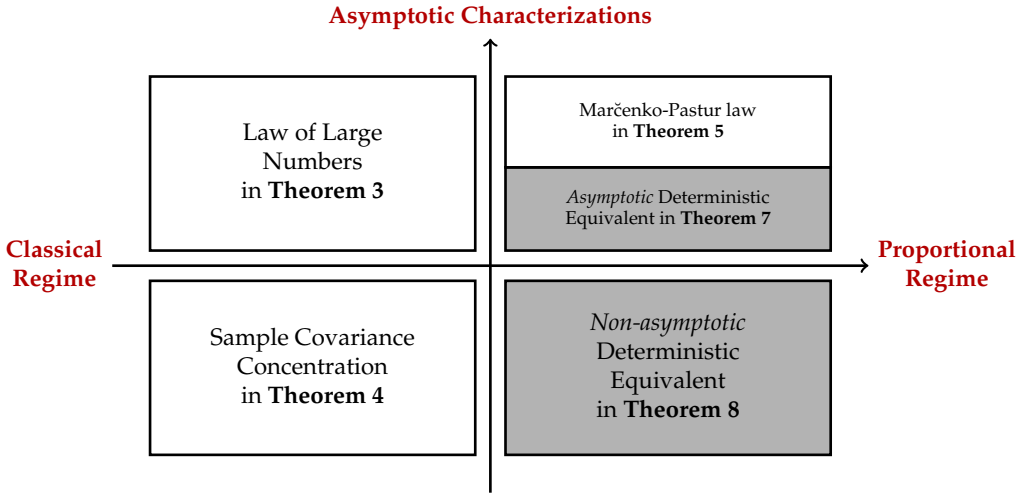


Figure: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$.

A modern RMT approach via deterministic equivalents for resolvent

- we have seen the **resolvent-based** approach as a **unified** analysis approach to **matrix spectral functionals**
- e.g., interested in the spectral behavior of a random matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$ from n samples, in the proportional $n \sim p \gg 1$ regime, more convenient to work with its resolvent $\mathbf{Q}_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I}_n)^{-1}$
- in particular, scalar observations $F\colon \mathbb{R}^{p\times p} \to \mathbb{R}$ of $\mathbf{X}$ and $\mathbf{Q}_{\mathbf{X}}(z)$ converge/concentrate, and there exists deterministic $\bar{\mathbf{Q}}(z)$ such that

$$F(\mathbf{Q}(z)) - F(\bar{\mathbf{Q}}(z)) \to 0, \tag{5}$$

  as $n, p \to \infty$.
- such $\bar{\mathbf{Q}}(z)$ is a **Deterministic Equivalent** of the random (resolvent) matrix $\mathbf{Q}$.
- so, our general recipe:

eigenspectral functional of large random matrix $\mathbf{X}$

$\downarrow$

more convenient to work with $\mathbf{Q}_{\mathbf{X}}(z)$

$\downarrow$

find its **Deterministic Equivalent**

Deterministic equivalent for RMT: intuition and a few words on the proof

What is actually happening for **Deterministic Equivalent**?

- while the random matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ remains random as the dimension p grows, in fact even "**more**" random due to the growing degrees of freedom;
- scalar observation $F(\mathbf{Q})$ of $\mathbf{Q}$ becomes "more concentrated" as $p \to \infty$;
- the random $F(\mathbf{Q})$, if concentrates, must concentrated around its expectation $\mathbb{E}[F(\mathbf{Q})]$;
- as $p \to \infty$, more randomness in $\mathbf{Q} \Rightarrow \mathrm{Var}[F(\mathbf{Q})] \to 0$ sufficiently fast (in p)
- if the functional $F\colon \mathbb{R}^{p\times p} \to \mathbb{R}$ is linear, then $\mathbb{E}[F(\mathbf{Q})] = F(\mathbb{E}[\mathbf{Q}])$.
- So, to propose a DE, suffices to evaluate $\mathbb{E}[\mathbf{Q}]$:
- **however**, $\mathbb{E}[\mathbf{Q}]$ may be hardly accessible, due to integration and nonlinear matrix inverse $\mathbf{Q}(z) = (\mathbf{X} - z\mathbf{I}_p)^{-1}$
- find a **simple** and **more accessible** deterministic $\bar{\mathbf{Q}}$ with $\bar{\mathbf{X}} \simeq \mathbb{E}[\mathbf{Q}]$ in some sense for p large, e.g., $\|\bar{\mathbf{Q}} - \mathbb{E}[\mathbf{Q}]\|_2 \to 0$ as $p \to \infty$; and
- show variance or higher-order moments of $F(\mathbf{Q})$ decay sufficiently fast as $p \to \infty$.

Deterministic Equivalent: definition

Definition (Deterministic Equivalent)

We say that $\bar{\mathbf{Q}} \in \mathbb{R}^{p\times p}$ is an $(\varepsilon_1, \varepsilon_2, \delta)$-*Deterministic Equivalent* for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ if, for a deterministic matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of unit norms (spectral and Euclidean, respectively), we have, with probability at least $1-\delta(p)$ that

$$\left|\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}})\right| \leq \varepsilon_1(p), \quad \left|\mathbf{a}^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b}\right| \leq \varepsilon_2(p), \tag{6}$$

for some non-negative functions $\varepsilon_1(p), \varepsilon_2(p)$ and $\delta(p)$ that decrease to zero as $p \to \infty$. Denote

$$\mathbf{Q} \overset{\varepsilon_1,\varepsilon_2,\delta}{\longleftrightarrow} \bar{\mathbf{Q}}, \text{ or simply } \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}. \tag{7}$$

An asymptotic Deterministic Equivalent for resolvent

Theorem (An asymptotic Deterministic Equivalent for resolvent, [CL22, Theorem 2.4])

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ *be a random matrix having i.i.d. sub-gaussian entries of zero mean and unit variance, and denote* $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ *the resolvent of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ *for* $z \in \mathbb{C}$ *not an eigenvalue of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$*. Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$*, the deterministic matrix* $\bar{\mathbf{Q}}(z)$ *is a Deterministic Equivalent of the random resolvent matrix* $\mathbf{Q}(z)$ *with*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{8}$$

with $m(z)$ *the unique valid Stieltjes transform as solution to*

$$czm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{9}$$

- The equation of $m(z)$ is quadratic and has two solutions defined via the complex square root
- **only one** satisfies the relation $\Im[z] \cdot \Im[m(z)] > 0$ as a "valid" Stieltjes transform
- this leads to the Marčenko-Pastur law

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x - E_-)^+ (E_+ - x)^+}\, dx,} \tag{10}$$

for $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$.

[2] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022

A non-asymptotic Deterministic Equivalent for resolvent

Theorem (A non-asymptotic Deterministic Equivalent for resolvent)

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ be a random matrix having i.i.d. sub-gaussian entries with zero mean and unit variance, and denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $z < 0$. Then, there exists universal constants $C_1, C_2 > 0$ depending only on the sub-gaussian norm of the entries of $\mathbf{X}$ and $|z|$, such that for any $\varepsilon \in (0,1)$, if $n \geq (C_1+\varepsilon)p$, one has

$$\|\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)\|_2 \leq \frac{C_2}{\varepsilon}\cdot n^{-\frac{1}{2}}, \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{11}$$

for $m(z)$ the unique positive solution to the Marčenko-Pastur equation $czm^2(z) - (1-c-z)m(z) + 1 = 0, c = p/n$.

- this is a deterministic characterization of the **expected resolvent**
- to get DE, it remains to show **concentration** results for trace and bilinear forms: more or less standard

Proof via leave-one-out and self-consistent equation

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the ith column of $\mathbf{X} \in \mathbb{R}^{p\times n}$ (so that $\mathbf{x}_i$ has i.i.d. sub-gaussian entries of zero mean and unit variance), and let $\mathbf{X}_{-i} \in \mathbb{R}^{p\times(n-1)}$ denote the random matrix $\mathbf{X}$ *without* its ith column $\mathbf{x}_i$. Define similarly $\mathbf{Q}_{-i}(z) = \left(\frac{1}{n}\mathbf{X}_{-i}\mathbf{X}_{-i}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}$ so that

$$\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}_{-i}\mathbf{X}_{-i}^\mathsf{T} + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - z\mathbf{I}_p\right)^{-1} = \left(\mathbf{Q}_{-i}^{-1}(z) + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)^{-1}. \tag{12}$$

First note that by definition,

$$\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \left(\frac{1}{1+cm(z)} - z\right)^{-1}\mathbf{I}_p, \tag{13}$$

for $c = p/n$, so that for $z < 0$,

$$\frac{1}{1+cm(z)}\|\bar{\mathbf{Q}}\|_2 \le 1. \tag{14}$$

Similarly, one has

$$\|\mathbf{Q}(z)\|_2 \le \frac{1}{|z|}, \quad \left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right\|_2 \le 1, \quad \left\|\mathbf{Q}(z)\frac{1}{\sqrt{n}}\mathbf{X}\right\|_2 = \sqrt{\left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}(z)\right\|_2} \le \frac{1}{\sqrt{|z|}}. \tag{15}$$

A few useful lemmas

Lemma (Resolvent identity)

For invertible matrices $\mathbf{A}$ *and* $\mathbf{B}$*, we have* $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$.

Lemma (Woodbury)

For $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p\times n}$*, such that both* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T}$ *are invertible, we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{V}^\mathsf{T}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\mathsf{T}\mathbf{A}^{-1}.$$

In particular, for $n = 1$*, i.e.,* $\mathbf{U}\mathbf{V}^\mathsf{T} = \mathbf{u}\mathbf{v}^\mathsf{T}$ *for* $\mathbf{U} = \mathbf{u} \in \mathbb{R}^p$ *and* $\mathbf{V} = \mathbf{v} \in \mathbb{R}^p$*, the above identity specializes to the following Sherman–Morrison formula,*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\mathsf{T}\mathbf{A}^{-1}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}, \quad \text{and } (\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}.$$

And the matrix $\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T} \in \mathbb{R}^{p\times p}$ *is invertible if and only if* $1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u} \neq 0$.

A few useful lemmas

Letting $\mathbf{A} = \mathbf{M} - z\mathbf{I}_p$, $z \in \mathbb{C}$, and $\mathbf{v} = \tau\mathbf{u}$ for $\tau \in \mathbb{R}$ in Woodbury identity leads to the following rank-one perturbation lemma for the resolvent of $\mathbf{M}$.

Lemma ([SB95, Lemma 2.6])

For $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p\times p}$ symmetric and nonnegative definite, $\mathbf{u} \in \mathbb{R}^p$, $\tau > 0$ and $z < 0$,

$$\left|\operatorname{tr}\mathbf{A}(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T} - z\mathbf{I}_p)^{-1} - \operatorname{tr}\mathbf{A}(\mathbf{M} - z\mathbf{I}_p)^{-1}\right| \leq \frac{\|\mathbf{A}\|_2}{|z|}.$$

Proof

It follows from the resolvent identity that

$$\begin{aligned}
\mathbb{E}[\mathbf{Q}-\bar{\mathbf{Q}}] &= \mathbb{E}\left[\mathbf{Q}\left(\frac{\mathbf{I}_p}{1+cm(z)}-\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}\right)\right]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}}-\frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^{\mathsf{T}}]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}}-\sum_{i=1}^{n}\frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}]\bar{\mathbf{Q}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}}-\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}}{1+\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i}\right]\bar{\mathbf{Q}}, \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}}-\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]\bar{\mathbf{Q}}}{1+cm(z)}+\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}d_i\right]\bar{\mathbf{Q}}}{1+cm(z)} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+cm(z)}\bar{\mathbf{Q}}-\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]\bar{\mathbf{Q}}}{1+cm(z)}+\frac{\mathbb{E}\left[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]\bar{\mathbf{Q}}}{1+cm(z)}
\end{aligned}$$

with $\boxed{d_i=\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i/n-cm(z)}$, so that $\mathbb{E}[\mathbf{Q}-\bar{\mathbf{Q}}]=(\mathbb{E}[\mathbf{Q}-\mathbf{Q}_{-i}])\,\frac{\bar{\mathbf{Q}}}{1+cm(z)}+\frac{\mathbb{E}\left[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right]\bar{\mathbf{Q}}}{1+cm(z)}$.

Let

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2, \quad T_2 = \left\|\mathbb{E}\left[d_i \mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\right\|_2, \tag{16}$$

we then have $\|\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}]\| \leq T_1 + T_2$.
For the first term T_1, it follows from Sherman–Morrison that

$$0 \preceq \mathbb{E}[\mathbf{Q}_{-i} - \mathbf{Q}] = \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right] \preceq \frac{1}{n}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}] = \frac{1}{n}\mathbb{E}\left[\mathbf{Q}_{-i}^2\right] \tag{17}$$

so

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2 = O(n^{-1}). \tag{18}$$

For T_2,

$$\begin{aligned}
T_2 &= \left\|\mathbb{E}\left[d_i \mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\right\|_2 \\
&= \sup_{\|\mathbf{u}\|=1,\|\mathbf{v}\|=1} \mathbb{E}\left[d_i \mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{v}\right] \\
&\leq \sqrt{\mathbb{E}[d_i^2]} \cdot \sup_{\|\mathbf{u}\|=1,\|\mathbf{v}\|=1} \sqrt{\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{v})^2]} \\
&\leq \underbrace{\sqrt{\mathbb{E}[d_i^2]}}_{T_{2,1}} \cdot \underbrace{\sup_{\|\mathbf{u}\|=1} \sqrt[4]{\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i)^4]}}_{T_{2,2}} \cdot \underbrace{\sup_{\|\mathbf{v}\|=1} \sqrt[4]{\mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{v})^4]}}_{T_{2,3}}.
\end{aligned}$$

For the term $T_{2,2}$. Note that

$$\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{x}_i)^4] = \mathbb{E}\left[\frac{(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4}{(1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4}\right] \leq \mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4] = \mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^2],$$

with

$$\|\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\|_2 = \mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u} \leq |z|^{-2}, \tag{19}$$

for $\|\mathbf{u}\| = 1$.
By Hanson–Wright inequality (concentration of quadratic form), there exists $C, C' > 0$ such that

$$\begin{aligned}\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4] = \mathbb{E}\left[\mathbb{E}[(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)^4|\mathbf{Q}_{-i}]\right] &\leq \mathbb{E}_{\mathbf{Q}_{-i}}\left[\int_0^\infty 2t\cdot\mathbb{P}\left(\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i \geq t\right)dt\right]\\ &\leq 2C'\cdot\mathbb{E}_{\mathbf{Q}_{-i}}\left[\int_0^\infty t\exp\left(-Ct/(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u})\right)dt\right]\\ &= 2C'\mathbb{E}\left[\frac{(\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u})^2}{C^2}\right] \leq (Cz^2)^{-2}.\end{aligned}$$

This allows us to conclude that $T_{2,2} = O(1)$, and analogously that $T_{2,3} = O(1)$.
We thus have

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq T_1 + T_2 \leq T_1 + T_{2,1}\cdot T_{2,2}\cdot T_{2,3} \leq C_1 n^{-1} + C_2\sqrt{\mathbb{E}[d_i^2]}, \tag{20}$$

for some universal constants C_1, C_2 and recall $\boxed{d_i \equiv \mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i/n - cm(z)}$.

Now, note that

$$
\begin{aligned}
d_i^2 &= \left(\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i - cm(z)\right)^2 \\
&= \left(\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i - \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] + \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2 \\
&\le 2\left(\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i - \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}]\right)^2 + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2 \\
&= 2\left(\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i - \frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i} + \frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i} - \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}]\right)^2 + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2,
\end{aligned}
$$

so that

$$
\frac{1}{2}\mathbb{E}[d_i^2] \le \underbrace{\mathbb{E}\left(\frac{1}{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{x}_i - \frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}\right)^2}_{D_1} + \underbrace{\mathbb{E}\left(\frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i} - \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}]\right)^2}_{D_2} + \left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2.
$$

- $D_1 \le Cn^{-2}$ by the same line of arguments as the term $T_{2,2}$
- D_2 that characterizes the **concentration property** of the resolvent trace $\operatorname{tr}\mathbf{Q}_{-i}$, using a martingale difference argument via **Burkholder inequality**.

Lemma

Under the notations and settings above, we have

$$\mathbb{E}\left[\left(\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\right)^2\right] \le Cn^{-1} \text{ and } \mathbb{E}\left[\left(\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\right)^4\right] \le Cn^{-2}, \tag{21}$$

for any $\mathbf{A} \in \mathbb{R}^{p\times p}$ of unit norm and some constant $C > 0$, and thus in particular for $\mathbf{A} = \mathbf{I}_p$.

Thus,

$$\mathbb{E}[d_i^2] \le 2(D_1 + D_2) + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2 \le Cn^{-1} + 2\left(\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}_{-i}] - cm(z)\right)^2, \tag{22}$$

for some universal constant $C > 0$. Putting together and by the trace rank-one update result,

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \le C_1 n^{-\frac{1}{2}} + C_2\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right|. \tag{23}$$

Finishing the proof

We "close the loop" by noting that by definition $\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}} = \frac{p}{n}m(z) = cm(z)$, so that

$$\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right| \leq \frac{p}{n}\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{p}{n}\left(C_1 n^{-\frac{1}{2}} + C_2\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right|\right), \tag{24}$$

and therefore for any $\epsilon > 0$ and $n > (C_2 + \varepsilon)p$, one has

$$\left|\frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}] - cm(z)\right| \leq \frac{C_1}{\varepsilon}\cdot n^{-\frac{1}{2}}, \tag{25}$$

and thus

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{C}{\varepsilon}\cdot n^{-\frac{1}{2}}, \tag{26}$$

for some universal constant $C > 0$. This concludes the proof.

Remark: extension to $z = 0$

- assume above $z < 0$ so that the bound on the random resolvent $\|\mathbf{Q}_{\hat{\mathbf{C}}}(z)\|_2 \leq 1/|z|$
- this, however, does not exploit the information in the random sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p\times n}$ on, e.g., how it concentrates around its population counterpart $\mathbf{C} = \mathbb{E}[\hat{\mathbf{C}}]$
- to extend the result above to, say, an inverse SCM of the type $\mathbf{Q}(z=0) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}$ with $z = 0$, first needs to ensure the inverse is **well-defined** for sub-gaussian $\mathbf{X}$ and for a specific choice of p, n
- can be obtained, e.g., per concentration of SCM $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ around its expectation.
- it follows from standard SCM concentration (Theorem 4) that there exists universal constant $C > 0$ such that for $n \geq C(p + \ln(1/\delta))$, one has, with probability at least $1 - \delta$, $\delta \in (0, 1/2]$ that

$$\left\| \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \mathbf{I}_p \right\|_2 \leq \frac{\mathbf{I}_p}{2}, \tag{27}$$

  and therefore $\|\mathbf{Q}(z)\|_2 \leq \frac{1}{1/2 - z} \leq 2$ for any $z \leq 0$
- allows for a control of the spectral norm $\|\mathbf{Q}(z)\|_2 \leq 2$ independent of $z \leq 0$ and holds with probability at least $1 - \delta$
- do everything else **conditioned on this high-probability event**, to get a bound on the **conditional expectation** $\mathbb{E}[\mathbf{Q} \,|\mathcal{E}]$, with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$

Remark: as extensions to results in the classical regime

(i) In the "easy" **classical regime**, with $n \gg p$ (and thus $p/n \to c = 0$), one has that $\hat{\mathbf{C}} \equiv \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \to \mathbb{E}[\hat{\mathbf{C}}] = \mathbf{I}_p$ as $n \to \infty$, so that

$$(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1} \simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1} = (1-z)^{-1}\mathbf{I}_p = \bar{\mathbf{Q}}(z). \tag{28}$$

(ii) In the "harder" and more general **proportional regime**, for $n \sim p$ with $p/n \to c \in (0, \infty)$, one has instead

$$\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)] \equiv \mathbb{E}[(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1}] \not\simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1}. \tag{29}$$

In this case, a Deterministic Equivalent $\bar{\mathbf{Q}}(z)$ can be very different from $(\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1}$.

- this is not surprising, consider the scalar case where $\mathbb{E}[1/x] \neq 1/\mathbb{E}[x]$ in general, unless $x \simeq C$ for some constant C

Remark: Deterministic Equivalents for Gaussian inverse SCM

- consider the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z}$ and positive definite $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\mathbf{Z} \in \mathbb{R}^{p\times n}$ having i.i.d. standard Gaussian entries
- the inverse $\hat{\mathbf{C}}^{-1}$ is known to follow the inverse-Wishart distribution [MKB79] with p degrees of freedom and scale matrix $\mathbf{C}^{-1}$, such that

$$\mathbb{E}[\hat{\mathbf{C}}^{-1}] = \frac{n}{n-p-1}\mathbf{C}^{-1} \tag{30}$$

  for $n \geq p+2$.
- On the other hand, it follows from our non-asymptotic result above by taking $z = 0$ that

$$\mathbb{E}[\mathbf{Q}(z)] \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \frac{n}{n-p}\mathbf{I}_p \tag{31}$$

  with $m(z) = \frac{1}{1-c} = \frac{n}{n-p}$.
- **note**: Deterministic Equivalents **are not unique**: could replace the "-1" in denominator by any constant $C' \ll n, p$ to propose another equally correct Deterministic Equivalent.

[3] Kanti Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. 1st ed. Probability and Mathematical Statistics. Academic Press, Dec. 1979

Some thoughts on the "leave-one-out" proof

- **in essence**: propose $\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)]$ (in spectral norm sense), but simple to evaluate (via a quadratic equation)
- leave-one-out analysis of large-scale system: $\mathbf{Q}(z) \simeq \mathbf{Q}_{-i}(z)$ for n, p large.
- low complexity analysis of large random system: joint behavior of p eigenvalues $\overset{\text{RMT}}{\to}$ a single deterministic (quadratic) equation
- **Side Remark**: another (as well) systematic and convenient RMT proof approach: **Gaussian method**, as the combination of

(1) Stein's lemma (Gaussian integration by parts)

(2) Nash–Poincaré inequality (a bound on the variance of smooth scalar observation of multivariate Gaussian random vector)

(3) interpolation from Gaussian to non-Gaussian, see [CL22, Section 2.2.2] for details.

Proof of MP law with Gaussian method

Theorem (Stein's Lemma)

Let $x \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \to \mathbb{R}$ a continuously differentiable function having at most polynomial growth and such that $\mathbb{E}[f'(x)] < \infty$. Then,

$$\mathbb{E}[xf(x)] = \mathbb{E}[f'(x)]. \tag{32}$$

In particular, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $f : \mathbb{R}^p \to \mathbb{R}$ a continuously differentiable function with derivatives having at most polynomial growth with respect to p,

$$\mathbb{E}[[\mathbf{x}]_i f(\mathbf{x})] = \sum_{j=1}^{p} [\mathbf{C}]_{ij} \mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_j}\right], \tag{33}$$

where $\partial/\partial[\mathbf{x}]_i$ indicates differentiation with respect to the i-th entry of $\mathbf{x}$; or, in vector form $\mathbb{E}[\mathbf{x}f(\mathbf{x})] = \mathbf{C}\mathbb{E}[\nabla f(\mathbf{x})]$, with $\nabla f(\mathbf{x})$ the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$.

First observe that $\mathbf{Q} = \frac{1}{z}\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q} - \frac{1}{z}\mathbf{I}_p$, so that $\mathbb{E}[\mathbf{Q}_{ij}] = \frac{1}{zn}\sum_{k=1}^n \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] - \frac{1}{z}\delta_{ij}$, in which $\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \mathbb{E}[xf(x)]$ for $x = \mathbf{X}_{ik}$ and $f(x) = [\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}$.
Therefore, from Stein's lemma and the fact that $\partial\mathbf{Q} = -\frac{1}{n}\mathbf{Q}\partial(\mathbf{X}\mathbf{X}^\mathsf{T})\mathbf{Q}$,[1]

$$\begin{aligned}\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] &= \mathbb{E}\left[\frac{\partial[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}}{\partial\mathbf{X}_{ik}}\right] = \mathbb{E}[\mathbf{E}_{ik}^\mathsf{T}\mathbf{Q}]_{kj} - \mathbb{E}\left[\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{Q}(\mathbf{E}_{ik}\mathbf{X}^\mathsf{T} + \mathbf{X}\mathbf{E}_{ik}^\mathsf{T})\mathbf{Q}\right]_{kj} \\ &= \mathbb{E}[\mathbf{Q}_{ij}] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ki}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}\right] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{X}]_{kk}\mathbf{Q}_{ij}\right]\end{aligned}$$

for $\mathbf{E}_{ij}$ the indicator matrix with entry $[\mathbf{E}_{ij}]_{lm} = \delta_{il}\delta_{jm}$, so that, summing over k,

$$\frac{1}{z}\frac{1}{n}\sum_{k=1}^n \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}_{ij}] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}_{ij}\,\mathrm{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T})] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ij}.$$

[1] This is the matrix version of $d(1/x) = -dx/x^2$.

Proof of MP law with Gaussian method

We have

$$\frac{1}{z}\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\top\mathbf{Q}]_{kj}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}_{ij}] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}_{ij}\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\top)] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\top\mathbf{Q}]_{ij}.$$

The term in the second line has vanishing operator norm (of order $O(n^{-1})$) as $n, p \to \infty$. Also, $\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\top) = np + zn\operatorname{tr}\mathbf{Q}$. As a result, matrix-wise, we obtain

$$\mathbb{E}[\mathbf{Q}] + \frac{1}{z}\mathbf{I}_p = \mathbb{E}[\mathbf{X}_{\cdot k}[\mathbf{X}^\top\mathbf{Q}]_{k\cdot}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}] - \frac{1}{z}\frac{1}{n}\mathbb{E}[\mathbf{Q}(p + z\operatorname{tr}\mathbf{Q})] + o_{\|\cdot\|}(1),$$

where $\mathbf{X}_{\cdot k}$ and $\mathbf{X}_{k\cdot}$ is the k-th column and row of $\mathbf{X}$, respectively.
As the random $\frac{1}{p}\operatorname{tr}\mathbf{Q} \to m(z)$ as $n, p \to \infty$, "take it out of the expectation" in the limit and

$$\mathbb{E}[\mathbf{Q}](1 - p/n - z - p/n \cdot zm(z)) = \mathbf{I}_p + o_{\|\cdot\|}(1),$$

which, taking the trace to identify $m(z)$, concludes the proof.

Nash–Poincaré inequality and Interpolation trick

Theorem (Nash–Poincaré inequality)

For $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $f\colon \mathbb{R}^p \to \mathbb{R}$ continuously differentiable with derivatives having at most polynomial growth with respect to p,

$$\mathrm{Var}[f(\mathbf{x})] \leq \sum_{i,j=1}^{p} [\mathbf{C}]_{ij} \mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_i} \frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_j}\right] = \mathbb{E}\left[(\nabla f(\mathbf{x}))^{\mathsf{T}} \mathbf{C} \nabla f(\mathbf{x})\right],$$

where we denote $\nabla f(\mathbf{x})$ the gradient of $f(\mathbf{x})$ with respect to $\mathbf{x}$.

Theorem (Interpolation trick)

For $x \in \mathbb{R}$ a random variable with zero mean and unit variance, $y \sim \mathcal{N}(0,1)$, and f a $(k+2)$-times differentiable function with bounded derivatives,

$$\mathbb{E}[f(x)] - \mathbb{E}[f(y)] = \sum_{\ell=2}^{k} \frac{\kappa_{\ell+1}}{2\ell!} \int_0^1 \mathbb{E}[f^{(\ell+1)} x(t)] t^{(\ell-1)/2} dt + \epsilon_k,$$

where κ_ℓ is the ℓ^{th} cumulant of x, $x(t) = \sqrt{t}x + (1-\sqrt{t})y$, and $|\epsilon_k| \leq C_k \mathbb{E}[|x|^{k+2}] \cdot \sup_t |f^{(k+2)}(t)|$ for some constant C_k only dependent on k.

Take-away of this section

- p-by-p SCM $\hat{\mathbf{C}}$ from n samples have different behavior in the **classical** ($n \gg p$) versus **proportional** ($n \sim p$) regime
- four ways to characterize SCM, asymptotic and non-asymptotic fashion
- "**old school**" results: (1) LLN and (2) matrix concentration in the classical regime, and (3) asymptotic Marčenko-Pastur law on SCM eigenvalues in the proportional regime
- **modern** approach of **deterministic equivalent for SCM resolvent**, both (4) asymptotic and (5) non-asymptotic
- proof via "leave-one-out" and self-consistent equation
- alternative proof via Gaussian method

Theorem (Wigner semicircle law)

Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be symmetric and such that the $\mathbf{X}_{ij} \in \mathbb{R}$, $j \geq i$, are independent zero mean and unit variance random variables. Then, for $\mathbf{Q}(z) = (\mathbf{X}/\sqrt{n} - z\mathbf{I}_n)^{-1}$, as $n \to \infty$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_n, \tag{34}$$

with $m(z)$ the unique Stieltjes transform solution to

$$m^2(z) + zm(z) + 1 = 0. \tag{35}$$

The function $m(z)$ is the Stieltjes transform of the probability measure

$$\mu(dx) = \frac{1}{2\pi}\sqrt{(4-x^2)^+}\,dx, \tag{36}$$

known as the Wigner semicircle law.

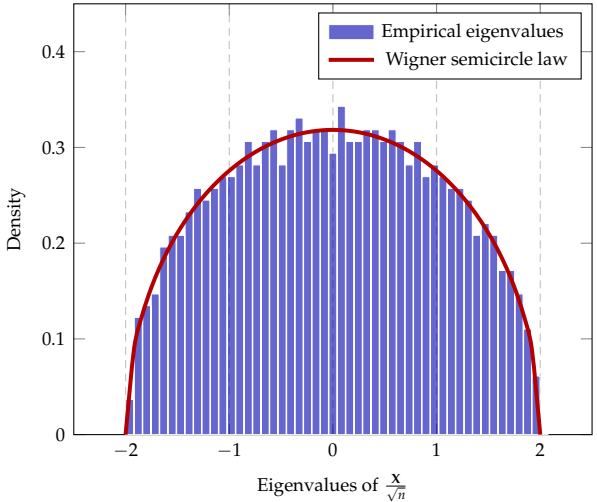


Figure: Histogram of the eigenvalues of $\mathbf{X}/\sqrt{n}$ versus Wigner semicircle law, for standard Gaussian $\mathbf{X}$ and $n = 1\,000$.

Theorem (General sample covariance matrix)

Let $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$ *with nonnegative definite* $\mathbf{C} \in \mathbb{R}^{p\times p}$, $\mathbf{Z} \in \mathbb{R}^{p\times n}$ *having independent zero mean and unit variance entries. Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$, *for* $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} - z\mathbf{I}_p)^{-1}$ *and* $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{X} - z\mathbf{I}_n)^{-1}$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}, \quad \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) = \tilde{m}_p(z)\mathbf{I}_n,$$

with $\tilde{m}_p(z)$ *unique solution to*

$$\tilde{m}_p(z) = \left(-z + \frac{1}{n}\operatorname{tr}\mathbf{C}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}\right)^{-1}. \tag{37}$$

Moreover, if the empirical spectral measure of $\mathbf{C}$ *converges* $\mu_{\mathbf{C}} \to \nu$ *as* $p \to \infty$, *then* $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}} \to \mu$, $\mu_{\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{X}} \to \tilde{\mu}$ *where* $\mu, \tilde{\mu}$ *admitting Stieltjes transforms* $m(z)$ *and* $\tilde{m}(z)$ *such that*

$$m(z) = \frac{1}{c}\tilde{m}(z) + \frac{1-c}{cz}, \quad \tilde{m}(z) = \left(-z + c\int \frac{t\nu(dt)}{1+\tilde{m}(z)t}\right)^{-1}. \tag{38}$$

A few remarks on the generalized MP law

- different from the explicit MP law, the generalized MP is in general implicit
- we have explicitness in essence due to with $\mathbf{C} = \mathbf{I}_p$, the implicit equation boils down to a quadratic equation that has explicit solution
- if $\mathbf{C}$ has discrete eigenvalues, e.g., $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_5)$, then becomes a (possibly higher-order) polynomial equation, which may admit explicit solution (up to fourth order) using radicals
- the uniqueness of (Stieltjes transform) solution is ensured within a certain region on the complex plane, there may exist solutions $\tilde{m}(z)$ with imaginary parts of wrong sign
- **numerical evaluation of** $\tilde{m}(z)$: note that the equation

$$\tilde{m}_p(z) = \left(-z + \frac{1}{n} \operatorname{tr} \mathbf{C} \left(\mathbf{I}_p + \tilde{m}_p(z) \mathbf{C} \right)^{-1} \right)^{-1} \tag{39}$$

naturally defines a fixed-point equation.

Matlab code

```
clear i % make sure i stands for the imaginary unit
y = 1e-5;
zs = edges_mu+y*1i;
mu = zeros(length(zs),1);

tilde_m=0;
for j=1:length(zs)
    z = zs(j);

    tilde_m_tmp=-1;
    while abs(tilde_m-tilde_m_tmp)>1e-6
        tilde_m_tmp=tilde_m;
        tilde_m = 1/( -z + 1/n*sum(eigs_C./(1+tilde_m*eigs_C)) );
    end

    m = tilde_m/c+(1-c)/(c*z);
    mu(j)=imag(m)/pi;
end
```

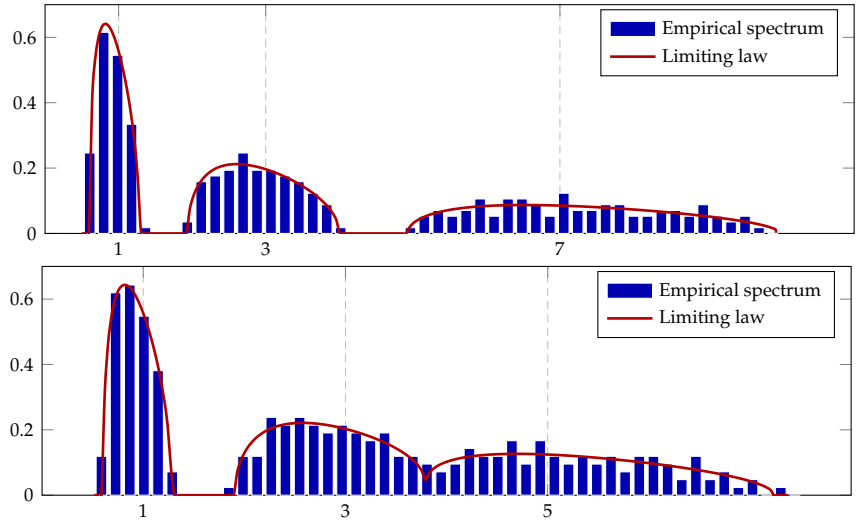


Figure: Histogram of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$, $[\mathbf{Z}]_{ij} \sim \mathcal{N}(0,1)$, $n = 3\,000$; for $p = 300$ and $\mathbf{C}$ having spectral measure $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_7)$ (**top**) and $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_5)$ (**bottle**).

Further comments on generalized SCM

- we know a lot more for the generalized SCM model: precise characterization of the support of its (limiting) eigenspectrum
- applications in **statistical inference**: given $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ SCM of the population covariance $\mathbf{C}$, infer eigenspectral functions of $\mathbf{C}$ using those of $\hat{\mathbf{C}}$ and wisely-chosen contour integration, etc.

Example: estimation of population eigenvalues of large multiplicity

Consider the following SCM inference,

$$\nu_{\mathbf{C}} = \frac{1}{p}\sum_{i=1}^{K} p_i \delta_{\ell_i} \to \sum_{i=1}^{K} c_i \delta_{\ell_i}$$

for $\ell_1 > \ldots > \ell_K > 0$, K fixed/small with respect to n, p, and $p_i/p \to c_i > 0$ as $p \to \infty$, i.e., each eigenvalue has a large multiplicity of order $O(p)$.

- **native** estimator: $\hat{\ell}_a = \frac{1}{p_a}\sum_{i=p_1+\ldots+p_{a-1}+1}^{p_1+\ldots+p_a} \lambda_i$
- **RMT-improved** estimator: $\hat{\ell}_a = \frac{n}{p_a}\sum_{i=p_1+\ldots+p_{a-1}+1}^{p_1+\ldots+p_a} (\lambda_i - \eta_i)$, with λ_i eigenvalues of $\hat{\mathbf{C}}$ and η_i eigenvalues of $\mathbf{\Lambda} - \frac{1}{n}\sqrt{\boldsymbol{\lambda}}\sqrt{\boldsymbol{\lambda}}^\mathsf{T}$, $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_i\}_{i=1}^p$ and $\sqrt{\boldsymbol{\lambda}} \in \mathbb{R}^p$ the vector of $\sqrt{\lambda_i}$s.

- see [CL22, Sections 2.3 and 2.4] for detailed derivations and discussions

Numerical results

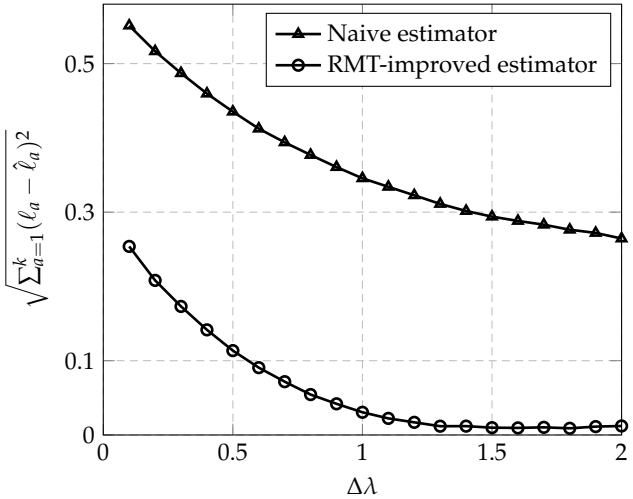


Figure: Eigenvalue estimation errors with naive and RMT-improved approach, as a function of $\Delta\lambda$, for $\ell_1 = 1$, $\ell_2 = 1 + \Delta\lambda$, $p = 256$ and $n = 1\,024$. Results averaged over 30 runs.

Separable covariance model: motivation

- data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ arise from a time series, each data vector is weighted by a coefficient
- SCM can be generalized to the so-called **bi-correlated** (or **separable covariance**) model

$$\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} = \frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^\mathsf{T}\mathbf{C}^{\frac{1}{2}} \tag{40}$$

  for $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $\tilde{\mathbf{C}} \in \mathbb{R}^{n\times n}$ two nonnegative definite matrices and $[\mathbf{Z}]_{ij}$ i.i.d. random variables with zero mean and unit variance.
- in particular, for $\mathbf{Z}$ Gaussian and $\tilde{\mathbf{C}}^{\frac{1}{2}}$ Toeplitz (i.e., such that $[\tilde{\mathbf{C}}^{\frac{1}{2}}]_{ij} = \alpha_{|i-j|}$ for some sequence $\alpha_0, \ldots, \alpha_{n-1}$), the columns of $\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}}$ model a first order auto-regressive process

Separable covariance model

Theorem (Bi-correlated model, separable covariance model, [PS09])

Let $\mathbf{Z} \in \mathbb{R}^{p\times n}$ be a random matrix with i.i.d. zero mean, unit variance and light tail entries, and $\mathbf{C} \in \mathbb{R}^{p\times p}$, $\tilde{\mathbf{C}} \in \mathbb{R}^{n\times n}$ be symmetric nonnegative definite matrices with bounded operator norm. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, letting $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^{\mathsf{T}}\mathbf{C}^{\frac{1}{2}} - z\mathbf{I}_p)^{-1}$ and $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\tilde{\mathbf{C}}^{\frac{1}{2}}\mathbf{Z}^{\mathsf{T}}\mathbf{C}\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}} - z\mathbf{I}_n)^{-1}$, we have

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \tilde{\delta}_p(z)\mathbf{C}\right)^{-1}, \quad \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) = -\frac{1}{z}\left(\mathbf{I}_n + \delta_p(z)\tilde{\mathbf{C}}\right)^{-1}$$

with $(z, \delta_p(z)), (z, \tilde{\delta}_p(z)) \in \mathcal{Z}(\mathbb{C} \setminus \mathbb{R}^+)$ unique solutions to

$$\delta_p(z) = \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}(z), \quad \tilde{\delta}_p(z) = \frac{1}{n}\operatorname{tr}\tilde{\mathbf{C}}\bar{\tilde{\mathbf{Q}}}(z).$$

In particular, if $\mu_{\mathbf{C}} \to \nu$ and $\mu_{\tilde{\mathbf{C}}} \to \tilde{\nu}$, then $\mu_{\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^{\mathsf{T}}\mathbf{C}^{\frac{1}{2}}} \xrightarrow{a.s.} \mu, \mu_{\frac{1}{n}\tilde{\mathbf{C}}^{\frac{1}{2}}\mathbf{Z}^{\mathsf{T}}\mathbf{C}\mathbf{Z}\tilde{\mathbf{C}}^{\frac{1}{2}}} \xrightarrow{a.s.} \tilde{\mu},$, where $\mu, \tilde{\mu}$ are defined by their Stieltjes transforms $m(z)$ and $\tilde{m}(z)$ given by

$$m(z) = -\frac{1}{z}\int \frac{\nu(dt)}{1+\tilde{\delta}(z)t}, \quad \tilde{m}(z) = -\frac{1}{z}\int \frac{\tilde{\nu}(dt)}{1+\delta(z)t}, \quad \delta(z) = -\frac{c}{z}\int \frac{t\nu(dt)}{1+\tilde{\delta}(z)t}, \quad \tilde{\delta}(z) = -\frac{1}{z}\int \frac{t\tilde{\nu}(dt)}{1+\delta(z)t}$$

[4] Debashis Paul and Jack W. Silverstein. "No eigenvalues outside the support of the limiting empirical spectral distribution of a separable

Take-away messages of this section

Asymptotic Deterministic Equivalent for resolvent results for

- symmetric $\mathbf{X}/\sqrt{n} \in \mathbb{R}^{n\times n}$: **Wigner semicircle law**, quadratic equation (again)
- **generalized SCM model** $\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^{\mathsf{T}}\mathbf{C}^{\frac{1}{2}}$: one self-consistent but integral equation
- application to **inference** of SCM eigenspectral functionals
- **bi-correlated model** or **separable covariance model** $\frac{1}{n}\mathbf{C}^{\frac{1}{2}}\mathbf{Z}\tilde{\mathbf{C}}\mathbf{Z}^{\mathsf{T}}\mathbf{C}^{\frac{1}{2}}$: two coupled self-consistent integral equations

Thank you! Q & A?